K.U.Leuven summarization system at DUC 2004

Roxana Angheluta, ∗ Rudradeb Mitra, Xiuli Jing, Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium
Interdisciplinary Center for Law & IT
∗ Department of Computer Science
{roxana.angheluta@law.kuleuven.ac.be, rudradeb_mitra@cs.kuleuven.ac.be,
xiuli.jing@student.kuleuven.ac.be, marie-france.moens@law.kuleuven.ac.be}

1 Introduction

This year at the Document Understanding Conference the K.U.Leuven participated in 3 tasks: very short single-document summarization (headlines), short multi-document summarization and short summarization focused by questions. In the preceding DUC competitions, the summaries were evaluated manually for coverage and quality. This year, an automatic score - ROUGE [5] - was suggested to replace the manual evaluation. For the multi-document summaries and the summaries answering a question, also the manual evaluation was performed for one submission/task/team.

The main conclusion from our experiments is that simple accurate techniques can be effective. Considering manual evaluation, we placed second in the question-focused summarization and fourth in the multi-document summarization. The headlines performed average. We did additional experiments for the headlines and question-focused summaries and we performed short studies on how people build them.

The article is organized as follows: for each of the tasks in which we participated, we present the methods used, the results and discussion of the results. For the headlines and for the question-focused summaries we add our observations about the manually made summaries. We end with conclusions.

2 Headlines

2.1 Methods

Headline generation is not trivial. Depending on the main focus of the headline, one can follow two paths: for a high coverage, picking out keywords seems a good approach; for a good readability, sentence compression techniques are more appropriate. The second method comes closer to the way humans construct headlines. Since each team was allowed more than one submission, we tried both approaches.

2.1.1 Keywords

We detected keywords using our topic segmentation module [10], [8]. This module builds a hierarchical table of content from a document, based on linguistic theories of sentence topic and focus (see figure 1 for an example). We used the topic terms augmented with their collocations as keywords that made up the summary.

```
Sihanouk government     0       1520
        deal    600     970
        senators Assembly       1112    1520
               Sen Senate      1237    1520
                    details 1369    1520
```

Figure 1: Example topic tree. Set d30001t, doc. APW19981124.0267.

2.1.2 Sentence compression techniques

When building headlines by compressing sentences, two steps are required: 1) pick the relevant sentences and 2) compress them.

1. Picking relevant sentences
   We considered for reduction the sentences which contained the topic terms detected with the topic segmentation module (see figure 2 for the way we picked the sentences).

```
start with an initial set of 10 topic terms
repeat
     order sentences by the number of topic terms they contain [a]
     pick the first sentence from the ordered set
     remove from the set of topic terms the ones contained in the sentence picked
until no topic terms are left
```

Figure 2: Algorithm for picking important sentences

[a] We used three heuristics in case of ties: proximity (sentences where the topic terms were close to each other were ranked higher), the number of words in the sentences (shorter sentences were ranked higher) and the position in the text (first sentences from the documents were ranked higher).

2. Compressing sentences
   We employed 2 algorithms for compression: substring selection and a statistical model based on the noisy-channel.

   Substring selection This algorithm selects in the set of good sentences the longest substring between 2 keywords inside a clause (see [2]). The clauses are detected from the parse trees of the sentences [1]. If the substring between the keywords has less than the desired length, next best matches - i.e. next longest substrings - between other keywords are appended to the output until the length is reached.[2] Finally, the determiners and auxiliaries are removed from the resulting headline. An example is given in figure 3.

