Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos

Text Summarisation based on Human Language Technologies and its Applications

PhD Thesis

Elena Lloret Pastor

Supervisor: Dr. Manuel Palomar Sanz

May 2011

This research work has been funded by the Spanish Government through the research program FPI (BES-2007-16268) associated to the project "TEXT-MESS: Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano" (TIN2006-1526-C06-01). Moreover, it has been also partially funded by projects "TEXT-MESS 2.0: Las Tecnologías del Lenguaje Humano ante los nuevos retos de la comunicación digital" (TIN2009-13391-C04), and "Desarrollo de técnicas inteligentes e interactivas de minería de textos" (PROMETEO/2009/199) from the Spanish and the Valencian Government, respectively.

"El resumen tiene dos funciones, una para quien lo hace y otra para quien lo lee. Pienso que hacerlo es mucho más importante que leerlo. El arte del resumen es importante y muy útil, y se aprende haciendo muchos resúmenes. Hacer resúmenes enseña a condensar las ideas. [...] Algo se perdía, naturalmente, pero el arte del resumen consiste también en eso, en saber qué se puede pasar por alto y en reconocer que algo que se dice en medio minuto no es lo mismo que se ha dicho en dos minutos, por lo cual es necesario decidir qué es lo verdaderamente importante, central."

Umberto Eco. Elogio del resumen. Quimera, n^o 51, pp. 13-15

Acknowledgements

Esta tesis es el fruto de un trabajo de investigación cuyos inicios se remontan al 3 de septiembre del 2007, día que entré a formar parte del Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información en mi universidad, la Universidad de Alicante. Si aquel día alguien me hubiera dicho que en estos cuatro años mi vida cambiaría, probablemente no lo hubiera creído del todo. Ciertamente, me equivocaba, puesto que este tiempo ha sido un sin parar de retos emocionantes y en ocasiones muy duros, que me han permitido crecer tanto profesional como personalmente... y que se han pasado volando. Echando la vista atrás, durante el desarrollo de la tesis he tenido la oportunidad de formarme como investigadora y realizar un montón de cursos de otra índole, complementando la formación adquirida. También he podido conocer a muchísimas personas de diferentes países y colaborar con otros grupos, durante mis estancias de investigación. Por último, a través de los congresos internacionales, he tenido la gran suerte de poder conocer otros países (Reino Unido, República Checa, Luxemburgo, Méjico, Estado Unidos, Argentina), y aprender un poco más de sus gentes y sus culturas. La combinación de todo ello me ha permitido poder "abrir la mente" (como me diría mi director de tesis) y ver las cosas desde una perspectiva más amplia.

Tengo que decir que no todo en el desarrollo de esta tesis ha sido un camino de rosas. Detrás de este trabajo, hay mucho esfuerzo, constancia y ganas de hacer las cosas bien. En más de una ocasión he pensado cosas como "quien me mandaría a mí meterme en esto", y alguna que otra vez, he tenido la tentación de tirar la toalla y volver a mi vida anterior. Aunque, sin duda alguna, prefiero quedarme con las cosas positivas que he podido vivir en estos cuatro años e intentar así cambiar mi pesimismo innato. Han sido muchas las personas que me han animado a continuar en este camino. Por esta razón, a todas ellas, les quiero agradecer todos los apoyos que me han transmitido. En especial, me gustaría agradecer:

- A Manuel Palomar, mi director de tesis, por haber confiado en mí y en mi trabajo desde el primer momento. Sus consejos, ideas, y opiniones, así como su preocupación en que yo estuviera a gusto en todo momento, han sido claves en el trabajo realizado.
- A Andrés Montoyo, Paloma Moreda, Patricio Martínez, Rafa Muñoz y al resto de los miembros del grupo de investigación que también siempre han estado ahí y me han ayudado en todo lo que me ha hecho falta.
- Al Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información, al Departamento de Lenguajes y Sistemas Informáticos y a la Universidad de Alicante, que me han permitido la realización de esta tesis doctoral.
- A todos y cada uno mis compañeros de laboratorio (los de antes y los de ahora), que me han "aguantado" todos los días y me han ayudado muchísimo; muchos de ellos, incluso, no se han librado de tener que evaluar resúmenes. Muy especialmente, me gustaría agradecer a José Manuel Gómez y a Ester Boldrini, por sus ánimos y apoyos constantes; y a Javi Fernández, por aportar su toque creativo a esta tesis[1].
- A Ahmet Aker, Laura Plaza, Hakan Ceylan, Rada Mihalcea y todas las personas que me han permitido colaborar con ellas en la parte científica, porque de todos ellos he aprendido algo y me ha servido, sin duda, para mejorar en muchos aspectos.
- A Constatin Orasan y Josef Steinberger, porque sus comentarios han servido para mejorar esta tesis.
- Al Dr. Ruslan Mitkov, el Dr. Horacio Saggion y la Dra. Mirella Lapata por haberme dejado pertenecer a sus grupos de investigación en la Universidad de Wolverhampton, la Universidad de Sheffield y la Universidad de Edimburgo, respectivamente, y poder así, mejorar la calidad científica de la investigación. Quisiera destacar la labor del Dr. Horacio Saggion y todo su apoyo y *feedback* que me ha seguido dando después. Tampoco me gustaría olvidarme de to-

[1] http://www.flickr.com/photos/17258892@N05/2588347668/

dos mis compañeros de las estancias: Laura Hasler, Iustina Ilisei, Georgiana Puscasu, Luz Rello, Natasha Ponomareva, Natali Konstantinova, Irina Temnikova, Silvia Pareti, Diego Frassinelli, Eva Hasler, Ioanis Konstas y Lidia Mora, que hicieron que mis estancias pasaran más rápido de lo normal; y de Carole y Brian Jackson, que me trataron como un miembro más de su familia cuando estuve en Edimburgo.

- A Andrea, Inma, María, Silvia y Teresa, por todos los cafés que tanta falta me hacían.
- A mis tíos, Alfonso y Rosi; mis primas, Celia y Mari Reyes; y a mi yaya, porque son una parte muy importante de mi familia.
- A mis padres y a mi hermano, que son los que me han soportado de verdad durante todo este tiempo. A mis padres, porque son un modelo a seguir; y a mi hermano, porque ha tenido mucha paciencia en más de una ocasión.
- Y por último, con muchísimo cariño, a Raúl, porque sabe sacarme una sonrisa en los malos momentos y porque él, sin duda alguna, se ha llevado la peor parte de todo esto y no ha rechistado en ningún momento. Espero que no me lo tengas muy en cuenta, que todavía nos queda mucho por vivir.

Abstract

In the current society, information plays a crucial role that brings competitive advantages to users, when it is managed correctly. However, due to the vast amount of available information, users cannot cope with it, and therefore new methods and approaches based on Human Language Technologies (HLT) are essential to process all the information in an effective and efficient manner. Text Summarisation (TS) is a research area in the context of HLT whose goal is to process, synthesise and present the information to users, avoiding the arduous task of having to read everything, as well as facilitating the process of guiding the user in what it is important in texts.

The research work carried out in this thesis focuses on TS, and more specifically in the task of generating summaries that are beneficial both for users, and HLT applications. Therefore, after analysing the main techniques and approaches in TS, as well as the existing evaluation methods, COMPENDIUM TS tool is proposed and developed. Our TS tool is based on a cognitive perspective (Van Dijk, 1980), (Van Dijk & Kintsch, 1983) that provides insights of how humans summarise, but in addition, it takes into account computational issues needed for its automation (Hovy, 2005).

COMPENDIUM is capable of generating different types of summaries. These are text summaries in English of a specific number of words or compression rate. Moreover, regarding the input, single- and multi-document summaries can be produced; as output, the summaries can be generated following an extractive or abstractive-oriented strategy. Finally, concerning their purpose, the summaries can be generic, query-focused or sentiment-based, and in all of the cases, their final aim is to be informative, thus providing information about the source document(s).

The proposed architecture for COMPENDIUM is divided in various stages, each of one focusing on a specific need. Moreover, a distinction between core and additional stages is made. The reason why this is done is to differentiate those stages aiming at producing generic extracts from those ones which enhance the capabilities of COMPENDIUM, specifically addressing the generation of other types of summaries, thus resulting in three more variants: query-focused; sentiment-based; and abstractive-oriented. In this way, the core stages are: i) *surface linguistic analysis*; ii) *redundancy detection*; iii) *topic identification*; iv) *relevance detection*; v) and *summary generation*, whereas the additional stages are: i) *query similarity*; ii) *subjective information detection*; and iii) *information compression and fusion*. Furthermore, novel methods and techniques are analysed within the aforementioned stages. In particular, textual entailment is suggested for detecting and removing redundant information, and the Code Quantity Principle (Givón, 1990), combined with the main topics identified by using term frequency, is employed for detecting relevant information in the documents. In addition, a word graph-based method is suggested for compressing and fusing information, thus producing abstractive-oriented summaries.

COMPENDIUM is evaluated both intrinsically and extrinsically, the latter through its integration into several HLT applications. On the one hand, concerning its intrinsic evaluation, different types of texts and domains are proposed: newswire, image captions, blogs and medical research papers. On the other hand, COMPENDIUM is evaluated extrinsically through its integration into three HLT applications (opinion mining, question answering and text classification). In particular, COMPENDIUM is integrated in opinion mining with the aim of generating better sentiment-based summaries, in contrast to take as a summary only those sentences that have higher opinion intensity values. In question answering, query-focused summaries are employed to find the answers of factual questions, instead of using the snippets retrieved by a search engine. Finally, for text classification (in the specific task of the rating inference), the use of generic, query-focused and sentiment-based summaries is compared with respect to the full document for predicting the correct rating associated to a review.

Therefore, the research conducted in this thesis shows the appropriateness of COMPENDIUM, either for being used on their own, as well as for improving the performance of HLT applications.

Resumen

La información juega un papel muy importante en la sociedad actual, puesto que si se procesa y maneja correctamente, proporciona grandes ventajas a los usuarios. Sin embargo, debido al crecimiento exponencial de la misma, los usuarios son incapaces de procesar toda esta información, y por tanto, las Tecnologías del Lenguaje Humano (TLH) son fundamentales para manejar dicha información de manera eficiente y efectiva, siendo de gran ayuda para los usuarios. La generación automática de resúmenes es un área de las TLH, cuyo objetivo es procesar, sintetizar y presentar al usuario la información de manera condensada, de tal manera que evita a los usuarios tener que leer multitud de documentos y extraer lo más importante de cada uno.

El trabajo de investigación que se ha desarrollado en esta tesis doctoral se centra en este área; en concreto, en la generación automática de resúmenes, demostrando que los resúmenes automáticos son beneficiosos tanto para los usuarios, como para otras aplicaciones de TLH. Después de realizar un análisis exhaustivo del estado de la cuestión tanto en enfoques para la generación de resúmenes como para su evaluación, se propone la herramienta de resúmenes COMPENDIUM. Esta herramienta sigue un enfoque cognitivo, que se basa en las teorías de (Van Dijk, 1980), (Van Dijk & Kintsch, 1983), que explican cómo generan resúmenes los humanos, pero también aporta una componente computacional (Hovy, 2005) que permite su automatización.

COMPENDIUM es capaz de generar distintos tipos de resúmenes de texto en inglés. La longitud de dichos resúmenes se determina en función de un número fijo de palabras o una tasa de compresión. Además, en lo que respecta a la entrada de la herramienta, se pueden generar resúmenes a partir de uno o de varios documentos (mono- o multi-documento, respectivamente). Como salida, los resúmenes siguen un paradigma extractivo (extractos) u orientado a abstractos. Final-

mente, en cuanto a su finalidad, éstos pueden ser resúmenes genéricos, orientados a un tópico, o resúmenes subjetivos, y en todos los casos, se pretende que puedan servir como sustituto del documento original, siendo informativos.

La arquitectura propuesta para COMPENDIUM se divide en dos tipos de etapas: las que forman el núcleo central de la herramienta, cuyo resultado son extractos genéricos y una serie de etapas adicionales, que sirven para generar tipos de resúmenes específicos: resúmenes orientados a un tópico, resúmenes subjetivos y resúmenes orientados a abstractos. Por un lado, las etapas que forman el núcleo de COMPENDIUM son: i) *análisis lingüístico*; ii) *detección de redundancia*; iii) *identificación del tópico*; iv) *detección de relevancia*; y v) *generación del resumen.* Por otro lado, las que etapas adicionales son: i) *similitud con la pregunta*; ii) *detección de información subjetiva*; y iii) *compresión y fusión de información.* Además, algunas de las etapas anteriormente citadas se basan en métodos y enfoques novedosos. En concreto, el uso del reconocimiento de la implicación textual como método para detectar y eliminar la redundancia de un documento, mientras que el principio de la cantidad de codificación se propone, junto con la frecuencia de las palabras, para identificar qué frases contienen la información más relevante. También se propone un método basado en grafos de palabras que permite combinar información extractiva y abstractiva, y que produce como resultado, resúmenes orientados a abstractos.

COMPENDIUM se ha evaluado de manera intrínseca y extrínseca. En lo que respecta a la evaluación intrínseca, se han usado distintos tipos de textos pertenecientes a diversos dominios: noticias periodísticas, descripciones de imágenes, blogs y artículos científicos del dominio médico. Para su evaluación extrínseca, COMPENDIUM se ha integrado en: minería de opiniones, búsqueda de respuestas y clasificación de textos. El objetivo de integrar COMPENDIUM en la primera de estas aplicaciones es mejorar la generación de resúmenes subjetivos con respecto a los enfoques que no tienen en cuenta técnicas de generación de resúmenes. Para la segunda aplicación, se han utilizado resúmenes orientados a un tópico, en vez de los *snippets* que devuelven los motores de búsqueda, para que un sistema de búsqueda de respuestas encuente de manera más eficaz las respuestas a preguntas factuales. Finalmente, en en la tercera, COMPENDIUM se ha usado para generar resúmenes que ayuden a predecir la puntuación asociada a un reseña, en lugar de procesar la reseña completa.

Por lo tanto, de todo ello se demuestra que los resúmenes automáticos generados con COMPENDIUM son adecuados para que se usen de manera individual o para que se integren en otra aplicaciones de TLH, con la finalidad de mejorar su rendimiento.

Contents

List of Tables

List of Figures

1. Introduction

Human Language Technologies (HLT) cover a broad range of activities with the final goal of enabling people to communicate with machines by using natural communication skills (Cole, 1997). All these activities face a common challenge: they have to deal with natural language, which is not a trivial issue. The difficulty resides in the specific nature of the language itself, as well as in the context it is developed. On the one hand, each language has its own specific structures and phenomena, which have to be taken into consideration. For instance, the ambiguity and anaphora are two well-known phenomena, that have been extensively studied in HLT due to their complexity when facing them from a computational point of view. On the other hand, it is important to have a general knowledge of the world in order to understand the ideas people express through the language. In light of this, HLT range from general tasks, which rely on whole documents (e.g. information retrieval) to more specific ones dealing with words (e.g. morphological analysers), which constitute the basis for building the more general ones.

The goal of this chapter is to place the research work carried out in this thesis in context. Since we deal with HLT, and in particular with Text Summarisation (TS), after having introduced the concept of HLT, the levels of language analysis, focusing on texts are explained (Section 1.1). Then, the most important HLT applications and the challenges they have to face are presented in Section 1.2. Section 1.3 focuses on TS, as it is the HLT application we carried out research into. The remaining of the chapter defines the objectives that will be achieved within the scope of this thesis (Section 1.4), how it is organised (Section 1.5), as well as the research projects this thesis is related to (Section 1.6).

1.1 Levels of Language Analysis

When dealing with natural language text it is necessary to analyse and process it in order to be able to understand the meaning behind it. According to (Jurafsky & Martin, 2000), natural language texts can be analysed from five distinct levels: phonetic and phonological; lexical-morphological, syntactic, semantic and pragmatic.

This section focuses on describing each of these levels of analysis. Moreover, the major problem, regardless of the level of analysis, is the ambiguity of language, a phenomenon difficult to tackle computationally. At the end of this section, we explain it and how it is reflected in each level, but first, we introduce the different levels of language analysis:

- **Phonetic and phonological analysis.** It comprises the study of the linguistic sounds and their composition into syllables, words, and phrases. A single letter or a group of them can be associated to different sounds. For instance, "ea" can be pronounced as "$/\varepsilon/$" (e.g. head) or "$/i:/$" (e.g. speak). Moreover, each sound has its own peculiarities. In this manner, for example, when the vowel "$/i:/$" is pronounced, our tongue is in a forward position in the mouth. Figure 1.1 illustrates graphically the word *believe* and the representation of its pronunciation.

believe
/bɪˈliːv/

Fig. 1.1. Phonetics of the word *believe*.

- **Lexical-morphological analysis.** Morphology studies how words are built up from smaller meaning-bearing components, such as suffixes or affixes. These components can carry information about the gender, number (singular or plural), the verbal form, whether the word has a negative meaning, etc. Moreover, this kind of analysis is capable of providing information about the type of word it is, such as verb, noun, or adjective (i.e., its part-of-speech). For instance, the word *happiness* (noun) is made of the adjective *happy* and the suffix *-ness*. Figure 1.2 illustrates this example.

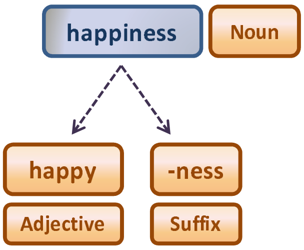


Fig. 1.2. Lexical-morphological analysis for the word *happiness*.

- **Syntactic analysis.** This level of analysis concerns the structural relationships between words; that is, which function a specific word has within a sentence: if it is the subject, the object, etc. An example of the syntactic structure of a sentence is provided in Figure 1.3.

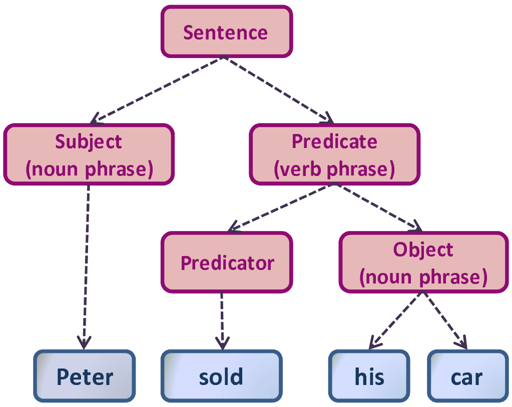


Fig. 1.3. Syntactic analysis for the sentence *"Peter sold his car"*.

- **Semantic analysis.** The aforementioned levels of analysis do not imply understanding the meaning of a sentence or group of sentences. This is the aim of the semantic analysis, which involves the study of the linguistic meaning. For instance, in the sentence *"Peter gives Mary a present"*, we know that *Peter* is a man; *Mary* a woman, and both are human beings. Moreover, we know that a person (*Peter*) is performing an action (*gives* – giving something to somebody) and another person (*Mary*) is receiving something (*a present*). Figure 1.4 shows this analysis graphically.

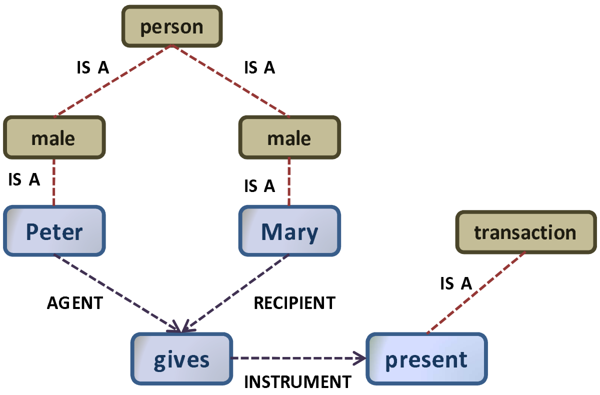


Fig. 1.4. Semantic analysis for the sentence *"Peter gives Mary a present"*.

- **Pragmatic analysis.** Pragmatics studies the relationships between language and context-of-use. For example, let's suppose the following context: It is snowing outside and a person enters a place and he/she feels it is not warm enough inside. Then, this person says: *'It is very cold here'*, but the real meaning behind his/her statement is that he/she wants the heating to be turned on. This example is illustrated in Figure 1.5.

In all these levels, there is a common phenomenon, which is an inherent property that language presents: ambiguity. This happens when a word, sentence, etc. may have several interpretations, thus needing extra information or knowledge to be able to distinguish which is the correct sense in a specific context. The ambiguity is a serious problem for the tools which deal with the automatic processing of natural language. However, this does not prevent such tools from being developed. In contrast, there are tools (e.g. part-of-speech taggers or

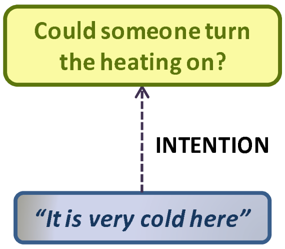


Fig. 1.5. Example of pragmatic analysis.

word sense disambiguators) and resources (e.g. WordNet (Fellbaum, 1998)) that address this problem, facilitating other applications to deal with the language. Depending on the level of language analysis, four types of ambiguity can be distinguished:

- **Lexical ambiguity.** It happens when a word is polysemous, i.e., it has more than one meaning. For example, the noun *bank* has 9 meanings[1]. Among them, we can find *"a business that keeps and lends money and provides other financial services"* or the *"land along the side of a river or lake"*.

  [**1**] *Paul went to the* ***bank*** *to open an account.*
  [**2**] *Paul went for a walk along the* ***bank*** *of the river Thames.*

  In these examples, we need to know the context the word belongs to in order to identify which is the correct sense in this case.

- **Syntactic ambiguity.** A sentence may be ambiguous due to different ways of parsing the same sequence of words. For instance:

  [**1**] *The man saw the boy with the binoculars.*

  In this example, the fragment *"with the binoculars"* leads to ambiguity due to the fact that we do not know if the person who carried the binoculars was the man or the boy.

- **Semantic ambiguity.** It happens when a sentence can be interpreted in different ways. Let's take a look at the following sentence:

[1] According to Longman Dictionary of Contemporary English (http://www.ldoceonline.com/dictionary/bank_1)

[1] *Mark gave a present to the children.*

Does it mean that each child received a present, or there was only one present for all the children? Again, in this case we need to have more information in order to decide which sense is the correct one.

- **Referential ambiguity.** This type of ambiguity occurs when a sentence contains referring expressions (e.g. pronouns) and it is unclear what they refer to. For instance:

  [1] *Ally hit Georgia and then she started crying.*

  In this example, without any additional information, it is difficult to know who is crying, because the pronoun *she* could refer to either Ally or Georgia.

1.2 Human Language Technologies Applications

The rapid growth of the Internet has resulted in a massive increase of available information in different formats (e.g. text, video, images) difficult to cope with. Consequently, intelligent applications based on HLT have to be developed in order to allow users to efficiently manage all this information. These applications can deal with language at different levels. There are applications that take into consideration words as the processing unit. For instance, that is the case of lemmatisers, which provide the root of a word; named entity recognisers, which identify the type of a proper noun (e.g. location, person, organisation); or part-of-speech taggers, which are able to provide information about the category of a word, i.e., if it is a noun, verb, adjective, adverb, etc. Another applications process sentences. For example, syntactic parsing aims at analysing the grammatical structure of a sentence, or semantic role labelling focuses on detection of the semantic arguments associated with the verb of a sentence and their classification into their specific roles (e.g. agent, patient, instrument, etc.). Finally, there are some applications which analyse the whole document. They usually make use of other tools that work with smaller text units, such as lemmatisers and syntactic parsers. In (Mitkov, 2003), the most common

HLT resources and applications are compiled. Next, several widespread HLT applications[2] are briefly introduced:

- **Information Retrieval** aims at finding documents, usually in text format, that satisfies an information need within large collections of documents (Manning *et al.*, 2008).
- **Question Answering** aims at answering simple or complex questions posed in natural language (Strzalkowski & Harabagiu, 2007).
- **Text Classification** aims at sorting a set of documents into categories from a predefined set (Sebastiani, 2002).
- **Text Summarisation** aims at producing a brief version of a document by condensing, selecting or generalising the important information in it (Spärck Jones, 1999).

All the aforementioned HLT applications process natural language in an automatic manner, and analyse it taking into consideration the different levels of analysis previously explained. Furthermore, apart from the ambiguity of the language, they all have to face specific challenges depending on the task they are concerned with. For example, question answering systems have to understand the specific information users are asking for, as well as to be able to find the right answer and provide it to them.

In this thesis, we focus on Text Summarisation, and for this reason, the next section introduces this HLT application in more detail.

1.3 Text Summarisation

Text Summarisation (TS) is a HLT application with the goal to automatically condense information keeping, at the same time, the most relevant facts or pieces of information. In order to be able to successfully perform this task, language has to be processed at different levels (e.g. lexical). Moreover, this task is especially challenging, because it requires the text to summarise to be understood, in order to

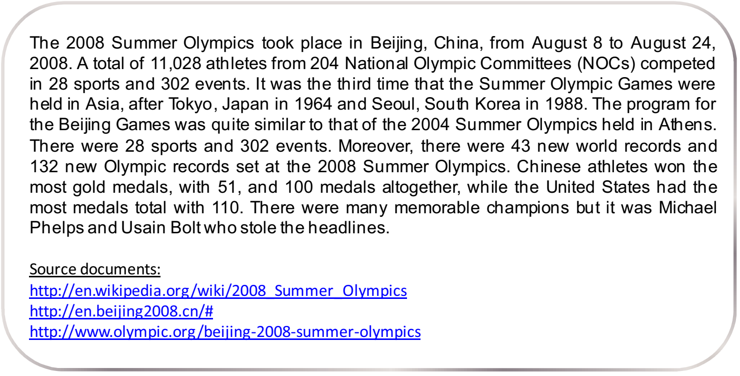
[2] We selected these HLT applications and not others because they are also mentioned in other chapters of this thesis.

distinguish between relevant and irrelevant information. When the information is taken from different documents, it is essential to identify similar information, so it can be included in the summary only once. Additionally, some fragments of information could be removed, combined and/or generalised when necessary. The final goal is to produce a coherent fragment of text (i.e., summary) with important information, that helps users to manage it more easily.

There are a considerable number of applications in which TS is extremely beneficial, providing competitive advantages in the current information society. For instance, if users are interested in knowing the most important facts about the Olympic Games in Beijing, they would use a search engine, that uses information retrieval techniques, to look for the specific information. However, as a result, they will have to cope with millions and millions of Web pages that contain information about the Olympic Games in Beijing[3]. They would not be able to read and process all of them, so they will probably only have a look at the first 10 Web pages, in order to check whether they contain the information they want. Figure 1.6 shows an illustrative example of a summary for the topic Olympic Games in Beijing. As it can be seen, the summary provides basic information about this topic. It also provides the links to the original documents from where the information included in the summary was extracted. In this way, if users are interested in more specific information, they can click into the links and read more about the Olympic Games at that year.

This is only an example of a common scenario where TS would be very beneficial. The same way we use search engines to seek for specific information, TS systems could provide personalised summaries, containing the most relevant information according to users' preferences and interests. Furthermore, TS is not only suitable for coping with large amounts information. There is also an educational part associated to this task, where summaries could serve for simplifying the content of documents, aiding people with learning difficulties to understand texts more easily.

[3] When this query was entered in Google (17-04-2011) we obtained 9.770.000 results.



The 2008 Summer Olympics took place in Beijing, China, from August 8 to August 24, 2008. A total of 11,028 athletes from 204 National Olympic Committees (NOCs) competed in 28 sports and 302 events. It was the third time that the Summer Olympic Games were held in Asia, after Tokyo, Japan in 1964 and Seoul, South Korea in 1988. The program for the Beijing Games was quite similar to that of the 2004 Summer Olympics held in Athens. There were 28 sports and 302 events. Moreover, there were 43 new world records and 132 new Olympic records set at the 2008 Summer Olympics. Chinese athletes won the most gold medals, with 51, and 100 medals altogether, while the United States had the most medals total with 110. There were many memorable champions but it was Michael Phelps and Usain Bolt who stole the headlines.

Source documents:
http://en.wikipedia.org/wiki/2008_Summer_Olympics
http://en.beijing2008.cn/#
http://www.olympic.org/beijing-2008-summer-olympics

Fig. 1.6. Example of a summary about the Olympic Games in Beijing.

1.4 Objectives of the Research Work

The main goal of the research carried out in this thesis is to analyse, develop and research into new techniques and approaches for Text Summarisation. In order to achieve this general goal, the following specific subobjectives are defined:

- To acquire a wide knowledge of Text Summarisation research area. This includes the types of summaries and their main characteristics, as well as the state-of-the-art techniques and approaches concerning both the generation and evaluation of summaries.
- To propose and analyse the appropriateness of different HLT-based techniques and methods for the automatic generation of summaries that contribute to the advances of Text Summarisation.
- To develop a Text Summarisation tool, COMPENDIUM, capable of generating different types of summaries.
- To carry out an exhaustive intrinsic evaluation of COMPENDIUM. This evaluation comprises the assessment of the proposed techniques, as well as the types of summaries it is able to generate, within different domains and scenarios.
- To study the integration of COMPENDIUM within specific HLT applications (opinion mining, question answering and text classification).

By achieving this subobjective, an extrinsic evaluation of COMPENDIUM is also performed.

1.5 Structure of the Thesis

This thesis is organised in several chapters, each of them focusing on a specific issue within the Text Summarisation (TS) research area.

- **Text Summarisation (Chapter 2)**. This chapter provides the state of the art in TS. This comprises the analysis of a wide range of methods and approaches for different summary types and scenarios, together with a review of existing corpora and the most relevant conferences with regard to this task.
- **Text Summarisation Evaluation (Chapter 3)**. This chapter contains the state of the art in the evaluation of summaries. It explains in detail the aspects concerning the evaluation of summaries, explaining the existing types for evaluating summaries, but focusing on the intrinsic assessment of summaries. The existing methods and tools for evaluating summaries, as well as their advantages and disadvantages are discussed. In addition, the use of crowdsourcing services in the context of TS evaluation is also explained.
- **COMPENDIUM Text Summarisation Tool (Chapter 4)**. This chapter provides the theoretical background behind the process of summarisation from a cognitive perspective. Taking this perspective as a basis, COMPENDIUM is proposed as a TS tool able to generate different types of summaries. A general overview of the suggested TS tool, as well as the stages involving COMPENDIUM are explained.
- **Evaluation and Experiments (Chapter 5)**. This chapter contains the evaluation environment, the experiments carried out, and the results obtained by COMPENDIUM. Firstly, the different methods and techniques COMPENDIUM employs are analysed. Then, the TS tool as a whole is evaluated in an intrinsic manner within different domains and types of summaries.
- **COMPENDIUM in Human Language Technology Applications (Chapter 6)**. This chapter analyses the influence of COMPENDIUM (i.e., text summaries) in other HLT applications, and

serves as the extrinsic evaluation of the tool. Specifically, we applied COMPENDIUM to opinion mining, question answering and text classification.

- **Conclusion and Work in Progress (Chapter 7)**. This chapter draws the main conclusions of this research work and the main contributions of this thesis. It also addresses the research that is being conducted at the moment, as well as some issues that will be faced in the future. Finally, a list of the relevant publications is also provided.
- **Síntesis en castellano (Annexe A)** provides a summary of the thesis in Spanish. This synthesis contains the main contributions and findings, as well as it explains the most relevant experiments carried out and the results obtained.

1.6 Context: Research Projects Related to the Thesis

The research work presented in this thesis is directly related to the activities developed in three research projects. We next provide a brief description of each one, together with their objectives, focusing on the aspects this thesis relates to them.

- **Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano**[4]

  This project was coordinated by Universidad de Alicante and it involved five more participants (Universidad de Jaén, Universidad Nacional de Educación a Distancia, Universidad Politécnica de Valencia, Universitat Politècnica de Catalunya, and Universitat de Barcelona).

  The goal of the project was to analyse, experiment, and develop intelligent, interactive and multilingual Text Mining technologies, as a key element of the next generation of search engines, systems with the capacity to find "the need behind the query".

  Three basic research lines were proposed to meet the overall goal of the project:

[4] TEXT-MESS (TIN2006-1526-C06-01)

- To develop Text Mining systems analysing interactive and multilingual issues, as well as the effectiveness of such systems on different types of texts (written, oral, images, etc.) and domains (open - the Web - and specific - tourism or biomedicine).
- To improve and adapt already existing resources and tools and to create new resources, techniques and tools required to undertake new applications based on HLT, by combining linguistic knowledge and machine learning techniques.
- To connect this project with the main international evaluation campaigns on search systems and HLT, to participate in those campaigns in order to contrast the results obtained with those of the main international groups, and to promote and coordinate some of the tasks, thus promoting research lines related to this project.

Within this project, this thesis was directly related to the activity 4: Clustering, visualisation, exploration and synthesis of search results of module 3: Intelligent, Interactive and multilingual Text Mining. In this activity, it was analysed how to cluster, display and facilitate to the user the exploration and synthesis of the searching results in the situations where it is necessary to analyse several sources, and to disregard, select, filter, summarise and analyse information, instead of just providing a ranking of Web pages, as search engines usually do. More specifically, it directly fits into task 3: Information synthesis, where automatic information synthesis and the building of interactive information synthesis assistants are analysed.

- **Las Tecnologías del Lenguaje Humano ante los nuevos retos de la comunicación digital**[5]

  This project, also coordinated by Universidad de Alicante and three more participants (Universidad de Jaén, Universidad Politécnica de Valencia, and Universitat de Barcelona), is a continuation of the previous project, with a particular emphasis on new textual genres born with the Social Web (or Web 2.0).

  The overall aim of this project centres on the study, development and experimentation with different techniques and systems based on HLT for developing the next generation of intelligent digital infor-

[5] TEXT-MESS 2.0 (TIN2009-13391-C04-01)

mation processing systems (modelling, retrieval, processing, comprehension and detection), in order to meet the present challenges posed by digital media. In this new scenario, systems have to incorporate the reasoning capability to ascertain the subjectivity of information in all contexts (spatial, temporal and emotional), while analysing the various dimensional uses (multilingualism, multimodality and register). In order to achieve this overall objective, the following courses of action are proposed:

- To create, adapt and improve resources, techniques and tools for modelling the language emerging from the current digital environment via machine learning approximations.
- To develop intelligent information processing systems (modelling, retrieval, processing, comprehension and detection), geared to the new communication formats, capable of interpreting and assessing the context of the message.
- To integrate HLT resources, tools and systems developed and compiled to form an intelligent information processing metasystem, the validity of which will be demonstrated in the specific application scenario of competitive intelligence in the socio-economic sphere. The TEXT-MESS 2.0 metasystem is illustrated in Figure 1.7.

In particular, the research work carried out in this thesis is related to the task 2: To Build a System of Automatic Opinion Synthesis of activity 4: Automatic Data Summarisation & Synthesis Systems of module 3: Information Comprehension and Detection Systems in Different Language Registers. This activity involves studying systems for automatically generating summaries as a mechanism for information synthesis.

- **Desarrollo de técnicas inteligentes e interactivas de minería de textos**[6]

  This project in which Universidad de Alicante is involved aims at developing intelligent and interactive techniques for text mining. The main objectives in this project are the following:

[6] PROMETEO/2009/199

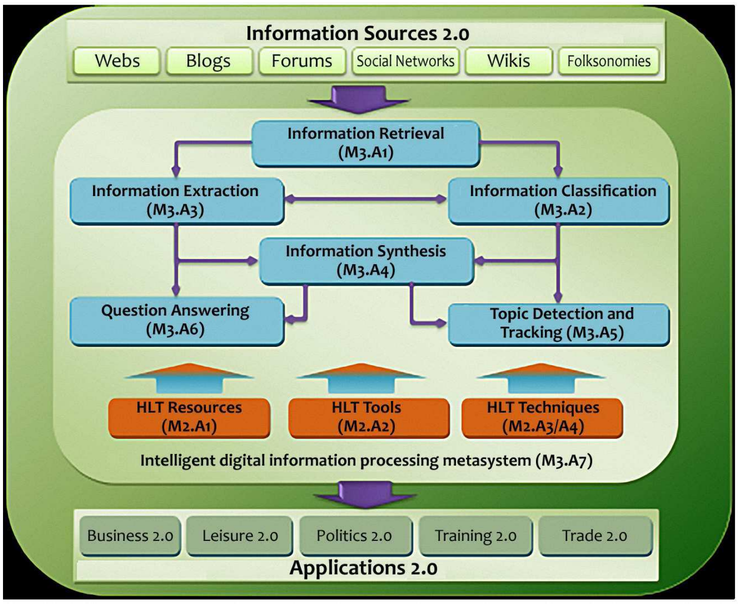

Information Sources 2.0
Webs
Blogs
Forums
Social Networks
Wikis
Folksonomies
Information Retrieval (M3.A1)
Information Extraction (M3.A3)
Information Classification (M3.A2)
Information Synthesis (M3.A4)
Question Answering (M3.A6)
Topic Detection and Tracking (M3.A5)
HLT Resources (M2.A1)
HLT Tools (M2.A2)
HLT Techniques (M2.A3/A4)
Intelligent digital information processing metasystem (M3.A7)
Business 2.0
Leisure 2.0
Politics 2.0
Training 2.0
Trade 2.0
Applications 2.0

Fig. 1.7. TEXT-MESS 2.0 metasystem.

- To develop text mining systems capable of finding, extracting, analysing, classifying and retrieving information. Within this objective, multi-lingual (giving emphasis on Spanish and Catalan), as well as interactive aspects are studied.
- To improve and adapt the existing tools and resources, and to create new ones in order to tackle current and future HLT applications.
- To assess the quality, reliability and reusability of the information using HLT techniques. In the current Information Society the manner in which humans interact has changed. Thanks to the new formats of communication, users play an active role in the generation of content for the Web. However, its underlying problem is that the Web ends up containing information with a poor quality, and not very reliable.

The research conducted in this thesis is related to the second objective. In this thesis, several HLT techniques are analysed for TS, proposing new ones and developing finally a TS tool, COMPENDIUM, that contributes to improve the state of the art in this research area.

2. Text Summarisation

2.1 Introduction

Although it started in the late fifties (Luhn, 1958), Text Summarisation (TS) has experienced a great development in recent years, and a wide range of techniques and paradigms have been proposed in this research field (Spärck Jones, 2007), (Lloret & Palomar, 2011b). However, to produce a summary automatically is very challenging. Issues such as redundancy, temporal dimension, coreference or sentence ordering, to name a few, have to be taken into consideration especially when summarising a set of documents (multi-document summarisation), thus making this field even more difficult (Goldstein *et al.*, 2000). Moreover, research attempting to overcome the lack of coherence that summaries often present has been fuelled in the last years, resulting in combined approaches that identify relevant content and merge it into new fragments of information (Barzilay & McKeown, 2005a), (Zajic *et al.*, 2008). It is also worth mentioning that as society changes, so does TS, adapting itself to new requirements. For instance, the Web 2.0 (social Web) has led to the emergence of new types of Websites, such as blogs, forums, or social networks, where anybody can express his/her opinions towards a topic, entity, product or service. This has resulted in a new type of summaries (sentiment-based) with the purpose of summarising users' opinions. Other types of summaries, such as update summaries, aim at providing users with the most recent information, assuming that they have a previous knowledge of the topic. When carrying out research into this area, it is essential to be aware of previous TS approaches, so that new or improved methods can be suggested in order to tackle the different types of summaries and their requirements.

Therefore, a review of the state of the art in TS is carried out in this chapter, describing the existing taxonomies and focusing on the approaches that have been developed in the last decade, the new types of summaries that have appeared in recent years, as well as the new scenarios. The aim of the chapter is to provide a comprehensive overview of this research area, emphasising recent summary types, and how summaries, although not perfect, can be of great help for other systems based on Human Language Technologies (HLT).

The chapter is organised as follows. Firstly, Section 2.2 provides general background in TS. Within this section there are four subsections: an analysis of the main taxonomies and types of summaries, as well as the stages involved in the process of TS (Subsection 2.2.1); a general overview of the main TS approaches developed in recent years (Subsection 2.2.2); a roadmap of the presence of TS in international evaluation campaigns (Subsection 2.2.3); and available corpora particularly developed for TS (Subsection 2.2.4). Secondly, Section 2.3 focuses on the approaches that have been proposed recently to tackle new types of summaries, and in new scenarios. In particular, TS techniques that have been employed for generating *personalised*, *update*, *sentiment-based*, *surveys* and *abstractive* summaries are explained in Subsection 2.3.1. Concerning the analysis of TS in other scenarios different from the traditional newswire texts, the approaches specifically proposed to address automatic summarisation of *literary texts*, *patent claims*, *image captioning*, and *Web 2.0 textual genres* are outlined in Subsection 2.3.2. Thirdly, Section 2.4 analyses the applicability of text summaries in other intelligent applications based on HLT, focusing on three specific tasks: information retrieval (Subsection 2.4.1), question answering (Subsection 2.4.2) and text classification (Subsection 2.4.3). Finally, Section 2.5 concludes the chapter, by providing some insights of the future directions of TS.

2.2 Text Summarisation Overview

The goal of this section is to provide a general overview on TS, and it is organised into four subsections. Subsection 2.2.1 discusses three taxonomies that have been proposed from different perspectives, and taking them as a basis, the most widespread types of summaries according to several factors are explained. In order to be able to generate different types of summaries, the stages of the TS process from a computational perspective are also outlined in this subsection. Then, a wide range of approaches are analysed (Subsection 2.2.2). These approaches have been grouped according to the predominant techniques employed (statistical-based, topic-based, graph-based, discourse-based, and machine-learning-based). Further on, a brief review of international forums related to TS is also provided in Section 2.2.3, together with an analysis of the TS trends that has been changing over the years which are reflected in the proposed conference tracks. Finally, the corpora specifically developed for generating summaries is described in Subsection 2.2.4.

2.2.1 Common Factors for Classifying Summaries

A wide range of summaries can be generated depending on different factors, such as the type of input/output, the purpose of the summary, or the type of reader. One of the most well-known existing taxonomies of summaries was proposed in (Spärck Jones, 1999), where three classes of context factors that influence summaries are taken into consideration: input, purpose and output factors. Input factors deal with aspects related to the source, such as genre, language, or register. The second ones, purpose factors, include audience and use, for example literary reviews or emergency alerts. Finally, output factors focus on the style and coverage, and are normally driven by purpose factors. Figure 2.1 shows the proposed taxonomy by Spärck Jones, where the context factors that affects the summarisation task can be seen in a schematic way.

The taxonomy proposed by (Spärck Jones, 1999) is not the only one proposed to classify summaries with respect to different aspects. Hovy and Lin (Hovy & Lin, 1999) suggested also a similar taxonomy (Figure 2.2). In the same way as the input, output and purpose factors described in the previous taxonomy, this classification distinguishes between characteristics of the source document, characteristics of the summary as a text, and characteristics of the summary usage. For instance, the characteristics of the source text comprise factors, such as the source size (whether the summary is generated from one or several documents), or the specificity (if the summary is general or domain-specific). The main difference between both taxonomies is the

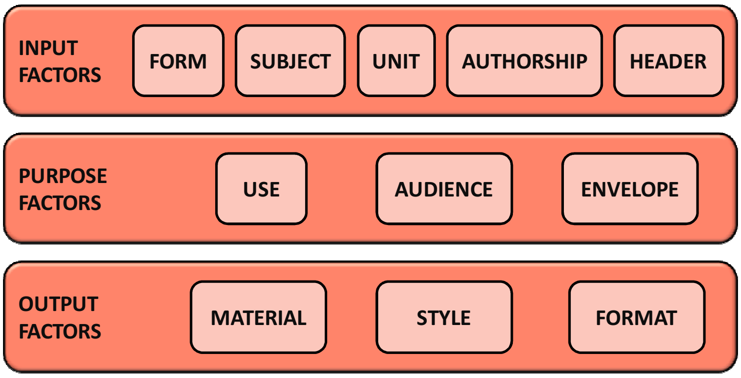


Fig. 2.1. Context factors proposed by Spärck Jones (1999).

fact that the latter takes into consideration specific factors concerning the coherence and the subjectivity level of the summary.

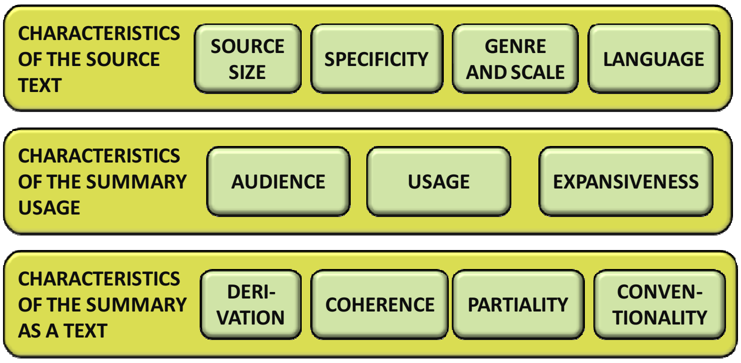


Fig. 2.2. Hovy and Lin's (1999) taxonomy.

Additionally, there is one more taxonomy suggested by Mani and Maybury (Mani & Maybury, 1999), where summarisation systems are classified with regard to the approach adopted to generate summaries. Different approaches can be tackled to produce summaries, facing the problem from three levels: surface-, entity- or discourse-level. Surface-level approaches aim at representing information in terms of shallow features which are then selectively combined together to determine a salient function used to extract the most important information of a document. Such features comprise *thematic features*, which take into consideration statistically salient terms, for example based on frequency counts; *location*, which accounts for the position of a specific unit (word, sentence, etc.) in a document (paragraph, section); *background* refers to the presence of terms from the title or headings in the text, the initial parts of the document, or a user's query; finally, *cue words* are expressions such as *"in summary"*, *"our investigation"*, *"in particular"*, or *"in conclusion"*. Entity-level approaches build an internal representation of the document or documents to model the entities and their relationships. These approaches tend to represent patterns of connectivity in the document and these relationships include, for example, *similarity* through vocabulary overlap; *proximity*, which refers to distance between text units; *thesaural relationships* among words like synonymy or part-of-relations; or *logical relations* such as contradiction, entailment or agreement. Lastly, discourse-level approaches model the global structure of the document. This includes features concerning the format of the document (such as document outlines or hypertexts), the threads of topics or subtopics developed in the document, or their attempt to capture the structure of different sorts of texts, for example, narrative or argumentative documents' structure. Figure 2.3 reflects these issues according to this taxonomy.

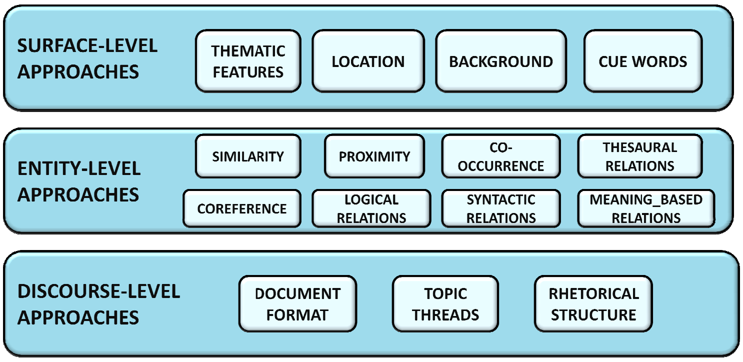


Fig. 2.3. Mani and Maybury's (1999) taxonomy.

The taxonomies proposed in (Spärck Jones, 1999) and (Hovy & Lin, 1999) deal with a very fine-grained granularity, in the sense that summaries are classified with respect to different criteria in accordance with their own nature, whereas the one suggested by Mani and Maybury (1999) groups summarisation approaches, concerning the type of features and techniques used to generate summaries. The main problem of having a taxonomy with such a fine-grained granularity arises when one wants to classify a summarisation system with regard to those criteria, since most systems may share several characteristics, thus increasing the difficulty in its classification and making it unclear. At the same time, the problem with Mani and Maybury's taxonomy is the purity of the suggested classification. Systems often rely on hybrid approaches (Mani & Maybury, 1999), combining for example, discourse- and surface-level features.

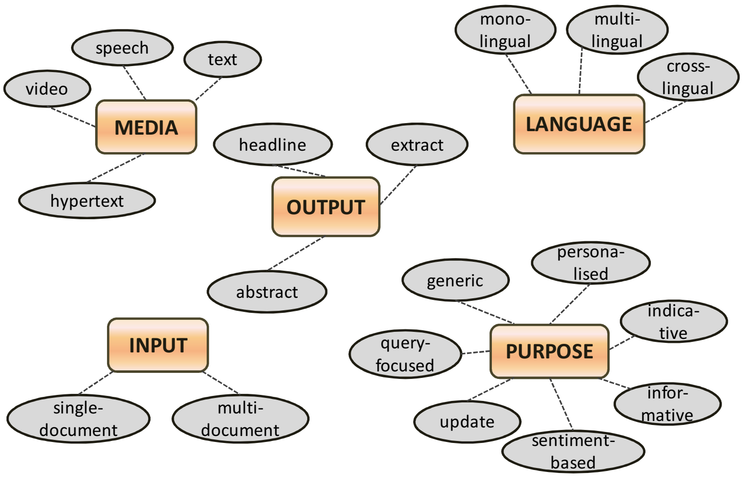


Fig. 2.4. Summarisation types according to several factors.

Therefore, taking into account the aforementioned taxonomies, summarisation approaches can be characterised according to many features. Although it has been traditionally focused on text, the input to the summarisation process can also be multimedia information, such as images (Fan *et al.*, 2008); video (He *et al.*, 1999) or audio (Zechner & Waibel, 2000), as well as on-line information or hypertexts (Sun *et al.*, 2005). Furthermore, we can talk about summarising only one document (*single-document summarisation*) or multiple ones (*multi-document summarisation*). Regarding the output, a summary may be an *extract* (i.e. when a selection of “significant” sentences of a document is shown), *abstract*, when the summary can serve as a substitute for the original document and new vocabulary is added, or even a *headline* (or title). It is also possible to distinguish between *generic* summaries and *query-focused* summaries (also known as user-focused or topic-focused). The first type of summaries can serve as a surrogate of the original text as they may try to represent all relevant facts of the source text. In the latter, the content of a summary is biased towards a user need, query or topic. Concerning the style of the output, a broad distinction is normally made between two types of summaries. *Indicative* summaries are used to indicate what topics are addressed in the source text. As a result, they can give a brief idea of what the original text is about. The other type, *informative* summaries, are intended to cover the topics in the source text and provide more detailed information. Apart from these two types of summaries, another one can also be taken into account, i. e. *critical evaluative abstracts.* This kind of summaries focuses on expressing the author’s point of view about a specific topic or subject, and they include reviews, opinions, feedback, recommendations, etc., with a strong dependence on cultural interpretation (Mani, 2001a). That is the reason why they are so difficult to produce automatically, and therefore most systems only attempt to generate either indicative or informative summaries, by just summarising what appears in the source document. In recent years, new types of summaries have appeared. For instance, the birth of the Web 2.0 has encouraged the emergence of new types of textual genres, containing a high degree of subjectivity, thus allowing the generation of *sentiment-based summaries.* Furthermore, *update summaries* are another example of new type of summary. They assume that users have already a background and they only need the most recent information about a specific topic. More types of summaries are explained in Section 2.3.1. Finally, concerning the language of the summary, it can be distinguished between *mono-lingual*, *multi-lingual*, and *cross-lingual* summaries, depending on the number of languages dealt with. The cases where the input and the output language is the same lead to mono-lingual summaries. However, if different languages are involved, the summarisation approach is considered multi-lingual or cross-lingual. For example, if a summarisation system produces a Spanish summary from one or more documents in Spanish, that is the case of a mono-lingual system. On the other hand, if it is able to deal with several languages, such as Spanish, English or German, and produces summaries in the same languages as the input document, we would have a multi-lingual summarisation system. Beyond these approaches, if the summary is in Spanish, but the original documents are in English, the summariser would deal with cross-linguality, since the input and output languages are different. The most common factors regarding summarisation are depicted in Figure 2.4.

All these types of summaries have to be created following a summarisation process, thus allowing to transform the source document or documents into a summary. According to (Hovy, 2005), three stages have to be taken into account for producing a summary from a computational point of view:

- *Topic identification.* It consists of determining the particular subject the document is about. It is usually approached by assigning each unit (words, sentences, phrases, etc.) a score which is indicative of its importance. In the end, the top score units up to a desired length are extracted.
- *Interpretation or topic fusion.* During this stage, the topics identified as important are fused, represented in new terms, and expressed using a new formulation, which includes concepts or words not found in the original text. This stage is what distinguishes extractive from abstractive summarisation.
- *Summary generation.* This stage only makes sense if abstractive summaries are generated. In these cases, natural language generation techniques (text planning, sentence planning, and sentence realisation) are needed to produce the final text of the summary.

In Figure 2.5 these stages are graphically depicted.

2.2.2 Common Approaches for Generating Summaries

As it was previously explained, the summarisation process can be decomposed into three main subtasks: *topic identification*, *topic interpretation*, and *summary generation* (Hovy, 2005). However, since the

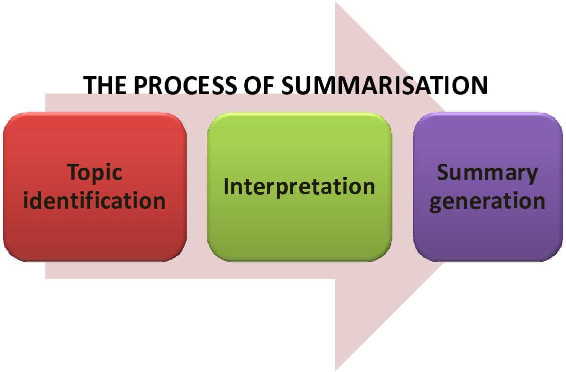


Fig. 2.5. Stages of the process of summarisation according to Hovy (2005).

summary generation stage is not easy to tackle, most approaches only focus on the first stages, by simply extracting the sentences as they appear in the documents, thus producing extracts as a consequence. Next, several extractive approaches are described. Specifically, we distinguish between five types of approaches suitable for TS. Such approaches are grouped depending on the nature of the techniques employed. In particular, these are: *statistical-based*; *topic-based*; *graph-based*; *discourse-based*; and *machine learning-based*.

Statistical-based approaches. Luhn (1958) used term frequency to produce summaries from scientific documents with the purpose of determining the relevance of a sentence in a document. The underlying assumption is that the most frequent words are indicative of the main topic of a document. However, not all the words are taken into consideration. For example, stop words, i.e. words without carrying any semantic information, such as “a” or “the”, are not used for computing the term frequency. Under the same assumption, a number of techniques based on term frequency counts have been employed in TS. For instance, (Lloret & Palomar, 2009) use the frequency of words in combination with the length of noun phrases to compute the relevance of a sentence, outperforming the results of the state-of-the-art in single-document summarisation for newswire domain.

In (McCargar, 2005), several statistical approaches, such as term frequency or inverse document frequency (*tf*idf*), are briefly analysed, as well as the potential problems this type of features may have. The idea behind *tf*idf* is that frequent terms in a document are important only if they are not very frequent in the whole collection. This technique has also been employed to score sentences, for instance in (Gotti *et al.*, 2007). In (Filatova & Hatzivassiloglou, 2004) it is claimed that these methods may be not sufficient for building high-quality summaries, and other types of knowledge, for instance events, semantic knowledge, or discourse information may be more appropriate to tackle TS. However, a deeper review of statistical techniques for TS is carried out in (Orăsan *et al.*, 2004) and (Orăsan, 2009), where it is shown that these techniques, despite being simple and not requiring a deep level of knowledge analysis, are appropriate for building good summaries. In addition to the aforementioned techniques, mutual information, information gain and residual inverse document frequency are also analysed. Mutual information is used to measure the dependency or the common information between two words, whereas information gain is a good metric for deciding the relevance of an attribute, and in this case, it could be perfectly applied to the terms or sentences in a document. Residual inverse document frequency is a variant of the inverse document frequency, which computes the term frequency according to a probabilistic distribution. Each technique establishes a manner of assigning weights to the words included in the document, and then, sentences are scored based on these weights, in order to determine their relevance.

