A Symbolic Summarizer with 2 Steps of Sentence Selection for TAC 2009

Pierre-Etienne Genest, Guy Lapalme
RALI-DIRO
Université de Montréal
P.O. Box 6128, Succ. Centre-Ville
Montréal, Québec
Canada, H3C 3J7

Luka Nerima, Eric Wehrli
Laboratoire d'Analyse et de
Technologie du Langage
Université de Genève
2, rue de Candolle
CH-1211 Genève 4, Switzerland

{genestpe,lapalme}@iro.umontreal.ca{Luka.Nerima,Eric.Wehrli}@lettres.unige.ch

Abstract

RALI developed the system NESS2 for the TAC 2009 summarization task, with the main goal of testing the hypothesis that performing sentence selection in *2 steps* improves the quality of the created summaries. The first step selects a number of top sentences, while the second step selects the best combination of the top scored sentences. We use the same small number of linguistic criteria to perform both evaluation steps. The two runs of NESS2, run IDs 8 and 10, ranked well in the competition, especially in linguistic quality and overall responsiveness, and the results show that the second step of sentence selection improves the performance of the system.

1 Introduction

For TAC 2009's summarization task, we developed a multi-document, topic-driven, update summarizer. The input documents were newswire articles from the collection AQUAINT-2 and they were guaranteed to be related to their given topic. The topics themselves represent "real-world questions" that the summaries should answer. Two clusters of 10 articles, referred to as *part A* and *part B*, were assigned to each topic and a 100-word summary was created for each part. Part B articles were more recent than part A articles, and the summary of the second cluster had to provide only an update about the topic, avoiding any repetition of information from the first cluster.

RALI proposes this year the second version of the NEws Symbolic Summarizer (NESS2), which is markedly different from last year's version (Genest et al., 2008). This year, our motivation was to test and quantify the effects of using a 2-step sentence selection scheme based on last year's best performing system of the TAC competition (Chen et al., 2008).

The first step consists of selecting the best 20 sentences from the cluster of articles, using a linear combination of 5 scoring criteria, 2 of which for update summaries only. The 20 top sentences are combined in all the possible was that they can form a 100-word or less summary, providing several *candidate summaries*. Then the second step uses the same criteria as before to select the best summary from all of the candidate summaries. The criteria include counting lemmas with high document frequency, counting lemmas in common with the topic, and sentence position. The update-only scoring criteria are the "new lemmas" from part B and the "lost lemmas" from part A; they serve to avoid repetition of known information. The analyses from the symbolic tagger FIPS are used to favor sentences in which important words appear in the subject or object positions.

We describe the approach for NESS2 in section 2. Section 3 presents and discusses the results that we obtained in the competition. The last section provides a conclusion.

2 Our Approach

2.1 Preprocessing

The preprocessing involves extracting the relevant information from the topic and the article files and performing some cleaning of the text provided to us. We also gather peripheral data that will be used during sentence evaluation, such as sentence segmentation, counting the document frequency of the words for each cluster, and obtaining sentence analyses that will be used to lemmatize the words and to identify the grammatical functions of important words when they appear.

For the topic, we only keep the `<title>` and `<narrative>`; and for the articles, we keep only the text. We extract the date of each article from the file names. All this information is kept in XML documents, one per cluster of articles to be summarized. Note that most of the information and modules of NESS2 are in XML and XSLT.

To make sure that the texts are compatible with our program's modules, we adjust the quotation marks, remove middle initials from names, and perform other low-level editing for convenience. The text of the articles is then segmented into sentences, for the purpose of extraction. Regular expressions are used to replace relative time references by the month and year in which the article containing the sentence was published. For example, "on Monday" could be replaced by "in January 2006" if the article in which the sentence with this expression appears was published in January 2006. There can obviously be mistakes if the news articles refer to events in previous or future months, but this seldom happens.

Finally, we gather information that will be useful later. FIPS (Wehrli, 2007) (Wehrli and Nerima, 2009), a robust multilingual symbolic parser and tagger based on generative grammar, is used to extract the lemmas of all the words in the topics and articles. FIPS is also used to tag grammatical functions in the articles. Finally, we compute the document frequency of each lemma, i.e. the number of documents of the cluster that include at least one instance of a lemma.

