A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization

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

We present SMMR, a scalable sentence scoring method for query-oriented update summarization. Sentences are scored thanks to a criterion combining query relevance and dissimilarity with already read documents (history). As the amount of data in history increases, non-redundancy is prioritized over query-relevance. We show that SMMR achieves promising results on the DUC 2007 update corpus.

1 Introduction

Extensive experiments on query-oriented multi-document summarization have been carried out over the past few years. Most of the strategies to produce summaries are based on an extraction method, which identifies salient textual segments, most often sentences, in documents. Sentences containing the most salient concepts are selected, ordered and assembled according to their relevance to produce summaries (also called extracts) (Mani and Maybury, 1999).

Recently emerged from the Document Understanding Conference (DUC) 2007[1], update summarization attempts to enhance summarization when more information about knowledge acquired by the user is available. It asks the following question: has the user already read documents on the topic? In the case of a positive answer, producing an extract focusing on only new facts is of interest. In this way, an important issue is introduced: redundancy with previously read documents (history) has to be removed from the extract.

A natural way to go about update summarization would be extracting temporal tags (dates, elapsed times, temporal expressions...) (Mani and Wilson, 2000) or to automatically construct the timeline from documents (Swan and Allan, 2000). These temporal marks could be used to focus extracts on the most recently written facts. However, most recently written facts are not necessarily new facts. Machine Reading (MR) was used by (Hickl et al., 2007) to construct knowledge representations from clusters of documents. Sentences containing "new" information (i.e. that could not be inferred by any previously considered document) are selected to generate summary. However, this highly efficient approach (best system in DUC 2007 update) requires large linguistic resources. (Witte et al., 2007) propose a rule-based system based on fuzzy coreference cluster graphs. Again, this approach requires to manually write the sentence ranking scheme. Several strategies remaining on post-processing redundancy removal techniques have been suggested. Extracts constructed from history were used by (Boudin and Torres-Moreno, 2007) to minimize history's redundancy. (Lin et al., 2007) have proposed a modified Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) re-ranker during sentence selection, constructing the summary by incrementally re-ranking sentences.

In this paper, we propose a scalable sentence scoring method for update summarization derived from MMR. Motivated by the need for relevant novelty, candidate sentences are selected according to a combined criterion of query relevance and dissimilarity with previously read sentences. The rest of the paper is organized as follows. Section 2

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

Coling 2008: Companion volume – Posters and Demonstrations, pages 23–26
Manchester, August 2008

introduces our proposed sentence scoring method and Section 3 presents experiments and evaluates our approach.

2 Method

The underlying idea of our method is that as the number of sentences in the history increases, the likelihood to have redundant information within candidate sentences also increases. We propose a scalable sentence scoring method derived from MMR that, as the size of the history increases, gives more importance to non-redundancy that to query relevance. 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 similarity measures and the scalable MMR scoring method.

2.1 A query-oriented multi-document summarizer

We have first started by implementing a simple summarizer for which the task is to produce query-focused summaries from clusters of documents. Each document is pre-processed: documents are segmented into sentences, sentences are filtered (words which do not carry meaning are removed such as functional words or common words) and normalized using a lemmas database (i.e. inflected forms "go", "goes", "went", "gone"... are replaced by "go"). An N-dimensional term-space Γ, where N is the number of different terms found in the cluster, is constructed. Sentences are represented in Γ by vectors in which each component is the term frequency within the sentence. Sentence scoring can be seen as a passage retrieval task in Information Retrieval (IR). Each sentence s is scored by computing a combination of two similarity measures between the sentence and the query. The first measure is the well known cosine angle (Salton et al., 1975) between the sentence and the query vectorial representations in Γ (denoted respectively $\vec{s}$ and $\vec{Q}$). The second similarity measure is based on the Jaro-Winkler distance (Winkler, 1999). The original Jaro-Winkler measure, denoted Jw, uses the number of matching characters and transpositions to compute a similarity score between two terms, giving more favourable ratings to terms that match from the beginning. We have extended this measure to calculate the similarity between the sentence s and the query Q:

$$Jw_e(s,Q) = \frac{1}{|Q|} \sum_{q \in Q} \max_{m \in S^{c}} Jw(q,m) \qquad (1)$$

where S^0 is the term set of s in which the terms m that already have maximized $Jw(q,m)$ are removed. The use of Jw_e smooths normalization and misspelling errors. Each sentence s is scored using the linear combination:

$$\begin{aligned} Sim_1(s,Q) = {} & \alpha \ cosine(\vec{s},\vec{Q}) \\ & + (1-\alpha) \ Jw_e(s,Q) \end{aligned} \qquad (2)$$

where $\alpha = 0.7$, optimally tuned on the past DUCs data (2005 and 2006). The system produces a list of ranked sentences from which the summary is constructed by arranging the high scored sentences until the desired size is reached.

2.2 A scalable MMR approach

MMR re-ranking 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:

$$\begin{aligned} MMR = {} & \arg\max_{s \in S} [\lambda \ Sim_1(s,Q) \\ & - (1-\lambda) \ \max_{s_j \in E} Sim_2(s,s_j) \,] \end{aligned} \qquad (3)$$

where S is the set of candidates sentences and E is the set of selected sentences. λ represents an interpolation coefficient between sentence's relevance and non-redundancy. $Sim_2(s,s_j)$ is a normalized Longest Common Substring (LCS) measure between sentences s and s_j. Detecting sentence rehearsals, LCS is well adapted for redundancy removal.