The approach suggested in (Mori, 2002) also employs information gain for determining the weight of document terms, and then use it for successfully summarising documents. The idea is to build clusters of documents first according to the similarity among them, and then compute the weight of each word in the clusters. The final summary is produced by selecting those highest scored sentences on the basis on the weight of words it contains, previously computed using information gain.

Topic-based approaches. In (Edmundson, 1969) summaries are produced by means of cue word identification. This technique consists in determining the relevance of a sentence by means of the phrases or words it contains. Sentences containing phrases like "*in conclusion*" or "*the aim of this paper*" may introduce new topics, and also be good indicators of relevant information. Moreover, other approaches, such as (Boguraev & Neff, 2000), (Neto *et al.*, 2000), (Angheluta *et al.*, 2002), or (Harabagiu & Lacatusu, 2005) take profit of the advantages of combining topics' identification and segmentation. Particularly, in (Harabagiu & Lacatusu, 2005), the topic structure is characterised in terms of topics themes, which are representations of events that are reiterated throughout the document collection, and therefore represent repetitive information. Five different ways of representing topics are analysed:

- Via *topic signatures*. This idea comes from (Lin & Hovy, 2000), where it is assumed that the topic of a document can be represented using a set of terms.
- Via *enhanced topic signatures*. This differs from the previous one in the fact that the aim now is to discover relevant relations between two topic concepts.
- Via *thematic signatures*. This is carried out by segmenting documents using the TextTiling algorithm (Hearst, 1997) first, and then assigning labels to themes to be able to rank them later.
- Via *modelling the content structure of documents*. The assumption here is that all texts describing a given topic are generated by a single content model (in this case a Hidden Markov Model).
- Via *templates*. This method follows the idea of the field of information extraction, identifying specific entities or facts to represent topics within a text.

Furthermore, in (Teng *et al.*, 2008), a single-document summarisation approach is suggested which combines local topic identification with term frequency. The proposed methodology computes the sentence similarity first, and then performs the topic identification by doing sentence clustering. In a second step, sentences from local topics are selected according to the term frequency value.

Moreover, not only topic words are used to detect relevant information within a document. In other approaches, (Kuo & Chen, 2008), for instance, informativeness and event words are also taken into consideration in order to produce multi-document summaries. The underlying idea is that this kind of words indicate the important concepts and relationships, and can be used to detect relevant sentences within a set of documents. Furthermore, a temporal resolution algorithm is used, so that dates and other temporal expressions can be translated into calendrical forms. The identification of multiple themes within an heterogeneous collection of documents is addressed in (Ando *et al.*, 2005) by means of vector space representations; in particular, *Iterative Residual Rescaling* is used. This method constructs a vector space model which indicates relationships among documents, topical phrases, and sentences.

Graph-based approaches. The use of graph-based ranking algorithms has also been shown to be effective in TS. Basically, the nodes of the graph represent text elements (i.e. normally words or sentences), whereas edges are links between those text elements, previously defined (for instance, semantic relations, such as synonymy). On the basis of the text representation as a graph, the idea is that the topology of the graph will reveal interesting things about the salient elements of the text, for example concerning the connectivity of the different elements. LexRank (Erkan & Radev, 2004) is a multi-document summarisation system, in which all candidate sentences that can be potentially included in the summary are represented in a graph. In this graph representation, two sentences are connected if the similarity between them is above a predefined threshold. Then, once the network is built, the system finds the most central sentences by performing a random walk on the graph. In (Mihalcea, 2004), an analysis of several graph-based algorithms is carried out, evaluating also their application to automatic sentence extraction in the context of TS.

Furthermore, in (Wan *et al.*, 2007), an approach based on affinity graphs, for both generic and query-focused multi-document summarisation, is suggested. The idea here is to extract sentences with high information richness and novelty. This is achieved by taking into consideration the similarity between each pair of sentences, incorporating topic information, differentiating intra-document and inter-document links between sentences, and finally penalising redundant information. In (Giannakopoulos *et al.*, 2008a) character and word n-gram graphs are used to extract relevant information from a set of documents, whereas in (Plaza *et al.*, 2008), (Plaza, 2011) graphs are built using concepts identified with Wordnet (Fellbaum, 1998) and *is-a* relationships, which are then used to build a graph representation of each sentence in a document. This approach has been proven successfully in different domains, such as newswire, biomedical documents or image captions.

Discourse-based approaches. Besides all the previous mentioned techniques, it is also possible to face the summarisation problem from a linguistic point of view, for instance exploiting discourse relations. Rhetorical Structure Theory (RST) proposed in (Mann & Thompson, 1988) served as a basis for the summarisation approach developed in (Marcu, 1999). In this approach, the rhetorical relations are extended, and this kind of discourse representation (nucleus and satellite relations, depending on how relevant the information is) is used to determine the most important textual units in a document. In (Khan *et al.*, 2005) the RST is combined with a generic summariser in order to add linguistic knowledge to the summarisation process. Although the results obtained for this mixed approach do not improved the ones obtained by the generic summariser, it was claimed that the drawback of this approach was mainly due to the parser, which could not detect all the RST relationships, otherwise linguistic knowledge could have improve the overall summarisation performance. Furthermore, in (Cristea *et al.*, 2005) an approach similar to RST is described, differing from the previous ones in the lack of relation names and the use of binary trees. This summarisation approach is intended to exploit the coherence and cohesion of a document.

Cohesion and coherence are two of the main challenging issues for TS. Some approaches rely on the identification of such relations in order to improve the quality of the generated summaries. Cunha et al. (2007) combine statistical and linguistic techniques to prove that results improve with respect to use only one type of them. In (Gonçalves *et al.*, 2008), coreference chains are used to deal with referential cohesion problems that are frequent in the extractive summarisation approach. A post-processing system is developed in order to rewrite referential expressions in the most possible coherent way, and it is applied after the summary is generated, obtaining considerable improvements in comparison to the original summaries. In order to guarantee the coherence of a summary, a widespread approach is to use lexical or coreference chains. However, the use of coreference chains is not novel in TS. The first approaches can be found in (Baldwin & Morton, 1998), and (Azzam *et al.*, 1999). The main assumption is that the longest coreference chain indicates the main topic of the document, and shorter chains represent subtopics. Therefore, one possible strategy for building summaries is to select only those sentences related to the longest chain. This strategy helps to maintain the coherence of the text.

A similar idea is to use *lexical chains*, which consist in determining sequences of semantic related words (for example, by concept repetition or synonymy relations). By using lexical chains, the main topics of a document can also be detected. This technique has been widely used in summarisation, and approaches like the ones described in (Barzilay & Elhadad, 1999), (Medelyan, 2007) or (Ercan & Cicekli, 2008) exploit them to produce summaries. It is worth mentioning that being able to identify all the entities that are connected within a document or across documents, prevent summaries from the common *dangling anaphora* phenomenon, thus producing more coherent resulting summaries (Elsner & Charniak, 2008). This phenomenon consists of having words in a text (mostly pronouns) without its correct antecedent included in the summary. For example, if a summary contains the pronoun “he”, but its antecedent (e.g. the president of Spain) is not mentioned in it, this would lead to an unclear summary with this specific type of problem, which would make the summary difficult to understand, or even incoherent. In order to reduce this problem some approaches combine the use of anaphora resolution in order to help TS (Orăsan, 2004), (Mitkov *et al.*, 2007). In these approaches, documents are first processed to resolve anaphoric pronouns, and then a summarisation system is run in order to produce a summary. Their final goal is to determine whether an anaphora resolution system improves the quality of automatic summaries or not. Due to the moderated performance of this kind of systems, this is hard to achieve, and contrary to the intuition, TS does not improve very much. However, in (Orăsan, 2007) an ideal anaphora resolution system was simulated, resolving the anaphoric relations in scientific documents manually, and, in this case, it was proven that summary’s results improve noticeably. In (Steinberger *et al.*, 2007), it was stated that the improvement associated to TS, when using an anaphora resolution system, not only depends on the lower performance of the anaphora resolver, but also in the way anaphoric relations are used. As a consequence, they used the anaphoric relations from two different perspectives: on the one hand, to improve the quality of summaries, and on the other hand, to check the coherence of a summary, once it was already generated. This was done by checking if the coreference chains of the summary were sub-chains of the ones identified in the source documents.

Machine Learning-based approaches. The approaches for generating summaries that are next explained are based on machine learning algorithms. The first machine learning methods used in TS include binary classifiers (Kupiec *et al.*, 1995), *Hidden Markov Models* (Conroy & O'Leary, 2001), (Schlesinger *et al.*, 2002), and *Bayesian* methods (Aone *et al.*, 1998). Apart from these, a wide range of machine learning techniques can be used for TS. NetSum (Svore *et al.*, 2007) bets on single-document summarisation and produces extracts from newswire documents based on neuronal nets, using RankNet (Burges *et al.*, 2005) as a learning algorithm to score the sentences and extract the most important ones. Besides the common features based on keywords and sentence position, a new set of features based on Wikipedia[1] and query logs are also used in a way that, for example, sentences containing query terms or Wikipedia entities, are indicative of important content. In (Schilder & Kondadadi, 2008), a query-focused multi-document summariser is presented, named as FastSum, where sentences are ranked using a machine learning technique called *Support Vector Regression* (SVR), and *Least Angle Regression* for feature selection. SVR was used in summarisation before, in the approach described in (Li *et al.*, 2007), where word-, phrase-, or semantic-based features, as well as sentence position or name entities were used to train the classifier automatically. Further on, the extracted features were combined, and then sentences were scored. In (Wong *et al.*, 2008), an extractive summarisation approach is presented, employing supervised and semi-supervised learning methods. The features involved are grouped into different types - surface, content, relevance and event features - which include sentence position, number of words in a sentence, centroid and high frequent terms, or similarity between sentences, among others. Regarding the supervised approach, a *Support Vector Machine* (SVM) algorithm is used, whereas for the semi-supervised approach, a probabilistic SVM and a *Naïve Bayesian* classifier are co-trained to exploit unlabelled data. SVM technique was also used in (Fuentes *et al.*, 2007) to detect relevant information to be included in a query-focused summary, where structural, cohesion-based and query-dependent features were used for training.

[1] http://www.wikipedia.org/

The advantage of using machine learning for TS is that it allows to easily test the performance of a high number of features, for instance lexical, syntactic, statistical, etc. using different machine learning paradigms for deciding which features perform the best. However, these approaches also need a big training corpus in order to be able to obtain conclusive results. In the case of TS, the corpus usually consists of a set of human-written summaries, or annotated source documents containing which sentences are important for the summary, and which are not.

2.2.3 Relevant Conferences and Workshops

At the end of the 1990's, the TIPSTER Text Summarisation Evaluation[2] (SUMMAC) was the first conference aimed at evaluating automatic summarisation systems, where text summaries were tested in document classification and question answering, in order to analyse whether they were suitable surrogates for full documents. A detailed explanation of this evaluation forum and how summaries were evaluated can be found in (Mani *et al.*, 2002). The National Institute for Informatics Test Collection for IR[3] (NTCIR) also developed a series of Text Summarisation Challenges (TSC) workshops, which included Japanese summarisation tasks in 2001 (TSC), 2002 (TSC2), and 2003 (TSC3). Besides these conferences, the important conferences that focused only in TS were the Document Understanding Conferences[4] (DUC) that were held yearly from 2001 to 2007. In these conferences, different tasks were proposed over the years, taking into account new challenges and requirements for TS, forcing also systems to be dynamic and adaptable. Along the editions of the DUC conferences, it can be seen how the summarisation systems have progressed, as well as the different evaluation methodologies that have been proposed to evaluate the corresponding summaries. These changed from a complete manual evaluation, where assessors used the SEE evaluation environment[5] to facilitate the comparison of automatic and human-made summaries' content, to a fully automatic evaluation of the content using ROUGE (Lin, 2004) and Basic Elements (Hovy *et al.*, 2006).

[2] http://www-nlpir.nist.gov/related_projects/tipster_summac/

[3] http://research.nii.ac.jp/ntcir/outline/prop-en.html

[4] http://www-nlpir.nist.gov/projects/duc/

[5] http://www.isi.edu/licensed-sw/see/

The tasks involved in the DUC conferences also changed over the editions, starting at the beginning with generic single-document summarisation, and continuing further on with query-focused multi-document summarisation. A comprehensive overview of the major summarisation conferences, focusing particularly in DUC, can be found in (Over *et al.*, 2007). More specifically, the overview for some DUC editions is also provided in (Over & Ligget, 2002) and (Dang, 2006). These types of conferences are very useful to evaluate and compare automatic systems, and at the same time, they also provide a good set of corpora, comprising documents and model summaries, which are freely available on demand[6]. Unfortunately, due to the fact that all the editions worked under the newswire domain, the data deals with a unique domain. Since 2008, DUC conferences have been no longer organised, because they have become part of the Text Analysis Conference[7] (TAC), within which a summarisation track is included. In TAC 2008, two different tasks were proposed within the TS track. The first followed the same idea as the update summarisation task in the DUC 2007, which consisted of building summaries containing updated information with respect to a given set of news documents, whereas the second one, was a pilot task whose aim was to generate opinion summaries from blogs. In TAC 2009, the update summarisation task was kept but, instead of the opinion summarisation task, a new one concerning the automatic evaluation of summaries was proposed (*Automatically Evaluating Summaries Of Peers* – AESOP[8]). The goal of this task was to automatically score a summary for a given metric that reflected summary content. In TAC 2010, the task concerning evaluation was maintained but, in contrast, update summaries have been changed into guided summaries. The idea under this new kind of summaries was to encourage systems to use deeper semantic knowledge, building summaries that contained specific information about different aspects of a topic. For instance, if a set of documents were about an accident, we might be interested in information concerning when it occurred, why, where, etc. In the current edition (TAC 2011), besides guided summarisation and the AESOP task, a multi-lingual pilot task is proposed. This is one of the novelties introduced in the present edition, since only in DUC 2004, a first attempt of cross-linguality was

[6] http://www-nlpir.nist.gov/projects/duc/data.html

[7] http://www.nist.gov/tac/

[8] http://www.nist.gov/tac/2009/summarisation/aesop.09.guidelines.html

carried out. The other novelty concerns the AESOP task, which in addition to focus on automatic metrics that can assess the content of a summary, readability aspects are also going to be taken into consideration this year.

Conference	**Summarisation task requirements**
SUMMAC [a]	single-document, query-focused, news
TSC [b](NTCIR)	query-focused, generic, news
TSC2 (NTCIR)	single and **multi-document**, generic, news
TSC3 (NTCIR)	multi-document, generic, news
DUC-01 [c]	single and multi-document, generic, news
DUC-02	single and multi-document, generic, news
DUC-03	multi-document, **query-focused**, news
DUC-04	single and multi-document, topic-oriented, news, **cross-lingual**
DUC-05	multi-document, query-focused, news
DUC-06	multi-document, query-focused, news
DUC-07	multi-document, **update**, query-focused, news
TAC-08 [d]	multi-document, update, query-focused, **sentiment-based**, news & **blogs**
TAC-09	multi-document, update, query-focused, news, **evaluation**
TAC-10	multi-document, **guided**, query-focused, news, evaluation
TAC-11	multi-document, **guided**, query-focused, news, evaluation, **multi-lingual**

[a] SUMMAC summary types and TIPSTER Text Summarisation Evaluation Conference.

[b] TSC: Text Summarisation Challenges.

[c] DUC: Document Understanding Conferences.

[d] TAC: Text Analysis Conferences.

Table 2.1. Features for the summarisation tasks in each conference.

Table 2.1 shows some features of the tasks involved in the previously mentioned conferences. Boldfaced words indicate the novelties introduced in the summarisation tasks over the years within each conference. In the first summarisation conference, SUMMAC (Mani *et al.*, 1999), query-focused single-document summaries from newswire documents were evaluated. To perform this, two extrinsic evaluation tasks, and an intrinsic one were proposed. Extrinsic evaluation judges the quality of the summarisation based on how it affects the completion of some other task, whereas intrinsic evaluation measures a summary on its own. In the extrinsic evaluation, an *adhoc* task was suggested in which indicative summaries were evaluated with regard to whether they allowed to quickly determine the relevance of a document focused on a specific topic or not. Moreover, a categorisation task was also proposed, whose aim was to determine if generic summaries could effectively present enough information to allow a person to correctly categorise a document. Regarding the intrinsic evaluation task (question-answering task), the goal was to measure the content of a summary with respect to which degree it contained answers to several topic-related questions.

The tasks involved in the TSC workshops within NTCIR conferences (Fukushima & Okumura, 2001), (Okumura *et al.*, 2004), (Hirao *et al.*, 2005) also dealt with the evaluation using intrinsic and extrinsic methods. In the first TSC, three tasks were proposed, where summaries of a specific length were produced. The differences between them were that in the first task, only important sentences had to be extracted whereas in the second, automatic generated summaries were compared to human-made ones. The third task involved extrinsic evaluation, and summaries were evaluated in the context of information retrieval. This task was very similar to the *adhoc* task of SUMMAC conference. Multi-document summarisation was first included in TSC2, in which single-document as well as multi-document tasks were defined. However, since then, multi-document summarisation became a central issue, and consequently, in the following conferences (TSC3), the proposed tasks did no longer address the generation of single-document summaries.

As time goes by, systems evolve and so does the requirements of summarisers, according to the needs of society. The changes of summarisation requirements and systems over the time line can also be seen at DUC conferences. At the beginning, the proposed tasks were aimed at producing generic summaries from a single or several input documents, but at the end, query-focused summaries were paid more attention to. One aspect to remark was the attempt to perform cross-lingual summarisation between English and Chinese in DUC 2004. This task consisted in producing either very short or short summaries in English from a set of documents that were previously automatic translated into English (its original language was Chinese). The concept of novelty and novel information was first addressed in the update task of DUC 2007. The goal of this task was to generate a summary from a cluster of related documents, but taking into account that some of those documents had already been read by users, so the information contained in them, did not need to appear in the summary. Regarding the domain of the documents, the DUC conferences were also focused in newswire documents.

Besides newswire documents, a well-known source of information on the Internet, i.e. blogs, was introduced to be dealt with, in the *Opinion Summarization* task at the TAC 2008 conference[9]. However, due to the difficulty involved in the task itself and type of data (blogs), this task was out of the scope of TAC 2009, keeping only the update summarisation task and introducing a new task concerning the automatic evaluation of summaries. Finally, as it was aforementioned, in TAC 2010, the task concerning evaluation was kept, and generation of guided summaries was introduced as a new task.

Conference	**Number of participant groups**
SUMMAC (1999)	16
TSC (2001)	9
TSC2 (2002)	8
TSC3 (2003)	9
DUC (2001)	15
DUC (2002)	17
DUC (2003)	21
DUC (2004)	22
DUC (2005)	31
DUC (2006)	34
DUC (2007)	32
TAC (2008)	38
TAC (2009)	39
TAC (2010)	32

Table 2.2. Number of participant groups for each conference.

Table 2.2 shows the number of participant groups for each conference edition. It is worth realising how this number has risen over the years, showing the increasing interest in the TS research area. Only in TAC 2010 the number of participants has dropped. The reason may

[9] http://www.nist.gov/tac/tracks/2008/summarisation/index.html

be the introduction of new tasks for summarisation, such as AESOP, which is really challenging.

Furthermore, apart from these conferences, specific workshops focusing only on TS are being organised within important conferences, such as the Workshop on Multi-source, Multilingual Information Extraction and Summarisation[10] (MIMIES), the Workshop on Language Generation and Summarisation[11] (UCNLG+Sum), the Workshop on Web Search Result summarisation and Presentation (WSSP)[12], the 1st International Workshop on Discovering, Summarising and Using Multiple Clusterings[13], the Workshop on Automatic Summarisation for Different Genres, Media and Languages[14] held in conjunction within the Association of Computational Linguistics 2011, or the Workshop on Summarisation organised by the Canadian Conference on Artificial Intelligence[15].

2.2.4 Available Corpora

When attempting to generate summaries automatically, one of the main challenges researchers have to face is the availability of human-written summaries (or model summaries). These are used for assessing the quality of the content selected by automatic summarisers, since most evaluation tools[16] rely on such model summaries to determine to what extent an automatic summary is good. Building such model summaries is not a trivial task. Therefore, it leads to great benefits for the TS research community when specific corpora and datasets are available. For instance, this allows different approaches to be fairly compared.

Next, we are going to provide a brief explanation of some of the best-known summarisation corpora that are available and can be freely used for academic and research purposes. The corpora used at DUC and TAC conferences deal mainly with newswire documents gathered

[10] http://doremi.cs.helsinki.fi/mmies2/

[11] http://www.nltg.brighton.ac.uk/ucnlg/ucnlg09/

[12] http://www.wssp.info/2009.html

[13] http://eecs.oregonstate.edu/research/multiclust/

[14] http://www.summarization2011.org/

[15] https://sites.google.com/site/ts11canai/

[16] Please refer to Chapter 3 for further details in the evaluation of Text Summarisation.

from several press agencies. The model summaries provided are either extracts or abstracts written by humans. These model summaries vary in content, depending on the proposed task in each conference edition (e.g. single-document or multi-document summarisation). Moreover, a new collection of documents pertaining to the *Blog06*[17] was used as corpora for generating summaries. In this case, instead of providing complete model summaries in TAC conferences, humans were asked to select fragments of information that were more relevant to the task, since summaries were evaluated using the Pyramid method[18].

The CAST Project Corpora (Hasler *et al.*, 2003) consist of 163 documents, comprising newswire and articles about popular science. This corpus differs from others in the sense that, apart from containing information about the importance of a sentence in a document, it also indicates which fragments of a sentence can be removed without affecting the sense of the sentence. This fine-grained annotation is of great help for evaluating the conciseness and coherence of the generated summaries.

The *AMI Meeting Corpus* (Carletta *et al.*, 2005) was developed as part of the AMI project[19] and it consists of 100 hours of meeting recordings in English. Although it is not specifically for TS, it can also be adapted for this type of summaries, and it provides abstractive and extractive human-written summaries as well.

The *Multilingual summary evaluation data* (Turchi *et al.*, 2010) is a set of documents related to four topics (genetic, the-conflict-between-Israel-and-Palestina, malaria, and science-and-society). Each cluster contains the same 20 documents in seven languages (Arabic, Czech, English, French, German, Russian and Spanish). In addition, the relevant sentences of each document are manually annotated, and as a consequence, this dataset is very appropriate for evaluating single- or multi-document, as well as multi-lingual extractive summarisation systems.

Also for multi-lingual summarisation, particularly for English and German but in the context of image captioning generation, Aker and Gaizauskas (2010b) developed a corpus of 932 human-written abstractive summaries that describe the most relevant facts of object types

[17] http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html

[18] Please see Chapter 3 for more details about this method.

[19] Project reference: FP6-506811

found in Wikipedia. For instance, given the object *zoo*, model summaries for *Edinburgh Zoo*, or *London Zoo* are provided. The model summaries were collected first for English and then automatically translated to German. In order to assure that the translation was correct, a manual post-editing process was carried out, where the wrong translated sentences were corrected.

The *ESSEX Arabic Summarisation Corpus* (El-Haj *et al.*, 2010) was created using a crowdsourcing service (i.e., Amazon's Mechanical Turk[20]). This corpus includes 153 Arabic articles and 765 human-written extractive summaries.

Table 2.3 provides the source where each of the corpora can be downloaded or requested.

Corpora	**Source**
DUC corpus	http://duc.nist.gov/data.html
TAC corpus	http://www.nist.gov/tac/data/index.html
CAST corpora	http://clg.wlv.ac.uk/projects/CAST/corpus/
AMI Meeting Corpus	http://corpus.amiproject.org/
Multilingual summary evaluation data	http://langtech.jrc.ec.europa.eu/JRC_Resources.html
Image Captioning Corpus	http://staffwww.dcs.shef.ac.uk/people/A.Aker/
ESSEX Arabic Summarisation Corpus	http://privatewww.essex.ac.uk/ melhaj/easc.htm

Table 2.3. Corpora for Text Summarisation.

2.3 Text Summarisation in the Current Context

Text Summarisation (TS) is a dynamic area of research that evolves according to new society requirements and/or users' needs. Therefore, once the most common approaches and techniques with respect to TS have been explained, it is worth explaining how this research area is addressed when new genres, domains and types of summaries appear.

[20] Please see Chapter 3 for more detail about research using crowdsourcing services.

The purpose of this section is to provide an overview of new types of summaries and scenarios into where TS has evolved in recent years. Apart from the classical summarisation types, such as single- or multi-document summaries, generic or query-focused, etc. there are several interesting novel types of summaries, where specific objectives are pursued (e.g. sentiment-based summaries). Consequently, a review of recent types of summaries is provided in Subsection 2.3.1. In addition, the emergence of new scenarios is also interesting for TS. Rather than carrying out research into the same datasets traditionally based on newswire or scientific documents, in recent years, novel domains, such as literary text, patents, or blogs have been paid a lot of attention to. Subsection 2.3.2 addresses these issues.

2.3.1 New Types of Summaries

In the remaining of this subsection, different types of summaries that have recently appeared are going to be described. It is worth stressing upon the fact that the types of summaries next described mostly focus on user's needs, or attempt to efficiently deal with vast amounts of information. Regarding the former, we have selected *personalised*, *update* and *sentiment-based* summaries, since their common goal is to produce a summary, the content of which is determined directly by user requirements (users have to delimit what type of information they are interested in). With respect to the latter, *surveys* and *abstractive* summaries are analysed, because they represent two good examples of summaries that have to be built employing techniques that go beyond the simple concatenation of sentences, and currently it seems that this is the tendency and final goal of TS systems.

Personalised summaries. Their purpose is to provide a summary containing the specific information a user is interested in. This means that different users may have different needs, so that summarisation systems have to determine the user profile before they select the relevant information that will be included in the final summary.

In (Agnihotri *et al.*, 2005) the user profile is created by means of a statistical mapping method from the users' personality traits identified using tests. The analysis performed over 59 users showed that only some traits (e.g. gender) and some features (e.g. text) were of help for personalising the summary. The main drawback of this approach is the limited data it is used for the experimental set-up and due to the difficulty of the task, the experiments in another environment are very hard to replicate. Regarding also personalised summaries, in (Díaz & Gervás, 2007) an approach to produce newswire summaries that contain relevant information for a given user profile is proposed. Their idea is to select those sentences that are the most relevant to a given user model. This is done by calculating the similarity between the user model for a specific individual and each one of the sentences in the document, so depending on which part of the model is chosen to compute the similarity, several possible personalised summaries can be obtained. The user-model is determined by the combination of the specific domain-features, a set of keywords, which reflects the information needs that do not change across the time, and a relevance feedback tier, which takes into account the changes given by users' feedback. The extensive set of experiments carried out showed the appropriateness of this type of summaries, achieving the summaries around 60% for the recall metric.

In (Kumar *et al.*, 2008), personalised summaries are generated based on the area of expertise and personal interests of a user. With this purpose, a user background model is developed taking as a basis the information found on the Internet with regard to a person, such as his/her personal Web page, or on-line publications. Once the user profile has been identified, the relevance of the sentences is determined according to this profile. Two scoring functions, one for generic information and one for user specific are proposed. The first one relies on term frequency to extract the most relevant generic sentences, whereas the second one computes the probability of the generic sentences to contain also user specific information. Then, in the summary generation stage, the top ranked sentences are selected and extracted. Although this approach is very interesting, its main difficulty is that name entity disambiguation should be performed when looking for specific information about a person on the Internet. This would be essential because it may happen that people share the same name but they are totally different (e.g. George Bush could refer either to the US president from 1989 to 1993, or to the US president from 2001 to 2009 – i.e., his son).

In (Berkovsky *et al.*, 2008), a preliminary user evaluation is conducted in order to assess different aspects of users' attitudes towards personalised TS. Three experiments are suggested with the purpose of analysing this issue. Firstly, whether the personalisation of summaries has the desired effect on users or not is evaluated. Then, the impact of summary length is analysed. Finally, the degree of faithfulness between the personalised summaries and the original documents is assessed. The conclusions derived from the analysis are very preliminary. It is shown that the more personalised information a summary contains, the better. It is also claimed that users prefer not too short nor too long summaries; however, no clues about which should be the optimal length are given.

Update summaries. Update summarisation attempts to generate summaries taking into consideration that users have a background knowledge of the topic they want to read about, so they are only interested in the most recent events related to that topic. This type of summaries emerged thanks to the task proposed at DUC and TAC conferences.

In order to generate update summaries, the approach described in (Sweeney *et al.*, 2008) consists in incorporating novelty to summaries, by minimising the content overlap between a summary sentence and a potential candidate one. In (Witte *et al.*, 2007), (Bellemare *et al.*, 2008) and (Li *et al.*, 2009a) update summaries are built on the basis of cluster graph data structures, which are based on the context and on the set of documents that are going to be summarised. A sentence ranking scheme is proposed depending on the overlap between sentences from clusters and the context, so in the end ranks are established and summaries are generated by selecting sentences from each rank. The approach suggested in (Li *et al.*, 2008), defines the concept of history (those documents already known by a reader), and introduces a new type of features (filtering features), which reflect that the summary is summarised with its history. Therefore, in such cases, filtering features can be calculated through two different similarity metrics to exclude those sentences which are similar to the history. One of these similarity metrics is based on the cosine distance formula, whereas the other one takes into consideration unigrams, bigrams and syntactic functions of the words and combines all of them linearly to obtain finally a similarity metric.

Machine learning algorithms are also exploited for generating update summaries. Schilder et al. (2008) rely on FastSum summarisation system (Schilder & Kondadadi, 2008) which uses SVM, but, in this new version, features related to new and old information, such as new/old entities, new/old word/document frequency are also taken into account. Such features penalise sentences that are similar to the previously selected ones. Moreover, in (Fisher *et al.*, 2009), similarity metrics are used as features within a supervised machine learning paradigm, a perceptron ranker, rather than being used to directly rank sentences. Together with these features, a discourse segmentation tool is also employed to determine potential sub-sentential units to be included in the summary as well.

In recent approaches, such as (Liu *et al.*, 2009) and (Nastase *et al.*, 2009), the background information is taken from Wikipedia articles. In the former approach, Wikipedia is used to produce a summary, taking the first paragraph of the entry related to a topic, and then computing the similarity between the potential summary sentences to the ones already contained in the Wikipedia-based summary. The sentences with lower similarity will be selected for the update summary. The latter uses Wikipedia to retrieve concepts that are discussed in the set of documents to summarise. This way, it is possible to predict which concepts are more likely to be found in well-formed summaries. Apart from Wikipedia knowledge, this approach also performs sentence compression to give the final summary an abstract nature. The results shown that Wikipedia is a useful resource to exploit due to the amount of information it contains and the way this information is structured.

Sentiment-based summaries. In recent years, the subjectivity appearing in documents has led to a new emerging type of summaries: sentiment-based summaries, which have to take into consideration the sentiment a person has towards a topic, product, place, service, etc. Consequently, TS and sentiment analysis, also known as opinion mining, have to be combined together in order to produce this type of summaries. Sentiment analysis provides the sentiment associated to a document at different levels (document, fragment, sentence or even word-level) (Pang & Lee, 2008), whereas TS identifies the most relevant parts of a document and build from them a coherent fragment of text (the summary). Regarding sentiment-based summaries, opinions have to be detected and classified first, according to their subjectivity (whether a sentence is objective or subjective, for instance), and then to their polarity (positive, negative or neutral). Further on, TS is in charge of determining which sentences will be included in the summary, thus generating the final summary. Sentiment-based summarisation systems that participated in the *Opinion Summarization Pilot Task* of TAC 2008 conference, such as (Conroy & Schlesinger, 2008), (He *et al.*, 2008a), (Balahur *et al.*, 2008), or (Bossard *et al.*, 2008) follow these steps.

However, out of the scope of the TAC competition, other interesting approaches can be found as well. For instance, in (Beineke *et al.*, 2004) machine learning algorithms are used to determine which sentences should belong to a summary, after identifying possible opinion text spans. The features found to be useful to locate opinion quotations within a text included location within the paragraph and document, and the type of words they contained. Similarly, in (Zhuang *et al.*, 2006) the relevant features (e.g. screenplay, actors for a movie) and opinion words together with their polarity (whether a positive or a negative sentiment) are identified, and then, after identifying all valid feature-opinion pairs, a summary is produced, but focusing only in movie reviews. Normally, on-line reviews contain also numerical ratings that users give when providing a personal opinion about a product or service. The approach described in (Titov & McDonald, 2008) proposed a *Multi-Aspect Sentiment* model. This statistical model uses aspect ratings to discover the corresponding topics and extract fragments of text. Moreover, in (Lerman & McDonald, 2009), an approach to produce contrastive summaries in the consumer reviews domain is suggested. Contrastive summarisation refers to the problem of generating a summary for two entities in order to highlight their differences, for example, different people's sentiments about several products. In order to produce this type of summaries, they adapt the *Sentiment Aspect Match* model described in (Lerman *et al.*, 2009), originally designed to generate single product sentiment-based summaries. This model determines which sentences to extract comparing the average sentiment of a sentence with respect to the average sentiment of the specific entity, thus selecting the closest ones.

Survey summaries. The goal of this type of summaries is to provide a general overview of a particular topic or entity. They are generally long rather than short, because they attempt to capture the most important facts concerning a person, for instance. Next, we focus on biographical summaries, survey summaries and Wikipedia articles as three of the most recently TS types that can be categorised within this group.

The challenge of producing summaries from biographies was presented in (Zhou *et al.*, 2004). The idea behind multi-document biography summarisation is to produce a piece of text containing the most relevant aspects of a specific person, answering questions, such as *"Who is Barack Obama?"*. To accomplish this task, several machine learning algorithms are used to classify sentences (*Naïve Bayes*, SVM, and *Decision Trees*). Moreover, redundant information is removed in a later stage. A similar biographical summarisation system, which also employs machine learning techniques, is described in (Biadsy *et al.*, 2008). The difference with the aforementioned one is that a binary classifier is used to discriminate between biographical and non-biographical sentences, and then a SVM regression model is trained to reorder biographical sentences extracted using Wikipedia as a corpus. The final stage of this approach is to employ a rewriting heuristic to create the final summaries.

Another interesting approach is to use citations from articles. In (Kan *et al.*, 2002) it was shown that from bibliographical entries it was possible to produce an indicative summary. The main idea behind this assumption is that such entries contain informative as well as indicative information. For example, details about the resource or metadata, such as the author or purpose of the paper. In their research, a big annotated corpus (2,000 annotated entries) is developed. Following the idea of generating summaries from this type of input information, in (Qazvinian & Radev, 2008) citations are analysed to produce a single-document summary from scientific articles. The final objective is to generate summaries about a specific topic. Also, the work described in (Mohammad *et al.*, 2009) addresses this topic and consequently presents some preliminary experiments of the usefulness of citation text to automatically generate technical surveys. Three kinds of input are used (full papers, abstracts and citation texts), and already existing summarisation systems are taken into consideration to create such surveys, for instance LexRank (Erkan & Radev, 2004), Trimmer (Zajic *et al.*, 2007), and C-RR and C-LexRank (Qazvinian & Radev, 2008). Among the conclusions drawn from the experiments, it was shown that multi-document technical survey creation benefits considerably from citation texts.

Different from these approaches, Sauper and Barzilay (2009) suggest the automatic creation of Wikipedia articles using domain-specific templates which are induced from human-generated documents. For producing such articles, a search engine is employed to retrieve documents related to a topic, which are considered as the input of the summarisation process. Following the same structure of a Wikipedia article, the appropriate information for each section is determined by machine learning techniques, training excerpts based on how representative they are to a selected topic.

Abstractive summaries. To simplify the problem of summarisation, most approaches follow an extractive paradigm, by outputting the most relevant sentences of a document/s without doing any changes. Although it is a widespread method, the resulting summaries often present several problems with respect to their quality, such as the lack of coherence or *"dangling anaphor"*. The abstractive paradigm can solve these limitations, since it attempts to produce new text from the fragments of information or concepts identified as relevant. Despite not being a novel issue, in recent years, research in abstraction has been fuelled, due to the fact that information is repeated across documents, and specific ways of conveying and presenting such information are required.

In addition, several analysis have been conducted to understand how humans summarise (Jing & McKeown, 2000), (Jing, 2002). As a consequence, the basic operations to transform source information into summary information are analysed. For instance, (Hasler, 2007) claims that the technique humans often do is to copy and paste the same material present in the source documents. However, some slightly changes are applied in most of the cases, and two types of operations, atomic and complex, are identified, involving deletion, insertion, replacement, reordering or merging (the first two are atomic operations while the last three are complex). From the evaluation carried out in terms of coherence, the results showed that 78% of the abstracts were more coherent than extracts. In (Fiszman *et al.*, 2004) an approach to generate abstracts from biomedical documents is proposed.

The main idea is to identify semantic predicates using SemRep[21] and then produce the summary in a schematic way. The summarisation process comprises the identification of such predicates, and the connectivity between them. Further on, the novelty and the salience of each predication is computed based on term frequency counts. Saggion (2009) suggests a novel approach to combine different fragments of information that have been extracted from one or more documents. From a predefined vocabulary (e.g. *to address*, *to indicate*, *to report*, etc.) the algorithm is able to decide which of these expressions is more appropriate for a sentence, depending on the content and the partial abstract generated. The motivation under this research is to study to what extent the addition of extra information not present in the source documents is useful and benefits the abstraction process. Using machine learning techniques and experimenting with different types of classifiers (e.g. *Decision Trees*), results showed that the best classifier was able to correctly predict 60% of the cases. This classifier was based on summarisation features, including linguistic, semantic, cohesive, discourse or positional information.

Furthermore, sentence compression (Zajic *et al.*, 2007), (Clarke & Lapata, 2006), and sentence fusion (Barzilay & McKeown, 2005b), (Marsi & Krahmer, 2005) are also quite employed when attempting to generate abstractive summaries. In particular, graph-based algorithms used for such purpose have been proven to be very successful (Filippova & Strube, 2008), (Filippova, 2010). Regarding sentence fusion, in (Filippova & Strube, 2008) related sentences are represented by means of dependency graphs, and then the nodes of such graphs are aligned taking into account their structure. Then, Integer Linear Programming (Schrijver, 1986) is used to generate a new sentence, where irrelevant edges of the graphs are removed, and an optimal sub-tree is found employing structural, syntactic and semantic constraints. For sentence compression, Filippova (2010) suggests a method based on word graphs, where the shortest path is computed to obtain a very short summary (only one sentence) from a set of related sentences. In (Liu & Liu, 2009), the authors attempt to transform an extractive summary into an abstractive one in the context of meeting summarisation by performing sentence compression.

[21] http://skr.nlm.nih.gov/papers/index.shtml#SemRep

Natural Language Generation (NLG) is also applied for producing abstractive summaries. In (Yu *et al.*, 2007) very short summaries are produced from large collections of numerical data. The data is presented in the form of tables, and new text is generated for describing the facts that such data represent. Firstly, the data has to be analysed, and understood before generating the descriptions, and in the last step a NLG module is used, which specifically accounts for three types of information to generate: background information, overall description, and most significant patterns found in the data collection. Belz (2008) also proposed a TS approach based on NLG to generate weather forecasts automatically, but focusing mainly on the NLG stage.

Other abstractive approaches rely on the use of templates to structure the information that has been previously identified, for instance using an information extraction system. In (Kumar *et al.*, 2009) an attempt to generate reports from the event information stored in databases from different domains (biomedical, sports, etc.) is presented. Human-written abstracts are used to determine the information to include in a summary, where some templates are generated and patterns to fill in such templates are identified in the texts. Similarly, in (Carenini & Cheung, 2008) patterns are also identified, but since the aim is to generate contrastive summaries, discourse markers indicating contrast such as *"although"*, *"however"*, etc. are also added to make the summary sound more naturally.

2.3.2 New Scenarios for Text Summarisation

Although most of the research work in TS has traditionally been focused on newswire (Gotti *et al.*, 2007), (Nenkova *et al.*, 2005), (Nenkova, 2005), scientific documents (Jaoua & Hamadou, 2003), (Teufel & Moens, 2002), or even legal documents (Saravanan *et al.*, 2006), (Cesarano *et al.*, 2007), these are not the unique scenarios in which TS approaches have been tested on. Next, several new scenarios in which TS has been also applied are described. From the analysis of TS in such scenarios, it is worth stressing upon the fact that, although the nature of the documents is totally different across domains, and it may seem that each domain would need a different manner to tackle the TS process, in practice, the techniques employed do not experiment great changes. It can be seen that statistical or positional features are the preferred ones. In some cases, specific vocabulary is added, but generally minor changes are performed to adapt TS approaches to other scenarios. Hence, the following discussion may arise: *"Would it be better to develop generic systems for a wide range of scenarios, although with moderate performance, or to build very specific systems that could obtain a higher performance?"*. Moreover, since some features, such as term frequency or inverse-document frequency can be considered domain-independent, an appropriate approach could be to combine this type of features with domain-specific ones within the same TS process. This would allow that, for each domain, the process could benefit from issues such as specific vocabulary or the structure of the documents, thus increasing its performance in the specific scenario with respect to a generic TS on their own.

In the remaining of this subsection, TS generated in the context of literary texts, patent claims, image captioning, and the new born Web 2.0 textual genres is explained.

Literary text. Attempts to summarise literary texts, either short stories (Kazantseva, 2006) or longer texts, i.e. books (Mihalcea & Ceylan, 2007) have also been addressed in recent years. In (Mihalcea & Ceylan, 2007), the difficulties of TS when addressing book summarisation are analysed, building a benchmark, where the evaluation of book's summaries is specifically targeted. Moreover, several techniques for book summarisation are suggested as well, for instance text segmentation, suggesting a summarisation approach based on the already existing system MEAD (Radev *et al.*, 2001), with some particular changes, in order to adapt the system to long-document summarisation. Further on, in (Ceylan & Mihalcea, 2009), two kinds of summaries are generated: objective and interpretative summaries. The former contains information about the events occurring in the books and their plot, whereas the latter attempts to capture the author's ideas and thoughts. From this analysis, it is found that approximately 48% of the objective summaries can be reconstructed by cut-and-paste operations from the original document. However, for interpretative summaries, this number decreases to only 25%.

For short stories (Kazantseva, 2006), indicative summaries are generated in order to help the user to decide whether to read or not the whole document. The relevant information to include is determined based on linguistic traits, such as grammatical tenses, temporal expressions, voice, meaning of the verb, and type of speech (direct or indirect).

Patent claims. The particular writing style of patents, although difficult to process due to the kind of language employed, has been also targeted for TS. An interesting approach performing at approximately 60% (F-measure) can be found in (Mille & Wanner, 2008) where a multi-lingual summarisation system for Spanish, English and French is developed taking advantage of the structure of the patent claims and employing discourse and semantic features as well as dependency patterns for the summarisation process. They also perform linguistic simplification in order to give the resulting summaries an abstractive nature. A complete text mining approach for patent analysis, including a TS stage is proposed in (Tseng *et al.*, 2007). With respect to TS, the extractive process ranks sentences based on the frequency of keywords, similarity to the title of the patent claim, and the cue words it contains. Also, positional features are considered. Finally, all these features are combined in a linear way and the highest weighted sentences, up to a desired length, are selected to form the final summary. Key phrases are also identified and used as features for determining the relevance of a sentence in (Trappey & Trappey, 2008). Moreover, clustering techniques are employed to obtain the information density of a sentence. However, different from the previous approaches, the novelty of this approach relies on incorporating domain-specific features based on phrases and topic sentences for a given patent document. Trappey et al. (2009) extend the previous work with ontological knowledge in order to retrieve the domain-specific keywords and phrases using concept hierarchies and semantic relationships. Results are evaluated in terms of compression and retention ratios. The compression ratio is the ratio of the word counts between the summary and its original document, whereas the retention ratio indicates the average value of the recall ratio and the precision ratio. Results show that the best compression ratio is 20%, which is line of state-of-the-art (Morris *et al.*, 1992). In addition, it is proven that the use of ontologies improves the retention ration.

Image captioning. The need of producing short descriptions of images can also be seen as a TS problem, where a summary is produced from a set of related documents referring to an image annotated with geographical information. In (Deschacht & Moens, 2007) image captions using the associated information related to an image are produced. They benefit from the immediate context of the image to extract such information, for instance, text in HTML tags. Their main purpose is to correctly detect and classify entities appearing in images, and then, calculate the salience of such entity with the final goal to produce a short annotation for the image. Similarly, in (Feng & Lapata, 2008) an image annotation model which is able to learn image captions from auxiliary documents and noisy annotations is suggested. The auxiliary documents are very useful for this task since they can provide important information related to the image, thus allowing to generate more accurate image descriptions.

Aker and Gaizauskas (2009) propose an approach based on language models (n-grams) to generate 250 word-length summaries for image captions using the corpora described in (Aker & Gaizauskas, 2010b). The results obtained are very encouraging, being later improved by means of dependency patterns (Aker & Gaizauskas, 2010a), which performed very close to the ones obtained using the first paragraph of Wikipedia articles as summary baseline. Furthermore, in (Plaza *et al.*, 2010), two TS approaches based on statistical features (term frequency and noun-phrase length) and semantic-graphs using WordNet concepts, are also tested within the same corpus. Results for both approaches were acceptable, obtaining around 10% in recall according to ROUGE-SU4 metric, and improving the language models approach originally proposed in (Aker & Gaizauskas, 2009).

Web 2.0 textual genres. Summaries of new textual genres, such as blogs (Balahur-Dobrescu *et al.*, 2009), (Lloret *et al.*, 2009), reviews (Balahur & Montoyo, 2008b), (Zhuang *et al.*, 2006) or threads (Zajic *et al.*, 2008), (Balahur *et al.*, 2009) can also be found in the literature. The summarisation techniques used within these approaches are in the line of the ones presented in Section 2.3.1 (sentiment-based summaries), being the integration of sentiment analysis techniques essential for generating summaries from these new genres born with the social Web. Focusing on TS, in (Balahur *et al.*, 2009), the techniques employed are based on term frequency counts, whereas in (Balahur-Dobrescu *et al.*, 2009) summaries are generated using Latent Semantic

Analysis[22]. Other approaches, such as the one proposed in (Zhuang *et al.*, 2006) simply rely on the output of a sentiment analysis system to group sentences according to their polarity without taking into consideration any TS technique.

2.4 Combining Text Summarisation with HLT Applications

The goal of this section is to present how summaries can help other systems, therefore analysing the applicability of TS within other intelligent systems. This can be considered as a manner of indirectly evaluating summaries, also known as extrinsic evaluation. Summaries can be a good way to allow systems to spend less processing time, if they are used instead of the whole document. Moreover, at the same time, summaries can be suitable for removing noisy information, thus keeping only the really important one. These aspects are derived from the definition of a summary itself, where a summary is a brief but accurate representation of the contents of a document or a set of them. In light of this, summaries can help both applications and users to save time. In this section, we are interested in carrying out a deep analysis of the usefulness of TS for other HLT applications, such as the ones explained next. In particular, we focus on *information retrieval* (Subsection 2.4.1), *question answering* (Subsection 2.4.2) and *text classification* (Subsection 2.4.3).

2.4.1 Text Summarisation with Information Retrieval

The goal of Information Retrieval (IR) is to find material (usually documents) of an unstructured nature (usually text) that satisfies an information need within large collections (usually stored on computers) (Manning *et al.*, 2008). TS has been combined with IR from a double perspective. Several approaches use summaries to benefit IR, for example at the indexing stage, improving the time to retrieve documents, as well as its performance, whilst others take as the input

[22] Latent Semantic Analysis is a technique that is used to analyse relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

for TS the output of IR systems (i.e. the documents retrieved by the IR system). Moreover, in some cases, instead of the traditional snippets provided by IR systems, summaries are presented as output. For instance, in (Kan & Klavans, 2002), summaries are employed as an alternative visualisation of the documents coming from a standard IR framework. Moreover, the optimal length that a summary should have in order to be useful for users is analysed in (Kaisser *et al.*, 2008), concluding that the preferred length by the users depends on the type of the query. However, the most common approach is to combine IR and TS in the following manner: the documents related to a topic are retrieved first, and then, a summary taking into account these documents is generated. Therefore, IR helps to gather only relevant documents to a query, while TS selects the most important information from them. Radev and Fan (2000) propose a domain-independent multi-document summariser that generates summaries from Web search results. Similarly, SWEeT (Steinberger *et al.*, 2008) relies on a search engine to retrieve relevant documents to a query from the Web, and then summarisation techniques based on Latent Semantic Analysis are used to identify and extract the most important sentences from the retrieved documents using, at the same time, cosine similarity to avoid redundancy in the final summaries. The QCS system (Dunlavy *et al.*, 2007) also integrates a IR module but, instead of retrieving documents directly from the Internet, it does so from a static document collection. Once the relevant documents have been retrieved, the system clusters them according to their main topic, and finally a summary is produced for each cluster. The summarisation process is performed in two steps. Firstly, a single-document summary is generated for each document cluster, and then those extracted summary sentences are taken into account to produce the final summary. The way sentences are selected to become part of the summaries is by using *Hidden Markov Models*, computing the probability of a sentence with regard to whether it is a good summary sentence or not.

Less research has been carried out to analyse how text summaries can be beneficial for the IR process. In (Sakai & Spärck Jones, 2001) it was proven that generic summaries with a compression rate ranging from 10% to 30% were the most appropriate ones for the indexing stage in IR, concluding that a summary index was as effective as the full text index, for precision-oriented search. In (Szlávik *et al.*, 2006), whether or not summarisation is useful in interactive XML retrieval is investigated, thus providing summaries from XML elements in order to allow users to browse and judge XML documents more easily.

2.4.2 Text Summarisation with Question Answering

The objective of Question Answering (QA) is to automatically answer simple or complex questions posed in natural language (Strzalkowski & Harabagiu, 2007). Specific research where different TS approaches have been integrated into a QA system can be found in the literature. The approach suggested in (Mori *et al.*, 2004) analyses the effectiveness of topic signatures in the multi-document QA summarisation context for a particular type of questions. The generated summaries were about people and contained the answer to the question *"Who is X?"*, where X is a person. It was found that, although topic signatures were able to capture information emphasised in the corresponding source texts, this was not sufficient, as human-written summaries also contained some details that were mentioned, despite not being emphasised.

An interesting approach to QA is presented in (Demner Fushman & Lin, 2006), where TS and IR techniques are combined to provide answers to questions belonging to the medical domain. Questions like *"What is the best drug treatment for X?"* are tackled by identifying the drugs from a set of citations first, and then clustering the corresponding abstracts, so that a short extractive summary can be produced for each of them. The summaries are generated by outputting the title of the abstract, the main intervention, and the top-scoring sentence, which is determined using supervised machine learning techniques. Also in the medical domain, the BioSquash system (Shi *et al.*, 2007) summarises multiple biomedical documents answering a specific question. The system, based on a generic summariser, has four main components. The *Annotator* module annotates the documents and the question with syntactic and shallow semantic information, and then the relations between concepts in the documents and the questions are determined by the *Concept Similarity* module. The remaining modules, the *Extractor* and *Editor* modules, focus on content selection and linguistic readability, respectively. The final result is a fluent summary relevant to a question concerning the biomedical domain.

Finally, the QAAS system (Torres-Moreno *et al.*, 2009) has resulted from the integration of a TS system with a QA system. In this approach, a generic multi-document summariser of several compression rates is coupled with a QA system, thus allowing the document search space to be reduced, compared to the whole document. The results obtained show that the number of correct answers returned by the combined system increase. The generic summariser is also used to identify informative textual zones in documents. However, the limitations of using generic summaries for QA are identified, and for this reason, the generic summarisation system is adapted to a query-focused one by doing query expansion and re-scoring the sentences selected by the generic summariser according to the terms of the question.

Apart from these approaches, the inverse combination, QA applied to TS can also be found in the literature. In (Mori *et al.*, 2005), a QA engine is used to help determine sentence importance, which is calculated based on the scores produced by the QA system and a set of queries. The final objective is to generate a summary, and for this purpose, the QA engine is finally integrated into a generic multi-document summariser.

2.4.3 Text Summarisation with Text Classification

Text Classification (TC), also known as text categorisation, aims at automatically sorting a set of documents into categories from a predefined set (Sebastiani, 2002). In (Ker & Chen, 2000), summarisation features (i.e. position and word frequency) are used to categorise news according to different categories (e.g. "money") reaching 82% for the precision value. Taking text summaries instead of full documents is the approach suggested in (Shen *et al.*, 2004), under the assumption that they may be a good noise filter. Since Web pages contain too much irrelevant information which can be detrimental for TC, summaries can extract the most important information, producing a new text, which is then used for classification purposes. The TS approaches used are based on term frequency and Latent Semantic Analysis. They carry out a large experimentation with more than 150,000 Web pages and 64 categories. The results obtained show that the proposed summarisation-based classification algorithm improves approximately 8.8% compared to the full documents. The same idea and dataset is analysed in (Shen *et al.*, 2007) where, in addition, the optimal compression rate for summaries is studied. Summaries of 20% and 30% reach the best results.

However in the range from 10% to 40%, it is proved that text summaries can improve the classification performance to some extent over the full documents.

The rating-inference task can be seen as a particular type of TC. Its goal is to identify the author's evaluation of an entity, product, service, etc. with respect to an ordinal-scale based on his/her textual evaluation of the entity (Pang & Lee, 2005). Therefore, it can be considered as an opinion classification problem. Usually, opinions are classified with regard to two or three dimensions, subjective vs. objective, or positive vs. neutral vs. negative, respectively (Wilson *et al.*, 2005). However, it is frequent to find texts where users give a score, depending on how much they liked or not a product, movie, restaurant, hotel, service, etc. which is normally associated with a rating scale (1=worst,...5=best). When tackling this task with short documents (reviews containing at most three sentences), the classification process achieves good results (74%) (Saggion & Funk, 2009), whilst it only reaches 32% when dealing with longer texts. As a consequence, and taking as a basis the aforementioned assumption concerning the suitability of TS for filtering noise, in (Lloret *et al.*, 2010) and (Saggion *et al.*, 2010), a preliminary set of experiments is carried out for predicting the rating of a review using text summaries instead of full documents. These experiments comprise the analysis of a wide range of summarisation types of different compression rates. In particular, generic, query-focused and sentiment-based summaries are studied, together with several types of baselines, including the first and the last sentences of a review[23]. Although it is claimed that query-focused and sentiment-based summaries may be more appropriate for the rating inference task, this analysis has two main limitations. On the one hand, this task is very complex, since they work at a very fine-grained granularity, and for instance, the differences between a text rated as 4 from another one rated as 5 can be very subtle. On the other hand, the dataset used is very small (only 89 reviews) to obtain strong evidence on what type of summaries could be more appropriate.

[23] Please see Chapter 6, Section 6.4 for more details about this research.

