A Comparative Study of the Impact of Statistical and Semantic Features in the Framework of Extractive Text Summarization

Tatiana Vodolazova, Elena Lloret, Rafael Muñoz, and Manuel Palomar

Dept. Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apdo de correos, 99, E-03080, Alicante, Spain
{tvodolazova,elloret,rafael,mpalomar}@dlsi.ua.es

Abstract. This paper evaluates the impact of a set of statistical and semantic features as applied to the task of extractive summary generation for English. This set includes word frequency, inverse sentence frequency, inverse term frequency, corpus-tailored stopwords, word senses, resolved anaphora and textual entailment. The obtained results show that not all of the selected features equally benefit the performance. The term frequency combined with stopwords filtering is a highly competitive baseline that nevertheless can be topped when semantic information is included. However, in the selected experiment environment the recall values improved less than expected and we are interested in further investigating the reasons.

1 Introduction

The research in Text Summarization (TS) is still mainly focused on extractive approaches, that attempt to determine the most relevant segments of the input document by computing their weights based on different techniques and features. These features commonly include the segment position within the original text [12], presence of the cue phrases [3], term frequency of the topic terms [5] and length of the segment [2] among others. In recent years the research in this area has been directed towards incorporating more semantic knowledge in the set of analysis features. A graph-based method combined with the word sense disambiguation (WSD) was proposed in [11]. The LeLSA+AR is a Latent Semantic Analysis-based system for TS supplied with anaphoric information [13]. Textual Entailment (TE) as well was reported to benefit the TS task [8]. However, to the best of our knowledge, there has been no research exploring in detail the impact of each of the aforementioned features and their combinations in the identical evaluation environment.

The present paper reports on the initial results of the ongoing work investigating the impact of different semantic and statistical features on the performance of extractive summary generation system. Those are represented by term frequency, inverse term and sentence frequencies, word senses, resolved anaphora, textual entailment and corpus-tailored stopword list. Although the system performance improved less than expected, the obtained results clearly show the relative importance of each of the selected features.

This paper is organized as follows. Section 2 introduces the features selected for evaluation with the brief explanations and related work reporting their usage. Section 3

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 306–313, 2012.

describes the evaluation environment. The results are reported in Section 4. Finally, the conclusions together with the future work can be found in Section 5.

2 Selected Features and Related Work

2.1 Term Frequency

Term frequency (TF) is one of the very first features used for automatic TS [9]. The impact of TF isolated from all the other features was analyzed in [10] . It was shown that the likelihood of a word appearing in a human summary depends on the word's frequency in the original text. Thus it can be hypothesized that the sentences with the highest number of most frequent words should be selected for the final summary.

For a sentence $S_j = \{t_1, t_2, ..., t_m\}$ with m tokens, its score based on the TF will be calculated using the following formula:

$$Sc_{tf(S_j)} = \frac{\sum_{i=1}^{m} tf_i}{n} \tag{2.1}$$

where tf_i is the frequency of the i_{th} token (or its stem/lemma) in the text j; n is the number of sentence tokens (stopwords removed).

2.2 Inverse Term and Sentence Frequencies

TFIDF is a common keyword identification method successfully used in the information retrieval. It involves calculating term frequencies, that as it has been hypothesized in the previous section represent a reliable measure of the sentences to be selected for the final summary. In the context of single document summarization the keywords of each single document do not depend on the other documents in the collection. The three language models were compared in [1]: Inverse Document Frequency(IDF), Inverse Sentence Frequency(ISF) and Inverse Term Frequency(ITF). It was speculated on using ISF for the systems identifying sentences and ITF for the systems identifying terms as their smallest compositional unit. Both ISF and ITF can be used for the single document summarization task, as the extractive summaries are commonly composed of the sentences of the original text and term counts proved to be a reliable method for the summary sentence identification. Roughly based on [1] the ISF and ITF measures were calculated the following way:

$$isf_{tf} = \log \frac{|S|}{|\{s \in S : t \in s\}|} \tag{2.2}$$

where $|S|$ is the total number of sentences in the document; $|\{s \in S \ t \in s\}|$ number of sentences where the term t appears.

$$itf = \log \frac{|V|}{|\{t \in V : t \in d\}|} \tag{2.3}$$

where $|V|$ is the vocabulary size of the document; $|\{t \in V \ t \in d\}|$ number of times where the term t appears in the document, i.e. term frequency.