   Statistical compression The second algorithm used for sentence compression is a variant of the noisy channel algorithm presented in [4]. Shortly described, the algorithm tries to eliminate subtrees from the parse tree of a sentence aiming to arrive to a valid reduction of the original. It is a probabilistic model, using 4 types of probabilities:

   - probabilities referring to the shape of a tree ($P(S->NP \quad VP)$)
   - probabilities referring to valid reductions ($P(S->NP \quad VP$ is reduced to $S->VP)$)
   - probabilities of bigrams ($P(on$ follows $rely)$)
   - prior probabilities assigned to the words of the sentences (addition to [4])

   The first three types of probabilities are learned from training corpora, the fourth is set empirically. These probabilities are combined, giving a score for each possible compression of a chunk (and ultimately of a sentence). The algorithm was designed for sentence compression in general, but we assumed that a good condensed sentence can serve as a headline. It combines into one model both information about grammaticality of a sentence and importance of the words for reduction. We used the same training corpora as described in [4] [3]. The original algorithm made the assumption that for each sentence there is only one valid reduction. This is generally not true, as can be seen in the following example: *John Doe, who used to live in this building, returned to his home town after the accident.* In a context speaking about John Doe, a good reduction might be *John Doe returned to his home*, while in a context speaking about the building, a good reduction might be *John Doe used to live in this building.* The example shows that context information need to be taken into account for reduction. Therefore we introduced high prior probabilities to the keywords detected with the topic segmentation algorithm, to increase their chances of remaining in the final headline. The method proposed in [4] is able to generate many compressed sentences, but we need only one output. We therefore selected the compression with the highest

[1] We used Charniak's parser [3].
[2] Keywords alone are the "next best match" if no other clauses with at least 2 keywords exist.
[3] We thank Prof. Marcu for providing us the training corpus for learning the second type of probabilities.

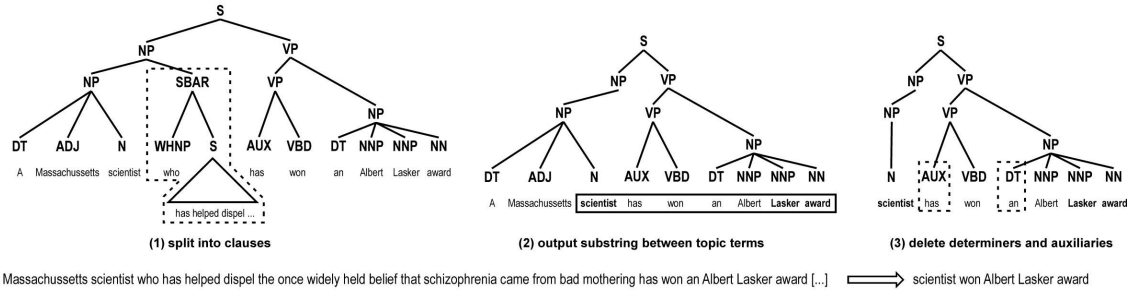

Figure 3: Selecting the longest substring spanned by keywords inside a clause. The keywords present in this sentence are *scientist, Lasker, award*.

a) HUN SEN; OPPOSITION; PRINCE NORODOM RANARIDDH; VICTORY; INVESTIGATION
b) Sam Rainsy and number of opposition figures have been under court investigation for grenade attack on Hun Sen
c) Sam Rainsy said they could not negotiate freely in Cambodia

Figure 4: Examples of headlines for the document APW19981016.0240, set d30001t submitted for task 1: a) keywords extracted from the topic tree; b) substring selection; c) statistical compression.

score and with length lower than 12 [4]. Like in the substring method, a postprocessing step has been applied in the end, removing the determiners and the auxiliaries.

Note: Although the input to this algorithm is, like in the case of substring selection, a set of sentences, in practice only the first sentence from that set is compressed.

Example headlines with each of the three methods described above are in figure 4.

2.2 Results and discussion

Last year, because of the big effort implied by the manual evaluation, each team was allowed to send only one run per task. We therefore combined two methodologies for the headline construction: substring selection with a shallow first sentence compression (for the cases in which the headline output had too few words). We obtained very good results. This year, however, each team was allowed to submit up to 3 runs per task and our purpose was to evaluate the three methods discussed above. The results are presented in figure 5.

The algorithms' performance was average (based on the ROUGE scores) and, for the sentence compression runs, the errors can be due to two facts: 1) we did not pick the good sentence for compression or 2) the compression algorithms were not effective.

