The LIA summarization system at DUC-2007

Florian Boudin[♭], **Fréderic Béchet**[♭], **Marc El-Bèze**[♭], **Benoît Favre**[♭], **Laurent Gillard**[♭], **Juan-Manuel Torres-Moreno**[♭,♯]
[♭] LIA - UPRES n° 931, Université d'Avignon et des Pays de Vaucluse, France
[♯] École Polytechnique de Montréal, Montréal (Québec), Canada
{florian.boudin, frederic.bechet, marc.elbeze}@univ-avignon.fr
{benoit.favre, laurent.guillard, juan-manuel.torres}@univ-avignon.fr

Abstract

This paper presents the LIA[1] summarization systems participating to DUC 2007. This is the second participation of the LIA at DUC and we will discuss our systems in both main and update tasks. The system proposed for the main task is the combination of seven different sentence selection systems. The fusion of the system outputs is made with a weighted graph where the cost functions integrate the votes of each system. The final summary corresponds to the best path in this graph. Our experiments corroborate the results we obtained at DUC 2006, the fusion of the multiple systems always outperforms the best system alone. The update task introduces a new kind of summarization, the over the time update summarization. We propose a cosine maximization-minimization approach. Our system relies on two main concepts. The first one is the cross summary redundancy removal which tempt to limit the redundancy between the update summary and the previous ones. The second concept is the novelty detection in a cluster of documents. In the DUC 2007 main and update evaluations, our systems obtained very good results in both automatic and human evaluations.

[1]Laboratoire Informatique d'Avignon, http://www.lia.univ-avignon.fr

1 Introduction

The 2007 Document Understanding Conference (DUC) organized by NIST have introduced a new pilot task besides the main real-world complex question answering task. The pilot task is to produce short (100 words maximum) multi-document up-to-date summaries of newswire articles, the time notion is added to the task by using different document clusters representing the corpus at different times. This is the second participation of LIA to the DUC workshop, the system we use for the main task is obviously the same as for 2006, but enhanced by adding more summarizers. The main originality of the LIA system is its use of a fusion process for combining the outputs of different summarization systems developed by our team and based on widely different sentence selection algorithms. In the framework of the main task, we will try to see if the 2006 results on system fusion are confirmed using 2007 data (section 2). For the update task, we will focus on developing a simple yet efficient approach that can be used as a base for further improvements (section 3).

2 Main task: query oriented multi-document summarization

For DUC 2007, we kept the same approach as for DUC 2006: generating a summary combining the outputs of multiple systems. Following the extractive summarization paradigm, each summarizer generates a list of sentences ranked according to the user query. Then, a fusion process builds a sentence selection that fits the 250 word limit while reflecting

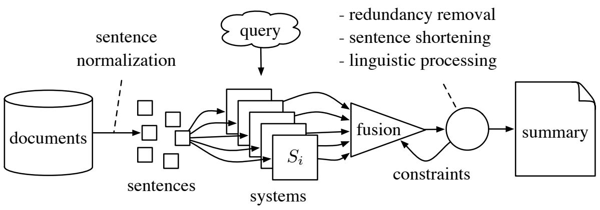

Figure 1: Main scheme of the fusion of multiple sentence selection systems.

ranked lists agreement. Linguistic quality and non redundancy are improved in a post-processing step that back-propagates constraints to the fusion. Figure 1 illustrates the whole process. Details on the different modules can be found in our last year paper (Favre et al., 2006). Our approach has proved in 2006 that the fusion would improve individual system performance and that involving more systems (with decent performance) would limit over-training risk on previous years data.

2.1 System description

The pre-processing step includes tokenization, word normalization, recapitalization of proper names, decapitalization of sentence head-words and link-word removal. The sentences are used as a common input of the different systems. Text segmentation in sentences now take heed of document structures to better detect sentence boundaries. In our implementation, sentence selection is mainly based on information retrieval and unsupervised summarization models. This year, we have gathered seven systems that focus on various models:

(S1) MMR-LSA: Maximal Marginal Relevance (Goldstein et al., 2000) using similarity between sentences in a Latent Semantic space (reduced co-occurrence matrix). User query is interpolated with global information distribution.

