NEO-CORTEX: a performant user-oriented multi-document summarization system

Florian Boudin and Juan Manuel Torres Moreno

Laboratoire Informatique d'Avignon
BP 1228 F-84911 Avignon Cedex 9 FRANCE
{florian.boudin,juan-manuel.torres}@univ-avignon.fr
http://www.lia.univ-avignon.fr

Abstract. This paper discusses an approach to topic-oriented multi-document summarization. It investigates the effectiveness of using additional information about the document set as a whole, as well as individual documents. We present NEO-CORTEX, a multi-document summarization system based on the existing CORTEX system. Results are reported for experiments with a document base formed by the NIST DUC-2005 and DUC-2006 data. Our experiments have shown that NEO-CORTEX is an effective system and achieves good performance on topic-oriented multi-document summarization task.

Key words: Automatic summarization, Statistical methods, Text mining, Query guided summaries

1 Introduction

The Big companies, civil services and laboratories are confronted with a challenge: manage the mass of unstructured electronic textual documents. How to quickly find relevant information? How to display the information in a simple and fast way? The notion of automatic text summarization becomes one of the big subjects of Natural Language Processing (NLP). Rather than to diffuse whole documents, is it not preferable to diffuse only summaries containing the relevant information? Indeed, it is easier to read some lines than to read a huge number of pages to find out if the information wanted is there. In this paper, we present NEO-CORTEX, a system for summarizing multiple documents concerning a given topic. NEO-CORTEX was one of the five sentence selection systems used by the LIA-THALES system at the NIST Document Understanding Conference (DUC) 2006. This paper is organized as follows: Section 3 presents the overall system, Section 4 describes the adaptations made for the DUC 2006 task. In Section 5, we analyze the results of NEO-CORTEX system and Section 6 concludes and shows future work.

2 Background and related works

This paper describes an approach to topic-oriented multi-document summarization (MDS). Builds on previous work in single-document summarization (SDS), this approach uses additional information about the document set as a whole, as well as individual documents. Generating an effective summary requires the summarizer to select, evaluate, order and aggregate items of information according to their relevance to a particular subject or purpose [1, 2]. Introduced by Luhn [3] at the end of the fifties with the text-span deletion summarization system, automatic summarization is a process to transform source texts into a reduced target text in which the relevant information is preserved. Most of the works in sentence extraction applied statistical techniques (frequency analysis, overlap, etc.) to linguistic units such as terms, sentences, etc. Other approaches are based on the structure of the document (cue words, structural indicators) [4, 5], the combination of information extraction and language generation, machine learning [6, 7] to find patterns in text, lexical chains [8, 9] or Rhetorical Structure Theory (RST) [10]. Previous works showed that researchers have extended various aspects of SDS approaches to apply to MDS. Our approach is based on the same principle but differs from these in several ways. It attempts to use a topic-independent SDS based mainly on statistical processing and to generate a query-relevant summary.

3 System overview

The ***CO**ndensation et **R**ésumés de **TEX**tes* [11] (CORTEX) is a performant and language independent SDS system [11–13]. The challenge was it's adaptation to a user-oriented MDS by introducing new features. The idea of CORTEX is to represent the text in an appropriate space and apply numeric treatments. In order to reduce the complexity, some reductions and filtering preprocessing are applied. Deletion of stop-words, words in high and very low frequency, text in brackets, figures and symbols. Each word is replaced by the stemming form of it's lemma to maximize coverage of relevant terms. The stemming algorithm used was the Porter stemmer [14]. The choice of combining lemmatization and stemming (see table 1) was done to overcome the problem of an incomplete lemma database (i.e. not containing all inflected and derived forms of words).

Word	Lemma	Stem	Lemma + Stem
being	*be*	*be*	**be**
was	*be*	*wa*	**be**
natural	·	*natur*	**natur**

Table 1. Examples of lemmatization and stemming preprocessing. The third example shows the possible problem of incomplete lemma database (the word "natural" considered as non present in the lemma database).