2.2.1 Evaluating selection of the sentences

As previously mentioned, although our compression algorithms get as input a set of sentences, in practice only the first one is usually reduced [5]. Therefore we tested the ability of our keyword-based selection algorithm to pick the best sentence for compression. We considered for each document one ideal sentence to be compressed and we computed how many of these ideal sentences were picked by our sentence selection program as the most important sentence - accuracy score (see table 1). The ideal sentences were picked as the ones containing most of the words - except stopwords - of the manual headlines (one sentence for each manual headline). There were 8 summarizers, each making headlines for half of the documents, resulting in 4 manual headlines per document. As an average ideal sentence per document we considered the first ideal sentence picked by at least two summarizers, if it existed (in 80.8% of the cases) or the first ideal sentence appearing in the document. We computed also the accuracy for the baseline case: the first sentence of each document.

[4] Empirically set so that after the postprocessing step we remain with ca. 10 words.

[5] Always in the statistical method and most of the time in the substring method, when it is applicable (see section 2.2.2 to know when it is applicable.)

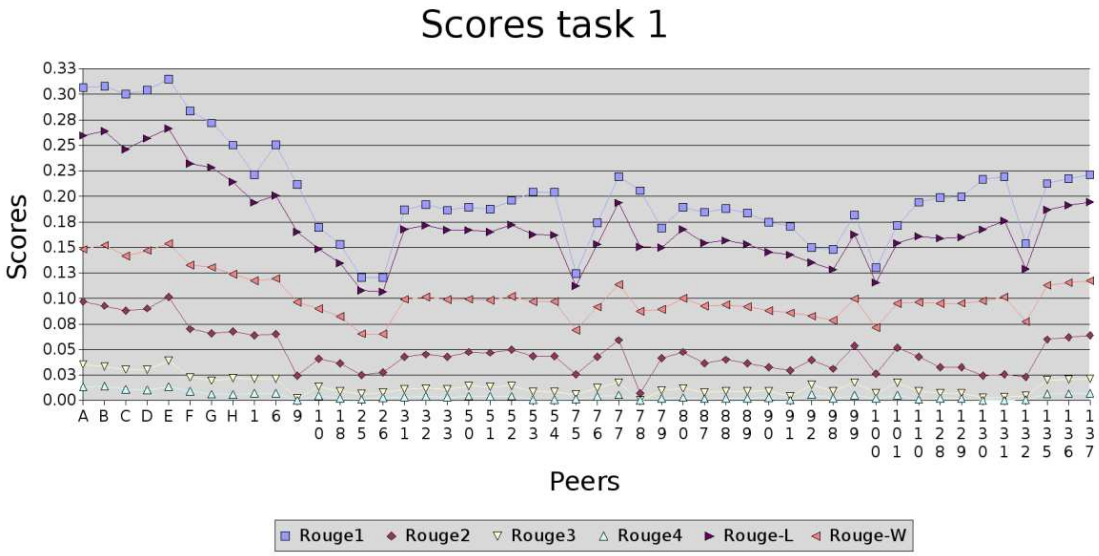

Figure 5: Results task 1. The different curves correspond with different ROUGE scores. The codes of our team are 90 (substring selection), 91 (keywords) and 92 (statistical compression). The codes for the manual submissions are A-H.

	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	All summarizers
keywords-based selection	39.2%	33.2%	37.6%	41.2%	43.6%	31.6%	30.4%	27.2%	47.2%
first sentence	58%	48%	52.8%	58.4%	60%	42%	42.4%	33.2%	71.8%

Table 1: Accuracy obtained by the keyword-based sentence selection - first line and the first sentence - second line with respect to the manual good sentences. The columns 2-9 correspond with each of the human summarizers. The last column combines all summarizers' opinions.

This methodology is questionable, because it makes two assumptions: 1) a good headline can be made by compressing only one sentence and 2) the words from the original sentence are kept in the headline. The second assumption is reasonable if the summarizer does not use a lot of abstracting. Based on a small study, we have indications that the first assumption is true in ca. 70% of the cases. In this study, a student was asked to analyze each of the manual headlines made for the DUC 2003 corpus and write down if at least one of them could have been obtained by compressing only one sentence from the document (keeping the meaning and all the information, but not necessarily using the same words). In 70.8% of the cases the headline could have been obtained by condensing one sentence (in 64.4% of the cases this was the first sentence).

