On the Effect of Stemming Algorithms on Extractive Summarization: A Case Study

Eleni Galiotou
Department of Informatics
Technological Educational Institute of Athens
GR-122 10 Athens, Greece
+30 210 5385824
egali@teiath.gr

Nikitas Karanikolas
Department of Informatics
Technological Educational Institute of Athens
GR-122 10 Athens, Greece
+30 210 5385736
nnk@teiath.gr

Christodoulos Tsoulloftas
Department of Informatics
Technological Educational Institute of Athens
GR-122 10 Athens, Greece
+30 210 5385824
cortone@gmail.com

ABSTRACT
In this paper, we discuss the efficiency of stemming algorithms and their contribution to the improvement of shallow summarization methods. We describe a benchmarking experiment on the use of two stemming algorithms and two different sets of stopwords in combination with two approaches to extractive summarization in Greek texts. The results of our experimentation show the limits of extractive summarization and stemming algorithms.

Categories and Subject Descriptors
*I.2.7 [**Artificial Intelligence**]: Natural Language Processing – text analysis.*

General Terms
Algorithms, Performance, Experimentation.

Keywords
Extractive summarization, stemming, stopwords.

1. INTRODUCTION
Different approaches to extractive summarization are in fact statistics-based methods aiming at selecting a significant subset of sentences for quick understanding. This procedure is followed by a synthesis stage in order to render the extracted sentences into a coherent set [5]. Moreover, the extraction procedure is preceded by a language dependent word normalization phase which comprises a term extraction phase, which will contribute to the selection of sentences for automatic summarization. In this paper, we describe a benchmarking experiment on the effect of stemming algorithms and stopword lists on extractive summarization on Greek texts. The paper is structured as follows: In section 2, we provide a quick overview of stemming algorithms for the Greek language and stopword lists containing words that are excluded from the processing. In section 3 we describe the shallow summarization methods that are used in our experiment and we argue for the choice of the particular approaches. Section 4 describes the benchmarking tool that combines the proposed summarization algorithms with two different stemmers for Greek and two different stopword lists. Then, in section 5 an evaluation of the alternative methods is conducted by comparing the results of the experimentation with different configurations of the system with summaries provided by three specialists in the field. Finally, in section 6, we draw conclusions and we point to future work.

PCI 2013, September 19 - 21 2013, Thessaloniki, Greece

http://dx.doi.org/10.1145/2491845.2491889

2. STEMMING AND STOPWORDS

2.1 Stemmers
Stemmers are modules used in various text processing tasks, including search engines, document/text summarizers, document/text classifiers, etc. Stemmers produce normalized forms of words in order to handle all the inflected word-forms of a word existing in documents as one attribute of the document collection. Alternatively, for the same purpose, one can use lemmatizers that conflate a set of words in their etymological root. However, lemmatizers demand extended resources, which are not always available. Stemmers are algorithmic approaches, using minimal resources, that elaborate suffix replacement (or suffix removal) and conflate all the inflected forms of words in a single stem. Stemmers rarely result into linguistically sound forms but, for the mentioned applications, this is not considered to be a problem.

2.2 Stemming Algorithms for Greek
The available (published) stemmers for the Greek language are Rule-Based (vs Dictionary-Based and Corpus-Based) stemmers. Two well known stemmers are proposed by Kalamboukis [6] and Ntais [13]. Both systems elaborate a set of rules that remove (or replace) suffixes under certain conditions. However, they have different levels of granularity for the application of the rules. The Ntais stemmer uses a set of 158 suffixes grouped into 22 rules and applies the suitable rule at a single level. The Kalamboukis stemmer removes totally 65 suffixes in a two level approach. On the first level, the suitable inflectional suffix is removed. On the second level the suitable derivational suffix is removed. Derivational suffixes are classified into nominal, adjectival and verbal ones.