The system uses an optimal decision algorithm that combines several metrics (up to $\mathit{\Gamma} = 13$ metrics [12]) resulting from processing statistical and informational algorithms to the document vector space representation (represented as a term/sentence matrix γ and a presence matrix ξ (1), only terms of frequency greater than two appears). The value $\gamma_{y,x}$ means 0 if the word x is in the sentence y and a positive value otherwise (can be boolean or frequency). N is the word set cardinality of the document and M is the sentence number.

$$\gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,N} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{M,1} & \gamma_{M,2} & \cdots & \gamma_{M,N} \end{pmatrix}; \qquad \xi_{j,m} = \left\{ \begin{matrix} 1 & \text{if } \gamma_{j,m} \text{ exists} \\ 0 & \text{elsewhere} \end{matrix} \right\} \tag{1}$$

The decision algorithm relies on all the normalized metrics (between $[0,1]$) combined in a sophisticated way and calculates the score ($Score_s^{cortex}$) for each sentence s. Two averages are calculated: the positive tendency, that is $\lambda_s > 0.5$, and the negative tendency, for $\lambda_s < 0.5$ (the case $\lambda_s = 0.5$ is ignored [1]). To calculate this average, we always divide by the total number of metrics $\mathit{\Gamma}$ and not by the number of "positive" or "negative" elements (real average of the tendencies). So, by dividing by $\mathit{\Gamma}$, we have developed an algorithm more decisive than the simple average and even more realistic than the real average of the tendencies. Here is the decision algorithm that allows to include the vote of each metrics:

$$\sum^{s} \alpha = \sum_{v=1}^{\mathit{\Gamma}} (||\lambda_s^v|| - 0.5); \quad ||\lambda_s^v|| > 0.5 \tag{2}$$

$$\sum^{s} \beta = \sum_{v=1}^{\mathit{\Gamma}} (0.5 - ||\lambda_s^v||); \quad ||\lambda_s^v|| < 0.5 \tag{3}$$

v is the index of the metrics, $\sum_s^{\mathit{\Gamma}}$ is the sum of the absolute differences between $\|\lambda\|$ and 0.5, $\sum^s \alpha$ are the "positive" normalized metrics, $\sum^s \beta$ the negative normalized metrics and $\mathit{\Gamma}$ the number of metrics used. The value attributed to every sentence is calculated in the following way:

$$\begin{aligned} &\text{if } (\textstyle\sum^s \alpha > \sum^s \beta) \\ &\text{then } Score_s^{cortex} = 0.5 + \textstyle\sum^s \alpha/\mathit{\Gamma} : \text{retain the sentence } s \\ &\text{else } Score_s^{cortex} = 0.5 - \textstyle\sum^s \beta/\mathit{\Gamma} : \text{not retain the sentence } s \end{aligned} \tag{4}$$

$\mathit{\Lambda}^s$ is the value used for the final decision whether or not to retain the sentence s. In the end, N_P sentences are sorted according to this value $\mathit{\Lambda}^s; s = 1, \cdots, N_P$.

[1] Simple average may be ambiguous if the value is close to 0.5, but the decision algorithm eliminates the sentences that their score is 0.5.

In order to summarize multiple documents, we have introduced two new parameters. A global parameter, the similarity between a document and the topic and a local parameter, the word overlap between a sentence and the topic.

3.1 Similarity

The CORTEX scores of each sentences are calculated for a single document, the score scale must be normalized to take into account the relevance degree of each document to the topic. Indeed, a relevant sentence of a document can have a lower score than a non relevant sentence of another document. This is due to the inter-document independency of the scores calculated by CORTEX. The similarity parameter (5) is a cosine similarity [15] and allows us to compute the similarity of two vectors, which are in our case the whole document $\boldsymbol{\nu_d} = (\nu^1, \nu^2, \cdots, \nu^n)$, $d = 1 \cdots Nb_{doc}$; Nb_{doc} is the total number of document and the topic $\boldsymbol{\omega_t} = (\omega^1, \omega^2, \cdots, \omega^n)$, $t = 1 \cdots \tau$; τ is the total number of topics. The dimension n is the number of different terms contained in the document and the topic. Similarity is then calculated as:

$$Sim(\boldsymbol{\nu_d}, \boldsymbol{\omega_t}) = \frac{\sum \boldsymbol{\nu_d} . \boldsymbol{\omega_t}}{\sqrt{\sum \boldsymbol{\nu_d}^2 + \sum \boldsymbol{\omega_t}^2}} \tag{5}$$

We use the $tf.idf$ [16] measure (term frequency, inverse document frequency) to obtain the weight of a term. This weight is a statistical measure used to evaluate how important a term is to a document. The importance increases proportionally to the number of times a term appears in the document but is offset by how common the term is in all of the documents in the collection. The *idf* measure was computed on the whole DUC document collection[2].

$$tf.idf_{\boldsymbol{\nu_d},j} = tf_{\boldsymbol{\nu_d},j} \times \log \left(\frac{Nb_{doc}}{n_j} \right) \tag{6}$$

$tf_{\boldsymbol{\nu_d},j}$ is the frequency of the term j in the document $\boldsymbol{\nu_d}$, n_j is the number of documents in which the term j is present. Similarity values are normalized in $[0, 1]$.