(S2) NEO-CORTEX: many statistical features are combined using an optimality criterion. This system is described in (Boudin and Torres-Moreno, 2007). It includes informations about the document set as a whole and rescales sentence scores according to their relevance in individual documents.

(S3) Variable length insertion gap n-term model: topic words, lemma and stems and aligned to sentences to compute a coverage rate. This score is scaled with sentence position information.

(S4) Vector Space Model (Buckley et al., 1996): similarity between a sentence and the topic is computed using the LNU*LTC metric.

(S5) Okapi similarity (Robertson et al., 1996).

(S6) Prosit similarity (Amati and Van Rijsbergen, 2002).

(S7) Compactness score: developed for the answer extraction component of the LIA Question Answering System (Gillard et al., 2006), the main idea is that density and closeness of important words found in a question can help to extract the best answer candidate. It allows to score each sentences based on the closeness and density of the important words of the topic that appeared inside the sentence.

Systems S1, S2, S3 and S7 are very similar to the ones used in 2006. S4, S5 and S6 are quick implementations of retrieval models to ensure diversity in the fusion process. These models follow the description in (Savoy and Abdou, 2006) using similar parameters. Stemming and stop-word stripping have been applied. The 7 systems generate ranked sentence lists according to the user query that will be merged in a fusion process.

For the fusion, we build a sentence graph with every valid summary (approximatly 250 words) from the ranked lists. Sentences are weighted according to their ranks and scores from the systems. Heuristics have been integrated to limit relative references (pronouns, time...) and reflect constraints of the post-processing.

Post-processing include a person name rewriter and an acronym rewriter. The first occurrence of acronyms and persons use full forms while next occurrences are replaced by shorted forms. Post-processing also includes simple redundancy removal using a textual inference baseline: word-overlap. Sentences bringing less than a percentage of new content words to the summary are considered redundant and marked. Sentences that contain long

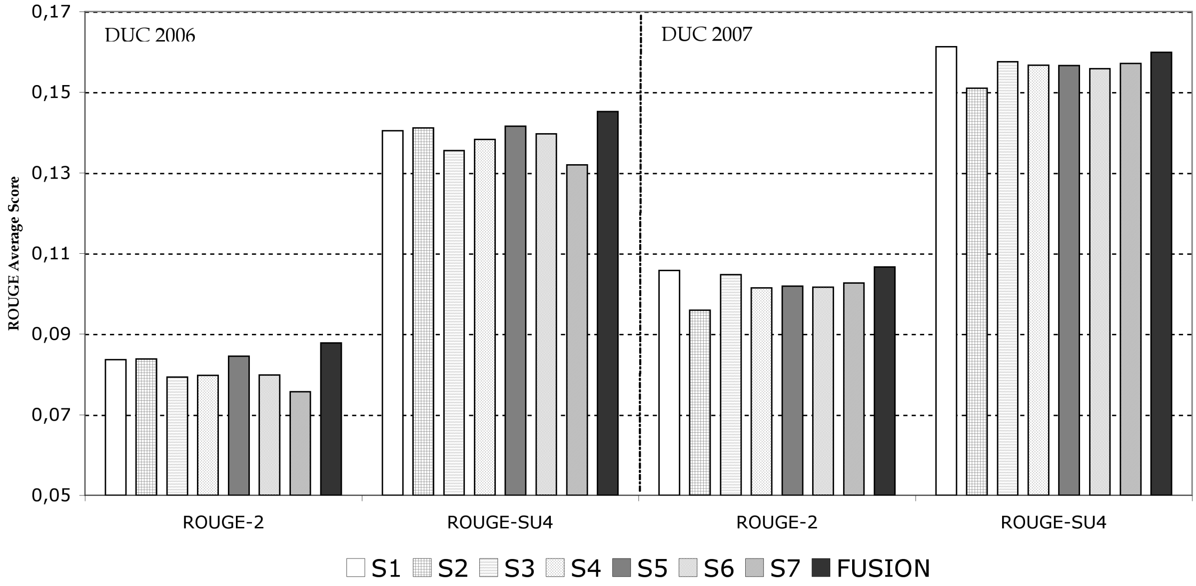

Figure 2: Recall ROUGE-2 and SU4 results for the LIA seven systems and the fusion on DUC 2006 and 2007 corpora.

forms of acronyms or person names are also marked. New sentence lengths, redundancy and rewriting constraints are back-propagated to a second pass of fusion to generate the final summary.

