K.U.Leuven summarization system at DUC 2004

Roxana Angheluta, Rudradeb Mitra, Xiuli Jing, Marie-Francine Moens
Katholieke Universiteit Leuven
Interdisciplinary Center for Law & IT
Tiensestraat 41, B-3000 Leuven, Belgium
{roxana.angheluta@law.kuleuven.ac.be, rudradeb_mitra@rediffmail.com,
xiuli.jing@student.kuleuven.ac.be, marie-france.moens@law.kuleuven.ac.be}

1 Introduction

This year at the Document Understanding Conferences we 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 [8] - 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 question-focused summarization and fourth in the multi-document summarization. The headlines performed average. We performed additional experiments for the headlines and question-focused summaries, and we are currently researching other metrics for evaluating headlines.

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, we suggest alternate evaluation criteria. 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. Since each team was allowed more than one submission, we tried both approaches.

For selecting keywords, we used our topic segmentation module [7]. 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.

For sentence compression, we had two separate runs, considering for reduction the sentences containing most of the keywords detected from each document with the topic segmentation module. The first algorithm selects from all sentences in our document the longest substring between 2 keywords inside a clause (see [1]). If the

```
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.

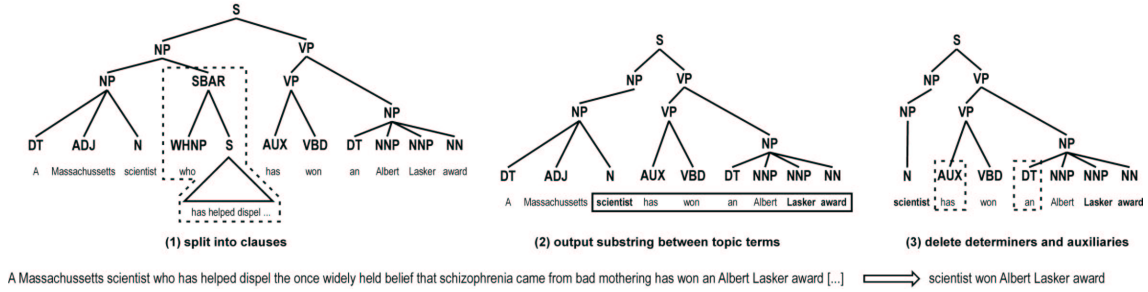

Figure 2: Selecting the longest substring spanned by keywords inside a clause. The keywords present in this sentence are *scientist, Lasker, award.*

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 3: Examples of headlines for the document APW19981016.0240, set d30001t submitted for task 1: a) the keywords extracted from the topic tree; b) longest substring between keywords; c) noisy channel for sentence compression.

best such substring match has less than the desired length, next best matches of other keywords are appended to the output until the length is reached.[1] Finally, the determiners and auxiliaries are removed from the resulting headline. An example is given in figure 2. In 45.6% of the documents, the longest substring between keywords had less than 5 words. In these cases, the final headline looked like a keyword headline.

The second algorithm used for sentence compression was a variant of the noisy channel algorithm presented in [3]. 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 3 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 $rely)$)

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 [3] [2]. By using a corpus with pairs of original/reduced sentences, the 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 assigned higher prior probabilities to the keywords detected with the topic segmentation algorithm, to increase their chances of remaining in the final headline.

Example headlines with each of the three versions described above are in figure 3.

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

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

Accuracy	S 1	S 2	S 3	S 4	S 5	S 6	S 7	S 8	All summarizers
program	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 summarizers. The last column combines all summarizers' opinions.

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 clean techniques, so we did not use the first sentence combined with the substring method. The results are presented in figure 4.

