

A study of chunk-based and keyword-based approaches for generating headlines*

Enrique Alfonseca¹, José María Guirao², and Antonio Moreno-Sandoval³

¹ Computer Science Dep., Universidad Autonoma de Madrid, 28049 Madrid, Spain
Enrique.Alfonseca@ii.uam.es

² Department of Computer Science, Universidad de Granada, 18071 Granada, Spain
jmguirao@ugr.es

³ Department of Linguistics, Universidad Autonoma de Madrid, 28049 Madrid, Spain
sandoval@maria.111f.uam.es

Abstract. This paper describes two procedures for generating very short summaries for documents from the DUC-2003 competition: a chunk extraction method based on syntactic dependences, and a simple keyword-based extraction. We explore different techniques for extraction and weighting chunks from the texts, and we draw conclusions on the evaluation metric used and the kind of features that are more useful. Two preliminary versions of this procedure ranked in the 12th and 13th positions with respect to unigram recall (ROUGE-1) at DUC-2004 (out of 39 runs submitted).

1 Introduction

Headline generation is the problem of generating a very short summary of a document, which condenses the main ideas discussed in it. This paper describes two different procedures tested on the collection provided for the 2003 Document Understanding Conference (DUC-2003), and a hybrid approach that combines them. In all the experiments described here, we can identify two separate steps: firstly, an identification and extraction of the most important sentences from the document. Secondly, the extraction of relevant keywords and phrases from those sentences. The purpose of the first step is to restrict as much as possible the search space for the second step, thereby simplifying the selection of fragments for the headline. We have evaluated the procedures using the ROUGE-1 score [1], and we also explore some of the characteristics of this metric. The words *summary* and *headline* will be used indistinctly throughout this paper.

A popular approach for generating headlines consists in first identifying the most relevant sentences, and then applying a compaction procedure. Sentence-extraction procedures are already well-studied [2], so we shall focus on the differences in the compaction step. Some of the techniques are (a) deletion of all subordinate clauses; (b) deletion of stopwords (determiners, auxiliary verbs, etc.) [3,4,5]; (c) extracting fragments from the sentences using syntactic information, e.g. the verb and some kind of arguments, such as subject, objects or negative particles [6,7,8,9]; and (d) using pre-defined templates [10]. A different approach consists in extracting, from the document, a list

* This work has been sponsored by CICYT, project number TIC2001-0685-C02-01.

of topical keywords, collocations, noun phrases [3,11,12,13]. Using this procedure, the resulting headline will not be grammatical, but it may provide a useful description of the topic of the article.

In Section 2 we describe the experimental settings for evaluating the system, and Section 3 briefly summarises the general architecture of the system. Next, Section 4 describes the procedures for sentence selection, and Sections 5, 6 and 7 describe all the experiments performed for generating headlines. Finally, Section 8 describes the conclusions we can draw from the results obtained, and discusses possible lines for future work.

2 Experimental settings

The purpose of the work is to generate very short headlines from documents. We can describe this task using Mani’s classification of automatic summarisation systems [2], which takes into account the following characteristics: **compression rate** is typically very high (a few words or bytes); the **audience** is generic, as the headlines do not depend on the user; the **function** is indicative, as it must suggest the contents of the original document without giving away details; and they should be coherent, and generated from single documents. In the experiments, the genre used is newswire articles, written in a single language (English).

All the experiments have been tested on the data provided for task 1 in DUC-2003. It is a set of 624 documents, grouped in sixty collections about some topics, such as schizophrenia or floodings in the Yangtze river. For each of the documents, NIST has provided four hand-written summaries to be used as gold standard. Throughout this work, we use a 75-byte limit, but we apply it in a lenient way: if a word or a chunk selected for a summary exceeds the limit, it will not be truncated.

2.1 ROUGE as evaluation metric

ROUGE [1,14] is a method to automatically rank summaries by comparing them to other summaries written by humans. The original idea for the ROUGE-N metric is basically an n-gram recall metric, which calculates the percentage of n-grams from the reference summaries appear in the candidate summary:

$$\frac{\sum_{S \in Refs} \sum_{gram_n \in S} |\{gram_n : gram_n \in Cand\}|}{\sum_{S \in Refs} \sum_{gram_n \in S} |\{gram_n\}|}$$

Note that if an n-gram appears in several references at the same time, it is counted as many times, which makes sense because an n-gram for which there is consensus between the humans should receive a higher weight. The procedure has been extended with additional calculations in order to improve its accuracy [14].