2.3 Word Sense Disambiguation

Although the approaches based on TF result in a rather competitive baseline, they fail to capture the semantics of a document. The cases of synonymy like between the nouns "a ship" and "a boat" can be captured employing a Word Sense Disambiguation (WSD) module and using concepts instead of terms. The TF is substituted by the concept frequency and Formulas (2.1), (2.2) and (2.3) will be modified the following way:

$$Sc_{cf(S_j)} = \frac{\sum_{i=1}^{m} cf_i}{n} \tag{2.4}$$

where S_j is the j_{th} sentence $S_j = \{t_1, t_2, ..., t_m\}$ with m tokens; cf_i is the frequency of the WordNet[1] synset id in the document that the sense of the i_{th} term belongs to; n is the number of sentence tokens (stopwords removed).

$$isf_{cf} = \log \frac{|S|}{|\{s \in S : c \in s\}|} \tag{2.5}$$

where $|S|$ is the total number of sentences in the document; $|\{s \in S \;\; c \in s\}|$ number of sentences s where the WordNet synset id c in the document appears, for all $c \in d$.

$$icf = \log \frac{|V|}{|\{t \in V : c \in d\}|} \tag{2.6}$$

where $|V|$ is the vocabulary size of the document as measured in the number of different WordNet synset ids appearing in the text; $|\{c \in V \;\; c \in d\}|$ number of times where the concept c appears in the document, i.e. concept frequency.

A graph-based method for TS combined with WSD is described in [11]. The authors report the recall values up to 0.4651 for ROUGE-1 on the DUC 2002 data which improves significantly over the DUC baseline of 0.4113 and outperforms the best DUC 2002 system. The important implementation detail of this system is that the concepts were used only for nouns. The use of verbs in the graph showed the decrease in the system performance. The final set of features considered in the present paper includes concept frequencies both for nouns and verbs, nouns and adjectives.

2.4 Anaphora Resolution

Taking into account the significance of most frequent words for the final summary generation mentioned in the Section 2.1, it becomes particularly important to resolve anaphora in the original document. The previous work on including anaphora resolution (AR) reports some increase in performance. [13] achieve the improvement of around 1.5% over their summarization system based on the lexical LSA by incorporating anaphoric information into it. The performance was tested on the DUC 2002 data using the ROUGE evaluation toolkit. The authors also mention two strategies for including anaphoric relations: (1) addition, when anaphoric chains are treated as another kind of terms for the input matrix construction; (2) substitution, when each representative

[1] WordNet `http://wordnet.princeton.edu/`

of an anaphoric chain in the text is substituted by the chain's first representative. The evaluation results show that the substitution approach performs significantly worse than the addition approach and in some tests even worse than the same system without including anaphora resolution.

2.5 Textual Entailment

The task of Textual Entailment (TE) is to capture the semantic inference of texts. In the framework of TS it was used for different purposes: (1) to segment the input document into subtopics [14]; (2) when evaluating the set of generated summaries to identify the summary that can be best deduced from the original document [6]; (3) to eliminate redundant information from the final summary. In [8] TE was used in its latter application which yielded the increase in the ROUGE-1 values over the TF baseline on the task of single document summarization. The TE module similar to [8] and [4] was employed for the present research. The sentences of the original document are being handled sequentially by the TE module which determines whether the sentence that is currently being processed can be inferred from the stack of already processed ones.

2.6 Corpus-Tailored Stopwords

Inspecting the first summaries we discovered that the words like "just", "near", "away", "ago", "today" among others were not in the standard stopword list. It was decided to extend the 350-words stopword list by the missing words from the list of 245 most frequent words of the DUC 2002 corpus[2] that was extracted using the Lucene 3.5.0[3] indexing module.

3 Evaluation Environment

3.1 Evaluation Corpus and Metrics

For our experiments we used the Document Understanding Conferences 2002 data for single-document summarization task. The corpus consists of 567 newswire articles covering a wide range of topics. Each newswire article is accompanied by one or more abstractive model summaries. The model summaries are approximately 100 words long and were created manually by humans.

The system's summaries were evaluated against the model ones using the ROUGE metrics [7], which is a standard measure for evaluation by now. ROUGE-N is a family of metrics based on the overlap n-grams between a system summary and one or more model summaries, where N stands for the n-gram length. For the present paper we computed the ROUGE-1 recall values (see Section 4) that were shown to correlate with human assessment [13].