The results show two things: 1) in news documents, picking the first sentence for compression works better than selecting it based on the keywords and 2) using our keyword-based selection of one sentence to be compressed, a headline generation algorithm could output a perfect valid headline in at most ca. 47.2% of the cases [6]. Considering the high percentage of the cases in which the first sentence was the correct one, an obvious strategy for the news stories is to compress the first sentence (result previously mentioned in [11]). This observation is confirmed also by the dropping in ranking comparative with last year, when we combined substring selection with first sentence compression.

Experimentally we run the compression algorithms considering the first sentence from each document as the only relevant sentence. The ROUGE scores obtained are presented in table 2 (last two lines). For comparison, we output also the scores obtained by the substring method and noisy channel variant on the sentence picked by the keyword-based selection algorithm (first two lines). Compressing the first sentence leads in both cases to better results than compressing the sentence with the most keywords.

2.2.2 Evaluating the compression algorithms

Substring selection In almost half of the cases (45.6%), the longest substring between keywords had less than 5

[6] rough estimate from the last column in table 1.

Algorithm	Rouge 1	Rouge 2	Rouge 3	Rouge 4	Rouge L	Rouge W
Substring	0.17460	0.03253	0.00945	0.00280	0.14552	0.08799
Noisy channel	0.15010	0.03949	0.01559	0.00042	0.13502	0.08272
Substring first sentence	0.19553	0.04668	0.01316	0.00383	0.16565	0.09987
Noisy channel first sentence	0.17557	0.05265	0.01989	0.00685	0.15751	0.09606

Table 2: Results task1, applying the compression algorithms on the sentence containing most of the keywords (first 2 lines, corresponding with teams 90, 92 from figure 5) or on the first sentence in each document (last two lines).

words. In these cases, the final headline looked like a keyword headline, which was not our original idea. Analyzing the results of this year and last year, it seems necessary to combine substring selection with another method. However, where it is applicable, i.e. when it doesn't lead to a keyword headline, the method works well and has some advantages: a generic character, no knowledge needed (rules or training) and no parameters to be tuned. We obtained slightly better ROUGE-1, ROUGE-L and ROUGE-W scores with this method comparatively with the other 2 methods, but the results are not conclusive for such small differences.

Statistical compression Sometimes the compressed sentences were not grammatical or did not contain the crucial information from the original sentences. We see two main reasons for this:

- We had two constraints which sometimes did not get along well for the long sentences: compress the sentence to just a few words, ideally keeping the keywords (which can be far apart) in the compression.
- Assigning a high prior probability to the keywords does not always guarantee their presence in the output.

The main problem with this algorithm is the fact that the relative contribution of each type of probability needs to be tuned. We learned that giving them equal importance did not yield to very good results. We tried to set them empirically, but, as the results show, this is not the best way to proceed. A more sophisticated technique (eventually learning the relative importance of each type of probability from a corpus) is needed.

2.3 A note on how people build headlines

Based on the literature [7] and on statistical studies of the DUC 2003 manual headlines corpus on what constitutes a good headline, we classified the headlines according to their syntactic characteristics [13]. We looked at features like the structure, the types of clauses and the articles present in the headline.

Below there is a more detailed description of the categories and - in the brackets - their frequencies in the DUC 2003 manual headline corpus. In total there were 2496 manual headlines, the most frequent being verbal and multi-structure headlines.