In addition to the above mentioned stemmers, the NNK stemmer was developed by N. Karanikolas and was incorporated both into commercial [11] and research [7, 9, 10] applications. The stemmer in question uses 200 suffixes and performs suffix stripping (or replacement) in six steps (levels). In addition, the stemmer in question, performs a light part-of-speech (POS) tagging. This is accomplished using the suffixes on the second level. These suffixes can specify exactly one POS tag or can decrease the number of suitable POS tags, in other words resolve or reduce the ambiguity produced by the POS tagging.

This more granular (and more sophisticated) approach is further exploited on the third level where suffixes are applied for removal (or replacement) and also for determining verb characteristics.

2.3 Stopword Lists

Stopwords are words that appear very frequently and/or act as a glue for forming tuneful (sonorous) sentences but they don't have any contribution to the meaning of a document (text). Stopwords are usually pronouns, prepositions, etc. Consequently, they do not contribute to the discrimination between documents. Removing stopwords decreases the indexing overhead without any loss in the accuracy of a text retrieval (or similar systems - like summarization, classification, etc). In most cases stopwords are list of trivial words, in their inflected form, which are removed during some preprocessing stage. This is also the case for the CELEX list of stopwords [1] that is applied at a preprocessing stage before the Ntais stemmer. In other cases, there is more than one level for stopwords removal. For example, the NNK stemmer elaborates two lists of stopwords. The first list of stopwords contains inflected forms of trivial words, which are removed in a very early stage of preprocessing. The second list contains stemmed stopwords (most of them are trivial verbs in their conflated form) which are removed after the stemming process. Obvioiusly, the NNK stemmer has a more granular approach even for stopword removal.

3. EXTRACTIVE SUMMARIZATION

Our experimentation was based on the following approaches to extractive summarization:

3.1 Sentence Weighting

*3.1.1 The TF*IDF factor*

The well-known TF*IDF factor which is used in IR applications in order to measure similarity between a document and a query, can also be used in order to extract sentences for summarization. [12]. In this case the the weight of a sentence is given by the following equation:

$$T(S_{ik}) = \sum_j t_{ij} \cdot q_j \qquad (1)$$

$T(S_{ik})$ is the weight of the k^{th} sentence in document D_i, for each term j existing in the k^{th} sentence of document D_i t_{ij} is given by the equation (2), and q_j is given by the equation (3).

$$t_{ij} = \frac{F_{ij}}{\sum F_i} \quad (2)$$

t_{ij} is the weight of term j in document D_i, F_{ij} is the frequency of term j in document D_i, and ΣF_i is the sum of frequencies of the index terms existing in document D_i.

$$q_j = \log_{10}\left(\frac{N}{DocFreq_j}\right) = -\log_{10}\left(\frac{DocFreq_j}{N}\right) \quad (3)$$

q_j is the weight of term j in the collection, N is the number of documents existing in the collection and $DocFreq_j$ is the number of documents where the term j occurs.

The use of sentence weighting based on the TF*IDF factor aims ta extracting sentences for summarization and at the same time minimizing the number of words in common across the summaries of all the texts in the collection.

*3.1.2 The TF*ISF Factor*

An alternative method for sentence extraction for summarization takes into account the TF*ISF factor which is given by the following equation:

$$T'(S_{ik}) = \sum_j t_{ij} \cdot isf_{ij} \quad (4)$$

where t_{ij} is given by the equation (2) and isf_{ij} is given by the equation (5):

$$isf_{ij} = \log_{10}(ns_i / ns_{ij}) \quad (5)$$

where isf_{ij} is the inverse sentence frequency of term j in document D_i, ns_i is the number of sentences in document D_i, ns_{ij} is the number of sentences of document D_i that contain the term j. Note that ISF does not use statistical measurements from the collection of documents.

The use of the the TF*ISF factor for summarization aims at extracting sentences which contain a large number of the most frequent terms of the document [2].

