TEXT SUMMARIZATION : AN OVERVIEW *

Elena Lloret
Dept. Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
elloret@dlsi.ua.es

Abstract

This paper presents an overview of *Text Summarization*. *Text Summarization* is a challenging problem these days. Due to the great amount of information we are provided with and thanks to the development of Internet technologies, needs of producing summaries have become more and more widespread. Summarization is a very interesting and useful task that gives support to many other tasks as well as it takes advantage of the techniques developed for related *Natural Language Processing* tasks. The paper we present here may help us to have an idea of what *Text Summarization* is and how it can be useful for.

Keywords: automatic text summarization; extracts and abstracts

*This paper has been supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01)

1 Introduction

The World Wide Web has brought us a vast amount of on-line information. Due to this fact, everytime someone searchs something on the Internet, the response obtained is lots of different Web pages with many information, which is imposible for a person to read completely. Although the attempts to generate automatic summaries began 50 years ago [40], in the recent years the field of automatic *Text Summarization (TS)* has experienced an exponential growth [25] [27] [46] due to these new tecnologies.

This paper addresses the current state-of-the-art of Text Summarization. Section 2 gives an overview of the field TS and we present the factors related to it. Section 3 explains the different approaches to generate summaries. In section 4 we present a number of Text Summarization systems existing today. Section 5 presents the common measures to evaluate those systems, whereas section 6 exposes the tendency adopted these days in Text Summarization. Finally, section 7 concludes this paper and discusses future work.

2 What is TEXT SUMMARIZATION?

2.1 Definition and types

A **summary** can be defined as a *text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s)* [23]. According to [39], text summarization is *the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or user)and task (or tasks).*

When this is done by means of a computer, i.e. automatically, we call this *Automatic Text Summarization.* Despite the fact that text summarization has traditionally been focused on text input, the input to the summarization process can also be multimedia information, such as images, video or audio, as well as on-line information or hypertexts. Furthermore, we can talk about summarizing only one document or multiple ones. In that case, this process is known as Multi-document Summarization (MDS) and the source documents in this case can be in a single-language (*monolingual*) or in different languages (*translingual or multilingual*).

The output of a summary system may be an *extract* (i.e. when a selection of "significant" sentences of a document is performed) or *abstact*, when the summary can serve as a substitute to the original document [15]. We can also distinguish between *generic* summaries and *user-focused* summaries (*a.k.a query-driven*). The first type of summaries can serve as surrogate of the original text as they may try to represent all relevant features of a source text. They are text-driven and follow a bottom-up approach using *IR techniques.* The *user-focused* summaries rely on a specification of a user information need, such a topic or query. They follow a top-down approach using *IE techniques.*

Concerning the style of the output, a broad distinction is normally made between *indicative* summaries, which are used to indicate what topics are addressed in the source text and they can give an brief idea of what the original text is about, and the *informative* summaries, which are intended to cover the topics in the source text [40][46].

2.2 Process of Automatic Text Summarization

Traditionally, summarization has been decomposed into three main stages [23] [40][53]. We will follow the Sparck Jones [53] approach, which is:

- ***interpretation*** of the source text to obtain a text representation,
- ***transformation*** of the text representation into a summary representation, and,

- finally, ***generation*** of the summary text from the summary representation

Effective summarizing requires an explicit and detailed analysis of context factors. Sparck Jones in [53] distinguishes three classes of context factors: input, purpose and output factors. We will described them briefly. For furhter information, see [53].

- **Input factors.** The features of the text to be summarized crucially determine the way a summary can be obtained. These fall into three groups, which are: *text form* (e.g. document structure); *subjet type* (ordinary, specialized or restricted) and *unit* (single or multiple documents as input).

- **Purpose factors.** These are the most important factors. They fall under three categories: *situation* refers to the context within the summary is to be used; *audience* (i.e. summary readers) and *use* (what is the summary for?).

- **Output factors.** In this class we can group: *material* (i.e. content) ; *format* and *style.*

3 Approaches to Text Summarization

Although many different approaches to text summarization can be found in the literature [46], [55], in this paper we will only take into account the one proposed by Mani and Marbury (1999) [40]. This classification, based on the level of processing that each system performs, gives an idea of which traditional approaches exist. This can be suitable as a reference point from which many techniques can be developed. Based on the traditional approaches, summarization can be characterized as approaching the problem at the *surface*, *entity*, or *discourse* levels [40].

Surface level
This approach inclines to represent information taking shallow features and then selectively combining them together in otrder to obtain a salience function that can be used to extract information. Among these features, we have:

- **Thematic features** rely on word (*significant words*) occurence statistics, so that sentences containing words that occur frequently in a text have higher weight than the rest. That means that these sentences are the important ones and they are hence extracted. Luhn (1958) [37] , who used the term frequency technique in his work, followed this approach. Before doing term frequency, a filtering task must be done using a stop-list words which contains words such as pronouns, prepositions and articles. This is the classical statistical approach. However, from a point of view of a corpus-based approach *td*idf* measure (commonly used in information retrieval) is very useful to determine *keywords* in text.

