A Cosine Maximization-Minimization approach for User-Oriented Multi-Document Update Summarization

Florian Boudin[♮] and Juan-Manuel Torres-Moreno[♮,♭]
[♮]Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228,
84911 Avignon Cedex 9, France.
{florian.boudin,juan-manuel.torres}@univ-avignon.fr
[♭] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.

Abstract

This paper presents a User-Oriented Multi-Document Update Summarization system based on a maximization-minimization approach. Our system relies on two main concepts. The first one is the cross summaries sentence redundancy removal which tempt to limit the redundancy of information between the update summary and the previous ones. The second concept is the newness of information detection in a cluster of documents. We try to adapt the clustering technique of bag of words extraction to a topic enrichment method that extend the topic with unique information. In the DUC 2007 update evaluation, our system obtained very good results in both automatic and human evaluations.

Keywords

User-Oriented Multi-Document Summarization, Question Focused Summarization, Update Summarization, Statistical approach, Detection of Newness, DUC evaluation, Cross summaries redundancy removal

1 Introduction

The seventh edition of the Document Understanding Conference[1] (DUC) has introduced a pilot task in counterpart to the question-focused multi-document summarization main task. Named update task, it's goal is to produce short update summaries of newswire articles under the assumption that the user has already read a set of earlier articles. This is the first time, as far as we know, that an update summarization task is evaluated. We have chosen to relies our system's approach on two main concepts: cross summaries sentence redundancy removal and newness of information detection using a bag of words extraction method for topic enrichment. The rest of the paper is organized as follows. Section 2 describes the previous works and section 3 the update task of DUC 2007. The section 4 introduces the two main ideas of our approach quote above. The section 5 gives an overview of the experiments and section 6 the performance of the system at the DUC 2007. Section 7 concludes this paper and examines possible futher work.

[1] Document Understanding Conferences are conducted since 2000 by the National Institute of Standards and Technology (NIST), http://www-nlpir.nist.gov

2 Background and related work

Interest in multi-document summarization of newswire started with the on-line publishing and the constant growth of internet. Introduced by Luhn [5] and Rath et al. [12] in the 50s-60s with single-document summarizers (SDS), research on automatic summarization can be qualified as a long tradition. However, the first automatic Multi-Document Summarizers (MDS) were developed only in the mid 90s [9]. Lately, DUC 2007 conference introduced the over-the-time update MDS evaluation. Most of work in automatic summarization apply statistical techniques to linguistic units such as terms, sentences, etc. to select, evaluate, order and assemblate them according to their relevance to produces summaries [6]. In general, summaries are constructed by extracting the most relevant sentences of documents. Automatic summarization systems can be divided in two categories: single document summarizers and more complex multidocument summarizers. Multi-document systems can be viewed as fusionning SDS systems outputs by using additionnal information about the document set as a whole, as well as individual documents [1]. MDS perform the same task as SDS but increase the probability of information redundancy and contradictions. Previous works comparing the redundancy techniques [10] have shown that using a simple *zero knowledge* vector based cosine similarity [15] for measuring sentence similarities make no difference in performance with more complex representation, such as Latent Semantic Indexing [2]. Au contraire to redundancy removal, precious little researchers have focused on time-based summarization. A natural way to go about time-based summarization is to extract the temporal tags [7] (dates, elapsed times, temporal expressions, ...) or to automatically construct the timeline from the documents [14]. For the last technique, the well known χ^2 measure [8] is

used to extract unusual words and phrases from documents. Our approach is based on the same principle of term extraction but differs from these in several ways. Our system relies on the simple idea that the most important unique terms of a cluster are suitable for representing the unique and unseen information.

3 Description of the DUC 2007 pilot task

The DUC 2007 update task goal is to produce a short (∼100 words) multi-document update summaries of newswire articles under the assumption that the user has already read a set of earlier articles. The purpose of each update summary will be to inform the reader of new information about a particular topic. Given a DUC topic and its 3 document clusters : A, B and C, create from the documents three brief, fluent summaries that contribute to satisfying the information need expressed in the topic statement.

