The LIA Update Summarization Systems at TAC-2008 (Draft)

Florian Boudin [♮] and **Marc El-Bèze** [♮]
[♮] Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228,
84911 Avignon Cedex 9, France.
florian.boudin@univ-avignon.fr
marc.elbeze@univ-avignon.fr

Juan-Manuel Torres-Moreno [♮,♭]
[♭] École Polytechnique de Montréal
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.
juan-manuel.torres@univ-avignon.fr

Abstract

For the third participation of the LIA to the DUC–TAC conferences, two summarizers were developed. The first is based on the SMMR sentence scoring algorithm described in (Boudin et al., 2008). The second summarizer is a fusion between two sentence scoring methods: SMMR and a variable length insertion gap n-term model (Favre et al., 2006; Boudin et al., 2007). We compare our two summarizers using the manual and automatic TAC's assessments. The fusion achieves better automatic scores but lower manual scores than the SMMR system alone. It is likely due to an overfitting problem owing to a small training corpus (DUC 2007 update).

1 Introduction

Recently emerged from the Document Understanding Conference (DUC) 2007, update summarization attempts to enhance summarization when more information about knowledge acquired by the user is available. It uses the fact that the user has already read documents about a particular topic and accordingly do not want to dispose of information about *old* facts. In this way, an important issue is introduced: redundancy with previously read documents (history) has to be removed from the extract.

The main originality of the LIA summarization system is its use of a fusion process for combining the outputs of two sentence scoring methods. These methods use different similarity measures between the topic and the sentences. After presenting these two systems, section 2 presents the fusion process, section 3 describes the linguistic post-processing, section 4 gives an overview of our results and section 5 concludes this paper.

2 Method

We define H to represent the previously read documents (history), Q to represent the query and s the candidate sentence. The following subsections formally define the two sentence scoring methods and the fusion strategy.

2.1 System 1: SMMR

Maximal Marginal Relevance (MMR) algorithm has been successfully used in query-oriented summarization (Ye et al., 2005). It strives to reduce redundancy while maintaining query relevance in selected sentences. The summary is constructed incrementally from a list of ranked sentences, at each iteration the sentence which maximizes MMR is chosen:

$$\text{MMR} = \arg\max_{s \in S} [\ \lambda \cdot Sim_1(s, Q) - (1 - \lambda)\ \cdot \max_{s_j \in E} Sim_2(s, s_j)\ ] \quad (1)$$

where S is the set of candidates sentences and E is the set of selected sentences. λ represents an interpolation coefficient between relevance and redundancy.

We propose an interpretation of MMR to tackle the update summarization issue. Since Sim_1 and Sim_2 are ranged in $[0, 1]$, they can be seen as probabilities even though they are not. Just as rewriting (1) as (NR stands for Novelty Rele-

vance):

$$\text{NR} = \arg\max_{s \in S} [\ \lambda \cdot Sim_1(s, Q) + (1 - \lambda)\ \cdot (1 - \max_{s_h \in H} Sim_2(s, s_h))\ ] \quad (2)$$

We can understand that (2) equates to an OR ($\vee$) combination. But as we are looking for a more intuitive AND ($\wedge$) and since the similarities are independent, we have to use the product combination. Sentences are scored thanks to a double maximization criterion in which the best ranked one will be the most relevant to the query AND the most different to the sentences in H.

$$\text{SMMR}(s) = Sim_1(s, Q) \cdot \left(1 - \max_{s_h \in H} Sim_2(s, s_h)\right)^{f(H)} \quad (3)$$

Decreasing λ in (1) with the length of the summary was suggested by Murray et al. (2005) and successfully used in the DUC 2005 by Hachey et al. (2005), thereby emphasizing the relevance at the outset but increasingly prioritizing redundancy removal as the process continues. Similarly, we propose to follow this assumption in SMMR using a function denoted f that as the amount of data in history increases, prioritize non-redundancy ($f(H) \rightarrow 0$). Details on this sentence scoring method can be found in (Boudin et al., 2008).

Parameter settings

Sim_1 is the well known $cosine$ angle measure and Sim_2 is a normalized Longest Common Substring (LCS) measure between sentences. Detecting sentence rehearsals, LCS is well adapted for redundancy removal. The fudge factor f is set to 1 for the cluster A and 0.5 for the cluster B.

2.2 System 2: Variable length insertion gap n-term model

This system relies on the simple idea that a term sequence found in a topic may be encountered in a document with some other words between the term members. By word term, we also mean *inflected forms*, lemmas or stems. From the topic, patterns are generated corresponding to three different models: the n-gram, the n-lemma and the n-stem. Pattern matching is then combined to other features to assign a score to each sentence. Details on this sentence scoring method can be found in (Boudin et al., 2007).