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 (3) as (NR stands for Novelty Relevance):

$$\begin{aligned} NR = {} & \arg\max_{s \in S} [\lambda \ Sim_1(s,Q) \\ & + (1-\lambda) \ (1 - \max_{s_h \in H} Sim_2(s,s_h)) \,] \end{aligned} \qquad (4)$$

We can understand that (4) equates to an OR combination. But as we are looking for a more intuitive AND and since the similarities are independent, we have to use the product combination. The scoring method defined in (2) is modified into a double maximization criterion in which the best ranked sentence will be the most relevant to the query AND the most different to the sentences in H.

$$\textsc{Smmr}(s) = \mathrm{Sim}_1(s, Q) \quad 1 - \max_{s_h \text{⍰} H} \mathrm{Sim}_2(s, s_h)^{f(H)} \tag{5}$$

Decreasing λ in (3) 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 ($f(H) \to 0$).

3 Experiments

The method described in the previous section has been implemented and evaluated by using the DUC 2007 update corpus[2]. The following subsections present details of the different experiments we have conducted.

3.1 The DUC 2007 update corpus

We used for our experiments the DUC 2007 update competition data set. The corpus is composed of 10 topics, with 25 documents per topic. The update task goal was to produce 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 (10 documents), B (8 documents) and C (7 documents), the task is to 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. Our system generates a summary for each cluster by arranging the high ranked sentences until the limit of 100 words is reached.

3.2 Evaluation

Most existing automated evaluation methods work by comparing the generated summaries to one or more reference summaries (ideally, produced by humans). To evaluate the quality of our generated summaries, we choose to use the Rouge[3] (Lin, 2004) evaluation toolkit, that has been found to be highly correlated with human judgments. Rouge-n is a n-gram recall measure calculated between a candidate summary and a set of reference summaries. In our experiments Rouge-1, Rouge-2 and Rouge-su4 will be computed.

3.3 Results

Table 1 reports the results obtained on the DUC 2007 update data set for different sentence scoring methods. cosine + Jw_e stands for the scoring method defined in (2) and NR improves it with sentence re-ranking defined in equation (4). Smmr is the combined adaptation we have proposed in (5). The function $f(H)$ used in Smmr is the simple rational function $\frac{1}{H}$, where H increases with the number of previous clusters ($f(H) = 1$ for cluster A, $\frac{1}{2}$ for cluster B and $\frac{1}{3}$ for cluster C). This function allows to simply test the assumption that non-redundancy have to be favoured as the size of history grows. Baseline results are obtained on summaries generated by taking the leading sentences of the most recent documents of the cluster, up to 100 words (official baseline of DUC). The table also lists the three top performing systems at DUC 2007 and the lowest scored human reference.

As we can see from these results, Smmr outperforms the other sentence scoring methods. By ways of comparison our system would have been ranked second at the DUC 2007 update competition. Moreover, no post-processing was applied to the selected sentences leaving an important margin of progress. Another interesting result is the high performance of the non-update specific method (cosine + Jw_e) that could be due to the small size

[2] More information about the DUC 2007 corpus is available at http://duc.nist.gov/.

[3] Rouge is available at http://haydn.isi.edu/ROUGE/.

of the corpus (little redundancy between clusters).

	Rouge-1	Rouge-2	Rouge-su4
Baseline	0.26232	0.04543	0.08247
3rd system	0.35715	0.09622	0.13245
2nd system	0.36965	0.09851	0.13509
cosine + JW$_e$	0.35905	0.10161	0.13701
NR	0.36207	0.10042	0.13781
Smmr	0.36323	0.10223	0.13886
1st system	0.37032	0.11189	0.14306
Worst human	0.40497	0.10511	0.14779

Table 1: Rouge average recall scores computed on the DUC 2007 update corpus.

4 Discussion and Future Work

In this paper we have described Smmr, a scalable sentence scoring method based on MMR that achieves very promising results. An important aspect of our sentence scoring method is that it does not requires re-ranking nor linguistic knowledge, which makes it a simple and fast approach to the issue of update summarization. It was pointed out at the DUC 2007 workshop that Question Answering and query-oriented summarization have been converging on a common task. The value added by summarization lies in the linguistic quality. Approaches mixing IR techniques are well suited for query-oriented summarization but they require intensive work for making the summary fluent and coherent. Among the others, this is a point that we think is worthy of further investigation.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche, France, project RPM2.

References

Boudin, F. and J.M. Torres-Moreno. 2007. A Cosine Maximization-Minimization approach for User-Oriented Multi-Document Update Summarization. In Recent Advances in Natural Language Processing (RANLP), pages 81–87.

Carbonell, J. and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 335–336. ACM Press New York, NY, USA.

Hachey, B., 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).

Hickl, A., K. Roberts, and F. Lacatusu. 2007. LCC's GISTexter at DUC 2007: Machine Reading for Update Summarization. In Document Understanding Conference (DUC).

Lin, Z., T.S. Chua, M.Y. Kan, W.S. Lee, L. Qiu, and S. Ye. 2007. NUS at DUC 2007: Using Evolutionary Models of Text. In Document Understanding Conference (DUC).

Lin, C.Y. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In Workshop on Text Summarization Branches Out, pages 25–26.

Mani, I. and M.T. Maybury. 1999. Advances in Automatic Text Summarization. MIT Press.

Mani, I. and G. Wilson. 2000. Robust temporal processing of news. In 38th Annual Meeting on Association for Computational Linguistics, pages 69–76. Association for Computational Linguistics Morristown, NJ, USA.

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

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620.

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

Winkler, W. E. 1999. The state of record linkage and current research problems. In Survey Methods Section, pages 73–79.

Witte, R., R. Krestel, and S. Bergler. 2007. Generating Update Summaries for DUC 2007. In Document Understanding Conference (DUC).

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