1. A summary of documents in cluster A.
2. An update summary of documents in B, under the assumption that the reader has already read documents in A.
3. An update summary of documents in C, under the assumption that the reader has already read documents in A and B.

Within a topic, the document clusters must be processed in chronological order; i.e., we cannot look at documents in cluster B or C when generating the summary for cluster A, and you cannot look at the documents in cluster C when generating the summary for cluster B. However, the documents within a cluster can be processed in any order.

4 A Cosine Maximization-Minimization approach

This paper proposes a statistical method based on a maximization-minimization of cosine similarity measures between sentence vectors. The main motivation behind this approach is to find a way to quantify the newness of information contained in an document cluster assuming a given topic and a set of already "known" document clusters but at the same time minimize the possible redundant information. The main advantage of this approach is that *zero knowledge* is required and that makes the system fully adjustable to any language. The following subsections formally define the measures formulas and the method to apply it to the update summarization task.

4.1 Back to basics: a simple User-Oriented MDS

We have first started by implementing a *baseline* system for which the task is to produce topic focused summaries from document clusters. Standard preprocessings are applied to the corpora, sentences are filtered (words which do not carry meaning are removed such as functional words or common words) and stemmed using the Porter algorithm [11]. An N-dimensional termspace Ξ, where N is the number of unique terms found in the corpus, is constructed. Sentences of a document are represented in Ξ by a vector. Similarity measures between sentences are calculated by using the angle cosine. The smaller the angle, the greater is the similarity. The system scores each sentence of a document by calculating the cosine similarity angle measure [13] (defined in formula 1 and illustrated by figure 1 with the θ_t) between the topic vector and the sentence vector using the well known $tf \times idf$ measures as weights. tf is the term frequency in the document and idf is the inverse document frequency. idf values are calculated on the whole DUC 2007 corpus (main and update task).

$$\cos(\vec{s}, \vec{t}\,) = \frac{\vec{s} \cdot \vec{t}}{\sqrt{\| \vec{s}\,\| + \| \vec{t}\,\|}} \tag{1}$$

In our case, $\vec{s}$ is the vectorial representation of the candidate sentence and $\vec{t}$ of the topic.

4.2 Redundancy removal techniques

Sentences coming from multiple documents and assembled together to generate a summary theoretically engender redundancy problems for classified document cluster. In practice, sentences of a cluster are all scored by calculating an angle regarding a particular topic, accordingly all high scored sentences are syntactically related. We have to deal with two different redundancy problems in our update MDS system, the within summary syntactical sentence redundancy and the cross summaries redundancy. The first one refers to the detection of duplicate sentences within a summary. We choose to measure the sentence similarity between the sentences already contained in the summary and the candidate sentences and remove them if this similarity is greater than a threshold T_o, empirically fixed. The second problem is more specific to the task, candidate summaries are generated assuming that other summaries have previously been produced. Therefore they have to contain different information about the same topic and inform the reader of new facts. Formally, n_p early summaries are represented as a set of vectors $\Pi = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_{n_p}\}$ in the termspace Ξ. Our sentence scoring method (formula 2) calculates a ratio between two angles: the sentence $\vec{s}$ with the topic $\vec{t}$ and the sentence with the all previous n_p summaries. The smaller value $\eta(\vec{s}, \vec{t}\,)$ and the higher value $\phi(\vec{s}, \Pi)$ produces the greater score $R(\bullet)$:

$$R(\vec{s}, \vec{t}, \Pi) = \frac{\eta(\vec{s}, \vec{t}\,)}{\phi(\vec{s}, \Pi) + 1} \tag{2}$$

where: $\eta(s, \vec{t}\,) = \cos(\vec{s}, \vec{t}\,)$

$\phi(\vec{s}, \Pi) = \sqrt{\sum_{i=1}^{n_p} \cos(\vec{s}, \vec{p}_i)^2}$

$0 \leq \eta(\bullet)\ ;\ \phi(\bullet) \leq 1$

Therefore:

$$\max R(s) \Longrightarrow \begin{cases} \max \eta(\bullet) \\ \min \phi(\bullet) \end{cases} \quad (3)$$

The highest scored sentence $\vec{s}$ is the most relevant assuming the topic $\vec{t}$ (i.e. $\eta \to 1$) and, simultaneously, the most different assuming the previous summaries Π (i.e $\phi \to 0$).