3.2 Overlap

We have introduced this measure believing that the selected sentences have to share the same information as the topic. In order to quantify the shared information, we have chosen the number of common words between the topic and a sentence s. The Overlap, calculated for each sentence, is the normalized cardinality of the intersection between the sentence word set S and the topic word set T. This measure forces high ranking for sentences containing topic

[2] See section 4 for more informations about the DUC Conference.

words and overcome the problem of high ranked sentence not containing any word of the topic.

$$Overlap(s, \boldsymbol{\omega_t}) = \frac{card\{S \bigcap T\}}{card\{T\}} \tag{7}$$

$card\{\bullet\}$ represents the cardinality of the set $\{\bullet\}$. $s = 1 \cdots N_L$, N_L is the total number of sentences. The Overlap values are normalized in $[0, 1]$.

3.3 Final sentence ranking

Similarity and Overlap parameters are used to refine the CORTEX scores. The final *Score* of a sentence s of a document $\boldsymbol{\nu_d}$ and a topic $\boldsymbol{\omega_t}$ is the linear combination:

$$Score = \alpha_0 \cdot CORTEX(s, \boldsymbol{\nu_d}) + \alpha_1 \cdot Overlap(s, \boldsymbol{\omega_t}) + \alpha_2 \cdot Sim(\boldsymbol{\nu_d}, \boldsymbol{\omega_t}) \ ; \quad \sum_i \alpha_i = 1 \tag{8}$$

The α_i values are empirical weights associated with the intermediate scores [3] of a sentence. The summary is generated with the Λ sentences of high score. Λ is fixed by the user, it can be a ratio of the initial size of all documents or a fixed number of sentences.

The NEO-CORTEX system (see figure 1) is resulting from the application of Similarity and Overlap parameters over the CORTEX system.

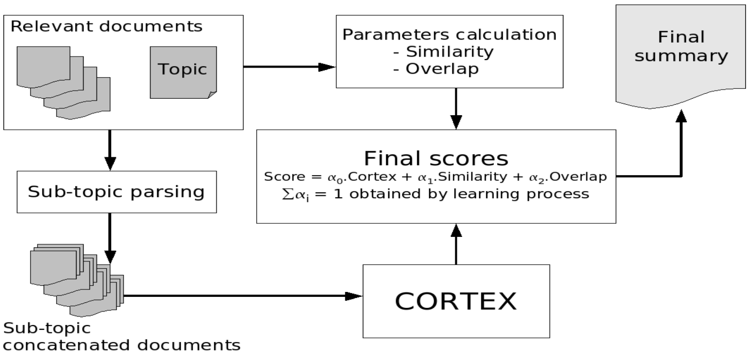
Relevant documents
Topic
Parameters calculation
- Similarity
- Overlap
Final summary
Final scores
Score = α_0.Cortex + α_1.Similarity + α_2.Overlap
$\sum\alpha_i$ = 1 obtained by learning process
Sub-topic parsing
CORTEX
Sub-topic concatenated documents

Fig. 1. General architecture of the NEO-CORTEX system, the process is applied for each couple of topic and relevant collection of documents.

[3] We called intermediate $CORTEX(s, \boldsymbol{\nu_d})$ the score calculated in formula (4), $Sim(.)$ the score calculated by (5) and $Overlap(.)$ the score calculated by (7).

3.4 Evaluating summary quality

The evaluation of the summaries is a difficult task, it can be achieved by evaluating independently the summary (intrinsic way) or by evaluating the summary in a specific task such as Question Answering (extrinsic way). The summaries are evaluated as either manually or semi-automatically. The first approach requires high human time cost (each summary has to be read, evaluated and appreciated) and is very subjective (divergence between judges can be considerable). The second approach is more standardized and has the ability to be exactly repeatable but requires human-produced reference documents. Several different approaches for semi-automatic evaluation exist such as Pyramid [17] or Basic Elements (BE) [18]. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [19] semi-automatic approach was used for our experiments. Two ROUGE Recall measures was computed for our evaluations, ROUGE-2 (bigram co-occurrence), ROUGE-SU4 (skip bigram with unigram and maximum gap length of 4) an Basic Elements BE. They are officially used and adopted for the DUC task. All the ROUGE results of this paper are obtained with light post processing and hard cut at 250 words.

