A Text Summarization Approach under the Influence of Textual Entailment*

Elena Lloret, Óscar Ferrández, Rafael Muñoz, and Manuel Palomar

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante
San Vicente del Raspeig, Alicante 03690, Spain
{elloret, ofe, rafael, mpalomar}@dlsi.ua.es

Abstract. This paper presents how text summarization can be influenced by textual entailment. We show that if we use textual entailment recognition together with text summarization approach, we achieve good results for final summaries, obtaining an improvement of 6.78% with respect to the summarization approach only. We also compare the performance of this combined approach to two baselines (the one provided in DUC 2002 and ours based on word-frequency technique) and we discuss the preliminary results obtained in order to infer conclusions that can be useful for future research.

Keywords: text summarization, textual entailment, extract, word-frequency.

1 Introduction

Text Summarization has become a very popular Natual Language Processing (NLP) task in recent years. Due to the vast amount of information, especially since the growth of the Internet, automatic summarization has been developed and improved in order to help users manage all the information available these days. There are many other NLP tasks, such as Information Retrieval (IR), Information Extraction (IE), Question Answering (QA), Text Categorization (TC) or Textual Entailment (TE), which can interact together with the purpose of improving their performance and obtaining better results. In this paper we explore the possibility of using Textual Entailment to help text summarization task. The goal is to study how text summarization can be influenced by Textual Entailment.

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)* [1]. Summarization systems can be characterised according to many features. Following the Sparck Jones approach [2], there are three classes of context factors that influence summaries: *input*, *purpose* and *output factors*.

* This research has been supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01) and partially subsidized by the QALL-ME consortium, 6^{th} Framework Research Programme of the European Union (EU), contract number: FP6-IST-033860.

This allows summaries to be characterised by a wide range of properties. For instance, summarization has traditionally been focused on text, but 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 (single-document summarization) or multiple ones (multi-document summarization). Regarding the output, a summary may be an *extract* (i.e. when a selection of "significant" sentences of a document is performed), *abstract*, when the summary can serve as a substitute to the original document or even a *headline* (or title). It is also possible to distinguish between *generic* summaries and *user-focused* summaries. 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. The *user-focused* summaries rely on a specification of a user information need. Concerning also 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 an brief idea of what the original text is about. The other type, the *informative* summaries, are intended to cover the topics in the source text [3, 4].

On the other hand, Textual Entailment has been proposed recently as a generic framework for modelling semantic variability in many Natural Language Processing (NLP) applications. An entailment relation consists in determining whether the meaning of one text snippet (the hypothesis, H) can be inferred by another one (the text, T) [5]. Several approaches have been proposed, being the *Recognising Textual Entailment Challenges* (RTE) [6] the most referred sources for determining which ones are the most relevant.

The following examples extracted from the development corpus provided by the *Third RTE Challenge* show a true and false entailment relation between two text snippets:

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.

Both the text and the hypothesis have to be coherent expressions written in natural language, and depending on the linguistic complexity of the sentences, a shallower or deeper linguistic analysis will be required in order to verify the entailment inference.

This paper focuses on generic single-document summarization to produce informative extracts for newswire stories in English. We believe that, if other neighbour fields of research, such as textual entailment, are integrated as features in summarization systems to generate partial or final summaries, this can lead to good improvements. Taking this as our hypothesis, this paper suggests a novel approach that combines textual entailment with summarization to tackle the task. We show some promising preliminary results that can be useful for future research.

This paper is organized as follows: an overview of the background in the field of summarization and the existing work combining summarization and textual entailment is given in Section 2. In Section 3, the approach adopted in this research is explained, followed by the preliminary results and discussion in Section 4. Finally, Section 5 concludes this paper and discusses future work.

2 Background

Research on *Text Summarization* began in the late sixties, when Luhn [7] and Edmundson [8] started to study how to produce summaries automatically by means of a computer with no human intervention. Since then, many different techniques have been developed and used in text summarization and different approaches can be found in the literature [4, 9]. One that has been used as a point of reference from which many techniques have been developed is the one suggested by Mani and Maybury in [3]. This classification is based on the level of processing that each system performs and from this point of view, summarization can be characterized as approaching the problem at the *surface*, *entity*, or *discourse* level [3].