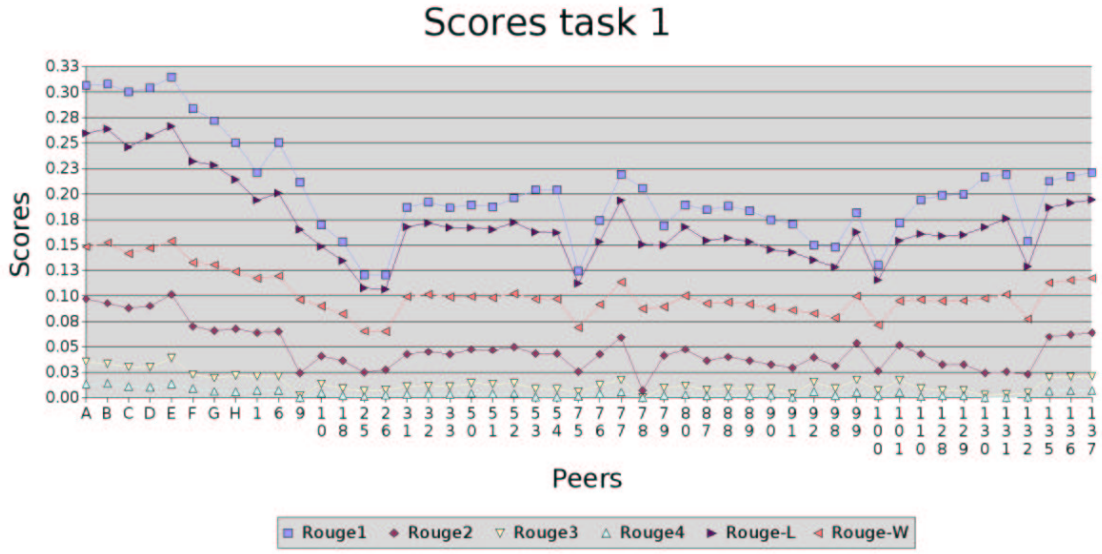

Figure 4: Results task 1. The different curves correspond with different ROUGE scores. The codes of our team are 90 (substring method), 91 (keywords) and 92 (variant of noisy channel). The codes for the manual submissions are A-H.

The algorithms' performance was average (if we can rely on the ROUGE scores) and, for the sentence compression runs, the errors are due to two facts: 1) we did not pick the good sentence for compression or 2) the compression algorithm did not work as expected.

In order to quantify the first type of errors, we considered for each document an ideal sentence to be compressed and we computed how many of these ideal sentences were picked by the program - 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.

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. For 64.4% of the documents that sentence 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 the sentence to be compressed, a headline construction algorithm could output a perfect valid headline in at most ca. 47.2% of the cases. 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. 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 on the first 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 containing most of the keywords (first two lines). Compressing the first sentence leads in both cases to better results than compressing the sentence with the most keywords.

We are currently performing an error analysis for the compression algorithms.

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 reduction algorithms on the sentence containing most of the keywords (first 2 lines) or on the first sentence in each document (last two lines).

2.3 Alternate evaluation

Based on literature [6] and statistical studies on what constitutes a good headline we found that 3 types of evaluation criteria can be used as measures to evaluate headlines: syntactic, length-oriented and semantic.

Regarding the syntax of a headline, we studied the DUC 2003 manual headlines corpus, looking at features like the structure, the types of clauses and the articles present in the headline. For each of these features we classified the headlines from the corpus into categories. Having a new headline, if it's grammatical, one can categorize it into one of the classes and score it with the probability of its class.

Below there is a more detailed description of the categories and their frequencies in the DUC 2003 manual headline corpus:

- Syntactic structure

  1. verbal headline: 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: 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: a prepositional phrase: *inside the villages of horror* (5); an adverb followed by: a prepositional phrase: *back with a sparkle* (3), infinitive phrase: *how to please the farmers by Mr. Crabtree* (4), conjunctional clause: *just when he thought the problem* (7); a noun phrase: *midnight in Ulster* (15)
  4. headline with more than one structure: 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 grater challenge* (24); verbal + adjective: *Iran's conservatives win decisive victory in national elections; unsurprising* (15); other multi-structure: *Mozambique aid late; search and resque underway; more help needed* (67)

5. headlines composed by key-words: *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

1. 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: *Taxing the living instead of the dead*; in a preposition phrase of the form nominal+of+nominal: *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 inherent phrases: *Fears of Y2K may be the biggest part of problem*; when a/an has a numerical meaning: *Assailants kill three soldiers and a civilian in East Timor*

The length-oriented criteria refer to the length of the headline comparing with the length of an ideal headline. In DUC 2003 corpus, the ideal headline had 10 words.

The semantic criteria refer to:

- the degree of catching the main idea of the original text (coverage score)
- how well each word/chunk acts as a headline word

Beside evaluation purposes, the syntactic categories could be used by a headline generation program: having multiple outputs, the program could chose the one whose category is the most frequent.

3 Multi-document summaries

3.1 Methods

For the multi-document summaries, we cluster the term vectors of sentences of the single-document with the covering method. In contrast with last years [1], [2], when we built the single-document summaries by picking sentences based on the level of the topic terms they contained in the topic tree computed with our topic segmentation algorithm [7], this year we select them based on the number of topic terms they contained.