3.5 Tuning the parameters for the DUC task

We have tuned the α_i parameters of NEO-CORTEX using the DUC 2005 dataset. In order to find the optimal repartition of Overlap in the final sentence score, we have settled the Similarity parameter to 0 and realized a precise scanning (in step of 0.05) by increasing the Overlap until we obtained the optimal repartition. The optimal ROUGE-2 score is obtained with $\alpha_1 \approx 0.4$ (see figure 2).

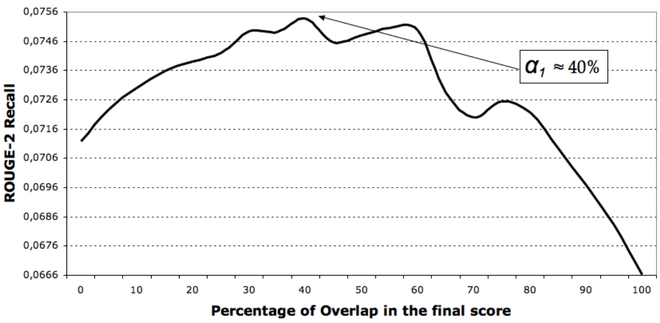

Fig. 2. NEO-CORTEX ROUGE-2 recall scores depending on the percentage of Overlap α_1 on the DUC 2005 dataset. The Similarity factor $\alpha_2 = 0$ and the $CORTEX$ factor $\alpha_0 = 1 - \alpha_1$ (Overlap). The optimal score is obtained with $\alpha_1 \approx 40\%$.

The optimal Similarity parameter α_2 is obtained by a similar way. The Overlap α_1 and the $CORTEX$ α_0 are settled to the previous optimal repartition

($\alpha_0 = 0.6$ and $\alpha_1 = 0.4$). The figure 3 shows two peaks (optimal values for α_2 parameter). As the DUC 2005 data-set is not enough important and in order to avoid errors due to the particularity of one corpus, we have empirically chosen the first peak, $\alpha_2 = 0.11$ (see figure 3) of the total repartition.

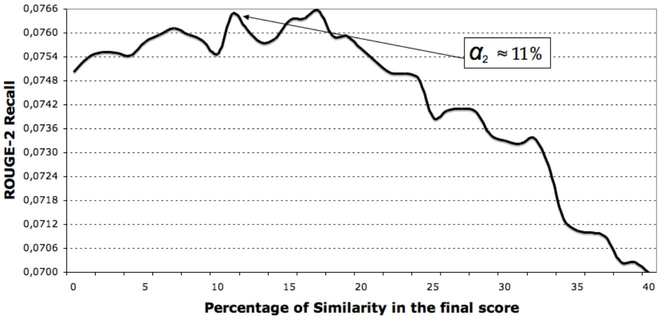

Fig. 3. NEO-CORTEX ROUGE-2 recall scores depending on the percentage of similarity α_2 on the DUC 2005 data-set. The Similarity factor $\alpha_2 = 1 - (\alpha_0(CORTEX) + \alpha_1(\text{Overlap}))$. The optimal score is obtained with $\alpha_2 \approx 11\%$.

Previous experiments showed that the Overlap is more important than the Similarity. This is why we have firstly tuned the Overlap parameter. We have normalized the α_i values and found the optimal repartition of the parameters for the DUC 2005 data-set: $\alpha_0(CORTEX) = 0.54$ (0.6 → 0.54) , α_1 (Overlap) = 0.36 (0.4 → 0.36) and α_2 (Similarity) = 0.10 (0.11 → 0.10). Further experiments confirmed that the parameters found are optimal.

4 Adaptations for DUC 2006 Task

The system task for DUC 2006 [4] is to model real-world complex Question Answering, in which a question cannot be answered by simply stating a name, date, quantity, etc. Given a topic and a set of 25 relevant documents [5], the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic statement.