- Syntactic structure
  1. verbal headline (1509): finite verbal headlines: *Ebay hosts 1.8 million auctions at any given time* (926); verbal headline with omitted auxiliary: *157 homeless dead in year in San Francisco* (459); non-finite verbal headline: *Talking while driving dangerous and rude* (81); subject+locative adverbial headlines: *TV ratings of Fiesta Bowl below expectations* (18); coordinated verbal headlines: *Space shuttle captures Hubble telescope and prepares for repair mission* (25)
  2. nominal headlines (102): premodified nominal headlines: *slower money growth* (10); postmodified nominal headlines: *Groundbreaking for Charles M. Schulz Museum and Research Center* (28); nominal headlines with both pre- and post- modifications: *click tricks by a top model* (57); coordinated and appositional nominal headlines: *Dr. Susan and the man who came back to live* (7)
  3. adverbial headlines (34): a prepositional phrase: *inside the villages of horror* (5); an adverb followed by: a prepositional phrase: *back with a sparkle* (3), an infinitive phrase: *how to please the farmers by Mr. Crabtree* (4), a conjunctional clause: *just when he thought the problem* (7); a noun phrase: *midnight in Ulster* (15)
  4. headline with more than one structure (611): verbal + verbal: *Anti-abortion web site notes Slepian's murder; claims non-involvment* (394); verbal + nominal: *India, Bangladesh discuss long-standing issues: water sharing, trade imbalance, transit* (62); nominal + verbal: *Massive Three Georges Project: Chinese confident knotty problems solved* (49); nominal + nominal: *Inuit territory of Nunavut: a great victory and*

greater challenge (24); verbal + adjective: *Iran's conservatives win decisive victory in national elections; unsurprising* (15); other multi-structure: *Mozambique aid late; search and rescue underway; more help needed* (67)

5. headlines composed by key-words (240): *Investigation, 400.000, scholarships, 2002 Winter Games, Hodler, bribe, World Court* (240)

- Types of clauses

1. coordinated structures: *Charlie Brown and friends retire but continue life in reprints* (72)
2. quotation + comment clause: *British defense minister says Czech Republic army still not NATO compatible* (79)
3. main clause + dependent clause: modifying clause: *Million soldiers fight flood that have already killed 1.268* (20), complementary clause: *House of Representatives sues Commerce Department over census taking methods* (72), adverbial clause: *Chinese to finish anti-flood projects before Yangtze floods begin* (98), nominal clause: *Mozambicans complain that western governments flood relief is slow arriving* (50)

- The article or determiner

  – should remain: in direct speech quotations: *Manila declaration calls Y2K bug "a social management problem"*; in questions, commands, exclamations and independent wh-headlines: *What the shopping clock says*; when it is a structural marker (followed by a word which can belong to multiple word-classes or in the structure of an apposition): *Taxing the living instead of the dead*; in a prepositional phrase of the form nominal+of+nominal (at least one of the nouns should be preceded by an article): *Study links schizophrenia to deficit in the sense of smell*; for some proper nouns: *Exchange rate established for Greek and Irish currencies' entry into the European Union*; in expressions/idioms: *Recent typhoons could cause inflationary pressures in Philippines for a while*; when a/an has a numerical meaning: *Assailants kill three soldiers and a civilian in East Timor*

3 Multi-document summaries

3.1 Methods

When dealing with multi-document summaries, one has to devote special attention to redundancy elimination. For this purpose, we cluster the term vectors of the important sentences of the single-documents using the covering method [12], [9]. In contrast with the last years [2], [9], when we selected the important sentences based on the level in the topic tree of the topic terms they contained, this year we selected them based on the number of keywords they contained (the same approach like in the case of the headlines - algorithm from figure 2).

3.1.1 Results and discussion

The results improved in comparison with the ones of last year considering the ranking obtained after the manual evaluation. We placed 4^{th} using the coverage score (see figure 6) [7]. The approach we have used to pick important sentences is not new. Even since the fifties Luhn [6] picked important sentences based on the significant words they contained. By obtaining a good result with this method, we show that simple techniques are still effective.

4 Multi-document summaries answering a question

4.1 Methods

The question-focused task requires summaries answering questions of the type *Who is X?*, where X is the name of a person. Our system consists of a succession of filters/sentence selection modules: selecting indicative sentences for the input person, intersecting them with sentences which are important for the whole document and filtering out indirect speech. Finally, to eliminate redundant content while fitting into required length, we cluster the resulting

[7] The automatic results do not correspond with the official ones. Due to an encoding which was not understood by the email client when it received our submission, the summaries we sent for evaluation contained noise (strings "= " were inserted in the text), affecting the results. In figure 6 the automatic results were obtained after removing the noise from the summaries.