The covering method is a non-hierarchical (partitioning) method that is based on the selection of representative objects (i.e., medoids). A candidate medoid attracts the most similar sentences from the set of remaining sentences based on a criterion or constraint of cluster goodness. The problem that the algorithm tries to solve can be seen as an optimization problem. The mathematical models for the algorithm uses the following notations:

- The set of n objects (i.e., sentences) to be clustered is denoted by $X = x_1, x_2, ..., x_n$.
- The similarity between objects x_i and x_j (also called objects i and j) is denoted by $s(i, j)$. In our implementation it was computed as the cosine between the term vectors of the sentences. Feature words can be selected based on their POS class. In our experiments we restrict the words to nouns, verbs and adjectives.

A solution to the model is determined by two types of decisions:

- The selection of objects as representative objects in clusters: y_i is defined as a 0-1 variable, equal to 1 if and only if object i is selected as the representative object ($i = 1, ..., n$).
- The assignments of each object j to one of the selected representative objects: z_{ij} is a 0-1 variable, equal to 1 if and only if object j is assigned to the cluster of which i is the representative object.

In the covering method the objective is to minimize the number of clusters (equivalent to minimizing the number of medoids) so that the similarity between the medoid and the objects of each cluster is greater than a threshold. This can be represented in a mathematical model as:

$$\text{minimize} \sum_{i=1}^{n}\sum_{j=1}^{n} s(i,j)z_{ij}$$

$$\text{where}$$

$$\sum_{i=1}^{n} z_{ij} = 1, j = 1, 2, ..., n$$

$$z_{ij} <= yi, , i, j = 1, 2, ..., n$$

$$s(i,j) >= S \text{ where } z_{ij} = 1, i, j = 1, 2, ..., n$$

The complexity of the algorithm grows exponentially with the number of objects to be clustered. Therefore we implemented a good, but not optimal solution for the clustering when the number of clusters exceeds a threshold value. Individual subsets of objects are clustered and clusters are merged when the similarity of their medoids exceeds the required threshold similarity value (S) and the medoid of the merged cluster is recomputed. The medoid sentences make up the summary.

We implemented the algorithm in such a way that the output can flexibly adapt to the required summary length. It allows choosing a minimum ($Smin$) and a maximum ($Smax$) threshold similarity. Between these two limits, the threshold can take a variable number of values. We split the interval [$Smin$,$Smax$] in smaller intervals (their number is a parameter in the program) and for each small interval a solution is computed. The solution that best fits the required length of the summary is picked.

3.1.1 Results and discussion

The results improved comparative with last year considering the ranking obtained after the manual evaluation. We placed 4^{th} using the coverage score (see figure 5) [3]. The approach we have used to construct the single summaries is not new. Even since the fifties Luhn [5] 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, filtering out sentences which are not important for the whole document and filtering out indirect speech. Finally, to eliminate redundant content while fitting into required length, we cluster the resulting sentences from all the documents in a set with the covering method (see section 3.1).

The sentences indicative for the input person were detected using an open source coreference resolution program [4], 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 used this feature

[3]The automatic results do not correspond with the official ones. Due to a encoding which was not understood by the email client when it received our submission, the summaries we sent for evaluation contained noise (strings "= " inserted in the text), affecting the results. In figure 5 the automatic results were obtained after removing the noise from the summaries.

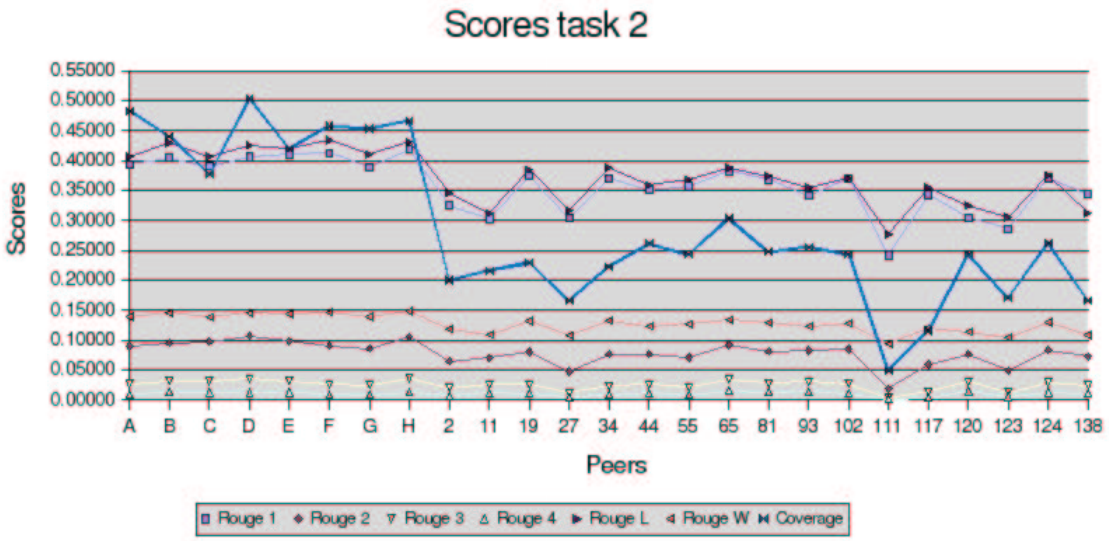

Figure 5: Results task 2. The thin curves correspond with the different ROUGE scores. The thick curve corresponds with the manual evaluation. Our team has code 93. The codes for the manual submissions are A-H.