4.1 Managing the topics

A topic is composed by two parts, the title and the narrative part (containing the questions). In the same way as a human would make, we have parsed the

[4] http://www-nlpir.nist.gov/projects/duc/index.html

[5] Documents source: AQUAINT corpus. Articles from the Associated Press and New York Times (1998-2000) and Xinhua News Agency (1996-2000).

topic to create sub-topics (see table 2). Indeed, each question of the narrative part needs to be answered, so we have chosen to create sub-topics that are the concatenation of the title and one of the topic's question of the narrative part. For each relevant document of the topic set, ζ document to be handle by CORTEX are created, ζ is the number of sub-topics.

Number and Title	**Question(s)**
D0603C Wetlands value and protection	Why are wetlands important? Where are they threatened? What steps are being taken to preserve them? What frustrations and setbacks have there been?
Sub-topic 1: wetland value protection important **Sub-topic 2:** wetland value protection threat **Sub-topic 3:** wetland value protection step preserve **Sub-topic 4:** wetland value protection frustration setback	
D0606F Impacts of global climate change	What are the most significant impacts said to result from global climate change?
Sub-topic 1: impact global climate change significant	

Table 2. Examples of DUC 2006 topics (D0603C, D0606F) and sub-topics resulting from their parsing (the sub-topics have been filtered and lemmatized).

4.2 Finding the best metrics for DUC 2006

The CORTEX system can use up to 13 metrics [12] to evaluate the sentence's pertinence, we have tested empirically a wide range of combinations and finally choose three metrics:

- Angle between a title and a sentence (A): Cosinus of the normalized scalar vector product between the sentence and the topic vector.

These two other metrics use a Hamming matrix H, a square matrix $N_L \times N_L$, in which every value $H[i,j]$ represents the number of sentences in which exactly one of the terms i or j is present.

$$H_{m,n} = \sum_{j=1}^{N_P} \left\{ \begin{array}{ll} 1 & \text{if } \xi_{j,m} \neq \xi_{j,n} \\ 0 & \text{elsewhere} \end{array} \right\} \quad \text{for} \quad \begin{array}{l} m \in [2, N_L] \\ n \in [1, m] \end{array} \tag{9}$$

The Hamming matrix is a lower triangular matrix where index m represents the line and index n the column, corresponding to the index of words ($m > n$). The main idea is that if two important words (may be synonyms) are in the same sentence, this sentence must certainly be important. The importance of every pair of words directly corresponds to the value in the Hamming matrix H.

- Hamming weight heavy (L): Among the sentences containing the same set of important words, how do we know which one is the best? i.e. wich one of these sentences is the more informative? The solution is to choose the one that contains the biggest part of the lexicon. Π = Sum of Hamming weight of words per segment × the number of different words in a sentence.
- Sum of Hamming weights of words by frequency (O): The sum of the Hamming weights of the words by frequencies uses the frequencies as factor instead of the presence. The sentences containing the most important words several times will be favored. O = The sum of the Hamming weights of the words × word frequencies.

We have tested a lot of metrics combinations as well as single metrics by trying to maximize the ROUGE measures (see figure 4). The other metrics [11] used in CORTEX system are: H for Perplexity; X for Sentence shape; B for partial $tf.idf$ (uses terms of frequency greater than one); F for Term Frequency (tf); P for Hamming weights of segments; D for Sum of probability frequencies; Y for Hamming distances; E for Entropy; T for Sum of Hamming weights of words per segments; I for Interactivity of segments.

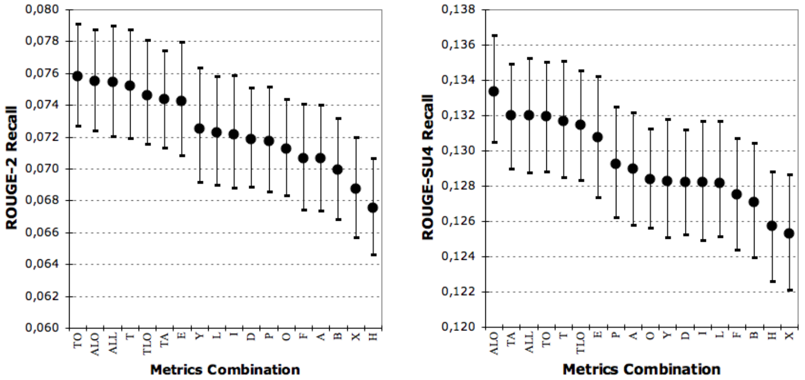

Fig. 4. NEO-CORTEX ROUGE scores in the DUC 2005 task depending on the metrics combination used. The ALL combination means all metrics of CORTEX.