2.5 Overall Discussion

This chapter presented the state of the art in TS, focusing especially in the new types of summaries and scenarios raised in the last years. In order to provide some basic background information about this research area, a brief review of well-known summarisation approaches was also conducted, grouping these approaches into different groups as far as the techniques or algorithms employed was concerned. Moreover, international forums and conferences with regard to TS that have been taken place along the years were described, in order to provide a historical perspective of this research field, highlighting how it has evolved, and the attention that the research community has paid to it. It is very interesting to see the impact of TS in the community research, since it can be concluded that, despite its age and difficulty, it attracts more and more researchers. Additionally, available corpora specific for TS was described. Also, we provided a review of new scenarios and types of summaries, and finally the combination of TS with other HLT-based systems, in particular, IR, QA or TC was also analysed, since text summaries can improve the performance of other applications, being very appropriate to use them in combination with summaries.

The analysis conducted in this chapter allows us to have a basic background about the past of TS, the current state of the art, and possible trends for the future. As far as the TS approaches is concerned, it is worth mentioning that over the years, existing approaches are changing. For instance, new machine learning algorithms are proposed for tackling TS; however, the features used do not change too much with respect to the ones already existing (e.g. term frequency, part-of-speech, sentence position). What it seems to be changing fast over the years is the types of summaries, as society has to adapt to new user requirements. Whereas at the beginning generic and single-document summarisation were two of the most important types, currently multi-document and even multi-lingual summarisation have gained great importance due to the vast amount of information we have to deal with in different languages. This can also be seen in the international forums devoted to TS, where, year after year, the proposed tracks are updated.

From the TS approaches described in this chapter, some conclusions about the tendencies of TS for the future can be drawn. As we previously said, the society requirements change, and the information grows at an exponential rate, which forces TS to adapt to new necessities. For the next years, multi-document and multi-lingual summarisation will be essential, since the same information can appear in a high number of documents but also in different languages. It is worth mentioning that this information has to be presented in a coherent way, which forces systems to advance beyond the concatenation of sentences. Therefore, abstractive paradigms or at least hybrid ones will become one of the main challenges to face. Hybrid approaches would be capable of identifying and selecting relevant fragments of information, and then merge, compress or delete such information in order to generate new summary information. As a consequence, it is possible to take into account the benefits of extractive and abstractive approaches together. Moreover, sentiment-based summaries, personalised and update summaries will be also very important, because users play a crucial role on the Internet and therefore, a summary should provide the exact information they require. The presentation of such information is another issue that is becoming more and more important in the sense that, traditionally, the input and the output of a TS system is text. However, this tendency is changing and we can find many approaches summarising other types of input, such as meetings, or video, or even producing the output in a format different from text. For instance, this would consist in taking text as input, but presenting the summary in another format such as by means of statistics, tables, graphics or visual rating-scales. This would allow users to visualise the results immediately, and find the information they are interested in more quickly. Besides, these visual representations could be also complemented by text summaries.

Despite having more that 50 years old, TS is still alive with a great interest among the research community. Indeed, this research field is very dynamic, since it is continuously adapting itself to the new needs and challenges. Although the performance of TS is still moderate and the generated summaries are far from perfect, it can be seen how the combination with other systems leads to the improvement of the overall performance of the combined system, helping in the development of more intelligent systems. All the possibilities that TS offers together with the extensive application it has in the real world make it an interesting research area to conduct research into. Therefore, the acquisition of general background to TS, which was the main goal of this chapter, is essential.

3. Text Summarisation Evaluation

3.1 Introduction

A lot of research effort has been made to suggest novel approaches capable of generating good summaries using a wide range of techniques, for instance statistical features (Teng *et al.*, 2008), graph-based algorithms (Plaza *et al.*, 2008), or machine learning techniques (Kumar *et al.*, 2009). Furthermore, advances have been partially supported by well-known international evaluation campaigns, such as DUC[1], NT-CIR[2], or TAC[3]. However, research concerning the evaluation of generated summaries is less developed. *How can a summary be assessed? Which aspects have to be taken into consideration to determine the quality of a summary? How can a good summary be distinguished from a bad one? Are there methods and tools that allow us to do this automatically?* These questions are very difficult to answer even by humans. The inherent subjectivity associated to what different humans may think about what a good summary is, may differ greatly. This fact has been proven in (Donaway *et al.*, 2000), (Mani, 2001b), where the agreement among human judges with respect to the evaluation of summaries showed a great variability.

Automatic evaluation of summaries is, indeed, a very challenging task. This is reflected in the advances achieved in this particular area, which are slower than in other tasks also based on Human Language Technologies (HLT). Whereas many researchers focus their attention on which are the best approaches for identifying the most relevant sentences in a document, methods accounting for the quality of a summary

[1] http://duc.nist.gov/
[2] http://research.nii.ac.jp/ntcir/outline/prop-en.html
[3] http://www.nist.gov/tac/

are not that much exploited, although in recent years there has been a surge interest in this issue.

Therefore, the objective of this chapter is to provide an overview of Text Summarisation (TS) evaluation, describing the main types, methods and existing resources that contribute to automate this process. This will allow us to envisage the necessities for the future concerning the automatic evaluation of summaries.

The structure of the chapter is as follows. In Section 3.2 the two broad types for evaluating any HLT application are introduced. Then, Section 3.3 provides an overview of the current evaluation methods for TS, distinguishing between those that focus on the informativeness of the summaries (Subsection 3.3.1), from the ones addressing their quality (Subsection 3.3.2). Moreover, the use of crowdsourcing services specifically being used for the evaluation of summaries are explained in Subsection 3.4. Finally, this chapter is concluded in Section 3.5, where some insights about the tendencies of TS evaluation for the nearest future are provided.

3.2 Types of Evaluation

Methods for evaluating systems based on HLT can be broadly classified into two categories: intrinsic and extrinsic (Spärck Jones & Galliers, 1996). In the context of TS, the former evaluates a summary itself, for example according to its information content, whereas the latter tests the effectiveness of a TS system on other HLT application (e.g. question answering). In this case, the summary would be beneficial for other applications, despite not being appropriate for serving as a surrogate of a document to be directly used by humans. Therefore, it is worth stressing upon the fact that evaluating a summary intrinsically is more challenging than carrying out an extrinsic evaluation.

As far as intrinsic evaluation is concerned, there are different criteria that can be taken into account to evaluate a summary. According to (Mani, 2001a), it can be distinguished between evaluating the informativeness or the quality of a summary. Apart from these two, another one is proposed, fidelity to the source, which determines summary informativeness in the context of the source document, that is, if the summary contains the same or similar relevant concepts as the source document. The most widespread intrinsic methodologies focus on evaluating the informativeness of a summary by comparing its content to a human-written one, also known as model summary or gold standard. However, due to the inherent subjectivity associated to summaries, it is very difficult to build a fair model summary. In (Donaway *et al.*, 2000) and (Mani, 2001b) it was shown how the recall value varied, depending on which human summary was taken as model for comparison with the automatic one. In contrast, other evaluation approaches are more concerned with a qualitative evaluation, which aims at evaluating the quality of a summary with respect to different criteria, such as grammaticality or coherence.

With respect to the extrinsic evaluation methods, several scenarios have been proposed as methods for TS evaluation inspired by different disciplines (Hassel, 2007). Examples of these scenarios are: *The Shannon Game*, which aims at quantifying information content by guessing tokens, so that the original document can be recreated; *The Question Game*, which tests the readers' understanding of the summary and his/her ability to convey the main concepts; *The Classification Game*, which consists of determining the category either for original documents or for summaries, by measuring the correspondence between them; and *Keyword Association*, in which a list of keywords is provided and the goal of the task is to check whether summaries contain such words or not. Also, in (Mani, 2001a) different extrinsic evaluation methods are outlined. For instance, *relevance assessment*, in which subjects are asked to determine the relevance of a topic, either in a summary or source document; or *reading comprehension*, which involves answering multiple-choice tests having read the summary or the whole document.

Moreover, as it was analysed in Chapter 2, TS has been successfully proven to help other HLT applications, such as information retrieval (Sakai & Spärck Jones, 2001), (Szlávik *et al.*, 2006); question answering (Shi *et al.*, 2007), (Torres-Moreno *et al.*, 2009), or text classification (Shen *et al.*, 2007), (Saggion & Funk, 2009). In these approaches, summaries are employed to improve the capabilities of such applications, and although a summary could be imperfect in its nature, it could be successfully used to improve their accuracy.

3.3 Review of the Evaluation Methods

As it was previously introduced, a summary can be evaluated either intrinsically or extrinsically. Due to the fact that intrinsic evaluation is more challenging and it has gained great attention in the last years, this chapter will only focus on this type of evaluation, describing the most widespread approaches capable of assessing the content of a summary with respect to the information it contains (Subsection 3.3.1) and its quality (Subsection 3.3.2).

3.3.1 Informativeness Assessment

According to Hovy (2005), a summary must fulfil two basic requirements: its length must be shorter than the source document and it must contain the most important information. These issues can be captured by means of the **compression** and **retention ratios**. The former computes the length of the summary with respect to the whole document, whereas the latter determines how much information is kept in the summary. Apart from these measures, the informativeness of a summary has been commonly assessed employing **recall**, **precision** and **F-measure** (Van Rijsbergen, 1981), adapting these metrics for the context of TS. In this way, an automatic summary (peer summary) is compared to a human-written one, and the common sentences between them are measured. Following Nenkova (2006), recall evaluates which portion of the sentences selected by a human are also identified by a summarisation system, whereas precision is the fraction of these sentences identified by the summarisation system that are correct. F-measure is a combination of both precision and recall. However, by using these metrics, it is possible that two equally valid summaries are judged very differently. This would happen in the cases where summary sentences do not match with the sentences contained in the model summary.

Relative Utility (Radev & Tam, 2003) was then proposed to overcome this shortcoming. This method allows multiple judges to rank each sentence in the source document with a score, giving it a value ranging from 0 to 10, which determines its suitability for a summary. Therefore, the higher the sentence is ranked, the more suitable for a summary is. Further on, these weights are used to score each sentence in the summary. **Factoid score** (Teufel & Halteren, 2004) is another proposed metric to identify atomic information units which represent the meaning of a sentence (i.e., factoids). The idea is to use several model summaries as gold standard and measure the information overlap among them, identifying the associated factoids and assigning them a weight based on the degree of agreement found. After that, an automatic summary is evaluated with respect to the number of factoids it contains, and their associated weights are employed to score the summary. The **Pyramid method** (Nenkova *et al.*, 2007) follows a similar philosophy. Its goal is to identify information with the same meaning across different model summaries, called *Summary Content Units* (SCUs). Each SCU has a weight depending on the number of human assessors who expressed the same information, allowing important content to be distinguished form less important one. However, the main drawback of these methods – Relative Utility, factoids, or the Pyramid – is the fact that human annotations are needed in order to identify important content, thus resulting in a time-consuming and a very laborious task. In order to avoid the laborious task to manually match fragments of text in the peer summaries to the SCUs in the Pyramid method, an attempt to automatically perform this part of the process is suggested in (Fuentes *et al.*, 2005). However, the manual effort needed to detect factoids or SCUs along a collection of model summaries, is still one main disadvantage.

QARLA, an evaluation framework proposed in (Amigó *et al.*, 2005) was developed under the assumption that the best similarity metric to assess a summary should be the one that best discriminates between manual and automatically generated summaries. Having a set of model summaries, a set of peer ones, and a wide range of similarity metrics, this framework provides the following types of measures: QUEEN, which is an estimation of the quality of an automatic summary; KING, which estimates the quality of a similarity metric; and finally, JACK, which is used to indicate the reliability of the automatic summaries set. QARLA provides a total number of 59 different similarity metrics, including for instance recall, precision, sentence length, or frequency and grammatical distribution metrics.

Relying on the vocabulary overlap (n-grams) between the peer and model summary for deciding the goodness of the former, **ROUGE** (Lin, 2004) is a tool to automatically evaluate a summary. Its name stands for *Recall-Oriented Understudy for Gisting Evaluation* and it was inspired by BLEU (Papineni *et al.*, 2002), which is a method for automatically evaluating the output of a machine translation system. The hypothesis of ROUGE is that if two texts have a similar meaning, they must also share similar words or phrases. As a consequence, it relies on n-gram co-occurrence, and the idea behind it is to compare the content of a peer summary with one or more model summaries, and compute the number of n-gram of words they all have in common. Different types of n-grams can be obtained, such as unigrams (ROUGE-1), bigrams (ROUGE-2), the longest common subsequence (ROUGE-L), or bigrams with a maximum distance of four words in-between (ROUGE-SU4), and based on them, values for recall, precision and F-measure in the most recent version of ROUGE (ROUGE-1.5.5) are finally obtained, thus determining how good the summary is (the higher recall, precision and F-measure values, the better). This tool was shown to have a high correlation with human evaluation (around 80% and 90% depending on the types of n-grams tested); however, this tool has some drawbacks in the way it assesses summaries. In (Sjöbergh, 2007), it was shown that a very poor summary could easily get high ROUGE scores. In order to prove this claim, a simple summarisation method was developed, using a greedy word selection strategy. Although the generated summaries were not good from a human's point of view, they obtained good results for some ROUGE values, for example a recall score of 41% for ROUGE-1, which is acceptable in the state-of-the-art in TS. In addition, the correlation between ROUGE and model summaries was shown to be lower than it was claimed, especially in other types of summarisation, for instance in speech summarisation (Liu & Liu, 2008). Another drawback of ROUGE is that model summaries are needed beforehand. In order to overcome with the difficulty of obtaining a set of model summaries, He et al. (2008b) suggest an alternative method based on ROUGE (**ROUGE-C**) that allows to evaluate a summary comparing it directly to the source document, given that some query-focused information is also provided.

Moreover, in order to address the shortcomings derived from comparing fixed n-gram words, **AutoSummENG** (Giannakopoulos *et al.*, 2008b) is another automatic n-gram based method recently developed which has been proven to have higher correlation with human judgements than ROUGE. This method differs from the other n-gram methods in three main aspects: (1) the type of statistical information extracted (n-gram characters); (2) the representation chosen for this extracted information (graph), and (3) the method used for calculating the similarity between summaries. Here, the comparison between summaries is carried out by building n-gram character graphs first, and then comparing their representations in order to establish a degree of similarity between the graphs. Moreover, its methodology is language-independent, so it works in other languages as well.

Other evaluation methods rely on dependency parsing for representing the information in peer and model summaries with the purpose of being more flexible when comparing the common information contained in both. **Basic Elements** (BE) (Hovy *et al.*, 2006) was proposed as a new evaluation methodology. The underlying idea of this method is to split a sentence into very small units of content in order to allow greater flexibility for matching different equivalent expressions. The small units are called basic elements and are defined as triplets of words consisting of a head, a modifier or argument, and their relationship between both (head – modifier – relation). An improved version of this evaluation tool was later developed in (Tratz & Hovy, 2008). It was called **Basic Elements with Transformations for Evaluation** (BEwT-E) and its philosophy was the same as for BE. However, whereas BE used a predefined and static list of paraphrases for matching equivalent expressions, BEw-T-E automates this stage of the process proposing a set of rules capable of identifying abbreviations, prepositional phrases, nominalisations or synonyms, among others. The main drawback of this method concerns the use of several language-dependent preprocessing modules for parsing and cutting the sentences. As a consequence, parser resources in other languages rather than English would be a requirement for using it when summaries in different languages have to be evaluated. **DEPEVAL(summ)** (Owczarzak, 2009) is also a dependency-based metric. The idea here is similar to BE, and similarly, it compares dependency triples extracted from automatic summaries against the ones from model summaries. The main difference with BE is that a different parser is employed. Whereas BE uses Minipar[4], DEPEVAL(summ) is tested with the Charniak parser[5].

[4] http://webdocs.cs.ualberta.ca/ lindek/minipar.htm
[5] http://ftp.cs.brown.edu/pub/nlparser/

Furthermore, in order to address the shortcomings of different but equally good expressions, some tools for identifying paraphrases can also be suitable for assessing summaries. This is the case of **ParaEval** (Zhou *et al.*, 2006). Its objective is to provide a summarisation evaluation method, facilitating the detection of paraphrase matching. The paraphrase detection is performed according to a strategy based on three levels. First, multi-word paraphrases between phrases in the model summaries and the automatic summaries are identified. Then, for those fragments that are not matched, the method tries to find synonyms between single-words, and if this also fails, simple lexical matching is finally performed.

Instead of relying on n-grams or dependency parsing, **GEMS** (Generative Modelling for Evaluation of Summaries) (Katragadda, 2010) suggests the use of signatures terms to analyse how they are captured in peer summaries. Signature terms (also known as topic signatures) are word vectors related to a particular topic. They are calculated on the basis of part-of-speech tags, such as nouns or verbs; query terms and terms of model summaries. The distribution of the signature terms is computed first in the source document and then the likelihood of a summary being biased towards such signature terms is obtained to determine how informative the peer summary is. The main difficulty associated to this approach is to have lists of signature terms belonging to a topic that could serve to determine the important content of the source document, and consequently be used to assess the information contained in the peer summary.

FuSE (Fuzzy Summary Evaluator) and **DeFuSE** (Dictionary-Enhanced Fuzzy Summary Evaluator) are two methods proposed in (Ravindra *et al.*, 2006). These methods also evaluate the informativeness of a summary by comparing it to a model one. However, they rely on a different representation of the text through fuzzy sets. In the former method, FuSE, each sentence in the peer summary is given a value indicating its similarity degree with each of the sentences in the model summaries. This similarity is based on Hamming distance between collocations. The latter, DeFuSE, is an improved version of FuSE which also accounts for WordNet[6] relationships, in particular for synonyms. In this way a higher number of concepts can be matched and the similarity can be captured at a more abstractive level.

[6] http://wordnet.princeton.edu/

Different from the aforementioned methods, an implementation of Van Dijk's theories about discourse analysis (Van Dijk, 1972) is presented in (Branny, 2007). This approach relies on **text grammars**. A text grammar is a way of describing a valid text structure in a formal way, and it takes into consideration the surface and deep structure of sentences by means of their relationships (microstructures) and the structure of the text as a whole (macrostructure), respectively. Under the assumption that vocabulary overlapping is not enough to measure the informativeness of a summary, this approach identifies first a list of propositions. Then, humans have to decide whether each proposition is relevant or not for a summary. Further on, three scores are proposed, based on: i) informativity (how many propositions are present in the summary); ii) misinformation (misleading statements of the summary are detected); and iii) t-grammaticality (which is related to the correctness of the sentences based on orthographical or grammatical issues, as well as coherence problems). The application of this method on model and peer summaries shows that human summaries get higher scores than automatic, as it would have been expected. The main shortcoming of this method is that it is not possible to know how well it would correlate with human evaluation. Moreover, human intervention is required for identifying propositions and evaluating the amount of misinformation and ungrammaticality summaries have, which is very costly and time-consuming. Finally, due to the complexity of the method, it would not be easily scalable.

Although most methods have been developed for English, other evaluation methodologies have been proposed specifically for languages such as Chinese or Swedish. **HowNet**[7] (Dong & Dong, 2003) is an electronic knowledge resource for English and Chinese similar to WordNet, but differing from it in the way in which word similarity is computed. Moreover, HowNet provides richer information and each concept is represented unambiguously by its definition and association links to other concepts. It is a well-known resource for Chinese, and has been used in many approaches also for TS evaluation. In (Wang *et al.*, 2008), an approach for evaluating summaries based on HowNet is proposed. Despite the fact that this method is also based on n-gram co-occurrence statistics similar to ROUGE, its main novelty is the use of HowNet to compute word similarity, so that synonyms can also be taken into

[7] http://www.keenage.com

consideration. In addition, the authors also claim that this approach could be also used for detecting a few quality metrics to some extent, such as conciseness or sequence ordering.

Saggion et al. (2002) suggested a framework for evaluating different types of summaries both in English and Chinese. The methods used only relied on vocabulary overlap by means of cosine similarity. Moreover, model summaries were also needed in order to be compared with peer summaries.

Specific evaluation tools and resources for Scandinavian languages (mostly Swedish and Norwegian) have been also developed. Dalianis and Hassel (2001) developed a newswire corpus useful for evaluating summaries in Swedish (KTH extract corpus) which contains a set of documents together with the corresponding extracts manually written. In addition, Hassel (2004) proposed an evaluation framework (KTH extract tool). This tool is capable to compute some statistics with regard to the source documents and the summaries. For instance, how close is a summary with respect to a model one, or which text units appear more frequently in model summaries. In a similar way, a corpus and a set of evaluation resources for the Norwegian language are suggested in (Liseth, 2004).

In the last years, the *Automatically Evaluating Summaries of Peers* track proposed at TAC since 2009 has encouraged research into the automatic evaluation of summary's informativeness. The objective of this task is to develop automatic metrics that accurately measure summary content. Improved versions of the aforementioned approaches were proposed in the context of this task, such as (Giannakopoulos & Karkaletsis, 2009) or (Conroy *et al.*, 2009), which experimented with different approaches based on ROUGE ideas in an attempt to establish novel appropriate metrics and methods for assessing the informativeness of a summary automatically.

3.3.2 Quality Assessment

Although the previously explained evaluation methods are useful to assess the content of a summary, they only provide information regarding its informativeness. The way the information contained in the summary is presented is also very important, since it determines a summary's quality. This is crucial for determining how helpful a summary is when a user reads it. The evaluation concerning the quality of a summary taking into consideration issues such as *redundancy*, *coherence*, or *grammaticality*, has always been in mind of the researchers. However, automatic methods capable of performing successfully with respect to human evaluations are much more challenging to develop. Next, we are going to describe several approaches that have been suggested to address this type of evaluation.

The **FAN** and **MLUCE** protocols (Minel *et al.*, 1997) were among the first attempts to assess the quality of an abstract independently from the source text and the information it contained. Four criteria were proposed in the FAN protocol: (1) number of anaphora deprived of referents; (2) rupture of textual segments; (3) presence of tautological sentences; and (4) legibility of the abstract. However, all these criteria were evaluated manually by two jurors. In order to facilitate the process, the idea behind the MLUCE protocol was to enable potential users to evaluate summaries, depending on what they wanted the summary for. The MLUCE has the same limitation as the FAN protocol, i.e., all the evaluation process has to be carried out manually.

The evaluation of **summary indicativeness** and **sentence acceptability** was also addressed in (Saggion & Lapalme, 2000). On the one hand, *indicativeness* measures whether the summary is able to extract the topics of the document. The authors focus on scientific papers, and therefore, *indicativeness* is computed by comparing the terms appearing in the summary to the ones included in the abstract this type of document already contains. Using the abstracts already given in the document avoids the costly task of producing model summaries again; but there is a limitation regarding this issue, since not all documents contain an abstract, so in these cases human need would be necessary. On the other hand, *acceptability* determines if a selected sentence by a summarisation system is adequate compared to what humans would have selected. In this case, human intervention is needed to evaluate this criterion.

Conroy and Dang (2008) address the need of having tools which assess the content as well as other linguistic aspects in summaries. For this reason, **ROSE** (ROUGE Optimal Summarization Evaluation) was developed. This tool is based on ROUGE, but in order to account for linguistic aspects, the idea behind it is to find which ROUGE metrics better correlate with the overall responsiveness criteria manually evaluated in DUC and TAC conferences. This criterion reflects a combination of the content and the readability of a summary. In DUC and TAC conferences, the readability of summaries is evaluated with respect to five linguistic quality questions (grammaticality, non-redundancy, referential clarity, focus, and structure and coherence) which do not involve any comparison with a model summary. In TAC conferences, responsiveness and overall responsiveness were also included in the readability evaluation. However, this type of evaluation is manually performed by expert human assessors, who normally score the quality of a summary according to a five-point scale. Moreover, in the existing TS conferences, such as DUC or TAC, expert judges are asked to evaluate summaries using normally a 5- point scale, consisting of qualitative values with respect to a question or statement. For example, the following question was used in the evaluation performed at DUC 2001:

To what degree does the summary say the same thing over and over again?
1. Quite a lot; most sentences are repetitive
2. More than half of the text is repetitive
3. Some repetition
4. Minor repetitions
5. None; the summary has no repeated information

The problem associated to this type of questions and answers is that humans could understand answers such as “some repetitions”, but this would be very difficult for computers. It would be possible to map the outermost values into a quantitative scale (i.e., “Quite a lot”, and “None”). For instance, “none” would mean no repetition at all, but the boundaries in the middle are very subtle. Moreover, this sort of statements contains a degree of subjectivity, which is not possible to capture automatically. All these issues make the task of evaluating a summary’s quality really challenging and difficult to tackle from an automatic point of view.

Attempts to automate some of these criteria can be found in the literature. For instance, we can find some studies to predict text quality through the analysis of various **readability** factors (Pitler & Nenkova, 2008). The idea here is to analyse the quality of a text by means of different criteria including vocabulary, syntax, or discourse, in order to account for the correlation between those factors and human readability ratings previously gathered from existing evaluation forums. Each criterion is modelled in a different way, for instance, vocabulary is represented in terms of unigrams, and syntax is modelled via features, such as average number of noun-phrases or average number of verb-phrases. Results show that when combining all proposed readability factors, the prediction obtains an accuracy value close to 90%, and therefore this idea could be applied and extended to evaluate the quality of a summary.

Text coherence is also an essential characteristic that summaries should account for. However, it is very difficult to correctly measure it. Attempts to find automatic approaches to model and evaluate the coherence of a text can be found in (Barzilay & Lapata, 2005), (Lapata & Barzilay, 2005), or (Hasler, 2008). These approaches range from the Centering Theory (Grosz *et al.*, 1995) to the development of syntactic and semantic models, in order to capture the distribution of entities in the text, or the degree of connectivity across sentences.

Furthermore, attempts to automatically evaluate the **grammaticality** of a summary have been explored in (Vadlapudi & Katragadda, 2010b). N-gram models, in particular unigrams, bigrams, trigrams and the longest common subsequence are used for capturing this aspect. In addition, this problem is considered as a classification problem, where summary sentences are classified into classes on the basis of their acceptability. The acceptability parameter is estimated using trigrams. The proposed methods are evaluated in the same way as summaries were evaluated in DUC or TAC. Results obtained correlate well (85% at most) with respect to the already existing manual evaluations. Furthermore, in (Vadlapudi & Katragadda, 2010a), structure and coherence aspects are also investigated on the basis of lexical chains and the semantic relatedness of two entities. Results achieve a 70% when measuring the Spearman's correlation.

Some psycholinguistic studies focused on trying to identify what a manual summary evaluation involves, as well as potential features that can influence humans on their decisions when assessing summaries (Attali & Burstein, 2006). This study concluded that prior knowledge of the assessor, interest and motivation, as well as reading ability are factors that highly influence, thus explaining to some extent the great variability in opinions when different judges evaluate the same summary. Moreover, (Ravindra *et al.*, 2006) suggest ***complexity score*** as a novel parameter to evaluate summaries. This parameter accounts for the difficulty of performing summarisation for a specific task. However, to automate the evaluation process is very difficult, since it is not obvious to find out a priori what information may be the most relevant nor specific linguistic traits that help to decide a summary's quality automatically.

Furthermore, Latent Semantic Analysis (LSA) (Landauer *et al.*, 1998) is a technique that has been widely used for scoring essays automatically. This technique is able to identify relationships between a set of documents and the terms they contain. Scoring a student's essay can be related somehow to the evaluation of summaries with respect to its quality. In both cases, the generated text must be clear and coherent. It is claimed that using LSA to assess written essays enables grade ranges similar to those awarded by human graders. For this reason, many researches have focused on this topic (Foltz *et al.*, 1999), (Landauer *et al.*, 2003), (Attali & Burstein, 2006). However, employing LSA for evaluating summaries is less researched, contrary to what happens when generating them, where LSA is a well-known technique. Only in (Venegás, 2009), LSA is used for specifically evaluating summaries written in Spanish.

Despite the challenges involved in automating quality criteria for evaluating summaries, the number of approaches attempting to automate some of these criteria, such as grammaticality or coherence, has increased considerably. Consequently, research in the TS evaluation is advancing beyond the content assessment only.

3.4 Crowdsourcing for Text Summarisation Evaluation

One of the limitations of the current evaluation approaches is the need of model summaries written by humans for determining how good a peer summary is. Building model summaries is not a trivial task. On the contrary, it is a very tedious and time-consuming task. Moreover, when dealing with large collections of documents, this task becomes hardly feasible, since lots of humans would be needed for generating such summaries within a reasonable time interval. Furthermore, in the case manual evaluation is carried out, this process has to be repeated every time a new summary is generated.

For all these reasons, crowdsourcing[8] services can be suitable for evaluating summaries, since they provide an environment where users can perform different tasks, getting some money for the job done as a reward. Therefore, the goal of this section is to describe how crowdsourcing services have been used for TS.

In particular, Amazon provides a service, called **Mechanical Turk**[9] (Mturk) that allows users (requesters) to define and upload Human Intelligence Tasks (HITS). These HITS will be then performed by other humans (turkers), who will be rewarded with the corresponding amount of money associated to the task. Mturk is very appropriate for carrying out those tasks that are simple for humans, but very difficult for computers. For instance, to obtain model summaries would be relatively easy and fast using Mturk. However, it has been shown in (Gillick & Liu, 2010) that one has to be very careful with the annotations provided by Mturk, since they are not always as good as they should be. The quality of the results has to be checked and therefore, when using this type of services, it is very important to ensure that turkers are suitable for the task, as well as to check that they do not give random answers. For this reason, Mturk, itself, provides a facility qualification to assist quality control. Requesters can attach various requirements to their task in order to force turkers to meet such requirements before they are allowed to work (Tang & Sanderson, 2010). For instance, the percentage of the accepted tasks a turker has completed can be used in order to decide if it is worthy to allow such turker to perform the tasks.

Focusing on TS, Mturk has been not as explored as for other applications, such as machine translation (Callison-Burch, 2009). Although Mturk should be an easy way to gather model summaries as well to evaluate them, the difficulty of obtaining the same readability results for peer summaries as in TAC 2009 with non-expert judges in contrast to expert ones was shown in (Gillick & Liu, 2010). Quality control policies were first established, in order to assure that only turkers with a 96% HIT approval could perform the task and, in addition, if the task

[8] Crowdsourcing is related to the act of outsourcing a particular task, but asking and allowing large amounts of people to perform it.

[9] https://www.mturk.com/mturk/welcome

was finished under 25 seconds, their work was rejected. Concerning the amount of money it was paid, different compensation levels were analysed, finding out that lower compensations ($0.7 per HIT) obtained higher quality results. It seemed that this compensation level attracted turkers less interested in making money and more conscious of their work. Regarding the results obtained, the Mturk evaluation presented high variability. Whereas TAC assessors could roughly agree on what makes a good summary, obtaining a standard deviation of 1.0, the standard deviation computed for turkers' results was doubled, obtaining a value of 2.3. As can be noticed, in this case, non-expert evaluation differed a lot from the official one, and therefore, Mturk was not of great help. However, (El-Haj *et al.*, 2010) showed the appropriateness of using Mturk for collecting a corpus of single-document model summaries from Wikipedia and newspaper articles in Arabic. These summaries were produced by extracting the most relevant sentences of the documents and not taking more than half of the sentences in the source documents. Finally, 765 model summaries were gathered. These summaries were then used to evaluate the corresponding automatic ones produced by several existing Arabic summarisation systems using different evaluation approaches, such as ROUGE, or AutoSummENG. In this case, Mturk facilitated the process of gathering a big number of model summaries.

3.5 Overall Discussion

This chapter presented an overview of the relevant issues concerning the evaluation of summaries. In particular, two existing types of evaluation, intrinsic and extrinsic, were explained, further focusing on the intrinsic one. The most widespread approaches for assessing either the content of summaries as well as their quality have been analysed, outlining their advantages and disadvantages. In addition, this chapter also introduced recent crowdsourcing services, such as Mechanical Turk, which can be of help for evaluating summaries or collecting large amounts of data in a relatively easy, fast, and cheap way.

Evaluating a summary, either manually or automatically, is not a trivial issue. The manual evaluation involves human effort for determining to what extent a summary is good with respect to specific criteria (information contained, grammaticality, coherence, etc.). This is very costly and time-consuming, especially if lots of summaries have to be evaluated. In addition, the subjective nature of manual evaluation may lead to different summary results depending on the assessor, even though strict guidelines are provided to carry out the evaluation process.

Most of the evaluation methods presented before rely on model summaries, that have been written by humans. Then, these model summaries are compared to the automatic generated ones. Different studies have proven that, if humans had to decide the most relevant sentences from documents in order to produce summaries, they would disagree in which sentences best represent the content of a document. Therefore, the low agreement between humans is a problem. Another problem is the semantic equivalence between different nouns, for example by means of synonymy, or expressions, when there are various ways to express the same idea, is another drawback of the evaluation. Most methods only perform a superficial analysis, and do not take into consideration the semantic meaning of phrases. This means that it is possible to have several valid summaries, although different in content. Consequently, research into other ways of evaluating the content of a summary would be useful.

With respect to the quality evaluation, the suggested methods are still at their early stages. Ideally, this type of evaluation is independent of the source documents, but performing it is really difficult. Developing good metrics that correlate well with human assessment is also very challenging.

Crowdsourcing services, such as Mturk, can be used for evaluating a summary; for instance, asking humans either to write model summaries, or evaluate existing summaries. Although they can provide fast and relative inexpensive mechanisms to carry out tasks that are simple for humans but very difficult for computers and required a lot of human effort, there are also some disadvantages related to these services. Some issues concerning the quality of the task performed by the turkers arise, since some turkers will probably enrolled in a task only for the money, providing non-sense answers in order to decrease the time their spend with the task, but at the same time, increasing their rate of pay, being able to finish more tasks. Regarding this, research on how to account for the quality of the results provided by these services, as well as methods for ensuring such quality would be needed.

Despite the considerable progress of the evaluation of TS in recent years, there is still a lot of room for improvement, especially for the quality-oriented approaches. The inherent subjectivity associated to the evaluation process poses greater challenges to this research subfield. State-of-the-art approaches mainly focus on intrinsic evaluation, in particular, in novel methods to assess either a summary content or its quality. To fully automate this process is very difficult and for this reason new research about this topic can be considered as emerging research. However, as long as semantic methods improve, it will be more feasible to account for equivalent expressions, and approaches would not rely on model summaries as much as they currently do for evaluating the informativeness of a summary. Furthermore, to be aware of the techniques and approaches existing for other HTL areas, such as essay scoring, can also help to achieve improvements for TS evaluation. Similar to what occurred with ROUGE, which was inspired in a method for evaluating machine translation systems (i.e., BLEU), the success of the techniques employed for scoring essays could be perfectly investigated for evaluating also text summaries.

4. COMPENDIUM Text Summarisation Tool

4.1 Introduction

The process of Text Summarisation (TS) can be seen from several perspectives. From a cognitive point of view, it tries to explain the mental mechanisms involved in the reading comprehension of texts (Van Dijk, 1980), (Van Dijk & Kintsch, 1983). Instead, from a computational point of view, the process does not attempt to go into such detail. On the contrary, it is more concerned with the techniques and approaches suitable for generating summaries automatically (Hovy, 2005).

In this chapter, COMPENDIUM TS tool is presented. COMPENDIUM produces different types of summaries and it is based on a cognitive perspective. In addition, it takes into account the computational perspective that it is needed for its automation.

The types of summaries COMPENDIUM is able to produce can be classified with respect to different factors, as it was explained in Chapter 2. In this way, concerning the *input*, COMPENDIUM can either take one or several texts, and produce single- or multi-document summaries. Regarding the *purpose* of the resulting summaries, these can be generic, query-focused or sentiment-based, and their aim is to provide information about the source document(s), thus being informative. As *output*, the final summaries can be extracts or abstractive-oriented summaries (i.e., a combination of extractive and abstractive information). Finally, it is important to mention that COMPENDIUM is a mono-lingual TS approach, working only for one *language*, i.e., English.

As far as the architecture of COMPENDIUM is concerned, an architecture based on specific stages is proposed, following Van Dijk's theories (Van Dijk, 1980), (Van Dijk & Kintsch, 1983). Moreover, some of the specific techniques suggested also rely on a cognitive point of view, such as the Code Quantity Principle (Givón, 1990). In particular, COMPENDIUM relies on five core stages that constitute the backbone of the TS process (*surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*). The result of applying these stages is a generic extract from a single o multiple documents, that fits into a specific compression rate[1] or number of words. Furthermore, a series of additional stages can be integrated into the core ones, thus enhancing the capabilities of COMPENDIUM, generating also query-focused, sentiment-based or abstractive-oriented summaries. Specifically, the additional stages are: *query similarity*; *subjective information detection*; and *information compression and fusion.*

Therefore, this chapter is structured as follows. Section 4.2 explains the cognitive perspective behind the TS process, which serves as the basis of our TS approach. Section 4.3 describes our proposed TS tool: COMPENDIUM. In this section, we first provide an overview of its main characteristics and architecture (Subsection 4.3.1), and then we focus on the stages it comprises (Subsections 4.3.2 and 4.3.3, for the core and additional stages, respectively). Further on, the equivalence between COMPENDIUM and other TS processes is explained in Section 4.4. In this way, the idea behind COMPENDIUM can be better understood, and the relationship between each stage of our TS approach with the phases of the cognitive and computational TS processes can be established. The last section (Section 4.5) concludes this chapter, analysing the main contributions.

4.2 The Process of Summarisation from a Cognitive Perspective

It is well known that the theory proposed by (Van Dijk & Kintsch, 1983) is one of the most influential theories about the process of reading and comprehension of texts. They explain how humans understand a text from a cognitive perspective. According to the authors, when a person reads a text, an “understanding” of it is created in the reader’s mind. In this comprehension process, a complex mental mechanism is involved, where three different representations of the

[1] The compression rate defines the proportion of text contained in the summary with respect to the original document.

text are produced in the reader's mind: 1) a textual microstructure; 2) a semantic macrostructure; and 3) an schematic superstructure. The first one consists of a verbatim representation of the text, and it refers to the coherence and cohesive relationships between the units of the text (i.e., words, phrases, clauses, sentences, and connections between sentences), resulting in a list of propositions derived from the text. After that, the list of propositions is transformed into a network connecting all of them (if the text is coherent), producing as a result the macrostructure of the text. Finally, the schematic superstructure is the global structure that characterises the type of text (e.g. narrative, descriptive, expository or argumentative), and it is independent of its content. Figure 4.1 shows these representations.

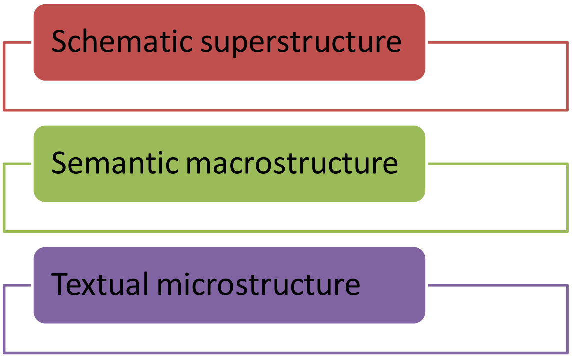


Fig. 4.1. Mental mechanisms involved in the comprehension process.

From this point to the end of the section, we will focus on the concept of ***macrostructure***. This concept refers to the global meaning of a discourse[2] (Van Dijk, 1980), and it may have different connotations depending on the research discipline. For instance, the macrostructure will be directly related to the topic of a text in the theory of discourse, whereas in the psycholinguistics of discourse processing, it will be connected with the type of information a person will be able to remember after a period of time (general information rather than specific details).

[2] Although the term *discourse* can refer both to oral or written discourse, in this thesis, we will refer to text every time the term is mentioned.

Focusing on the theory of discourse, the macrostructure will be taken into account for the global meaning of a text: its topic, theme, or gist. According to Van Dijk (1980), this means that some rules are needed to relate meanings of words and sentences to their semantic macrostructure. Moreover, the macrostructure of a text is also related to its global coherence.

Therefore, a summary defined as *"a coherent text that contains the overall gist of a document"* (Mani, 2001a) is directly related to the concept of macrostructure (the global meaning of discourse). In particular, **a summary is the explicit representation of the macrostructure of a text** (i.e., its overall meaning) (Álvarez Angulo, 2002). This representation contains the most relevant elements stated in the text from a semantic point of view, which is captured by means of the knowledge of the world a reader has, as well as the level of importance he/she assigns to each discourse element. Moreover, it also involves that users understand the text. Regarding this issue, each user understands a text in a different manner. Thus, the process of summarisation is subjective; there is not a unique possible summary of a text. In fact, due to its subjectivity, two readers may produce different summaries from the same document.

Van Dijk (1980) also states that the macrostructure of a text consists of a set of macropropositions which can be derived from a sequence of propositions of a text. This process is achieved by applying some kinds of rules that link textual propositions with the macropropositions used to define the global topic of a text fragment. Indeed, they derive macrostructures from microstructures. These rules are called macrorules, and they can have both a reductive and a constructive nature. On the one hand, some of them are employed to remove information, whereas on the other hand, they also allow certain elements to be combined in new, more complex units of information. In particular, there is a distinction between four major macrorules:

- **Deletion**: This macrorule is the simplest as well as the most general. It consists in deleting all information in the text that is not relevant. In other words, this rule operates on the irrelevant details that do not contribute to the construction of the topic. Among the information that is not relevant, redundant information is also included. The following example will be used along the explanation of the four macrorules to illustrate each of them with a specific example.

[1] *Mary was playing with a doll.*
[2] *She was wearing a red dress.*
[3] *A few days ago, she was ill.*
[4] *She missed several days of school.*
[5] *Peter was building figures with Lego blocks.*
[6] *John enjoyed himself riding a bicycle.*
[7] *Paul told their children to come in because the dinner was ready.*

By applying the *deletion* rule, we can remove the second one (*"She was wearing a red dress"*), because it is not crucial for the interpretation of the remaining sentences. Consequently, sentence [2] would not appear in the macrostructure of the text (the summary).

- **Strong deletion**: This macrorule is a stronger variant of the previous one, in which other details that can be locally relevant for a sentence, but not for the global meaning of the text, are also deleted. Normally, only one *deletion* rule is used, so the first and the second macrorules could be grouped in a general deletion. Moreover, these rules can also be seen as positives (selection and strong selection, respectively), in the sense that they allow us to select which information is relevant.

  [1] *Mary was playing with a doll.*
  [3] *A few days ago, she was ill.*
  [4] *She missed several days of school.*
  [5] *Peter was building figures with Lego blocks.*
  [6] *John enjoyed himself riding a bicycle.*
  [7] *Paul told their children to come in because the dinner was ready.*

  From the remaining sentences after applying the first macrorule, the *strong deletion* one would discard sentences [3] and [4]. Although sentence [3] is necessary to understand proposition [4], in the remaining of the text, anything else about the illness of Mary is mentioned, so these sentences give some details that are not essential to understand the gist of the text.

- **Generalisation**: The purpose of this macrorule is to generalise the information stated in several propositions into a more general one: the hyponyms are usually substituted by their corresponding hypernym. An example of this macrorule is shown below:

  [**1**] *Mary was playing with a doll.*
  [**5**] *Peter was building figures with Lego blocks.*
  [**6**] *John enjoyed himself riding a bicycle.*
  [**7**] *Paul told their children to come in because the dinner was ready.*

  In the above example, the generalisation process is simple for sentences [1], [5], and [6] into the new sentence *"The children were playing"*. It is worth underlying that the level of the generalisation has to be taken into account; otherwise, too general propositions could be obtained (e.g. *"The children were doing something"*), and the meaning of the whole sequence of information could not be correctly represented. In order to avoid this, (Van Dijk, 1980) states that this macrorule involves the least possible generalisation, for instance, by taking the immediate superset of predicates.

- **Construction**: By applying the fourth macrorule, a new proposition is built, which is the result of a joint sequence of propositions that denotes the overall meaning of them. This new proposition is the macroproposition. Taking as a basis the same example as before, once the previous macrorules have been applied:

  [**New**] *The children were playing.*
  [**7**] *Paul told their children to come in because the dinner was ready.*

  The corresponding macroproposition for the initial set of sentences would be:
  [**Macroproposition**] *The children were playing and their father told them to come in to have dinner.*

Apart from these four macrorules, there are some cases where the proposition and the macroproposition are identical. This leads to a **zero macrorule**, where everything that is said in a proposition is equally relevant and therefore, such proposition is left intact. For instance, sentence [7] in the above examples ("Paul told their children to come in because the dinner was ready.") could have been left unchanged.

One important aspect about the application of these macrorules is the order in which they should be applied. Whether to apply *deletion* or *construction* macrorules first is difficult to answer (Van Dijk, 1980). Applying *deletion* in the first place could lead to a loss of important information that we could later need when applying the *construction* macrorule. In these cases, we might not have enough information to apply the *construction* macrorule. It seems more reasonable to apply the *construction* macrorule first to see which propositions can be joint together, then delete the irrelevant information (*deletion*) and finally apply the *generalisation* macrorule. However, from a cognitive point of view, each reader will apply these macrorules in a different way, depending on his/her interests, knowledge, purpose, etc., so not a specific order can be established in advance.

The recursively application of these macrorules will end up with the macrostructure of a text, which is only a part in the whole discourse comprehension process (Figure 4.1). However, as it was previously explained, the macrostructure will provide us with the global meaning of the text (its topic), and therefore, obtaining the macrostructure of a text is equivalent to obtain the summary, since the summary is the direct expression of the macrostructure of discourse (Álvarez Angulo, 2002). The process of summarisation according to the cognitive perspective (Van Dijk, 1980), (Van Dijk & Kintsch, 1983) is schematically depicted in Figure 4.2.

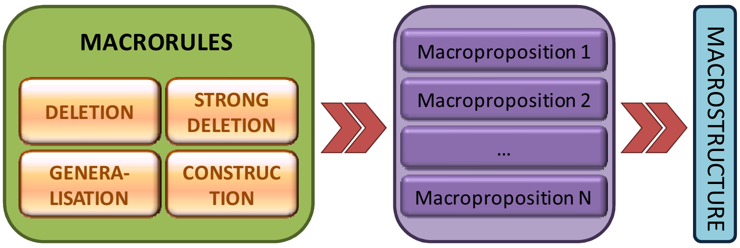


Fig. 4.2. The process of summarisation from a cognitive perspective.

4.3 COMPENDIUM Text Summarisation Tool

In this section, COMPENDIUM TS tool is described in detail. The aim of COMPENDIUM is to produce different types of summaries automatically. Therefore, we first explain the main characteristics of the summaries generated, as well as an overview of its proposed architecture in Subsection 4.3.1. Then, we go into detail and we explain its stages, distinguishing between a set of core ones, that are the most important in COMPENDIUM (Subsection 4.3.2), and the additional stages (Subsection 4.3.3), which are used to generate different types of summaries (query-focused, sentiment-based and abstractive-oriented summaries). It is worth noting that COMPENDIUM is based on a cognitive perspective (Van Dijk, 1980) and (Van Dijk & Kintsch, 1983), and it also takes into account the computational issues stated in (Hovy, 2005). Therefore, its equivalence between each of its stages and the different macrorules, as well as other TS process is provided in Subsection 4.4.

4.3.1 The Approach

Taking the schema depicted in Figure 2.4 of Chapter 2 as a basis, the summaries generated with COMPENDIUM according to the proposed factors (media, input, output, purpose and language) can be characterised. Next, each of these factors are explained in the context of COMPENDIUM.

- **Media**. It refers to the nature of the object to be summarised (video, text, audio, etc). In particular, COMPENDIUM only works with texts.
- **Input**. Regarding this factor, COMPENDIUM takes one or several texts, thus being able to produce single- and multi-document summaries.
- **Output**. The type of summaries generated can be either extracts (i.e., the most important sentences are selected) or abstractive-oriented summaries (in our case, information is compressed and fused to generate new sentences different from the original ones, and then these sentences are combined with the previously extracted ones).
- **Purpose**. The summaries generated with COMPENDIUM aim at being substitutes for the original documents. Therefore, they must

contain the most relevant facts, thus being informative. In addition, specific purposes according to user's interests are taken into account. This leads to generic, query-focused and sentiment-based summaries. Generic summaries provide a general overview of the document; query-focused summaries are biased towards a user need, question or topic; and finally, sentiment-based summaries contain a high degree of subjective information, reflecting the opinions of users about a topic.

- **Language**. COMPENDIUM is a mono-lingual TS tool, which work with English, both for the input and output.

Figure 4.3 illustrates the types of summaries COMPENDIUM is able to generate according to such factors.

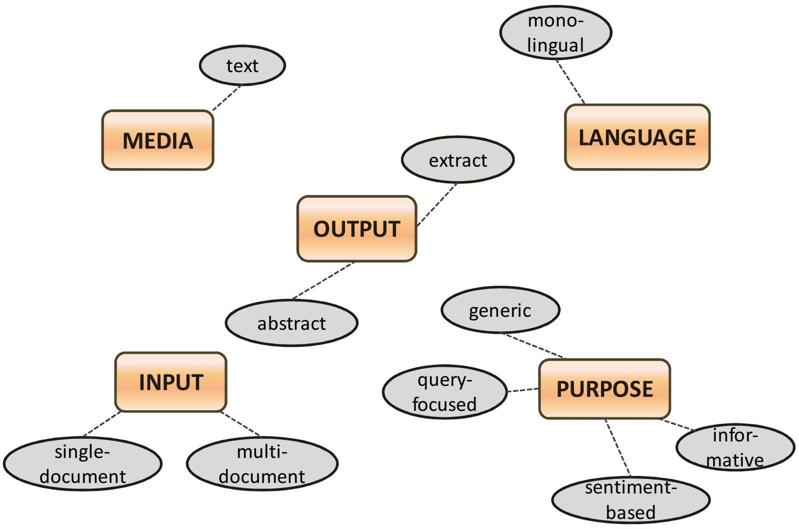


Fig. 4.3. Characteristics of the summaries generated with COMPENDIUM.

Concerning the stages of COMPENDIUM, five core stages that constitute the backbone of the TS process are distinguished. These stages are:

- **Surface linguistic analysis**, which pre-processes the input text.
- **Redundancy detection**, which identifies and removes repeated information.
- **Topic identification**, which determines the main topics of the document/s to be summarised.
- **Relevance detection**, which identifies the most relevant sentences of the document/s.
- **Summary generation**, which extracts the most relevant sentences in the same order as they were initially.

Furthermore, we suggest three additional stages in order to increase the capabilities of COMPENDIUM, allowing it to generate query-focused, sentiment-based, or abstractive-oriented summaries. Specific stages for achieving each of these types of summaries are:

- **Query similarity**, needed for generating query-focused summaries.
- **Subjective information detection**, necessary for identifying subjective information and producing sentiment-based summaries.
- **Information compression and fusion**, crucial for generating new information that will appear in the summary, thus resulting in an abstractive-oriented summary, rather than an extract.

Figure 4.4 depicts the general architecture of COMPENDIUM. In this figure, the core stages are represented within a big rectangle with rounded borders, whereas dotted rectangles correspond to the additional stages. By applying only the core stages, the resulting summaries are single- or multi-document generic informative extracts. In contrast, by taking into consideration the additional stages, query-focused or sentiment-based extracts, as well as abstractive-oriented summaries from a single or several documents can be generated. The remaining of this chapter focuses on the explanation of the core and additional stages.

4.3.2 Core Stages

As it was aforementioned, the core stages are: a) *surface linguistic analysis*; b) *redundancy detection*; c) *topic identification*; d) *relevance*

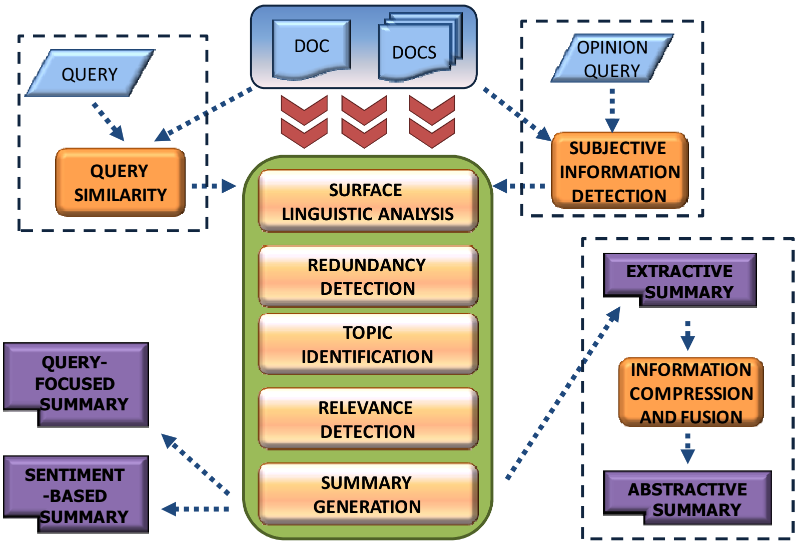


Fig. 4.4. General architecture of COMPENDIUM.

detection; and e) *summary generation.* Since COMPENDIUM is based on the cognitive theories of (Van Dijk, 1980) and (Van Dijk & Kintsch, 1983), some of the stages correspond to a specific macrorule. In particular, the redundancy detection stage corresponds to the first macrorule (deletion), whereas the topic identification and relevance detection are considered as a "strong deletion". Finally, the last core stage, summary generation, is equivalent to the construction macrorule. Next, we explain each of stages in detail, and we leave the correspondence between COMPENDIUM stages and the macrorules for Section 4.4.

Surface Linguistic analysis. This stage aims at carrying out a basic linguistic analysis on the input document, thus preparing it for further processing. In order to carry out this analysis, external state-of-the-art tools and resources are used. In particular, for this stage we propose:

- **Sentence segmentation.** The text is splitted into sentences, which are the textual units considered for generating the summary. For this purpose, the sentence segmentation tool provided at DUC evaluation campaigns[3] is used.

- **Tokenisation.** A tokeniser allows us to identify each word in the document, since we will need to compute for instance the frequency of each word, or distinguish between stop words and non stop words in later stages. In order to be able to identify each word of the text, a tokeniser is used. In particular, we employ *Word Splitter*[4].
- **Part-of-speech tagging.** A part-of-speech tagger assigns each word with its corresponding morphological category (noun, verb, adjective, preposition, adverb, determiner, pronoun, and conjunction). This process is useful for distinguishing between types of words, since some of them (e.g. nouns or verbs) can be more important than others (e.g. determiners). This tool will be used in the additional stage of *information compression and fusion*. In particular, TreeTagger[5] was used as a part-of-speech tagger, because it is very easy to use, and it can be used for different languages, not only for English.
- **Stemming.** This process consists in reducing words to their stem form. It is very useful in the cases where there is no need to differentiate between two inflected words that belongs to the same family (e.g. *running* and *runs*, both come from the verb *run*). The *Porter Stemmer*[6] is employed for performing this task, which will be necessary for considering all terms sharing a common stem as a single one, for later computing their frequency.
- **Stop word identification.** Stop words are words which appear very frequent in documents, but they do not carry any semantic information. They are normally not used for further processing, because they are not relevant (e.g. articles: *the*, *a*; conjunctions: *and*, *or*, etc.). In our case, this process is essential; otherwise when computing the frequency of a word for determining the topic of a document (please see the *topic identification* stage), words such as *"and"*, *"is"*, or *"the"* could be wrongly identified as document's topics. For

[3] http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz
[4] http://cogcomp.cs.illinois.edu/page/tools_view/8
[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[6] http://tartarus.org/ martin/PorterStemmer/

carrying out this task, a list of stop words is needed. In particular, we use an English stop word list[7].