- **Location** refers to the position in text, paragraph or any other particular section in the sense that they contain the target sentences to be included in the summary. This is usually genre-dependent, but there are two basic general methods, which are *lead-method* and the *title-based method*. The first one consists of extracting only the first sentences, assuming that these are the most relevant ones, whereas the second considers that words in the headings or titles are positive relevant to summarization. Edmundson (1969) in [15] used this approach together with *cue-word* method which is explained later.

- **Background** assumes that the importance of meaning units is detemined by the presence of terms from the title or headings, initial part of the text or a user's query.

- **Cue words** and phrases, such as "in conclusion", "important", "in this paper", etc. can be very useful to determine signals of relevance or irrelevance. These kind of units can be detected automatically as well as manually.

Entity level

This approach attemps to build a representation of the text, modeling text entities and their relationships. The objective is to help to determine what is salient. This relations between entities include:

- **Similarity** occurs for example, when two words share a common stem, i.e. whose form is similar. This can be extended for phrases or paragraphs. Similarity can be calculated by vocabulary overlap or with linguistic techniques.

- **Proximity** refers to the distance between texts units. With that information is possible to establish entity relations.

- **Co-ocurrence**: meaning units can be related if they occur in common texts.

- **Thesaural relatioships among words** can be described as relationships like synonymy, hypernymy, part-of-relations (meronymy).

- **Coreference**. The idea behind coreference is that referring expressions can be linked so that, coreference chains can be built with coreferring expressions.

- **Logical relations** such as agreement, contradiction, entailment, and consistency.

- **Syntactic relations** are based on parse trees.

- **Meaning representation-based relations**, establishing relations between entities in the text as for example, predicate-argument relations.

Discourse level

The target of discourse level approaches is to model the global structure of the text and its relations in order to achieve communicative goals. The information that can be exploited at this level includes:

- **Format** of the document, such as hypertext markup or document outlines.

- **Threads of topics** as they are revealed in the text.

- **Rethorical structre of text**, representing argumentative or narrative structure. The idea behind this deals with the possibility to build the coherence structure of a text, so that the 'centrality' of textual units will reflect their importance.

  To applied all these methods, two different appoaches can be taken. These techniques described so far can be developed by using linguistic knowledge or by applying marchine learning techniques. Last ones have a support role, for example, in identifying the information to be applied at specific process stages such as interpretation or generation (for instance, they seem useful for training output sentence order).

4 Text Summarization Systems

Approaches presented so far are examples of pure techniques to apply, in order to develop summarization systems. The predominant tendency in current systems is to adopt a **hibryd approach** and combine and integrate some of the techniques mentioned before (e.g. cue phrases method combined with position and word frequency based methods in [24], or position, length weight of sentences combined with similarity of these sentences with the headline in [21]). As we have given a general overview of the classical techniques used in summarization and there is a large number of different techniques and systems, we are going to describe in this section only few of them briefly, considering systems as wholes. However, in table 1 some more systems are shown as well as their main features. To finish this section, the most recent approaches concerning summarization are mentioned.

Systems which have been selected to be broader described are the following: **MEAD** [51] , **WebInEssence** [50] , **NeATS** [36] , **GISTExter** [20] , and **NetSum** [54]

- ***MEAD*** [51]: This system was developed at the University of Michigan in 2001. It can produce both single and multi-document extractive summaries. The idea behind it is the use of the ***centroid***-based feature. Moreover, two more features are used: ***position*** and ***overlap with the first sentence***. Then, the linear combination of the three determines what sentences are most salient to include in the summary.

  The system works as follows: first of all, MEAD uses the CIDR Topic Detection and Tracking system to identify all the articles related to an emerging event. CIDR produces a set of clusters. From each cluster a centroid is built. Then, for each sentence, three values are computed: the centroid score, which measures how close the sentence to the centroid is; the position score indicates how far is the sentence with respect to the beginning of a document; and finally, the overlap with the first sentence or tittle of the document by calculating tf*idf between the given sentence and the first one. Then all these measures are normalized and sentences which are too similar to others are discarded (based on a cosine similarity measure). Any sentence that have not been discarded would be included in the summary.

- ***WebInEssence*** [50]: This system was also developed at the University of Michigan in 2001. It is more than a summarization system. It is a search engine to summmarize clusters of related Web pages which provide more contextual and summary information to help users explore retrieval results more efficiently. A version of MEAD [51] was used in the development of this Web-based summarizer, so that the features used to produce extracts are the same as the ones used in MEAD. The overall architecture of the system can be decomposed into two main stages: the first one behaves as a Web-spider that collects URLs from the Internet and then it groups the URLs into clusters. The second main stage is to create a multi-document summary from each cluster using the MEAD centroid-algorithm.