4.3 Managing the sentence length

The summary word limit, for the DUC 2005 and 2006 tasks, is 250 words. The NEO-CORTEX system was not able to choose between two sentences of same score but with different lengths. Is a n or $10n$ words sentence important for short summary ? We have introduced a smoothing of the CORTEX scores depending of the sentence length by dynamically calculate, for each document, a gaussian. Further experiments showed that using a sigmoid based smoothing instead of the gaussian would improve significantly the ROUGE scores.

5 Results

In this section we will compare the ROUGE scores of NEO-CORTEX and CORTEX systems. We have compared the overall performances of NEO-CORTEX and CORTEX with the seven best ROUGE score metrics combinations on the DUC 2005 data-set. The ROUGE scores of all metrics combinations (see figure 5) are improved.

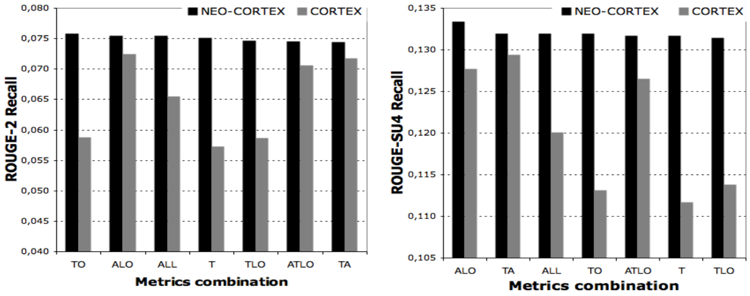

Fig. 5. ROUGE scores for NEO-CORTEX vs CORTEX in the DUC 2005 task.

The NEO-CORTEX system was also compared to the other participants of the DUC 2005 evaluation (see figure 6). Our system achieves very good performance (best system for all ROUGE scores). The fact is that the training data-set used for tuning NEO-CORTEX was the DUC 2005 data-set. NEO-CORTEX is optimally tuned for the DUC 2005 evaluation, this explain why it is very performant.

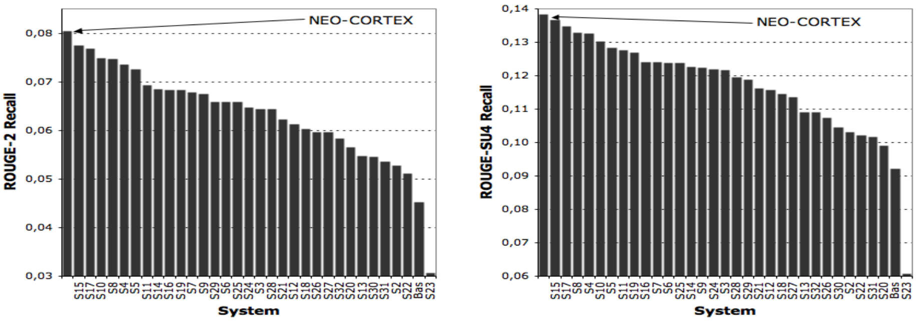

Fig. 6. ROUGE-2 recall (left) and ROUGE-SU4 recall (right) scores of NEO-CORTEX vs all participating systems in the DUC 2005 task.

In order to quantify the real perfomance of our system, we have also compared it to the participants of the DUC 2006 evaluation. The evaluation criteria in DUC 2006 remained same as DUC 2005, our summarization system performed well in the automatic evaluation (see figure 7).

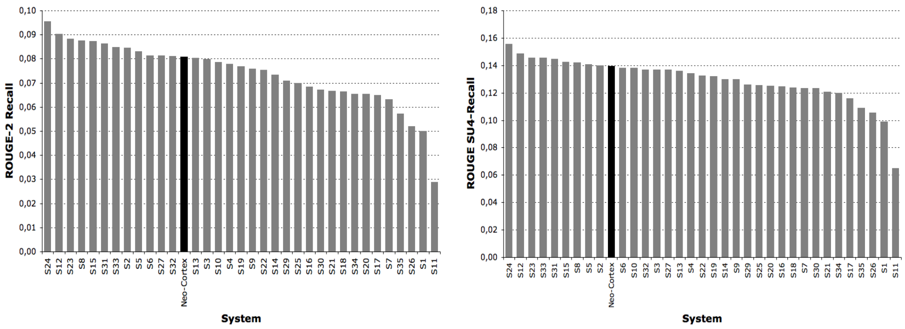

Fig. 7. ROUGE-2 recall (left) and ROUGE-SU4 recall (right) scores of NEO-CORTEX vs all participating systems in the DUC 2006 task. Neo-Cortex is ranked 13^{th} in ROUGE-2 and 10^{th} in ROUGE-SU4 over 35 systems.