Redundancy Detection. The aim of this stage is to identify redundant information in the source documents, in order not to include it in the summary. For this purpose, Textual Entailment (TE) is employed, since its objective is to determine whether the meaning of a text snippet, also known as hypothesis (H), can be inferred from another one, called the text (T) (Glickman, 2006). For instance, the following examples, taken from the RTE corpora[8], show a true and false entailment, respectively.

Pair id=50 (entailment = true)
T: Edison decided to call "his" invention the Kinetoscope, combining the Greek root words "kineto"(movement), and "scopos" ("to view").
H: Edison invented the Kinetoscope.
Pair id=18 (entailment = false)
T: Gastrointestinal bleeding can happen as an adverse effect of non-steroidal anti-inflammatory drugs such as aspirin or ibuprofen.
H: Aspirin prevents gastrointestinal bleeding.

The main idea behind the use of TE for detecting redundancy is that those sentences whose meaning is already contained in other sentences can be discarded, as the information has been previously mentioned. Therefore, by applying TE we can obtain a set of sentences from the text which do not hold an entailment relation with any other, and then keep this set of sentences for further processing. To illustrate how TE is employed, let's assume that a document consists of the following six sentences:

$S_1\ S_2\ S_3\ S_4\ S_5\ S_6$

Then, in order to come up with a set of non-redundant sentences, the TE is performed in this manner[9]:

[7] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[8] http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/
[9] $NRsent$ is the set of non-redundant sentences.

$NRsent = \{S_1\}$
$NRsent \longrightarrow entails \longrightarrow S_2 \Rightarrow NO$
$NRsent = \{S_1, S_2\}$
$NRsent \longrightarrow entails \longrightarrow S_3 \Rightarrow NO$
$NRsent = \{S_1, S_2, S_3\}$
$NRsent \longrightarrow entails \longrightarrow S_4 \Rightarrow YES$
$NRsent = \{S_1, S_2, S_3\}$
$NRsent \longrightarrow entails \longrightarrow S_5 \Rightarrow YES$
$NRsent = \{S_1, S_2, S_3\}$
$NRsent \longrightarrow entails \longrightarrow S_6 \Rightarrow NO$
$NRsent = \{S_1, S_2, S_3, S_6\}$

As a consequence, in this example, S_4 and S_5 are discarded from the text, and only the non-entailed sentences (i.e S_1, S_2, S_3 and S_6) are kept for further stages, thus forming the final set of non-redundant sentences (NRsent). To compute such entailment relations we have used the TE approach presented in (Ferrández, 2009). This TE system relies on lexical (e.g. cosine similarity, Leveshtein distance), syntactic (e.g. dependency trees) and semantic measures based on WordNet (Fellbaum, 1998).

In this sense, the use of TE for detecting and removing redundancy in TS is novel. On the one hand, attempts to study the influence of TE on TS have been focused on the evaluation of summaries (Harabagiu *et al.*, 2007) to determine which candidate summary, among a set of them, better represents the content in the original document depending on whether the summary entails it or not. On the other hand, TE has been combined with TS to generate a summary directly from the entailment relations found in a text (Tatar *et al.*, 2008), or by extracting the highest scored sentences of a document, where the score of each sentence is computed as the number of sentences of the text that are entailed by it.

Topic Identification. In this stage, the most relevant topic/topics of the document are identified by means of their frequency of appearance in the text. Term Frequency (TF) calculation is employed for achieving this goal, because it has been shown in previous work that high frequent words are indicative of the topic of a document (Luhn, 1958), (Montiel Soto & García-Hernández, 2009), and, what is more, they are very likely to appear in human-written summaries (Nenkova *et al.*, 2006).

Therefore, the frequency of a word, without considering stop words (they were previously identified in the *stop word identification* stage) is computed and it is used in the next stage of the TS process for determining the overall relevance of a sentence.

Relevance Detection. The relevance detection stage assigns a weight to each sentence, depending on how relevant it is within the text. This weight takes the TF computed in the previous stage, by means of which the main topics have been identified, and it adds another feature based on *The Code Quantity Principle* (CQP) (Givón, 1990). This principle has a cognitive background, which states that: (1) a larger chunk of information is given a larger chunk of code; (2) the less predictable information, the more coding material; and (3) the more important information, the more coding material. In other words, the most important information within a text will contain more lexical elements, and therefore it will be expressed by a high number of units (for instance, syllables, words or phrases). This is also in line with *The Code Quantity, Attention and Memory Principle* (Givón, 1990) where it is stated that there is a proportional relation between the relevance of information and the amount of quantity through it is coded, since the more salient and different coded information in a text, the more easily the reader's attention will be caught. As a result, readers will retain, keep and retrieve this kind of information more efficiently. Furthermore, *The Code Quantity Principle* has been proven to hold true in written texts (Ji, 2007).

On the basis of this, a coding element can range from characters to phrases. A noun-phrase is the syntactic structure which allows more flexibility in the number of elements it can contain (pronouns, adjectives, or even relative clauses), and is able to carry more or less information (words) according to what the writer wants to express. In addition, the longer a noun-phrase is, the more information it carries. For instance, let's take these two sentences as example:

S_1: *The Spanish Academy of Motion Pictures Arts and Sciences presented an honorific award for the best actor.*

S_2: *The Academy presented an honorific award.*

In this case, S_1 contains more information than S_2. Although at a first sight, the second sentence might be more appropriate for TS, since it reflects the same facts of the first one but in a shorter manner, the first one contains more details, and this would lead to more informative summaries.

Under these assumptions, CQP can be combined with TF to decide on which sentences of a document contain more relevant information. The lexical units considered as encoding elements for the CQP are words inside noun-phrases, without taking into account the stop words. To select noun-phrases as coding units seems appropriate, since in (Mittal *et al.*, 1999) it was found that the average length of complex noun-phrases in summary sentences was more than twice as long than those in non-summary sentences. In addition, Lloret and Palomar (2009) carried out a preliminary study of the percentage of noun-phrases contained in both source documents and model summaries of a corpus of newswire and another one of fairy tales. This analysis showed that words belonging to noun-phrases were predominant over other types of words in the documents as well as in the summaries, representing on average more than 70% of all content words (i.e., without taking stop words into account), and approximately 30% of the total words of documents.

In order to identify noun-phrases within a sentence the *BaseNP Chunker*[10], is employed. One important aspect to take into consideration is that the use of a chunker (as well as any other tool based on Human Language Technologies) can introduce some error rate. This tool achieves recall and precision rates of roughly 93% for base noun-phrase chunks, and 88% for more complex chunks (Ramshaw & Marcus, 1995), thus being suitable for our purposes.

The method by which the relevance of a sentence is computed taking into account CQP and TF is shown in formula 4.1.

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \tag{4.1}$$

where:
r_{s_i} = is the relevance of sentence i,
$\#NPi$ = number of noun-phrases contained in sentence i,

[10] This resource is free available in ftp://ftp.cis.upenn.edu/pub/chunker/

tf_w = frequency of word w that belongs to the sentence's noun-phrase.

Summary Generation. The objective of this stage is to generate a summary of a specific length. This length can be expressed in words or in the form of the compression rate (i.e., the percentage of information the summary contains with respect to the source document). However, the way this stage is carried out, strongly depends on the type of summary we want to produce. Consequently, four main types of summaries can be distinguished: 1) generic; 2) query-focused; 3) sentiment-based; and 4) generic abstractive-oriented. Type 1 is directly produced when the core stages of COMPENDIUM are applied, whereas for types 2, 3 and 4 the additional stages of *query similarity*, *subjective information detection* and *information compression and fusion* are also required, respectively. Next, we briefly explain each of these strategies for generating the final summary.

1. **Generic summaries (COMPENDIUM$_E$).** Once the final score of a sentence is computed by means of the *relevance detection* stage, the most important sentences (i.e. the ones with highest scores) are selected and extracted to form the final summary up to a desired length or compression rate. In this case, the final summary is a generic extract.
2. **Query-focused summaries (COMPENDIUM$_{QE}$).** Having computed the two different weights for each sentence (its relevance –r_{s_i}–, and its similarity with regard to the query – $qSim_{s_i}$ –, in the *relevance detection* and *query similarity* stages[11], these two values are combined within the same formula (Formula 4.2), where β can be assigned different weights between 0 and 1, depending on whether we want to give more importance to the relevance or to the query similarity weight, when producing query-focused summaries.

$$Sc_{s_i} = (1+\beta^2)\frac{r_{s_i} * qSim_{s_i}}{\beta^2 * r_{s_i} + qSim_{s_i}} \tag{4.2}$$

where:
Sc_{s_i} = is the score associated to sentence i,

[11] Please see the description of the *query similarity* stage in Section 4.3.3.

r_{s_i} = is the relevance of the sentence *i*,
$qSim_{s_i}$ = is the similarity between the query and sentence *i*.

Taking into account the size restrictions, the top-ranked sentences will be selected and extracted, forming the final query-focused extract.

3. **Sentiment-based summaries (COMPENDIUM$_{SE}$)**. For generating this type of summaries, the same strategy as for the generic ones is followed. The main difference between them is that in this case we only focus on subjective information, which is identified and processed in the *subjective information detection* stage[12]. As a result, extracts containing only subjective information are produced.
4. **Generic abstractive-oriented summaries**. These summaries (**COMPENDIUM$_{E-A}$**) combine extractive and abstractive TS strategies, and as a consequence, the sentences of the final summary are selected following a strategy that maximises the similarity between each of the new sentences generated in the *information compression and fusion* stage (Section 4.3.3), and the ones that have been selected as the most important in the *relevance detection* stage, given that the similarity[13] between them is above a predefined threshold. In the cases where a sentence in the extract has an equivalent in the set of the new generated sentences, the former will be substituted for the latter; otherwise, the sentence as it appears in the extract will be kept. The resulting summaries are abstractive-oriented.

4.3.3 Additional Stages

Apart from the core stages, there are three additional stages (*query similarity*, *subjective information detection*, and *information compression and fusion*) that, when integrated with the former, enhance the capabilities of COMPENDIUM. These stages allow the generation of other types of summaries (i.e., query-focused, sentiment-based and a

[12] This stage is explained in detail in Section 4.3.3.

[13] We use the cosine similarity measure to compute the similarity between two sentences.

abstractive-oriented summaries). It is important to mention that one of this stages, *information compression and fusion*, corresponds to the generalisation macrorule proposed in (Van Dijk, 1980) and (Van Dijk & Kintsch, 1983). Each of these additional stages are described below.

Query Similarity. When query-focused summaries have to be produced, a query is usually associated to the source documents, in order to specify the kind of information the user is interested in. From an extractive point of view, the summary should contain the most relevant sentences in the document that also contains the information in the query. Therefore, the goal of this stage is to take into account the information expressed in a given query to tailor the contents of the final summary to such information.

In order to determine which sentences may be potentially related to the given query, the cosine similarity between each sentence and the query is computed, using the Text Similarity package[14]. Formula 4.3 shows how the query similarity weight is calculated.

$$qSim_{s_i} = CosineSimilarity(S_i, Query) \tag{4.3}$$

where:
$qSim_{s_i}$ = is the similarity between the query and sentence i.

Subjective Information Detection. The objective of this stage is to detect and process subjective information, with the purpose of producing sentiment-based summaries. This has to be performed before the relevance of sentences is assigned. In order to be able to carry out such task, a tool capable of analysing and classifying the sentiment associated to a fragment of text (e.g. words, sentences, documents, etc.) is necessary. This manner, it is possible to know whether a fragment of text is subjective or objective, and in addition, if it is positive or negative about a specific topic, entity, product, etc. In particular, COMPENDIUM uses the external opinion mining tool described in (Balahur-Dobrescu *et al.*, 2009). It assigns to each sentence in the document one of these three different scores: i) $score > 0$, if the sentence has a positive nature; ii) $score < 0$, if it has a negative nature; and iii) $score = 0$, if the sentence is neutral (i.e., its an objective sentence). Once all the sentences have been analysed according to their associated

[14] http://www.d.umn.edu/ tpederse/text-similarity.html

sentiment, only the ones that are not neutral are selected for further processing.

Information Compression and Fusion. This stage aims at generating new sentences in one of these forms: either a compressed version of a longer sentence, or a new sentence containing information from two individual ones. The main steps involved in this stage are:

- **Word graph generation**: for generating new sentences, we rely on word graphs adopting a similar approach to the one described in (Filippova, 2010). Specifically in our approach, we first generate an extractive summary in order to determine the most relevant content for being included in the summary. Then, a directed weighted word graph is built taking as input the generated extract, where the words represent the nodes of the graph, and the edges are adjacency relationships between two words. The weight of each edge is calculated based on the inverse of the frequency of co-occurrence of two words, taking also into account the importance of the nodes they link through the Pagerank algorithm (Brin & Page, 1998). Once the extract is represented as a word graph, a pool of new sentences is created by identifying the shortest path between nodes (e.g. using Dijkstra's algorithm).

  Figure 4.5 illustrates the idea of using word graphs to represent text and how a new sentence can be obtained.

- **Incorrect paths filtering**: this stage is needed since not all of the sentences obtained by the shortest paths are valid. For instance, some of them may suffer from incompleteness (e.g. *"Therefore the immune system"*). Consequently, in order to reduce the number of incorrect generated sentences, we define a set of rules, so that sentences not accomplishing all the rules are not taken into account. Three general rules are defined after analysing manually a set of generated sentences derived from a small data set, which are:

  – The minimal length for a sentence must be 3 words[15].

  – Every sentence must contain a verb.

[15] We assume that three words (i.e., subject+verb+object) is the minimum length for a complete sentence.

S1: The student goes to the campus by bike.
S2: The campus of the University of Alicante is very nice.

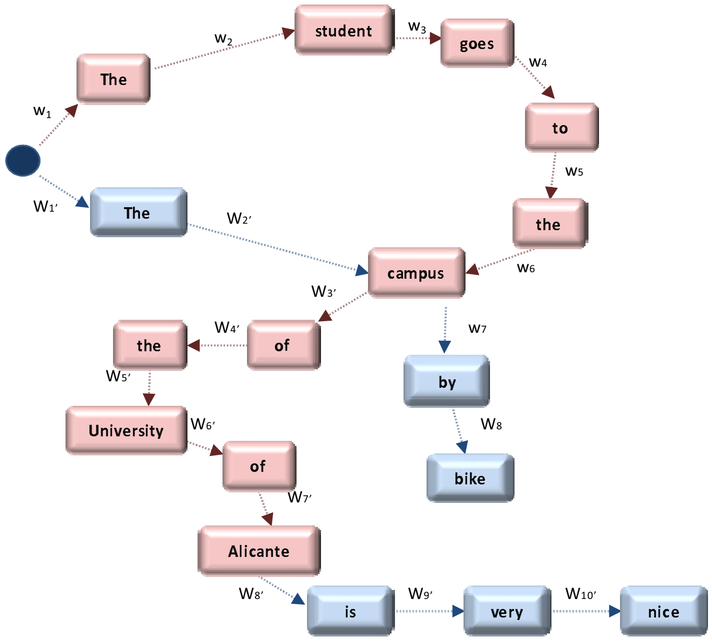


Fig. 4.5. Example of word graph representation.

– The sentence should not end in an article (e.g. a, the), a preposition (e.g. of), an interrogative word (e.g. who), nor a conjunction (e.g. and).

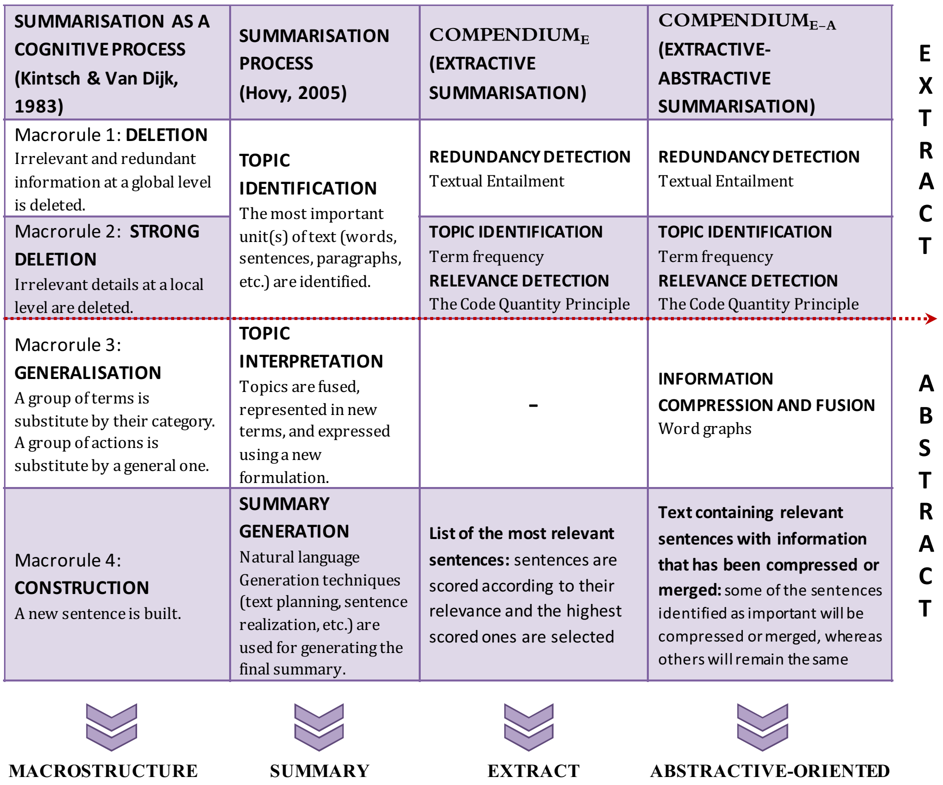
Once the incorrect sentences have been removed, we can use the new sentences instead of the original ones (see "generic abstractive-oriented summaries" in the *summary generation* stage). It is important to stress upon the fact that, in COMPENDIUM, this stage takes place after the *relevance detection* and before the *summary generation* stages. This is because we need to be aware of the most relevant sentences of the document first. However, this is not the only possible strategy to adopt, and other alternatives are being also studied with respect to this issue (please see Chapter 7 which explains the work in progress).

4.4 COMPENDIUM in relation to Cognitive and Computational Summarisation Processes

As it can be seen, the TS process may differ depending on the approach we adopt. From a cognitive perspective (Van Dijk, 1980), (Van Dijk & Kintsch, 1983), the application of the macrorules is essential to be able to build the summary, whereas from a more computational point of view (Hovy, 2005), the stages are not as theoretical as the former. Despite their differences, a correspondence between them can be established. Moreover, such correspondence is also possible in COMPENDIUM, in the sense that one or several stages of our proposed TS tool can be mapped to different macrorules or phases in the TS process defined by the previous authors.

In this manner, all the core stages (except the surface linguistic analysis) together with the *information compression and fusion* correspond to a different level on the TS process previously mentioned.

Figure 4.6 shows a schema representing the equivalence between COMPENDIUM and the TS processes from two perspectives. On the one hand, the TS process from a cognitive point of view is shown in the first column. This process is based on (Van Dijk, 1980) and (Van Dijk & Kintsch, 1983), whereas on the other hand, the second column shows the TS stages according to Hovy (2005). The last two columns focus on COMPENDIUM in its two most important variants: the extractive (COMPENDIUM$_E$) and the abstractive-oriented, i.e., the one which combines extractive with abstractive information (COMPENDIUM$_{E-A}$). In this way, the correspondence of each proposed stage in our TS tool and the other TS processes can be analysed. As it can be seen the first two macrorules (deletion and strong deletion) can be considered equivalent to the topic identification stage defined in (Hovy, 2005). Equivalent to them are the redundancy detection and the topic identification and relevance detection stages of COMPENDIUM. Compared to Hovy (2005), they all can be included in the topic identification stage, since by applying such stages the most relevant sentences will be determined in both approaches (extractive and abstractive-oriented). However, from the cognitive point of view, a distinction is made between the redundancy detection and the topic identification and relevance detection stages. All of them involve “deletion”, but the former only accounts for repeated information, whereas the “deletion” that is carried out



SUMMARISATION AS A COGNITIVE PROCESS (Kintsch & Van Dijk, 1983)	SUMMARISATION PROCESS (Hovy, 2005)	COMPENDIUM$_{E}$ (EXTRACTIVE SUMMARISATION)	COMPENDIUM$_{E-A}$ (EXTRACTIVE-ABSTRACTIVE SUMMARISATION)
Macrorule 1: **DELETION** Irrelevant and redundant information at a global level is deleted.	**TOPIC IDENTIFICATION** The most important unit(s) of text (words, sentences, paragraphs, etc.) are identified.	**REDUNDANCY DETECTION** Textual Entailment	**REDUNDANCY DETECTION** Textual Entailment
Macrorule 2: **STRONG DELETION** Irrelevant details at a local level are deleted.		**TOPIC IDENTIFICATION** Term frequency **RELEVANCE DETECTION** The Code Quantity Principle	**TOPIC IDENTIFICATION** Term frequency **RELEVANCE DETECTION** The Code Quantity Principle
Macrorule 3: **GENERALISATION** A group of terms is substitute by their category. A group of actions is substitute by a general one.	**TOPIC INTERPRETATION** Topics are fused, represented in new terms, and expressed using a new formulation.	-	**INFORMATION COMPRESSION AND FUSION** Word graphs
Macrorule 4: **CONSTRUCTION** A new sentence is built.	**SUMMARY GENERATION** Natural language Generation techniques (text planning, sentence realization, etc.) are used for generating the final summary.	**List of the most relevant sentences:** sentences are scored according to their relevance and the highest scored ones are selected	**Text containing relevant sentences with information that has been compressed or merged:** some of the sentences identified as important will be compressed or merged, whereas others will remain the same
MACROSTRUCTURE	SUMMARY	EXTRACT	ABSTRACTIVE-ORIENTED

Fig. 4.6. COMPENDIUM in relation to cognitive and computational summarisation processes.

in the other two is stricter, since it is necessary to identify the most relevant sentences in the document. The generalisation macrorule is equivalent to the topic interpretation stage. The idea is to generate new information from the concepts presented in the documents. In the case of COMPENDIUM, it is worth noting that COMPENDIUM$_E$ does not specifically address this issue; therefore, it produces extracts as output. On the contrary, COMPENDIUM$_{E-A}$ tackles this by means of the *information and compression fusion* stage. This stage takes a set of sentences as input, and it generates new information that can be either a compressed version of a specific sentence or a new sentence containing information from others. For this reason, the *information and* *compression fusion* stage corresponds to the third macrorule as well as the topic identification stage. The last macrorule (construction) is at the same level as the summary generation. This is the final stage of the TS process. In our case, this corresponds to the summary generation stage of COMPENDIUM, but for clarity reasons, a short explanation of the stage for COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$ is provided, since each approach addresses the stage of summary generation taking into account specific issues. Whereas in the former, only a list of the most relevant sentences is extracted maintaining the order in which they were in the original document, in the latter, the information contained in the summary is a combination of new sentences generated in the *information and compression fusion* stage, and sentences previously extracted with COMPENDIUM$_E$.

4.5 Conclusion

This chapter presented COMPENDIUM, a TS tool that is able to generate different summary types. In particular, according to the most common factors, COMPENDIUM generates summaries only from text data and in English. The summaries can be generated from a single or multiple documents, as far as the input is concerned. Regarding the output, the summaries are informative and can be either extracts or abstractive-oriented summaries. Finally, depending on what a user wants the summary for, COMPENDIUM is capable of generating generic, query-focused and sentiment-based summaries. In order to be capable of generating such types of summaries, an architecture which comprises two kinds of stages: core and additional was proposed. On the one hand, the core stages (*surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*) are the main stages in the process, and through them generic informative extracts from one or more documents can be produced. On the other hand, the additional stages (*query similarity*; *subjective information detection*; and *information compression and fusion*) are specifically developed for generating query-focused, sentiment-based or abstractive-oriented summaries, and thanks to them the capabilities of COMPENDIUM increase.

Furthermore, it is worth mentioning that COMPENDIUM is based on a cognitive perspective (Van Dijk, 1980), (Van Dijk & Kintsch, 1983), which was also explained in the chapter. By applying this approach through the different proposed macrorules (deletion, strong deletion, generalisation and construction) the macrostructure of a text (i.e., its global meaning) can be obtained, which corresponds to the summary of such text. In this way, we showed the equivalence of each of the stages in COMPENDIUM with the macrorules proposed from the cognitive perspective. Moreover, this approach takes also into account the computational point of view (Hovy, 2005), thus leading to the automation of the TS process.

Regarding the advantages of COMPENDIUM, it is worth stressing upon the fact that its modular architecture allows the integration of new stages in an easy way, that can improve the TS tool as well as extending its functionalities. Moreover, another positive issue is the fact that it is capable of generating the most common types of summaries. This means that COMPENDIUM can be applied to a wide range of contexts, depending on the user's needs. However, it also has some limitations. On the one hand, it does not deal with multi-linguality, since it is only able to produce summaries in English. Due to the fact that some of the external tools our TS tool employs have specifically been developed for English, to adapt it to other languages (e.g. Spanish) cannot be done straightaway; however, it would be feasible to do so, provided that we find equivalent tools for the target language. On the other hand, the techniques suggested do not take into account semantic information. Semantic information would allow COMPENDIUM to perform a deeper understanding of the text to be summarised. In this way, important concepts (instead of words) together with their relationships could be identified, as well as specific knowledge could be also taken into consideration (e.g. ontologies). As a consequence, the process of TS would be enriched, and a stage of information generalisation could be integrated to COMPENDIUM.

To recapitulate, the main contributions of COMPENDIUM with respect to the state of the art TS systems and approaches are:

- **A TS tool which also takes into consideration the process of summarisation from a cognitive point of view.** We follow (Van Dijk, 1980) and (Van Dijk & Kintsch, 1983) discourse theory, where each macrorule has its corresponding stage in COMPENDIUM.

The fact that an automatic TS tool attempts to model the TS process from this perspective is a novelty.

- **The proposal and development of COMPENDIUM TS tool.** The architecture proposed relies on a set of core stages, which are the backbone of the tool. Specific methods and techniques are suggested for each of the stages (*surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*). Briefly, the surface linguistic analysis comprises basic linguistic tasks, such as tokenisation, stemming, or part-of-speech tagging, in order to pre-process the input. Then, redundant information is removed by using textual entailment in the redundancy detection stage. Further on, the main topics of the document/s are identified by means of the term frequency technique. The information about the topics is needed in the relevant detection stage, which consists of determining the most relevant sentences. This is achieved by computing a score for each sentence based on the topics it has and a cognitive-based technique, the Code Quantity Principle. Finally, the last stage extracts the highest score sentences in the same order as they appear in the initial document/s. Moreover, it allows the integration of specific modules into its core stages (the so-called additional stages), thus extending the types of summaries it is able to generate. In particular, these stages are *query similarity*; *subjective information detection*; and *information compression and fusion*, through which query-focused, sentiment-based and abstractive-oriented summaries can be produced, respectively.

  Within this contribution, we would like to remark the following novelties that have been addressed in COMPENDIUM:

  - **The use of textual entailment as a redundancy detection method.** This method combines lexical, syntactic and semantic information in order to detect entailment relationships between sentences. And although textual entailment and TS had been used before to generate and evaluate summaries, it was not employ for dealing with the problem of redundancy.

  - **The use of linguistic theories directly related to how humans remember the information.** We propose a novel feature based on the *Code Quantity Principle* for detecting relevant information in texts.

- **Address the challenge of abstract generation**, through the combination of extractive and abstractive techniques that leads to abstractive-oriented summaries. Although the graph-based method employed for generating new sentences followed the ideas of Filippova (2010), the novelty is the manner the abstractive-oriented summaries are generated. A pool of new sentences are created from the extract, and then they are used to substitute those sentences in the extract that express the same information. In the case that no equivalence is found, the sentences are not replaced.

5. Evaluation and Experiments

5.1 Introduction

Text Summarisation (TS) research area involves the automatic generation of summaries, as well as their evaluation. However, although evaluating a summary is a hard task; it is necessary. Without carrying out any type of assessment, one cannot guarantee that generated summaries meet the purposes they were generated for nor to what extent users find them useful.

The objective of this chapter is to conduct an intrinsic evaluation of COMPENDIUM. Therefore, we focus on assessing the generated summaries on their own, i.e., with respect to their content and quality. The exhaustive evaluation performed comprises three sub-objectives. Firstly, we ensure that the proposed techniques in each stage of COMPENDIUM are appropriate. For this reason, textual entailment is compared to other methods that are often employed for detecting and removing redundant information (cosine similarity and maximal marginal relevance). A manual evaluation is carried out in order to evaluate the amount of redundant information contained in a summary. Then, term frequency and the code quantity principle techniques are evaluated as single features as well as in combination. In this way, we analyse to what extent they are suitable for the topic identification and the relevance detection stages, respectively. Secondly, COMPENDIUM is evaluated in different domains and for the different types of summaries it is able to generate. Contrary to most of the summarisation research, our evaluation does not only focus on newswire, but also we analyse the use of COMPENDIUM in blogs, image captions, and medical research papers. Moreover, within each context several types of summaries taking into account different factors are evaluated. Concerning the input, single and multi-document summaries are evaluated. Regarding the purpose of the summaries, generic, query-focused, and sentiment-based summaries COMPENDIUM is able to produce are also evaluated; and finally, abstractive-oriented summaries will be also assessed, as far as the output is concerned. The evaluation methodology mostly employed is ROUGE (Lin, 2004), since it is automatic and it has been widely adopted by the research community for the evaluation of summaries. However, we are aware of its limitations as a tool, and therefore, in some cases, the evaluation carried out is complemented with manual evaluation according to quality criteria or user satisfaction. Finally, apart from evaluating the summaries generated using COMPENDIUM, we also compared the results obtained with respect to the state-of-the-art approaches, thus being able to analyse potentials and limitations the proposed TS tool has.

Therefore, in order to cover the main objective and its subobjectives, the structure of the chapter is as follows. Section 5.2 explains the evaluation methodology employed. Then, the description of the different corpora used is provided in Section 5.3. Finally, the experiments and the results obtained are presented in Section 5.4. Previous to provide an evaluation of COMPENDIUM as a whole, the appropriateness of textual entailment as a redundancy method is measured (Subsection 5.4.1), and then the techniques proposed for topic identification and relevance detection are assessed (Subsection 5.4.2). After that, its performance is evaluated in a wide range of domains, comprising newswire, image captions, blogs and medical research papers (Subsection 5.4.3). The evaluation will provide some insights of the advantages and limitations of COMPENDIUM, which will be discussed in Section 5.5, as well as the main conclusions drawn.

5.2 Evaluation Environment

A number of experiments with various corpora are conducted in order to assess the performance of COMPENDIUM. We select ROUGE[1] as the tool for automatically evaluating our summaries, since it is a widespread TS evaluation tool that has been shown to correlate well with human evaluations (Lin & Hovy, 2003). ROUGE is able to evaluate how informative an automatic summary is by comparing its content

[1] http://berouge.com/default.aspx

to one or more reference summaries. Such comparison is made in terms of *n-gram* co-occurrence. The most well-known ROUGE metrics are: ROUGE-1 and ROUGE-2, which compute the number of overlapping unigrams and bigrams, respectively; ROUGE-SU4, which measures the overlap of skip-bigrams an automatic summary contains with respect to a model one, with a maximum distance of four words between them; and finally, ROUGE-L, which calculates the longest common subsequence between two summaries. However, as we explained in Chapter 3, ROUGE presents some disadvantages. For instance, we need to have at least one reference summary in order to be able to use this tool. An important issue is that ROUGE can give an idea of how a summary performs regarding its content with respect to another; however, it does not ensure that the evaluated summary is good or bad. Good summaries can obtain low ROUGE results if their vocabulary do not match with the one included in the reference summaries; or bad summaries can perform good according ROUGE, being illegible from a human point of view.

In the cases where we do not have model summaries, we perform a manual evaluation, taking into account different criteria concerning the quality of the generated summaries, such as grammaticality, redundancy or focus. In other cases, the user satisfaction is evaluated, in order to determine the usefulness of an automatic summary from a human point of view. In these cases, we establish a rating scale where different degrees of goodness are analysed. For instance, a 3-level scale may comprise the values "low", "medium", and "high", whereas in a 5-level Likert scale the degrees to measure the agreement with respect to a specific issue are established ("strongly agree", "agree", "neither agreee nor disagree", "disagree", and "strongly disagree").

5.3 Corpora Used

As corpora, we use several data sets belonging to different domains, depending on the experiment performed and the type of summary we want to evaluate. The advantages of not focusing only on one specific corpus is that COMPENDIUM can be assessed from a broader perspective, and its strengthens and weaknesses can be found out. In particular, COMPENDIUM is evaluated for the following corpora:

- **Newswire**. It consists of a collection of English newswire documents provided by DUC[2]. On the one hand, we use the news documents within the DUC 2002 conference. These are 567 documents grouped in 59 clusters, where each cluster represents a set of topic related documents. Moreover, model summaries for each news are also provided. Specifically, for each document, the number of reference summaries range from 1 to 2 (1,112 model summaries in total). This data is appropriate for carrying out single- and multi-document generic TS. On the other hand, for multi-document summarisation, the data provided in DUC 2003 and DUC 2004 are also suitable. They consist of 30 and 50 clusters of news documents, for DUC 2003 and 2004 respectively, containing approximately 10 documents each, and 4 model summaries for each cluster.

- **Image captions**. This corpus was created by Aker and Gaizauskas (2010b), and it contains 308 different images with manually assigned place names. Each image has 10 documents in English related to it that have been retrieved using a search engine, where the name of each place has been set as the query. Additionally, each image has up to 4 model summaries (932 in total) which were created manually using the on-line social site, VirtualTourist[3]. This corpus is especially suitable for generating query-focused summaries of multiple input documents.

- **Blogs**. This corpus consists of 51 blogs together with their comments extracted from the Web. They deal with five different topics (economy, science and technology, cooking, society and sport). The blogs and their corresponding comments are written in English and all of them have the same structure: the authors create an initial entry containing a piece of news and their opinion on it and subsequently, bloggers reply expressing their opinions about the topic. We use this corpus for generating sentiment-based summaries.

- **Medical research papers**. This corpus consists of a collection of 50 research articles from specialised journals of medicine that were gathered directly from the Web[4]. Each article contains a human-written abstract, that can be considered as a model summary.

[2] http://www-nlpir.nist.gov/projects/duc/data.html

[3] http://www.virtualtourist.com/

[4] http://www.elsevier.com/wps/product/cws home/622356

This collection of journal articles are appropriate to perform single-document summarisation.

Table 5.1 shows an overview of all the data sets we have worked with. Additionally, some properties are also provided. In particular, for each corpus, the table shows: the number of clusters, the number of total documents within each corpus, the average size of the documents (in number of words), the number of model summaries available for each corpus, and finally the average length of the model summaries (in number of words, as well).

Corpus	**Num. clusters**	**Num. docs**	**Avg.length (docs)**	**Num.model sum.**	**Avg.length (model sum)**
Newswire DUC 2002	59	567	630	1,112	100
Newswire DUC 2003	30	298	669	120	100
Newswire DUC 2004	50	500	601	200	100
Image captions	308	3,080	690	932	200
Blogs	-	51	5,547.55	-	-
Medical research papers	-	50	2,060	51	162.7

Table 5.1. Overview of the corpora used and their properties.

5.4 Experiments and Results

In this section, the experiments conducted for evaluating the different types of summaries COMPENDIUM is able to generate are described. In particular, we evaluate: single- and multi-document summaries; generic, query-focused and sentiment-based summaries; and extracts and abstractive-oriented summaries. Moreover, two additional experiments are also taken into consideration. The performance of textual entailment as a redundancy detection method is assessed, whereas term frequency and the Code Quantity Principle are evaluated to decide whether or not they are good techniques to identify the topics of a document, and determine the relevance of a sentence, respectively. The motivation behind the use of these techniques was explained in the previous chapter, where COMPENDIUM TS tool was presented. Briefly, textual entailment was selected as a redundancy method because it can identify whether a sentence can be deduced from another. This fact indicates that both sentences share some content, despite being written with different vocabulary. Term frequency was chosen since it was proven to be very useful for TS, and it has been widely taken into account in lots of approaches. Finally, the idea behind the Code Quantity Principle is to have more information about how humans retain the information when they read it, and approach this perspective with methods based on HLT. Moreover, the reasons why the proposed experiments were carried out are twofold. On the one hand, we want to ensure that the techniques proposed for COMPENDIUM are appropriate, and on the other hand, we do not want to restrict our TS tool to a particular type of summary or domain. On the contrary, our aim is to analyse if the techniques proposed within COMPENDIUM are appropriate for generating the most common types of summaries of well-known text types (newswire, image captions, blogs, and medical research papers). Moreover, by carrying out these experiments, we are also able to compare our results with other TS systems. The results obtained for all these experiments are shown in different subsections (Subsection 5.4.1, 5.4.2, and 5.4.3), and a discussion is provided within each of them.

Table 5.2 provides a general perspective of the experimental setup. In this table it is shown the type of summary assessed, the evaluation method employed, the corpus used, as well as the section where the evaluation of such experiment can be found.

5.4.1 Assessing Textual Entailment as a Redundancy Detection Method

COMPENDIUM relies on Textual Entailment (TE) as a redundancy detection method. In order to evaluate whether it is a good choice with respect to other approaches, we compared it to two well-known methods for avoiding redundancy in TS: i) Cosine Similarity (CoSim) and ii) Maximal Marginal Relevance (MMR). The reason why such methods are selected for comparison are explained next. On the one hand,

Assessment	**Corpus**	**Sum. length**	**Eval. method**	**Section**
Redundancy detection method	Newswire DUC 2002, 2003, and 2004	100	Human eval.	5.4.1
Features for topic identification and relevance detection	Newswire DUC 2002	100	ROUGE	5.4.2
Single-doc, generic extracts	Newswire DUC 2002	100	ROUGE	5.4.3
Multi-doc, generic extracts	Newswire DUC 2002, 2003, and 2004	100	ROUGE	5.4.3
Multi-doc, generic extracts	Image captions	200	ROUGE	5.4.3
Multi-doc, query-focused extracts	Image captions	200	ROUGE	5.4.3
Single-doc, sentiment-based extracts	Blogs	10%, 15% and 20%	Human eval.	5.4.3
Single-doc, generic abstracts	Medical research papers	162	ROUGE and human eval.	5.4.3

Table 5.2. Overview of the experiments performed.

CoSim is one of the most widespread approaches to tackle redundancy in TS (Radev *et al.*, 2001), (Toutanova *et al.*, 2007), (Stokes *et al.*, 2007), since it is a fast and simple method. However, it has some limitations due to the fact that it only detects lexical similarity (word overlapping) between sentences. On the other hand, MMR (Carbonell & Goldstein, 1998) attempts to maximise the relevance of a sentence for including in the summary the most relevant ones, but at the same time, it reduces the chances of adding a sentence with information that has been already included.

For assessing the performance of TE with regard to CoSim and MMR, we used the clusters of news provided in DUC editions of 2002, 2003, and 2004. COMPENDIUM is employed to generate summaries with TE, and an external TS system, MEAD (Radev *et al.*, 2004), is selected for producing summaries with cosine similarity[5] and MMR. Also, two baselines are also analysed: LEAD and RANDOM. In the first one, summaries are produced by selecting the first sentence of each doc-

[5] By default, the similarity threshold for cosine similarity is set to 0.7 in MEAD.

ument in the cluster, then, the second sentence, and so on until the desired summary size is reached (in our case, 100 words). The latter selects random sentences from documents.

To carry out the human evaluation, a group of seven undergraduate and postgraduate students was given specific guidelines about how summaries should be evaluated. The same guidelines as in the evaluation of DUC conferences[6] are followed, but focusing only on the non-redundancy aspect. The goal of this evaluation is to determine the amount of repeated information in the generated summaries, so they were asked to provide a score in a five-point scale for each summary, according to the non-redundancy quality criteria: *"There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentence that is repeated, or repeated facts, or the repeated use of a noun or noun phrase when a pronoun would suffice."* Specifically, the humans were told to rate a summary according this exact question:

> **To what degree does the summary say the same thing over and over again?**
> *1. Quite a lot; most sentences are repetitive*
> *2. More than half of the text is repetitive*
> *3. Some repetition*
> *4. Minor repetitions*
> *5. None; the summary has no repeated information*

The evaluation process was completely blind, in the sense that humans did not have any information about which summary was generated with which method. Once all the summaries were evaluated, each numeric rating was mapped into a qualitative scale of three possible values: *very good*, *acceptable*, and *very bad*. The first one, *very good*, grouped ratings 4 and 5; *acceptable* corresponded to value 3; and finally *very bad* was bound to values 1 and 2. At the end, each summary pertained to one of this three categories. The reason for doing this was that sometimes the barrier between one category and another was very fuzzy, so in order to reduce ambiguities, the initial fine-grained scale was mapped to this new one.

[6] http://www-nlpir.nist.gov/projects/duc/

Approach	%	DUC-02	DUC-03	DUC-04	Mean
LEAD	Very good	17.0	13.3	30.0	20.1
	Acceptable	39.0	30.0	44.0	37.7
	Very bad	44.1	56.6	26.0	42.2
RANDOM	Very good	76.1	76.7	86.0	79.6
	Acceptable	22.0	23.3	12.0	19.1
	Very bad	1.7	0	2.0	1.2
CoSim	Very good	55.9	73.3	36.0	55.1
	Acceptable	39.0	13.3	50.0	34.1
	Very bad	5.1	13.4	14.0	10.8
MMR	Very good	55.9	61.7	38.0	51.9
	Acceptable	37.3	26.7	20.0	28.0
	Very bad	6.8	11.6	42.0	20.1
TE	Very good	**86.4**	**85.0**	**98.0**	**89.8**
	Acceptable	10.2	15.0	2.0	9.1
	Very bad	3.4	0	0	1.1

Table 5.3. Analysis of the redundancy found in summaries with different methods (CoSim=cosine similarity within MEAD; MMR=maximal marginal relevance within MEAD; and TE = textual entailment within COMPENDIUM).

Table 5.3 shows the percentage of summaries in each category and for each data set. As it can be seen, most summaries that have used TE as a redundancy detection method were evaluated as *"very good"*, being the percentage of *"very bad"* summaries almost non-existent. Although the performance of the TE module used is around 60% (Ferrández *et al.*, 2009), it is able to identify sufficient redundancy, preventing summaries from containing repeated information. Therefore, TE is an appropriate method for detecting redundant information. Regarding the other analysed approaches, it is worth mentioning that both CoSim and MMR under the MEAD system perform quite similar. Finally, with regard to the baselines (LEAD and RANDOM), the RANDOM baseline performs very good, and despite selecting sentences randomly, the corresponding summaries do not contain much redundant information. In contrast, the LEAD baseline does not get satisfactory results concerning redundancy. This was expected since it was produced by taking the first sentence of the first document, then the first sentence of the second, etc. so, the chances that similar information is stated in the first sentence of several related documents increase.

5.4.2 Assessing the Features for Topic Identification and Relevance Detection

In this evaluation, we aim at evaluating the appropriateness of the features proposed for topic identification (term frequency) and relevance detection (the Code Quantity Principle) that COMPENDIUM employs. The evaluation carried out comprises the analysis of these features on their own, as well as their combination, as it was explained in the *relevance detection* stage in Chapter 4.

In order to analyse to what extent the proposed features are suitable for TS, we generate generic extracts of 100 words, using the newswire DUC 2002 corpus. To produce such summaries, we follow one of these approaches:

- **Only Term Frequency (TF)**. We select as summary sentences those ones containing only the important topics identified with TF. This corresponds to the *topic identification* stage of COMPENDIUM, where TF is used to identify important topics. In order to build a summary from these topics, we rank sentences, according to Formula 5.1, and the top-ranked ones form the summary.

  $$Sc_{s_i} = \frac{\sum_{i=1}^{n} tf_j}{n} \; . \tag{5.1}$$

  where
  Sc_{s_i} = Score of sentence i
  tf_j = frequency of word j, i.e, number of times that j appears in the source document
  n = length of the sentence without considering stop words.

- **Only The Code Quantity Principle (CQP)**. We select as summary sentences those ones which their relevance has been determined using only CQP, without taking into account the *topic identification* stage. Formula 5.2 shows how we rank sentences taking into account this feature on its own. Again, the top-ranked sentences according to this formula are extracted.

  $$Sc_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |w| \; . \tag{5.2}$$

where:
Sc_{s_i} = Score of sentence i
#NPi = number of noun-phrases contained in sentence i,
|w |= 1, when a word belongs to a noun-phrase.

- **Combination of TF and CQP (TF+CQP)**. This corresponds to the stages *topic identification* and *relevance detection* of COMPENDIUM explained in Chapter 4. We next remember the formula used for ranking sentences (Formula 5.3).

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \tag{5.3}$$

  where:
  r_{s_i} = is the relevance of sentence i,
  $\#NPi$ = number of noun-phrases contained in sentence i,
  tf_w = frequency of word w that belongs to the sentence's noun-phrase.

For the evaluation, we computed the values for ROUGE-1, ROUGE-2, ROUGE-SU4, and ROUGE-L[7]. Table 5.4 shows the results obtained for each of the three approaches analysed. The results represent the F-measure value ($\beta = 1$), since it combines both precision and recall, and gives an idea of how good the summaries are with respect to the model summaries also provided in the corpus. A paired t-test was carried out in order to account for the statistical significance of the results obtained. The results with a statistical significance at a 95% confidence level are marked with a star.

Approach	ROUGE-1	ROUGE-2	ROUGE-SU4	ROUGE-L
TF	0.42951	0.16970	0.19618	0.38951
CQP	0.41751	0.17127	0.19261	0.38039
TF+CQP	**0.44153***	**0.18565***	**0.20795***	**0.39920***

Table 5.4. Analysis of the features used in COMPENDIUM.

[7] The parameters for ROUGE were fixed to: -n 2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -l 100 -d

As it can be seen in Table 5.4, the combined approach (TF+CQP) obtains better results than the other two, improving the performance of TF and CQP, respectively, by 4% and 6.80% on average, for all ROUGE metrics. Regarding the analysis of the individual features, it is worth mentioning that the summaries generated using only TF perform better than the ones produced only with CQP. This may be due to the fact we do not take into consideration any technique to identify the most important terms or topics of a document. Therefore, it can happen that the noun-phrases of the selected sentences do not contain relevant terms, and as a consequence, the performance is affected. However, when combined the two proposed features in the same approach (TF+CQP), results improved being statistical significant over the remaining approaches. Hence, the proposed features are appropriate for TS, and in particular for COMPENDIUM.

5.4.3 Assessing COMPENDIUM in Different Domains and Contexts

In this Section, we evaluate the performance of COMPENDIUM as a whole TS approach. As we explained in the previous chapter (Chapter 4), we defined some stages of the TS process as core, whereas other were proposed as additional stages, for enhancing the capabilities of COMPENDIUM. Table 5.5 shows the different variants of COMPENDIUM, together with the kinds of summary generated and the stages involved in the process.

Moreover, we group the different evaluations conducted with respect to the type of corpora used (newswire, image captions, blogs, and medical research papers). We next described in detail the experiments performed and the results obtained within each domain.

- **Newswire corpus**. Using the different DUC corpora about news, COMPENDIUM$_E$ is evaluated for single- and multi-document summarisation. For both evaluations, we use ROUGE-1, ROUGE-2 and ROUGE-L[8], and we report the values for F-measure ($\beta = 1$). On the one hand, for single-document summarisation, we use the DUC

[8] We follow the guidelines corresponding to DUC 2002 (task 1 and 2), DUC 2003 (task 2), and DUC 2004 (task 2). Moreover, the parameters for ROUGE are fixed to: -n 2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -l 100

Short name	Type of summaries	TS stages involved
COMPENDIUM$_E$	single- or multi-doc, generic extracts	- surface linguistic analysis - redundancy detection - topic identification - relevance detection - summary generation
COMPENDIUM$_{QE}$	single- or multi-doc, query-focused extracts	- surface linguistic analysis - redundancy detection - query similarity - topic identification - relevance detection - summary generation
COMPENDIUM$_{SE}$	single- or multi-doc, sentiment-based extracts	- surface linguistic analysis - redundancy detection - subjective information detection - topic identification - relevance detection - summary generation
COMPENDIUM$_{E-A}$	single- or multi-doc, generic abstractive oriented summaries	- surface linguistic analysis - redundancy detection - topic identification - relevance detection - information compression and fusion - summary generation

Table 5.5. Variants of COMPENDIUM.

2002 corpus, whereas on the other hand, this corpus together with the ones for DUC 2003 and DUC 2004 are used for generating multi-document summaries.

Table 5.6 shows the results COMPENDIUM$_E$ obtains for the different newswire corpora. As it can be seen, single-document summaries achieve better results (around 45% for ROUGE-1) than multi-document ones (30% on average). The difference in performance may be due to the fact that we do not employ any specific technique for tackling multi-document summarisation. In contrast, we merge all related documents into a single one, and this is used as input for COMPENDIUM$_E$.

		Single-document	**Multi-document**
DUC 2002	ROUGE-1	0.45611	0.30137
	ROUGE-2	0.20252	0.05327
	ROUGE-L	0.41382	0.26373
DUC 2003	ROUGE-1	-	0.28977
	ROUGE-2	-	0.05481
	ROUGE-L	-	0.25399
DUC 2004	ROUGE-1	-	0.31091
	ROUGE-2	-	0.06316
	ROUGE-L	-	0.27633

Table 5.6. COMPENDIUM$_E$ results for single- and multi-document summarisation for the newswire domain, F-measure ($\beta = 1$).

An example of a single- and a multi-document summary generated by COMPENDIUM$_E$ is shown in Table 5.7.

Furthermore, we carry out a comparison of our results with respect to the best scoring system in the respective DUC editions, as well as a Lead baseline that builds the summary taking the first sentence of each document in the case of single-document summarisation, and it takes the first sentence of the first document, then the first sentence of the second document, and so on, for multi-document. In these cases, we report the recall value for ROUGE-1. Table 5.8 shows such comparison.

The single-document summaries achieve very good performance compared to the best system at DUC 2002, and the Lead baseline. In this case, COMPENDIUM$_E$ outperforms the best system by approximately 8%, and an increase of 12% is obtained over the baseline. On the contrary, results for multi-document summarisation are not as good. The Lead baseline is improved by a 33% and 40% for DUC 2002 and 2003 data, respectively, but there is a marginal increase for DUC 2004 (0.2%). Despite these improvements, the performance of COMPENDIUM$_E$ for multi-document summarisation does not surpass the best system at DUC. This difference is due to the fact that we approach multi-document summarisation as a single-document summarisation where all related documents are considered as a single one. This is an indication that this type of summarisation may require a more elaborate processing instead.

COMPENDIUM$_E$ (single-document summary):
What do Charlie Chaplin, Greta Garbo, Cary Grant, Alfred Hitchcock and Steven Spielberg have in common?
They have never won Academy Awards for their individual achievements.
Oscar's 60-year history is filled with examples of the film world's highest achievers being overlooked by the Academy of Motion Picture Arts and Sciences.
The honorary award has also proved useful to salve the Academy's conscience.
Douglas Fairbanks, Judy Garland, Noel Coward, Ernst Lubitsch, Fred Astaire, Gene Kelly, Harold Lloyd, Greta Garbo, Maurice Chevalier, Stan Laurel, Cary Grant, Lillian Gish, Edward G. Robinson, Groucho Marx, Howard Hawks and Jean Renoir are others who have received honorary awards.
COMPENDIUM$_E$ (multi-document summary):
Oscar, manufactured by the R.S. Owens Co., Chicago, is made of Britannia metal, copper plate, nickel plate and gold plate.
They have never won Academy Awards for their individual achievements.
Oscar's 60-year history is filled with examples of the film world's highest achievers being overlooked by the Academy of Motion Picture Arts and Sciences.
The honorary award has also proved useful to salve the Academy's conscience.
How long was the longest Oscar ceremony?
Free enterprise has collided with the Academy Awards, and everybody's trying to pick up the pieces.

Table 5.7. Single- and multi-document summaries generated by COMPENDIUM$_E$ for document AP880325-0239 (DUC 2002, cluster d078b).

	DUC 2002		**DUC 2003**	**DUC 2004**
TS system	Single	Multi	Multi	Multi
Best DUC participant	0.42776	**0.35151**	**0.37980**	**0.38232**
COMPENDIUM$_E$	**0.46008**	0.30341	0.29355	0.31362
Lead baseline	0.41132	0.22771	0.20967	0.31293

Table 5.8. Comparison of COMPENDIUM$_E$'s performance with other text summarisation systems (recall value).

- **Image caption corpus**. Generic and query-focused summaries were produced using COMPENDIUM$_E$ and COMPENDIUM$_{QE}$, respectively. In total, 308 summaries of 200 words each were generated. To evaluate both approaches, we use ROUGE-2 and ROUGE-SU4[9], and compare the automatic summaries against the model summaries also provided in this corpus.

[9] ROUGE parameters: -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -l 200

Furthermore, as baseline the first 200 words of the Wikipedia article describing each image were selected. Also, for comparison purposes, we generated summaries employing different state-of-the-art TS systems. In particular, these were: SummGraph (Plaza *et al.*, 2008), MEAD (Radev *et al.*, 2004), and SUMMA (Saggion, 2008). In this way, the performance for generic and query-based summarisation for such systems was also compared. The results are given in Tables 5.9 and 5.10, respectively.

TS system	**ROUGE-2**	**ROUGE-SU4**
Wikipedia baseline	0.09632	0.14203
SummGraph	0.08950	0.14290
MEAD	0.08866	0.13769
COMPENDIUM$_E$	0.08551	0.13371
SUMMA	0.06423	0.10919

Table 5.9. Results for generic summarisation in the image caption corpus (recall value).

TS system	**ROUGE-2**	**ROUGE-SU4**
Wikipedia baseline	0.09632	0.14203
SummGraph	0.10075	0.15430
MEAD	0.10192	0.15353
COMPENDIUM$_{QE}$	0.08864	0.13892
SUMMA	0.06532	0.10946

Table 5.10. Results for query-focused summarisation in the image caption corpus (recall value).

From the results obtained, we can conclude that query-focused summarisation is more appropriate for this type of data. If we observed the results for the summarisers, all ROUGE scores in Table 5.10 (query-focused summaries) are higher than the ones in Table 5.9 (generic summaries). Moreover, a Wilcoxon signed-rank test is computed for assessing the significance of the results. For all summarisers, except SUMMA, the query-focused summaries are significantly better than the generic ones.

Regarding our TS approach, COMPENDIUM, the results for the query-focused approach (COMPENDIUM$_{QE}$) increase by approximately a 4% on average with respect to the generic one (COMPENDIUM$_{E}$). Wikipedia summaries are a difficult goal to achieve, as it can be seen from the results where only two summarisers, when producing query-focused summaries (SummGraph and MEAD), obtain higher results. The reasons why it is so difficult to perform better than the Wikipedia baseline are: 1) these articles have been created by humans; 2) the first paragraph in a Wikipedia article is usually a summary of the entire document content; and 3) Wikipedia articles almost exclusively contain salient information to the subject matter, and do not present other information somehow related to the topic but not important (e.g. nearby hotels, or transport services).