Two different ways can be adopted to tackle any NLP task and consequently summarization: a knowlegde-based or a corpus-based approach. The former uses linguistic knowledge, such as *topics signature* [10], *rethorical structure* of texts [11] or *centroid-based* feature [12], whilst the latter focuses on machine learning algorithms, for instance, *support vector machine SVM* as in the case of NTT system [13] or *neuronal nets* in NetSum [14].

Some previous work related to the background of summarization can be found in the literature. Sparck Jones in [15] carries out a review of summarization in the last decade. Furthermore, *Alonso et al.* in [4, 16] give a general overview of summarization systems providing also a description of their main features and techniques. These papers give an idea of all the different resources and approaches existing today to deal with summarization, and we can also realise that most systems combine several features instead of using only one. For example, NeATS [10] combines techniques such as *sentence position*, *term frequency*, *topics signature* whereas MEAD [12] relies on *centroid* score or *overlap with the first sentence*, among other features.

On the other hand, attempts to study the influence of textual entailment on summarization have been focused on the evaluation of summaries [17] 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. However, very little effort has been done to consider both fields together to produce extracts. Only in [18] approaches to combine summarization and textual entailment can be found, where a summary is generated either direclty from the entailment relations that appears in a text, or extracting the highest scored sentences of a document. The score of each sentence is computed as the number of sentences of the text that are entailed by it.

In contrast to the previous work, in this paper we have opted for a knowledge-based approach for summarization which incorporates textual entailment recognition into a summarization baseline system as a pre-processing tool extracting the meaningful

sentences to make the final summary construction more accurate, so we can discuss later the encouranging results of this pilot study.

3 Text Summarization Approach

A knowledge-based aproach for summarization, which uses two different resources to generate extracts of single documents: **word-frequency** and **textual entailment**, has been developed. The aim of this paper is to show a preliminary study of the influence that textual entailment has on text summarization. The output of our baseline system produces extracts[1] from single documents of newswire stories, although it can be extended to any other domain. For further details of the experiments performed, see Section 4.

– **WORD-FREQUENCY:** The core of this approach employs a technique based on the word's frequency which assumes that the more times a word appears in a document, the more relevant become the sentences that contain this word. Therefore higher scored sentences will be extracted to produce the final summary. Let's suppose that a sentence S consists of a group of tokens *t*:

$$S_1 = t_1\ t_2\ t_3\ t_4\ t_5\ t_6$$

then, the score for S_1 would be

$$Sc_{s_1} = \frac{\sum_{i=1}^{n} tf_i}{n}\ . \tag{1}$$

where

tfi = frequency of word ***i****, i.e, number of times that* ***i*** *appears in the source document*
n = length of the sentence without considering stopwords.

For example, let's have a look at these two sentences appearing in a text from DUC 2002.[2] The frequency for each word in the whole text is shown in brackets (note that for stopwords, frequency is not considered).

$S_a : Tropical(2)\ Storm(6)\ Gilbert(7)\ formed(1)\ in(0)\ the(0)\ eastern(1)$
$Caribbean(1)\ and(0)\ strengthened(1)\ into(0)\ a(0)\ hurricane(7)\ Saturday(4)$
$night(2).$
$S_b : There(0)\ were(0)\ no(0)\ reports(1)\ of(0)\ casualties(1).$

Considering that the total amount of frequencies for these sentences are 32 and 2 respectively, and the first sentence has a length of 10 words and the length for the second one is 2, the final score (Sc) for each sentence would be:

[1] The extracts have been truncated to 100-word length approximately so that we can compare them to the reference summaries provided by DUC 2002 data.
[2] Sentences have been taken from AP880911-0016 document, which belongs to cluster *d061j*.

$Sc(S_a) = 3.20$
$Sc(S_b) = 1.00$

Therefore, from those two sentences the first one would be extracted because it has higher score than the second one.

– **TEXTUAL ENTAILMENT:** The main idea here is to make up a preliminary summary by the sentences of the text that does not hold an entailment relation. Let's assume that a document consists of a list of sentences:

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

and we perform the entailment experiment as follows:

$SUM = \{S_1\}$

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

Therefore, the final summary obtained by the processed entailment inferences comprises the sentences that are not entailed by the accumulated summary of the previous non-entailed sentences (i.e S_1, S_2, S_3 and S_6 regarding the above example). To compute such inferences we have used the entailment engine presented in [19] trained with the corpora provided by the *Third Recognising Textual Entailment Challenge* [6].