- ***NeATS*** [36] was first developed in 2001 by the University of Southern California's Information Sciences Institute. It is tailored to the genre of newspaper news. Its architecture consists of three main components: ***content selection***, ***content filtering*** and ***content presentation***. The goal of content selection is to identify important concepts mentioned in a document collection. The techniques used at this stage are *term frequency*, *topic signature* or *term clustering*. For content filtering three different filters are used: *sentence position*, *stigma words* and *redundancy filter*. To achieve the latter, NeATS uses a simplified version of MMR [9] algorithm. To ensure coherence of the summary, NeATS outputs the final sentences in their chronological order. From this system, iNeATS [35], i.e., an interactive multi-document summarization system that provides a user control over the summarization process, was later developed.

- ***GISTexter*** [20]: This system was developed in 2002 and produces single and multi-document extracts and abstracts by template-driven IE. The system performs differently depending on working with single document or multi-document summarization. For single-documents, the most relevant sentences are extracted and compressed by rules learned from a corpus of human-written abstracts. In the final stage, reduction is performed to trim the whole summary to the legnth of 100 words. When multi-document summarization has to be done, the system, based on ***Information Extraction (IE) techniques***, uses IE-style templates, either from a prior set (if the topic is well-known) or by ad-hoc generation (if it is unknown). The templates generated by CICERO IE system are then mapped into text snippets from the texts, in which anaphoric expressions are resolved. These text snippets can be used to generate coherent, informative multi-documents summaries.

- ***NetSum*** [54]. Different from the other approaches previous shown, NetSum, developed in 2007 by Microsoft Research Department, bets on single-document instead of multi-document summarization. The system produces fully automated single-document extracts of newswire articles based on neuronal nets. It uses ***machine learning techniques*** in this way: a train set is labeled so that the labels identify the best sentences. Then a set of features is extracted from each sentence in the train and test sets, and the train set is used to train the system. The system is then evaluated on the test set. The system learns from a train set the distribution of features for the best sentences and outputs a ranked list of sentences for each document. Sentences are ranked using RankNet algorithm [8].

After the brief description of the former systems, the reader can take a look at table 1, where a few more systems can be found. In order to understand what each column means, the following information is provided. In the first column (*SYSTEM [REF.], YEAR*) the name of the system with its reference and year is written; the second column (*# INPUTS*) distinguish between single document or multi-document summarization (both values are also possible. That means the system can perform both inputs). Next column (*DOMAIN SYSTEM*) indicates the genre of the input, that is, whether it is designed for domain-specific topics or for non-restricted domain. The fourth column (*FEATURES*) lists the main characteristics and techniques used in each system, and finally, last column (*OUTPUT*) represents whether the summary generated is either an extract or an abstract (with its variants. For some particular systems both values are also possible).

Table 1: Text Summarization Systems

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
Luhn [37], 1958	single-document	domain-specific (technical articles)	- term filtering and word frequency is carried out (low-frequency terms are removed) - sentences are weighted by the significant terms they contained - sentence segmentation and extraction is performed	extracts[1]
Edmundson [15], 1969	single-document	domain-specific (articles)	- techniques used in this approach are: *word frequency*, *cue phrases*, *title and heading* words and *sentence location* - it uses a corpus-based methodology	extracts
ADAM [48], 1975	single-document	domain-specific (chemistry)	- Cue phrases - term frequencies - sentence selection and rejection	indicative abstracts[2]
ANES [7], 1995	single-document	domain-specific (news)	- term and sentence weighting (*tf*idf*) - non-anaphora resolution - first sentences are added to the summary	extracts
Barzilay & Elahadad [4], 1997	single-document	*unknown*	- topic identification of the text by grouping words into *lexical chains*[3] - sentence extraction is helped by strong chains identification - non-anaphora resolution	extracts

continue on next page

[1]Although the output in Luhn's work is called *abstract*, it is more correct to say *extract*, as sentences from the source document take part into the summary.

[2]Output sentences are edited to produce somewhat different to the original ones, but not new sentences are generated

[3]Lexical chains are sequences of related terms grouped together by text cohesion relationships (e.g. synonymy or holonymy)

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
Boguraev & Kennedy [6], 1997	single-document	domain-independent	- linguistic techniques to identify salient phrasal units (*topic stamps*) - content characterisation methods to reflect global context (*capsule overview*) - anaphora resolution	*capsule overview*[4]
DimSum [3], 1997	single-document	*unknown*	- it uses corpus-based statistical NLP techniques -multi-word phrases are extracted automatically - conceptual representation of the text is performed - discourse features of lexical item within a text (name aliases, synonyms, and morphological variants) are exploited	extracts
Marcu [41], 1997	single-document	domain-specific (news)	- it uses text coherence models - RST[5] trees are built - this kind of representation is useful to determine the most important units in a text	extracts
SUMMARIST [24], 1998	single-document	domain-specific (news)	- symbolic concept-level world knowledge is combined with NLP processing techniques - stages for summarization are divided in: *topic indentification*, *interpretation* and *generation* - it is a multi-lingual system	extracts

continue on next page

[4]A capsule overview is not a conventional summary, i.e. it does not attemp to output document content as a sequence of sentences. It is a semi-formal representation of the document