Next, we try to elucidate the rearons why COMPENDIUM performs lower than SummGraph and MEAD. Analysing the type of documents and the resulting summaries, we have realised that one of the main problems resides in the nature of the corpus. Most documents in the corpus contain sentences with a high number of noun-phrases, but which are unrelated to the topic (e.g ‘‘*Mahogany, Maple, crown mouldings, multiple Viking ovens, Sub-Zero refrigerators, antique...*”). Since COMPENDIUM gives more importance to sentences containing longer noun-phrases, according to the CQP feature, these types of sentences are scored higher. Therefore, they are wrongly considered relevant, and they are final incorporated in the summary. In these cases, the quality of the final summaries is directly affected by these sentences.

Finally, two examples of summaries for this corpus is shown in Table 5.11. The summary at the top corresponds to a generic summary produced using COMPENDIUM$_{E}$, whereas the one at the bottom is a query-focused one (COMPENDIUM$_{QE}$).

COMPENDIUM$_E$ (generic summary):
Nou Camp in both Spanish and English is a football stadium in Barcelona, Spain.
The stadium has been the home of FC Barcelona since The stadium's construction in 1957.
The stadium is a UEFA 5-star rated stadium, and has hosted numerous international matches at senior level, and UEFA Champions League finals, the most recent being in 1999.
The stadium has a capacity of 98,772, making The stadium the largest stadium in Europe, and the eleventh largest in the world.
The stadium's official name was Estadi del FC Barcelona FC Barcelona Stadium until 2000, when the club membership voted to change the official name to the popular nickname, Camp Nou.
Across Camp Nou is the Palau Blaugrana, the stadium for indoor sports and adjacent is the Ice Rink, the stadium for ice-based sports.
Just behind the complex is the Mini Estadi, the stadium where the FC Barcelona Atlètic plays the FC Barcelona Atl tic's games.
By the early 1950s, Barcelona had outgrown Barcelona's old stadium, Camp de Les Corts which had held 60,000 supporters.
With the outlawing of standing sections at the stadium in the late 1990s, stadium's capacity settled to just below 99,000.
COMPENDIUM$_{QE}$ (query-focused summary):
The Camp Nou "new field", Catalan pronunciation : ['kam 'now], often erroneously called the "Nou Camp" in both Spanish and English is a football stadium in Barcelona, Spain.
The stadium's official name was Estadi del FC Barcelona FC Barcelona Stadium until 2000, when the club membership voted to change the official name to the popular nickname, Camp Nou.
Across Camp Nou is the Palau Blaugrana, the stadium for indoor sports and adjacent is the Ice Rink, the stadium for ice-based sports.
By the early 1950s, Barcelona had outgrown Barcelona's old stadium, Camp de Les Corts which had held 60,000 supporters.
The Camp Nou, built between 1954 and 1957, was designed by architects Francesc Mitjans-Mir, Lorenzo Garc a Barbon and Josep Soteras Mauri.
FC Barcelona won Eulogio Martínez first game at Camp Nou in impressive fashion, a 4-2 victory against Legia Warsaw, with Eulogio Martínez scoring the first goal at the new stadium.
On September 18, 2007, British architect Norman Foster and Foster's company were selected to "restructure" the Camp Nou.
The upper ring lowers to give way to a roof structure above the main tribune.
The height of the stadium offers a great sensation of majesty.

Table 5.11. Generic and query-focused summaries generated by COMPENDIUM for the place *Camp Nou*.

- **Blog corpus**. The comments associated to each blog in the corpus are used to produce sentiment-based summaries (COMPENDIUM$_{SE}$). Different from the previous experiments, instead of generating summaries of a fixed length, we experimented with three different compression rates (10%, 15% and 20% of the document).

  The evaluation conducted also varies from the previous ones in the sense that ROUGE is not employed. Instead, a qualitative evaluation is conducted, where summaries are evaluated manually. It is worth stressing upon the fact that we focus more on the quality of the summaries rather than on their content, since the content would depend on the specific need a user has at a particular moment. Moreover, for this corpus we do not have model summaries, and to produce them manually is a difficult and time-consuming task.

  In particular, the criteria proposed for evaluating the opinion summaries are the following: *redundancy*, *grammaticality*, *focus* and *difficulty*. *Redundancy* measures the presence of repeated information in a summary. *Grammaticality* accounts for the number of spelling or grammatical errors that a summary presents. *Focus* evaluates whether it is possible or not to understand the topic of the summary, that is, the main subject of the text; and finally, *difficulty* refers to the extent to which a human can understand a summary as a whole or not.

  If we have a look at the criteria proposed in DUC and TAC conference, we will realise that we adopt more or less the same, except from the difficulty criteria which is non-conventional. The reason why this criterion is included is that it provides an idea of the summary as a whole, regarding its readability.

  Furthermore, three different degrees of goodness are established for each evaluated criteria. These were *non-acceptable*, *understandable* and *acceptable*. In this classification, *acceptable* means that the summary meets the specific criterion and therefore is good, whereas *non-acceptable* means that the summary would not be good enough with respect to a criterion. When assessing the *difficulty*, the summaries were classified with regard to *high*, *medium* and *low*, being low, the better. Table 5.12 shows the results for this evaluation.

  Analysing the results obtained, a set of interesting conclusions can be drawn. As far as the grammaticality criterion is concerned, the re-

		Compression Rate		
Criterion	**%**	**10%**	**15%**	**20%**
Redundancy	Non-acceptable	26	0	4
	Understandable	45	6	10
	Acceptable	29	94	86
Grammaticality	Non-acceptable	4	2	0
	Understandable	22	27	55
	Acceptable	74	71	45
Focus	Non-acceptable	33	26	14
	Understandable	43	29	47
	Acceptable	24	45	39
Difficulty	High	35	18	8
	Medium	28	35	51
	Low	37	51	41

Table 5.12. Results for COMPENDIUM$_{SE}$ within the blog corpus.

sults show a decrease of grammaticality errors as the size of the summary lowers. We can see that the number of acceptable summaries varies from 74% to 45%, for a compression rate of 20% and 10%, respectively. This is obvious, because the longer the summary, the more chances there are for containing orthographic or grammatical errors. Due to the informal language used in blogs, we thought a priori that summaries would contain many spelling mistakes. Contrary to our expectations, generated summaries are quite well-written, only 4% of them, at most, being non-acceptable. Another important fact that can be inferred from the results is related to how the summaries deal with the topic. According to the percentages shown in the tables presented previously, the number of summaries that have correctly identified the topic and have therefore been evaluated as acceptable, changes considerably with respect to the different summary sizes, increasing when we change from 10% to 15%, but decreasing when changing from 15% to 20%. However, as a general trend, we can see that when taking into account the number of summaries that have not performed correctly in the focus parameter, there is a decreasing trend, reducing the incorrect summaries from 33% to 14%. This means that for longer summaries, the topic may be stated along the summary, although not necessarily at the beginning of it, whereas for shorter summaries, there is no such flexibility, and as a consequence, if the topic does not appear at the beginning, the most probable thing is that it does not appear in the summary at all. Finally, regarding redundancy, results indicate that summaries of 15% and 20% contain less repeated information than shorter ones. What can be seen from the results is that the summaries of 20% compression rate obtain the best results on average over the rest of the size experimented with. This is due to the fact that this compression rate achieves higher percentage (for the *understandable* and *acceptable* degrees of goodness) in two (grammaticality and focus) out of the three criteria proposed. Only the 15% compression rate summaries obtained better results in the redundancy criterion. On the other hand, as far as the difficulty criteria concerned, results are also encouraging. According to the evaluation performed, the longer the summaries, the easier they are to be understood in general. Grouping the percentages of summaries, we obtained that 65%, 86% and 92% of the summaries of size 10%, 15% and 20%, respectively, have, either medium or low level of difficulty, which means that they could be understood as a whole without serious difficulties. Again, for this criterion, the 20% summaries achieve the best results. This has also been proven by previous research work, which demonstrated that this compression rate is more suitable for an acceptable quality of summaries (Morris *et al.*, 1992). It is worth mentioning that this criterion is rather subjective and depends to a large extent on different factors, such as the knowledge the person who reads the summaries, the number of grammatical errors the text contains, or the connectedness of the sentences. Moreover, it is reasonable to think that long summaries can be more difficult to understand, but our experiments show that is it actually the other way around, because longer summaries may contain more information than short ones, which allows the user to have more awareness of the content and what the summary is about.

In general terms, while evaluating the summaries obtained, we noticed some recurrent mistakes. The first one is the punctuation; in some cases we noticed some commas missing or instead of having a comma, contain a full stop. (e.g. "So. One option...") Also, in some cases, apostrophes are missing, in examples such as *dont.* The second error is that in some cases the summaries start with a sentence containing a coreference element that we cannot resolve, because the antecedent has been deleted or sentences that imply some concept previously mentioned in the original text have not been selected. It is also worth mentioning that some of the grammatical errors are due to users' misspellings, for example "I thikn". The third error concerns the spelling mistakes found in the summaries, which are directly transferred from the initial blog posts, which also contains such kind of errors (e.g. calender). Finally, we also found some void sentences, that do not contribute to the general meaning of the summary as for example, "I m an idiot", "Just an occasional visitor', or "welcome back!!!".

COMPENDIUM$_{SE}$ (sentiment-based summary):
Clothilde, I love the wallpapers!
They keep everything tasty and fresh!
Thanks a lot for the gorgeous calender desktop background.
What a great idea and beautiful photo.
I've just started recreating some of the easier and more attainable recipes.
Another lovely calendar! Clotilde, have you discontinued your "Bonjour mois" newsletter?
I'm terribly late this month but was enjoying the cheese so much that I just forgot! The peas are another winner of course.
My only quibble would be about the name.

Table 5.13. Sentiment-based summary generated by COMPENDIUM$_{SE}$ for blog 29.

Table 5.13 shows an example of an automatic summary for blog 29 with a compression rate of 10%. In this case, the summary contains mostly positive opinions, having only the last sentence a negative charge (*"My only quibble would be about the name"*).

As it can be seen, only opinions have been considered and these are presented grouped into positives, on the one hand, and negatives, on the other. We considered it as good due to the fact that there are no objectives or useless sentences. The system presents subjective sentences with an emotional charge, and as a consequence this summary meets our purposes.

- **Medical research papers corpus**. With this set of documents, we want to analyse the capabilities of COMPENDIUM for generating abstractive-oriented summaries. In this particular evaluation, we generate summaries of 162 words and our goal is to analyse to what extent resulting abstracts are valid. Therefore, we set up a comparison between COMPENDIUM$_{E}$ and COMPENDIUM$_{E-A}$.

Table 5.14 shows an example of two summaries generated with COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$, respectively. It is worth mentioning that these approaches produce generic summaries, and for generating them neither the keywords of the original article nor the information in the titles or in the abstract were taken into consideration. As it can be seen both summaries share some of the sentences, whereas others have been shorten in the latter.

Specifically, we set up three different types of evaluation. In the first one, we use the model abstracts of the articles and we compare them with the ones generated by COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$, for our extractive and abstractive-oriented approach, respectively. The second evaluation aims at determining to what extent our generated summaries contain the main topics of the research articles. Finally, in the third evaluation, we assess the summaries with regard to user satisfaction with respect to a 5-level Likert scale. Next, each of these types of evaluation is explained in more detail.

- **Comparison with human abstracts.**

  In this evaluation, the summaries generated by COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$ are assessed with respect to the human abstracts provided in the original articles. We use ROUGE-1[10] and we report the values of recall, precision and F-measure ($\beta = 1$). We also compare the results obtained with a state of the art summariser: MS-Word 2007 Summarizer[11], and we run a t-test to account for statistical significance of the results at a 95% confidence level (statistical significant results are marked with a star).

  Table 5.15 shows the results for the first evaluation. In this evaluation, COMPENDIUM$_{E-A}$ summaries are evaluated with respect to the human abstracts, and compared to COMPENDIUM$_E$ and MS-Word 2007.

  As it can be seen, our both TS approaches are comparable with respect to the state of the art TS tool (i.e., MS-Word 2007 Summarizer). Regarding COMPENDIUM$_{E-A}$, it is worth mentioning that the precision obtained is higher and statistically significant compared to the remaining approaches, thus meaning the information contained is the right one. However, its recall is

[10] ROUGE parameters:-n 1 -m -x -c 95 -r 1000 -f A -p 0.5 -t 0 -d -l 160

[11] http://www.microsoft.com/education/autosummarize.aspx

COMPENDIUM$_E$ (extractive summary):
Histologic examination of lesions plays a key role in the diagnostics of cutaneous lupus erythematosus LE .
LE has a broad spectrum of histological signs which are related to the stages of the lesions, but some signs apply to all stages e.g. mucin deposition.
Histologic findings of skin lesions are essentially identical for systemic lupus erythmatosus SLE and cutaneous LE.
From the histological standpoint, LE can be classified only into early, fully developed, late LE, and special manifestations of LE.
The early histologic findings of LE lesions are sparse superficial perivascular lymphocytic infiltrates, neutrophils and sometimes nuclear dust immediately beneath the dermoepidermal junction.
Few individual necrotic keratinocytes and focal vacuolar alteration of basal cells may occur.
Fully developed lesions are characterized by moderately dense to dense perivascular and periappendageal lymphocytic infiltrates in the papillary and reticular dermis with abundant mucin deposition in the reticular dermis.
According to Kuhn et al. the presence of even slight epidermal or junctional involvement should exclude LE tumidus.
COMPENDIUM$_{E-A}$ (abstractive oriented summary):
LE lesions plays a key role in the diagnostics.
LE has a broad spectrum of histological signs which are related to the stages of the lesions, but some signs apply to all stages e.g. mucin deposition.
LE lesions are essentially identical for systemic lupus erythmatosus SLE.
LE can be classified only into early histologic.
LE lesions are sparse superficial perivascular lymphocytic infiltrates neutrophils and sometimes nuclear dust immediately beneath the dermoepidermal junction.
Few individual necrotic keratinocytes and focal vacuolar alteration of basal cells may occur.
Fully developed lesions are characterized by moderately dense to dense perivascular and periappendageal lymphocytic infiltrates in the papillary and reticular dermis with abundant mucin deposition in the reticular dermis.
According to Kuhn et al the presence of even slight epidermal or junctional involvement should exclude.

Table 5.14. Example of summaries generated with COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$ for the 4th medical document in the corpus.

lower, so in the end the final value of F-measure is negatively affected, being COMPENDIUM$_E$ statistical significant with respect to COMPENDIUM$_{E-A}$ for these specific measures. This is due to the fact that for this TS approach we rely on the sentences detected as important in the relevance detection stage, and we compress or merge some information within them. Therefore, the result-

TS Approach	**Recall**	**Precision**	$\mathbf{F}_{\beta=1}$
COMPENDIUM$_E$	**0.44022***	0.40525	**0.42201***
COMPENDIUM$_{E-A}$	0.38658	**0.41809***	0.39533
MS-Word 2007	0.43610	0.40456	0.41974

Table 5.15. ROUGE-1 results for COMPENDIUM$_{E-A}$ and its comparison with other TS approaches.

ing summaries are shorter than the extracts, and since no extra information is added, the recall value will be never higher than it is for COMPENDIUM$_E$. One possible solution to address this issue would be to rely on the source document and generate the new sentences from it instead of the most relevant sentences. Another strategy would be to include in the COMPENDIUM$_{E-A}$ summary the next highest ranked sentence in the document according to the relevance detection stage that were not included in the extract, because of summary length restrictions. Regarding COMPENDIUM$_E$, it achieves slightly better results than MS-Word 2007; however, there are not statistical differences between them. Only statistically significant results are obtained with respect to COMPENDIUM$_{E-A}$.

- **Topic identification.**

  The objective of this evaluation is to assess the generated summaries with respect to the topics they contain. Together with the content of the article and the abstract, a number of keywords were also included (5 on average). These keywords usually reflect the most important topics dealt in the article. Consequently, we analyse to what extent such keywords appear in the summaries generated by COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$. If our summaries are able to contain such keywords, it will mean that they are indicative of the content of the source document, and therefore, they will be appropriate to provide an idea of what the article is about. In order to compute the number of keywords a summary contains, we calculate how many of them are included in the summary, and we divide this result by the total number of keywords the corresponding article has. Table 5.16 shows the results obtained when we evaluate the topics included in the summaries.

	% Correct Topics			
	< 25%	< 50%	< 75%	75-100%
COMPENDIUM$_E$	5%	12.5%	47.5%	35%
COMPENDIUM$_{E-A}$	7.5%	17.5%	42.5%	32.5%

Table 5.16. Percentage of topics resulting summaries contain.

As it can be seen, a considerable percentage of summaries are able to reflect at least half of the topics of the articles (82.5% and 75%, for COMPENDIUM$_E$ and COMPENDIUM$_{E-A}$, respectively). It is worth stressing upon the fact that our approaches produce generic summaries and in none of the cases, the keywords provided in the article were taken into account in the summarisation process. Some of summaries generated employing the COMPENDIUM$_{E-A}$ approach do not contain as many topics as the ones for COMPENDIUM$_E$. This occurs because in the former approach the resulting summaries contain sentences that may have been compressed, so in some of these cases, there is a loss of information, although minimal.

- **User satisfaction study.** In the last evaluation, we aim at assessing the user satisfaction with respect to the generated summaries. For this purpose, we perform a qualitative evaluation and we asked 10 humans to evaluate our summaries[12] according to a 5-level Likert scale (1= strongly disagree...5=strongly agree). For each summary, humans were asked to respond to three questions concerning the appropriateness of the summaries. The questions asked, as well as the percentage of summaries for each question in a scale 1 to 5 are shown in Table 5.17 and Table 5.18, respectively.

Q1: The summary reflects the most important issues of the document.
Q2: The summary allows the reader to know what the article is about.
Q3: After reading the original abstract provided with the article, the alternative summary is also valid.

Table 5.17. Qualitative questions to evaluate the summaries.

[12] The humans were also provided with the original articles and their abstracts.

As it can be seen from the results shown in Table 5.18, our abstractive-oriented approach (COMPENDIUM$_{E-A}$) obtains better results than the extractive one (COMPENDIUM$_{E}$). Although the evaluation concerning the information contained in the summaries generated with COMPENDIUM$_{E-A}$ was not as good as for the extractive approach, taking into consideration their quality from a human point of view, the abstractive-oriented summaries are much better than the extractive ones. When we have a look at the different percentages of summaries that have been rated in one of each categories, we observe that there is a higher percentage of abstractive-oriented summaries that humans agree with, compared to the extractive summaries for the same rating. Moreover, it is worth stressing upon the fact that, analogously, the percentage of summaries with lower ratings (strongly disagree and disagree) also decrease when COMPENDIUM$_{E-A}$ is employed.

%	**TS Approach**	**Q1**	**Q2**	**Q3**
1. Strongly disagree	COMPENDIUM$_{E}$	9.76	19.51	19.51
	COMPENDIUM$_{E-A}$	2.44	0	2.44
2. Disagree	COMPENDIUM$_{E}$	41.46	19.51	34.15
	COMPENDIUM$_{E-A}$	31.37	21.95	31.71
3. Neither agree nor disagree	COMPENDIUM$_{E}$	24.39	29.27	26.83
	COMPENDIUM$_{E-A}$	21.95	29.27	26.83
4. Agree	COMPENDIUM$_{E}$	21.95	21.95	7.32
	COMPENDIUM$_{E-A}$	41.46	39.02	34.15
5. Strongly agree	COMPENDIUM$_{E}$	2.44	9.76	12.20
	COMPENDIUM$_{E-A}$	2.44	9.76	4.88

Table 5.18. User satisfaction results for the different text summarisation approaches.

Furthermore, concerning the average individual scoring results (between 1 to 5), COMPENDIUM$_{E-A}$ achieves at most 3.37 for Q2 and 3.1 for Q1 and Q3, whereas the maximum average value for COMPENDIUM$_{E}$ is 2.83 for Q2, the remaining questions obtaining values lower than 2.60. In light of the results obtained, it has been proved that the combination of extractive and abstractive techniques is more appropriate and leads to better summaries than extracts.

5.5 Conclusion

This chapter presented the exhaustive evaluation, results and discussion of COMPENDIUM. On the one hand, this evaluation consisted in assessing the individual features COMPENDIUM relies on, i.e., textual entailment for detecting and removing redundancy; term frequency for identifying the topics of a document; and the Code Quantity Principle for determining the relevance of a sentence. On the other hand, the different types of summaries COMPENDIUM is able to generate (single- and multi-document, generic, query-focused, sentiment-based, and abstractive-oriented summaries) were also evaluated taking into account different domains and textual genres; in particular, newswire, image captions, blogs and medical research papers.

To summarise, the main contributions of this chapter were to:

- **Assess the appropriateness of the proposed techniques in COMPENDIUM.** In particular, textual entailment, term frequency and the Code Quantity Principle were shown to be useful for generating summaries.
- **Assess the capabilities of COMPENDIUM for generating different types of summaries belonging to different domains and textual genres.** COMPENDIUM was evaluated in a wide range of corpora, in order to analyse its performance across domains and types of texts. In addition, different kinds of summaries were produced, showing the suitability of COMPENDIUM as a TS tool.
- **Show the competitiveness of COMPENDIUM in the state of the art of TS.** The results obtained and the comparison made against other systems prove that COMPENDIUM obtains competitive results compared to other TS approaches, outperforming them in some of the cases.

Furthermore, the evaluation performed in this chapter allow us to be aware of the advantages and limitations of COMPENDIUM. Concerning its advantages, it is important to mention its capability of producing a wide range of summaries (single- and multi-document, generic, query-focused, sentiment-based and abstractive-oriented). At the moment, the tool is mono-lingual, and therefore it has been only evaluated for English. In order to adapt it to other languages, such as Spanish, we would need specific linguistic resources for the target language, for instance, a list of stop words, or noun-phrases detector. As far as the redundancy detection method is concerned, it is worth noting that when using textual entailment to address this problem, the entailment is only detected in one direction and it is not reciprocal. However, having shown that it is appropriate for detecting redundant information, it could be feasible to compute the entailment also in the opposite direction, as simulating a paraphrases detection tool. Another possibility would be to carry out an analysis on how a paraphrase detection tool performs in removing redundant information and conduct a comparison between both to select the best approach. Regarding the main limitations of COMPENDIUM, there are two priority issues that need to be faced in light of the results obtained. On the one hand, it is important to analyse different strategies to face the multi-document summarisation problem, that allows us to decide which one to follow. Therefore, multi-document summarisation would be addressed differently as it is in COMPENDIUM so far (only by merging all input documents into a single-one). On the other hand, in order to improve the quality of the resulting summaries, incorporating semantic knowledge to the systems would be essential. This knowledge will allow us to determine, extract, generalise and generate new information from the information contained in the source documents. For facing this challenge we should employ semantic resources, such as WordNet (Fellbaum, 1998) or ontologies.

6. COMPENDIUM in Human Language Technology Applications

6.1 Introduction

Automatic summaries, despite not being perfect, can be very useful for being integrated into other HLT applications (e.g. question answering). In this way, a summary or a Text Summarisation (TS) tool can also be evaluated extrinsically, not focusing on the content of a summary, but also how good it is to help other applications to meet their goals[1].

Therefore, the objective of this chapter is to analyse to what extent the use of COMPENDIUM is appropriate when it is integrated into HLT applications. In particular, COMPENDIUM is studied under three major HLT applications: opinion mining, question answering, and text classification (in the specific task of the rating inference). The aim of its integration into opinion mining is to generate better sentiment-based summaries. In question answering, the goal is to employ query-focused summaries to find the answers of factual questions, and for text classification, the use of generic, query-focused and sentiment-based summaries is analysed for predicting the correct rating associated to a review. The purpose of combining our TS tool with these applications is to study if the use of summaries, instead of the full document, is beneficial for them, thus improving their final performance with respect to the initial versions without taking summaries into consideration. This will also allow us to evaluate the summaries generated with COMPENDIUM in an extrinsic manner.

This chapter is structured in three sections, each one addressing the analysis of COMPENDIUM in a different HLT application. First, Section 6.2 reports the study when COMPENDIUM is applied to opinion mining. The aim here is to analyse the benefits of integrating both

[1] Please see Chapters 2 and 3 for more information about the combination of summaries with other HLT applications and the extrinsic evaluation, respectively.

HLT applications within the same approach. The starting point is our participation in the *Opinion Summarization Pilot* task in TAC 2008 and, consequently its guidelines, the proposed methods as well as the results obtained are discussed in Subsection 6.2.1. Then, an improved version of the TAC 2008 approaches where COMPENDIUM is integrated is provided in Subsection 6.2.2. The second section (Section 6.3) deals with the application of TS in question answering. Subsection 6.3.1 explains how COMPENDIUM is integrated into a Web-based QA system, providing information about each stage involved in the proposed combined approach. The remaining subsection (Subsection 6.3.2), describes all issues concerning the experiments and the results obtained when the proposed architecture is tested using a set of factual questions. In the third section (Section 6.4) of the chapter, COMPENDIUM is integrated into text classification. Specifically, summaries are used as a "noise filtering" for the rating inference task, which is a specific type of fine-grained granularity text classification. In the first subsection, the methods and tools needed for setting up the experimental framework are explained (Subsection 6.4.1), whereas in the last, the experiments developed and the results obtained are shown (Subsection 6.4.2). Finally, the main conclusions drawn are provided in Section 6.5, where a brief discussion of the overall performance of COMPENDIUM in the HLT applications selected is provided.

6.2 COMPENDIUM in Opinion Mining

Opinion mining (also known as sentiment analysis) is the research field which deals with the computational treatment of opinion, sentiment, and subjectivity in text (Pang & Lee, 2008).

Currently, available information containing users' opinions (e.g. forums, reviews, blogs, etc.) has considerably increased on the Internet. Such information is very useful for the decision making process. For instance, when a person wants to buy a laptop, he/she would like to know in advance the opinions of other users who own the same laptop. Within this scenario, sentiment analysis is essential to automatically detect and classify subjective text. Nevertheless, a human will still not be able to read and process all the available information concerning a specific topic, product, etc. Therefore, TS techniques are of great help in order to provide users a summarised fragment of text containing all the requested information, thus not having to read the individual opinions.

In this section, we integrate COMPENDIUM into an opinion mining approach and we carry out an analysis of the benefits it leads to. As a starting point, we report our participation in the TAC 2008 *Opinion Summarization* task (Subsection 6.2.1), and further on, we enhance the approach presented in this competition with COMPENDIUM. We finally compare both approaches and we present the results obtained (Subsection 6.2.2).

6.2.1 Participation in TAC 2008: *Opinion Summarization* Task

In order to analyse COMPENDIUM within the opinion mining research area, we take as a basis our participation in the *Opinion Summarization* task organised within the Text Analysis Conference[2] (TAC). Therefore, we are going to briefly describe the objectives of this task and the proposed approach.

Description of the task. In 2008, TAC conferences were first organised as a successor of DUC conferences. Among the proposed tasks, one of them, the *Opinion Summarization* task[3], consisted in generating summaries from blogs, according to specific opinion questions provided by the TAC organisers. Specifically, given a set of blogs from the Blog06 collection[4] and a list of questions, participants had to produce a summary that answered these questions. The questions generally required determining the opinion expressed on a target, each of which dealt with a single topic (e.g. George Clooney). Additionally, a set of text snippets were also provided, which contained the answers to the questions. These snippets were provided by real Question Answering applications, and the participants in the *Opinion Summarization* task could either use them or choose to perform themselves the retrieval of the answers to the questions in the corresponding blogs. In total, there were 609 blogs grouped into 25 targets, being at most two questions associated to each one. Table 6.1 depicts an example of target, question, and optional snippets.

[2] www.nist.gov/tac/

[3] http://www.nist.gov/tac/2008/summarization/op.summ.08.guidelines.html

[4] http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html

Target:	George Clooney
Questions:	Why do people like George Clooney? Why do people dislike George Clooney?
Snippets:	1050 BLOG06-20060205-018-0000647869 Yes George Clooney is a spoiled punk he judges his own people while having a silver spoon in his mouth. 1050 BLOG06-20060209-006-0013539097 he's a great actor

Table 6.1. Example of target, question, and optional snippets from the TAC 2008 data.

TAC 2008 approach. For participating in the *Opinion Summarization* task, we considered two different approaches, which mainly differed in the use of the optional text snippets provided by the TAC organisation. The first approach (*Snippet-driven*) used these snippets, whereas the second (*Blog-driven*) found the answers directly in the corresponding blogs by computing the similarity between the question and each sentence in the blog.

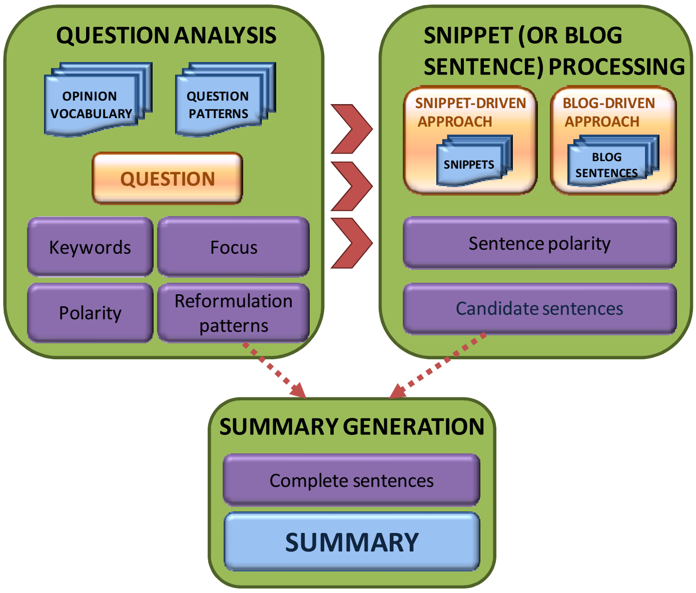


Fig. 6.1. Proposed architecture for TAC 2008.

Figure 6.1 depicts the proposed generic architecture for both approaches (*Snippet-driven* and *Blog-driven*). In this architecture, three main stages can be clearly distinguished: i) question analysis; ii) snippets processing (or, in the case we do not use such snippets, the sentences related to the question found directly in the blogs); and iii) summary generation. Next, each of this stages is explained in more detail.

- **Question analysis**

  For each question, we extract its topic, keywords and polarity. The topic refers to all named entities[5] that the question contains. If no named entities are found, the topic will be identical to the keywords. Keywords are all the words of the question, except the stop words. The polarity of the question, that is, if the question has a positive or negative nature, is determined extracting its nouns, verbs, adverbs, adjectives together with their determiners and classifying them using different resources, such as WordNet Affect (Strapparava & Valitutti, 2004) or emotion triggers (Balahur & Montoyo, 2008a). Furthermore, a set of reformulation patterns are also obtained. These patterns will be used to link sentences and give coherence to the final summaries. For instance, the pattern *"People like X because"* is extracted from the question *"Why do people like X?"*.

  At the end of this stage, we obtain the reformulation patterns together with the topic, keywords, and polarity of each question. Table 6.2 shows an example of the information obtained in this stage for the question *"Why do people like George Clooney?"*.

- **Snippet (or blog sentence) processing**

  Depending on whether we use the provided snippets or not, we will address this stage differently. On the one hand, if the snippets are used, the whole sentence where this snippet comes from, and the topic of the sentence together with its polarity will be detected and extracted. On the other hand, if we do not use such snippets, we will compute the similarity[6] between the question and all the sentences in the blogs in order to find the most similar ones. Previous to this

[5] Freeling (http://garraf.epsevg.upc.es/freeling/) is employed for this task.

[6] Text Similarity is employed: http://www.d.umn.edu/t̃pederse/text-similarity.html

Question:	Why do people like George Clooney
Keywords:	people like George Clooney
Focus:	George Clooney
Polarity:	positive
Reformulation patterns:	People like George Clooney because One reason why people like George Clooney is Another reason people like George Clooney is It is said that people like George Clooney because A further motivation people like George Clooney

Table 6.2. Information obtained in the question analysis stage.

step, we will preprocess the blogs, transforming them to plain text, splitted into sentences.

Once we have a set of candidate sentences containing the answer to the question, the next step is to extract the topic and polarity of each sentence in the same manner we extracted them for the questions. Then, they have to be mapped onto the corresponding question, since one target may have more that one question, being one positive and other negative. Four rules were then created to map each snippet/sentence with the corresponding question:

1. If there is only one question made on the topic, determining its polarity is sufficient for making the correspondence between the question and the snippets retrieved; the retrieved snippet must simply obey the criteria that it has the same polarity as the question.
2. If there are two questions made on the topic and each of the questions has a different polarity, the correspondence between the question and the answer snippets can simply be done by classifying the snippets retrieved according to their polarity.
3. If there are two questions that have different focus but different polarities, the correspondence between the questions and the answer snippets is done using the classification of the answer snippets according to focus and polarity.
4. If there are two questions that have the same focus and the same polarity, the correspondence between the questions and the answer snippets is done using the order of appearance of the entities in

focus, both in the question and in the possible answer snippet retrieved, simultaneously with the verification that the intended polarity of the answer snippet is the same as that of the question.

Once each sentence is mapped with the correct question, the next stage is to generate the summary, which is detailed next.

- **Summary generation**

  This is the last stage of the approach. Basically, it consists of grouping together the sentences with the same polarity ranked in decreasing order, and use the different reformulation patterns to link all the sentences until the maximum length is reached. However, it can occur that some of the sentences are incomplete. In order to avoid incorporating such sentences into the final summary, we first carry out a syntactic analysis[7] to discard those ones which are not complete (i.e., not containing subject and verb).

  It is worth stressing upon the fact that the TS techniques employed by our approaches in the *Opinion Summarization* task are minimal, since we ordered the sentences according to their polarity score and extracted the top scored ones, adding at the beginning of each sentence one reformulation pattern.

  Table 6.3 shows two examples of summaries following our proposed approaches (Snippet- and Blog-driven). The summaries correspond to the target *"George Clooney"*, which includes the following questions: *"Why do people like George Clooney?"* and *"Why do people dislike George Clooney?"*. Due to the fact that summaries are very long (each summaries contains on average 14,000 non-white spaces characters), the whole summaries cannot be reproduced verbatim.

TAC 2008 Results. A total of 19 groups participated in the *Opinion Summarization* task and since each team could submit up to three different approaches, 45 runs were received but only the first two for each team were evaluated.

Table 6.4 shows the final results obtained by our approaches: *Snippet-driven* and *Blog-driven*. Also, the rank among the 36 participating systems is shown in brackets for each evaluated criteria. In particular, the criteria evaluated were: grammaticality; non-redundancy;

[7] MINIPAR is used for this task: http://www.cs.ualberta.ca/ lindek/minipar.htm

Snippet-driven Approach:
One reason people like George Clooney is Clooney is a fantastic actor and it would make your wife very happy, because she is a big fan of him.
Another reason people like George Clooney is people have fallen in love with him.
People dislike George Clooney because politik Ditto: George Clooney is an idiot.
A further motivation people dislike George Clooney is they can tell their first-hand, that George get his pretty boy, elitiest attitude naturally.
It is said that people dislike George Clooney because not one penny of them will go for any movie featuring George Clooney. [...]
Blog-driven Approach:
It is said that people like George Clooney because George Clooney mocked lobbylist at awards show.
People like George Clooney because Clooney was part of that machine, and figured it out partway through.
A further motivation people like George Clooney is Clooney is hot.
People like George Clooney because there is a Read more in George Clooney – terrorist lover.
One reason people dislike George Clooney is as for George Clooney, they's a well-known fact that he's a bad actor who gets by on his good looks and charm.[...]

Table 6.3. Examples of summaries generated with the Snippet- and the Blog-driven approaches, respectively.

structure and coherence; overall fluency and readability; and overall responsiveness. These criteria were manually assessed by a group of expert judges, who gave a score between 1 (the worst) to 10 (the best). In addition, F-measure ($\beta = 1$) was obtained employing the Pyramid Method for evaluating the content of the summaries with respect to a list of references nuggets (i.e., short fragments of text stating a relevant piece of information), each of them associated to a score, depending on how relevant it was.

As it can be noticed from the results obtained, our system performed well regarding F-measure ($\beta = 1$), the *Snippet-driven* approach being classified 7th among the 36 evaluated. As far as the structure and coherence is concerned, the results were also good, placing the *Snippet-driven* approach in the fourth. Also worth mentioning is the good performance obtained regarding its overall responsiveness, it ranked 5th. Generally speaking, the results for the *Snippet-driven* approach showed well-balanced among all the criteria evaluated, except for non-

	Approach	
	Snippet-driven	**Blog-driven**
Grammaticality	4.727 (8/36)	3.545 (36/36)
Non-redundancy	5.364 (28/36)	4.364 (36/36)
Structure and coherence	3.409 (4/36)	3.091 (13/36)
Overall fluency and readability	3.636 (16/36)	2.636 (36/36)
Overall responsiveness	5.045 (5/36)	2.227 (28/36)
F-measure ($\beta = 1$)	0.357 (7/36)	0.155 (23/36)

Table 6.4. TAC 2008 *Opinion Summarization* task results.

redundancy and overall fluency and readability. For the *Blog-driven* approach, results were generally poor. In this approach, the method employed for identifying sentences containing the answer to each question was very simple. It only took into account the similarity between the question and each sentence in the blog, so answers were not as accurate as the ones provided by TAC organisers. Therefore, this had a negative influence in the results obtained in this approach.

When comparing our approaches separately, in both cases, they did not perform very well with respect of the non-redundancy criterion, nor the overall fluency and readability. None of our approaches developed a specific step for tackling redundancy. Instead, this was addressed in a very naïve manner: only identical sentences were considered as redundant. Regarding the overall fluency and readability, the results show that we should improve this criterion. Our *Blog-driven* approach performed poorer than the *Snippet-driven.* In the case of the Blog-driven approach, the summaries did not obtain satisfactory results as far as the grammaticality criterion is concerned. The spelling errors in the summaries were derived in most of the cases from the original blogs. An analysis carried out in (Lloret & Palomar, 2011a) showed that 100% of the blogs contained spelling errors (other factors were also studied), and therefore, summaries relying on sentence extraction had high probabilities of suffering also from spelling errors. However, an interesting thing that is worth mentioning concerning the results obtained is that the use of reformulation patterns, in order to generate sentences for completing the summaries was appropriate, leading to very good positions according to the structure and coherence criterion.

6.2.2 Enhancing the TAC 2008 Approach with COMPENDIUM

In light of the results obtained in the *Opinion Summarization* task, we decided to analyse whether the use of TS techniques using COMPENDIUM, which takes also into consideration the redundancy problem, can improve the previous results obtained at TAC. Therefore, instead of carrying out the summary generation stage as previously described, a variant of COMPENDIUM[8] was employed to generate the final summary. In particular, such variant[9] used only textual entailment to avoid redundancy and term frequency to account for the topics of the documents, which also served to score each candidate sentence, in order to select and extract the top ranked ones.

In Table 6.5, an example of a summary obtained when COMPENDIUM is integrated into the *Snippet-driven* approach is provided.

Snippet-driven Approach + COMPENDIUM:
I'm sure all of us remember the tasteless remarks George Clooney made about Charlton Heston and his suffering from Alzheimer's disease when Mr. Clooney was receiving an award from the National Board of Reviews in 2003.
Seems this has been the worst year of his life.
George Clooney definitely gave a stellar performance and the movie was beautifully directed and he's still totally cute and I'd still marry him - but he has a long way to go when it comes to writing and telling a good story.
Not one penny of mine will go for any movie featuring George Clooney.
But George is a Hollywood actor, so arrogance and narcissism come with the territory.
Mark Nicodemo: George Clooney with Inflated Sense of Self Mark Nicodemo: George Clooney with Inflated Sense of Self I think he s a dingbat sometimes, but he s also smarter than the typical Hollyweird Leftist.
Clooney holds no regrets about Abramoff joke Wire Report LOS ANGELES - George Clooney, who may be giving speeches again at next month's Academy Awards, says he has no regrets about making an off-color joke about disgraced lobbyist Jack Abramoff during last month's Golden Globes.

Table 6.5. Example of a sentiment-based summary with COMPENDIUM.

The resulting summaries were evaluated following the same guidelines of TAC 2008, i.e., using the Pyramid nuggets provided in the

[8] Please see Chapter 4 for more detail about this text summarisation tool.

[9] This variant correspond to the one explained in Chapter 5, Section 5.4.2, “ Only Term Frequency (TF)” part.

competition. Table 6.6 shows the results obtained with COMPENDIUM in comparison with the *Snippet-* and *Blog-driven* approaches, and the two top ranked systems according to the F-measure ($\beta = 1$) value. The table shows the values for recall and precision as well.

Approach	**Recall**	**Precision**	$\mathbf{F}_{\beta=1}$
Best TAC system	-	-	0.534
Second best TAC system	-	-	0.490
Snippet-driven	0.592	0.272	0.357
Blog-driven	0.251	0.141	0.155
Snippet-driven+COMPENDIUM	**0.684**	**0.630**	**0.639**
Blog-driven+COMPENDIUM	0.292	0.282	0.262

Table 6.6. Results for COMPENDIUM in the *Opinion Summarization* task (TAC 2008).

As it can be seen from the results obtained, the F-measure ($\beta = 1$) improves for both approaches (80% and 70%, for the Snippet-driven and the Blog-driven, respectively). Moreover, the use of a more elaborate redundancy detection method is beneficial, and results are not negatively affected. Unfortunately, we were not able to carry out a manual assessment of the remaining criteria also evaluated at TAC (grammaticality, non-redundancy, etc.), so in this respect we cannot know whether these criteria might have also improved or not. What we have proven is that use of a TS tool, like COMPENDIUM, combined with an opinion mining approach is appropriate and leads to better results than carrying out opinion mining on its own.

6.3 COMPENDIUM in Question Answering

Question Answering (QA) applications allow users to formulate questions in natural language and provide them with the exact information required, also in natural language. They differ from information retrieval applications in that, instead of focusing on the retrieval of relevant documents given a query made by a user, QA enables users to pose precise questions and obtain specific answers, with no extra information. A special type of QA applications are those based on the Internet (Web-based QA applications) to find answers directly in the brief texts that Internet search engines attached to the results and provide to users, also known as snippets, but one of the main drawbacks of relying on these snippets is associated to their quality. On the one hand, such snippets are incomplete fragments of text in most of the cases. For instance, if one enters the question *"who is the* ***head*** *of* ***Spanish football team***[10] *?"* in Google, the snippets retrieved will be in the form of *"The Spanish national football team represents Spain in international football and is . . . In his second World Cup as Spain's coach, Clemente led his team . . . "*. On the other hand, the snippets do not always contain the answer of such questions, especially if the keywords found in the question do not appear close within the text, as for instance in the following snippet *"12 Jul 2010 . . . The latest news on Spain national* ***football team****, from thousands of sources . . .* ***head*** *as they celebrate on a stage set up for the* ***Spanish*** *team . . . "*.

We therefore propose the use of TS techniques using the query-focused version of COMPENDIUM (COMPENDIUM$_{QE}$[11]), for solving the limitations snippets present. Additionally, we analyse the influence summaries have on finding the correct answers with respect to snippets. These aspects are developed in Subsection 6.3.1. Moreover, a broad experimental set-up and evaluation is also carried out in Subsection 6.3.2.

6.3.1 Integrating COMPENDIUM in a Web-based QA approach

COMPENDIUM$_{QE}$ is combined with the Web-based QA approach developed in (Moreda *et al.*, 2008). This Web-based QA approach uses search engine snippets to extract answers to specific questions. However, when COMPENDIUM$_{QE}$ is integrated, it was adapted to substitute summaries for the snippets. Consequently, the resulting *Web-based QA+*COMPENDIUM$_{QE}$ approach collects the full texts of the resulting Web pages and summarises them in order to be used for finding the answer to a question. The architecture of the combined approach is illustrated in Figure 6.2.

As shown in this figure, the approach is divided into three sequential modules: i) *question analysis*, ii) *information retrieval* iii) *summarisa-*

[10] The keywords for this question are shown in boldface.

[11] This approach is explained in Chapter 4.

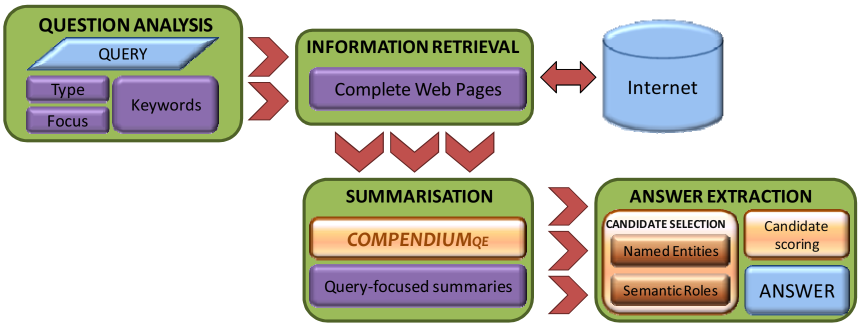


Fig. 6.2. Web-based QA approach with COMPENDIUM$_{QE}$.

tion and iv) *answer extraction*. All these modules are very important in the QA process because the success of the approach as a whole strongly depends on the individual success of each one. Next, we explain each one in detail.

- **Question analysis.** The goal of the *question analysis* module is to extract all relevant information from the question. Specifically, this module extracts the *question type*, the *focus* and the *keywords*. In order to extract the question type, the approach uses a hand-crafted set of rules, and for extraction of the focus and the keywords it uses the syntactic dependency parsing information in the question[12]. The focus has been defined as the noun depending on the "what/which" question words. The keywords are defined as the heading elements of the noun phrases and the main verbs of the question. Then, the focus, if it exists, is omitted from the keywords. For example, in the question *"Which city is the capital of France?"*, *"city"* is the focus and *"is"*, *"capital"* and *"France"* are the keywords.
- **Information retrieval.** The *information retrieval* module obtains the relevant information from the previously extracted question keywords. Typically, this information consists of a set of documents or document passages retrieved from a limited and predefined collection. However, in this approach we use an Internet search engine

[12] MINIPAR (http://webdocs.cs.ualberta.ca/l̃indek/minipar.htm) has been used to obtain the dependency tree

(i.e. Google[13]) to retrieve the first 20 documents (Web pages) associated to a specific question. These Web pages are the input to the summarisation stage next explained.

- **Summarisation.** In this stage, we integrate COMPENDIUM$_{QE}$ into the Web-based QA approach. The whole HTML content of each of the 20 documents obtained in the previous stage is extracted and then used as input for COMPENDIUM$_{QE}$ taking into consideration different lengths. It is worth stressing upon the fact that the original Web-based QA approach used the snippets attached to the first 20 results retrieved by Google to find the answers to the questions. By integrating COMPENDIUM$_{QE}$, such answers will be sought in the summaries instead of in the snippets. The generated summaries are single-document, query-focused and extractive.
- **Answer extraction.** Finally, the aim of the *answer extraction* module is to find candidate answers for the question formulated by the user within the retrieved information. These answers must be exact, not containing any extra information (e.g. "Which is the capital of Spain?"[Madrid]). Regarding the candidate answer selection process, the Web-based QA system relies on two approaches:
  - **Answer extraction based on Named Entities (NE-based QA**: In this approach, a NE recogniser[14] is applied to the text and all entities corresponding to the previously detected question type are selected as candidate answers.
  - **Answer extraction based on Semantic Roles (SR-based QA)**: In this approach, a SR labelling tool (Moreda *et al.*, 2007) is applied to the text and selects as candidate answers the arguments with specific roles contained in sentences that match specific role patterns depending on the detected question type.

  Once the candidate answers to a question are selected, in order to determine the final answer, we score them with regard to the following criteria: i) their distance to the keywords; and ii) the relevance of their n-grams (unigrams[15] and bigrams[16]). Finally, the best scored

[13] http://www.google.com/

[14] LingPipe: http://alias-i.com/lingpipe/

[15] The relevance of unigrams is calculated using the probability of a n-gram in the text.

[16] The relevance of bigrams is measured using the mutual information, which measures the mutual dependence of the two random words.

answered is returned as the answer, provided that its score is greater than a predefined threshold, which is automatically determined with respect to the number of retrieved results. This avoids the system to return a low scored answer.

6.3.2 Experiments and Results

In order to assess the benefits of incorporating COMPENDIUM$_{QE}$ into a Web-based QA approach, we compare it with respect to a Web-based QA approach that uses the retrieved snippets for finding the correct answer to a specific question. Furthermore, we evaluate the impact of COMPENDIUM$_{QE}$ within the two approaches that were proposed to find the answers to a question (NE-based QA and SR-based QA). With this purpose, the evaluation environment and the results obtained are next explained in detail.

- **Corpus.** A set of 100 factual questions of four different types was collected[17] (person, location, temporal and organisation). Within these types, two types of answers can be found: named entities and common nouns. Some of these questions were taken from the TREC data sets[18], whereas those ones whose answer was a common noun were manually created by external users. Table 6.7 shows two examples of each type of question together with their corresponding answers.

  The whole Web was then used as a corpus for collecting documents. In particular, Google search engine was used and, as it was previously mentioned, the top 20 results were taken into account. We collected the snippets attached to these first 20 results for the Web-based QA approach, and we produce summaries of the documents retrieved using COMPENDIUM$_{QE}$, for the Web-based QA+COMPENDIUM$_{QE}$ approach. Table 6.8 shows an example of these two types of texts.
- **Summary's length.** Concerning the length of the summaries, in order to carry out a fair comparison both snippets and text summaries must have a similar length. Therefore, we first conducted an analysis of the retrieved snippets for each type of question, finding

[17] 25 questions for each type.

[18] TREC-8: http://trec.nist.gov/data/qa/t8_qadata.html and TREC-9: http://trec.nist.gov/data/qa/t9_qadata.html

Question Type	**Question**	**Answer**
Location	Where is Ocho Rios located?	Jamaica
	Where is pancreas located?	abdomen
Organization	Which company is Steve Jobs CEO of?	Apple
	Which Spanish company was founded by Amancio Ortega in 1975?	Zara
Temporal	When was Mozart born?	January 27, 1756
	When was the first Barbie produced?	1959
Person	Who is the voice of Miss Piggy?	Frank Oz
	Who was the lead actress in Sleepless in Seattle?	Meg Ryan

Table 6.7. Examples of questions.

Snippet	It located in the Kalahari Desert. The population was 6591 in 2001 census.... Orapa. Orapa. Orapa is a town located in east-central Botswana. ....
Text Summary	The Kalahari Desert is a large, arid desert area in southwestern Africa is the world's second-largest and second most-populous continent, after Asia.

Table 6.8. Example of snippet and text summary for the question *"Where is the Kalahari desert located?"*.

that this length was on average 31 words (26, 27, 35, and 36 words, for location, temporal, organization and person question types, respectively). We initially established this same average length for the summaries, thus being referred to as COMPENDIUM$_{QE}$ *short*. With the purpose of experimenting with a different length, and analyse its consequences, summaries at twice as the previous length for each type of question (i. e., 52, 54, 70, and 72 words, respectively) were also generated. These summaries are referred to as COMPENDIUM$_{QE}$ *long* in our comparison framework.

Finally, 2,000 snippets (500 for each question type) were collected and 4,000 summaries (1,000 summaries for each type of question), were generated (20 summaries for each question x 100 questions x 2 summary lengths).

- **Evaluation methodology.** For evaluating the performance of both Web-based QA applications (snippets vs. COMPENDIUM$_{QE}$), answers are judged as *Correct*, *Incorrect* or *Not answered* by two human assessors. A question is considered:

  – *Correct*, if the answer retrieved by the QA system matches with the correct answer;

  – *Incorrect*, if the answer obtained by the QA system is not *Correct*;

  – *Not answered*, if the approach does not give any answer.

  Taking into account such values, we use recall, precision and F-measure ($\beta = 1$) to evaluate each of the approaches.

- **Results.** We next show the results obtained for each of the Web-based QA approaches (NE-based QA and SR-based QA), comparing the use of snippets versus the summaries generated with COMPENDIUM$_{QE}$. In addition, a Lead baseline was also defined. This baseline extracted the first sentences of each Web page, which contained the topic of the question, up to a specific length which, in our case, corresponded to the same summary length generated for each type of question dealt with, which was previously explained. In the end, we had two baselines: Lead-short and Lead-long.

  – **Results for NE-based QA**

  Table 6.9 contains the results for the Web-based QA approach, when a NE recogniser is used in order to identify candidate answers. For each of the possible summary lengths (i.e., short and long summaries, with average mean of 31 and 62 words, respectively), we compare the performance of COMPENDIUM$_{QE}$ with respect to the Lead baseline and the original snippets.

  As it can be seen, when using COMPENDIUM$_{QE}$, the results improve with respect to any other approach. Considering the average results for F-measure ($\beta = 1$), the best approach is COMPENDIUM$_{QE}$ for short summaries, which improves by approximately 12% the results obtained with the snippets and both baselines, and by 5% COMPENDIUM$_{QE}$ the results for long summaries. Furthermore, as far as the type of question is concerned, COMPENDIUM$_{QE}$ for short summaries performs the best with temporal questions, and then, with location, person and organization questions.

Approach		**Question type**				
Name	**%**	**person**	**organization**	**temporal**	**location**	**Avg.**
Snippets	Pre	60.0	54.5	55.0	62.5	58.0
	Rec	48.0	50.0	44.0	60.0	50.5
	$F_{\beta=1}$	53.3	52.2	48.9	61.2	53.9
Lead baseline short	Pre	57.9	47.4	66.7	64.0	59.0
	Rec	44.0	36.0	56.0	64.0	50.0
	$F_{\beta=1}$	50.0	40.9	60.9	64.0	53.9
COMPENDIUM$_{QE}$ short	Pre	61.9	60.0	75.0	66.7	65.9
	Rec	52.0	48.0	60.0	64.0	56.0
	$F_{\beta=1}$	**56.5**	**53.3**	**66.7**	**65.3**	**60.5**
Lead baseline long	Pre	54.5	42.9	70.0	62.5	57.5
	Rec	48.0	36.0	56.0	60.0	50.0
	$F_{\beta=1}$	51.1	39.1	62.2	61.2	53.4
COMPENDIUM$_{QE}$ long	Pre	59.1	60.0	63.2	66.7	62.2
	Rec	52.0	48.0	48.0	64.0	53.0
	$F_{\beta=1}$	55.3	53.3	54.5	65.3	57.6

Table 6.9. Comparison between NE-based QA approach with snippets and COMPENDIUM$_{QE}$.

Now, focusing on the size of the summaries, it is worth stressing that short summaries offer better F-measure ($\beta = 1$) average results than long ones. This happens in the Lead baselines, as well as for COMPENDIUM$_{QE}$. The reason for this is that there is less noise, that is to say less useless extra information in the shorter summaries and therefore in the answers that are selected by the Web-based QA approach. Web pages often include information concerning advertisements, links, unrelated tables and pictures, etc. which is not directly related to the specific information the Web-based QA approach needs. The aim of TS is to distinguish between relevant and irrelevant content, but it can happen that the irrelevant content also includes query terms, thus increasing the difficulty associated to the TS task. In these cases, the longer the summaries, the higher chances they have to include useless information, and therefore the results of our final Web-based QA approach when TS is integrated will be affected, as it has been shown in the results (COMPENDIUM$_{QE}$ *short* performs better than COMPENDIUM$_{QE}$ *long*).

If the results for F-measure ($\beta = 1$) using the snippets (53.9% on average) is compared to the corresponding ones obtained integrating COMPENDIUM$_{QE}$ in the Web-based QA approach, it can be seen that the latter are better than those of the snippets, regardless the summary size (60.5% and 57.6%, for COMPENDIUM$_{QE}$ short and long, respectively).

Hence, these results support our initial hypothesis that using text summaries has considerable advantages as far as finding the correct answer in a Web-based QA approach is concerned, and overcomes the limitations of using only search engines snippets.