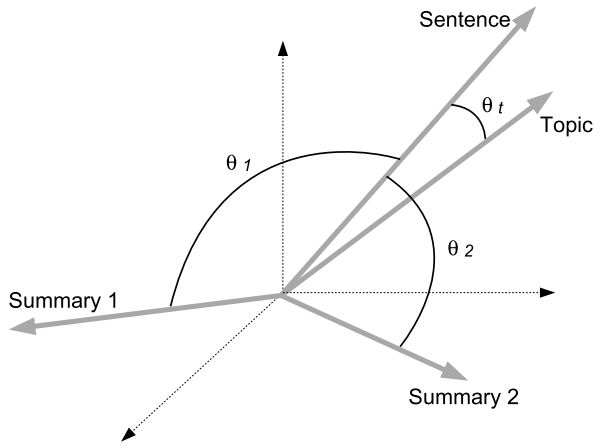

Fig. 1: *Cosine Maximization - Minimization illustration example, the case of two previous summaries: for each sentence, minimize the angle θ_t and maximize the angles θ_1 and θ_2*

4.3 Newness of information

The detection of the newness of information is a critical point in the update summarization process. Indeed, how to detect, quantify and "blend" unseen information into an existing MDS system are challenging questions that we try to answer with our approach. In the same way that several previous works in document clustering use a list of high $tf \times idf$ terms as topic descriptors, we have chosen to represent the most important information of a document cluster X by a bag of word B_X of the n_t highest $tf \times idf$ words. The newness of information of a document cluster A in relation to already processed clusters is the difference of it's bag of words B_A and the intersection of B_A with all the previous cluster's bags of words (see formula 3). The system uses the terms of B_X to enrich the topic t of the cluster X, the topic is extented by a small part of the unique information contained in the cluster. Selected sentences are not only focused on the topic but also on the unique information of the cluster.

$$B_X = B_X \setminus \bigcup_{i=1}^{i=n_p} B_i \quad (4)$$

4.4 Summary construction

The final summary is constructed by arranging the most high scored sentences until the limit size of 100 words is reached. As a consequence the last sentence have a very high probability to be truncated. We propose a last sentence selection method to improve the summary's reading quality by looking at the next sentence. This method is applied only if the remaining word number in greater than 5 otherwise we just produce a non-optimal size summary. The after last sentence is prefered to the last if it's length is almost 33% shorter and to avoid noise if it's score is greater than a threshold of 0.15. In all cases the last summary sentence is truncated of 3 words maximum. We try to protect the sentence grammaticality by removing only stop-words and very high frequency words. A set of about fifty re-writing patterns and a dictionnary based name redundancy removal system have been specially created for the DUC update task. The figure 2 is a global overview of the main architecture of our system.

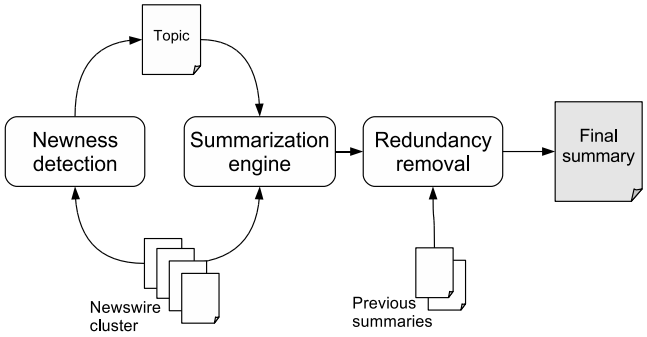

Fig. 2: *General architecture of the update summarization system.*

5 Experiments

The method described in the previous section has been implemented and evaluated. The following subsections present some details of the different parameter settings experiments.