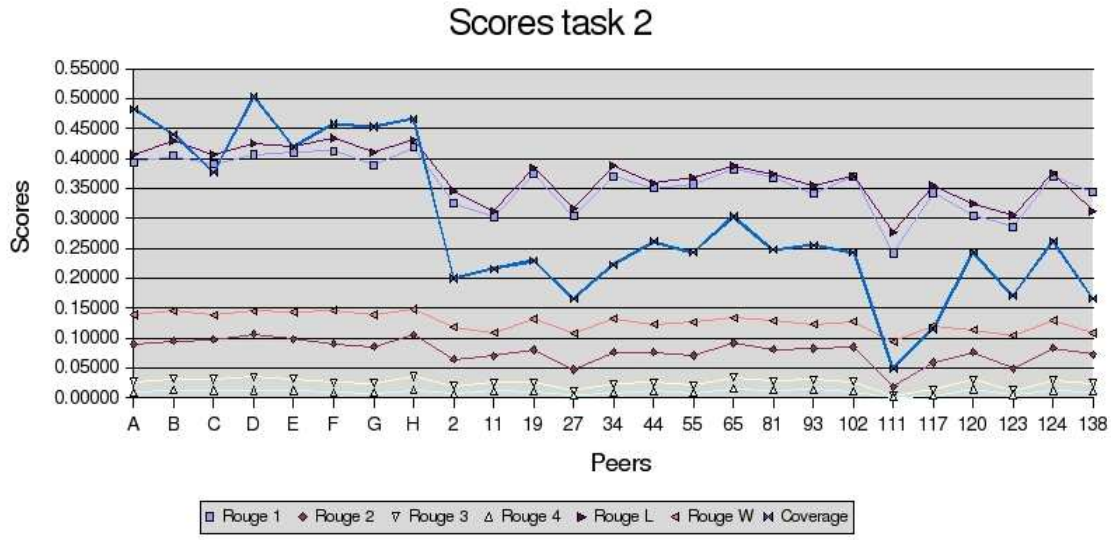

Figure 6: Results task 2. The thin curves correspond with the different ROUGE scores. The thick curve (Coverage) corresponds with the manual evaluation. Our team has code 93. The codes for the manual submissions are A-H.

sentences from all the documents in a set with the covering method. The documents are considered in the order of their dates and the sentences in the reading order in the documents.

The sentences indicative for the input person are detected using an open source coreference resolution program [1], trained on news stories. The program includes a statistical named-entity detector which accepts a user dictionary that forces all instances of a phrase to be tagged as the specified input type. We use this feature of the program to tag the input entity X from the question *Who is X?* as a person. We pick only the sentences in which the coreferent appears before the verb (approximating sentences in which the coreferent is in subject position). The sentences indicating the core topics of the whole document are detected with our topic segmentation module. A manually built list of verbs signaling verbal actions (e.g. *said, told, responded, added*) is used to filter quotations and indirect speech. In an attempt to improve the coherence of the final summaries, we order the sentences by the type of the coreferent word they contain, making sure that the first sentence contained the full name. The workflow of the program is presented in figure 7.

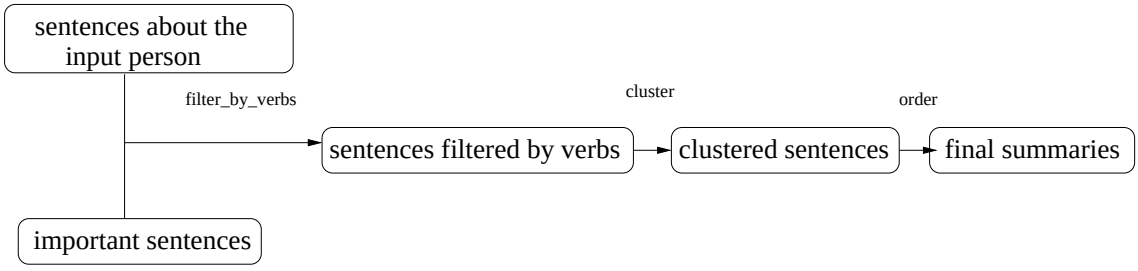

Figure 7: Workflow task 5

4.2 Results and discussion

In the 5^{th} task we got the second place conform with the manual evaluation in terms of coverage. The results are presented in figure 8 [8].