6 Conclusion and future work

We have presented NEO-CORTEX, a multi-document summarization system based on the CORTEX system, and the participation in DUC 2006 task. Our experiments have shown that NEO-CORTEX is an effective system and achieves good performance on topic-oriented multi-document summarization task. NEO-CORTEX is however sensitive to the sentence segmentation, ROUGE scores have increased throughout our research time according to the segmentation quality enhancement. The ability of the system to be language independent is key point. Our participation in DUC 2006 was an excellent opportunity to evaluate the flexibility of the CORTEX system on a new and different task. In DUC 2006 the LIA-Thales fusion of five summarization systems among NEO-CORTEX, obtained very good results in the automatic evaluations (ranked 5th in ROUGE-SU4, 6th in ROUGE-2, 6th in BE and 6th in Pyramid) and achieved good performance in human evaluations (ranked 8th in the Resp-Overall) [20] . As always, there is a room for improvement and future work. In NEO-CORTEX, we would like to focus on improving our performance in metrics combinations, which we believe would enhance summaries quality. To that end, we are currently experimenting an incremental process, which in each step tries to find a different metrics combination. We would also like to use machine-learning to dynamically find the optimal α_i parameters of the sentence scoring and automatically adapt the system.

References

1. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. The MIT Press (1999)
2. Mani, I.: Automatic Summarization. John Benjamins Publishing company (2001)
3. Luhn, P.: Automatic creation of literature abstracts. IBM Journal of Research and Development (1958) 155–164
4. Edmundson, H.: New Methods in Automatic Extracting. Journal of the ACM (JACM) **16**(2) (1969) 264–285
5. Paice, C.D.: Constructing literature abstracts by computer: techniques and prospects. Inf. Process. Manage. **26**(1) (1990) 171–186
6. Mani, I., Bloedorn, E.: Machine Learning of Generic and User-Focused Summarization. Arxiv preprint cs (CL/9811006) (1998)
7. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (1995) 68–73
8. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization (1997) 10–17
9. Stairmand, M.: A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval. Unpublished PhD Thesis. UMIST Computational Linguistics Laboratory (1996)
10. Mann, W., Thompson, S.: Rhetorical Structure Theory: A Theory of Text Organization. University of Southern California, Information Sciences Institute (1987)
11. Torres-Moreno, J.M., Velázquez-Morales, P., Meunier, J.: Cortex: un algorithme pour la condensation automatique de textes. ARCo **2** (2001) 365
12. Torres-Moreno, J.M., Velázquez-Morales, P., Meunier, J.: Condensés de textes par des méthodes numériques. JADT **2** (2002) 723–734
13. Abdillahi, N., Nocera, P., Torres-Moreno, J.M.: Boîtes à outils TAL pour les langues peu informatisées: Le cas du somali. JADT (2006) 697 – 705
14. Porter, M.: An algorithm for suffix stripping. Program **14**(3) (1980) 130–137
15. Salton, G. In: Automatic text processing. Addison-Wesley Publishing Co. (1989)
16. Salton, G., McGill, M.: Introduction to modern information retrieval. Computer Science Series McGraw Hill Publishing Company (1983)
17. Passonneau, R., Nenkova, A., McKeown, K., Sigleman, S.: Applying the Pyramid Method in DUC 2005. Proc. of DUC 2005 at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP) (2005)
18. Hovy, E., Lin, C., Zhou, L.: Evaluating DUC 2005 using Basic Elements. Proc. of DUC 2005 at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP) (2005)
19. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Technical report, Information Sciences Institute (2002)
20. Favre, B., Béchet, F., Bellot, P., Boudin, F., El-Bèze, M., Gillard, L., Lapalme, G., Torres-Moreno, J.M.: The LIA-Thales summarization system at DUC-2006. http://www-nlpir.nist.gov/projects/duc/index.html (2006)