[5]Rhetorical Structure Theory

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
SUMMONS[6] [44], 1998	multi-document	domain-specific (online news)	-its input is a set of templates from the *Message Understanding Conference*[7] - *key sentences* from an article are extracted using statistical techniques and measures - planning operators[8] such as *contradiction*, *agreement* or *superset* are used to synthesize a single article	extracts and abstracts
FociSum [28], 1999	sinlge-document	domain-independent	- it merges information extraction (IE) with sentence extraction techniques - the topic of the the text (called *foci* in this system) is determined dynamically from name entities and multiwords terms	extracts
MultiGen [5], 1999	multi-document	domain-specific (news)	- it identifies and synthesizes similar elements across related text from a set of multiple documents - it is based on *information fusion* and *reformulation* - sets of similar sentences are extracted (*themes*)	abstracts
Chen & Lin [26], 2000	multi-document	domain-specific (news)	- it produces multilingual (only English and Chinese) news summaries - monolingual and multilingual clustering is done - meaning units detection such as topic chains or linking elements is performed - similarity between meaning units is measured	extracts

continue on next page

[6]SUMMarizing Online NewS articles
[7]http://www-nlpir.nist.gov/related_projects/muc/index.html
[8]for more detailed information see[44]

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
CENTRIFUSER [30] [29], 2001	multi-document	domain-specific (health-care articles)	- it produces query-driven summaries - clustering is applied by SIMFINDER tool - it is based on *document topic tree* (each individual document is represented by a tree data strucutre) - *composite topic trees are designed* (they carry topic information for all articles) - query mapping using a similarity function enriched with structural information from the topic trees is done	extracts
Cut & Paste [22], 2001	single-document	domain-independent	- it uses sentence reduction and sentence combination techniques - *Key sentences* are identified by a sentence extraction algorithm that covers this techniques: *lexical coherence*, *tf*idf* score, *cue phrases* and *sentence positions*	abstracts[9]
MEAD [51], 2001	multi-document	domain-specific (news)	- it is based on sentence extraction through the features: *centroid* score, *position* and *overlap with the first sentence* - sentences too similar to others are discarded -experiments with CST[10] and Cross-document subsumption have been made	extracts

continue on next page

[9]In this case, summaries are generated by reformulating the text of the original document

[10]Cross-Document Structural Relationships proposes a taxonomy of the informational relationships between documents in clusters of related documents. This concept is similar to Rhetorical Structure Theory (RST)

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
NeATS[11] [36], 2001	multi-document	domain-specific (news)	- to select important content it uses: *sentence position*, *term frequency*, *topic signature*, *term clustering* - to avoid redundancy it employs *MMR*[12] technique [9] - to improve cohesion and coherence *stigma words* and *time stamps* are used	extracts
NewsInEssence [49], 2001	multi-document	domain-specific (online news)	- clusters are built through *CIDR Topic detection and tracking* component -it is based on the *Cross-document Structure Theory* (CST) - its summaries are produced by MEAD [51]	personalized extracts
WebInEssence [50], 2001	multi-document	domain-independent	- it is a Web-based summarization and recommendation system - it employes *centroid-based* sentence extraction technique - it uses similar techniques to NewsInEssence [49] but applied to Web documents	extracts and personalized summaries
COLUMBIA MDS [43], 2002	multi-document	domain-specific (news)	- it is a composite system that uses different summarizers depending on the input: *MultiGen* for single events or *DEMS*[13] for multiple events or biographical documents - statistical techniques are used	extracts and abstracts

continue on next page

[11]Next Generation Automated Text Summarization

[12]Maximal Marginal Relevance

[13]Dissimilarity Engine for Multidocument Summarization

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
Copeck et al. [12], 2002	single-document	domain-specific (biographies)	- it uses *machine learning* techniques - *keyphrases* are extracted and ranked - text is segmented according to sentences that talk about the same topic	extracts
GISTexter [20], 2002	single and multi-document	domain-specific (news)	- for single-document summarization, it extracts key sentences automatically using the technique of single-document decomposition - for multi-document summaries, it relies on CICERO IE system to extract relevant information by applying templates that are determined by the topic of the collection	extracts and abstracts
GLEANS [13], 2002	multi-document	*unkown*	- it performs document mapping to obtain a database-like representation that explicits their main entities and relations - each document in the collection is classified into one of these categories: *single person* , *single event*, *multiple event* and *natural disaster* - the system is IE based	headlines, extracts and abstracts
NTT [21], 2002	single-document	*unknown*	- it employs the *Support Vector Machine* (SVM) machine learning techique to classify a sentence into relevant or non-relevant - it also uses the following features to described a sentence: *position*, *length*, *weight*, *similarity* with the headline and *presence* of certains verbs or prepositions	extracts