[8]The automatic results do not correspond with the official ones. Due to an encoding which was not understood by the email client when it received our submission, the summaries we sent for evaluation contained noise (strings "= " were inserted in the text), affecting the results. In figure 8 the automatic results were obtained after removing the noise from the summaries.

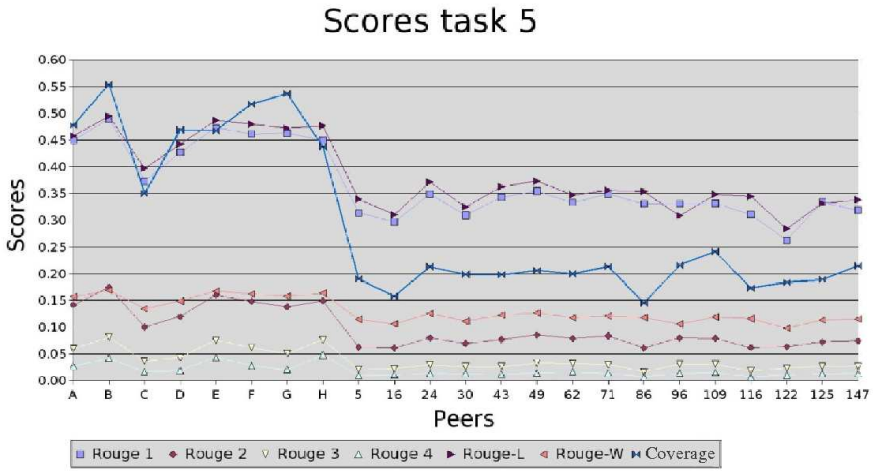

Figure 8: Results task 5. The thin curves correspond with the different ROUGE scores. The thick curve (Coverage) corresponds with the manual evaluation. Our team has code 96. The codes for the manual submissions are A-H.

We evaluated the coreference resolution module separately, on a subset of 30 documents. The precision and the recall for the input person were very high: 96.29% and 85.6% respectively [9]. We also quantified the contribution of each module for the reduction (percentage relative to the input to that module): 63% for the coreference, 89% for the topic segmentation, 15% for the verb filtering and 38% for the clustering. The coreference and the topic segmentation module have the highest influence.

4.2.1 Short analysis of the manual question-focused summaries

In order to get an insight into how people build summaries, we classified the sentences from the manual summaries into 4 categories. Such a classification might be used to develop a strategy for summarizing automatically. We looked at 5 sets from the total of 50 sets.

- Sentences form the original text (no example in our subset)
- Clauses from the original text:
  Manual: On Oct. 20, 1999 he was ousted in a nonviolent military coup
  Original: **Sharif**, *who* **was ousted in a nonviolent military coup on Oct. 20**, *was convicted of hijacking for temporary ordering the diversion of a Karachi-bound commercial plane*
- Combination of sentences from the original text (cut and paste) (no example in our subset)
- Reformulation/Generalization/Inference

  1. where cut and paste would do:
     -from one sentence:
     Manual: Anglican Archbishop Desmond Tutu won the Nobel Peace prize in 1984 for his two decades opposing apartheid in South Africa.
     Original: **The** *69-year old* **Anglican archbishop, who won the Nobel Peace Prize in 1984 for his role in fighting white rule** *which ended with all-race election in 1994, subsequently chaired a commission which probed apartheid-era abuses and gave amnesty to those who confessed of their involvement.*

     -from multiple sentences:
     Manual: Nawaz Sharif was installed as Pakistani's prime minister twice, first in November 1990 and again in February 1997.
     Original: **Sharif first assumed his premiership in November 1990** *when the then president Ghulam Ishaq Khan announced fresh polls after the dismissal of the Pakistan People's Party-led government by*

[9] Thanks to the fact that we could reliably classify the input person and its variants into the semantic class person.