– **Results for SR-based QA**

These results concerned the Web-based QA approach when a SR labelling tool is employed for determining the potential candidate answers to a question. As for the previous approach, we compare the performance of COMPENDIUM$_{QE}$ against the Lead baseline and the original approach based solely on snippets. In Table 6.10 the results obtained by the SR-based QA approach are shown.

As shown in the the table, the results obtained for the SR-based version of the QA approach, are similar but not identical to the ones obtained using the NE-based version. Again COMPENDIUM$_{QE}$ *short* obtains the best results focusing on the F-measure value ($\beta = 1$), increasing the results by 53% with respect to the use of snippets, and by 44% when compared to the Lead baseline also for short summaries. Regarding the type of questions, location type achieves the best average F-measure results for all the analysed approaches, followed by temporal, person and organization.

Focusing on the size of the summaries, it can be seen that short summaries offer better average results than do longer ones in most of the cases, except for the Lead baseline for long summaries, whose F-measure average results are slightly higher than the same baseline with shorter size (Lead baseline short). In this manner, the previous analysis done in the NE-based QA approach is verified as far as summary length is concerned as well. As it happened with the NE-based QA approach, Web pages usually contain extra information not directly related to its main content. Therefore, the longer the summary, the higher the probability of including such useless content, thus influencing negatively in the overall results.

Approach		**Question type**				
Name	**%**	**person**	**organization**	**temporal**	**location**	**Avg.**
Snippets	Pre	56.3	33.3	50.0	55.0	48.6
	Rec	36.0	20.0	16.0	44.0	29.0
	$F_{\beta=1}$	43.9	25.0	24.2	48.9	35.5
Lead baseline short	Pre	46.7	50.0	62.5	70.6	57.5
	Rec	28.0	20.0	20.0	48.0	29.0
	$F_{\beta=1}$	35.0	28.6	30.3	57.1	37.7
COMPENDIUM$_{QE}$ short	Pre	68.8	77.8	80.0	65.2	72.9
	Rec	44.0	28.0	48.0	60.0	45.0
	$F_{\beta=1}$	**53.7**	**41.2**	**60.0**	**62.5**	**54.3**
Lead baseline long	Pre	50.0	62.5	56.3	50.0	54.7
	Rec	32.0	20.0	36.0	48.0	34.0
	$F_{\beta=1}$	39.0	30.3	43.9	49.0	40.6
COMPENDIUM$_{QE}$ long	Pre	64.7	50.0	64.7	62.5	60.5
	Rec	44.0	28.0	44.0	60.0	44.0
	$F_{\beta=1}$	52.4	35.9	52.4	61.2	50.5

Table 6.10. Comparison between SR-based QA approach with snippets and COMPENDIUM$_{QE}$.

Now, comparing the results obtained using the snippets with the ones obtained with COMPENDIUM$_{QE}$, it is worth mentioning that both sizes for the summaries generated with COMPENDIUM$_{QE}$ (short and long) are better than the former in all the question types evaluated, increasing by 53% and 42% respectively, as far as the F-measure value ($\beta = 1$) is concerned. Again, this verifies the analysis performed for NE-based approach.

Apart from verifying the main points of the analysis conducted for the NE-based QA approach, there is an important difference in the results obtained for the SR-based QA approach. This difference is the magnitude of the improvement achieved using summaries generated with COMPENDIUM$_{QE}$ rather than snippets. While for the NE-based QA approach the average improvement obtained is around 10%, the average improvement obtained in the SR-based QA approach is 48%. This difference occurs because when applying SR-based QA the quality of the text that has to be processed to find the answer is crucial. The text has to be analysed at different language levels and specifically at the semantic level using a semantic roles labelling tool. When using snippets, which many times consist of parts of incomplete sentences or sentences without a verb, the SR-based methods find it very difficult to process them at that high language-level analysis. For this reason, when using query-focused summaries generated with COMPENDIUM$_{QE}$, in which the sentences are not only complete but also the most relevant sentences in the text regarding the keywords much better results are obtained.

6.4 COMPENDIUM in Text Classification: the Rating Inference Task

As it was explained in Chapter 2 (Section 2.4.3), the goal of text classification is to automatically identify the category of a document, chosen from a predefined set of categories (Sebastiani, 2002). Within this task, we can speak about the *rating inference*, as the task of identifying the author's evaluation of an entity with respect to an ordinal-scale based on the author's textual evaluation of the entity (Pang & Lee, 2005). In the Web 2.0, it is frequent to find texts where users give a score, depending on how much they liked or not a product, movie, restaurant, hotel, service, etc. which is normally associated with a rating scale (1=worst..5=best). The interest of the rating inference task is to automatically predict such rating, instead of classifying the text according to its nature (subjective vs. objective) or its polarity (positive vs. neutral vs. negative), goals of the sentiment analysis task.

The *rating inference* task is in general considered a difficult problem because of the fuzziness of mid-range ratings (Mukras *et al.*, 2007), and when addressing this task, most approaches derive features for the classification task from the full document. In contrast to these approaches, our objective is to analyse the capabilities of different variants of COMPENDIUM for correctly associating a fine-grained rating (1=worst,...5=best) to a review. These variants consist of generic summaries – COMPENDIUM$_{E}$; query-focused summaries – COMPENDIUM$_{QE}$; and sentiment-based summaries – COMPENDIUM$_{SE}$. In particular, we want to investigate whether extracting features from document summaries could help a classification system, since text summaries are meant to contain the essential content of a document.

In the following subsections we explain how the rating inference task is approached, and how summaries are used (Subsection 6.4.1). Furthermore, the evaluation carried out and the results obtained together with a discussion is also provided (Subsection 6.4.2).

6.4.1 Methods and Tools for the Rating Inference Process

Different types of summaries generated using COMPENDIUM are used to extract features to serve as input for a classification system based on Support Vector Machine (SVM) (Joachims, 1998). In order to give an overview of the whole process carried out, we need to employ different methods and HLT-based tools, which are next explained in detail.

- **Linguistic analysis**

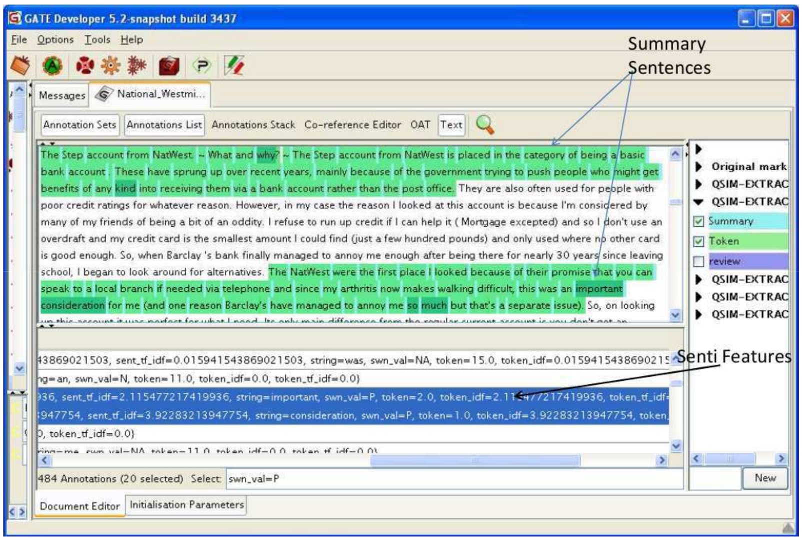
  Linguistic analysis of textual input (either the full document or the corresponding summary) is carried out using the General Architecture for Text Engineering (GATE) – a framework for the development and deployment of language processing technology in large scale (Cunningham *et al.*, 2002).

  Different GATE components, such as tokenisation, part of speech tagging, or morphological analysis are used to produce annotated documents with this kind of information. From the annotations we produce a number of features for document representation. In particular, these features are: *string* – the original, unmodified text of each token; *root* – the lemmatised, lower-case form of the token; *category* – the part-of-speech (POS) tag, a symbol that represents a grammatical category such as determiner, present-tense verb, past-tense verb, singular noun, etc.; *orth* – a code representing the token's combination of upper- and lower-case letters. In addition to these basic features, "sentiment" features based on a lexical resource are computed as explained below.

  Figure 6.3 shows an example of the linguistic analysis carried out using GATE for a summary.

- **Sentiment features extraction**

  When dealing with subjective texts, it is very important to take also into consideration the vocabulary that contains a subjective charge, since it can give clues about the general feeling of the whole text.



The Step account from NatWest. ~ What and why? ~ The Step account from NatWest is placed in the category of being a basic bank account . These have sprung up over recent years, mainly because of the government trying to push people who might get benefits of any kind into receiving them via a bank account rather than the post office. They are also often used for people with poor credit ratings for whatever reason. However, in my case the reason I looked at this account is because I'm considered by many of my friends of being a bit of an oddity. I refuse to run up credit if I can help it (Mortgage excepted) and so I don't use an overdraft and my credit card is the smallest amount I could find (just a few hundred pounds) and only used where no other card is good enough. So, when Barclay 's bank finally managed to annoy me enough after being there for nearly 30 years since leaving school, I began to look around for alternatives. The NatWest were the first place I looked because of their promise that you can speak to a local branch if needed via telephone and since my arthritis now makes walking difficult, this was an important consideration for me (and one reason Barclay's have managed to annoy me so much but that's a separate issue). So, on looking up this account it was perfect for what I need. Its only main difference from the regular current account is you don't get an

```
43869021503, sent_tf_idf=0.015941543869021503, string=was, swn_val=NA, token=15.0, token_idf=0.0159415438690215
ng=an, swn_val=N, token=11.0, token_idf=0.0, token_tf_idf=0.0}
936, sent_tf_idf=2.115477217419936, string=important, swn_val=P, token=2.0, token_idf=2.115477217419936, token_tf_idf=
947754, sent_tf_idf=3.92283213947754, string=consideration, swn_val=P, token=1.0, token_idf=3.92283213947754, token_
0, token_tf_idf=0.0}
ring=me, swn_val=NA, token=11.0, token_idf=0.0, token_tf_idf=0.0}
```

484 Annotations (20 selected) Select: swn_val=P

Fig. 6.3. Linguistic analysis of a summary using GATE.

Therefore, we also considered this issue, and extract from the reviews a sentiment-based feature, using SentiWordNet (Esuli & Sebastiani, 2006). SentiWordNet is a lexical resource in which each synset (set of synonyms) of WordNet (Fellbaum, 1998) is associated with three numerical scores *obj* (how objective the word is), *pos* (how positive the word is), and *neg* (how negative the word is). Each of the scores ranges from 0 to 1, and their sum equals 1. Since we are interested in the "general sentiment" of a word (i.e., if the word is generally "positive", generally "negative" or generally "neutral"), we compute it in the following manner: given a word w we compute the number of times the word w is more positive than negative (positive $>$ negative), the number of times is more negative than positive (positive $<$ negative) and the total number of entries of word w in SentiWordNet. For example, a word such as "good" has many more entries where the positive score is greater than the negativity score, while a word such as "unhelpful" has more negative occurrences than positive. We use this aggregated scores in our classification experiments. It is worth noting that we do not apply any word sense disambiguation procedure here.

- **Machine learning algorithm**

For this purpose, we adopt a SVM learning paradigm not only because it has recently been used with success in different tasks in natural language processing (Isozaki & Kazawa, 2002), but it has been shown particularly suitable for text categorisation (Kumar & Gopal, 2009) where the feature space is huge, as it is in our case. We rely on the SVM implementation distributed with the GATE[19] system (Li *et al.*, 2009b) which hides from the user the complexities of feature extraction and conversion from documents to the machine learning implementation. The tool has been applied with success to a number of datasets for opinion classification and rating inference (Saggion & Funk, 2009).

In particular, different features are used to train the SVM classifier with 10-fold cross validation, using the whole review: the *root* of each word, its *category*, the sentiment-based feature *SentiWordNet* (SentiWN), as well as their combinations.

- **Generating Summaries using COMPENDIUM**

  Since we want to analyse to what extent summaries are good to predict the correct rating of a review, COMPENDIUM is used to generate different types of summaries of different compression rates (from 10% to 50%). In particular, we focus on generic, query-focused, and sentiment-based. In Chapter 4, each of these approaches were explained in detail. Here, we only provide a general outline of them.

  – **Generic summarisation**. Generic summaries are those which contain the main ideas of one or various documents. We use COMPENDIUM$_E$ to address this task, which first removes redundant information from texts, and then determines the relevance on a sentence based on the topics it contains which are included in longer noun-phrases. As a result, the most relevant sentences are selected and extracted.

  – **Query-focused summarisation**. Query-focused summarisation generates summaries which contain the most important facts associated to a specific query, entity or topic. COMPENDIUM$_{QE}$ is capable of generating such sort of summaries, by employing an additional stage which computes the similarity of the query with each sentence of the document. It is important to mention that in this case, since we specifically focus on bank reviews, we consider

[19] http://gate.ac.uk/

the name of the bank as the query for each review. The result is a query-focused extractive summary.

- **Sentiment-based summarisation**. Sentiment-based summaries take into account the subjective information within a document, thus containing only positive information, negative, or a combination of both. This type of summaries are generated with COMPENDIUM$_{SE}$ which makes use of an external opinion mining tool (Balahur-Dobrescu *et al.*, 2009) to identify and classify the subjective information. Then, from the set of subjective sentences, the most relevant ones are extracted in order to produce the sentiment-based summary. It is worth noting here that the resulting summary is not biased towards only positive or negative information, but also a combination of both.

6.4.2 Experiments and Results

We next explain the evaluation environment and experimental framework that is set out to test the appropriateness of COMPENDIUM for the rating inference task.

- **Corpus.** We used a set of 89 reviews of several English banks (Abbey, Barcalys, Halifax, HSBC, Lloyds TSB, and National Westminster) gathered directly from the Internet. In particular, the documents were collected from *Ciao*[20], a Website where users can write reviews about different products and services, depending on their own experience. Each review is associated to a star (1=worst..5=best), which reflects the user opinion towards the bank.

  Table 6.11 lists some of the statistical properties of the data. It is worth stressing upon the fact that the reviews have on average 2,603 words, which means that we are dealing with long documents rather than short ones, making the rating inference task even more challenging. The shortest document contains 1,491 words, whereas the longest document has more than 5,000 words.

  Furthermore, an analysis concerning the number of reviews for each class (i.e., 1 to 5 stars) is also carried out. This allows us to have an idea whether or not we work with a balanced distribution. Table

[20] http://www.ciao.co.uk/

Num. of reviews	**Avg. length**	**Max. length**	**Min. length**
89	2,603	5,730	1,491

Table 6.11. Bank reviews corpus statistics.

6.12 shows this information. It is worth mentioning that one-third of the reviews belong to the 4-star class. In contrast, we have only 9 reviews that have been rated as 3-star, consisting of the 10% of the corpus, which is a very low number.

Star-rating	**Num. of reviews**	**%**
1-star	17	19
2-star	11	12
3-star	9	10
4-star	28	32
5-star	24	27

Table 6.12. Class distribution.

- **Evaluation methodology.** F-measure ($\beta = 1$) is used to assess the accuracy of the correct rating prediction. However, since we deal with a fine-grained classification problem (our categories range from 1 to 5 stars), we also take into consideration the *Mean Square Error* (MSE) (Urdan, 2005), which is capable of capturing the deviation of the prediction from the true class label. The MSE can be defined as

$$MSE = \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}{n}$$

where:
n is the total number of samples,
Y_i is the true class label,
$\widehat{Y}_i$ is the predicted class.

This measure has been previously used for text classification purposes (Mukras *et al.*, 2007) and it is also appropriate for our experiments because we are interested in analysing which summarisation approaches minimise the error in the star-rating prediction. As a consequence, if the original rating of the review belong to the 5-star class, and when using summaries for its rating prediction, it is rated as 4 stars, the MSE would be lower than if it was rated as 1 star. Other evaluation metrics such as F-measure ($\beta = 1$) would not account for this fact, and they would have only focused on the number of reviews correctly rated. In the aforementioned example, if we only used F-measure, both results would have obtained the same performance (i.e., 0), although it would be much better to be as close as possible to the original rating. For this reason, when evaluating the summaries in this task, we employ both measures (F-measure and MSE) in the evaluation.

- **Feature analysis.** Before using the summaries for predicting the correct rating of each review, we analyse the performance of the individual extracted features, as well as their combination using the full reviews. Specifically, we analysed the *root* of each word, its *category*, the sentiment-based feature *SentiWordNet*, as well as their combinations. Moreover, as a baseline for the full review, we took into account a totally uninformed approach with respect to the class with higher number of reviews, i.e. considering all documents as if they were scored with 4 stars, since this is the most frequent class. The different results according different features can be seen in Table 6.13.

Feature	$\mathbf{F}_{\beta=1}$
4-star rating baseline	0.300
root	0.378
category	0.367
sentiWN	0.333
root+category	0.356
root+sentiWN	0.333
category+sentiWN	0.389
root+category+sentiWN	**0.413**

Table 6.13. Classification results (F-measure) with the full review.

A first analysis derived from the results obtained in Table 6.13 makes us be aware of the difficulty associated to the rating inference task. As it can be seen, a baseline without any information from the document at all (by considering all reviews the most probably rating, i.e.

4 stars), performs around 30%, which compared to the remaining approaches is not a very bad number. However, we assumed that dealing with some information contained in documents, the classification algorithm will do better in finding the correct star associated to a review. This was the reason why we experimented with different features alone or in combination.

From these experiments, we obtained that the combination of linguistic and sentiment-based features leads to the best results, obtaining a F-measure value ($\beta = 1$) of 41%, thus increasing the results over the baseline by 38% approximately. If sentiment-based features are not taken into account, the best feature is the root of the word on its own, which performs 26% better than the baseline. Therefore, we extract the *root*, *category* and *sentiWN* features from the summaries, and we use them in an attempt to predict the correct rating of a review, instead of using the full review.

- **Results.** We next show and discuss the results obtained when using the summaries generated with COMPENDIUM for predicting the star of a review. We experimented with five different types of compression rates for summaries (ranging from 10% to 50%). Apart from comparing the performance of COMPENDIUM with respect to the full review, we defined an additional baseline: *Lead*, which produces a summary taking the first sentences according to a specific compression rate.

  As far as the features for training the summaries is concerned, it is worth mentioning that we take into account the best performing individual feature (i.e., the *root* of the words), and the one which combines all the features (*root+category+sentiWN*), since it was the best performing for correctly predicting a review using the full text.

  We next show the results obtained with respect to the F-measure ($\beta = 1$) and MSE when the feature *root* is used for training the summaries. Then, we carry out the same experiments but with the combination of all features (*root+category+sentiWN*) and we also report the results obtained.

  – **Results for *root*.**

  Table 6.14 and Table 6.15, shows the results for the F-measure ($\beta = 1$) and MSE, respectively when using the *root* feature extracted from summaries for training the SVM classifier. Results that improve the full review as well as the 4-star baseline are emphasised in boldface. The best approach (COMPENDIUM$_{SE}$ for 10% compression rate) improves the results obtained by the full review by a 18% approximately as far as F-measure ($\beta = 1$) is concerned.

Approach		**Compression Rate**				
Full Document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	$F_{\beta=1}$	0.378	0.378	0.378	0.378	0.378
4-star rating baseline	$F_{\beta=1}$	0.300	0.300	0.300	0.300	0.300
Summarisation approach						
Lead	$F_{\beta=1}$	**0.411**	**0.378**	0.367	0.311	0.322
COMPENDIUM$_E$	$F_{\beta=1}$	**0.422**	0.356	0.333	0.300	0.322
COMPENDIUM$_{QE}$	$F_{\beta=1}$	0.322	0.322	0.367	0.367	0.356
COMPENDIUM$_{SE}$	$F_{\beta=1}$	**0.446**	0.334	0.358	0.292	0.369

Table 6.14. Classification results (F-measure) using *root* as feature for the rating classification.

Approach		**Compression Rate**				
Full document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	MSE	2.59	2.59	2.59	2.59	2.59
4-star rating baseline	MSE	2.58	2.58	2.58	2.58	2.58
Summarization approach						
Lead	MSE	3.10	3.00	3.10	3.30	3.10
COMPENDIUM$_E$	MSE	2.70	2.93	3.00	3.10	2.71
COMPENDIUM$_{QE}$	MSE	**2.11**	**2.51**	**2.28**	**2.40**	**2.10**
COMPENDIUM$_{SE}$	MSE	2.69	3.21	3.46	2.90	2.93

Table 6.15. Classification results (MSE) using *root* as feature for the rating classification.

Whereas in Table 6.14 only COMPENDIUM$_E$ and COMPENDIUM$_{SE}$ for a 10% compression rate are superior to the full review, it does not happen the same when computing the MSE. In this case, there is a clear tendency that COMPENDIUM$_{QE}$ is more appropriate than the remaining approaches for all compression rates, because the error made in the classification process is lower than the full review.

As a first intuition, one could expect the sentiment-based summaries to perform better than the query-focused ones. However, according to the F-measure ($\beta = 1$), this occurs only for some compression rates (e.g. 10%), and with regard to MSE, all query-focused summaries obtained lower error rates. The reason why this may happen is the following: although the sentences are classified according to their polarity between positive or negative, this does not mean that one specific sentence is about the topic we are interested in (in our case, a particular bank). However, when generating query-focused summaries, this information is taken into consideration, giving also more importance to those sentences talking about the specific bank. Therefore, these summaries may contain more specific information about the given bank than sentiment-based summaries. We think that a good strategy would be the combination between query-focused and sentiment-based summaries, so that a summary would contain subjective information regarding a specific bank.

- **Results for *root+category+sentiWN.***

  Table 6.16 and Table 6.17 show the results for the F-measure ($\beta = 1$) and MSE, respectively, when the combination of all the linguistic and sentiment-based features (*root+category+sentiWN*) have been extracted and taken into account for training the SVM classifier. Results that improve the full review as well as the 4-star baseline are emphasised in boldface.

Approach		**Compression Rate**				
Summarisation method		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	$F_{\beta=1}$	0.413	0.413	0.413	0.413	0.413
4-star rating baseline	$F_{\beta=1}$	0.300	0.300	0.300	0.300	0.300
Summarization approach						
Lead	$F_{\beta=1}$	0.275	**0.422**	**0.422**	0.378	0.322
COMPENDIUM$_E$	$F_{\beta=1}$	**0.444**	0.411	0.411	0.311	0.322
COMPENDIUM$_{QE}$	$F_{\beta=1}$	0.356	0.378	0.356	0.367	0.356
COMPENDIUM$_{SE}$	$F_{\beta=1}$	**0.436**	**0.413**	**0.425**	0.359	0.324

Table 6.16. Classification results (F-measure) using *root*, *category* and *SentiWordNet*, as features for the rating classification.

Approach		Compression Rate				
Full document		**10%**	**20%**	**30%**	**40%**	**50%**
Full review	MSE	2.46	2.46	2.46	2.46	2.46
4-star rating baseline	MSE	2.58	2.58	2.58	2.58	2.58
Summarization method						
lead	MSE	2.63	2.62	2.72	2.86	2.60
COMPENDIUM$_E$	MSE	3.02	2.87	2.83	3.18	2.53
COMPENDIUM$_{QE}$	MSE	2.60	2.83	2.52	2.44	2.53
COMPENDIUM$_{SE}$	MSE	2.77	**2.31**	**2.14**	2.82	2.47

Table 6.17. Classification results (MSE) using *root*, *category* and *SentiWordNet* as features for the rating classification.

When using the combination of *root*+*category*+*SentiWN* as features for the rating inference process, the results obtained are better than for the previous experiment (when the root of a word was used as a single feature for training the SVM system).

Concerning the results obtained for the F-measure value ($\beta = 1$), it is worth stressing upon the fact that sentiment-based summaries generated with COMPENDIUM$_{SE}$ achieve better results than the full review for compression rates ranging from 10% to 30%. So does the Lead baseline for 20% and 30%, performing in those cases slightly better than our summarisation approaches. However, our best approach is COMPENDIUM$_E$ for a 10% compression rate, which increases the results of the full review by 8%.

Regarding the MSE results, we also obtain better results than when using only *root* for the full review. Some compression rates for sentiment-based summarisation (COMPENDIUM$_{SE}$) obtain better results than the remaining approaches. In particular, sentiment-based summaries of 20% and 30% are the best performing ones. This is explained by the fact that the classification process takes profit of sentiment-based features (*sentiWN*) employed.

Although we cannot claim that there exists a particular feature or strategy that contributes the most to the rating inference problem, this is partly due to the difficulty associated to this task. As it was previously explained, even when using the full review results are rather poor. In some cases (i.e., COMPENDIUM$_{QE}$ and COMPENDIUM$_{SE}$), summaries obtains at least similar results as the full review. Despite the results obtained by summaries are not very high, the use of TS in this task (in particular COMPENDIUM) for reducing noise in the full reviews has been shown appropriate. This is a positive thing, because it means that we do not have to process the whole text. Instead, a short summary (up to 30% of the text) would be useful, thus containing sufficient information to be able to perform the rating inference task.

6.5 Conclusion

This chapter presented the integration of COMPENDIUM in HLT applications. Specifically, these were: i) opinion mining; ii) question answering; and iii) text classification through the rating inference task. In light of the results obtained, it is worth mentioning that COMPENDIUM is appropriate for being integrated into other applications, improving the initial results obtained when summaries were not taken into account. The performance of COMPENDIUM in the different proposed HLT applications varied, obtaining the highest improvements in opinion mining (80% compared to the initial Snippet-driven approach), and the lowest in the rating inference task (only a moderate improvement for the F-measure value and MSE, with respect to the full review). Table 6.18 summarises the main results when COMPENDIUM is applied to different tasks.

	Initial Approach	**Improvement with COMPENDIUM**
Opinion mining	Snippet-driven	80%
	Blog-driven	70%
Question answering	NE-based QA	12%
	SR-based QA	48%
Rating inference	*root*	18%
	root+category+SentiWN	8%

Table 6.18. Summary of the improvements obtained by COMPENDIUM within each HLT application.

Therefore, the main contributions of this chapter can be divided into two groups:

- **Analyse the integration of COMPENDIUM to other HLT applications.** It has been shown that the integration of COMPENDIUM into other applications is beneficial, leading to improvements with respect to not using it. In this sense, we have proven that summaries extract the most relevant information of texts. For instance, in the case of opinion mining, from all the potential subjective sentences to be included in the summary, COMPENDIUM was able to extract the most important ones. Then, for question answering, very short summaries were generated in order to substitute the original snippets, and we showed that they were able to contain the answer to a question. Finally, for the rating inference task, we used summaries instead of full documents to correctly predict the star associated to a review. In this case, some types of summaries (e.g. sentiment-based) up to a 30% compression rate increased the results of the classification and minimised the error with respect to the full document.

- **Evaluate extrinsically the summaries generated with COMPENDIUM.** At the same time COMPENDIUM is integrated in other HLT applications, it is being evaluated extrinsically, thus providing us with an idea of how good the resulting summaries are for increasing the performance of external applications; in our case, opinion mining, question answering and text classification through the rating inference task. This type of evaluation complements the intrinsic one carried out in Chapter 5, thus verifying the appropriateness of the summaries generated by COMPENDIUM TS tool, for being used on their own, as well as in combination with HLT applications.

7. Conclusion and Work in Progress

Text Summarisation (TS) started in the late 1950's, and since then, summarisation methodologies and systems have experienced great advances. New approaches have been developed, taking also into account linguistic aspects, which allows an automatic summary to be something else than a simple joining of sentences. Furthermore, these advances have led to new types of summaries (e.g. update or personalised summaries) and new scenarios where summaries play a crucial role (e.g. patent claims, blogs, reviews). However, there is still a lot of room for improvement, especially due to the large amounts of available data in different formats, and the rapid development of the technology, which brings new challenges for this research field, such as multi-document, multi-lingual or multimedia summarisation. Another challenging issue is the evaluation, which raises questions not answered yet, such as *"is it possible to evaluate a summary in an objective way?"*; *"is it fair to compare automatic summaries against human-written?"*; *"how could the quality of a summary be assessed automatically?"*.

Nevertheless, being aware of the state of the art in TS is as important as knowing how summaries are produced from different perspectives. In Chapter 2, apart from providing a comprehensive review of different TS approaches, we explained the types of summaries according to different aspects. Moreover, the process of TS from a computational point of view (Hovy, 2005) was also described, whereas in Chapter 4, this process was explained from a cognitive perspective (Van Dijk, 1980), (Van Dijk & Kintsch, 1983). Following the underlying principles and statements of the cognitive and computational processes, we suggested COMPENDIUM TS tool which is able to generate different kinds of summaries for English, thus being a mono-lingual TS tool. The length of the summaries is given by a specific compression rate or a fixed number of words. Concerning the *input*, COMPENDIUM produces single- or multi-document summaries; with regard to the *purpose* of the resulting summaries, generic, query-focused or sentiment-based summaries can be generated. Moreover, all of them are informative, since their goal is to provide information about the source document. Finally, as *output*, the final summaries can be extracts or abstractive-oriented summaries, through a combination of extractive and abstractive information. Regarding COMPENDIUM's architecture, two types of stages can be distinguished: core and additional. The core stages constitute the backbone of the TS process, and they are: *surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*, whilst the additional stages can be integrated to the core ones, thus enhancing the capabilities of the TS tool. These are: *query similarity*; *subjective information detection*; and *information compression and fusion.*

It is worth noting that within the *redundancy detection* stage, textual entailment is proposed as a novel method to tackle this problem. In the *relevance detection* stage, the main novelty is to rely on the Code Quantity Principle, after identifying the most important topics using the term frequency technique, in order to determine important information in documents. The proposed techniques were shown appropriate for the generation of summaries. Furthermore, the assessment of COMPENDIUM both intrinsically (i.e., with respect to the content and quality of the generated summaries) with different types of texts (Chapter 5), as well as extrinsically, by integrating it into Human Language Technology (HLT) applications (Chapter 6), proved that the resulting summaries were good enough to be used on their own, as well as for being beneficial to other HLT applications.

In the next sections, the main contributions in the form of conclusions (Section 7.1), the work in progress and for the future (Section 7.2), and the list of relevant publications related to this thesis (Section 7.3) are provided.

7.1 Main Contributions

Summing up all the research work carried out in this thesis, the main contributions were:

- **Analysis of the state of the art with respect to the approaches and methods for generating summaries**

  From the exhaustive analysis performed, we were able to come up with some insights concerning the directions of TS for the next years. This is related to the Web 2.0 and all the types of new textual genres that have appeared: reviews, forums, wikis, blogs, etc. It will be of crucial importance to study methods and approaches that are capable of providing summaries from such types of texts. Moreover, the large amount of information provided by people with different backgrounds, and in different languages will lead to the urgent need of carrying out research into abstractive summarisation, as well as multi-lingual and cross-lingual summarisation techniques. This will allow TS approaches to manage the available information more efficiently, and enrich the resulting summaries with related information stated in different languages, formats, and with different intentions.

- **Analysis of the state of the art in the evaluation of summaries**

  Despite the large number of methods and tools for carrying out TS, at present, the evaluation of summaries is still a challenging issue and very difficult to tackle, even for humans. The inherent subjectivity associated to the TS process means that two humans can assess the same summary differently. The reason why this occurs is because each user has a specific interest or a different knowledge. In recent years, approaches have started to focus on the automatic evaluation of several aspects more related to the quality of the summary and not so much in its content.

- **Research into novel techniques and approaches for TS**

  In this thesis, we have proposed and study different novel methods for generating summaries. Concerning the redundancy problem, we proved that textual entailment is appropriate for identifying and discarding repeated information. Further on, we analysed the Code Quantity Principle as a feature for detecting relevant information in documents, obtaining successful results once the topics of the documents have been identified in the first place. Finally, we analysed the use of word graphs to compress and fuse information, and we suggest a new type of summaries (abstractive-oriented) which combines extractive and abstractive information and goes beyond the simple selection of sentences.

- **Proposal and development of COMPENDIUM TS tool**

  In light of the aforementioned techniques that were shown appropriate for the generation of summaries, we developed COMPENDIUM TS tool, which relies on different stages and it is able to generate several types of summaries. The stages involved in COMPENDIUM can be grouped into two categories. On the one hand, there are a set of core stages, which constitute the central part of COMPENDIUM. These stages are: *surface linguistic analysis*; *redundancy detection*; *topic identification*; *relevance detection*; and *summary generation*, and they are mainly responsible for detecting and removing redundant information, determining the topic or topics of the document, and finally, identifying relevant information. On the other hand, there are several additional stages (*query similarity*; *subjective information detection*; and *information compression and fusion*), which specifically deal with a type of summary: query-focused; sentiment-based; and abstractive-oriented. For instance, the query similarity stage aims at identifying potential sentences related to the query, in order to produce a query-focused summary.

- **Evaluation of COMPENDIUM**

  In order to verify the appropriateness of the proposed TS method, the content and the quality of summaries generated with COMPENDIUM have to be evaluated. It has been shown that COMPENDIUM achieves competitive results with respect to the state of the art in TS. Furthermore, from the extensive evaluation carried out we can draw some interesting conclusion about the strong and weak points of COMPENDIUM.

  Regarding its strengthens, COMPENDIUM generates different kinds of summaries, by employing the same core stages, and adding specific additional stages. In addition, the proposed techniques have been proven to work successful in a variety of domains and types of texts, such as newswire, blogs, image captions and medical research papers. In contrast, its main limitations involve two issues. On the one hand, multi-document summarisation has to be approached employing a different strategy, since it has been proved that the current one does not lead to very good results. On the other hand, it is crucial to incorporate semantic information in the process of TS, in order to be able to generalise and obtain new information, that can be later used for abstract generation. Finally, it is worth stressing upon the fact that, although at the moment, COMPENDIUM is mono-lingual, it would be feasible to extent it to other languages, given that the language-specific resources involved in the process of TS are available for the target language.

- **Integration of COMPENDIUM in HLT applications: opinion mining, question answering and text classification**

  Once the summaries generated with COMPENDIUM have been evaluated, it is important to analyse to what extent it can be integrated into other HLT applications. In light of this, we showed the benefits of combining our TS tool (i.e., COMPENDIUM) with opinion mining, question answering and text classification.

  Regarding the first task, COMPENDIUM was used to improve the summaries generated only using an opinion mining tool, without taking into account any summarisation technique. In the second task, question answering, it was found that when using query-focused summaries generated with COMPENDIUM instead of search engine snippets in a Web-based question answering approach, its performance increased. Finally, concerning text classification, we employed summaries (i.e., COMPENDIUM) in order to filter noisy information from documents, and to be able to correctly predict the rating associated to a review.

Therefore, the objectives proposed in Chapter 1 have been achieved through the research conducted in this thesis.

7.2 Work in Progress

This thesis constitutes only a small part in the TS research area. Therefore, there are still some issues that are being currently tackled, and some aspects that will be addressed in the future.

On the one hand, the aspects that are currently being investigated can be divided into the following groups:

- **Analysing other alternatives for generating abstractive summaries.** Taking as a basis the word graph algorithm proposed, we want to study to what extent generating the compressed or merged sentences directly from the original document rather than from the extract would be more suitable. In this case, we first generate the new sentences, and then we apply COMPENDIUM to select the most important ones. The results obtained at the moment, shows that for TS it seems to be better to select first important information, and then try to merge or compress it. However, we are analysing some strategies that allow us to select the best generated sentences, in order to see if the initial hypothesis still holds true.

- **Enriching COMPENDIUM with semantic knowledge.** This line of research broadens the research carried out into abstractive-oriented summarisation. The objective of this research is to incorporate semantic knowledge to COMPENDIUM by means of concept graphs or other semantic techniques. Semantic knowledge allows a higher level of abstraction. In this sense, compared to the word graphs algorithms, where only sentence compression or fusion was possible, the study of concept graphs is more appropriate for generating abstracts, since different concepts can be grouped into a more general one.

  For carrying out this, we are going to use semantic resources, such as WordNet, where the concepts of the document are mapped to the concepts of WordNet and then, their hypernyms are taken into account, in order to generalise several related concepts into the same.

- **Studying the integration of TS techniques into a opinion information retrieval application.** Currently, we are analysing the benefits of combining COMPENDIUM with information retrieval and opinion mining, in order to build a unified framework for dealing with new textual genres (blogs, reviews, etc.). This research is motivated by the good results achieved when applying COMPENDIUM to opinion mining. The proposed approach is capable of retrieving subjective content from the Web (for instance, blogs), analysing and classifying the opinions found in them, and finally producing a summary, that contains the specific information a user is looking for. Moreover, we are also analysing the contribution and influence of each of the individual tasks to the whole process. Different approaches and techniques for each of the tasks are analysed as well. For instance, in the case of information retrieval, we experiment with two passage lengths (i.e. passages of length 1 and 3).

The preliminary results obtained at the moment indicate that the best performing approach is the one which integrates information retrieval, opinion mining and TS and a length passage of 3 is taken into consideration for the information retrieval component. This approach reaches a precision around 30% when using ROUGE for evaluating the resulting summaries. Furthermore, the longer the summary, the better. In this sense, summaries of 50% compression rate obtain higher results. Although it is a moderate performance, it outperforms the average results obtained by TAC 2008 participants[1] (23%), as well as it increases the results by 66%, when compared to a baseline without taking into consideration an information retrieval component.

Our current aim focuses on improving the performance of the suggested approach, once the reasons of the possible causes that may be negatively affecting such performance are analysed. For instance, one issue to take into account is how to filter fragments of noisy information, because when working with blogs, information which is not directly related to the main content of it is also included (e.g. advertisements).

- **Extending the analysis of the suitability of COMPENDIUM into the rating inference task.** In our analysis concerning the use of summaries for the rating inference task (please see Chapter 6, Section 6.4), it seemed that sentiment-based summaries were more suitable to help classification systems to predict the correct rating associated to a review; however, the results obtained were not very conclusive. Therefore, we are still working in this issue, in order to come up with more robust results that can support our hypothesis. In light of this, we are working with an alternative corpus about movie reviews. This data set consists of approximately 5,000 reviews written by four different people. The average length for each review is 750 words, the longest one containing almost 3,000 words whereas the shortest one only 160 words. Each review is also associated with two types of ratings: the first one ranges from 0 to 2, whereas the other from 0 to 4. In this manner, we can extend the experiments

[1] For the experiments and evaluation we used the same guidelines as in the *Opinion Summarization Task* at TAC 2008.

into a wider data set and analyse other types of granularity for the classification task (i.e., 3 star-class).

The results obtained at the moment are still very preliminary. In general terms, when using the full review for training we have noticed that depending on who wrote the review, results vary, obtaining in some cases better results when using only the root of the word as a feature compared to the combination of the root of the word, its part of speech, and sentiment-based features (for instance, 59% vs. 29%), whereas in other cases, the latter combination performs better than the former (40% vs. 24%). Regarding the summaries generated with COMPENDIUM, although in most of the cases the accuracy do not surpass the full review, the distance between the predicted rating and the correct one, computed by means of the mean square error, is lower with respect to that of the full review.

On the other hand, for the long term, we will continue to carry out research into semantic TS techniques and we will focus in the evaluation of summaries. Therefore, we propose two more lines of action:

- **To analyse the influence of a domain-specific ontology to the process of TS.** This aspect is related to the research line of enriching COMPENDIUM with semantic knowledge. Ontologies are able to represent the knowledge as a set of concepts within a domain, and the relationships between those concepts. In particular, OntoFIS (Romá, 2009) is a pharmacotherapeutic ontology, and we will analyse whether taking it into account in the TS process leads to better domain-specific summaries.
- **To develop a qualitative TS evaluation framework.** A challenging topic in which we are also interested concerns the evaluation of summaries. A preliminary research about this topic was already conducted in (Lloret & Palomar, 2010), where a qualitative evaluation framework was outlined, proposing some criteria that would help to assess the quality of a summary (e.g. coherence, topic identification) instead of its content by comparing it to gold standard. The idea behind this kind of evaluation would be to set up an independent summarisation evaluation environment capable of testing a summary's quality, and deciding the degree of quality a summary would have. However, due to the complexity of this task, we think that machine learning algorithms would be the best way to approach

it. In this case, different stages would be distinguished: i) analyse and select an individual qualitative criteria; ii) learn the features in documents or summaries that indicate that such criterion is accomplished; iii) train and test the extracted features in a corpus in order to obtain the performance of the proposed criteria; iv) correlate the results obtained with already human evaluation to assess to what extent the suggested criterion could be determined automatically.

7.3 Relevant Publications

Although some publications have already been cited throughout this thesis, the following list groups all of them by the topic they are related to.

- Publications concerning the state of the art:
  - Elena Lloret and Manuel Palomar. Text Summarisation in Progress: A Literature Review. Journal of Artificial Intelligence Review. *In press.*
  - Elena Lloret and Manuel Palomar. 2011. Current Trends in the Evaluation of Text Summarization. Emerging Trends in Natural Language Processing: Concepts and New Research. Book chapter. *Submitted.*
- Publications regarding COMPENDIUM summarisation tool and its intrinsic evaluation:
  - Elena Lloret and Manuel Palomar: A Text Summarization System for Generating Abstracts of Research Papers. 16th International Conference on Applications of Natural Language to Information Systems, 2011. *In press.*
  - Ahmet Aker, Laura Plaza, Elena Lloret and Robert Gaizauskas: Towards automatic image description generation using multi document summarization techniques. Multi-source Multilingual Information Extraction and Summarization (MMIES). Book chapter. *In press.*

- Elena Lloret and Manuel Palomar: Tackling Redundancy in Text Summarization through Different Levels of Language Analysis, Information Processing and Management. *Submitted*
- Elena Lloret and Manuel Palomar: Resúmenes de textos: nuevos retos en la Web 2.0. Subjetividad y Procesos Cognitivos, 14 (2), ISSN 1852-7310, 2010.
- Laura Plaza, Elena Lloret, and Ahmet Aker: Improving Automatic Image Captioning Using Text Summarization Techniques. 13th International Conference on Text, Speech and Dialogue (TSD), 2010, Brno, Czech Republic.
- Elena Lloret and Manuel Palomar: Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. International Journal of Informatica, 34 (2), ISSN 0350-5596, 2010.
- Alexandra Balahur, Elena Lloret, Ester Boldrini, Andrés Montoyo, Manuel Palomar, and Patricio Martínez: Summarizing Threads in Blogs using Opinion Polarity. Proceedings of the Events in Emerging Text Types Workshop of the RANLP, 2009, Borovets, Bulgaria.
- Elena Lloret, and Manuel Palomar: A Gradual Combination of Features for Building Automatic Summarisation Systems. 12th International Conference on Text, Speech and Dialogue (TSD), 2009, Pilsen, Czech Republic.
- Elena Lloret, Óscar Ferrández, Rafael Muñoz, and Manuel Palomar: Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 41, pp. 183-190, ISSN 1135-5948, 2008, Leganés, Spain.
- Elena Lloret, Óscar Ferrández, Rafael Muñoz, and Manuel Palomar: A Text Summarization Approach Under the Influence of Textual Entailment. In Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) in conjunction with the 10th International Conference on Enterprise Information Systems (ICEIS 2008), pp. 22-31, 2008, Barcelona, Spain.

- Publications related to the integration of COMPENDIUM into HLT applications:
  - Elena Lloret, Héctor Llorens, Paloma Moreda, Estela Saquete and Manuel Palomar: Text Summarization Contribution to Semantic Question Answering: New Approaches for Finding Answers on the Web. International Journal of Intelligent Systems. *Submitted.*
  - Horacio Saggion, Elena Lloret, and Manuel Palomar: Using Text Summaries for Predicting Rating Scales. 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), 2010, Lisbon, Portugal.
  - Elena Lloret, Horacio Saggion, and Manuel Palomar: Experiments on Summary-based Opinion Classification. Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 2010, Los Angeles, California, USA.
  - Elena Lloret, Alexandra Balahur, Manuel Palomar, and Andrés Montoyo: Towards Building a Competitive Opinion Summarization System: Challenges and Keys. Proceedings of the North American Chapter of the ACL (NAACL), 2009, Boulder, Colorado, USA.
  - Alexandra Balahur, Elena Lloret, Óscar Ferrández, Andrés Montoyo, Manuel Palomar, and Rafael Muñoz: The DLSIUAES Team's Participation in the TAC 2008 Tracks. Proceedings of the Text Analysis Conference (TAC), 2008, Gaithersburg, USA.

A. Generación de resúmenes de textos basados en Tecnologías del Lenguaje Humano y su aplicación

Este anexo contiene un resumen extendido en castellano de la investigación llevada a cabo en esta tesis doctoral. En él se presentan los objetivos perseguidos, los métodos y sistemas más relevantes tanto para la generación de resúmenes como para su evaluación, nuestra contribución a la tarea a través de la herramienta de generación de resúmenes propuesta (COMPENDIUM), así como su evaluación en distintos dominios y para distintos tipos de resúmenes, y su integración a otras aplicaciones de Tecnologías del Lenguaje Humano (TLH). Finalmente se destacan las conclusiones más importantes y los trabajos que se están desarrollando en la actualidad, y las líneas de investigación futuras.

A.1 Introducción

Las Tecnologías del Lenguaje Humano (TLH) se encargan de procesar el lenguaje humano de forma automática. Este área de investigación es una subdisciplina de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos para facilitar la interrelación hombre-máquina, permitiendo una comunicación mucho más fluída y menos rígida que los lenguajes formales (Moreno Boronat *et al.*, 1999).

A pesar de su sencilla definición, su puesta en práctica resulta sumamente compleja. Su dificultad radica por la naturaleza misma del lenguaje natural y por el contexto socio-cultural en el que se enmarca. Por un lado, se debe tener constancia de las estructuras propias de una lengua concreta y de los fenónemos y propiedades que en dicho lenguaje se producen. La ambigüedad, propiedad inherente en todas las lenguas naturales, o mecanismos de economía lingüística como la elipsis, son dos ejemplos de ello, ampliamente tratados debido a su complejidad de procesamiento automático. Por otro lado, se debe disponer también de un conocimiento general acerca del mundo para comprender las ideas que se pretenden transmitir a través del lenguaje. Por esta razón, el campo de investigación de las TLH puede abarcar el tratamiento del lenguaje desde documentos completos, hasta las unidades que forman las palabras, por ejemplo los morfemas. Esto da lugar a un amplio abanico de subtareas, que comprende desde aplicaciones más generales, como la *recuperación de información*, *búsqueda de respuestas*, *extracción de información*, *generación de resúmenes*, *clasificación de textos*, etc., hasta aplicaciones intermedias, tales como *analizadores morfológicos*, *analizadores sintácticos*, *desambiguadores del sentido de las palabras*, *reconocedores de entidades*, etc. que constituyen los pilares básicos permitiendo el desarrollo de aplicaciones generales.

A.2 Generación de resúmenes: motivación

La generación de resúmenes no es una tarea nueva, ya que los primeros intentos de producir resúmenes automáticos se llevaron a cabo a finales de los años 50. Sin embargo, ha experimentado una gran evolución en la última década, sobre todo desde el rápido crecimiento de Internet. La cantidad de información disponible en formato electrónico crece de manera exponencial, dando lugar a millones de documentos cuya magnitud dificulta en gran medida su manejo. Debido a esto, la generación de resúmenes es de gran utilidad para procesar dicha información y presentarla de forma resumida y sencilla, de modo que ofrezca al usuario la posibilidad de gestionar la información de una forma más eficiente.

Según la Real Academia Española (RAE), "resumir" es *"reducir a términos breves y precisos, o considerar tan solo y repetir abreviadamente lo esencial de un asunto o materia"* (DRAE, 22ª edición). De esta definición se deduce que un resumen, si es elaborado correctamente, puede servir como sustituto del documento completo y ahorrar así, el trabajo de leerlo en su totalidad. La realización de un resumen requiere la lectura del documento en cuestión, saber extraer los conceptos e información más relevante y finalmente, reescribir toda esa información de manera que se obtenga un texto de menor tamaño que el original. Esto no es un proceso inmediato, sino que requiere tiempo y esfuerzo por parte de las personas que efectúen el resumen. En cambio, la obtención de dichos resúmenes de forma automática implicaría que apenas bastarían pocos segundos para resumir grandes cantidades de documentos. Ante los millones de documentos existentes en la web, supondría una gran ventaja disponer de este tipo herramientas automáticas. Lamentablemente, debido a todos los retos que presenta la tarea de generación automática de resúmenes, el conseguir resúmenes perfectos (igual que los haría un humano) es todavía una utopía, ya que la "inteligencia" que pueda tener un ordenador no es comparable a la humana. Sin embargo, gracias a los esfuerzos realizados por la comunidad científica, esta tarea se va perfeccionando consiguiendo cada vez, resúmenes mejores y de diversos tipos, según el cometido perseguido. Así, puesto que la tarea de generación de resúmenes conlleva una gran aplicación práctica y de gran utilidad, es necesario investigar en técnicas que permitan identificar la información más relevante de uno o varios documentos, y presentarla en forma de resumen.

A.3 Objetivos

El principal objetivo de esta tesis doctoral es el análisis, desarrollo e investigación de nuevas técnicas y enfoques para la generación automática de resúmenes de textos basados en la generación de extractos y abstractos. De este objetivo se extraen las siguientes líneas de investigación:

- Estudiar el estado de la cuestión de los sistemas de generación automática de resúmenes, abarcando también la evaluación de los mismos, a través del análisis de los métodos existentes en la actualidad.
- Proponer y analizar la influencia que determinadas técnicas ejercen sobre la generación de resúmenes, examinando las ventajas e inconvenientes que cada una de ellas aporta.
- Proponer una nueva herramienta para la generación automática de resúmenes, COMPENDIUM, que aplique y combine las distintas técnicas previamente estudiadas, y pueda generar diversos tipos de resúmenes en diferentes escenarios.
- Evaluar COMPENDIUM en distintos escenarios y para distintos tipos de resúmenes.

- Integrar COMPENDIUM en aplicaciones específicas de TLH (minería de opiniones, búsqueda de respuestas y clasificación de textos), que permitan evaluar, además, de manera extrínseca los resúmenes generados..

A.4 Organización del trabajo

Este capítulo está estructurado del siguiente modo: en la sección A.5 se presenta brevemente el estado de la cuestión tanto para los métodos y enfoques para abordar la tarea de generación de resúmenes como para su evaluación. Las siguientes secciones (A.6, A.7 y A.8) hacen referencia a nuestra propuesta para la generación de resúmenes: COMPENDIUM. En la sección A.6 se presentan las etapas que forman el proceso de generación de resúmenes junto con las técnicas y métodos involucrados en cada una de ellas. En la sección A.7 se exponen los principales experimentos y se describen las principales líneas de evaluación de la herramienta desarrollada. Además, la sección A.8 muestra la integración de COMPENDIUM en tareas de TLH, con el objetivo final de demostrar que se pueden obtener mejoras en los resultados de tareas tales como minería de opiniones, búsqueda de respuestas y clasificación de textos si se utilizan resúmenes automáticos. Finalmente, la sección A.9 contiene las conclusiones más importantes de esta tesis, así como los trabajos que se están llevando a cabo en la actualidad y los que se pretenden realizar en un futuro como continuación de esta tesis.

A.5 Estado de la cuestión

Según la definición dada en (Spärck Jones, 2007), el objetivo de la tarea de generación de resúmenes es obtener una versión reducida del documento o documentos fuente, reduciendo su contenido de tal forma que se seleccionen y queden presentes en el resumen los conceptos más importantes de dichos documentos. Por lo tanto, de la definición de esta tarea se deduce que un resumen debe contener la información más significativa de uno o varios documentos, teniendo un tamaño considerablemente inferior al del documento(s) fuente.

Los sistemas de generación de resúmenes se pueden clasificar en base a múltiples factores. Por ejemplo, relacionado con los factores de entrada, podemos producir un resumen a partir de un solo documento (mono-documento) o a partir de un conjunto de ellos (multi-documento). Además, no siempre debe tratarse de documentos textuales, ya que podemos realizar resúmenes a partir de otros tipos de texto, como páginas web, blogs, imágenes, vídeos, reuniones, etc. En lo que respecta al resumen generado, éste se puede conseguir siguiendo dos posibles estrategias: extractiva o abstractiva. Si se sigue una estrategia extractiva, se seleccionarán y extraerán, literalmente, las frases más importantes sin realizar ninguna modificación sobre ellas. Sin embargo, si se opta por llevar a cabo una estrategia abstractiva para producir un resumen, será necesario realizar algún tipo de transformación en las frases seleccionadas o en los conceptos que se deseen que formen parte del resumen, de tal manera que la información que aparezca en el resumen estará expresada de una forma distinta a la que aparece en el documento original. Otro aspecto a tener en cuenta en un resumen es si se trata de un resumen genérico o bien si debe estar centrado u orientado a algún tema o necesidad específica. Otra distinción posible, referente a los tipos de resumen que los sistemas automáticos suelen abordar, es la distinción entre resúmenes indicativos o informativos. Dependiendo de la finalidad que se le quiera dar al resumen, optaremos por uno u otro tipo. Los resúmenes indicativos son aquellos que simplemente proporcionan unas pequeñas pinceladas acerca de un documento, de manera que podemos tener una idea de si vale la pena acudir o no al documento completo, para encontrar en él la información que buscamos. Por el contrario, si producimos un resumen informativo, lo que pretendemos es que éste contenga la máxima información relevante posible, de tal manera que el resumen por sí solo nos proporcione la información requerida, sirviendo como sustituto del documento original. Finalmente, otro factor interesante a tener en cuenta es el idioma de los documentos origen y del resumen, ya que esto también puede dar lugar a diferentes tipos de resumen. Diremos que un sistema de resúmenes es monolingüe si únicamente es capaz de producir resúmenes para un determinado idioma (por ejemplo, el español). En este caso, los documentos fuente y el resumen estarán en el mismo idioma. Si por el contrario, un sistema produce resúmenes para diferentes idiomas (inglés, español, italiano, etc.) nos encontramos ante un sistema multilingüe, pero tanto los documentos fuente como el resumen siguen estando en el mismo idioma. Ahora bien, si el sistema no sólo no es capaz de tratar varios idiomas, sino que también puede generar resúmenes en un idioma determinado (por ejemplo, español) aunque los documentos a partir de los que tiene que generar el resumen estén en un idioma distinto (por ejemplo, italiano), estaremos ante el caso de un sistema *cross-lingual*.

A.5.1 Generación de resúmenes

A continuación se describen los métodos más comunes para generar resúmenes de textos.

- **Metódos estadísticos.**

  Las investigaciones en resúmenes de textos comenzaron a finales de los años cincuenta. Los primeros investigadores en estudiar posibles técnicas para poder generar resúmenes de forma automática fueron Luhn (Luhn, 1958) y Edmundson (Edmundson, 1969), que propusieron técnicas como la frecuencia de las palabras, la posición de las frases en un documento o la identificación de *cue words*. Esta última técnica consiste en calcular la relevancia de una frase en función de si tiene locuciones como "en conclusión", "el objetivo de este artículo", etc. que pueden ser indicadoras de que dicha frase pueda contener información significativa. Además de estas técnicas, en la literatura se pueden encontrar diversos enfoques basados en conocimiento para generar resúmenes tanto mono- como multi-documento. Por ejemplo, en (Qiu *et al.*, 2007) se proponen diferentes estrategias para producir resúmenes multi-documento, dependiendo de la finalidad que se persiga (por ejemplo, centrado en personas, lugares, eventos, objetos u otros), combinando varias técnicas en cada una de ellas. La idea subyacente es encontrar la combinación de técnicas más adecuada para cada tipo de resumen, y esto se consigue analizando los documentos de entrada. Las características que se proponen incluyen la frecuencia de palabras, la longitud y posición de las frases, o la importancia de las entidades nombradas. Otro sistema clásico que también se basa en la combinación de características es MEAD (Radev *et al.*, 2001). Este sistema es capaz de producir resúmenes tanto mono- como multi-documento mediante un enfoque extractivo, basándose en las siguientes fuentes de conocimiento: posición de la frase, solapamiento de una frase con respecto a la primera, y medidas para calcular la similitud de una oración respecto de la oración *centroide*. Las frases que formarán parte del resumen final serán seleccionadas en base al resultado de combinar linealmente las características previamente expuestas.