Moreover, in order to assess both the entailment engine on summarization task and how the recognition of entailment relations can influence positively the overall performance of a summarization system, we propose two different evaluations: (i) on the one hand, we evaluate the summary directly obtained from the word-frequency approach; and (ii) on the other hand, we evaluate a final summary built from the highest scored sentences belonging to the preliminary entailment summary and according to the word-frequency calculus. Therefore, the aim of this pilot study is to develop an incremental system which integrates different types of knowledge. Particularly, this paper shows the performance obtained with word-frequency approach first, and then combining word-frequency and textual entailment. Results achieved for each experiment will be described in detail in the next section.

4 Experiments and Dicussion

In this section we describe the evaluation performed and the results we have obtained. In summarization we can take two evaluation methodologies depending on whether we use *intrinsic* or *extrinsic* methods [20]. Among these two types of evaluation, we have chosen the intrinsic one, so that we can evaluate how good the automatic extract is, by comparing it against human-made summaries using the ROUGE[3] tool [21]. This tool allows us to obtain the Precision, Recall and F-measure for every automatic summary (peer) compared to one or more reference summaries (models).

4.1 Evaluation Environment

As test data set, we have taken the DUC 2002 test documents and human generated summaries for single-document task.[4] That year was the last year in which single-document summarization evaluation of informative summaries was performed. The data consisted of 59 different clusters with non fixed number of documents in each one. All the documents are related to newspaper stories and those ones inside the same cluster deal with the same kind of topic. The total number of documents is 567, and two different assesors produced a 100-word manual summary for each document.[5] However, some documents are duplicated and included in more than one cluster. Therefore, summaries written for those documents were assigned to different humans depending on the cluster. For each document, we have not considered the tags that reference the author's name, document's name, title or graphics so we have processed all documents as a previous step deleting these kinds of tags. The reason why we did it was that these types of tags introduced a lot of noise in our intermediate or final summaries and although some of those tags have strong information within the summary, such as the title, in documents talking about news, this information is usually reflected in the first sentence as well. At the end, without considering repeated documents, we have dealt with 530 documents in total with at least one reference summary for each of them, having 1079 reference summaries.

Futhermore, in order to evaluate our system with a different type of document (not only with newswire documents) we have used "The Koan" [6] text in two different variants: the original text and the text incorporating manual anaphora resolution, which consisted in replacing personal pronouns with their reference. We want to show how our system can be extended to other domains, not only for newswire texts. Summaries generated by our system are compared with two human-made summaries as gold standards. One of them is the Long variant of the human extract that can be found in [18]. The other one has been written by ourselves.

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

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

[5] Except for two particular clusters where only one human-made summary was written for the documents in them.

[6] This text is available at http://www.cs.ubbcluj.ro/~dtatar/nlp/Koan-fara-anaph.txt and it consists of 65 sentences. We would like to thank Professor Doina Tatar for providing "The koan" corpus and its human-made summaries.

4.2 Results Analysis

All the experiments described in this section were evaluated using the ROUGE tool[7] [21]. We computed ROUGE-1, ROUGE-2 values as well as ROUGE-L and ROUGE-SU4 and we obtained recall, precision and F-measure on average for the sytem's performance. ROUGE-1 and ROUGE-2 compute an n-gram recall between a candiate summary and a set of reference summaries, where the length of the n-gram, in this particular case, are 1 and 2 respectively. ROUGE-L relies on the Longest Common Subsequence (LCS) between two texts and ROUGE-SU4 measures the overlap of skip-bigrams between a candidate summary and a set of reference summaries with a maximum skip distance of 4. Two independent evaluations were performed. For the first one the DUC 2002 data were used, whereas for the second evaluation "The Koan" text was used as input data for the system.