[2] DUC Conferences: `http://duc.nist.gov/`

[3] Apache Lucene `http://lucene.apache.org/`

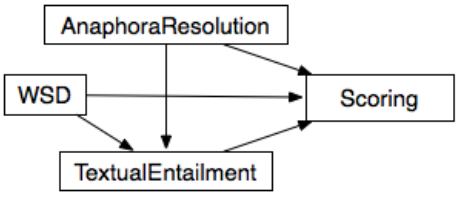

Fig. 1. Interaction of semantic components

3.2 System Settings

The core of our system is the scoring module (see Figure 1). In its basic setting it computes the score of each sentence in the original document based on the TF as in Formula (2.1). The scoring module can be set to filter out the stopwords from a custom stopword list. The second scoring strategy involves WSD and rates the sentences based on the CF as in Formula (2.4). The WSD was performed using the built-in function of Freeling[4]. Freeling package provides a number of WSD algorithms. We experimented with the most frequent sense (MFS) and the PageRank algorithm (UKB). Another setting for the WSD module is to disambiguate only nouns. This was motivated by the results provided in [11]. The more sophisticated versions of TF and CF involve computing ISF and ITF as described in Formulas (2.2), (2.3), (2.5) and (2.6). All the scoring module settings except for the stopwords, ITF and ISF can be found in the results Table 1 in the column names[5]. Before rating the sentences the original text can be processed using AR, TE, or both. When AR is applied, each anaphor is substituted by its antecedent and then the text is sent to the scoring module. The JavaRAP[6] was used to solve this task. TE is used to remove semantically redundant sentences. It in its turn can also be combined with the WSD module replacing each representative of a WordNet synset with one and the same synset member. Finally, AR can be combined together with the TE module. The AR can be applied either only for redundancy detection, i.e. the resolved text is being consumed by the TE module, but the original version of the remaining sentences sent further for scoring, or the sentences with the resolved anaphoric expressions will be submitted for scoring.

The following settings have been evaluated so far:

- **SSW**: using the standard stopword list
- **ASW**: using the standard and the extended stopword list
- **ASW ITF**: ASW combined with the ITF
- **ASW ISF**: ASW combined with the ISF
- **ASW AR**: ASW combined with the AR
- **ASW TE**: ASW combined with the TE
- **ASW WSD TE**: use ASW, replace the words of the selected parts of speech with the same member of the WordNet synset, then process the result using the

[4] Freeling `http://nlp.lsi.upc.edu/freeling/`

[5] N stands for nouns, NVA stands for nouns, verbs, adjectives – the parts of speech WSD is applied to.

[6] JavaRAP `http://aye.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html`

TE module. The scoring module is applied to the original version of remaining sentences, thus making it possible to evaluate the impact of both MFS and UKB algorithms on the resulted data.

– **ASW AR WSD TE**: is similar to the previous one with the difference that AR is applied before the WSD.The scoring module is again applied to the original version of the remaining sentences.

Each of the listed settings is further combined with either TF or CF scoring strategy as shown in Table 1.

The scoring module arranges the sentences in the descending order. The top N sentences with the common length not exceeding the size of 100 words are selected for the final summary.

4 Results and Discussion

Table 1 shows the ROUGE-1 recall values for the system settings listed in Section 3.2. The system components interact in a complex way. This way the name conventions for the columns cannot be associated exclusively with the scoring module strategies. For example, the system setting ASW WSD TE uses WSD prior to scoring, so the particular WSD algorithms used during this step is also indicated in a column name. As well as the stopwords filtering was used not only during the scoring procedure, but also before applying WSD.

Table 1. ROUGE-1 Recall values

	ROUGE-1 Recall				
	TF	**CF-MFS**		**CF-UKB**	
		NVA	**N**	**NVA**	**N**
SSW	**0.40906**	0.40869	0.41717	0.41765	0.41396
ASW	0.41779	0.40976	0.41466	0.41785	0.41462
ASW ITF	0.36924	0.36828	0.36668	0.36890	0.36719
ASW ISF	0.38126	0.37985	0.37894	0.37924	0.37562
ASW AR	0.38945	0.38873	0.39077	0.39146	0.38991
ASW TE	0.41804	0.41807	0.41596	**0.41897**	0.41665
ASW WSD TE	0.41807	0.41796	0.41627	**0.41894**	**0.41843**
ASW AR WSD TE	**0.43235***	**0.43050***	**0.42963***	**0.43196***	**0.43017***