- **Métodos basados en la detección de tópicos.**

  La estrategia que siguen otros sistemas es la de localizar el tema principal en el documento, y ver cómo va cambiando a lo largo de éste, o analizar cómo se va descomponiendo en subtemas, de tal manera que en función de dichos cambios, se seleccionan las frases más adecuadas para el resumen. Existe una gran variedad de métodos mediante los cuáles se puede determinar el tema principal de un documento. Además, el poder usar técnicas que segmentan un documento en función de los cambios temáticos que se producen en un documento también es muy útil para la tarea de resúmenes. En la literatura, encontramos ejemplos de sistemas que utilizan estas técnicas para generar resúmenes, como es el caso de (Neto *et al.*, 2000), (Angheluta *et al.*, 2002), (Boguraev & Neff, 2000), (Harabagiu & Lacatusu, 2005). En este último enfoque, la estructura de temas se caracteriza en términos de lo que se denominan *topic themes*, que son representaciones de eventos o estados que se repiten a lo largo de un documento. En esta aproximación se analizan cinco maneras diferentes para representar el tema de un documento: (1) *topic signatures.* Las *topic signatures* (Lin & Hovy, 2000) se basan en representar un tema en función de un conjunto de términos relacionados con ese tema. (2) Representar el tema mediante la identificación de relaciones entre las *topic signatures.* (3) Identificar el tema de un documento segmentando el texto con el algoritmo *TextTiling* (Hearst, 1997) para ver en qué puntos se produce una ruptura temática. (4) Representar el tema del documento modelando el contenido del mismo, utilizando modelos ocultos de Markov. (5) Finalmente, el último mecanismo está basado en el uso de plantillas (*templates*), como si de una tarea de extracción de información se tratase.

  Por otra parte, en (Teng *et al.*, 2008) se propone un sistema para producir resúmenes mono-documento, identificando en una primera etapa, los posibles temas locales del documento. Esto se realiza agrupando las oraciones en función de su similitud, y una vez hecho esto, mediante la técnica de la frecuencia de las palabras, se extraen las frases más relevantes de cada uno de los grupos que se han creado, para representar los diferentes temas del documento.

- **Métodos basados en grafos.**

  Por otra parte, el uso de los algoritmos basados en grafos aplicados a la tarea de generación automática de resúmenes ya sea para producir, bien resúmenes genéricos o bien resúmenes centrados en un tema concreto, también ha sido muy utilizado recientemente. Esto se debe a que son muy efectivos y obtienen buenos resultados. Básicamente, la idea es representar un documento en forma de grafo, de manera que los nodos del mismo representen elementos del texto (normalmente, palabras o frases), mientras que las aristas son las relaciones entres dichos elementos (por ejemplo, relaciones de similitud entre frases, o de sinonimia, hiperonimia, etc. entre palabras). Una vez que tenemos el documento representado en forma de grafo, la idea es que dicha representación nos ayudará a determinar los elementos más importantes del documento, por ejemplo atendiendo al grado de salida de cada nodo del grafo. En (Mihalcea, 2004) se realiza un análisis de varios algoritmos basados en grafos para generar resúmenes. Además, en (Wan *et al.*, 2007) se propone un enfoque basado en grafos de afinidad, para producir tanto resúmenes genéricos como orientados a una necesidad específica. La idea aquí consiste en extraer frases que contengan una gran riqueza informativa, a la vez que incorporen información nueva. Esto se consigue a través de la similitud entre cada par de frases, teniendo en cuenta también información relacionada con el tema del documento y, diferenciando también entre enlaces dentro de un mismo documento y entre documentos diferentes. Finalmente, se elimina la información redundante.

- **Métodos basados en técnicas de aprendizaje automático.**

  Los sistemas que a continuación se presentan, realizan el proceso de generación de resúmenes basándose en algoritmos de aprendizaje automático. Las primeras técnicas de aprendizaje automático que se usaron en esta tarea incluyeron clasificadores binarios (Kupiec *et al.*, 1995), Modelos Ocultos de Markov (Conroy & O'Leary, 2001), (Schlesinger *et al.*, 2002) y redes bayesianas (Aone *et al.*, 1998). Pero no son los únicos algoritmos que se pueden utilizar para entrenar un conjunto de datos y extraer características. En los últimos años, sistemas como *NetSum* (Svore *et al.*, 2007) apuestan por el uso de redes neuronales, basándose en el algoritmo de aprendizaje *RankNet* (Burges *et al.*, 2005). Además de utilizar características como son las palabras clave y la posición de la frase, se propone también, un nuevo conjunto de características basadas en la Wikipedia[1], que da preferencia a las frases que contengan términos que estén contemplados en dicho recurso.

En *FastSum*, sistema propuesto por (Schilder & Kondadadi, 2008), se seleccionan las frases usando los algoritmos *Support Vector Regression* (SVR), y *Least Angle Regression* para seleccionar y extraer características. SVR ya se utilizó previamente para generar resúmenes en (Li *et al.*, 2007), donde se usaban características a nivel de palabra y frase, así como también la posición de las frases y la presencia de entidades nombradas, para entrenar el sistema de forma automática y puntuar cada frases del documento. En (Wong *et al.*, 2008), se realiza la generación de resúmenes utilizando métodos de aprendizaje automático supervisados (utilizando *Support Vector Machines (SVM)*) y semi supervisados (combinando SVM con clasificadores *Naïve Bayes*). Las características utilizadas se agrupan en distintos tipos e incluyen, entre otras, la longitud de la frase, la frecuencia de las palabras, o la similitud entre las frases.

- **Métodos para generar resúmenes de tareas concretas.**

  En los últimos años, los resúmenes multi-documento y orientados a usuarios están cobrando especial importancia, puesto que Internet se está convirtiendo en una fuente de conocimiento muy utilizada por todos los usuarios, dónde acuden cuando necesitan encontrar algún tipo de información específica, generalmente utilizando algún motor de búsqueda existente. Por tanto, debido a la gran cantidad de información que podemos encontrar en la web, lo que algunos sistemas de resúmenes intentan es ayudar en cierta medida, a los usuarios, facilitándoles el trabajo de tener que navegar documento por documento hasta encontrar la información concreta que buscan, y proporcionando un breve resumen informativo, que permita al lector ahorrar tiempo en leer la gran cantidad de información recuperada, teniendo sólo que echar un vistazo a los resúmenes, en vez de a los documentos completos. Tal es el caso de los sistemas *SWEeT* (Steinberger *et al.*, 2008) y el descrito en (Yang & Liu, 2008) que tienen su punto de partida en los documentos devueltos por un motor de búsqueda, y a partir de dicho conjunto de documentos, realizan los resúmenes sobre cada uno de ellos (resumen mono-

[1] http://www.wikipedia.org

documento) o sobre clusters (resumen multi-documento), utilizando diversas técnicas, como *Latent Semantic Analysis* en el primer caso, y grafos semánticos en el segundo, para identificar así, la relevancia de cada frase. Otras aproximaciones de sistemas basadas en la web, como la versión mejorada de *NetSum* (Svore *et al.*, 2008), producen resúmenes muy cortos a partir de documentos de noticias periodísticas, con el objetivo de extraer la información más destacada del texto (*highlights*).

En lo que respecta a resúmenes orientados a usuario, en (Díaz & Gervás, 2007) se plantea un sistema de resúmenes para el dominio periodístico capaz de contener información relevante de acuerdo a un perfil de usuario determinado. Para ello, se necesita tener primero un modelo de usuario para un individuo con el fin de, posteriormente, poder calcular la similitud entre cada una de las frases del documento y dicho perfil de usuario. Podremos encontrar diferentes resúmenes personalizados para un mismo usuario, dependiendo de la parte del perfil del usuario que se tome como referencia. A veces, no queremos que los resúmenes estén centrados en los usuarios, sino en temas concretos. El sistema *QCS* (Dunlavy *et al.*, 2007) es un sistema complejo que integra tareas de recuperación de información, agrupación y resúmenes de textos. Centrándonos en la tarea de resúmenes, que es la que nos interesa, el sistema presentado realiza primero un resumen de cada documento recuperado, y a partir del conjunto de resúmenes generados para cada grupo de documentos relacionados, se realiza el resumen de todo el cluster.

Por otro lado, la generación de resúmenes que contengan solamente información novedosa o más actual respecto a un tema (conocidos en inglés como *update summaries*) es también un reto dentro de la tarea de resúmenes. La idea que se persigue es generar resúmenes, pero teniendo en cuenta que los lectores de los mismos tienen un conocimiento previo sobre dicho tema, y por tanto sólo desean conocer los acontecimientos más recientes y no una visión general de dicho tema. Por ejemplo, ante un tema de actualidad como puede ser el brote de gripe aviar, si conocemos información acerca del tema, posiblemente no nos interese conocer las causas, ni las vías de contagio puesto que ya lo sabremos, sino que si queremos estar actualizados, posiblemente sea más útil tener un sistema de resúmenes que nos proporcione diariamente el número de nuevos casos afectados en cada país. Un ejemplo de sistema que se basa en esta idea lo podemos encontrar en (Sweeney *et al.*, 2008), en el que las frases que contienen novedades se determinan calculando la similitud entre las frases que ya forman parte del resumen y las posibles frases candidatas, y escogiendo las que menos similitud tengan. En cambio, en (Witte *et al.*, 2007) y (Bellemare *et al.*, 2008), este tipo de resúmenes se generan a partir de grafos de clusters que se basan en el contexto del conjunto de documentos de los que queremos obtener el resumen. Es decir, se propone un método para determinar el ránking de las oraciones de los documentos en base a la cantidad de vocabulario común entre las frases pertenecientes a los clusters y el contexto, de tal forma, que finalmente se seleccionan las frases dependiendo de su posición en dicho ránking. En (Li *et al.*, 2008), se propone el concepto de "historia", que hace referencia a aquellos documentos que el lector ya conoce. Además, se introducen métricas para filtrar frases mediante medidas de similitud, tales como el coseno, que evitan incorporar al resumen aquellas frases que guardan cierta similitud con alguna frase contenida en el histórico de documentos.

Otra aplicación de la tarea de resúmenes es generar resúmenes biográficos a partir de un conjunto de documentos referentes a una persona (Zhou *et al.*, 2004). La idea subyacente es producir pequeños resúmenes que contengan aspectos relevantes sobre una determinada persona, por ejemplo respondiendo a preguntas como *"¿Quién es Barack Obama?"*. En el sistema mencionado, para lograr esta tarea se utilizaron diversas técnicas de aprendizaje automático (redes bayesianas, SVM y árboles de decisión) para clasificar las oraciones y determinar cuáles eran las más apropiadas para formar parte del resumen, eliminando la información redundante en una etapa posterior. Otro sistema muy parecido para generar resúmenes biográficos lo encontramos en (Biadsy *et al.*, 2008), que está basado, al igual que el anterior, en algoritmos de aprendizaje automático. La diferencia entre ambos radica en que en este caso se utiliza un clasificador binario para decidir si una frase es de tipo biográfica o no, y además se utiliza la Wikipedia como corpus.

Por otro lado, aunque el dominio que más se ha utilizado para generar resúmenes ha sido el dominio periodístico (Nenkova *et al.*, 2005), (Nenkova, 2005), el científico (Jaoua & Hamadou, 2003), (Teufel & Moens, 2002) o incluso podemos encontrar algunos trabajos relacionados con el dominio legal (Saravanan *et al.*, 2006), (Cesarano *et al.*, 2007), algunos sistemas realizan resúmenes sobre documentos que pertenecen al dominio literario, ya se trate de relatos breves (Kazantseva, 2006) o de libros. En (Mihalcea & Ceylan, 2007) se exponen los problemas derivados al intentar producir resúmenes de libros. Se proponen, además, varias técnicas que pueden ser útiles para resumir este tipo de documentos, tomando como referencia el sistema MEAD (Radev *et al.*, 2001), con algunos cambios para adaptarlo a la generación de resúmenes de documentos con mayor extensión.

Recientemente, una de las aplicaciones de la tarea de generación de resúmenes es combinar esta tarea con la de minería de opiniones (también conocida como *Opinion Mining*) para producir resúmenes subjetivos orientados a opiniones, y por tanto expresen sentimientos o argumentos a favor o en contra de un determinado producto, lugar, persona, etc. En lo que respecta a este tipo de resúmenes, primero se deben detectar qué frases expresan opiniones y determinar el carácter de la opinión, es decir, si dicha opinión está a favor o en contra de algo. Una vez que las opiniones han sido clasificadas y agrupadas, habrá que construir el resumen. Algunos de los sistemas que participaron en la tarea piloto del TAC 2008[2] (*Opinion Summarization Pilot task*) seguían estos pasos (Conroy & Schlesinger, 2008), (He *et al.*, 2008a), (Balahur *et al.*, 2008) o (Bossard *et al.*, 2008). Por otro lado, fuera del ámbito de la competición también podemos encontrar enfoques interesantes para abordar esta tarea. Por ejemplo, en (Beineke *et al.*, 2004) se propone un sistema con una filosofía similar, en el que una vez identificados los fragmentos de texto que expresan opiniones, se utilizan técnicas de aprendizaje automático para seleccionar las frases que pertenecerán al resumen. De manera similar, en (Zhuang *et al.*, 2006) se identifican las palabras que expresan opinión y el tipo de las mismas (bien expresando una opinión positiva o negativa), para componer posteriormente el resumen, pero orientado solamente a reseñas de películas de cine. Generalmente, las opiniones que se encuentran en la web vienen acompañadas muchas veces de índices numéricos que representan el grado de satisfacción de un usuario ante un determinado producto o servicio. En (Titov & McDonald, 2008) se propone un modelo (*Multi-*

[2] Text Analysis Conference: www.nist.gov/tac/

Aspect Sentiment Model) para identificar el carácter de las opiniones expresadas en dicho formato y poder extraer los fragmentos de textos correspondientes para generar automáticamente un resumen.

A.5.2 Evaluación de resúmenes

Los métodos para evaluar cualquier tarea basada en TLH se pueden clasificar en dos grandes grupos: métodos intrínsecos y extrínsecos (Spärck Jones & Galliers, 1996). Concretamente, aplicados a la tarea de generación de resúmenes, los métodos intrínsecos evalúan el resumen en sí, atendiendo al contenido del mismo. En cambio, los extrínsecos se centran en evaluar cómo de buenos son los resúmenes generados, para poder cumplir el cometido de otra tarea externa, como por ejemplo tareas de recuperación de información. En lo que respecta a la evaluación intrínseca (que es en la que nos centraremos en esta sección), existen varios métodos que pueden ser útiles a la hora de evaluar un resumen automático. Según (Mani, 2001a), dentro de los métodos intrínsecos, se puede hacer una subclasificación entre evaluar la calidad del resumen, y realizar, así, una evaluación cualitativa, o bien evaluar el grado de información que contiene (*informativeness*), y por lo tanto realizar una evaluación cuantitativa. Además, se puede evaluar también el grado de similitud respecto al documento o los documentos origen, para determinar si el resumen cubre los mismos conceptos relevantes que el documento fuente. El problema de este método es cómo determinar la importancia de los conceptos en los documentos fuentes. Sin embargo, a pesar de existir diferentes metodologías para llevar a cabo una evaluación intrínseca, la más común es evaluar la información del resumen, comparando el contenido del mismo frente a un conjunto de resúmenes elaborados por humanos que se toman como referencia.

Para lograr esta tarea, se han propuesto varios métodos y se han desarrollado varias herramientas a lo largo de estos años. Sin embargo, debido a la subjetividad inherente que presenta la tarea de generación de resúmenes, no se puede establecer de forma objetiva un resumen de referencia que sirva como modelo (*gold standard*), puesto que podemos encontrar para un mismo documento, diferentes resúmenes igual de válidos. Distintas personas pueden variar en sus opiniones acerca de lo bueno que puede ser un resumen, y además un resumen depende mucho de la finalidad que se pretenda conseguir.

Una de las herramientas más conocidas y usadas para la evaluación automática de resúmenes es ROUGE (Lin, 2004). Esta herramienta[3] permite obtener los valores de precisión, cobertura y medida-F para un resumen generado automáticamente, siempre y cuando tengamos disponible al menos, un resumen de referencia (escrito por un humano). La idea subyacente es que los textos con un significado similar, deben contener palabras o frases comunes. ROUGE se basa en la comparación de n-gramas entre ambos resúmenes, y para ello se establecen diferentes tipos de n-gramas, como unigramas (ROUGE-1), bigramas (ROUGE-2), subsecuencia común más larga (ROUGE-L), etc., siendo ROUGE-1 y ROUGE-2 las más utilizadas.

Otra herramienta similar, pero que permite una mayor flexibilidad es *Basic Elements* (Hovy *et al.*, 2006), cuya idea se basa en segmentar una frase en unidades de contenido mínimo, definiendo tripletas formadas por un elemento principal (*head*), un modificador (*modifier*) y la relación (*relation*) entre ambos elementos. Su objetivo es permitir mayor flexibilidad a la hora de identificar expresiones comunes entre un resumen automático y su correspondiente conjunto de resúmenes de referencia.

AutoSummENG (Giannakopoulos *et al.*, 2008b) es otra herramienta desarrollada para evaluar resúmenes de forma automática. El método usado para esta herramienta difiere de los anteriores principalmente en tres aspectos: (1) el tipo de información estadística extraído; (2) la representación escogida para dicha información; y (3) la manera de calcular la similitud entre resúmenes. La comparación entre resúmenes manuales y automáticos se realiza a partir de grafos de n-gramas de caracteres. Una vez construidos estos grafos, se lleva a cabo la comparación entre ambas representaciones para poder establecer el grado de similitud entre ambos.

Por otra parte, también podemos encontrar diversas metodologías de evaluación para idiomas distintos del inglés, como por ejemplo para el chino. Tal es el caso de *HowNet*[4], que es una base de conocimiento para el inglés y el chino, y se diferencia de otros recursos similares, como WordNet[5] en la manera que se establecen las relaciones entre

[3] *Recall-Oriented Understudy for Gisting Evaluation*: http://haydn.isi.edu/ROUGE/

[4] http://www.keenage.com

[5] http://wordnet.princeton.edu/

las palabras que forman el recurso. Además, *HowNet* proporciona más información para cada uno de los conceptos, asociando a cada uno de ellos una definición y relaciones con otros conceptos de forma no ambigua. Este recurso ha sido muy usado para el chino en diferentes tareas basadas en TLH. Por ejemplo, en (Wang *et al.*, 2008) se propone un método para evaluar resúmenes de forma automática haciendo uso de este recurso, donde además de calcular la similitud entre pares de resúmenes en base a la co-ocurrencia de n-gramas, se tienen en cuenta también las palabras que guardan alguna relación de sinonimia entre ellas.

A.6 La herramienta de resúmenes COMPENDIUM

COMPENDIUM es nuestra contribución para la tarea de generación automática de resúmenes de textos. Se trata de una herramienta que es capaz de producir distintos tipos de resúmenes. COMPENDIUM puede recibir uno o más documentos como entrada (mono- o multi-documento) y puede producir como salida resúmenes genéricos, orientados a un tema en concreto o resúmenes subjetivos. El objetivo de los resúmenes generados es que proporcionen al usuario las ideas más importantes de los correspondientes documentos fuente, y por lo tanto los resúmenes generados se pretende que sean informativos. En cuanto a la forma de construir dichos resúmenes, podemos generar resúmenes extractivos o bien combinar técnicas extractivas con abstractivas para generar resúmenes orientados a abstractos. Finalmente, es importante destacar el hecho de que, por el momento, COMPENDIUM ha sido desarrollado y evaluado sólo para el inglés, lo cual no quita que no se pueda extender y probar para otros idiomas, como el castellano.

El proceso de generación de resúmenes se basa en una perspectiva cognitiva (Van Dijk, 1980), (Van Dijk & Kintsch, 1983), pero también aporta una componente computacional (Hovy, 2005) que permite su automatización. La arquitectura que se propone para COMPENDIUM consta de dos tipos de etapas. Por un lado, el núcleo de la herramienta está formado por cinco etapas cuyo resultado serán resúmenes genéricos extractivos. Por otro lado, se proponen una serie de etapas adicionales, con el objetivo de incrementar las funcionalidades de COMPENDIUM, y que se encargarán de producir tipos de resúmenes concretos, como por ejemplo resúmenes subjetivos.

La figura A.1 muestra la arquitectura general de COMPENDIUM, dónde la parte central representa el núcleo de la herramienta, mientras que los procesos enmarcados en rectángulos con bordes discontinuos se refieren a las etapas adicionales.

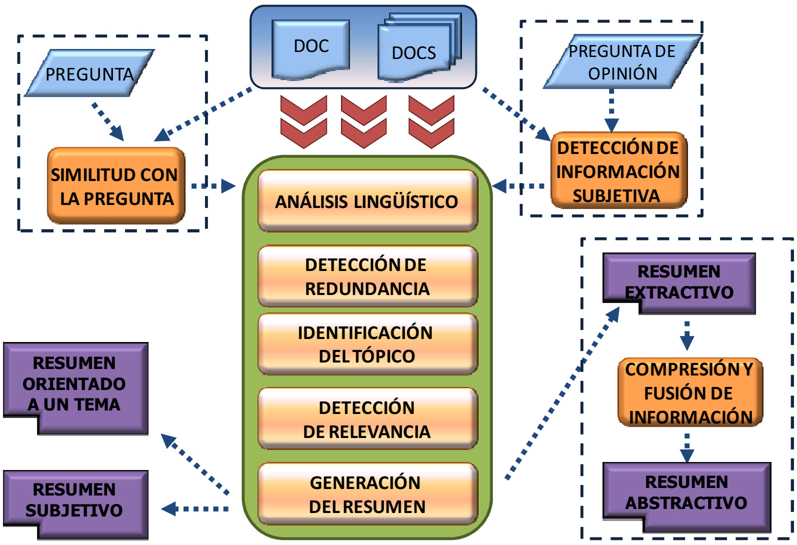


Fig. A.1. Arquitectura general de COMPENDIUM.

A continuación se hace una distinción entre las etapas que integran el núcleo de la herramienta y las que se consideran adicionales, y se explica con detenimiento cada una de ellas:

A.6.1 Núcleo de COMPENDIUM

Las etapas que forman el núcleo central de COMPENDIUM son: i) *análisis lingüístico*; ii) *detección de redundancia*; iii) *identificación del tópico*; iv) *detección de relevancia*; y v) *generación del resumen*.

Análisis lingüístico. En esta etapa se realiza un preprocesado básico del texto o de los textos de entrada utilizando recursos existentes de TLH, que consiste en:

- **Segmentación de oraciones.** El texto de entrada se segmenta en oraciones[6], ya que esta es la unidad que vamos a considerar para generar el resumen.
- **Segmentar en tokens.** Además de segmentar el texto de entrada en frases, también debemos segmentarlo en tokens[7] para poder luego calcular la frecuencia de aparición de cada uno, o bien distinguir si se trata de una *stop word* o no.
- **Realizar *stemming*.** Este proceso nos permitirá obtener la raíz de una palabra[8].
- **Identificación de *stop words*.** Este proceso nos permitirá descartar palabras como artículos, preposiciones, conjunciones, etc, que no son necesarias para determinar la relevancia de una oración en el texto. Para poder identificar correctamente las *stop words* de un documento, nos basamos en una lista predefinida de stop words para el inglés[9].

Detección de redundancia. El objetivo de esta fase es detectar y eliminar la información redundante de un documento, para evitar así que el resumen contenga información repetida. Para lograr este objetivo, nos basamos en un módulo de reconocimiento de la implicación textual (TE) (Ferrández, 2009), que nos indicará, dadas dos oraciones si una se puede deducir de la otra. Este sistema se basa en el cómputo de un conjunto de medidas léxicas (como por ejemplo, distancia de *Leveshtein*, *SmithWaterman*, similitud del coseno) y semánticas basadas en *WordNet 3.0*[10], aplicando un clasificador SVM con el objetivo de tomar la decisión final.

La idea principal que sostiene el uso de la implicación textual en tareas automáticas de resúmenes siguiendo el enfoque propuesto, reside en conseguir un conjunto preliminar de oraciones formado por aquéllas que no tienen relación de implicación con ninguna otra frase

[6] Para lograr esto, se utiliza la siguiente herramienta: http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz

[7] Para esto, utilizamos la herramienta *Word Splitter*: http://cogcomp.cs.illinois.edu/page/tools_view/8

[8] Utilizamos la herramienta The *Porter Stemmer*: http://tartarus.org/ martin/PorterStemmer/

[9] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[10] http://wordnet.princeton.edu/

del documento. La identificación de dichas relaciones de implicación ayudan a que el resumen final contenga la menor redundancia posible.

Identificación del tópico. A pesar de que la técnica basada en el cálculo de las frecuencias de las palabras fue de las primeras en utilizarse para generar resúmenes de forma automática (Luhn, 1958), todavía se sigue utilizando, ya que es una técnica sencilla de aplicar que obtiene muy buenos resultados.

Por lo tanto, usamos la frecuencia de las palabras (TF) para obtener el tema o temas principales de un documento, de tal manera que las palabras con mayor frecuencia en un documento (sin tener en cuenta las *stop words*, que han sido previamente eliminadas) indican cuáles son los conceptos más importantes del mismo.

Detección de relevancia. Esta etapa se encarga de asignar un peso a cada oración del documento, en función de su relevancia. Este peso se calcula combinando la frecuencia de cada término obtenido en la etapa anterior con el Principio de la Cantidad de Codificación (CQP) (Givón, 1990).

Este principio es de origen lingüístico y establece que mientras menos predecible (o más importante) es la información, más prominente, más evidente y larga será el medio de codificación que la represente. Esto significa que un elemento encargado de presentar una determinada información en un texto, recibirá una codificación que será más o menos larga, en función del grado de relevancia que tenga dicha información en el texto. En otras palabras, si la información es más importante, entonces recibirá una cantidad de codificación mayor, con lo que se codificará con mayor peso léxico. En cambio, si se trata de información menos importante, ésta se codificará con menor peso léxico. En (Ji, 2007), se ha demostrado que este principio se cumple en textos escritos y además, está directamente relacionado con otro principio de carácter cognoscitivo, que es el Principio de la cantidad, la atención y la memoria (Givón, 1990), cuyas premisas son: (1) la codificación más prominente y distinta atraerá más la atención del receptor, y (2) la información que atrae más la atención se memoriza, almacena y recupera de forma más eficiente.

Por todas estas razones, incluimos este principio para detectar información relevante en nuestra propuesta. Como unidad de codificación, decidimos seleccionar los sintagmas nominales, puesto que son capaces de contener más o menos información según lo que se quiera transmitir, al poder incorporar modificadores (determinantes, adjetivos, nombres, o incluso cláusulas de relativo) que permiten aclarar y dar más información sobre un determinado sustantivo. Otro motivo para considerar los sintagmas nominales lo encontramos en el análisis realizado en (Mittal *et al.*, 1999), donde se demostró que, en términos medios, la longitud de los sintagmas nominales de las oraciones que forman parte de un resumen es más del doble que las oraciones que no son seleccionadas para formar parte del resumen final.

Nuestra hipótesis de trabajo al utilizar este principio es que las oraciones que contengan sintagmas nominales más largos y que se compongan de palabras más frecuentes, harán que dichas frases reciban mayor peso para ser seleccionadas y formar parte del resumen final.

Para identificar los sintagmas nominales de una frase, utilizamos la herramienta *BaseNP Chunker*[11]. Esta herramienta, como cualquier otra herramienta de PLN, tiene errores, pero su rendimiento es bastante bueno, alcanzando valores de precisión y cobertura del 93% para sintagmas nominales simples, y 88% para los sintagmas nominales más complejos (Ramshaw & Marcus, 1995). Una vez que ya tenemos los sintagmas nominales identificados en las frases de los documentos, las frases serán clasificadas en función de un ránking obtenido mediante la fórmula A.1.

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \tag{A.1}$$

donde:
r_{s_i} = representa la relevancia de la oración i,
$\#NPi$ = número de sintagmas nominales que la oración i contiene,
tf_w = frecuencia de la palabra w que pertenece al sintagma nominal.

Generación del resumen. El objetivo de esta fase es generar un resumen de una determinada longitud. Dicha longitud puede estar expresada en número de palabras, o bien en forma de tasa de compresión respecto al documento fuente.

[11] Esta herramienta se encuentra disponible en: ftp://ftp.cis.upenn.edu/pub/chunker/

Podemos distinguir entre cuatro tipos de resúmenes que se pueden generar en esta fase: 1) resúmenes genéricos; 2) resúmenes orientados a un tópico; 3) resúmenes subjetivos; y 4) resúmenes orientados a abstractos. A continuación se explica brevemente cómo se genera cada uno de estos resúmenes.

1. **Resúmenes genéricos (COMPENDIUM$_E$).** Una vez calculada la relevancia de cada frase de los documentos, las frases cuyo valor de relevancia sea mayor serán seleccionadas para pertenecer al resumen final hasta alcanzar la longitud de resumen deseada.
2. **Resúmenes orientados a un tópico (COMPENDIUM$_{QE}$).** Para generar este tipo de resúmenes, tenemos que tener en cuenta, además de la etapa de detección de redundancia, la etapa de similitud con la pregunta, que se explicará posteriormente. Una vez obtenidos estos dos valores, se combinarán mediante la fórmula A.2), donde β puede obtener valores que oscilen entre 0 y 1, dependiendo de si se quiere dar más importancia a la relevancia de la frase o a la similitud con la pregunta.

$$Sc_{s_i} = (1+\beta^2)\frac{r_{s_i} * qSim_{s_i}}{\beta^2 * r_{s_i} + qSim_{s_i}} \tag{A.2}$$

   donde:
   Sc_{s_i} = es el valor final para la frase i,
   r_{s_i} = es la relevancia de la frase i,
   $qSim_{s_i}$ = es la similitud entre la pregunta (o tópico) y la frase i.

   Las frases que obtengan mayor puntuación serán las que seleccionen para formar el resumen final.
3. **Resúmenes subjetivos (COMPENDIUM$_{SE}$).** La particularidad de este tipo de resúmenes es que sólo contienen información subjetiva que ha sido identificada en la etapa de *detección de información subjetiva*, que se comentará más adelante.
4. **Resúmenes orientados a abstractos (COMPENDIUM$_{E-A}$).** Para generar este tipo de resúmenes, combinamos información extractiva y abstractiva. Una vez generada la información abstractiva mediante la etapa *compresión y fusión de información*, la estrategia que seguiremos para seleccionar qué frases se incluirán en el

resumen final, será calcular la similitud entre la nuevas frases generadas y las frases del extracto inicial. Para ello, determinaremos un umbral de similitud y en los casos que se supere, las frases extractivas se sustituirán por las abstractivas equivalentes; en otro caso, nos quedaremos con la frase extractiva original.

A.6.2 Etapas adicionales

Además del núcleo del sistema, tenemos tres etapas adicionales, que permiten generar distintos tipos de resúmenes. Estas etapas se describen a continuación.

Similitud con la pregunta. Para determinar qué frases están relacionadas con una pregunta o tópico en concreto, calculamos la similitud (usando la medida del coseno[12]), de acuerdo a la fórmula A.3:

$$qSim_{s_i} = SimilitudCoseno(S_i, Query) \tag{A.3}$$

donde:
$qSim_{s_i}$ = es la similitud entre en la pregunta (*Query*) y la frase *i*.

Detección de información subjetiva. Esta etapa sirve para detectar y clasificar la información subjetiva de un documento, y tiene lugar antes de que se calcule la relevancia de una oración. Para ello, se necesita una herramienta que permita realizar dicho análisis subjetivo, y que permita clasificar una frase como positiva, negativa o neutra. En concreto, COMPENDIUM usa la herramienta de minería de opiniones propuesta en (Balahur-Dobrescu *et al.*, 2009). Esta herramienta asigna a cada frase, uno de estos tres valores: i) $valor > 0$, si la frase es positiva; ii) $valor < 0$, si la frase es negativa; y iii) $valor = 0$, si la frase es neutra. Una vez realizado este proceso, solamente las frases positivas o negativas se seleccionarán para las siguientes etapas.

Compresión y fusión de información. En esta etapa se utilizan grafos de palabras para obtener una versión comprimida de una frase más larga, o bien una nueva frase que contiene información de varias frases.

[12] Disponible en la herramienta *Text Similarity*: http://www.d.umn.edu/ tpederse/text-similarity.html

Para trabajar con grafos de palabras seguimos un enfoque similar al que se propone en (Filippova, 2010). La idea subyacente al uso de grafos de palabras es que un texto se puede representar como un conjunto de vértices y aristas, donde los vértices son las palabras del documento y las aristas son las relaciones de adyacencia entre palabras. Concretamente, nosotros utilizamos un grafo dirigido pesado, en el que el peso de las aristas representa la inversa de la frecuencia de aparición de dos palabras en el texto, teniendo en cuenta también la importancia de los nodos que una dichas palabras, mediante el algoritmo Pagerank (Brin & Page, 1998). Una vez representado el texto, obtendremos los caminos más cortos entre los distintos nodos utilizando el algoritmo de Dijkstra. Puesto que de estos caminos, se pueden obtener frases que no estén completas o que no sean correctas, filtraremos los caminos incorrectos, en base a tres reglas: i) la oración debe contener al menos tres palabras; ii) la oración debe contener un verbo; iii) y la oración no puede acabar en un artículo, preposición, conjunción o partícula interrogativa. Una vez que se han filtrado las frases incorrectas, podemos combinar esas nuevas frases con las del extracto original generado. En este punto, es importante destacar que la etapa de *compresión y fusión de información* se realiza una vez generada la versión extractiva del resumen genérico con COMPENDIUM. La razón es que, de esta manera, nos aseguramos que el resumen contenga la información más importante del documento, y además, intentamos comprimir y combinar esa información para que el resumen final no sea solamente un conjunto de frases concatenadas.

A.7 Evaluación de COMPENDIUM

Para evaluar COMPENDIUM utilizamos varios corpus de datos de diferentes tipos y dominios para generar los resúmenes correspondientes. Además empleamos la herramienta ROUGE[13] con distintas métricas para evaluar la información contenida en los resúmenes, para aquellos casos que dispongamos de resúmenes modelo. Para los casos en los que estos resúmenes no estén disponible se llevará a cabo una evaluación manual de la calidad de los resúmenes, o incluso, en algunos tipos de resúmenes, se realizará un estudio de la satisfacción del usuario.

[13] http://berouge.com/default.aspx

A continuación vamos a mostrar para cada corpus de datos evaluados, los resultados obtenidos, junto a una discusión de los mismos.

- **Noticias periodísticas**. Se trata de varios conjuntos de noticias de diferentes periódicos con sus correspondientes resúmenes modelo, que se usaron en las competiciones DUC[14] durante los años 2002, 2003 y 2004. Estos documentos nos permitirán evaluar COMPENDIUM$_E$ para la generación de resúmenes genéricos mono- y multi-documento.

  El cuadro A.1 muestra los resultados obtenidos para COMPENDIUM$_E$. Como se observa, los resúmenes mono-documento obtienen mejores resultados (45% para ROUGE-1) que los multi-documento (30% de media). Esta diferencia puede ser debida a que para la generación de resúmenes multi-documento no se emplea ninguna técnica específica, sino que se consideran todos los documentos como uno solo.

		Mono-documento	**Multi-documento**
DUC 2002	ROUGE-1	0,45611	0,30137
	ROUGE-2	0,20252	0,05327
	ROUGE-L	0,41382	0,26373
DUC 2003	ROUGE-1	-	0,28977
	ROUGE-2	-	0,05481
	ROUGE-L	-	0,25399
DUC 2004	ROUGE-1	-	0,31091
	ROUGE-2	-	0,06316
	ROUGE-L	-	0,27633

Cuadro A.1. Resultados de COMPENDIUM$_E$ para mono- y multi-documento para el dominio periodístico (medida $F_{\beta=1}$).

  Comparando estos resultados con los obtenidos por el mejor sistema que participó en las diferentes ediciones de las conferencias DUC, obtenemos que COMPENDIUM$_E$ obtiene mejores resultados que el mejor sistema para mono-documento (con un incremento del 8%). Sin embargo, no sucede lo mismo para el caso multi-documento, obteniendo resultados ligeramente inferiores. Como ya comentamos antes, esto puede suceder a que no estamos abordando la generación

[14] http://www-nlpir.nist.gov/projects/duc/data.html

de resúmenes multi-documento de la manera más apropiada, ya que este tipo de resúmenes pueden requerir un procesamiento más específico.

- **Descripciones de imágenes**. Este corpus (Aker & Gaizauskas, 2010b) contiene 308 imágenes que representan a diferentes lugares. Cada imagen tiene asociada 10 documentos recuperados de Internet que guardan relación con la imagen en cuestión. También disponemos de resúmenes manuales (hasta cuatro para cada uno de los lugares). Utilizando este corpus compararemos las versiones genéricas y orientadas a tópico de COMPENDIUM (COMPENDIUM$_E$ y COMPENDIUM$_{QE}$, respectivamente).

  Los resultados se muestran en el cuadro A.2. De estos resultados se puede concluir que los resúmenes orientados a un tópico son más apropiados en este caso que los resúmenes genéricos, ya que los resultados para COMPENDIUM$_{QE}$ mejoran en un 4% y además esta mejora es estadísticamente significativa (test de Wilcoxon). Sin embargo, vemos que los resultados no mejoran los obtenidos utilizando las 200 primeras palabras de la Wikipedia como resumen. Esto se debe a que estos resúmenes considerados como *baseline* son difíciles de superar, debido principalmente a que estos artículos han sido creados por humanos, y además em las primeras líneas la información que contienen está directamente relacionada con el tópico.

	ROUGE-2	**ROUGE-SU4**
Baseline Wikipedia	0,09632	0,14203
COMPENDIUM$_E$	0,08551	0,13371
COMPENDIUM$_{QE}$	0,08864	0,13892

Cuadro A.2. Comparación de COMPENDIUM$_E$ y COMPENDIUM$_{QE}$ en el dominio de descripciones de imágenes (valor para la cobertura).

- **Blogs**. Este corpus nos servirá para generar resúmenes subjetivos (COMPENDIUM$_{SE}$), puesto que consiste en 51 blogs extraídos directamente de Internet, sobre varios temas (economía, cocina, deportes, ciencia y tecnología o sociedad). De este corpus nos interesarán los comentarios asociados a cada blog, a partir de los cuales, generaremos el resumen. Al no disponer de resúmenes modelo, la evaluación que llevamos a cabo consistió en evaluar de manera manual los resúmenes generados en base a los siguientes criterios: *redundancia*, *corrección gramatical*, *foco* y *dificultad*. Para cada uno de estos criterios se estableció una escala de tres valores (aceptable, legible y no aceptable; o bien, alta, media o baja, para el caso de la dificultad). A continuación se muestran los resultados obtenidos (Cuadro A.3).

		Ratio de compresión		
Criterio	**%**	**10%**	**15%**	**20%**
Redundancia	No aceptable	26	0	4
	Legible	45	6	10
	Aceptable	29	94	86
Correc. gramatical	No aceptable	4	2	0
	Legible	22	27	55
	Aceptable	74	71	45
Foco	No aceptable	33	26	14
	Legible	43	29	47
	Aceptable	24	45	39
Dificultad	Alta	35	18	8
	Media	28	35	51
	Baja	37	51	41

Cuadro A.3. Resultados para COMPENDIUM$_{SE}$.

De los resultados es importante destacar que, en líneas generales, los resúmenes de tamaño 15% o 20% del documento original obtienen mejores resultados. Por otro lado, obtenemos que el número de resúmenes que han sido evaluados como aceptables oscila entre el 45% al 74%. En general los resultados son bastante prometedores, a pesar de la dificultad de la tarea. Esta dificultad radica en que la mayoría de los blogs y los comentarios asociados contienen mucha información ruidosa, que afecta directamente a los resúmenes y es muy difícil de filtrar automáticamente, debido a que no hay una estructura uniforme. Así, encontramos en el propio corpus errores ortográficos y gramaticales (*“I thikn”*) o frases que no aportan nada al resumen (*“welcome back!!!”*).

- **Artículos científicos del dominio médico**. Este corpus contiene 50 artículos científicos. Cada artículo incluye un resumen generado por el autor o los autores del mismo (abstracto), que utilizaremos como resumen modelo. Este corpus nos servirá para evaluar los resúmenes orientados a abstractos generados usando COMPENDIUM$_{E-A}$ y compararlos con los obtenidos con la versión extractiva (COMPENDIUM$_E$).

En este caso realizamos tres tipos de evaluación diferentes. Primero evaluamos los resúmenes comparándolos con el abstracto original (cuadro A.4); después, evaluamos si los resúmenes contienen los tópicos adecuados, teniendo en cuenta las palabras claves asociadas a cada artículo (cuadro A.5); y finalmente realizamos un estudio de la satisfacción del usuario con respecto a los resúmenes generados con COMPENDIUM (cuadro A.6). Para este estudio, se propusieron tres preguntas en concreto:

- **P1:** El resumen proporcionado refleja los contenidos mas importantes del documento.
- **P2:** El resumen proporcionado permite al lector saber de que trata el texto.
- **P3:** Despues de haber leído el resumen hecho por humanos, el resumen alternativo proporcionado (el que estás evaluando) es también válido.

	Cobertura	**Precisión**	**F$_{\beta=1}$**
COMPENDIUM$_E$	**0,44022**	0,40525	**0,42201**
COMPENDIUM$_{E-A}$	0,38658	**0,41809***	0,39533

Cuadro A.4. Resultados obtenidos por COMPENDIUM$_{E-A}$ y su comparación con COMPENDIUM$_E$.

De los resultados obtenidos con ROUGE, podemos observar que los resúmenes orientados a abstractos obtienen mejor precisión (los resultados con asteriscos indican que la mejora es estadísticamente significativa) que los extractos puros. Esto nos indica que el método para comprimir y fusionar frases es adecuado. Sin embargo, el valor de la cobertura es inferior, y por lo tanto, el valor final para $F_{\beta=1}$ se ve también afectado. Esto se debe a que el resumen orientado a abstractos se genera a partir del resumen extractivo, por lo que al comprimir y fusionar frases, podemos estar perdiendo algo de información.

	% Topicos correctos			
	< 25%	< 50%	< 75%	75-100%
COMPENDIUM$_E$	5%	12,5%	47,5%	35%
COMPENDIUM$_{E-A}$	7,5%	17,5%	42,5%	32,5%

Cuadro A.5. Porcentaje de tópicos clave incluídos en los resúmenes.

En cuanto a los tópicos identificados en los resúmenes, un número considerable de resúmenes reflejan, por lo menos la mitad de los tópicos clave (82,5% y 75%, para los resúmenes generados con COMPENDIUM$_E$ y COMPENDIUM$_{E-A}$, respectivamente). Un aspecto que debemos tener en cuenta es que los resúmenes generados con COMPENDIUM en este caso son genéricos, lo que puede influir en que no aparezcan todos los tópicos que el autor del artículo ha considerado como clave.

%		**P1**	**P2**	**P3**
1. Totalmente en desacuerdo	COMPENDIUM$_E$	9,76	19,51	19,51
	COMPENDIUM$_{E-A}$	2,44	0	2,44
2. En desacuerdo	COMPENDIUM$_E$	41,46	19,51	34,15
	COMPENDIUM$_{E-A}$	31,37	21,95	31,71
3. Ni de acuerdo ni en desacuerdo	COMPENDIUM$_E$	24,39	29,27	26,83
	COMPENDIUM$_{E-A}$	21,95	29,27	26,83
4. De acuerdo	COMPENDIUM$_E$	21,95	21,95	7,32
	COMPENDIUM$_{E-A}$	41,46	39,02	34,15
5. Totalmente de acuerdo	COMPENDIUM$_E$	2,44	9,76	12,20
	COMPENDIUM$_{E-A}$	2,44	9,76	4,88

Cuadro A.6. Estudio de la satisfacción del usuario.

Finalmente, en el último análisis realizado, los usuarios muestran mayor satisfacción hacia los resúmenes orientados a abstractos, como se puede ver en el cuadro A.6, donde el porcentaje de resúmenes orientados a abstractos con los que los usuarios están de acuerdo para cada una de las preguntas (41,46%, 39,02% y 34,15%) es mayor que el de los extractos.

A.8 COMPENDIUM en aplicaciones de Tecnologías del Lenguage Humano

Además de la evaluación intrínseca de COMPENDIUM, su integración en otras aplicaciones de TLH sirve para evaluar también de forma extrínseca la herramienta. Concretamente, COMPENDIUM se ha aplicado a tareas de minería de opiniones, búsqueda de respuestas y clasificación de textos. En todos los casos, el uso de resúmenes, y en particular, de COMPENDIUM, es adecuado. Detallamos a continuación los resultados más significativos que se han obtenido.

- **Minería de opiniones.**

  En este caso concreto, aplicamos técnicas de generación de resúmenes (COMPENDIUM) a un sistema de generación de resúmenes subjetivos que solamente se basa en técnicas de minería de opiniones. Este sistema participó en la edición del TAC 2008, concretamente en la tarea de *Opinion Summarization*[15].

  Los resultados obtenidos se muestran en el cuadro A.7. Se observa que la aplicación de COMPENDIUM a los enfoques propuestos iniciales (Snippet-driven y Blog-driven) es muy beneficiosa, mejorando los resultados iniciales en un 80% y un 70%, respectivamente.

Enfoque	**Cobertura**	**Precisión**	**F$_{\beta=1}$**
Mejor sistema del TAC	-	-	0,534
Segundo mejor sistema del TAC	-	-	0,490
Snippet-driven	0,592	0,272	0,357
Blog-driven	0,251	0,141	0,155
Snippet-driven+COMPENDIUM	**0,684**	**0,630**	**0,639**
Blog-driven+COMPENDIUM	0,292	0,282	0,262

Cuadro A.7. COMPENDIUM$_E$ en la tarea de *Opinion Summarization* (TAC 2008).

- **Búsqueda de respuestas.**

  La finalidad que perseguimos al integrar COMPENDIUM en un sistema de búsqueda de respuestas que se basa en la Web es sustituir los *snippets* devueltos por un motor de búsqueda (en concreto, Google) por

[15] http://www.nist.gov/tac/2008/summarization/op.summ.08.guidelines.html

resúmenes de textos de las páginas Web de dónde se han obtenido dichos snippets. La motivación es intentar mejorar el rendimiento de los sistemas de búsqueda de respuesta semánticos, es decir aquellos que se basan, por ejemplo, en roles semánticos para obtener la respuesta correcta a una pregunta y cuyo principal problema viene derivado de la falta de completitud de los *snippets*, que en su mayoría son fragmentos incompletos de texto.

A partir de un conjunto de 100 preguntas de distintos tipos (persona, organización, lugar y tiempo), cuya respuesta se encontraba en Internet, generamos resúmenes de texto usando COMPENDIUM (COMPENDIUM$_{QE}$). La longitud para estos resúmenes era la misma que para los snippets iniciales, puesto que lo que nos interesa es ver si los resúmenes de texto mejoran el sistema de búsqueda de respuestas con respecto a los *snippets.* De los resultados obtenidos (cuadro A.8), se concluye que la integración de COMPENDIUM en un sistema de búsqueda de respuestas semántico es adecuada, puesto que los resultados mejoran en un 48% de media.

		Tipo de pregunta				
Name	**%**	**persona**	**organización**	**temporal**	**localidad**	**Avg.**
Snippets	Pre	56,3	33,3	50,0	55,0	48,6
	Rec	36,0	20,0	16,0	44,0	29,0
	F$_{\beta=1}$	43,9	25,0	24,2	48,9	35,5
COMPENDIUM$_{QE}$	Pre	68,8	77,8	80,0	65,2	72,9
	Rec	44,0	28,0	48,0	60,0	45,0
	F$_{\beta=1}$	**53,7**	**41,2**	**60,0**	**62,5**	**54,3**

Cuadro A.8. Resultados de COMPENDIUM$_{QE}$ para un sistema de búsqueda de respuestas basado en roles semánticos.

- **Clasificación de textos.**

  La última tarea donde aplicamos COMPENDIUM es en la de *“rating inference”*. Esta tarea, que consiste en predecir automáticamente la puntuación asociada a un documento (por ejemplo, una reseña), es una tarea particular de la clasificación de textos. El objetivo es analizar y estudiar si, mediante el uso de resúmenes de texto, se pueden mejorar los resultados obtenidos usando el documento completo. De esta manera, podremos decir que los resúmenes son capaces de filtrar la información innecesaria.

Sobre un conjunto de 89 reseñas de bancos ingleses, generamos distintos tipos de resúmenes con COMPENDIUM (COMPENDIUM$_E$, COMPENDIUM$_{QE}$ y COMPENDIUM$_{SE}$) y extrajimos, tanto de los resúmenes como del documento completo, la raíz de las palabras, su categoría morfológica y su polaridad usando el recurso SentiWordNet (Esuli & Sebastiani, 2006). Cada reseña tenía una puntuación asociada comprendida entre 1 y 5, siendo 1 el peor valor y 5, el mejor. Es importante destacar que se trata de una tarea compleja, puesto que tenemos cinco categorías diferentes, y los límites entre algunas de ellas (sobre todo en los valores centrales, 2, 3 y 4) no están del todo claros. Debido a este hecho, decidimos evaluar los resultados en base al error cuadrático medio (MSE) que nos indica en cuánto se ha equivocado el clasificador entre la clase predecida y la real. El cuadro A.9, muestra los resultados obtenidos para el documento completo y los distintos tipos de resúmenes.

		Ratio de compresión				
Doc. completo		**10%**	**20%**	**30%**	**40%**	**50%**
Reseña completa	MSE	2,46	2,46	2,46	2,46	2,46
Método de resumen						
COMPENDIUM$_E$	MSE	3,02	2,87	2,83	3,18	2,53
COMPENDIUM$_{QE}$	MSE	2,60	2,83	2,52	2,44	2,53
COMPENDIUM$_{SE}$	MSE	2,77	**2,31**	**2,14**	2,82	2,47

Cuadro A.9. Resultados (MSE) empleando la raíz de la palabra, su categoría morfológica y su polaridad basada en SentiWordNet como características de entrenamiento.

A pesar de que no existe una tendencia general por la que podamos decir que los resúmenes funcionan mejor en todos los casos, si que podemos observar que para algunos tamaños de resúmenes (20% y 30%) y para un tipo concreto de resúmenes, los resúmenes subjetivos (COMPENDIUM$_{SE}$), el error cuadrático medio disminuye con respecto a utilzar el documento completo (2,31 ó 2,14 vs. 2,46). Una de las posibles razones por la que no se obtienen resultados muy elevados, es por la propia complejidad de la tarea. Otra razón puede estar asociada a las características seleccionadas o bien al tipo de resúmenes generado. En este último caso, igual nos hubiera convenido más generar resúmenes subjetivos pero orientados a un tema concreto, en vez de generar cada uno de los tipos por separado.

A.9 Conclusión y trabajo en progreso

Esta tesis se centra en el estudio de la tarea de generación automática de resúmenes desde el punto de vista de las TLH. A continuación se exponen las principales conclusiones y contribuciones que aporta esta tesis, que se pueden resumir en los siguiente puntos:

- **Análisis del estado de la cuestión en la tarea de generación de resúmenes de textos**.

  De la revisión del estado de la cuestión en métodos para la generación de resúmenes, se observan algunas tendencias para los próximos años. Debido al crecimiento de la Web, y en particular, de la Web 2.0, la aparición y el aumento de distintos géneros textuales, como por ejemplo, blogs, reseñas, foros, wikis, etc. hace que cada vez estos nuevos tipos de textos cobren mayor relevancia en la sociedad. Por tanto, es imprescindible que los sistemas de resúmenes sean capaces de procesarlos correctamente. Además el uso de técnicas abstractivas y multilingües serán fundamentales para poder tratar estos nuevos tipos de textos, ya que habrá que combinar en un mismo resumen, opiniones y declaraciones de distintos tipos de personas, fuentes, idiomas, etc.

- **Análisis del estado de la cuestión en lo que respecta a la evaluación automática de resúmenes**

  La evaluación automática de resúmenes continúa siendo un reto en la actualidad. A pesar de las distintas herramientas que se ha propuesto a lo largo de todos estos años, el hecho de que la mayoría de ellas, sigan comparando el contenido de un resumen automático con otro hecho por humanos (cuando los resúmenes generados por humanos sobre un mismo documento son distintos) hace que todavía quede mucho por investigar en esta tarea. Si bien en los últimos años, la investigación en este tema se ha centrado también en la evaluación de otros aspectos más relacionados con la calidad del resumen, y no tanto en su contenido, proponiendo métodos automáticos para evaluar la corrección gramatical o la coherencia de un resumen, todavía se trata de investigaciones muy preliminares.

- **Propuesta de métodos y técnicas novedosas para la generación de resúmenes**

  En esta tesis se han propuesto y analizado diversos métodos y técnicas novedosas para la generación de resúmenes. Una de ellas, es el uso del reconocimiento de la implicación textual como método para detectar y eliminar la redundancia de un documento. Otra, es la utilización del principio de la cantidad de codificación para, junto con la frecuencia de las palabras, identificar qué frases contienen información más importante. Finalmente, se ha estudiado un método basado en grafos de palabras que permite combinar información extractiva y abstractiva, y que produce como resultado, resúmenes orientados a abstractos.

- **Propuesta y desarrollo de una herramienta de resúmenes de textos: COMPENDIUM**

  Las técnicas propuestas anteriormente han servido de base para el desarrollo de COMPENDIUM, nuestra propuesta de herramienta de resúmenes. Esta herramienta se compone de diferentes etapas y, como ya se explicó anteriormente, es capaz de generar resúmenes de distintos tipos y para distintos dominios.

  En concreto, COMPENDIUM es una herramienta de resúmenes monolingüe (sólo produce resúmenes a partir de textos en inglés) y los tipos de resúmenes que es capaz de generar son resúmenes mono y multi-documento, en cuanto al tipo de entrada se refiere. En relación a la finalidad que se pretende, los resúmenes son informativos, es decir, contienen la información más importante del documento, de tal manera que el usuario, solamente leyendo el resumen pueda disponer de información básica sobre lo que trata el documento origen. Atendiendo también a la finalidad del resumen, COMPENDIUM puede generar resúmenes genéricos, orientados a un tema concreto y subjetivos. Finalmente, en relación al tipo de salida, los resúmenes son extractos u orientados a abstractos.

  Por otro lado, en cuanto a su arquitectura, se distinguen dos tipos de etapas: las que constituyen el núcleo central de la herramienta (análisis lingüístico; detección de redundancia; identificación del tópico; detección de relevancia; y generación del resumen), y una serie de etapas adicionales (similitud con la pregunta; detección de información subjetiva; y compresión y fusión de información).