5.1 Experimental Settings

We used for our experimentations the DUC 2007 update task data set, the task is described in section 3. The corpus is composed of 10 topics, with 25 documents per topic. For each topic, the documents will be ordered chronologically and then partitioned into 3 sets : A, B and C, where the time stamps on all the documents in each set are ordered such that time(A) < time(B) < time(C). There is approximately 10 documents in set A, 8 in set B, and 7 in set C. Tuning the system parameters requires to find a way of automatically evaluate the quality of the produced summaries and producing reliable and stable scores. All existing automated evaluation methods work by comparing the systems output summary to one of more reference summaries (ideally, produced by humans). The ROUGE [4] ant Basic Elements [3] automated performance measures are considered relevant and will be used for our experiments.

5.1.1 Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE [4] is a word n-gram recall between a candidate summary and a set of reference summaries. In our experiments the two ROUGE-2 and ROUGE-SU4 measures will be computed. ROUGE-2 measure which is based on bigram of words is defined in equation 5. $Count_{match}$ stands for the maximum number of bigrams co-occurring in a candidate summary and a set of reference summaries R_S. The ROUGE-2 is a recall-related measure because of the denominator of the equation is the total sum of the number of bigrams occurring in the reference summaries.

$$\text{ROUGE-2} = \frac{\sum_{s \in R_S} \sum_{bigram \in s} Count_{match}}{\sum_{s \in R_S} \sum_{bigram \in s} Count} \quad (5)$$

The ROUGE-SU4 measure is a also a word bigram recall but extended to take into account the unigrams and allowing for arbitrary gaps of maximum length 4. For example the sentence *"why using text summarization"* has $Count(4,2) = 6$ skip-bigrams which are: *"why using", "why text", "why summarization", "using text", "using summarization", "text summarization"*. We calculated the count of skip-bigrams with an arbitrary gap γ an we it defined in equation 6.

$$Count(k,n) = C\binom{n}{k} - \sum_{0}^{k-\gamma}(k-\gamma)\ ;\ \gamma \geq 1 \quad (6)$$

where n is the n-gram length and k the sentence length in words.

5.1.2 Basic Element (BE)

Basic Element [3] is a specific evaluation method using very small units of content, called Basic Element, that adress some of the shortcomings of n-grams. The problem of the ROUGE evaluation is that multi-word units (such as "United Mexican States") are not treated as single items, thereby skewing the scoring, and that relatively unimportant words (such as "from") count the same as relatively more important ones. The Basic Element evaluation tempt to solve this problems by using a syntactic parser to extract just the valid minimal semantic units, called BEs.

5.2 Newness of information

One of the major difficulties is to evaluate and optimize the quantity of "newness" terms extracted from the clusters. If too much terms are extracted the produced summaries will be away from the point considering the topic. Otherwise, if too few terms are extracted, summaries readability will decrease due to the high information redundancy. We can observe in figure 3 that the topic enrichment always decreases automatic evaluation scores. This is due to the "noise" introduced by the newness of information terms extracted. Our experiments have also shown that the newness of information enrichment considerably enhances the readability and the intrinsic quality of the produced summaries. The information containing in the summaries is more heterogeneously spread, syntactical redundancy decrease and so readability and general quality enhance.

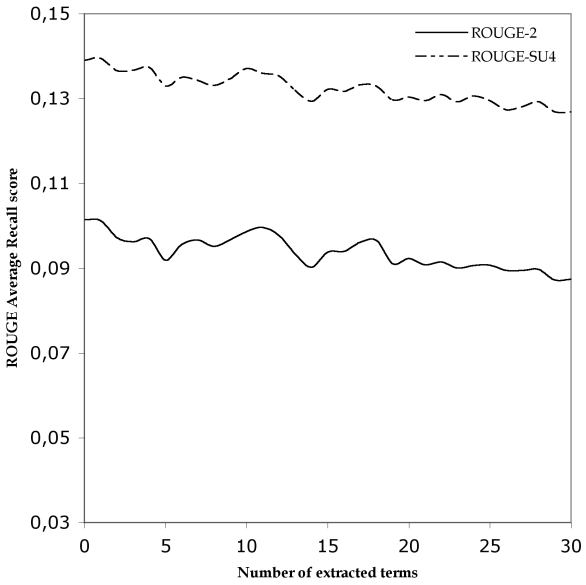

Fig. 3: *ROUGE average recall scores in comparison to the number of extracted terms for the topic enrichment.*