Lin and Hovy’s experiments [1] indicate that ROUGE-1 correlates well with human evaluations of automatic headlines. In fact, given the availability of four hand-written summaries for each document, ROUGE has been used for evaluating the summaries produced by the participant teams in DUC-2004. Therefore, we have chosen to evaluate our system with the ROUGE-1 metric.

		Document 1								Genotype						Score
Words		w_1	w_2	w_3	w_4	w_5	w_6	Words		w_1	w_2	w_3	w_4	w_5	w_6	
Model 1		1	0	1	1	0	0	Candidate 1		1	0	1	1	0	1	0.8571
Model 2		0	0	1	1	1	1	Candidate 2		1	0	1	0	0	1	0.5714
Model 3		0	0	1	1	0	0	Candidate 3		1	0	0	0	0	1	0.3571
Model 4		2	1	0	1	0	1	Candidate 4		0	0	0	0	1	0	0.0714
Frequency		3	1	3	4	1	2	Candidate 5		1	1	1	1	1	1	0

Table 1. Example of the procedure for finding the combination of words for which the unigram recall is maximised.

2.2 Upper bound of ROUGE-1 score in 75-bytes summaries

Before using ROUGE-1 to evaluate the summaries, it would be interesting to discover which is the range of scores that a system can obtain in this particular task. ROUGE-1 has been used to rank existing summarisation systems in the DUC-2004 competition. Although we know the score obtained by human summarizers, between 0.25017 and 0.31478 in DUC-2004 for 75-byte summaries, to our knowledge, we are not aware of the highest score that can possibly be obtained with this score.

In a first experiment, we study which is the range of values that can be obtained using ROUGE-1 when comparing a candidate summary to four manual headlines.

We shall use, as in DUC, four reference summaries for each documents. When evaluating a candidate headline, ROUGE-1 can be considered as the unigram recall of the candidate. If the candidate and all the references have the same length, it is obvious that, unless all the references have the same words, a candidate summary will never contain every word from every reference (which would mean a recall of 1).

The experiment for discovering the highest possible score has been designed in the following way:

- A. For each document,
 1. Take its four hand-written headlines.
 2. Collect all the words that appear in them, excluding closed-class words.
 3. Count the frequency of each word.
 4. Look for the combination of words that maximises the ROUGE-1 score and has less than 75 bytes altogether.
- B. Calculate the average of this score for all the documents.

Step 4 is the most costly step. A good approximation can be obtained by choosing the words with the highest frequencies in the model summaries. Still, that does not guarantee that the obtained summary will be the best one, as it may be better to substitute a long word with a large frequency for two short words which altogether have a higher frequency. A brute force approach would require too much computational time, and therefore we opted for a genetic algorithm to find the combination of words that maximises the unigram recall. Table 1 illustrates how the search is performed:

- The upper part of the table represents the reference headlines (the models) for a couple of documents, and the frequency of each word in each model. For instance, the fourth model contains w_1 twice, and w_2 , w_4 and w_6 once.

find babies may be more schizophrenia possibly brain development dopamine
cesarean babies be more schizophrenia possibly brain development dopamine

1. Researchers find Cesarean babies may be more susceptible to schizophrenia.
2. Natural childbirth possibly instrumental in brain development. Cesareans associated with schizophrenia
3. Canadian rat research links caesarean birth with schizophrenic dopamine reactions.
4. Cesarean, babies, susceptible, schizophrenia, Boksa, El-Khodor, dopamine, amphetamines, brain, development

Fig. 1. Two of the best scoring summaries for a document from collection d100 (DUC-2003 data), and the four gold-standard headlines.

- The next line, labelled *Frequency*, contains the sum of frequencies of each word in all the models, for each of the documents.
- We encode a candidate summary, shown below in the table, as a boolean vector of a length equal to the total number of words in all the models.
- The fitness function for the genetic algorithm is 0 if any summary has more than 75 bytes (e.g. Candidate 5), and the ROUGE-1 metric otherwise (all the others).
- The genetic algorithm evolves a population of boolean vectors, using the mutation and crossover operators, until for a large number of generations there is no improvement in the fitness of the population.