2.2 Results

Figure 2 shows the ROUGE scores obtained by our seven systems on DUC 2006 and DUC 2007 corpora. The fusion of all systems is also displayed. The fusion process always improves the scores over ther best system alone. These results corroborate the fact that the combination of several systems outperforms the best system and prevents overfitting on the training corpora. In other words, assembling very different sentence selection algorithms is a good strategy. Indeed, the reliability of our systems is low. We can observe that S2 was very performant in DUC 2006 but in DUC 2007 was the worst system. The fusion strategy allows to overcome these kind of stability issues.

The rest of this section presents the results obtained by our system (id is 3) at the DUC 2007 main evaluation. Among the 30 participants, our system ranks 9^{th} in ROUGE-2 and 11^{th} in Basic Elements evaluation, 8^{th} in ROUGE-SU4 evaluation and 8^{th} in manual evaluation. Figure 7 shows the position of our system in the ROUGE automatic evaluations comparing to the other 29 participants and the two baselines (ids are 1 and 2). For ROUGE-2, our system scored 0.106 where the mean of all systems was 0.0948 with standard deviation of 0.0188. For ROUGE-SU4 our system scored 0.159, which is above the mean of all systems that was 0.1474 with standard deviation of 0.021.

Figure 4 refers to the average content Responsiveness score. This score is an integer between 1 (very poor) and 5 (very good) and is based on the amount of information in the summary that satisfies the user information need. The average responsiveness score obtained by our system was 2.933, which is above the mean (2.61 with standard deviation of 0.462).

In term of linguistic quality, our system has good performances given the fact that we didn't improve linguistic processing. The 2007 results (presented in table 1) are similar to the 2006 results with a bad score in linguistic non-redundancy (use of pronouns and form diversity) and average score in other areas. The non-redundancy accident is characteristic of the voting process that selects highly relevant sentences that usually contain the most frequent named entities. A post-processing step to detect and reformulate these entities should improve the non-redundancy score.

To conclude about the main task, we have re-

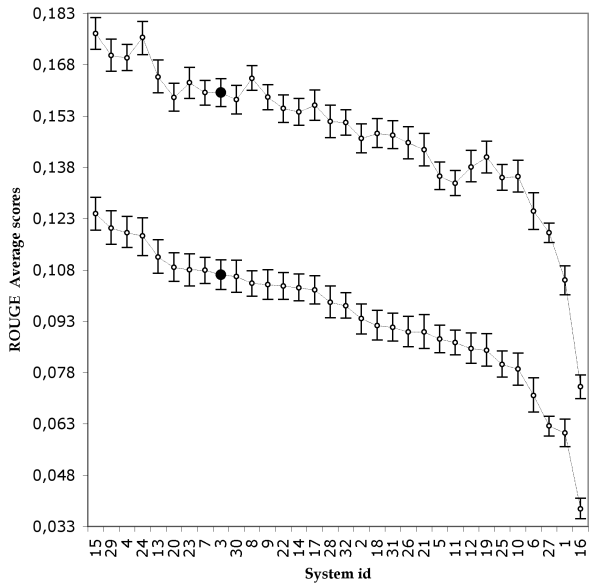

Figure 3: Recall ROUGE-2 and SU4 for the 32 systems at DUC 2007. Our system id is 3 (marked in the figure by a dark square), the systems 1 and 2 are two baselines.