continue on next page

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
Karamuftuoglu [31], 2002	single-document	*unkown*	- it is based on the *extract-reduce-organize* paradigm - as a pattern matching method it uses *lexical links* and *bonds*[14] - sentences are selected by SVM technique	extracts
Kraaij et al. [33], 2002	multi-document	*unkown*	- it is based on probabilistic methods: *sentence position*, *length*, *cue phrases* -for headline generation, noun phrases are exracted	headlines and extracts
Lal & Reuger [34], 2002	single-document	*unkown*	- it is built within the GATE[15] framework - it uses simple Bayes classifier to extract sentences - it resolves anaphora using GATE's ANNIE[16] module	extracts
Newsblaster [42], 2002	multi-document	domain-specific (online news)	- news articles are clustered using *Topic Detection and Tracking* (TDT) - it is a composite summarization system (it uses different strategies depending on the type of documents in each cluster) - it uses similar techniques to [43] - thumbnails of images are displayed	extracts
SumUM [52], 2002	single-document	domain-specific (technical articles)	- shallow syntactic and semantic analysis - concept identification and relevant information extraction - summary representation construction and text regeneration	abstracts

continue on next page

[14] A lexical link between two sentences is a word that appears in both sentences. When two or more lexical links between a pair of sentences occur, a lexical bond between them is constituted

[15] General architecture for Text Engineering, Univerity of Sheffield

[16] A Nearly New Information Extraction System

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
Alfonseca & Rodríguez[1], 2003	single-document	domain-specific (articles)	- relevant sentences identification (with a genetic algorithm) - relevant words and phrases from identified sentences are extracted - coherence is keept for the output	very short extracts (10 words)
GISTSumm [47], 2003	single-document	domain-independent	- it is based on the gist[17] of the source text - it uses statistical measures: *keywords* to determine what the *gist sentence* is - by means of the gist sentence, it is possible to build coherent extracts	extract
K.U. Leuven [2], 2003	sigle and multi-document	*unkown*	- *topic segmentation* and *clustering* techniques are used for multi-document task - topic segmentation, *sentence scoring* (weight, position, proximity to the topic) and *compression* are used for single-document summarization	extracts
Univ. Lethbridge [10], 2003	single and multi-document	*unkown*	- it performs topic segmentation of the text - it computes lexical chains for each segment - sentence extraction techniques are performed - it uses heuristics to do some surface repairs to make summaries coherent and readable	extracts
LAKE [14], 2004	single-document	domain-specific (news)	- it exploits keyphrase extraction methodology to identify relevant terms in the document - it is based on a supervised learning approach - it considers linguistic features like name entity recognition or multiwords	very short extracts

continue on next page

[17]The most important passage of the source text

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
MSR-NLP Summarizer [56], 2004	multi-document	domain-specific (news)	- its objetive is to identify important events as opposed to entities - it uses a graph-scoring algorithm to identify highly weighted nodes and relations - summaries are generated by extracting and merging portions of logical forms	extracts
CATS [16], 2005	multi-document	domain-specific (news)	- it analyzes which information in the document is important in order to include it in the summary - it is an Answering Text Summarizer - statistical techniques are used to computed a score for each sentence as well as temporal expresion and redundancy are solved	extracts
CLASSY [11], 2005	multi-document	domain-specific (news)	- it is a query-based system - it is based on *Hiden Markov Model* algorithm for sentence scoring and selection - it classifies sentences into two sets: those ones belonging to the summary and those ones which not	extracts
QASUM-TALP [17], 2005	multi-document	domain-specific (news)	- it is a query-driven system - summary content is selected from a set of candidate sentences in relevant passages - summaries are produced in four steps: (1) collection pre-processing, (2) question generation, (3) relevant information extraction (4) and summary content selection	extracts
ERRS [57], 2007	single and multi-document	domain-specific (news)	- it is basically a heurisitic-based system - all kinds of summaries are generated with the same data structure: *Fuzzy Coreference Cluster Graph*	extracts

continue on next page

Table 1 – continued from previous page

SYSTEM [REF.], YEAR	# INPUTS	DOMAIN SYSTEM	FEATURES	OUTPUT
FemSum [18], 2007	single and multi-document	domain-specific (news)	- the system is aimed to provide answers to complex questions - summaries are produced taking into account a syntactic and a semantic representation of the sentences - it uses graph-representation to establish relations between candidate sentences - it is composed of three language independent components: *RID* (Relevant Information Detector), *CE* (Content Extractor), *SC* (Summary Composer)	extracts
GOFAISUM [19], 2007	multi-document	domain-specific (news)	- it is only based on a symbolic approach - basically, the techniques used are *tf·idf* and syntactic pruning - sentences with the highest score are selected to build the summary	extracts
NetSum [54], 2007	single-document	domain-specific (news)	- it is based on machine learning techniques to generate summaries - it uses a neuronal network algorithm to enhance sentence features - the three sentences that best matches the document's highlights are extracted	extracts

From the systems described above, it is possible to notice that each system performs different methodologies to produce summaries. Furthermore, the inputs and the genre can be different too. That gives us an idea of how developed the state-of-the-art is and the number of different approaches that exist to tackle this field of research.