Finalmente, es importante destacar que el proceso de generación de resúmenes propuesto en COMPENDIUM está fundamentado desde una perspectiva cognitiva, tomando como base las ideas de (Van Dijk, 1980), (Van Dijk & Kintsch, 1983) y además, aporta una componente computacional, siguiendo las etapas descritas en (Hovy, 2005), permitiendo la automatización de este proceso.

- **Evaluación de COMPENDIUM**

  Por una parte, se ha evaluado COMPENDIUM de acuerdo a la información que contienen los resúmenes generados respecto a resúmenes considerados como modelos. Por otra parte, en otros casos se ha evaluado su calidad en función de diferentes criterios, como la redundancia, corrección gramatical o satisfacción de usuario.

  En todos los casos se ha obtenido resultados competitivos con respecto al estado de la cuestión, lo que demuestra que la herramienta COMPENDIUM es adecuada para la generación de resúmenes de texto. Entre las fortalezas de la herramienta, tenemos que destacar la flexibilidad y adaptabilidad de la misma, puesto que, al tratarse de una herramienta que permite la integración de nuevos módulos, se pueden desarrollar módulos específicos dirigidos a una determinada tarea. En lo que se refiere a las debilidades de la herramienta, la principal limitación de COMPENDIUM es que no incorpora información semántica, y por tanto, los resúmenes generados son en su mayoría extractos. Como una primera aproximación de resúmenes orientados a abstractos, se ha propuesto el uso de grafos de palabras para comprimir y fusionar información.

- **Integración de COMPENDIUM en aplicaciones de TLH**

  La integración de COMPENDIUM en otra aplicaciones de TLH, concretamente a la minería de opiniones, la búsqueda de respuestas, y la clasificación de textos, demuestran cómo los resúmenes generados con COMPENDIUM pueden mejorar e incrementar las capacidades de éstas.

  En líneas generales, la tarea que mayor se beneficia de la aplicación de COMPENDIUM es la de minería de opiniones, debido a que es fundamental utilizar técnicas de generación de resúmenes si queremos agrupar y presentar la información subjetiva al usuario de manera coherente. Se ha demostrado también que los resúmenes influyen positivamente en los sistemas de búsqueda de respuestas que se basan en métodos semánticos, como por ejemplo, roles semánticos. En cambio, las mejoras menos significativas han tenido lugar en la tarea de clasificación de textos, en la que no podemos concluir todavía que exista un tipo de resumen concreto que influya positivamente en todos los casos. Esto puede ser debido a la complejidad de la tarea en sí, cuyo objetivo es predecir la puntuación asociada a un documento en una escala de 1 a 5. Sí que es cierto que se ha observado que los resúmenes subjetivos para algunos ratios de compresión disminuyen el error cometido usando el documento completo, lo que nos lleva a pensar que el uso de resúmenes puede ser adecuado en algunos casos.

Como trabajos que se están llevando a cabo actualmente como continuación de esta tesis, y los que se pretenden realizar en un futuro, podemos destacar las siguientes líneas de investigación a corto, medio y largo plazo:

- **Integrar COMPENDIUM en un sistema completo que sea capaz de recuperar información subjetiva, procesarla, clasificarla y resumirla.** Se trata de un trabajo en marcha que se centra en la recuperación de información contenida en blogs y analiza la influencia de cada módulo (recuperación de información, minería de opiniones, y generación de resúmenes) con respecto al sistema completo. Los resultados preliminares obtenidos hasta el momento indican que la combinación de los tres módulos en la que obtiene mejores resultados, y que la precisión obtenida supera en un 23% al promedio de los sistemas similares que participaron en TAC 2008, ya que tanto para la experimentación como la evaluación se siguieron las mismas directrices que en la tarea de *Opinion Summarization Task* del TAC 2008.
- **Incorporar técnicas semánticas a COMPENDIUM.** En la actualidad también estamos investigando nuevas técnicas que permitan la generalización de información mediante el uso de grafos de conceptos y recursos semánticos como WordNet. De esta manera, podríamos generalizar a un mayor nivel, y analizar si este tipo de generalización es más adecuado para generar resúmenes abstractivos.

- **Analizar la influencia de una ontología de dominio farmacoterapéutico (OntoFIS) en la generación de resúmenes.** Este objetivo, que se pretende realizar a corto o medio plazo, propone el estudio de la influencia de una ontología en COMPENDIUM. De esta manera podríamos investigar si la ontología aporta beneficios a la hora de generar los resúmenes, ya que a través de ésta nos permitirá conocer vocabulario específico del dominio de la ontología, así como las relaciones entre los distintos conceptos. Esto último también nos permitiría generalizar distintos conceptos en otros más abstractos, favoreciendo así, la producción de resúmenes abstractivos.

- **Desarrollar un marco cualitativo para la evaluación de resúmenes.** Este trabajo se plantea a largo plazo, y su objetivo es investigar diferentes criterios que sirvan para medir la calidad de un resumen de forma cualitativa y analizar su correlación con las evaluaciones realizadas por expertos humanos. Para ello, para cada criterio propuesto, tendríamos que: i) analizarlo en detalle; ii) extraer las características de los textos o el resumen que hacen que se cumpla dicho criterio; iii) entrenar el sistema con textos que cumplan ese criterio y testearlo con textos de prueba; iv) ver si los resultados guardan una correlación alta con las evaluaciones realizadas por humanos.

References

Agnihotri, Lalitha, Kender, John R., Dimitrova, Nevenka, & Zimmerman, John. 2005. User Study for Generating Personalized Summary Profiles. *Pages 1094–1097 of: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME).*

Aker, A., & Gaizauskas, R. 2009. Summary Generation for Toponym-Referenced Images using Object Type Language Models. *In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2009).*

Aker, A., & Gaizauskas, R. 2010a. Generating Image Descriptions Using Dependency Relational Patterns. *In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.*

Aker, A., & Gaizauskas, R. 2010b. Model Summaries for Location-related Images. *In: Proceedings of the 7th Language Resources and Evaluation Conference.*

Álvarez Angulo, Teodoro. 2002. *El resumen como estrategia de composición textual y su aplicación didáctica.* Ph.D. thesis, Universidad Complutense de Madrid.

Amigó, Enrique, Gonzalo, Julio, Peñas, Anselmo, & Verdejo, Felisa. 2005. QARLA: a Framework for the Evaluation of Text Summarization Systems. *Pages 280–289 of: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.*

Ando, Rie, Boguraev, Branimir, Byrd, Roy, & Neff, Mary. 2005. Visualization-enabled Multi-document Summarization by Iterative Residual Rescaling. *Natural Language Engineering*, **11**(1), 67–86.

Angheluta, Roxana, De Busser, Rik, & Moens, Marie-Francine. 2002. The Use of Topic Segmentation for Automatic Summarization. *Pages 66–70 of: In Proceedings of the ACL-2002 Post-Conference Workshop on Automatic Summarization.*

Aone, Chinatsu, Okurowski, Mary Ellen, & Gorlinsky, James. 1998. Trainable, Scalable Summarization Using Robust NLP and Machine Learning. *Pages 62–66 of: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics.*

Attali, Y., & Burstein, J. 2006. Automated Essay Scoring With e-rater V.2. *Journal of Technology, Learning, and Assessment*, **4**(3).

Azzam, S., Humphreys, K., & Gaizauskas, R. 1999. Using Coreference Chains for Text Summarization. *In: Proceedings of the ACL'99 Workshop on Coreference and its Applications.*

Balahur, A., & Montoyo, A. 2008a. An Incremental Multilingual Approach to Forming a Culture Dependent Emotion Triggers Database. *In: Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE 2008).*

Balahur, A., Lloret, E., Ferrández, O., Montoyo, A., Palomar, M., & Muñoz, R. 2008. The DLSIUAES Team's Participation in the TAC 2008 Tracks. *In: Proceedings of the Text Analysis Conference (TAC).*

Balahur, Alexandra, & Montoyo, Andrés. 2008b. Multilingual Feature-Driven Opinion Extraction and Summarization from Customer Reviews. *Pages 345–346 of: Proceedings of 13th International Conference on Applications of Natural Language to Information Systems.*

Balahur, Alexandra, Lloret, Elena, Boldrini, Ester, Montoyo, Andrés, Palomar, Manuel, & Martinez-Barco, Patricio. 2009. Summarizing Threads in Blogs Using Opinion Polarity. *Pages 5 – 13 of: Proceedings of the International Workshop on Events in Emerging Text Types (eETTs).*

Balahur-Dobrescu, Alexandra, Kabadjov, Mijail, Steinberger, Josef, Steinberger, Ralf, & Montoyo, Andrés. 2009. Summarizing Opinions in Blog Threads. *Pages 606–613 of: Proceedings of the Pacific Asia Conference on Language, INformation and Computa-*

tion Conference.

Baldwin, Breck, & Morton, Thomas S. 1998. Dynamic Coreference-based Summarization. *In: In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing.*

Barzilay, Regina, & Elhadad, Michael. 1999. Using Lexical Chains for Text Summarization. *Pages 111–122 of: Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization. MIT Press.

Barzilay, Regina, & Lapata, Mirella. 2005. Modeling Local Coherence: An Entity-Based Approach. *Pages 141–148 of: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05).*

Barzilay, Regina, & McKeown, Kathleen R. 2005a. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, **31**(3), 297–328.

Barzilay, Regina, & McKeown, Kathleen R. 2005b. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, **31**, 297–328.

Beineke, Philip, Hastie, Trevor, Manning, Christopher, & Vaithyanathan, Shivakumar. 2004. An exploration of Sentiment Summarization. *In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.*

Bellemare, S., Bergler, S., & Witte, R. 2008. ERSS at TAC 2008. *In: Proceedings of the Text Analysis Conference (TAC).*

Belz, Anja. 2008. Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-space Models. *Natural Language Engineering*, **14**(4), 431–455.

Berkovsky, Shlomo, Baldwin, Timothy, & Zukerman, Ingrid. 2008. Aspect-Based Personalized Text Summarization. *Pages 267–270 of: Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems.*

Biadsy, Fadi, Hirschberg, Julia, & Filatova, Elena. 2008. An Unsupervised Approach to Biography Production Using Wikipedia. *Pages 807–815 of: Proceedings of the 46th Annual Meeting of the Associ-*

ation of Computational Linguistics: Human Language Technologies.

Boguraev, Branimir K., & Neff, Mary S. 2000. Discourse Segmentation in Aid of Document Summarization. *Page 3004 of: HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 3.*

Bossard, A., Généreux, M., & Poibeau, T. 2008. Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. *In: Proceedings of the Text Analysis Conference (TAC).*

Branny, Emilia. 2007. Automatic Summary Evaluation based on Text Grammars. *Journal of Digital Information*, **8**(3).

Brin, Sergey, & Page, Lawrence. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks ISDN Systems*, **30**, 107–117.

Burges, Chis, Shaked, Tal, Renshaw, Erin, Lazier, Ari, Deeds, Matt, Hamilton, Nicole, & Hullender, Greg. 2005. Learning to Rank Using Gradient Descent. *Pages 89–96 of: ICML '05: Proceedings of the 22nd international conference on Machine learning.*

Callison-Burch, Chris. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality using Amazon's Mechanical Turk. *Pages 286–295 of: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.*

Carbonell, Jaime, & Goldstein, Jade. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. *Pages 335–336 of: SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.*

Carenini, Giuseppe, & Cheung, Jackie C. K. 2008. Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. *Pages 33–40 of: Proceedings of the Fifth International Natural Language Generation Conference, ACL 2008.*

Carletta, Jean, Ashby, Simone, Bourban, Sebastien, Flynn, Mike, Guillemot, Mael, Hain, Thomas, Kadlec, Jaroslav, Karaiskos, Vasilis, Kraaij, Wessel, Kronenthal, Melissa, Lathoud, Guillaume, Lincoln, Mike, Lisowska, Agnes, McCowan, I., Post, Wilfried, Reidsma, Dennis, & Wellner, Pierre. 2005. The AMI Meetings Corpus. *In:* Noldus, L. P. J. J., Grieco, F., Loijens, L. W. S., & Zimmerman, Patrick H. (eds), *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior".*

Cesarano, Carmine, Mazzeo, Antonino, & Picariello, Antonio. 2007. A System for Summary-document Similarity in Notary Domain. *Pages 254–258 of: International Workshop on Database and Expert Systems Applications.*

Ceylan, Hakan, & Mihalcea, Rada. 2009. The Decomposition of Human-Written Book Summaries. *Pages 582–593 of: Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '09).*

Clarke, James, & Lapata, Mirella. 2006. Models for Sentence Compression: A Comparison across Domains, Training Requirements and Evaluation Measures. *Pages 377–384 of: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL.*

Cole, Ron (ed). 1997. *Survey of the State of the Art in Human Language Technology.* Cambridge University Press.

Conroy, J.M., & Schlesinger, J.D. 2008. CLASSY at TAC 2008 Metrics. *In: Proceedings of the Text Analysis Conference (TAC).*

Conroy, John M., & Dang, Hoa Trang. 2008. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. *Pages 145–152 of: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008).* Manchester, UK: Coling 2008 Organizing Committee.

Conroy, John M., & O'Leary, Dianne P. 2001. Text Summarization via Hidden Markov Models. *Pages 406–407 of: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Conroy, John M., Schlesinger, Judith D., & O'Leary, Dianne P. 2009. CLASSY 2009: Summarization and Metrics. *In: Proceedings of the Text Analysis Conference (TAC).*

Cristea, Dan, Postolache, Oana, & Pistol, Ionut. 2005. Summarisation Through Discourse Structure. *Pages 632–644 of: Proceedings of the Computational Linguistics and Intelligent Text Processing, 6th International Conference.*

Cunha, Iria Da, Fernández, Silvia, Velázquez-Morales, Patricia, Vivaldi, Jorge, SanJuan, Eric, & Moreno, Juan Manuel Torres. 2007. A New Hybrid Summarizer Based on Vector Space Model, Statistical Physics and Linguistics. *Pages 872–882 of: MICAI 2007: Advances in Artificial Intelligence.*

Cunningham, Hamish, Maynard, D., Bontcheva, K., & Tablan, V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *In: Proceedings of the* 40^{th} *Annual Meeting of the Association for Computational Linguistics (ACL'02).*

Dalianis, H., & Hassel, M. 2001. *Development of a Swedish Corpus for Evaluating Summarizers and other IR-tools.* Tech. rept. Technical report TRITA-NAP0112, IPLab-188, NADA, KTH.

Dang, Hoa Trang. 2006. Overview of DUC 2006. *In: The Document Understanding Workshop (presented at the* HLT/NAACL*).*

Demner Fushman, Dina, & Lin, Jimmy. 2006. Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering. *Pages 841–848 of: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics.*

Deschacht, Koen, & Moens, Marie-Francine. 2007. Text Analysis for Automatic Image Annotation. *Pages 1000–1007 of: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.*

Díaz, Alberto, & Gervás, Pablo. 2007. User-model based Personalized Summarization. *Information Processing & Management*, **43**(6), 1715–1734.

Donaway, Robert L., Drummey, Kevin W., & Mather, Laura A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. *Pages 69–78 of: Proceedings of NAACL-ANLP 2000 Workshop on Automatic Summarization.*

Dong, Z., & Dong, Q. 2003. HowNet – a Hybrid Language and Knowledge Resource. *Pages 820–824 of: Proceedings of Natural Language Processing and Knowledge Engineering conference.*

DRAE. 22$^{\text{a}}$ edición. *Diccionario de la Lengua Española.* http://rae.es.

Dunlavy, Daniel M., O'Leary, Dianne P., Conroy, John M., & Schlesinger, Judith D. 2007. QCS: A System for Querying, Clustering and Summarizing Documents. *Information Processing & Management*, **43**(6), 1588–1605.

Edmundson, H. P. 1969. New Methods in Automatic Extracting. *Pages 23–42 of: Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization. MIT Press.

El-Haj, Mahmoud, Kruschwitz, Udo, & Fox, Chris. 2010 (may). Using Mechanical Turk to Create a Corpus of Arabic Summaries. *In: Proceedings of the Seventh conference on International Language Resources and Evaluation.*

Elsner, Micha, & Charniak, Eugene. 2008. Coreference-inspired Coherence Modeling. *Pages 41–44 of: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies, Short Papers.*

Ercan, Gonenc, & Cicekli, Ilyas. 2008. Lexical Cohesion Based Topic Modeling for Summarization. *Pages 582–592 of: Proceedings of the 9th International Conference in Computational Linguistics and Intelligent Text Processing.*

Erkan, Güneş, & Radev, Dragomir R. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research (JAIR)*, **22**, 457–479.

Esuli, Andrea, & Sebastiani, Fabrizio. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Pages 417–422 of: Proceedings of the 5th Conference on Language Resources and Evaluation.*

Fan, Jianping, Gao, Yuli, Luo, Hangzai, Keim, Daniel A., & Li, Zongmin. 2008. A Novel Approach to Enable Semantic and Visual Image Summarization for Exploratory Image Search. *Pages 358–365 of: Proceeding of the 1st ACM international conference on Multimedia Information Retrieval.*

Fellbaum, C. 1998. *WordNet: An Electronical Lexical Database.* Cambridge, MA: The MIT Press.

Feng, Yansong, & Lapata, Mirella. 2008. Automatic Image Annotation Using Auxiliary Text Information. *Pages 272–280 of: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies.*

Ferrández, Óscar. 2009. *Textual Entailment Recognition and its Applicability in NLP Task.* Ph.D. thesis, Universidad de Alicante.

Ferrández, Óscar, Muñoz, Rafael, & Palomar, Manuel. 2009. Alicante University at TAC 2009: Experiments in RTE. *In: Proceedings of the Text Analysis Conference.*

Filatova, Elena, & Hatzivassiloglou, Vasileios. 2004. Event-Based Extractive Summarization. *Pages 104–111 of:* Marie-Francine Moens, Stan Szpakowicz (ed), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop.*

Filippova, Katja. 2010. Multi-Sentence Compression: Finding Shortest Paths in Word Graphs. *Pages 322–330 of: Proceedings of the 23rd International Conference on Computational Linguistics.*

Filippova, Katja, & Strube, Michael. 2008. Sentence Fusion via Dependency Graph Compression. *Pages 177–185 of: Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Fisher, Seeger, Dunlop, Aaron, Roark, Brian, Chen, Yongshun, & Burmeister, Joshua. 2009. OHSU Summarization and Entity Linking Systems. *In: Proceedings of the Text Analysis Conference (TAC).*

Fiszman, Marcelo, Rindflesch, Thomas C., & Kilicoglu, Halil. 2004. Abstraction Summarization for Managing the Biomedical Research Literature. *Pages 76–83 of:* Moldovan, Dan, & Girju, Roxana (eds), *HLT-NAACL 2004: Workshop on Computational Lexical Semantics.*

Foltz, P.W., Laham, D., & Landauer, T.K. 1999. Automated Essay Scoring: Applications to Educational Technology.

Fuentes, María, González, Edgar, Ferrés, Daniel, & Rodríguez, Horacio. 2005. QASUM-TALP at DUC 2005 Automatically Evaluated with a Pyramid Based Metric. *In: Proceedings of the Document Understanding Workshop.*

Fuentes, María, Alfonseca, Enrique, & Rodríguez, Horacio. 2007. Support Vector Machines for Query-focused Summarization trained and evaluated on Pyramid data. *Pages 57–60 of: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.*

Fukushima, Takahiro, & Okumura, Manabu. 2001. Text Summarization Challenge: Text Summarization Evaluation at NTCIR Workshop2. *Pages 9–13 of: Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization.*

Giannakopoulos, G., & Karkaletsis, V. 2009. N-GRAM GRAPHS: Representing Documents and Document Sets in Summary System Evaluation. *In: Proceedings of the Text Analysis Conference (TAC).*

Giannakopoulos, G., Karkaletsis, V., & Vouros, G. 2008a. Testing the Use of n-gram Graphs in Summarization Sub-tasks. *In: Proceedings of the Text Analysis Conference (TAC).*

Giannakopoulos, George, Karkaletsis, Vangelis, Vouros, George, & Stamatopoulos, Panagiotis. 2008b. Summarization System Evaluation Revisited: N-gram Graphs. *ACM Transactions on Speech and Language Processing*, **5**(3), 1–39.

Gillick, Dan, & Liu, Yang. 2010. Non-Expert Evaluation of Summarization Systems is Risky. *In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk.*

Givón, Talmy. 1990. *Syntax: A functional-typological introduction, II.* John Benjamins.

Glickman, Oren. 2006. *Applied Textual Entailment.* Ph.D. thesis, Bar Ilan University.

Goldstein, Jade, Mittal, Vibhu, Carbonell, Jaime, & Kantrowitz, Mark. 2000. Multi-document Summarization by Sentence Extraction. *Pages 40–48 of: NAACL-ANLP 2000 Workshop on Automatic Summarization.*

Gonçalves, Patricia Nunes, Rino, Lucia, & Vieira, Renata. 2008. Summarizing and Referring: Towards Cohesive Extracts. *Pages 253–256 of: DocEng '08: Proceeding of the 8th ACM Symposium on Document Engineering.*

Gotti, Fabrizio, Lapalme, Guy, Nerima, Luka, & Wehrli, Eric. 2007. GOFAISUM: A Symbolic Summarizer for DUC. *In: Proceedings of the Document Understanding Workshop.*

Grosz, Barbara J., Weinstein, Scott, & Joshi, Aravind K. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, **21**(2), 203–225.

Harabagiu, Sanda, & Lacatusu, Finley. 2005. Topic Themes for Multi-document Summarization. *Pages 202–209 of: Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval.*

Harabagiu, Sanda, Hickl, Andrew, & Lacatusu, Finley. 2007. Satisfying Information Needs with Multi-document Summaries. *Information Processing & Management*, **43**(6), 1619–1642.

Hasler, Laura. 2007. From Extracts to Abstracts: Human Summary Production Operations for Computer-aided Summarisation. *Pages 11–18 of: Proceedings of the RANLP 2007 Workshop on Computer-Aided Language Processing.*

Hasler, Laura. 2008. Centering Theory for Evaluation of Coherence in Computer-Aided Summaries. *In: Proceedings of the Sixth International Conference on Language Resources and Evaluation.*

Hasler, Laura, Orăsan, Constantin, & Mitkov, Ruslan. 2003. Building Better Corpora for Summarization. *Pages 309–319 of: Proceedings of Corpus Linguistics 2003.*

Hassel, M. 2004. *Evaluation of Automatic Text Summarization: A Practical Implementation.*

Hassel, Martin. 2007. *Resource Lean and Portable Automatic Text Summarization.* Ph.D. thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden.

He, Liwei, Sanocki, Elizabeth, Gupta, Anoop, & Grudin, Jonathan. 1999. Auto-summarization of Audio-video Presentations. *Pages*

489–498 of: Proceedings of the seventh ACM International Conference on Multimedia (Part 1).

He, T., Chen, J., Gui, Z., & Li, F. 2008a. CCNU at TAC 2008: Proceeding on Using Semantic Method for Automated Summarization. *In: Proceedings of the Text Analysis Conference (TAC).*

He, Tingting, Chen, Jinguang, Ma, Liang, Gui, Zhuoming, Li, Fang, Shao, Wei, & Wang, Qian. 2008b. ROUGE-C: A Fully Automated Evaluation Method for Multi-document Summarization.

Hearst, Marti A. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, **23**(1), 33–64.

Hirao, Tsutomu, Okumura, Manabu, Fukusima, Takahiro, & Nanba, Hidetsugu. 2005. Text Summarization Challenge 3 - Text Summarization Evaluation at NTCIR Workshop 4 -. *Pages 407–411 of: Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization.*

Hovy, Eduard. 2005. *The Oxford Handbook of Computational Linguistics.* Oxford University Press. Chap. Text Summarization, pages 583–598.

Hovy, Eduard, Lin, Chin-Yew, Zhou, Liang, & Fukumoto, Junichi. 2006. Automated Summarization Evaluation with Basic Elements. *In: Proceedings of the 5th International Conference on Language Resources and Evaluation.*

Hovy, E.H., & Lin, C-Y. 1999. *Automated Multilingual Text Summarization and its Evaluation.* Tech. rept. Information Sciences Institute, University of Southern California.

Isozaki, H., & Kazawa, H. 2002. Efficient Support Vector Classifiers for Named Entity Recognition. *Pages 390–396 of: Proceedings of the 19th International Conference on Computational Linguistics.*

Jaoua, Maher, & Hamadou, Abdelmajid Ben. 2003. Automatic Text Summarization of Scientific Articles Based on Classification of Extract's Population. *Pages 623–634 of: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing.*

Ji, Shaojun. 2007. A textual perspective on Givon's quantity principle. *Journal of Pragmatics*, **39**(2), 292–304.

Jing, Hongyan. 2002. Using Hidden Markov Modeling to Decompose Human-written Summaries. *Computational Linguistics*, **28**(4), 527–543.

Jing, Hongyan, & McKeown, Kathleen R. 2000. Cut and Paste based Text Summarization. *Pages 178–185 of: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference.*

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Pages 137–142 of: Proceedings of the 10th European Conference on Machine Learning.* Lecture Notes in Computer Science, no. 1398.

Jurafsky, Daniel, & Martin, James H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall.

Kaisser, Michael, Hearst, Marti A., & Lowe, John B. 2008. Improving Search Results Quality by Customizing Summary Lengths. *Pages 701–709 of: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies.*

Kan, Min-Yen, & Klavans, Judith L. 2002. Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. *Pages 36–45 of: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries.*

Kan, Min-Yen, Klavans, Judith L., & Mckeown, Kathleen R. 2002. Using the Annotated Bibliography as a Resource for Indicative Summarization. *Pages 1746–1752 of: Proceedings of the Language Resources and Evaluation Conference.*

Katragadda, Rahul. 2010. GEMS: Generative Modeling for Evaluation of Summaries. *Pages 724–735 of: Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing.*

Kazantseva, Anna. 2006. An Approach to Summarizing Short Stories. *In: Proceedings of the Student Research Workshop at the*

11th Conference of the European Chapter of the Association for Computational Linguistics.

Ker, Sue J., & Chen, Jen-Nan. 2000. A Text Categorization based on Summarization Technique. *Pages 79–83 of: Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval.*

Khan, Afnan Ullah, Khan, Shahzad, & Mahmood, Waqar. 2005. MRST: A New Technique For Information Summarization. *Pages 249–252 of: The Second World Enformatika Conference.*

Kumar, Chandan, Pingali, Prasad, & Varma, Vasudeva. 2008. Generating Personalized Summaries Using Publicly Available Web Documents. *Pages 103–106 of: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology.*

Kumar, M. A., & Gopal, M. 2009. Text Categorization Using Fuzzy Proximal SVM and Distributional Clustering of Words. *Pages 52–61 of: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.*

Kumar, Mohit, Das, Dipanjan, Agarwal, Sachin, & Rudnicky, Alexander. 2009. Non-textual Event Summarization by Applying Machine Learning to Template-based Language Generation. *Pages 67–71 of: Proceedings of the 2009 Workshop on Language Generation and Summarisation.*

Kuo, June-Jei, & Chen, Hsin-Hsi. 2008. Multidocument Summary Generation: Using Informative and Event Words. *ACM Transactions on Asian Language Information Processing*, **7**(1), 1–23.

Kupiec, Julian, Pedersen, Jan, & Chen, Francine. 1995. A Trainable Document Summarizer. *Pages 68–73 of: Proceedings of the 18th annual international ACM SIGIR Conference on Research and Development in Information Retrieval.*

Landauer, T. K., Foltz, P. W., & Laham, D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. 2003. *Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor.* Lawrence Erlbaum Associates, Inc. Pages 87–112.

Lapata, Mirella, & Barzilay, Regina. 2005. Automatic Evaluation of Text Coherence: Models and Representations. *Pages 1085–1090 of: Proceedings of the 19th International Joint Conference on Artificial Intelligence.*

Lerman, Kevin, & McDonald, Ryan. 2009 (June). Contrastive Summarization: An Experiment with Consumer Reviews. *Pages 113–116 of: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Lerman, Kevin, Blair-Goldensohn, Sasha, & McDonald, Ryan. 2009 (March). Sentiment Summarization: Evaluating and Learning User Preferences. *Pages 514–522 of: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.*

Li, S., Wan, W., & Wang, C. 2008. TAC 2008 Update Summarization Task of ICL. *In: Proceedings of the Text Analysis Conference (TAC).*

Li, Sujian, Ouyang, You, Wang, Wei, & Sun, Bin. 2007. Multi-document Summarization Using Support Vector Regression. *In: Proceedings of the Document Understanding Workshop.*

Li, Sujian, Wang, Wei, & Zhang, Yongwei. 2009a. TAC 2009 Update Summarization of ICL. *In: Proceedings of the Text Analysis Conference (TAC).*

Li, Y., Bontcheva, K., & Cunningham, H. 2009b. Adapting SVM for Data Sparseness and Imbalance: A Case Study in Information Extraction. *Natural Language Engineering*, **15**(2), 241–271.

Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. *Pages 74–81 of: Proceedings of Association of Computational Linguistics Text Summarization Workshop.*

Lin, Chin-Yew, & Hovy, Eduard. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. *Pages 495–501 of: Proceedings of the 18th Conference on Computational Linguistics.*

Lin, Chin-Yew, & Hovy, Eduard. 2003. Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics. *Pages 71–78 of: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human*

Language Technology.

Liseth, A. 2004. *En evaluering av NorSum en automatisk tekstsammenfatter for norsk. Hovedfagsoppgave.* Tech. rept. Universitetet i Bergen, Seksjon for lingvistiske fag.

Liu, Fei, & Liu, Yang. 2009. From Extractive to Abstractive Meeting Summaries: Can it be done by Sentence Compression? *Pages 261–264 of: Proceedings of the ACL-IJCNLP Conference Short Papers.*

Liu, Feifan, & Liu, Yang. 2008. Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. *Pages 201–204 of: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies, Short Papers.*

Liu, Maofu, Yu, Bo, Fang, Fang, & Sun, Hao. 2009. TAC 2009 Update Summarization Task of WUST. *In: Proceedings of the Text Analysis Conference (TAC).*

Lloret, Elena, & Palomar, Manuel. 2009. A Gradual Combination of Features for Building Automatic Summarisation Systems. *Pages 16–23 of: Proceedings of the 12th International Conference on Text, Speech and Dialogue.*

Lloret, Elena, & Palomar, Manuel. 2010. Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. *Informatica. Special Issue on Computational Linguistics*, **34**(1), 29–35.

Lloret, Elena, & Palomar, Manuel. 2011a. Resúmenes de textos: nuevos retos en la Web 2.0. *Subjetividad y Procesos Cognitivos*, **14**(2), 113–126.

Lloret, Elena, & Palomar, Manuel. 2011b. Text Summarisation in Progress: A Literature Review. *Artificial Intelligence Review.* In press.

Lloret, Elena, Balahur, Alexandra, Palomar, Manuel, & Montoyo, Andrés. 2009. Towards Building a Competitive Opinion Summarization System: Challenges and Keys. *Pages 72–77 of: Proceedings of the NAACL. Student Research Workshop and Doctoral Consortium.*

Lloret, Elena, Saggion, Horacio, & Palomar, Manuel. 2010. Experiments on Summary-based Opinion Classification. *Pages 107–115 of: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *Pages 15–22 of: Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization. MIT Press.

Mani, Inderjeet. 2001a. *Automatic Summarization.* John Benjamins Pub Co.

Mani, Inderjeet. 2001b. Summarization Evaluation: An Overview. *In: Proceedings of the North American chapter of the Association for Computational Linguistics. Workshop on Automatic Summarization.*

Mani, Inderjeet, & Maybury, Mark T. 1999. *Advances in Automatic Text Summarization.* The MIT Press.

Mani, Inderjeet, House, David, Klein, Gary, Hirschman, Lynette, Firmin, Therese, & Sundheim, Beth. 1999. The TIPSTER SUMMAC Text Summarization Evaluation. *Pages 77–85 of: Proceedings of the 9th conference on European Chapter of the Association for Computational Linguistics.*

Mani, Inderjeet, Klein, Gary, House, David, Hirschman, Lynette, Firmin, Therese, & Sundheim, Beth. 2002. SUMMAC: a Text Summarization Evaluation. *Natural Language Engineering*, **8**(1), 43–68.

Mann, William C., & Thompson, Sandra A. 1988. Rhetorical structure theory: Toward a functional Theory of Text Organization. *Text*, **8**(3), 243–281.

Manning, Christopher D., Raghavan, Prabhakar, & Schtze, Hinrich. 2008. *Introduction to Information Retrieval.*

Marcu, Daniel. 1999. Discourse Trees Are Good Indicators of Importance in Text. *Pages 123–136 of: Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization. MIT Press.

Marsi, Erwin, & Krahmer, Emiel. 2005. Explorations in Sentence Fusion. *In: Proceedings of the 10th European Workshop on Natural Language Generation.*

McCargar, Victoria. 2005. Statistical Approaches to Automatic Text Summarization. *Bulletin of the American Society for Information Science and Technology*, **30**(4), 21–25.

Medelyan, Olena. 2007. Computing Lexical Chains with Graph Clustering. *Pages 85–90 of: Proceedings of the Association of Computational Linguistics Student Research Workshop.*

Mihalcea, Rada. 2004. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. *Page 20 of: Proceedings of the Association of Computational Linguistics on Interactive poster and demonstration sessions.*

Mihalcea, Rada, & Ceylan, Hakan. 2007. Explorations in Automatic Book Summarization. *Pages 380–389 of: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.*

Mille, Simon, & Wanner, Leo. 2008. Multilingual Summarization in Practice: The Case of Patent Claims. *Pages 120–129 of: Proceedings of the 12th European Association of Machine Translation conference.*

Minel, Jean-Luc, Nugier, Sylvaine, & Piat, Gérald. 1997. How to Appreciate the Quality of Automatic Text Summarization? Examples of FAN and MLUCE Protocols and their Results on SERAPHIN. *Pages 25–30 of: Proceedings of Intelligent Scalable Text Summarization Workshop in conjunction with the European Chapter of the Association of Computational Linguistics.*

Mitkov, Ruslan. 2003. *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.).* Oxford University Press.

Mitkov, Ruslan, Evans, Richard, Orăsan, Constantin, Ha, Le An, & Pekar, Viktor. 2007. Anaphora Resolution: To What Extent Does It Help NLP Applications? *Pages 179–190 of: Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium.*

Mittal, Vibhu, Kantrowitz, Mark, Goldstein, Jade, & Carbonell, Jaime. 1999. Selecting Text Spans for Document Summaries: Heuristics and Metrics. *Pages 467–473 of: Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial intelligence Conference.*

Mohammad, Saif, Dorr, Bonnie, Egan, Melissa, Hassan, Ahmed, Muthukrishan, Pradeep, Qazvinian, Vahed, Radev, Dragomir, & Zajic, David. 2009. Using Citations to Generate surveys of Scientific Paradigms. *Pages 584–592 of: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Montiel Soto, R., & García-Hernández, R.A. 2009. Comparación de Tres Modelos de Texto para la Generación Automática de Resúmenes. *Sociedad Española para el Procesamiento del Lenguaje Natural*, **43**, 303–311.

Moreda, Paloma, Navarro, Borja, & Palomar, Manuel. 2007. Corpus-based Semantic Role Approach in Information Retrieval. *Data Knowledge Engineering*, **61**(3), 467–483.

Moreda, Paloma, Llorens, Hector, Saquete, Estela, & Palomar, Manuel. 2008. The Influence of Semantic Roles in QA: A Comparative Analysis. *Pages 55–62 of: Procesamiento del Lenguaje Natural (SEPLN)*, vol. 41.

Moreno Boronat, Lidia, Palomar Sanz, Manuel, Molina Marco, Antonio, & Ferrández Rodríguez, Antonio. 1999. *Introducción al procesamiento del lenguaje natural.* Universidad de Alicante.

Mori, Tatsunori. 2002. Information Gain Ratio as Term Weight: the Case of Summarization of IR Results. *Pages 1–7 of: Proceedings of the 19th International Conference on Computational linguistics.*

Mori, Tatsunori, Nozawa, Masanori, & Asada, Yoshiaki. 2004. Multi-answer-focused Multi-document Summarization using a Question-answering Engine. *Pages 439–445 of: Proceedings of the 20th International Conference on Computational Linguistics.*

Mori, Tatsunori, Nozawa, Masanori, & Asada, Yoshiaki. 2005. Multi-answer-focused Multi-document Summarization using a Question-answering Engine. *ACM Transactions on Asian Language Information Processing*, **4**(3), 305–320.

Morris, Andrew H., Kasper, George M., & Adams, Dennis A. 1992. The Effect and Limitations of Automatic Text Condensing on Reading Comprehension Performance. *Information Systems Research*, **3**(1), 17 – 35.

Mukras, Rahman, Wiratunga, Nirmalie, Lothian, Robert, Chakraborti, Sutanu, & Harper, David. 2007. Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution. *In: Proceedings of the Textlink workshop at IJCAI-07.*

Nastase, Vivi, Milne, David, & Filippova, Katja. 2009. Summarizing with Encyclopedic Knowledge. *In: Proceedings of the Text Analysis Conference (TAC).*

Nenkova, Ani. 2005. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. *Pages 1436–1441 of: Proceedings of the American Association fro Artificial Intelligence.*

Nenkova, Ani. 2006. Summarization Evaluation for Text and Speech: Issues and Approaches. *In: INTERSPEECH-2006, paper 2079-Wed1WeS.1.*

Nenkova, Ani, Siddharthan, Advaith, & McKeown, Kathleen. 2005. Automatically Learning Cognitive Status for Multi-document Summarization of Newswire. *Pages 241–248 of: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing.*

Nenkova, Ani, Vanderwende, Lucy, & McKeown, Kathleen. 2006. A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors that Influence Summarization. *Pages 573–580 of: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Nenkova, Ani, Passonneau, Rebecca, & McKeown, Kathleen. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, **4**(2), 4.

Neto, Joel Larocca, Santos, Alexandre, Kaestner, Celso A. A., & Freitas, Alex Alves. 2000. Generating Text Summaries through the

Relative Importance of Topics. *Pages 300–309 of: IBERAMIA-SBIA '00: Proceedings of the International Joint Conference, 7th Ibero-American Conference on AI.*

Okumura, Manabu, Fukusima, Takahiro, Nanba, Hidetsugu, & Hirao, Tsutomu. 2004. Text Summarization Challenge 2 Text Summarization Evaluation at NTCIR workshop 3. *SIGIR Forum*, **38**(1), 29–38.

Orăsan, Constantin. 2004. The Influence of Personal Pronouns for Automatic Summarisation of Scientific Articles. *Pages 127–132 of: Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium.*

Orăsan, Constantin. 2007. Pronominal Anaphora Resolution for Text Summarisation. *Pages 430–436 of: Proceedings of the Recent Advances in Natural Language Processing.*

Orăsan, Constantin. 2009. Comparative Evaluation of Term-Weighting Methods for Automatic Summarization. *Journal of Quantitative Linguistics*, **16**(1), 67–95.

Orăsan, Constantin, Pekar, Viktor, & Hasler, Laura. 2004. A Comparison of Summarisation Methods based on Term Specificity Estimation. *Pages 1037 – 1041 of: Proceedings of the Fourth International Conference on Language Resources and Evaluation.*

Over, Paul, & Ligget, Walter. 2002. Introduction to DUC: An Intrinsic Evaluation of Generic News Text Summarization Systems. *In: Proceedings of the Document Understanding Workshop.*

Over, Paul, Dang, Hoa, & Harman, Donna. 2007. DUC in Context. *Information Processing and Management*, **43**(6), 1506–1520.

Owczarzak, Karolina. 2009. DEPEVAL(summ): Dependency-based Evaluation for Automatic Summaries. *Pages 190–198 of: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*

Pang, Bo, & Lee, Lillian. 2005. Seeing stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *Pages 115–124 of: Proceedings of the Association of Computational Linguistics.*

Pang, Bo, & Lee, Lillian. 2008. Opinion mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. 2002 (July). BLEU: a Method for Automatic Evaluation of Machine Translation. *Pages 311–318 of: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics.*

Pitler, Emily, & Nenkova, Ani. 2008 (October). Revisiting Readability: A Unified Framework for Predicting Text Quality. *Pages 186–195 of: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.*

Plaza, Laura. 2011. *Uso de Grafos Semánticos en la Generación Automática de Resúmenes y Estudio de su Aplicación en Distintos Dominios: Biomedicina, Periodismo y Turismo.* Ph.D. thesis.

Plaza, Laura, Díaz, Alberto, & Gervás, Pablo. 2008. Concept-Graph Based Biomedical Automatic Summarization Using Ontologies. *Pages 53–56 of: Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing.*

Plaza, Laura, Lloret, Elena, & Aker, Ahmet. 2010. Improving Automatic Image Captioning Using Text Summarization Techniques. *In: Proceedings of the 13th International Conference on Text, Speech and Dialogue (TSD).*

Qazvinian, Vahed, & Radev, Dragomir R. 2008. Scientific Paper Summarization Using Citation Summary Networks. *Pages 689–696 of: Proceedings of the 22nd International Conference on Computational Linguistics.*

Qiu, Li-Qing, Pang, Bin, Lin, Sai-Qun, & Chen, Peng. 2007. A Novel Approach to Multi-document Summarization. *Pages 187–191 of: Proceedings of the 18th International Workshop on Database and Expert Systems Applications (DEXA 2007), 3-7 September 2007, Regensburg, Germany.*

Radev, Dragomir, Allison, Tim, Blair-Goldensohn, Sasha, Blitzer, John, Celebi, Arda, Drabek, Elliott, Lam, Wai, Liu, Danyu, Otterbacher, Jahna, Qi, Hong, Saggion, Horacio, Teufel, Simone, Topper, Michael, Winkel, Adam, & Zhang, Zhu. 2004. MEAD - A Platform for Multidocument Multilingual Text Summarization.

In: Proceedings of the 4th International Conference on Language Resources and Evaluation.

Radev, Dragomir R., & Fan, Weiguo. 2000. Automatic Summarization of Search Engine Hit Lists. *Pages 99–109 of: Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval.*

Radev, Dragomir R., & Tam, Daniel. 2003. Summarization Evaluation Using Relative Utility. *Pages 508–511 of: CIKM '03: Proceedings of the 12th international conference on Information and knowledge management.*

Radev, Dragomir R., Blair-Goldensohn, Sasha, & Zhang, Zhu. 2001. Experiments in Single and Multi-Document Summarization using MEAD. *Pages 1–7 of: First Document Understanding Conference.*

Ramshaw, Lance A., & Marcus, Mitchell P. 1995. Text Chunking Using Transformation-Based Learning. *In: Proceedings of the Third ACL Workshop on Very Large Corpora.*

Ravindra, G., Balakrishnan, N., & Ramakrishnan, K.R. 2006. Methods for Automatic Evaluation of Sentence Extract Summaries. *In: Proceedings of the 2nd International Conference on Universal Digital Library, Alexandria, Egypt.*

Romá, María Teresa. 2009. *OntoFIS: tecnología ontológica en el dominio farmacoterapéutico.* Ph.D. thesis.

Saggion, H. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Languages*, **49**, 103–125.

Saggion, H., & Funk, A. 2009. Extracting Opinions and Facts for Business Intelligence. *RNTI*, **E-17**, 119–146.

Saggion, H., Teufel, S., Radev, D., & Lam, W. 2002. Meta-evaluation of Summaries in a Cross-lingual Environment Using Content-based Metrics. *Pages 1–7 of: Proceedings of the 19th international conference on Computational linguistics.*

Saggion, Horacio. 2009. A Classification Algorithm for Predicting the Structure of Summaries. *Pages 31–38 of: Proceedings of the 2009 Workshop on Language Generation and Summarisation.*

Saggion, Horacio, & Lapalme, Guy. 2000. Selective Analysis for Automatic Abstracting: Evaluating Indicativeness and Acceptability. *Pages 747–764 of: Proceedings of Content-Based Multimedia Information Access.*

Saggion, Horacio, Lloret, Elena, & Palomar, Manuel. 2010. Using Text Summaries for Predicting Rating Scales. *In: Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.*

Sakai, Tetsuya, & Spärck Jones, Karen. 2001. Generic Summaries for Indexing in Information Retrieval. *Pages 190–198 of: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Saravanan, M., Ravindran, B., & Raman, S. 2006. Improving Legal Document Summarization Using Graphical Models. *Pages 51–60 of: Proceedings of the Nineteenth Annual Conference on Legal Knowledge and Information Systems - JURIX 2006.*

Sauper, Christina, & Barzilay, Regina. 2009 (August). Automatically Generating Wikipedia Articles: A Structure-Aware Approach. *Pages 208–216 of: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*

Schilder, Frank, & Kondadadi, Ravikumar. 2008 (June). FastSum: Fast and Accurate Query-based Multi-document Summarization. *Pages 205–208 of: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies, Short Papers.*

Schilder, Frank, Kondadadi, Ravi, Leidner, Jochen L., & Conrad, Jack G. 2008. Thomson Reuters at TAC 2008: Aggressive Filtering with FastSum for Update and Opinion Summarization. *In: Proceedings of the Text Analysis Conference.*

Schlesinger, J. D., Okurowski, M. E., Conroy, J. M., O'Leary, D. P., Taylor, A., Hobbs, J., & Wilson, H. 2002. Understanding Machine Performance in the Context of Human Performance for Multi-Document Summarization. *In: Proceedings of the DUC 2002 Workshop on Text Summarization.*

Schrijver, Alexander. 1986. *Theory of Linear and Integer Programming.* John Wiley & Sons, Inc.

Sebastiani, Fabrizio. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**(1), 1–47.

Shen, Dou, Chen, Zheng, Yang, Qiang, Zeng, Hua-Jun, Zhang, Benyu, Lu, Yuchang, & Ma, Wei-Ying. 2004. Web-page Classification through Summarization. *Pages 242–249 of: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Shen, Dou, Yang, Qiang, & Chen, Zheng. 2007. Noise Reduction through Summarization for Web-page Classification. *Information Processing and Management*, **43**(6), 1735 – 1747.

Shi, Zhongmin, Melli, Gabor, Wang, Yang, Liu, Yudong, Gu, Baohua, Kashani, Mehdi M., Sarkar, Anoop, & Popowich, Fred. 2007. Question Answering Summarization of Multiple Biomedical Documents. *Pages 284–295 of: Proceedings of the 20th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence.*

Sjöbergh, Jonas. 2007. Older Versions of the ROUGEeval Summarization Evaluation System were Easier to Fool. *Information Processing & Management*, **43**(6), 1500–1505.

Spärck Jones, Karen. 1999. Automatic Summarizing: Factors and Directions. *Pages 1–14 of: Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization. MIT Press.

Spärck Jones, Karen. 2007. Automatic summarising: The State of the Art. *Information Processing & Management*, **43**(6), 1449–1481.

Spärck Jones, Karen, & Galliers, Julia Rose (eds). 1996. *Evaluating Natural Language Processing Systems, An Analysis and Review.* Lecture Notes in Computer Science, vol. 1083. Springer.

Steinberger, Josef, Poesio, Massimo, Kabadjov, Mijail A., & Ježek, Kerel. 2007. Two Uses of Anaphora Resolution in Summarization. *Information Processing & Management*, **43**(6), 1663–1680.

Steinberger, Josef, Jezek, Karel, & Sloup, Martin. 2008. Web Topic Summarization. *Pages 322–334 of: Proceedings of the 12th International Conference on Electronic Publishing.*

Stokes, N., Rong, J., Laugher, B., Li, Y., & Cavedon, L. 2007. NICTAs Update and Question-based Summarisation Systems at DUC 2007. *In: Proceedings of the Document Understanding Workshop.*

Strapparava, C., & Valitutti, A. 2004. WordNet-Affect: An Affective Extension of WordNet. *Pages 1083–1086 of: Proceedings ofthe 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May.*

Strzalkowski, Tomek, & Harabagiu, Sanda. 2007. *Advances in Open Domain Question Answering.*

Sun, Jian-Tao, Shen, Dou, Zeng, Hua-Jun, Yang, Qiang, Lu, Yuchang, & Chen, Zheng. 2005. Web-page Summarization using Click-through Data. *Pages 194–201 of: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Svore, Krysta M., Vanderwende, Lucy, & Burges, Christopher J.C. 2007. Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. *Pages 448–457 of: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.*

Svore, Krysta Marie, Vanderwende, Lucy, & Burges, Christopher J. C. 2008. Using Signals of Human Interest to Enhance Single-document Summarization. *Pages 1577–1580 of: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17.*

Sweeney, Simon, Crestani, Fabio, & Losada, David E. 2008. 'Show me more': Incremental length summarisation using Novelty Detection. *Information Processing & Management*, **44**(2), 663–686.

Szlávik, Zoltán, Tombros, Anastasios, & Lalmas, Mounia. 2006. Investigating the Use of Summarisation for Interactive XML Retrieval. *Pages 1068–1072 of: Proceedings of the 2006 ACM Symposium on Applied Computing.*

Tang, Jiayu, & Sanderson, Mark. 2010. Evaluation and User Preference Study on Spatial Diversity. *In: Proceedings of the 32nd European Conference on Information Retrieval (ECIR).*

Tatar, Doina, Tamaianu-Morita, Emma, Mihis, Andreea, & Lupsa, Dana. 2008. Summarization by Logic Segmentation and Text Entailment. *Pages 15–26 of: Proceedings of Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008).*

Teng, Zhi, Liu, Ye, Ren, Fuji, Tsuchiya, Seiji, & Ren, Fuji. 2008. Single Document Summarization Based on Local Topic Identification and Word Frequency. *Pages 37–41 of: Proceedings of the Seventh Mexican International Conference on Artificial Intelligence.*

Teufel, S., & Halteren, H. Van. 2004. Evaluating Information Content by Factoid Analysis : Human annotation and stability. *Pages 419–426 of: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.*

Teufel, Simone, & Moens, Marc. 2002. Summarizing Scientific Articles: Experiments with relevance and Rhetorical Status. *Computational Linguistis*, **28**(4), 409–445.

Titov, Ivan, & McDonald, Ryan. 2008 (June). A Joint Model of Text and Aspect Ratings for Sentiment Summarization. *Pages 308–316 of: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies.*

Torres-Moreno, Juan-Manuel, St-Onge, Pier-Luc, Gagnon, Michel, El-Bèze, Marc, & Bellot, Patrice. 2009. Automatic Summarization System coupled with a Question-Answering System (QAAS). *NLP News. Computation and Language.*

Toutanova, Kristina, Brockett, Chris, Gamon, Michael, Jagarlamudi, Jagadeesh, suzuki, Hisami, & Vanderwende, Lucy. 2007. The PYTHY Summarization System: Microsoft Research at DUC 2007. *In: Proceedings of the Document Understanding Workshop.*

Trappey, Amy, Trappey, Charles, & Wu, Chun-Yi. 2009. Automatic patent Document Summarization for Collaborative Knowledge Systems and Services. *Journal of Systems Science and Systems Engineering*, **18**(1), 71–94.

Trappey, Amy J. C., & Trappey, Charles V. 2008. An R&D Knowledge Management Method for Patent Document Summarization. *Industrial Management and Data Systems*, **108**(2), 245–257.

Tratz, S., & Hovy, E. 2008. Summarization Evaluatin Using Transformed Basic Elements. *In: Proceedings of the Text Analysis Conference (TAC).*

Tseng, Yuen-Hsien, Lin, Chi-Jen, & Lin, Yu-I. 2007. Text Mining Techniques for Patent Analysis. *Information Processing and Management*, **43**(5), 1216–1247.

Turchi, Marco, Steinberger, Josef, Kabadjov, Mijail, & Steinberger, Ralf. 2010. Using Parallel Corpora for Multilingual (Multi-document) Summarisation Evaluation. *Pages 52–63 of: Multilingual and Multimodal Information Access Evaluation.* Lecture Notes in Computer Science, vol. 6360.

Urdan, Timothy C. 2005. *Statistics in Plain English.* Psychology Press.

Vadlapudi, Ravikiran, & Katragadda, Rahul. 2010a. On Automated Evaluation of Readability of Summaries: Capturing Grammaticality, Focus, Structure and Coherence. *Pages 7–12 of: Proceedings of the NAACL HLT 2010 Student Research Workshop.*

Vadlapudi, Ravikiran, & Katragadda, Rahul. 2010b. Quantitative Evaluation of Grammaticality of Summaries. *Pages 736–747 of: Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2010, Iasi, Romania.*

Van Dijk, T. 1980. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Van Dijk, T. A., & Kintsch, W. 1983. *Strategies of Discourse Comprehension.* New York: Academic Press, Inc.

Van Dijk, T.A. 1972. *Some Aspects of Text Grammars. A Study in Theoretical Linguistics and Poetics.* The Hague, Paris, Mouton.

Van Rijsbergen, C. J. 1981. *Information Retrieval.* Elsevier Science & Technology.

Venegás, R. 2009. Towards a method for assessing summaries in Spanish using LSA.

Wan, Xiaojun, Yang, Jianwu, & Xiao, Jianguo. 2007. Towards a Unified Approach Based on Affinity Graph to Various Multi-document Summarizations. *Pages 297–308 of: Proceedings of the 11th European Conference.*

Wang, Chan, Long, Lixia, & Li, Lei. 2008. HowNet Based Evaluation for Chinese Text Summarization. *Pages 82–87 of: Proceedings of the International Conference on Natural Language Processing and Software Engineering.*

Wilson, Theresa, Wiebe, Janyce, & Hoffmann, Paul. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. *Pages 347–354 of: Proceedings of the Empirical Methods for Natural Language Processing.*

Witte, René, Krestel, Ralf, & Bergler, Sabine. 2007. Generating Update Summaries for DUC 2007. *In: Proceedings of the Document Understanding Workshop.*

Wong, Kam-Fai, Wu, Mingli, & Li, Wenjie. 2008. Extractive Summarization Using Supervised and Semi-Supervised Learning. *Pages 985–992 of: Proceedings of the 22nd International Conference on Computational Linguistics.*

Yang, Xiao-Peng, & Liu, Xiao-Rong. 2008. Personalized multi-document summarization in information retrieval. *Machine Learning and Cybernetics, 2008 International Conference on*, **7**(July), 4108–4112.

Yu, Jin, Reiter, Ehud, Hunter, Jim, & Mellish, Chris. 2007. Choosing the Content of Textual Summaries of Large Time-series Data Sets. *Natural Language Engineering*, **13**(1), 25–49.

Zajic, David, Dorr, Bonnie J., Lin, Jimmy, & Schwartz, Richard. 2007. Multi-candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. *Information Processing & Management*, **43**(6), 1549–1570.

Zajic, David M., Dorr, Bonnie J., & Lin, Jimmy. 2008. Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression. *Information Processing & Management*, **44**, 1600–1610.

Zechner, Klaus, & Waibel, Alex. 2000. DiaSumm: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains. *Pages 968–974 of: Proceedings of the 18th Conference on Computational* *linguistics.*

Zhou, Liang, Ticrea, Miruna, & Hovy, Eduard. 2004. Multi-document Biography Summarization. *Pages 434–441 of: Conference on Empirical Methods in Natural Language Processing.*

Zhou, Liang, Lin, Chin-Yew, Munteanu, Dragos Stefan, & Hovy, Eduard. 2006. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. *Pages 447–454 of: Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference.*

Zhuang, Li, Jing, Feng, & Zhu, Xiao-Yan. 2006. Movie Review Mining and Summarization. *Pages 43–50 of: Proceedings of the 15th ACM International Conference on Information and Knowledge Management.*