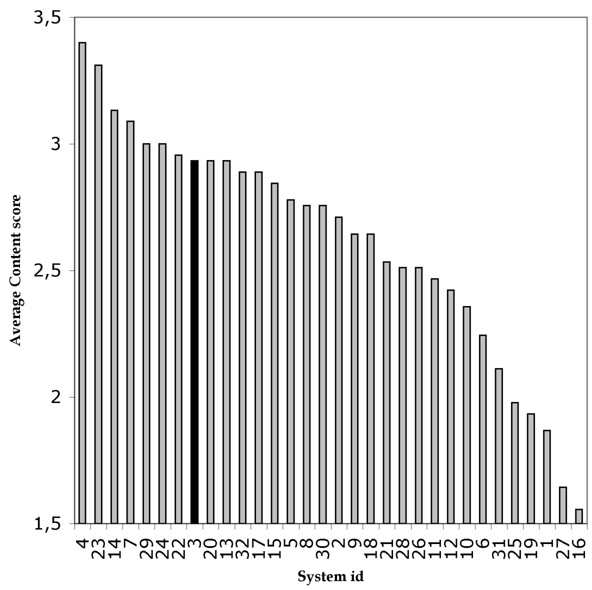

Figure 4: Average content responsiveness score of the 32 systems at DUC 2007. Our system id is 3 (marked in the figure by the dark bar).

produced 2006 results and maintained our overall rank in both responsiveness and linguistic quality. This confirms the benefits of the fusion in preventing over-fitting and coping with variable system performance.

Measure	2006	2007
Mean linguistic quality	3.57	3.42
Grammaticality	4.08	4.11
Non-Redundancy	3.84	3.62
Reference Clarity	3.42	3.36
Focus	3.74	3.56
Structure and Coherency	2.76	2.47

Table 1: Linguistic quality scores of our submission in 2006 and 2007. There is a small difference between the two evaluations since we didn't change the linguistic processing module.

3 The update task: A Cosine Maximization-Minimization approach

The update task of DUC 2007 is to produce short summaries (maximum 100 words) from three small document clusters representing the corpus through the time. The summarizer has to take into account the already read newswire articles and remove their information from the candidate summary. We propose a statistical method based on a maximization-minimization of simple similarity measures using the vector space model. The main motivation of this method is to minimize the redundancy of information between the different time summaries and in the same time maximize the accuracy of the information in relation to the given topic. A simple cosine similarity (Salton, 1989) scoring method is used as the core system to produce sentence scores (defined in formula 1), each sentence is compared to the topic using a cosinus similarity between the two vectors (the term weighting used is the well known $tf \times idf$ (McGill and Salton, 1983)). As the sentence of the whole cluster are scored according to the same topic, inter-sentence redundancy within the summary is an important problem. Thus, a sentence is added to the final summary only if all the cosine scores compared to the other sentences are lower than a threshold τ. Figure 5 give a global overview of the main architecture of our system. The following subsections formally define the measures and the methods that we have implemented in our summarizer.

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

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

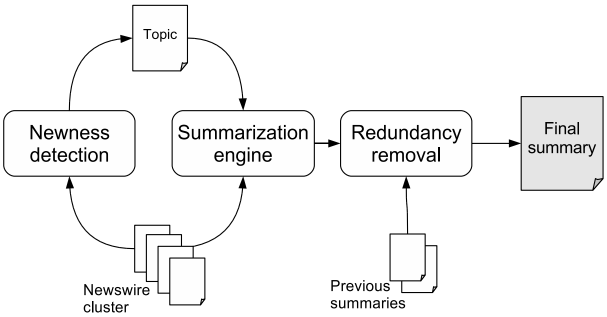

Figure 5: General architecture of the update summarization system.

3.1 Cross summary redondancy removal

The information about a particular topic contained by the candidate summary have to be different than the information of the previous time summaries and inform the reader of new facts. We propose to modify the sentence scoring method and minimize the score of sentences sharing duplicate information with previous summaries. 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 a termspace Ξ of dimension N (number of unique terms in the corpus). 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 $R(\bullet)$:

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

Where:

$$\begin{aligned} \phi(\vec{s}, \Pi) &= \sqrt{\textstyle\sum_{i=1}^{n_p} \cos(\vec{s}, \vec{p_i})^2} \\ \eta(s, \vec{t}) &= \cos(\vec{s}, \vec{t}); \;\; 0 \leq \eta(\bullet) \; ; \; \phi(\bullet) \leq 1 \end{aligned}$$

Therefore:

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

The highest scored sentence $\vec{s}$ is the most relevant assuming the topic $\vec{t}$ (i.e. $\eta \rightarrow 1$) and, simultaneously, the most different assuming the previous summaries Π (i.e $\phi \rightarrow 0$). Our approach is based on the principles that Maximal Marginal Relevance (MMR) but differs from these in several ways. Our system relies on the simple idea that a candidate sentence is high relevant if it is both relevant to the query and contains minimal similarity to previously produced summaries. Thus, the similarity is calculated between the sentence, the query and previous summaries instead of the query and previous sentences (the granularity is changed). Our scoring method remains much simple than MMR and adapted to this particular task.