using his constitutional powers under the eight amendment of the Constitution. ..... **Nawaz Sharif came into power for the second time** *when he took oath* **on February 17, 1997** *as Pakistan's top executive following the general elections in the same month.*

2. where cut and paste is not enough:
   Manual: Despite various ailments....
   Original: Helms, 78, has been diagnosed with **peripheral neuropathy** *and received treatment last fall at the Bethesda Naval Medical Center. .....Helms said he has recovered from earlier* **double knee-replacement** *surgery and feels no knee pain..... Over the last decade, Helms also has undergone surgery for* **prostate cancer and a quadruple heart bypass**..... *He has suffered from* **Paget's disease**, *a degenerative bone disorder, in his hip.*

Although we don't have frequencies for each category, our short inspection of the manual summaries confirmed the fact that naive extraction techniques are clearly not enough and information fusion and abstracting are needed in order to build good summaries. The fact that there are examples where cut-and-paste approaches lead to the same information as the one present in the summary in a comparable number of words is encouraging for the automatic techniques. The last category requires additional semantic knowledge and the current summarization systems are yet far of reaching similar performances.

5 Conclusions

We have presented our summarization system used for DUC 2004. We obtained very good results in the question-focused summarization task and good in the multi-document summarization. For the headline task, we performed average. We made additional experiments for the headlines and question-focused summaries. The main conclusion from our experiments is that simple accurate techniques can be effective. We have shortly studied the way in which people build headlines and we concluded that the best approach seems to combine keyword extraction with sentence compression techniques. For the question-focused summaries we have inspected 5 sets of manually made summaries and we came to the conclusion that the current extraction techniques are not enough to build good summaries, but that sentence fusion techniques based on cut and paste from multiple sentences might lead to acceptable results.

References

[1] Alias-i LingPipe *http://www.alias-i.com/lingpipe/index.html* (visited 21.04.2004).

[2] Angheluta R, Moens M-F & De Busser R (2003). *The K.U.Leuven Summarization System DUC-2003.* In Proceedings of the Document Understanding Conference (DUC-2003). National Institute of Standards and Technology, USA.

[3] Charniak E (2000) A Maximum-Entropy-Inspired Parser. In Proceedings of NAACL, 2000.

[4] Knight K & Marcu D (2001). *Statistical-Based Summarization Step One: Sentence Compression (2000).* In Proceedings of AAAI-2001.

[5] Lin C-Y (2003) *Cross-domain Study of N-gram Co-occurrence Metrics.* In Proceedings of the Workshop on Machine Translation Evaluation, Sept. 2003, New Orleans, USA.

[6] Luhn H P (1958) *The automatic creation of literature abstracts.* IBM Journal of Research Development, vol. 2, pp. 159-165, 1958.

[7] Mårdh I (1980). *Headlinese. On the Grammar of English Front Page Headlines.* CWK Gleerup, Gotab, Malmö, 1980.

[8] Moens M-F, Angheluta R, De Busser R & Jeuniaux P (2004). *Summarizing Text at Various Levels of Detail.* In Proceedings of RIAO 2004 Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval (forthcoming).

[9] Moens M-F, Angheluta R & Dumortier J (2004) *Generic Technologies for Single- and Multi-document Summarization.* Information Processing & Management (forthcoming).

[10] Moens M-F & De Buseer R (2001) *Generic Topic Segmentation of Document Texts.* In Proceedings of the 24th ACM SIGIR Annual International Conference on Research and Development in Information Retrieval (pp. 418-419). New York: ACM.

[11] Moens M-F & Dumortier J (2000) *Use of a Text Grammar for Generating Highlight Abstracts of Magazine Articles.* Journal of Documentation, 56 (5), 520-539.

[12] Moens M-F, Uyttendaele C & Dumortier J (1999) *Abstracting of Legal Cases: The Potential of Clustering Based on the Selection of Representative Objects.* Journal of the American Society for Information Science , 50 (2), 151-161.

[13] Xiuli J (2004) *The Evaluation of Headline Generation.* Master Thesis, Katholieke Universiteit Leuven.