In the first experiment four different tests were performed. In the first one, we took the results performed by the DUC 2002 baseline, which consisted in generating the summary with the first 100 words of a document. These results have been provided in [22], where all DUC 2002 participants have been evaluated again, with the ROUGE tool. For this evaluation we only have recall value. [8]

Next, we decided to run our baseline, based on *word-frequency* feature, on the original documents (non pre-processed documents) for DUC 2002 documents mentioned before. The third test was identical to the second one, except that we pre-processed all documents removing tags we considered noisy. Finally, in the last test we added the *textual entailment* feature as a previous step in summarization and we ran it on the same data set as before. To evaluate all tests, the output of each test, that is, the automatic extracts (peers) were compared to human summaries (models) using the ROUGE tool previously mentioned. Results are presented in Table 1.

Table 1. Results obtained for the DUC 2002 data.

		ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-SU4**
BASELINE DUC 2002	**Recall**	**41.132%**	**21.075%**	**37.535%**	**16.604%**
BASELINE	Recall	42.483%	17.912%	38.247%	20.014%
(word-frequency	Precision	40.567%	17.024%	36.529%	19.035%
original texts)	**F-measure**	**41.468%**	**17.442%**	**37.337%**	**19.495%**
BASELINE	Recall	43.741%	17.522%	39.575%	20.195%
(word-frequency	Precision	43.398%	17.362%	39.248%	20.016%
pre-processed texts)	**F-measure**	**43.538%**	**17.435%**	**39.388%**	**20.094%**
Textual Entailment +	Recall	45.428%	19.533%	41.264%	21.779%
Word Frequency	Precision	45.004%	19.324%	40.860%	21.553%
	F-measure	**45.184%**	**19.421%**	**41.038%**	**21.654%**

[7] ROUGE version (1.5.5) run with the same parameters as in [22] (ROUGE-1.5.5.pl -n 2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -l 100 -d).

[8] ROUGE version used in [22] (ROUGE-1.4.2) only computed recall measure.

As can be seen from Table 1, our three approaches obtain better results that DUC 2002 baseline in any ROUGE measure, except for ROUGE-2 value. Moreover, when combining summarization with textual entailment F-measure values increase by 6.78% on average with respect to applying summarization using word-frequency only. When applying the word-frequency approach to the pre-processed documents, results also increase by 4.14% with respect to the DUC 2002 baseline, regarding to recall value. However, when comparing our word-frequency baseline with the whole documents to the DUC 2002 baseline, we observe that the improvement achieved is not as significant as the other results, obtaining an increase of 2.68%. Therefore, this also reveals that transforming the documents first by removing some kinds of tags in the original HTML document does not lead necessarily to a loss of relevant information. In any case, textual entailment helps positively text summarization and performs the highest improvements among all the tests carried out in this experiment, which proves that the addition of textual entailment in summarization has been appropriate.

For the second evaluation we used "The Koan" text as an input for our system. We tested our system with the original text and with the one with the resolved anaphors. The results of this evaluation can be seen in Table 2 where "NO" means that anaphora has not been taken into account whilst "YES" means that manual anaphora resolution has been employed. In this evaluation we carried out two tests: in the first one, we ran our baseline based on word-frequency whereas in the second test we tested the textual entailment together with word-frequency approach. Results in percentages are shown in Table 2.

Table 2. Results obtained for "The Koan" text.

		ROUGE-1		**ROUGE-2**		**ROUGE-L**		**ROUGE-SU4**	
		NO	YES	NO	YES	NO	YES	NO	YES
BASELINE ***(word-frequency)***	Recall	40.404	55.000	16.327	33.838	34.343	45.500	17.820	34.846
	Precision	40.404	55.556	16.327	34.184	34.343	45.960	17.820	35.208
	F-measure	**40.404**	**55.277**	**16.327**	**34.010**	**34.343**	**45.729**	**17.820**	**35.026**
Textual Entailment + Word Frequency	Recall	43.939	64.500	22.959	44.444	37.879	54.000	22.751	44.349
	Precision	44.388	64.500	23.196	44.444	38.265	54.000	22.990	44.349
	F-measure	**44.162**	**64.500**	**23.077**	**44.444**	**38.071**	**54.000**	**22.870**	**44.349**

From the results shown in Table 2 we can notice that the system also achieves promising results when dealing with documents outside the news domain, obtaining better results when textual entailment is applied. We perform 22.46% better on average when using textual entailment together with word-frequency summarization (and lack of anaphora resolution taken into account) with respect to word-frequency only. Nevertheless, the same test considering an ideal case of anaphora resolution achieves an increase of 23% on average. On the other hand, in an ideal situation of anaphora resolution context, results obtained when using word-frequency approach increase by 43.71% compared to non-anaphora resolution, whilst combining textual entailment and

summarization in the same case, performs 68.60% better, which means a significant increase with respect to non-anaphora resolution.