3.2 Novelty boosting

The problem of the scoring method is that all scores are evaluated in relation to a particular topic. Removing information redundancy by the previously defined technique forces irrelevant sentences to enter the summary. To provide relief to the scoring method, we propose to enrich the topic with the most relevant information of the document cluster. In the same way as several works in document clustering use a list of high weighted terms as topic descriptors, we suggest to enrich the topic of a cluster X at time t_0 with a bag of words B_X of the n_t high unique weighted terms present in cluster X and not in clusters at time $t_i < t_0; i = 1, \cdots, n_p$. In other words, the novelty 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 (4) with all the previous cluster's bags of words.

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

The system uses the terms of B_X to enrich the topic t of the cluster X, the topic is extended by a small part of the unique information contained in the cluster so as to focus selected sentences not only on the topic but also to represent the unique information of the cluster.

3.3 Summary construction

The final summary is constructed by arranging the highest score sentences until the limit size of 100 words is reached, as a consequence the last sentence have a very high probability to be truncated. A best fit method (try to output exactly 100 words) may lead to poor results as there are not so many candidate sentences brining new information to the summary. 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 sentence after the last one is preferred to the last one if it's length is almost 33% shorter and, to avoid noise, if it's score is greater than a threshold of $\tau = 0.15$. In all cases the last summary sentence is truncated of 3 words maximum. By this, we try to protect the sentence grammaticality and remove only stop-words and very high frequency words from the 3 remaining words. A set of about fifty re-writing patterns and a dictionary based name redundancy removal system have been specially created for the DUC update task.

3.4 Results

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 explains why the parameters used for the system submission are not optimal. The following parameters have been used for the final evaluation: Bag of words size = 15, Redundancy threshold $\tau = 0.4$, minimal sentence length = 5. Among the 24 participants, our system ranks 4^{th} in both ROUGE-2 and Basic Elements 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 SU4) of the systems and their average responsiveness scores. 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 relatively low responsiveness score (ranked only 7^{th}) is maybe due to the poor sentence post-processing. For average Basic Elements

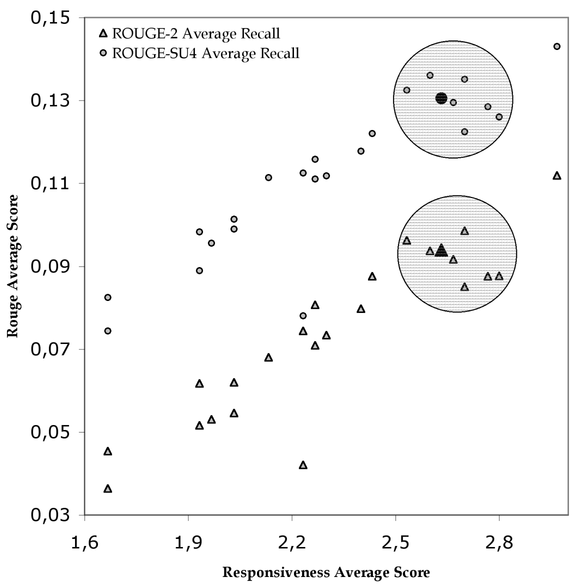

Figure 6: The correlation between ROUGE average recall scores and the responsiveness score for the 24 participants of the DUC 2007 update evaluation. Our system is represented by a dark circle (ROUGE-2) and a dark triangle (ROUGE-SU4).

(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.

Cluster	A	B	C
ROUGE-2	0.08170	0.08080	0.03670
ROUGE-SU4	0.08657	0.06826	0.02878

Table 2: Standard deviations of our system ROUGE scores in relation to the cluster used.