3.2 Sentence Weighting

In order to improve the coherence of summaries produced by sentence weighting methods, the location of sentences in a document must be taken into account. A promising approach to that matter is the News Article Algorithm [4] which aims at assigning a different weight to each sentence of the text according to the position of the sentence in the document and in the host paragraph. In order to assign a weight to a sentence, the algorithm uses equation (6):

$$((SP - P + 1) / SP) * ((SIP - SPIP + 1) / SIP) \quad (6)$$

where SP is the number of paragraphs in the document, P is the serial number of the paragraph under investigation, SIP is the number of sentences in the paragraph under investigation and $SPIP$ is the sentence position inside the paragraph.

3.3 Title Method

Another approach to term weighting assignment which is incorporated in our system is an adaptation of the Title Method which was proposed by Edmundson [3]. The method in question relies on the assumption that title words circumscribe the subject matter of a document. Therefore, positive weights are assigned to words that appear in the title and subheadings of a document and the "title weight" of a sentence is the sum of the "title weights" of its constituent words. In our system, a predefined constant is assigned to words of a sentence which also occur in the title. Therefore, the "title weight" of a sentence is the product of of the predefined constant multiplied by the number of title words in the sentence

During a previous benchmarking experiment on extractive summarization methods [8], the combination of the News Articles position weighting method with each of the two term weighting methods (TF*ISF and TF*IDF) gave the most promising results. As a consequence, we decided to use the two

combinations in our experimentation with stemming algorithms which is described in the following section:

4. THE EXPERIMENT

We have implemented a software tool for benchmarking of extractive summarization methods. The tool was also used in order to draw conclusions on the efficiency of the methods and their combination with different stemming algorithms for the Greek language. The system was designed so as to incorporate a combination the 3 categories of factors described in section 3 as it is given in the following equation:

$$w1 * ST + w2 * SL + w3 * TT \quad (7)$$

where ST is the sentence weighting based on terms, SL is the sentence location factor, and TT is the title terms factor. ST is computed using one of the equations (6), (8) or (10). SL is computed using the equation (11) and TT is computed as described in section 2.3 (Title words / Keywords). The weights w1, w2 and w3 are user defined with default values of 1.0 for each one.

Moreover, the system provides the user with the possibility to define his own "Compress rate" i.e. the percentage of most relevant sentences according to equation (7). The user can also define a "words per sentence threshold", i.e. the range of the number of words in a sentence, in order for it to be considered for participation in the summary [8].

In order to evaluate the influence of stemming algorithms and the efficiency of different stopword lists on the performance of summarizing methods the system was enriched with the NNK stemmer and the corresponding stopword list. Thus, the system provides the user with 3 possible configurations according to the stemming algorithms described in section 2:

[1] The Ntais stemmer and the CELEX stopword list (default option)

[2] The Ntais stemmer and the NNK stopword list

[3] The NNK stemmer and the NNK stopword list

A snapshot of our benchmarking system is provided in figure 1.