5.3 Within summary redundancy

We have implemented two similarity measures to deal with the within summary sentence redundancy problem. These measures are calculated between a candidate sentence and the sentences that are already considered as summary's sentences. The first one is a normalized symmetrical word overlapping measure whereas the second is a classic cosine similarity measure. A candidate sentence is accepted in the final summary only if it's similarity scores with the other summary sentences are lower than a threshold. Previous works [10] have shown that the classic cosine similarity measure (see equation 1) is the most performant measure for redundancy removal task. The two measures are binded by the fact that they use the terms has units of comparison so we decide to use only the classic cosine similarity. The sentence acceptance threshold has been tuned empirically using the ROUGE automatic evaluation as reference measure, ROUGE scores are increasing until the threshold is reaching 0.4 (see figure 4). In other words, the deletion of sentence with lower cosine score that 0.4 remove information from the candidate summary and a sentence is considered as increasing the summary redundancy if at least one of it's cosine scores with the other sentences is greater than 0.4.

5.4 Experiments on DUC 2007 data

The above sections delineate the tuning techniques using the DUC 2007 corpus as reference and so how we

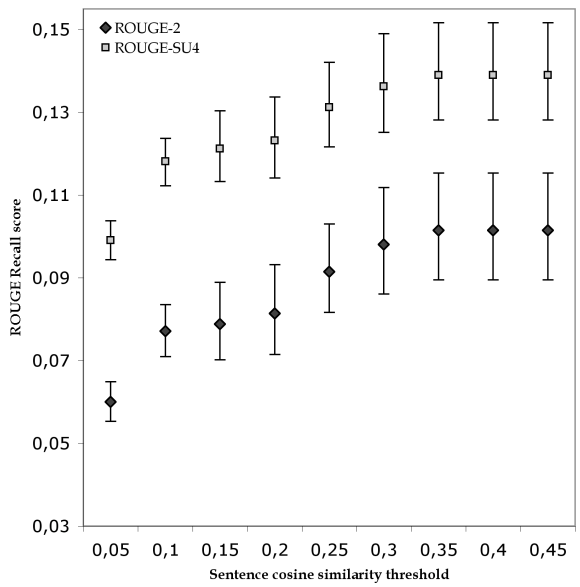

Fig. 4: *ROUGE average recall scores versus the redundancy similarity measure threshold.*

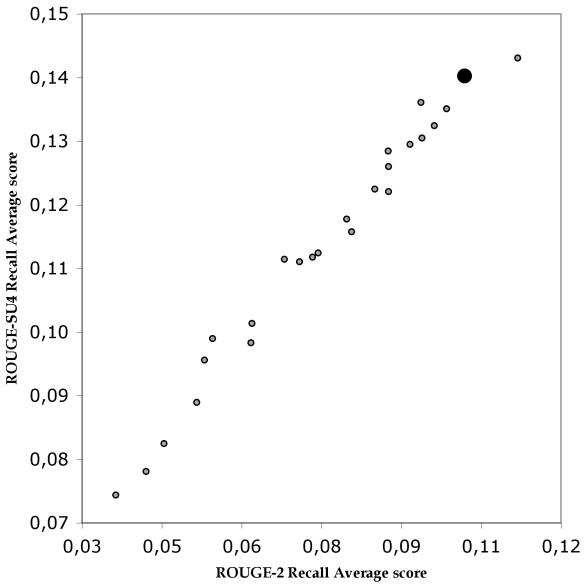

Fig. 5: *ROUGE-2 versus ROUGE-SU4 scores for the 24 participants of DUC 2007 update evaluation (our system is the dark circle).*