The summaries obtained with this procedure are simply a list of keywords. Figure 1 shows a couple of keyword choices that produce the best ROUGE-1 scores for a document in collection d100, and the four gold-standards used.

The best choice of keywords for the 624 documents in the data set has produced a mean ROUGE-1 score of 0.48735. Therefore, we may take it as the upper bound that can be obtained using this evaluation procedure with this data collection. This result is consistent to the evaluation done in DUC-2004, where the test set is very similar: newswire articles and four manual summaries for each one. In this case, all the human-made models have received a ROUGE-1 score between 0.25017 and 0.31478, which represents nearly 65% of the upper bound. Constraints such as grammaticality and the fact that the same idea can be expressed in many ways probably make it difficult to reach a higher score.

3 Our approach

Our system has been divided into two main steps:

1. Firstly, we select a few sentences from the document, so that there is much less information from where to extract the headline.
2. Secondly, from those sentences, we extract and rank either keywords or phrases.
3. The headline is finally built by concatenating the keywords or chunks extracted in the previous step, until we reach the length limit. As said before, if the last keyword exceeds the limit, we do not truncate the summaries.

The following three sections further elaborate these steps.

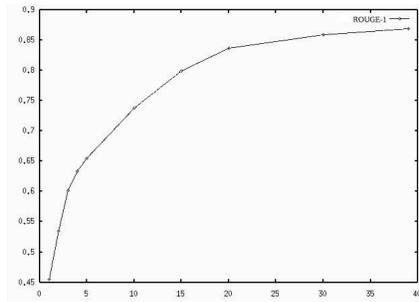


Fig. 2. ROUGE-1 results for a different number of sentences selected.

4 Selection of sentences

The first step of our system is the selection of the most relevant sentences. This is done to reduce the search space for finding the best chunks of texts with which to construct the headline. The sentence-extraction procedure we use is based on the Edmundsonian paradigm: as a linear combination of the value of several features. The features used are, among others, the sentence length, the position of the sentence in the paragraph, the position of the paragraph in the document, and the word overlapping between the sentences selected for the summary. Although some related work indicates that just by choosing the first sentences from the document can be equally useful for headline extraction [9,15], we have opted to continue using this paradigm.

In previous work we described a procedure for acquiring the weights for the linear combination [7,16]. It uses genetic algorithms, in a way which is very similar to the procedure used in the previous section: for each possible set of weights, we calculate the summary produced from those weights and evaluate it against the model headlines. The unigram recall of the summary is the fitness function of the set of weights. Finally, we keep the weights which select the summaries that maximise the unigram recall. The use of genetic algorithms for summarisation had been used previously by Jaoua and Ben Hamadou [17].

The hypothesis that is the basis for every sentence-extraction procedure is that there are a few sentences which hold the most relevant information, and a large number of sentences which elaborate those main ideas. Figure 2 shows the ROUGE-1 score of a summary in function of the number of sentences selected. As can be seen, with just the top three sentences, the ROUGE-1 score reaches around 0.60, using the same reference headlines as in the previous experiment. From that point onward, the slope of the curve slows down. The maximum score attained is around 0.87, when the complete documents are selected.

This indicates that just a few sentences have enough information for generating the summaries. The next step will be to reduce those sentences to no more than 75 bytes, trying to keep the ROUGE-1 score as near 0.35 as possible.

Sentences	
Portugal and Indonesia are mulling a new proposal to settle their dispute over East Timor, and talks between the two sides are at a crucial stage, according to a U.N. envoy.	
Envoy Jamsheed Marker said late Thursday that the U.N. proposal envisages broad autonomy for the disputed Southeast Asian territory.	
We've reached a very important, and I might even say critical, moment in the talks," Marker told reporters after a dinner with President Jorge Sampaio.	
Verb phrases	Filter
[Portugal and Indonesia] are mulling [a new proposal]	none.
to settle [their dispute over East Timor]	none.
are	1
[Envoy Jamsheed Marker] said [autonomy]	2
[We] 've reached [a very important]	3
[I] might even say	2
[Marker] told [reporters]	2

Table 2. Sentences selected from document APW19981106.0542, and verb phrases obtained from them.