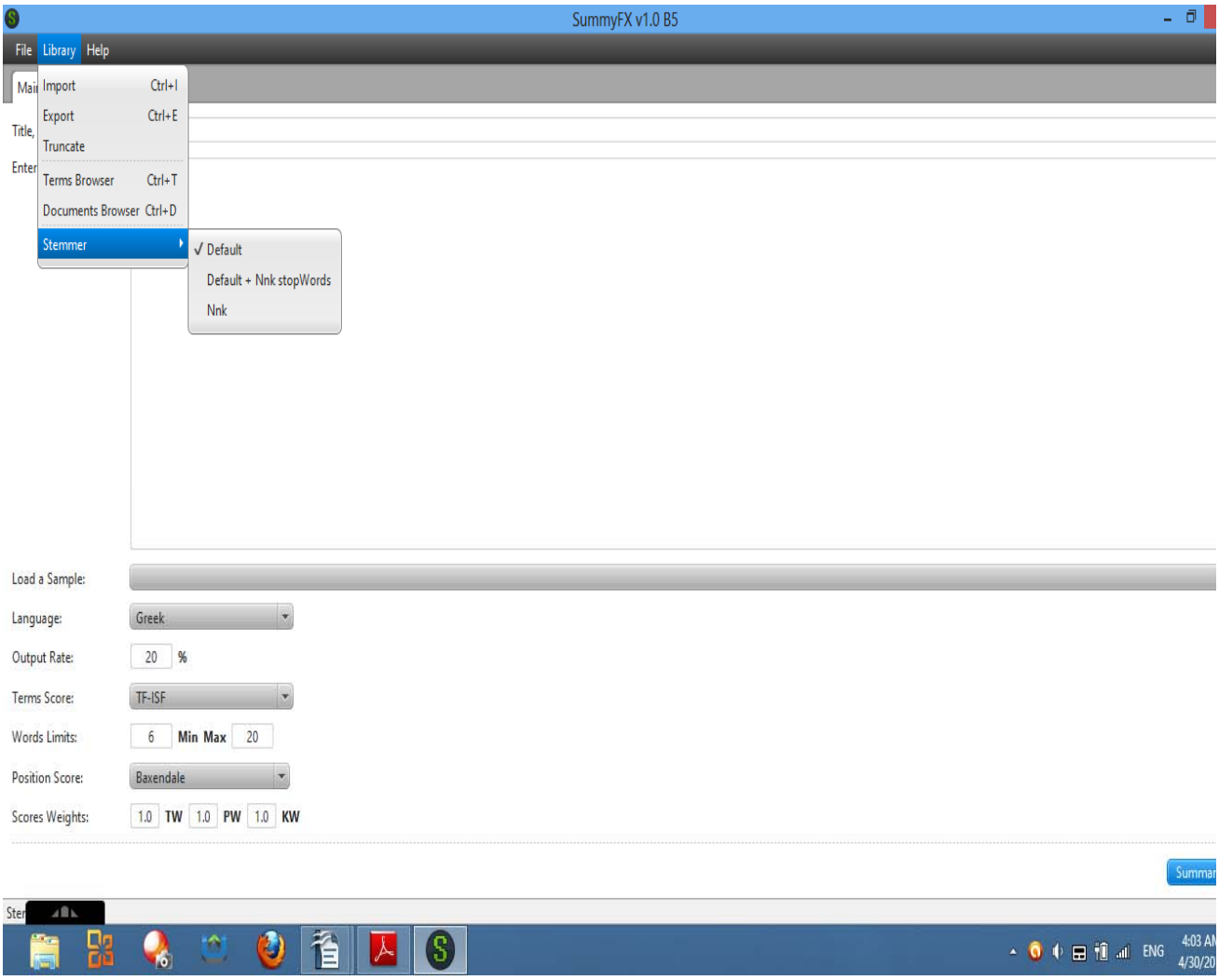

Figure 1: The summarization benchmarking software

We have experimented with 15 articles of Greek newspapers (**http://www.greek-language.gr**). We used a *Compress rate* of 20% (extraction of the 20% of the sentences) and the *Words per sentence thresholds* ranging from 2 to 25.

Thus, almost every sentence was candidate to summarization. Moreover, the *w1, w2,* and *w3* parameters of equation (7) were all assigned a value of 1.0. Thus, the *sentence weighting*, the *sentence location* and the *title factor* (*Sentence Score Weights*) were taken to equally contribute to the overall performance of the system. The above mentioned parameters were combined with the two combinations of term weighting methods (TF*ISF, IF*IDF) with the *News Articles* position weighting method. In this way, we experimented with 6 alternative system configurations.

5. EVALUATION

The results of our experimentation were evaluated against summaries provided by human summarizers. The 15 documents of the experiment were distributed to 3 native speakers of Greek. (2 short texts, 2 texts of medium length, 1 long text to each person). Then, we asked them to select the 20% more representative sentences of each text. Moreover, we asked them to select the next 20% most promising sentences to be included in the summary of the text.

Then, we counted the sentences in common between the summary produced by the system and the 20% top sentences selected by the user for each configuration of the system. Then, the procedure was repeated for the results of the system and the extended selection (20% top + 20% next more promising sentences) of the user.