found the optimal parameter combination by comparing our system automatic evaluation scores. This section will evaluate our system performance in the optimal parameter combination with the 24 participants of DUC 2007 update task (in which we participate with a non-optimal version of this system, the system's id is 47). An example of our system output for the topic D0726F is shown in the appendixes section. We observe in the figure 5 that our system is the second best system for the ROUGE automatic evaluation, this is a very good performance in view of the fact that the applied post-processings achieve poor performance and that they are not designed especially for the task but are more generic ones. An important margin of progress in improving these main post-processings appears. Sentence rewritting process in the specific kind of document used in the DUC conferences is not yet developed but we are currently investigate sentence reduction techniques.

6 The system at DUC 2007

This section present the results obtained by our system at the DUC 2007 update evaluation. No training corpus was, at the time of submission, available and there was, as far as we known, no equivalent corpora for training systems. Only manual evaluation of the output summaries was possible, this explain why the parameters used for the system submission are not the optimal ones. The following parameters have been used for the final evaluation : Bag of words size : 15, Redundancy threshold : 0.4, minimal sentence length: 5. Among the 24 participants, our system ranks 4^{th} in both ROUGE-2 and Basic Element evaluation, the 5^{th} in ROUGE-SU4 evaluation and the 7^{th} in overall responsiveness. The figure 6 shows the correlation between the average ROUGE scores (ROUGE-2 and ROUGE-SU4) of the systems and their average responsiveness scores. ROUGE-2 and ROUGE-SU4 scores were computed by running ROUGE-1.5.5 with stemming but no removal of stopwords. The input file implemented jackknifing so that scores of systems and humans could be compared. The content responsiveness evaluation assesses how well each summary responds to the topic. The content responsiveness score is an integer between 1 (very poor) and 5 (very good) and is based on the amount of information in the summary that helps to satisfy the information need expressed in the topic narrative. The average responsiveness score obtained by our system was 2.633, which is above the mean (2.32 with standard deviation of 0.35). Our system is contained in the group of the top 8 well balanced systems (It must be noticed that the value of the scores range in a small interval), the mean responsiveness score (ranked only 7^{th}) is due to the poor rewritting sentence post-processing (only less than fifty general rewritting regular expressions).

The figure 7 illustrates another automatic measure, the previously described Basic Element (BE) evaluation measure. Basic Elements (BE) scores were computed by first using the tools in BE-1.1 to extract BE-F from each sentence-segmented. The BE-F were then matched by running ROUGE-1.5.5 with stemming, using the Head-Modifier (HM) matching criterion. For average BE our system scored 0.05458, which is above the mean (0.04093 with standard deviation of 0.0139) and ranked 4^{th} out of 24 systems. We observe in the figure 8 that the average automatic scores are better for the last summary (cluster C) and most of all that the standard deviations extensively decrease (see table 1). The stability of our system enhance with the quantity of previous time documents, the light fall with the cluster B summaries may be due to the non-optimal enrichment done without enough previous extracted terms.

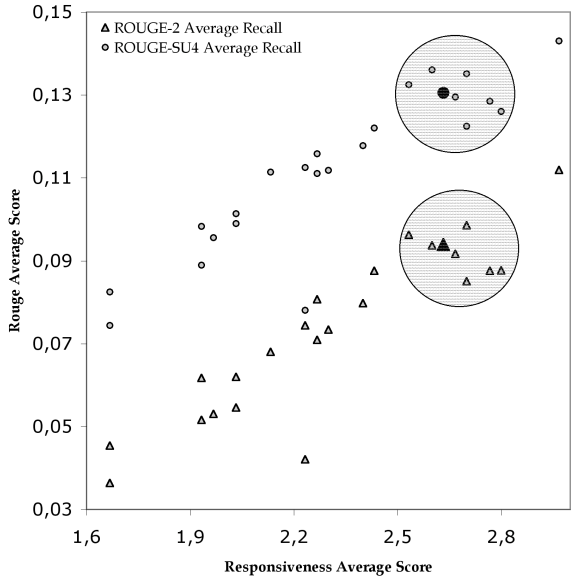

Fig. 6: *ROUGE versus responsiveness scores for the 24 participants of the DUC 2007 update evaluation. Our system is the dark circle for ROUGE-2 and the dark triangle for ROUGE-SU4.*