2.2 Evaluation Criteria

Five evaluation criteria are used in both sentence selection steps of our system, to compute a relevance score for sentences and candidate summaries respectively. To remain non-specific about whether we score sentences or candidate summaries, we refer to both as just a "candidate" in this section.

The scores for each criterion described below are normalized to values between 0 and 1 inclusively. The normalization is performed over all the sentences of a cluster of articles (a summarization instance) and separately over all the candidate summaries of a given summarization instance. The normalization factor of a criterion is the inverse of the highest score achieved by any candidate for that criterion.

The **high document frequency lemmas** score is computed by counting the number of distinct lemmas in the candidate, that have a document frequency of 6 or more. Document frequency is computed by counting the number of different articles of the cluster which contain at least one occurrence of the lemma.

The **topic similarity** score of the candidate is a count of the distinct topic lemmas that it contains.

The **sentence position** score is the only criterion that is not computed in the same way for either types of candidate. Sentences are given a score of 1 for the sentence position criterion if it appears first in the article that contains it, and 0 otherwise. Candidate summaries are given a score based on the ratio of its sentences that are in first position in their article.

The **new lemmas** score is used only for update summaries and relies on comparing the lists of lemmas with a high DF of 6 or more in parts A and B. The new lemmas score of a candidate is a count of its new lemmas that appear in cluster B's list of lemmas with high DF but not in cluster A's list.

Conversely, the **lost lemmas** score of a candidate is a count of its lost lemmas that have high DF in part A but that do not have a high DF in part B. The lost lemmas score is subtracted rather than added, thus decreasing the global score of a candidate.

For the scores of high DF lemmas, topic similarity, new lemmas and lost lemmas, the lemmas that play the grammatical function of subject or object at least once within the candidate are counted three times instead of once. Thus, candidates with important words playing important syntactic roles in its sentence or sentences have an increased score. This means that a sentence with its subject being a topic

lemma, for example, will receive a better topic similarity score than one containing two topic lemmas that do not have a grammatical function of subject or object.

2.3 First Sentence Selection Step

The first sentence selection step is usually the only step in standard extractive summarizers. It consists of evaluating all the sentences of the input documents and ranking them.

In our case, we compute a linear combination of the scores for each criterion described in section 2.2 to give a global relevance score to each sentence. The coefficients for each evaluation criterion is given in table 1. The global score is a weighted average of the five criteria that represent the degree of relevance of that sentence with the topic and the cluster of articles.

Evaluation Criterion	Coefficient
High DF Lemmas	45
Topic Similarity	15
Sentence Position	20
New Lemmas	2
Lost Lemmas	2

Table 1: Coefficients in the linear combination that computes the global score in both sentence selection steps.

2.4 Second Sentence Selection Step

The second sentence selection step consists of determining which of the top scored sentences, as evaluated in the first step, would make the best summary when combined together.

The first step provides us with a global score for all the sentences in the cluster of articles. We retrieve the 20 sentences with the best relevance score and then find *candidate summaries* from combinations of these top 20 sentences. Valid candidate summaries have these three properties: 1) they are made up of a combination of top 20 sentences; 2) their combined number of words does not exceed 100 words; and 3) no other sentences within the top 20 can be added while still remaining under the 100-word limit.

All the candidate summaries are then scored using the same criteria as step 1, as explained in section 2.2. A global relevance score is obtained by a linear combination of the evaluation criterion scores, which coefficients are given in table 1. The sentences contained in the candidate summary with the best global score are the ones selected for the summary.

2.5 Postprocessing

The postprocessing consists of ordering the sentences selected in the two-step process described above. The selected sentences are sorted for the summary in ascending chronological order of publication, with ties between resolved by choosing the sentence that appears the earliest in its source article.

2.6 Submitted Runs

We have submitted two runs in the competition, RALI1 (run ID 10) and RALI2 (run ID 8). RALI1 uses NESS2 as described above. RALI2 uses a more traditional one-step sentence selection scheme, which greedily selects the best scored sentences within the 100-word limit; RALI2 is otherwise identical to RALI1.

