Transductive Learning over Automatically Detected Themes for Multi-Document Summarization

Massih-Reza Amini
National Research Council Canada
Institute for Information Technology
283, boulevard Alexandre-Taché
Gatineau, QC J8X 3X7, Canada
Massih-Reza.Amini@nrc-cnrc.gc.ca

Nicolas Usunier
Université Pierre et Marie Curie
Laboratoire d'Informatique de Paris 6
4, Place de Jussieu
75252 Paris, cedex 05
Nicolas.Usunier@lip6.fr

ABSTRACT

We propose a new method for query-biased multi-document summarization, based on sentence extraction. The summary of multiple documents is created in two steps. Sentences are first clustered; where each cluster corresponds to one of the main themes present in the collection. Inside each theme, sentences are then ranked using a transductive learning-to-rank algorithm based on RankNet [2], in order to better identify those which are relevant to the query. The final summary contains the top-ranked sentences of each theme. Our approach is validated on DUC 2006 and DUC 2007 datasets.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation, Performance

Keywords

Mutli-document summarization, Learning to Rank

1. INTRODUCTION

Multi-document summarization (MDS) aims at reducing the information overload caused by the ever increasing number of documents on the same or similar topics, and hence has attracted significant research attention since the past decade. With the current web growth, there are increasingly more web-oriented summarization applications. MDS can be used with conventional search engines, for example to provide informative snippets to help users navigate through different parts of the result page [8]. It can also provide short summaries of documents initially clustered by e.g. a news aggregator to assist users in better understanding the different views presented in the news [5]. Another application is a Question & Answering system which, for each asked question, provides information about the answer in the form of a short extractive summary [3]. In this study, we consider query-biased MDS where we dispose of a set of queries and a set of relevant documents for each of these queries. For each query, an ideal multi-document summarizer consists in producing relevant information around key facets dealing with the query and which is present in the set of its relevant documents. A major issue for a MDS system is, therefore, to automatically detect these themes, and in each of these themes, to rank sentences relevant to the query.

This paper introduces a two-step method for query-biased MDS. Our approach first detects the main themes of the documents by clustering the sentences of all of the documents associated to a given query, where sentences are represented in a low-dimensional space obtained with LSI. For each cluster (or theme), the sentences are then ordered using a transductive learning-to-rank algorithm based on RankNet [2]. Experiments carried out on DUC 2006 and 2007 corpora show that we consistently improve over competing techniques.

SIGIR'11, July 24-28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

2. THE PROPOSED MODEL

We consider a setting similar to the TAC competitions[1], where each query consists of a title and a question, and, for each query, we dispose of a set of relevant documents. Our algorithm operates in two steps.

Step 1 - Theme detection: For each query and its set of relevant documents, we suppose that each theme inside this set, provides a partial answer to the query. Our theme detection scheme indirectly takes into account the query through this assumption. To find these themes, we first group syntagmatic similar words by applying a word-clustering algorithm, proposed in one of the top performing systems at DUC 2007 for query expansion [1]. Sentences are then parsed and, adjectives and verbs are extended using word clusters found before. The augmented sentences are coded in the bag-of-words space with TF features, and the word-sentence matrix is reduced using SVD, a reduction technique similar to LSI. Under this reduced representation, sentence clustering is finally performed using X-means [7]. This clustering algorithm is an extension of the well-known K-means algorithm in which the optimal number of clusters is found at the same time than centroid locations. This clustering allows us to treat each theme independently, with the goal of finding the most relevant sentences to the query in each of them. At the end of this step, sentences are then ranked using their bag-of-words similarity with the query.

[1]http://www.nist.gov/tac/

Step 2 - A transductive RankNet algorithm for MDS: As we shall see in the experiments, the ranking obtained at the end of the first step is rather crude. We propose to improve it with learning-to-rank techniques inspired by web search, adapted to the transductive setting. For each cluster, we first automatically define relevant (resp. irrelevant) sentences to the query by taking the top (resp. bottom) ranked sentences of the first step. This gives us a (artificially) labeled training set to initiate the learning process. By extracting a generic query for each theme defined as the most frequent terms of that theme, we then characterize sentences in the latter by taking 12 features used in the Letor datasets [6] as well as a feature produced by a bigram language model proposed in the top performing system at DUC 2006 [4]. Thus, for each theme, sentences have a representation that depends on the theme while the associated relevance judgment depends on the topic in hand. For each theme, we then iteratively train the RankNet algorithm [2] and assign pseudo-relevance judgments to sentences using the output of the learned ranking algorithm. With this new training set, we learn another ranking function using RankNet. This *transductive* learning scheme allows us to gradually improve the quality of the rankings.