In the latest ACL (*ACL'07*) conference[18] attempts to summarize entire books [45] have been made. The argument to support this idea is that most of studies have been focused in short documents, specially in news reports and very little effort has been done on summarization of long documents, like books. Generating book summaries can be very useful for the reader to choose or discard a book only by looking at the extract or abstract of that book. On the contrary, in the previous year, european ACL conferences *(EACL'06)*[19] summarization of short fiction stories were also investigated [32] arguing that summarization is the key issue to determine whether to read a whole story or not.

Techniques employed in recent years are very similiar to the classical ones but they have to be adapted to each particular kind of system and its objectives. Improvements in machine learning techniques have allowed that they can be used to train and develop summarization systems these days as well. NetSum [54] which was presented in ACL'07 is an example of a system that uses machine learning algorithms to perform summarization.

5 Measures of Evaluation

Methods for evaluating automatic text summarization can be classified into two categories: *intrinsic* or *extrinsic* methods [38]. The first one measures the system's performance on its own, whereas the *extrinsic* methods evaluate how summaries are good enough to accomplish the purpose of some other specific task, e.g. filtering in information retrieval or report generation. An assesment of a summary can be done in different ways. Several examples, like *Shannon Game* or *Question Game* can be found in [23]. In summary evaluation programmmes such as SUMMAC[20], DUC[21] or NTCIR[22] automatic generated summaries (extracts or abstract) are evaluated mostly *instrinsically* against human reference or gold-standard summaries (ideal summaries). The problem is to establish what an ideal summary is. Humans know how to sum up the most important information of a text. However, different experts may disagree in considering which information is the best to be extracted. Automatic evaluation programmes have therefore been developed to try to give

[18]The 45th Annual Meeting of the Association of Computational Linguistics was held in Prague, Czech Republic, June 23rd30th 2007, http://ufal.mff.cuni.cz/acl2007/

[19]The 11th Conference of European Chapter of the Associationfor Computer Linguistics was held in Trento, Italy, April 3rd-7th 2006, http://eacl06.itc.it/

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

[21]http://duc.nist.gov

[22]http://research.nii.ac.jp/ntcir/

an objective point of view of evaluation. Systems like SEE[23], ROUGE[24] or BE[25] have been created to help to this task.

6 The Evolution of Text Summarization Approaches

Throughout the recent years summarization has experienced a remarkable evolution. Due to the evaluation programmes that take place every year, the field of *Text Summarization* has been improved considerably. For example, the tasks performed in *The Document Understand Conferences* (DUC) have changed from simple tasks to more complex ones. At the beginning, efforts were done to generate simple extracts from single documents usually in English. Lately, the trend has evolved to generate more sophisticated summaries such as abstracts from a number of documents, not just a single one, and in a variety of languages. Different tasks have been introduced year after year so that, apart from the gereral main task, it is possible to find taks consisting of producing summaries from a specific question or user-need, or just to generate a summary from updated news. Finally, concerning to the evaluation, the tendency has moved on to extrinsic evaluation, i.e. how useful the task is in order to help other tasks, rather than intrinsic evaluation. However, this kind of evaluation is also important to measure linguistic quality or content responsiveness, so manual evaluation is still performed by humans, together with automatic evaluation systems like BE, SEE or ROUGE introduced in the previous section. The evolution of summarization systems has not finished yet. There is still a great effort to do to achieve good and high quality summaries, either extracts or abstracts.

7 Conclusion and Future Work

In this paper, we have described a general overview of automatic text summarization. The status, and state, of automatic summarising has radically changed through the years. It has specially benefit from work of other asks, e.g. information retrieval, information extraction or text categorization. Research on this field will continue due to the fact that text summarization task has not been finished yet and there is still much effort to do, to investigate and to improve. Definition, types, different approaches and evaluation methods have been exposed as well as summarization systems features and techniques already developed. In the future we plan to contribute to improve this field by means of improving the quality of summaries, and studying the influence of other neighbour tasks techniques on summarization.

[23]Summarization Evaluation Environment, http://www.isi.edu/publications/licensed-sw/see/

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

[25]Basic Elements, http://haydn.isi.edu/BE/

References

[1] Alfonseca, E., Rodríguez, P. *Descrip-*

tion of the UAM system for generating very short summaries at DUC-2003. In *HLT/NAACL Workshop on Text Summarization / DUC 2003*, 2003.