2.3 Fusing sentence scoring outputs

In the last two participations of our team to the DUC campaigns (Boudin et al., 2007; Favre et al., 2006), we have seen that fusing several summarizers prevent overfitting and outperform the best system alone. Although in a restrained way, we propose to follow this assumption by combining two summarizer outputs. Since each system uses different features and scoring functions to assign scores to sentences, combining scores linearly becomes hazardous because it depends on the value interval variation. Indeed, even if scores are commonly normalized in $[0, 1]$, value distribution is not homogeneous. One possible way to tackle this problem is to use ranks instead of scores. However, information contained in score deviations is lost. For example, once ordered, two consecutive sentences may have very different scores. This is the reason why we propose a method based on score deviations with the first rank ($\max$). One's complement of normalized score deviations with the first rank is used to assign scores. The score of a sentence s is given by:

$$score_{fusion}(s) = \alpha \cdot deviation_{S_1}(s) + (1 - \alpha) \cdot deviation_{S_2}(s) \quad (4)$$

with

$$deviation_{S_x}(s) = 1 - \left(\frac{\max - score_{S_x}(s)}{\max}\right)$$

Where α is a priority coefficient empirically tuned on the DUC 2007 update data that gives more weight to one or the other summarizer when it happens to achieve better results.

Parameter settings

3 Post-processing

3.1 Summary generation

Once sentences are selected to be assembled in the final summary, some linguistic treatments are ap-

Parameter	A	B
α	0.6	0.8

Table 1: Parameter settings of the fusion.

plied. Indeed, once out of their contexts, discursive forms are considerably decreasing summary's coherence. For example, two sentences one next to the other in the summary may be in opposition while not dealing with the same subject. Our rule based linguistic post-processing targeted sentence length reduction and coherency maximization. The process is composed by the following steps:

1. **Acronym rewriting**: first occurrence of an acronym is replaced by its complete form (acronym and definition), following ones only by their reduced forms. Definitions are automatically mined in the corpus by pattern matching.

2. **Date and number rewriting**: numbers are reformatted and dates are normalized to the US standard forms (MM/DD/YYYY, MM/YYYY and MM/DD).

3. **Temporal references rewriting**: time tags are used to replace fuzzy temporal references. For example "... *the end of next year, ...*" with temporal tag 1992_06_02 is replaced by "... *the end of 1993, ...*".

4. **Discursive form rewriting**: ambiguous discursive forms are deleted. For example "*But, it is ...*" is replaced by "*It is ...*".

5. Finally, say clauses and parenthesized content are removed and punctuation cleaned.

Sentences are ordered within the summary by original document order and temporal order of documents. Since these linguistic treatments are dependent to the sentence order and modify the sentence's length, several passes are required to generate the final summary.

3.2 Anaphora Resolution

In the summary generated, important phrases, probably containing anaphora, can be retained. For exemple in the summary (the numbers of sentences are in []):

[1] He said a study was carried out which indicated each airport would have to spend about 80 million dollars to accept the A380. [2] He said the figure should be even lower as none of the facilities will have to build a new runway.

In this case, summary's quality is poor because the pronoun "He" is unknown in this context. At this point, imagine that the sentence: *The cost will be relatively modest, according to* **Dick Marchi**, *an expert on infrastructure of airports*, that was not retained by the scoring algorithm, is before sentence [1]. The information of person [Dick_Marchi], can help to resolve the anaphora and the modified summary will be as following:

[1] Dick_Marchi said a study was carried out which indicated each airport would have to spend about 80 million dollars to accept the A380. [2] Dick_Marchi said the figure should be even lower as none of the facilities will have to build a new runway.

In order to increase the cohesion and linguistic quality of the summary, a resolution of anaphora has been implemented. Statistical approaches are suitable but they need large labeled ressources in order to learn probabilities (Ge et al., 1998). We developped a rule based algorithm to identify noun phrase antecedents of personal pronouns. We use the DUC-2007 pilot task documents as developpement corpus. Firstly, the summary is syntactically analysed using Treetagger (Schmid, 1995) (tool for annotating text with part-of-speech and lemma information). Secondly, terms with the lexical tags NN or NP are marked as anaphora candidates. The score of each candidate is computed as a fonction of distance to anaphoric reference. The most likely candidate is retained and the corresponding pronoun is replaced.