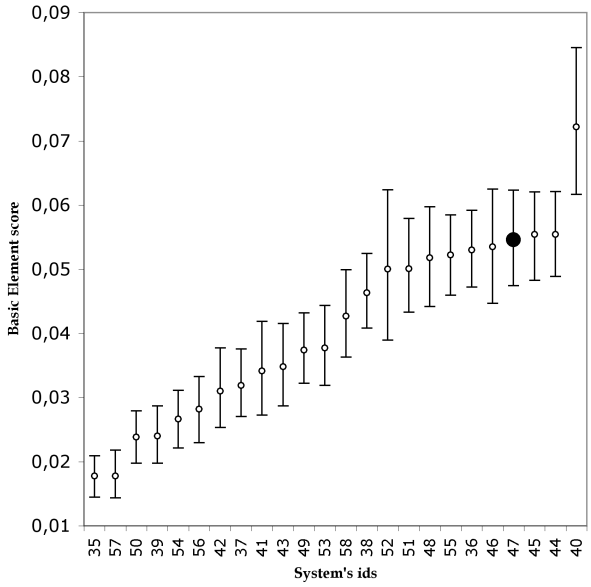

Fig. 7: *Basic Element (BE) scores of the 24 participants of the DUC 2007 update task, our system id is 47 (marked in the figure by the dark circle).*

Cluster	A	B	C
ROUGE-2	0,08170	0,08080	0,03670
ROUGE-SU4	0,08657	0,06826	0,02878

Table 1: *ROUGE scores standard deviations of our system for each document cluster used.*

After analysing all the figures, one system clearly stand out from the crowd (this system id is the 40),

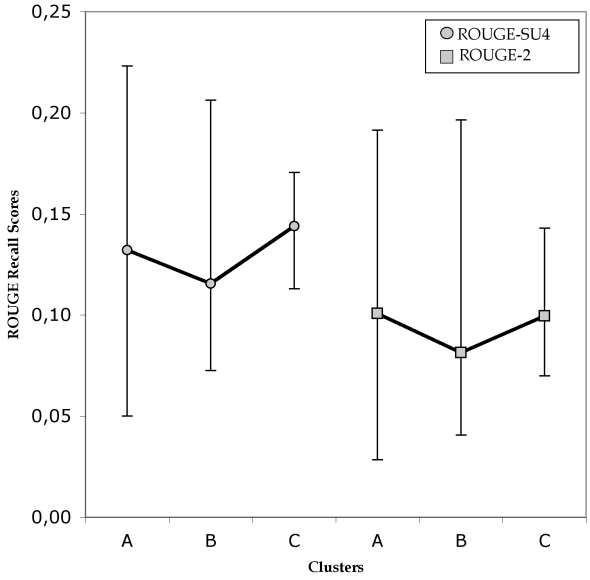

Fig. 8: *ROUGE recall scores (average and maximum - minimum deviations) for each document clusters (A∼10, B∼8 and C∼7 articles).*

this system ranks first in all the automatic and manual evaluations. Our system definitely is, in term of performance, in the pack leading group. We would like to say, in a word, that our system runs very fast, it only take ≈ 1 minute to compute the whole DUC 2007 update corpus on a 1.67Ghz G4 with 1.5Gb of RAM running MAC OS X 10.4.9.