We observe in the figure 7 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 2). The stability of our system enhances with the quantity of previous documents, the slight decrease with the cluster B summaries may be due to the non-optimal enrichment done without enough previous extracted terms. After analysing all the figures, our system definitely is, in term of performance, in the pack leading group. To conclude about the update task, we have proposed a simple and fast system. This system models redundancy from previous knowledge and boosts

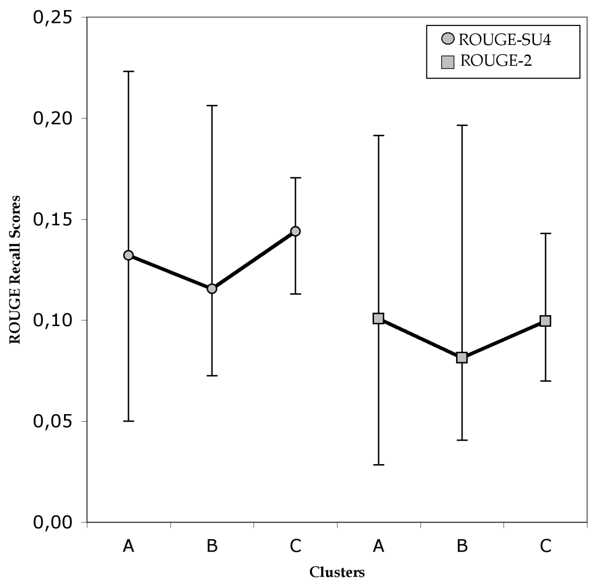

Figure 7: ROUGE recall scores (average and minimum-maximum) for each cluster of documents (A∼10, B∼8 and C∼7 articles).

new information using query expansion. To improve its performance, we will have to implement a better post-processing.

4 Discussion

In this paper, we have reported the LIA participation in DUC 2007. On the main summarization task, we extended our 2006 approach with new systems in the fusion process. The results confirmed that the fusion brings more stability and reduces the overfitting risk.We also participated to the new update task in which we had to include knowledge of previously seen documents as redundancy. We proposed a simple approach that appeared to be quite successful. The approach simultaneously selects sentences similar to the topic while dissimilar to the already known information. Then, new information is boosted by expanding the topic with words appearing only in the new documents. Future work include the porting of the fusion paradigm to the update task and implementation of sentence compression in this framework. In a more general way, the update task needs a specific evaluation of redundancy from the previous cluster (manual and/or automatic) as this is not reflected by current evaluation. In the long term, we push forward to evaluate speech summarization which is of great interest for us. Others applications are also possibles: we developed[2] a summarizer coupled with a QA system focused on Organic Chemistry. The system will allow user to spare time by reading only new facts and skip all already known information.

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

G. Amati and C.J. Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

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

C. Buckley, A. Singhal, M. Mitra, and G. Salton. 1996. New retrieval approaches using SMART: TREC 4. *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 25–48.

B. Favre, F. Béchet, P. Bellot, F. Boudin, M. El-Bèze, L. Gillard, G. Lapalme, and J.M. Torres-Moreno. 2006. The LIA-Thales summarization system at DUC-2006. *Proceedings of the 2006 DUC Workshop*. http://duc.nist.gov.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Ro ver. In *Proceedings of IEEE ASRU Workshop, Santa Barbara, USA*, pages 347–352.

L. Gillard, L. Sitbon, E. Blaudez, P. Bellot, and M. El-Bèze. 2006. The LIA at QA@CLEF-2006. *Working Notes of the CLEF 2006 Workshop*.

J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *ANLP/NAACL Workshop on Automatic Summarization*, page 4048.

M.J. McGill and G. Salton. 1983. *Introduction to modern information retrieval.* McGraw-Hill.

[2]In collaboration with the University of Namur, Belgium.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 1997. AT&T FSM Library - Finite State Machine Library. *AT&T Labs - Research*.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. In *Eurospeech 2005*.

S.E. Robertson, S. Walker, MM Beaulieu, M. Gatford, and A. Payne. 1996. Okapi at TREC-4. *Proceedings of the Fourth Text Retrieval Conference*, pages 73–97.

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

J. Savoy and S. Abdou. 2006. UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. *CLEF*.

Juan-Manuel Torres-Moreno, P. Velázquez-Morales, and J.G. Meunier. 2001. Cortex: un algorithme pour la condensation automatique de textes. *ARCo*, 2:365.