of the program to tag the input entity X from the question *Who is X?* as a person. We picked only the sentences in which the coreferent appeared before the verb (approximating sentences in which the coreferent is in subject position).

The sentences indicating the core topics of the whole document were detected using our topic segmentation module. A manually built list of verbs signaling verbal actions (e.g. *said, told, responded, added*) was used to filter quotations and indirect speech. In an attempt to improve the coherence of the final summaries, we ordered the sentences by the type of the coreferent word they contained (priorities: full name > last name > pronoun).

An alternate experiment replaced the coreference program with a simple baseline selecting sentences that contained the exact match of the input person. The workflow of the program is presented in figure 6.

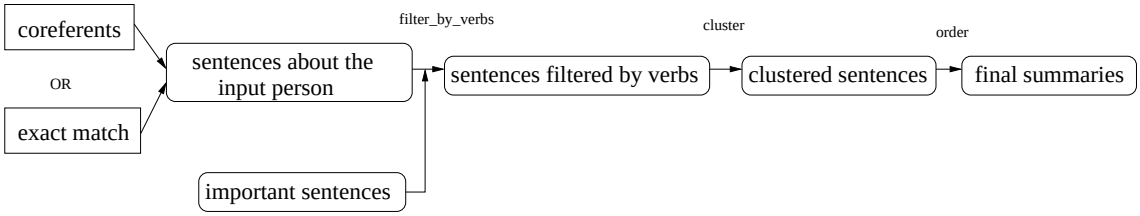

Figure 6: Workflow task 5

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 7 [4].

In the experiment considering the exact matches instead of the coreferents, evaluating the results with ROUGE 1 score, the scores drop, but not significantly, suggesting that the main role is played by the topic segmentation module.

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

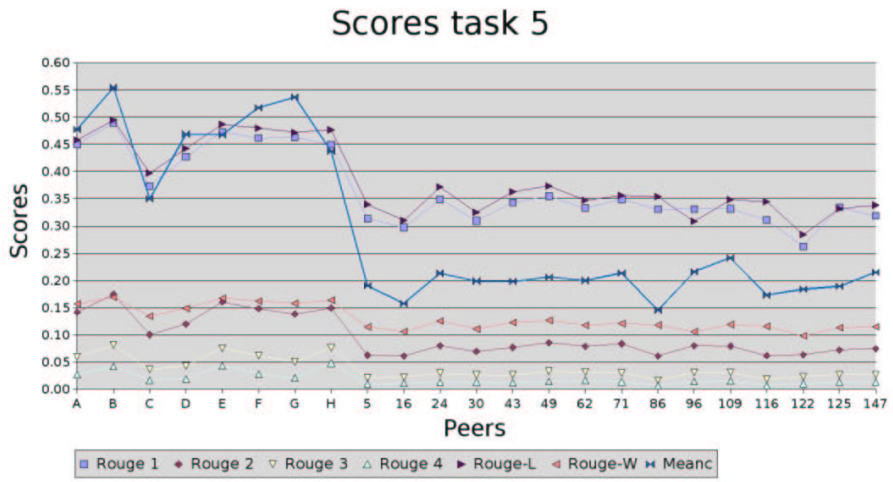

Figure 7: Results task 5. The thin curves correspond with the different ROUGE scores. The thick curve corresponds with the manual evaluation. Our team has code 96. The codes for the manual submissions are A-H.

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 defined a couple of criteria for a good headline based on the manually made DUC 2003 headlines.

References

[1] 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.

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

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

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

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

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

[7] 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).

[8] 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.