Table 1 tabulates the results of the default configuration of the system (Usage of the Ntais stemmer and the CELEX stopword list, the other parameters as described in section 5).

Table 1. Sentences in common between expert's selection and system's summary – usage of Ntais stemmer & CELEX stopword list

Default Stemmer		Sentences in common between expert's selection (20%) and system's summary (20%)		Sentences in common between expert's extended selection (40%) and system's summary (20%)	
Doc Id	#sent	TF-ISF & News Articles	TF-IDF & News Articles	TF-ISF & News Articles	TF-IDF & News Articles
3388	9	2	2	3	3
3385	10	1	1	1	2
3384	37	2	2	3	3
3343	46	2	2	3	3
3354	97	2	3	5	5
3381	13	1	1	1	1
3382	12	2	2	3	3
3339	49	4	4	9	9
3338	48	3	3	8	9
3320	64	6	5	9	11
3387	7	1	1	1	1
3386	6	1	1	1	1
3372	33	5	5	4	5
3358	37	3	4	4	4
3353	80	5	5	8	8
		40	**41**	**63**	**68**

The second experiment consisted in using the NNK stopword list and keeping the same stemmer and the other system parameters. The results are tabulated in Table 2.

Table 2. Sentences in common between expert's selection and system's summary – usage of Ntais stemmer & NNK stopword list

Ntais stemmer & NNK stopword list		Sentences in common between expert's selection (20%) and system's summary (20%)		Sentences in common between expert's extended selection (40%) and system's summary (20%)	
Doc Id	#sent	TF-ISF & News Articles	TF-IDF & News Articles	TF-ISF & News Articles	TF-IDF & News Articles
3388	9	2	2	3	3
3385	10	1	1	1	2
3384	37	2	2	3	3
3343	46	2	2	4	4
3354	97	3	3	7	6
3381	13	1	1	1	1
3382	12	3	3	3	2
3339	49	5	5	8	8
3338	48	5	4	8	8
3320	64	5	6	11	11
3387	7	1	1	1	1
3386	6	1	1	1	1
3372	33	4	4	5	5
3358	37	3	3	4	5
3353	80	5	5	8	8
		43	**42**	**68**	**68**

The performance of the system between the first and second experiment was improved when compared to the extended user selection (20% top + 20% next most promising sentences).

Then the third experiment was conducted where the NNK stemmer and stopword list were used. The stemmer in question when applied to the same subset of documents, decreased to 2/3 the number of produced stems (conflate more word forms with the same meaning) without fault conflations. Thus, a significant increase in the performance of the system was anticipated. The results of the experiment are tabulated in Table 3.

The performance of the system when compared to the extended user selection (20% top + 20% next most promising sentences) was further improved with the use of the NNK stemmer.

Table 3. Sentences in common between expert's selection and system's summary – usage of NNK stemmer and stopword list

NNK Stemmer & Stopword List		Sentences in common between expert's selection (20%) and system's summary (20%)		Sentences in common between expert's extended selection (40%) and system's summary (20%)	
Doc Id	#sent	TF-ISF & News Articles	TF-IDF & News Articles	TF-ISF & News Articles	TF-IDF & News Articles
3388	9	2	2	3	3
3385	10	1	1	1	1
3384	37	2	2	3	3
3343	46	2	2	3	3
3354	97	3	4	6	6
3381	13	2	1	2	1
3382	12	3	3	3	3
3339	49	5	6	8	9
3338	48	4	4	9	10
3320	64	5	5	10	10
3387	7	1	1	1	1
3386	6	1	1	1	1
3372	33	5	5	6	5
3358	37	2	2	6	6
3353	80	5	3	8	6
		43	**42**	**70**	**68**

6. CONCLUSIONS & FUTURE WORK

We have described an attempt to evaluate the influence of stemming algorithms on the performance of shallow summarization methods. Our approach consisted in combining two promising combinations for extractive summarization (TF*ISF with News Articles) and (TF*IDF with News Articles) with two stemmers ad stopword lists for Greek.