The TF combined with the SSW was selected as the baseline. This baseline is slightly lower than the DUC 2002 baseline of 0.41132 [13]. Only some of the system settings outperformed it. All of them involved TE. As expected, the best setting is the ASW AR WSD TE – the combination of all the semantic modules used sequentially applied one after the other. The scores for this setting were evaluated using the t-test and the statistically significant values are indicated with the asterisk in Table 1. As for the scoring strategies, the TE-based and the CF-UKB applied for NVA outperform the CF-MFS strategies. The worst results were obtained for the AWS ITF, AWS ISF and AWS

AR settings. For further experiments we plan to substitute the JavaRAP by another AR module and examine the difference.

The best results for the systems mentioned in Section 1 involving TE [8] and a WSD-based graph implementation of [11] are 0.4518 and 0.4651 respectively as reported in [11]. They outperform our best result of 0.43235. However, our system improved on the best score of the LeLSA+AR system [13]. This shows that combining TE, WSD and AR with graph-based method is worth experimenting with.

5 Conclusion and Future Work

This paper presented the initial results of the ongoing research on the impact of a set of statistical and semantic features on the TS task. In particular, we analyzed the impact of textual entailment, anaphora resolution, word sense disambiguation, term frequency, inverse term and sentence frequency and corpus-tailored stopword list. The results showed that each of these features alone do not improve the performance, moreover inverse sentence and term frequencies and anaphora resolution influence it negatively and we would like to further investigate these issues. But the combination of the semantic features that include anaphora resolution, textual entailment and word sense disambiguation outperforms the baseline. In the future we plan to experiment with a different anaphora resolution system, compute the scores for all the mentioned settings using the standard stopword list and extend our system to cover other languages.

Acknowledgements. This research work has been funded by the Spanish Government through the project TEXT-MESS 2.0 (TIN2009-13391-C04) and by the Valencian Government through projects PROMETEO (PROMETEO/2009/199) and ACOMP/2011/001.

References

1. Blake, C.: A Comparison of Document, Sentence, and Term Event Spaces. In: ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 601–608. ACL, Stroudsburg (2006)
2. Chuang, W.T., Yang, J.: Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms. In: Terano, T., Chen, A.L.P. (eds.) PAKDD 2000. LNCS, vol. 1805, pp. 454–457. Springer, Heidelberg (2000)
3. Fujii, Y., Kitaoka, N., Nakagawa, S.: Automatic Extraction of Cue Phrases for Important Sentences in Lecture Speech and Automatic Lecture Speech Summarization. In: INTERSPEECH, pp. 2801–2804 (2007)
4. Ferrández, O., Micol, M., Muñoz, R., Palomar, M.: A perspective-based approach for solving textual entailment recognition. In: ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 66–71 (2007)
5. Harabagiu, S., Lacatusu, F.: Topic Themes for Multi-document Summarization. In: 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 202–209 (2005)
6. Harabagiu, S., Hickl, A., Lacatusu, F.: Satisfying Information Needs with Multi-document Summaries. Information Processing & Management 43(6), 1619–1642 (2007)

7. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: ACL Text Summarization Workshop, pp. 74–81 (2004)
8. Lloret, E., Ferrández, O., Muñoz, R., Palomar, M.: A Text Summarization Approach Under the Influence of Textual Entailment. In: 5th International Workshop on NLPCS, pp. 22–31 (2008)
9. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2), 157–165 (1958)
10. Nenkova, A., Vanderwende, L., McKeown, K.: A Compositional Context Sensitive Multi-doscument Summarizer: Exploring the Factors That Influence Summarization. In: 29th SIGIR, pp. 573–580. ACM, New York (2006)
11. Plaza, L., Díaz, A.: Using Semantic Graphs and Word Sense Disambiguation. Techniques to Improve Text Summarization. Procesamiento del Lenguaje Natural 47, 97–105 (2011)
12. Saggion, H.: A Robust and Adaptable Summarization Tool. Traitement Automatique des Languages 49(2), 103–125 (2008)
13. Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K.: Two Uses of Anaphora Resolution in Summarization. Information Processing and Management 43(6), 1663–1680 (2007)
14. Tatar, D., Tamaianu-Morita, E., Mihis, A., Lupsa, D.: Summarization by Logic Segmentation and Text Entailment. In: 33rd CICLing, pp. 15–26 (2008)