[2] Angheluta, R., Moens, M, F., De Busser, R. *K.U.Leuven Summarization System.* In *DUC 2003, Edmonton, alberta, Canada*, 2003.

[3] Aonet, C., Okurowskit, M. E., Gorlinskyt, J., *et al. A Scalable Summarization System Using Robust NLP.* In *Proceedings of the ACL'07/EACL'97 Workshop on Intelligent Sclalable Text Summarization, pages 66-73*, 1997.

[4] Barzilay, R., Elhadad, M. *Using Lexixal Chains for Text Summarization.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[5] Barzilay, R, McKeown, K, Elhadad, M. *Information fusionn in the context of multi-document summarization.* In *Proceedings of ACL 1999*, 1999.

[6] Boguarev, B., Kennedy, C. *Salience-based content characterisation of text documents.* In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarization, pages 2-9, Madrid, Spain*, 1997.

[7] Brandow, R., Mitze, K., Rau, L. F. *Automatic condensation of electronic publications by sentence selection. Information Processing Management*, 31(5):675–685, 1995.

[8] Burges, C., Shaked, T., Renshaw, E., *et al.* Learning to rank using gradient descent. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 89–96, New York, NY, USA, 2005. ACM.

[9] Carbonell, J, Goldstein, J. *The use of MMR, diversity-based reranking for reordering documents and producing summaries.* In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[10] Chali, Y., Kolla, M, Singh, N, *et al. . The University of Lethbridge Text Summarizer at DUC 2003.* In *HLT/NAACL Workshop on Text Summarization /DUC 2003*, 2003.

[11] Conroy, J. M., Schlesinger, J. D., Goldstein, J. *CLASSY Query-Based Multi-Document Summarization.* In *the Document Understanding Workshop (presented at the* HLT/EMNLP *Annual Meeting), Vancouver, B.C., Canada*, 2005.

[12] Copeck, T., Szpakowicz, S., Japkowic, N. *Learning How Best to Summarize.* In *Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia*, 2002.

[13] Daumé III, H., Echihabi, A., Marcu, D., *et al. GLEANS: A Generator of Logical Extracts and Abstracts for Nice Summaries.* In *Workshop on Text Summarization (In conjunction with the ACL 2002*

and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia, 2002.

[14] D'Avanzo, E., Magnini, B., Vallin, A. *Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004.* In *the Document Understanding Workshop (presented at the* HLT/NAACL *Annual Meeting), Boston, USA*, 2004.

[15] Edmundson, H. P. *New Methods in Automatic Extracting.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[16] Farzindar, A., Rozon, F., Lapalme, G. *CATS A Topic-Oriented Multi-Document Summarization System at DUC 2005.* In *the Document Understanding Workshop (presented at the* HLT/EMNLP *Annual Meeting), Vancouver, B.C., Canada*, 2005.

[17] Fuentes, M., Gonzaelz, E., Ferres, D., *et al. QASUM-TALP at DUC 2005 Automatically Evaluated with a Pyramid Based Metric.* In *the Document Understanding Workshop (presented at the* HLT/EMNLP *Annual Meeting), Vancouver, B.C., Canada*, 2005.

[18] Fuentes, M., Rodríguez, H., Ferres, D. *FEMsum at DUC 2007.* In *the Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*, 2007.

[19] Gotti, F., Lapalme, G., Nerima, L., *et al. GOFAISUM: A Sympolic Summarizer for DUC.* In *the Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*, 2007.

[20] Harabagiu, S., Lacatusu, F. *Generating Single and Multi-Document Summaries with GISTEXTER.* In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization) Philadelphia, Pennsylvania, USA*, 2002.

[21] Hirao, T., Sasaki, Y., Isozaki, H., *et al. NTT's Text Summarization system for DUC-2002.* In *Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia*, 2002.

[22] Hongyan, J. *Cut-and-paste text summarization.* PhD thesis, 2001. Adviser-Kathleen R. Mckeown.

[23] Hovy, E. H. *Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics*, chapter 32, pages 583–598. Oxford University Press, 2005.

[24] Hovy, E., Lin, C. Y. *Automated Text Summarization in SUMMARIST.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[25] Hovy, E., Lin, C. Y., Zhou, L., *et al. Automated Summarization Evaluation with Basic Elements.* In *Proceedings of the 5th International Confer-*

ence on Language Resources and Evaluation (LREC). Genoa, Italy, 2006.

[26] Hsin-Hsi, C., Chuan-Jie, L. *A multilingual news summarizer.* In *Proceedings of the 18th conference on Computational linguistics*, pages 159–165, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[27] Jackson, P., Moulinier, I. *Natural language processing for online applications.* John Benjamins Publishing Company, 2002.