7 Discussion and applications

We have presented a cosine maximization - minimization technique for producing user-oriented update summaries. This summarization system achieves efficient performances in the Document Understanding Conference 2007 evaluation regarding to other participants: 4th in ROUGE-2 average recall and Basic Element average recall, 5th in ROUGE-SU4 average recall and 7th in responsiveness in relation to 24 participants. The results of our experiments point out several research questions and directions for future work. The detection of the newness of information in the document clusters introduces too much "noise" in the summaries, considering only the most relevant sentences for the term extraction have to enhance the responsiveness. We are currently working on a more precise similarity maximization in the redundancy removal process by changing the granularity (using the sentence instead of the whole summary). Applications to a domain of speciality, the Organic Chemistry, is currently in development with a Chemistry textbook questionning system. This system will allow users to spare time by reading only new facts and skip all already known informations.

Acknowledgment

This work was partially supported by the *Laboratoire de chimie organique de synthèse*, FUNDP (*Facultés Universitaires Notre-Dame de la Paix*), Namur, Belgium.

References

[1] F. Boudin and J. Torres-Moreno. NEO-CORTEX: A Performant User-Oriented Multi-Document Summarization System. In *Computational Linguistics and Intelligent Text Processing*, pages 551–562. CICLing, 2007.

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[3] E. Hovy, C. Lin, L. Zhou, and J. Fukumoto. Automated Summarization Evaluation with Basic Elements. *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC)*, 2006.

[4] C. Lin. Rouge: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26, 2004.

[5] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[6] I. Mani and M. Maybury. *Advances in Automatic Text Summarization*. MIT Press, 1999.

[7] I. Mani and G. Wilson. Robust temporal processing of news. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76, 2000.

[8] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[9] K. McKeown and D. Radev. Generating summaries of multiple news articles. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, 1995.

[10] E. Newman, W. Doran, N. Stokes, J. Carthy, and J. Dunnion. Comparing redundancy removal techniques for multi-document summarisation. *Proceedings of STAIRS*, pages 223–228, 2004.

[11] M. Porter. An algorithm for suffix stripping, 1980.

[12] G. Rath, A. Resnick, and T. Savage. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143, 1961.

[13] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[14] R. Swan and J. Allan. Automatic generation of overview timelines. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, 2000.

[15] C. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann Newton, MA, USA, 1979.

Appendix

This is an example of our system output for the topic D0726F of the DUC 2007 task. The title is *"Al Gore's 2000 Presidential campaign"* and the narrative part is *"Give the highlights of Al Gore's 2000 Presidential campaign from the time he decided to run for president until the votes were counted."*.

UPDATE DOCSUBSET="D0726F-A"

Vice President Al Gore's 2000 campaign has appointed a campaign pro with local Washington connections as its political director. Al Gore, criticized for not having enough women in his inner circle, has hired a veteran female strategist to be his deputy campaign manager for his 2000 presidential bid. Al Gore will take his first formal step toward running for president in 2000 by notifying the Federal Election Commission that he has formed a campaign organization, aides to the vice president said. Al Gore took his presidential campaign to a living room that helped launch Carter and Clinton into the White House.

UPDATE DOCSUBSET="D0726F-B"

Patrick Kennedy, D-R.I., endorsed Vice President Al Gore for the Democratic presidential nomination in 2000. Al Gore named a veteran of the Clinton-Gore presidential campaigns to be his campaign press secretary. Bradley retired from the Senate in 1996, briefly mulled an independent run for president, then spent time lecturing at Stanford University in California before deciding to challenge Gore for the Democratic presidential nomination. Klain was criticized by some Gore allies after President Clinton called a reporter for The New York Times and said Gore needed to loosen up on the campaign trail. Bill Bradley of New Jersey, Gore's sole competitor.

UPDATE DOCSUBSET="D0726F-C"

After hearing that Stamford-native Lieberman had been chosen as Al Gore's running mate, Marsha Greenberg decided to knit him a gift. Vice President Al Gore, who continues to reshuffle his struggling presidential campaign, has selected Donna Brazile to be his new campaign manager, officials said. Al Gore declared "a new day" in his presidential bid with a symbolic homecoming and the opening of a new campaign headquarters far from the constant political intrigue and daily odds-making of Washington. Coelho, Brazile and Carter Eskew, the media consultant hired to help develop Gore's campaign message, are already working out of the Nashville office.