3 Results and Discussion

There were four evaluation methods used to assess the quality of the summaries submitted to TAC. The *overall responsiveness* score is based on both the linguistic quality of the summary and the amount of information in the summary that helps to satisfy the information need expressed in the topic narrative, as judged by NIST evaluators. The *linguistic quality* score is based on grammaticality, non-redundancy, referential clarity, focus and structure and coherence. The *Pyramid* scores are an evaluation of summary content relying on a manual comparison of semantic units with manual models (Harnly et al., 2005). Finally, *ROUGE* scores come from an automatic comparison with reference (manual) summaries, based on repeated fragments such as n-grams (Lin, 2004).

Table 2 shows how well we did in the four evaluation metrics when compared to the 52 submissions of the competition.

The results of the first run are very competitive, arriving in 4^{th} and 7^{th} place for both the overall responsiveness and linguistic quality, respectively in

Part A	Rouge	Pyr.	Ling. Q.	Overall R.
RALI1	9	12	4	4
RALI2	19	15	8	12
Part B				
RALI1	19	11	7	7
RALI2	15	12	4	9

Table 2: Ranks for each evaluation metric of our two runs in the TAC 2009 competition, out of the 52 submitted automatic runs.

part A and part B. Its pyramid scores are also pretty good, being in the top 12 scores in both parts.

The results for the second run are similar, but globally lower than the results of the first run. In particular, the overall responsiveness of the first run is better than that of the second run in both part A and part B. RALI2 is a simple and straightforward approach but – perhaps because of good tuning or because straightforward techniques actually are competitive even today – it nevertheless ranks amongst the top systems in the competition.

The differences between the two runs come entirely from the 2-step approach used in RALI1 but not in RALI2 – there are no other differences in the two submissions. This is what we wanted to test in the competition and the results show that a 2-step sentence selection process provides a somewhat significant gain in performance, though mostly in part A. In both parts, the 2-step method received better scores of pyramid and overall responsiveness. This can be explained by the fact that, even though the second sentence selection step makes it so that "less relevant" sentences are included in the summaries – often even excluding the sentence with the best global score – when compared to using the one-step approach, it is nevertheless a strong way to avoid redundancy and to increase on average the coverage of the summaries created, by including more sentences and sentences that are more different amongst themselves. It is an alternative to the centroid approach and other techniques used to avoid redundancy that deserves to be explored, notably because it allows the top scoring sentence to be excluded from the summary when this provides a gain – the possibility to include more relevant sentences for instance.

4 Conclusion

With NESS2, we have successfully verified the hypothesis that completing sentence selection in two steps performs better than doing it in one step. The results show that a simple approach that uses only a few linguistic criteria to evaluate sentences for extraction can still be competitive, and that it can be further improved by a second round of evaluation based on candidate summaries. The 2-step sentence selection method is an interesting, simple technique to reduce the redundancy in summaries.

References

Shouyuan Chen, Yuanming Yu, Chong Long, Feng Jin, Lijing Qin, Minlie Huang, and Xiaoyan Zhu. 2008. Tsinghua university at the summarization track of tac 2008. In *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

Pierre-Etienne Genest, Guy Lapalme, Luka Nerima, and Eric Wehrli. 2008. A symbolic summarizer for the update task of tac 2008. In *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, USA. National Institute of Standards and Technology. http://www.nist.gov/tac/publications/.

Aaron Harnly, Ani Nenkova, Rebecca Passonneau, and Owen Rambow. 2005. Automation of summary evaluation by the pyramid method. In *Recent Advances in Natural Language Processing (TANLP)*, September.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 74–81.

Eric Wehrli and Luka Nerima. 2009. L'analyseur syntaxique fips. In *IWPT'09 ATALA Workshop: What French parsing system ?*, Paris, France, October. Association pour le traitement automatique des langues.

Eric Wehrli. 2007. Fips, a "deep" linguistic multilingual parser. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague, Czech Republic, June. Association for Computational Linguistics.