5 Conclusions and Future Work

In this paper the positive influence of textual entailment on summarization has been presented showing a successful approach combining both approaches. The preliminary results obtained revealed that the improvements achieved by applying textual entailment together with a summarization system, specially when incorporating some kind of anaphora resolution, encourage us to consider these two research lines (textual entailment and anaphora resolution) for further research.

The main problem to address for future research will be to extend the system for multi-document summarization. Future work can be also related to the development of a system that takes advantage of the techniques employed in textual entailment recognition, not only as a previous stage to summarization task, as well as an anaphora resolution module as one of the important tasks to take into consideration in the future. Another future research line could consist in adding more knowledge to the system, exploring new approaches based on semantic relations (for instance, WordNet relations such as synonymy or hyponymy) and graph-based relations.

References

1. Hovy, E. H.: Automated Text Summarization. In: The Oxford Handbook of Computational Linguistics. Oxford University Press (2005) 583–598
2. Spärck Jones, K.: *Automatic summarizing: factors and directions*. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1999) 1–14
3. Mani, I., Maybury, M. T.: Advances in Automatic Text Summarization. The MIT Press (1999)
4. Alonso, L., Castellón, I., Climent, S., Fuentes, M., Padró, L., Rodríguez, H.: *Approaches to Text Summarization: Questions and Answers*. Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601 (2004) 79–102
5. Glickman, O.: Applied Textual Entailment. PhD thesis, Bar Ilan University (2006)
6. Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B.: The third pascal recognizing textual entailment challenge. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, Association for Computational Linguistics (2007) 1–9
7. Luhn, H.P.: *The Automatic Creation of Literature Abstracts*. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1999) 15–22
8. Edmundson, H.P.: *New Methods in Automatic Extracting*. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1999) 23–42
9. Tucker, R.: Automatic summarising and the CLASP system. PhD thesis (1999) Cambridge.
10. Lin, C.Y., Hovy, E.: *Automated Multi-document Summarization in NeATS*. In: Proceedings of the Human Language Technology (HLT) Conference. San Diego, CA. (2002) 50–53
11. Marcu, D.: *Discourse Trees Are Good Indicators of Importance in Text*. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press (1999) 123–136
12. Radev, D.R., Blair-Goldensohn, S., Zhang, Z.: *Experiments in Single and Multi-Document Summarization using MEAD*. In: First Document Understanding Conference, New Orleans, LA. (2001) 1–7

13. Hirao, T., Sasaki, Y., Isozaki, H., Maeda, E.: *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)
14. Svore, K.M., Vanderwende, L., Burges, C.J.: *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). (2007) 448–457
15. Spärck Jones, K.: *Automatic summarising: The state of the art*. Information Processing & Management **43** (2007) 1449–1481
16. Alonso, L., Castellón, I., Climent, S., Fuentes, M., Padró, L., Rodríguez, H.: *Comparative Study of Automated Text Summarization Systems*. Technical report, Hermes Project Technical Report. Technical Report IMA 02-02-RR, Universitat de Girona.Barcelona (2002)
17. Harabagiu, S., Hickl, A., Lacatusu, F.: *Satisfying information needs with multi-document summaries*. Information Processing & Management **43** (2007) 1619–1642
18. Tatar, D., Tamaianu-Morita, E., Mihis, A., Lupsa, D.: *Summarization by Logic Segmentation and Text Entailment*. In: Proceedings of Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008). (2008) 15–26
19. Ferrández, O., Micol, D., Muñoz, R., Palomar, M.: A perspective-based approach for solving textual entailment recognition. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, Association for Computational Linguistics (2007) 66–71
20. 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)
21. Lin, C.Y.: *ROUGE: a package for automatic evaluation of summaries*. In: Proceedings of ACL Text Summarization Workshop. (2004) 74–81
22. Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K.: *Two uses of anaphora resolution in summarization*. Information Processing & Management **43** (2007) 1663–1680