The first stemmer, proposed by Ntais [13], uses a set of 158 suffixes incorporated in 22 rules and performs suffix stripping at a single level. The second one, proposed by Karanikolas [7] uses some 200 suffixes and performs suffix stripping or replacement in 6 steps (levels). The suffixes of the 2nd level are also used in order to perform a rudimentary POS tagging. In addition, we took into account two different stopword lists: the stopword list of the Greek version of the CELEX database and the stopword list proposed by Karanikolas which was extracted from the Triandafyllidis Grammar [14]. The use of the Karanikolas list in combination with the Ntais stemmer resulted in a slight improvement of the system performance. The performance of the system was further improved by the application of the NNK stemmer and stopword list during the 3rd phase of our experimentation. Nevertheless, the difference in the performance of the two stemmers during the preprocessing phase had led us to anticipate more significant improvement in the overall performance of the system.

We claim that for texts written in morphologically rich languages like Greek, more sophisticated tools like morphological analyzers are needed in order to obtain better results. In the near future we plan to conduct similar experiments where the stemmer will be replaced by a morphological analyzer that takes into account the particularities of the Greek language.

7. ACKNOWLEDGMENTS

This research has been co-funded by the European Union (Social Fund) and Greek national resources under the framework of the "Archimedes III: Funding of Research Groups in TEI of Athens" project of the "Education & Lifelong Learning" Operational Programme.

8. REFERENCES

[1] Bagola, H. 2004. *Informations utiles à l'intégration de nouvelles langues européennes*, (DIR/A-Cellule «Méthodes et développements», section «Formats et systèmes documentaires»), Technical Report. http://publications.europa.eu/pdf/fr/elarg-vl.pdf

[2] Chong, B. and Chen, Y.Y.. 2009. Text summarization for oil and gas news articles In *World Academy of Science, Engineering and Technology* , vol. 53

[3] Edmundson, H.P., 1969. New methods for automatic extracting. *J. ACM.* 16, 2, 264-285.

[4] Hariharan, S. 2010. Multi-document summarization by combinational approach, *Int. J. Comp. Cog*., 8, 4

[5] Hovy, E. and Lin C-Y. 1999. Automated text summarization in SUMMARIST, In *Advances in automatic text summarization*, I. Mani and M. Maybury, Eds.

[6] Kalamboukis T. Z. and Nikolaidis S. 1995. Suffix stripping with Modern Greek, *Program* 29, 313-321.

[7] Karanikolas, N. 2007. Low cost, cross-language and cross-platform Information Retrieval and Documentation tools. *J. Comp. Inf. Tech. (CIT)*, ISSN: 1330-1136, 15, 107.

[8] Karanikolas, N. Galiotou, E. and Tsoulloftas, C. 2012. A Workbench for Extractive Summarization Methods, In *Proceedings PCI'2012*, (Athens, Greece), IEEE, 454-458

[9] Karanikolas N. and Skourlas C. 2010. A parametric methodology for text classification. *J. Inf. Sci*., 34, 6, 421-442, 2010, doi:10.1177/0165551510368620

[10] Karanikolas, N. Vassilakopoulos, M. and Giannoulis, N. 2012. *A Software Tool for Building a Statistical Prefix Processor*. *BCI 2012: 5th Balkan Conference in Informatics*, (Novi Sad, Serbia, September 16-20, 2012), ACM Digital Library

[11] Moumouris, N. 1995. "The 'Erevinitis' Document Retrieval System", Greek CHIP, 11 (March 1995), 60-61

[12] Murray, G. and Renals, S. 2007. Term-Weighting for Summarization of Multi-Party Spoken Dialogues, *Proc. of MLMI 2007*

[13] Ntais, G. 2006. *Development of a Stemmer for the Greek Language*, Master Thesis, University of Stockholm

[14] Triandafyllidis, M. (1991/1941). *Modern Greek Grammar* (3rd revised ed.). Thessaloniki: Manolis Triandafyllidis Foundation (in Greek)