However, cohesion and linguistic quality does not mean automatically better ROUGE scores. Moreover, the anaphora could be wrongly resolved, making the summary incoherent. In fact, in our algorithm we avoid anaphoric resolution as post-processing.

4 Results

Table 2 shows the results obtained by our submissions at the update summarization task of TAC 2008. Our system achieved good results for Overall Responsiveness and Linguistic Quality but average ones for automatic evaluations. One interesting result is that the fusion achieves better automatic scores but lower manual scores than the system S_1 alone. This may be due to the fact that the fusion parameters were tuned by using automatic scores as reference.

Evaluation	Score (S_1)	Rank
Overall Resp.	2.32 (2.33)	23/58 (**-1**)
Linguistic Quality	2.56 (2.65)	16/58 (**-2**)
ROUGE-1	0.33831 (0.33611)	41/72 (**+1**)
ROUGE-2	0.07698 (0.07450)	32/72 (**+6**)
ROUGE-SU4	0.11634 (0.11581)	30/72 (**+2**)
Basic Elements	0.04792 (0.04574)	32/72 (**+3**)
Pyramids	0.254 (0.238)	26/58 (**+4**)

Table 2: Results of manual and automatic evaluations for the LIA system at the TAC 2008 update task. Results achieved by the system S_1 alone (SMMR) are shown in parenthesis.

Automatic scores for each method are often statistically indistinguishable from in the official evaluations considering the 95% confidence interval. However, enumerate systems that performs significantly better and lower than our approach can be done by studying confidence intervals from automatic evaluations. The table 3 shows these results for our system. Most of the scores achieved by our approach are above the average. At this point, it is worth noting that our approach is simple and do not uses any linguistic or knowledge resources.

5 Discussion

What we try and do not work:

- **Anaphora resolution**: Unfortunatly, results with anaphora resolution are disappointing. In fact, there are only few anaphoric pronouns in summaries, but they are very hard to resolve. As the resolution was wrong in most of cases, we decided to not include it in our final submission. A best algorithm for anaphoric resolution (with a deep analysis of sentences and context) must be used. A mix of linguistic and statistical approaches is actually under research and could be integrated.

- **Person Name rewriting**: We tried pattern matching approaches to automatically mine person names in the news articles. The very low precision of our approach makes us decide to not include this process in the final submission. Again, we are still working on this process by adding a POS-tagger.

Evaluation	Score $^{upper}_{lower}$	nb. $>$	nb. $<$
ROUGE-1	0.33831 $^{0.00564}_{0.00525}$	26 (**-7**)	27 (**=0**)
ROUGE-2	0.07698 $^{0.00425}_{0.00372}$	24 (**-6**)	30 (**+3**)
ROUGE-SU4	0.11634 $^{0.00336}_{0.00321}$	20 (**-11**)	31 (**+1**)
Basic Elements	0.04792 $^{0.00329}_{0.00312}$	21 (**-6**)	33 (**+6**)

Table 3: Automatic evaluations for our fusion system at the TAC 2008 update task with lower/upper limits for each score and the number of significantly better (nb. $>$) and lower (nb. $<$) systems. Difference with the system S_1 alone (SMMR) are shown in parenthesis.

References

Florian Boudin, Benoit Favre, Frédéric Béchet, Marc El-Bèze, Laurent Gillard, and Juan-Manuel Torres-Moreno. 2007. The LIA-Thales summarization system at DUC-2007. In *Document Understanding Conference (DUC)*, Rochester, USA, April.

Florian Boudin, Marc El-Bèze, and Juan-Manuel Torres-Moreno. 2008. A scalable MMR approach to sentence scoring for multi-document update summarization. In *Coling 2008: Companion volume: Posters and Demonstrations*, pages 21–24, Manchester, UK, August. Coling 2008 Organizing Committee.

Benoit Favre, Frédéric Béchet, Patrice Bellot, Florian Boudin, Marc El-Bèze, Laurent Gillard, Guy Lapalme, and Juan-Manuel Torres-Moreno. 2006. The LIA-Thales summarization system at DUC-2006. In *Document Understanding Conference (DUC)*, New York City, USA, June.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *In Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170.

B. Hachey, G. Murray, and D. Reitter. 2005. The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space. In *Document Understanding Conference (DUC)*, Vancouver, Canada, October.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive Summarization of Meeting Recordings. In *Ninth European Conference on Speech Communication and Technology*. ISCA.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

S. Ye, L. Qiu, T.S. Chua, and M.Y. Kan. 2005. NUS at DUC 2005: Understanding documents via concept links. In *Document Understanding Conference (DUC)*, Vancouver, Canada, October.