5 Chunk-based headline extraction

Most newswire articles describe events, which are usually (but not always) expressed with verbs. Therefore, we thought that a good idea for generating the headline was to select the most relevant verbs from the selected sentences, together with their arguments. The process is divided in three steps: verb extraction, verb-phrase ranking and headline generation.

To this aim, we process all the documents using a syntax analyser. The parser used is the *wraetic* tools [18]⁴. These include a Java PoS tagger based on TnT [19], a stemmer based on the LaSIE morphological analyser, [20], three chunkers written in C++ and Java (with ~94.5% accuracy when evaluated on the WSJ corpus), and a subject-verb and verb-object detector, written in Java *ad hoc* with hand-crafted rules. Multiword expressions have also been identified automatically with the following procedure: (a) eliminate stop-words and verbs from the text; (b) Collect bigrams and trigrams whose frequency is above a threshold (e.g. three times); (c) put again stopwords where necessary (e.g. in "*President of the United States*"). All the experiments reported in this section were evaluated on the whole DUC-2003 corpus for Task 1 (headline generation).

5.1 Verb-phrase extraction

Verbs are extracted in the following way: using our PoS tagger, we first obtain all the verbs from the document. With the partial parser, we markup each verb with its subject and arguments. Table 2 shows the sentences obtained from a document, and the verb phrases extracted from them. Note that the parser is not perfect: sometimes it cannot identify the arguments of a verb. Some errors are due to a poor grammar coverage, and others are due to mistakes in the PoS tagging. However, in many cases the arguments found are correct.

⁴ Available at www.ii.uam.es/~ealfon/eng/download.html

No. of sentences	All sentences		All VPs		Kept VPs	
	Mean length	ROUGE-1	Mean length	ROUGE-1	Mean length	ROUGE-1
1	198	0.44155	113	0.24635	73	0.20851
2	364	0.53025	194	0.30725	102	0.23756
3	527	0.59137	277	0.35974	129	0.26616
4	723	0.63480	333	0.38579	145	0.28082

Table 3. ROUGE-1 score obtained (a) by selecting sentences from the document; (b) by extracting the verb phrases (with their arguments) from them; and (c) after filtering.

Filtering heuristics A manual revision of the verb phrases showed that many of them contained information that most probably was not relevant enough as to be included in the headline. Some of these cases include the following:

1. If the parser has not been able to identify any argument of the verb phrase.
2. If the verb, in WordNet, is a hyponym of *communicate*, *utter*, or *judge*, because in many cases the information that was communicated is more relevant than the name of the person who stated it.
3. If the subject is in first or second person.
4. If the verb is either in passive form or a past participle, and the parser did not find its subject nor its agent. This is because in most of these cases the verb is functioning as an adjective and it was wrongly tagged by the PoS tagger.

The right column of Table 2 shows, for each verb, either the number of the filter because of which it was ruled out, or *none*, in the case that it passed all filters. As can be seen, only two verb phrases from this document have passed the set of filters.

Information kept in the VPs By extracting the verbs, the sentences selected in the previous step are reduced to a small set of verbs and their arguments. After applying the filters, this set is reduced even further, as in the example above. In Figure 2 we studied the amount of information, expressed with the ROUGE-1 score, that we still kept by selecting just two or three sentences from a document. We can do the same experiment now to see how large is the decrease in the ROUGE-1 score if we substitute the selected sentences with the list of verb phrases, and if we substitute this list with just the verb phrases that have passed all the filters.

The results are shown in Table 3. The first column shows the number of sentences that have been selected from the original document. The second column shows the ROUGE-1 score if the summary contains the complete sentences selected. These are the same values used for plotting Figure 2. The third column shows the ROUGE-1 score if we score not the complete sentences, but the list of verbs and their arguments, as printed in Table 2. Finally, the fourth column shows the score if we list the verb phrases after applying all the filters. It can be seen that the score decreases in the last two columns. However, the decrease is not proportional to the compaction level performed in each of these steps. Therefore, we know that we are removing mostly words that do not appear in the reference summaries.