For each topic, a summary is finally generated by taking the top ranked sentence given by different RankNet algorithms in each of the themes. The first top ranked sentence, of one of the themes, included in the summary is the one which appears in the most recent document of the set of documents associated to the topic. After including all the first ranked sentences, if the overall legnth of the summary does not exceed 250 words (as in DUC 2006 and DUC 2007 main task), we repeat the operation using the second ranked sentences until this summary length is reached.

3. EXPERIMENTAL RESULTS

We carried out experiments on DUC 2006 and DUC 2007 datasets[2]. DUC 2006 and DUC 2007 contain respectively 50 and 45 topics, each composed of a set of keywords (the title) and a question. Each topic is associated to 25 relevant documents from the AQUAINT corpus. For each topic, the dataset also has three reference summaries produced by human assessors. Since we do not need any *prior labeled* training data to run our algorithm, these reference summaries are only used for evaluation. In our experiments, we considered title keywords and non-stop words in the question as the query and employed the ROUGE toolkit applied by NIST for performance evaluation in DUC competitions. We compared our approach with two base-level summarizers, namely lead and random, and the top two performing systems in both competitions. The latter are those which achieved the highest ROUGE scores in that competition. In Table 1, these systems are denoted by system and their attributed numbers given in these competitions. The lead baseline returns all the leading sentences (up to 250 words) in the most recent document for each topic and the random baseline selects sentences in random. In order to show the contribution of each of the two steps; for each topic, we also generate summaries by extracting the most similar sentence to the topic question and title in each of the sentence clusters. The most similar sentences which appear in the most recent documents are added first. This strategy corresponds to our first step (before training) and is noted by step 1 in Table 1. We observe that on DUC 2006 and DUC 2007, the combination of step 1 and step 2, achieves the best results over other systems. These results indicate that the transductive approach is able to leverage from both the query dependent pseudo-relevance judgments and the theme dependent sentence representations to find an efficient combination of sentence features for the summary.

[2]http://www-nlpir.nist.gov/projects/duc/data.html

Table 1: Comparison results on DUC 2006 and DUC 2007.

	Method	ROUGE-2	ROUGE-L	ROUGE-SU4
DUC 2006	Random	0.04892	0.29384	0.10083
	Lead	0.05267	0.29726	0.10408
	Step 1	0.06842	0.31984	0.12454
	System 12	0.08990	0.37132	0.14753
	System 24	0.09513	0.37741	0.15478
	Step 1 + Step 2	**0.09855**	**0.38030**	**0.15739**
DUC 2007	Random	0.05968	0.30713	0.11001
	Lead	0.06490	0.31074	0.11278
	Step 1	0.08411	0.34001	0.13876
	System 15	0.12285	0.40561	0.17470
	System 24	0.11605	0.41033	0.17304
	Step 1 + Step 2	**0.12866**	**0.41897**	**0.17867**

4. CONCLUSION

We proposed a learning to rank approach for extractive summarization based on a transductive setting. Our approach allows to extract sentences from different themes of a document collection, which are relevant to the query. Our experiments on DUC 2006 and DUC 2007 show that our algorithm achieves the best results in terms of the ROUGE measures compared to state-of-the-art.

Acknowledgments

This work was supported in part by the IST Program of the EC, under the PASCAL2 Network of Excellence.

5. REFERENCES

[1] M.-R. Amini and N. Usunier. A contextual query expansion approach by term clustering for robust text summarization. In *Proceedings of DUC*, 2007.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In Proceedings of ICML, pages 89–96, 2005.

[3] T. Hirao, Y. Sasaki, and H. Isozaki. An extrinsic evaluation for question-biased text summarization on QA tasks. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 61–68, 2001.

[4] J. Jagarlamudi, P. Pingali, and V. Varma. Query Independent Sentence Scoring Approach to DUC 2006. In *Proceedings of DUC*, 2006.

[5] K. R. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg. Do summaries help? In *Proceedings of ACM SIGIR*, pages 210–217, 2005.

[6] T. Qin, T. Y. Liu, J. Xu, and H. Li. LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. Information Retrieval Journal, 2010.

[7] D. Pelleg, and A. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of ICML*, pages 727–734, 2000.

[8] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. Fast generation of result snippets in web search. In *Proceedings of ACM SIGIR*, pages 127–134, 2007.