[28] Kan, M. Y., McKeown, K. *Information extraction and summarization: Domainn independence through focus types.* Technical report, Computer Science Department, Columbia University.

[29] Kan, M. Y., McKeown, K. R., Klavans, J. L. *Applying Natural Language Generation to Indicative Summarization.* In *the 8th European Workshop on Natural Language Generation, Toulouse, France*, 2001.

[30] Kan, M. Y., McKeown, K. R., Klavans, J. L. *Domain-specific informative and indicative summarization for information retrieval.* In *the Document Understanding Conference, New Orleans, USA*, 2001.

[31] Karamuftuoglu, M. *An approach to summarizaion based on lexical bonds.* In *Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia*, 2002.

[32] Kazantseva, A. *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*, 2006.

[33] Kraaij, W., Spitters, M, Hulth, A. *Headline extraction based on a combination of uni- and multidocument summarization techniques.* In *Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia*, 2002.

[34] Lal, P.,Rueger, S. *Extract-based Summarization with Simplification.* In *Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia*, 2002.

[35] Leuski, A., Lin, C. Y., Hovy, E. *iNEATS: Interactive Multi-document Summarization.* In *ACL'03*, 2003.

[36] Lin, C. Y., Hovy, E. *Automated Multi-document Summarization in NeATS.* In *Proceedings of the Human Language Technology (HLT) Conference. San Diego, CA*, 2001.

[37] Luhn, H., P. *The Automatic Creation of Literature Abstracts.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[38] Mani, I. *Summarization evaluation: An overview.* In *Proceedings of the*

North American chapter of the association for computational linguistics (NAACL) workshop on automatic summarization, 2001.

[39] Mani, I., House, D., Klein, G., *et al.* *The TIPSTER SUMMAC Text Summarization Evaluation.* In *Proceedings of EACL*, 1999.

[40] Mani, I., Maybury, M. T., Ed. *Advances in Automatic Text Summarization.* The MIT Press, 1999.

[41] Marcu, D. *Discourse Trees Are Good Indicators of Importance in Text.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[42] McKeown, K, Barzilay, R, Evans, D, *et al. Tracking and summarizing news on a daily basis with the Columbia's Newsblaster.* In *Proceedings of the Human Language Technology (HLT) Conference. San Diego, CA*, 2002.

[43] McKeown, K., Evans, D., Nenkova, A., *et al. The Columbia Multi-Document Summarizer for DUC 2002.* In *Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia*, 2002.

[44] McKeown, K., Radev, D. R. *Generating Summaries of Multiple News Articles.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[45] Mihalcea, R., Ceylan, H. *Explorations in Automatic Book Summarization.* In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 380–389, 2007.

[46] Padró Cirera, L., Fuentes, M.J., Alonso, L., *et al. Approaches to Text Summarization: Questions and Answers. Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601*, (22):79–102, 2004.

[47] Pardo, T., Rino, L., Nunes, M. *GistSumm: A summarization tool based on a new extractive method.* In *6th Workshop on Computational Processing of the Portuguese Language – Written and Spoken. Number 2721 in Lecture Notes in Artificial Intelligence, Springer (2003) 210–218*, 2003.

[48] Pollock, J. J., Zamora, A. *Automatic Abstrating Research at Chemical Abstracts.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[49] Radev, D., Blair-Goldensohn, S., Zhang, Z., *et al. NewsInEssence: a system for domain-independent, real-time news clustering and multi-document summarization.* In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–4, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[50] Radev, D, Weiguo, F, Zhang, Z. *Webinessence: A personalized web-based multi-document summarization and recommendation system.* In *NAACL Workshop on automatic Summarization, Pittsburg*, 2001.

[51] Radev, R., Blair-goldensohn, S, Zhang, Z. *Experiments in Single and Multi-Docuemtn Summarization using MEAD.* In *First Document Understanding Conference, New Orleans, LA*, 2001.

[52] Saggion, H, Lapalme, G. *Generating Indicative-Informative Summaries with SumUM. Computational Linguistics*, 28(4), 2002.

[53] Spärck Jones, K. *Automatic summarizing: factors and directions.* In *Inderjeet Mani and Mark Marbury, editors,* Advances in Automatic Text Summarization. MIT Press, 1999.

[54] Svore, K., Vanderwende, L., Burges, C. *Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources.* In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 448–457, 2007.

[55] Tucker, R. *Automatic summarising and the CLASP system.* PhD thesis, 1999. Cambridge.

[56] Vanderwende, L., Banko, M., Menezes, A. *Event-Centric Summary Generation.* In *the Document Understanding Workshop (presented at the* HLT/NAACL *Annual Meeting), Boston, USA*, 2004.

[57] Witte, R., Krestel, R., Bergler, S. *Generating Update Summaries for DUC 2007.* In *the Document Understanding Workshop (presented at the* HLT/NAACL*), Rochester, New York USA*, 2007.