5.2 Verb-phrase ranking

We have now extracted a list of verbs from the selected sentences. In order to generate the headline, we would like to rank them according to some metric of relevance. Lin and Hovy [21] describe how topic signatures can be built automatically from collections, and shows an application to text summarisation. We have followed the same approach, but we have calculated the topic signatures both for collections and for single documents. The procedure is the following:

1. Collect all the words and their frequencies for each document or collection.
2. Collect all the words from the rest of the documents or collections. Consider this as the contrast set.
3. Calculate the weight of each of the words using a weight function.

There are several parameters whose values we can vary. There are many weight functions that we can apply. We have tried with the *likelihood ratio* [22], which was the one used by Lin and Hovy [21]; and the tf.idf, χ^2 , Mutual Information and t-score metrics. Furthermore, as indicated above, the signatures may be calculated either for each document separately, or for each collection.

With this procedure, for all the words in each document or collection, we can calculate their weight, which is a measure of the relevance of that word in the scope of the document or collection. The verb phrases that we had extracted and filtered can be weighted using the values from the topic signatures: each verb phrase may receive as weight the sum of the weights of all the words in the verb and its arguments.

5.3 Headline generation and results

To generate the headline, while the summary has a length lower than 75 bytes, we add the next verb phrase with the highest score. To keep the grammaticality, we do not truncate the summaries if they exceed the limit in a few bytes. In the example above, only two verb phrases remain after the filtering:

[Portugal and Indonesia] are mulling [a new proposal]
to settle [their dispute over East Timor]

These will be weighted and ranked next using a topic signature. The headline will be generated in the following way:

1. Firstly, the system chooses the most weighty verb phrases until their total length limit exceeds 75 bytes. Note that, if the limit is exceeded by a few bytes, we do not truncate the summaries, so as to keep them grammatical. In this example, both VPs will be selected.
2. Secondly, they will be put together in the order in which they appeared in the original document. If there was any conjunction linking them, it will be added to the summary so as to improve the readability.

The resulting summary in the example will be:

Portugal and Indonesia are mulling a new proposal to settle their dispute over East Timor.

(a)			(b)	
Function	Docs	Cols.	No. of sentences	ROUGE-1 score
Likelihood ratio	0.18298	0.19726	1	0.19363
tf.idf	0.18910	0.19231	2	0.19956
χ^2	0.19753	0.20105	3	0.20105
Mutual Information	0.18458	0.17933	4	0.19972
t-score	0.18599	0.19011		

Table 4. (a) Effect of using a different weight function for calculating the topic signatures in the final ROUGE-1 score. (b) Effect of selecting a different number of sentences for each document or for each collection in the final ROUGE-1 score.

We have evaluated the ten different configurations (choice of weight function and topic signatures for either documents and collections) using the ROUGE-1 score. We saw, at the beginning, that by choosing just three sentences from the original document we could reach a very high ROUGE-1 score, and the slope of the curve in Figure 2 slows down from that point onward, this experiment was done using three sentences from each document.

Table 4(a) shows the results. The best score is the one obtained with the χ^2 function and for collections; while the likelihood ratio for collections has attained the second best score (not statistically significant). Apart from tf.idf, the other weight functions have lower results, statistically significant at 0.95 confidence.

It can be observed here that the likelihood ratio needs a larger set of examples, because it is the one that scored worse if the signatures are obtained for single documents, but reaches the second place if we consider documents. Furthermore, when we calculate the signatures for each collection the results are slightly better than when we select the words that are more representative just for each document.

In order to check whether the choice of selecting three sentences at the beginning of the process was correct, we tried yet another experiment. We may think that the more sentences selected, the more verb phrases we have for generating the headline. On the other hand, if we have too many verb phrases, it will be more difficult to identify those that are more informative. Table 4(b) shows the results obtained by selecting a different number of sentences from each document in the sentence extraction step. This result also suggests that three sentences is a good choice, although the difference is not statistically significant.

Finally, Figure 3 shows the headlines obtained for the documents in the collection about East Timor. It can be seen that most of them are grammatical and easy to read, although the context of the news does not appear in the headlines, so it is not possible to know for most of them that the events occurred in East Timor.

6 Keyword-based headline extraction

After selecting the most relevant sentences from a document, at the beginning of the process, we can follow a completely different approach for headline generation which consists in extracting keywords about the topic discussed in the document. These headlines will not be grammatical, but they may also be informative for a reader.

Indonesia 's National Commission will investigate accusations.
the documents show a total, of Indonesian troops assigned the number.
to extradite former President Suharto; Suharto be extradited.
Stray bullets killed two villagers and police.
Habibie put an end.
Rebels were holding two soldiers; Three soldiers and one activist were killed.
Jakarta does not let six East Timorese asylum-seekers.
Portugal and Indonesia are mulling a new proposal; to settle their dispute over East Timor.
who to break the long-standing deadlock over East Timor
Assailants killed three soldiers and a civilian.

Fig. 3. Summaries generated for the documents in the collection about East Timor.

Setting no.	Doc.	Col.	ROUGE-1 score
1	freq	-	0.25997
2	-	freq	0.21689
3	freq	freq	0.27255
4	wei.	-	0.26810
5	-	wei.	0.20724
6	wei.	wei.	0.29643

Table 5. ROUGE-1 score for the several keyword selection strategies.

In our experiments, the keyword-based headlines have been generated in six ways:

1. By collecting the highest frequency words in the document.
2. By collecting the highest frequency words in the collection.
3. By alternating high frequency words from the document and its collection, i.e. we start with the highest-frequency word in the collection; next the highest-frequency word in the document, next the 2nd. highest-frequency word in the collection, etc.
4. By collecting the words with the highest χ^2 weights in the document.
5. By collecting the words with the highest χ^2 weights in the collection.
6. By alternating the highest weight words from the document and from its collection.

Table 5 shows the results obtained with each approach. As can be seen, using just the collections is not very useful, because all the documents from the same collection will have the same headline. As expected, the best results have been obtained with a combination of the words with the highest frequencies or weights from the document and its collection. In general, we can see that using weights is better than using frequencies. Finally, Figure 4 shows some headlines obtained for the collection about East Timor. From them, a reader can guess the topic of the document, but it is still difficult to grasp the main idea. However, the ROUGE-1 score is surprisingly high (approaching 0.30).

7 Mixed approach

We have seen that a keyword-based approach produces headlines difficult to read, but which are highly informative, as they receive very high ROUGE-1 scores. On the other hand, using the verb phrases produces grammatical headlines, but it is difficult to place

pro-independence, troops, portuguese, timorese, indonesian, timor, east, carrascalao
territory, timorese, portuguese, autonomy, indonesian, timor, east, document
affair, timorese, portugal, timor, indonesian, portuguese, east, extradite
marker, jakarta, pro-independence, timorese, portuguese, be, east, indonesian
timorese, protester, portuguese, activist, indonesian, timor, east, habibie

Fig. 4. Keyword-based summaries generated for the first five documents in the collection about East Timor.

timorese, east, timor, Indonesia 's National Commission will investigate accusations.
timor, the documents show a total, of Indonesian troops assigned the number.
portuguese, east, timor, to extradite former President Suharto; Suharto be extradited.
indonesian, east, timor, be, said Yacob Hamzah , a lawyer; is a Muslim province.
group, portuguese, activist, indonesian, east, timor; xanana, Habibie put an end.
timor, Rebels were holding two soldiers; Three soldiers and one activist were killed.
pro-independence, indonesian, Jakarta does not let six East Timorese asylum-seekers.
marker, Portugal and Indonesia are mulling a new proposal; to settle their dispute over East Timor.
have, say, indonesian, who to break the long-standing deadlock over East Timor.
portuguese, east, timor, indonesian, Assailants killed three soldiers and a civilian.

Fig. 5. Summaries generated for the documents in the collection about East Timor.

the contents of the headline in context, as most of the times there are no topical keywords. A mixed approach can combine the strongest points of both.

Our final approach consists of generating the headlines from the verb phrases of the documents, weighted with the χ^2 weight function. Most of these summaries have far less than 75 bytes, so we can complete them with other information in the following way:

- While the length of the summary is lower than 75,
 - Add the next word, from the document and the collection alternatively.

Furthermore, to check the impact of the keywords, we always add at least one keyword to the VP-based summary. When tested on the DUC-2003 data, this configuration attains a ROUGE-1 score of 0.28270, which is a large improvement from the highest score obtained by the verb phrases alone (0.20105). It is lower than the best mark obtained with only keywords, but the headlines are easier to read as a large part of the headline is formed with complete sentences. The summaries obtained for the collection on East Timor are shown in Table 5.

8 Conclusions and future work

We have developed a method for headline generation that combines a verb-phrase extraction phase and a keyword-extraction procedure. In general, we have observed that the keywords can be very useful for identifying the topic of the text. In fact, the addition of a few keywords boost the ROUGE-1 score from around 0.20 up to around 0.28. On the other hand, the verb phrases not only usually provide the main idea of the document, but also give the headline a more natural and readable shape than headlines formed of just keywords.

Another conclusion that we can draw from the results is that it is equally important to study both the separate documents alone, and the documents inside their collection. A combination of the topic signatures from the documents and from their collections is the one that has produced the best results. It can be argued that having the documents organised in collections is not natural. However, if we have a single document, this problem can in theory be overcome by automatically downloading similar documents from the Internet, or by clustering them to form automatically the collections.

Future work includes a deeper understanding of the meaning of ROUGE-1 and other possible metrics for evaluating automatic summaries.

References

- [1] Lin, C.Y., Hovy, E.H.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of HLT-NAACL 2003. (2003)
- [2] Mani, I.: Automatic Summarization. John Benjamins Publishing Company (2001)
- [3] Angheluta, R., Moens, M.F., Busser, R.D.: K. u. leuven summarization system - DUC 2003. In: Proceedings of DUC-2003. (2003)
- [4] Zhou, L., Hovy, E.: Headline summarization at ISI. In: Proceedings of DUC-2003. (2003)
- [5] Liang, S.F., Devlin, S., Tait, J.: Feature selection for summarising: The sunderland duc 2004 experience. In: Proceedings of DUC-2004, Boston, MA (2004)
- [6] Fuentes, M., Massot, M., Rodríguez, H., Alonso, L.: Mixed approach to headline extraction for DUC 2003. In: Proceedings of DUC-2003. (2003)
- [7] Alfonseca, E., Rodríguez, P.: Description of the UAM system for generating very short summaries at DUC-2003. In: Proceedings of the DUC-2003. (2003)
- [8] Dorr, B., Zajic, D., Schwartz, R.: Hedge Trimmer: A parse-and-trim approach to headline generation. In: Proceedings of Workshop on Automatic Summarization, Edmonton (2003)
- [9] Zajic, D., Dorr, B.J., Schwartz, R.R.: Bbn/umd at duc-2004: Topiary. In: DUC-2004. (2004)
- [10] Daumé III, H., Echihiabi, A., Marcu, D., Munteanu, D., Soricut, R.: GLEANS: A generator of logical extracts and abstracts for nice summaries. In: Proceedings of DUC-2003. (2003)
- [11] Bergler, S., Witte, R., Khalife, M., Li, Z., Rudzickz, F.: Using knowledge-poor coreference resolution for text summarization. In: Proceedings of DUC-2003. (2003)
- [12] Kraaij, W., Spitters, M., Hulth, A.: Headline extraction based on a combination of uni- and multidocument summarization techniques. In: Proceedings of DUC-2003. (2003)
- [13] Witte, R., Bergler, A., Li, Z., Khalifé, M.: Multi-erss and erss 2004. In: DUC-2004. (2004)
- [14] Lin, C.Y.: Rouge working note v. 1.3.1 (2004)
- [15] Erkan, G., Radev, D.R.: The university of michigan at duc 2004. In: DUC-2004. (2004)
- [16] Alfonseca, E., Guirao, J.M., Moreno-Sandoval, A.: Description of the UAM system for generating very short summaries at DUC-2004. In: Proceedings DUC-2004. (2004)
- [17] Jaoua, M., Hamadou, A.B.: Automatic text summarization of scientific articles based on classification of extract's population. In: Proceedings of CICLING-2003. (2003)
- [18] Alfonseca, E.: Wraetlic user guide version 1.0 (2003)
- [19] Brants, T.: Tnt – a statistical part-of-speech tagger. In: Proceedings of the 6th Applied NLP Conference, ANLP-2000, Seattle, WA, U.S.A (2000)
- [20] Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., Wilks, Y.: University of sheffield: Description of the lasie system as used for MUC-6. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann (1995) 207–220
- [21] Lin, C.Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In: Proceedings of the COLING conference. (2000)
- [22] Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19** (1993) 61–74