Incorporating Prior Knowledge into a Transductive Ranking Algorithm for Multi-Document Summarization
Amini, Massih-Reza; Usunier, Nicolas

National Research Council Canada

Conseil national de recherches Canada

Canada

Incorporating Prior Knowledge into a Transductive Ranking Algorithm 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
104, avenue du Président Kennedy
75016 Paris, France
Nicolas.Usunier@lip6.fr

ABSTRACT

This paper presents a transductive approach to learn ranking functions for extractive multi-document summarization. At the first stage, the proposed approach identifies topic themes within a document collection, which help to identify two sets of relevant and irrelevant sentences to a question. It then iteratively trains a ranking function over these two sets of sentences by optimizing a ranking loss and fitting a prior model built on keywords. The output of the function is used to find further relevant and irrelevant sentences. This process is repeated until a desired stopping criterion is met.

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 extracting information relevant to an implicit or explicit subject from different documents written about that subject or topic. MDS is generally a more complex task than single document summarization (SDS) as it aims to capture different themes inside a set of documents rather than to simply shorten the source texts [2]. An ideal multi-document summarizer attempts to produce relevant information around key facets dealing with the topic and present in the set of its relevant documents. A major issue for MDS is, therefore, to automatically detect these themes. In this paper we propose to incorporate prior knowledge induced from a set of keywords into a transductive algorithm to learn ranking functions with a minimal annotation effort for multi-document summarization. Learning with prior knowledge has become a wide field of research in the last years [4]. The emphasis here is to incorporate domain knowledge in the learning process rather

than labeled instances which, in some applications are impossible to gather. The summarizer we propose learns a linear ranking function from the bag-of-words representation of sentences. At this stage, the construction of ranking functions is hindered by two difficulties. First, the initial training set should be defined automatically, whereas typical machine learning methods require manually annotated data. Secondly, for each question the algorithm should be able to learn an accurate function with very few examples. To this end, our approach first defines a prior probability of relevance for every sentence using the set of keywords associated to a given question. It then iteratively learns a scoring function which fits the prior probabilities, and also minimizes the number of irrelevant sentences scored above the relevant ones. At each iteration, new relevant and irrelevant sentences are identified using the scores predicted by the current function. These sentences are added to the training set, and a new function is trained.

2. THE PROPOSED MODEL

We consider the case where for each given question, there are not any manually labeled sets of relevant and irrelevant sentences available. In order to learn, our approach first builds training sets automatically from the following common assumption that question words as well as their topically related terms are relevant to the summary. For a given question $q \in \mathcal{Q}$ and a set of documents $\mathcal{D}$, we thereby use a term-clustering technique proposed in [1] to first find *words that are topically related* to the question. This technique partitions terms that appear in the same documents with the same frequency. It has empirically been shown that words belonging to the same term-cluster are topically related. Following the assumption above and for each question q, we first create an initial training set by gathering two sets of relevant and irrelevant sentences to the summary. The relevant set is composed of the extended question, $\bar{q}$, containing question words and words that belong to the same term-clusters than each of the question words. The set of irrelevant sentences is constituted of sentences that do not contain any of the expanded question words.

Prior model of sentence relevance.

The prior model we propose takes then the form of a language model that computes conditional probability estimates $\pi(q \mid s)$, over the set of questions $q \in \mathcal{Q}$, for each sentence $s \in \mathcal{D}$. The model first defines, for any keyword w, a conditional probability of generating the question $\pi(q \mid w)$.

We further assume that all no-keyword terms are equiprobable to all questions: $\forall \bar{w} \notin \bigcup_{q \in \mathcal{Q}} k_q, \forall q, \pi(q \mid \bar{w}) = \frac{1}{|\mathcal{Q}|}$ and set an uniform prior distribution for questions: $\forall q, \pi(q) = \frac{1}{|\mathcal{Q}|}$, where k_q is the extended keywords set of q. Finally by making the naive Bayes assumption that sentence terms are conditionally independent given a question, we estimate the conditional probability $\pi(q \mid s)$ of a question q given a sentence s using Bayes' rule. An advantage of this model is that it provides fast probability estimates which are computed once before the training stage that we present in the following section.

Learning to rank with few examples.

Our ranking algorithm works in a *transductive setting*. In this case the whole set of sentences to be ranked is known *prior to learning*. This setting makes use of the unlabeled examples in the learning stage in order to compensate for the small size of the initial generated training set. The transductive summarizer algorithm is then composed of two main parts: (1) a *prior knowledge model* and (2) an iterative architecture that follows the self-learning paradigm [3] minimizing the following criterion

$$\mathcal{L}(f) = \frac{1}{|S^+||S^-|} \sum_{s^+ \in S^+} \sum_{s^- \in S^-} \log_2(1 + e^{-(f(s^+)-f(s^-))})$$
$$+ \frac{\lambda}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} kl(\pi(q \mid s)||P(f(s))) \qquad (1)$$

Where, the first term is a standard ranking loss and the second term is the Kullback-Leibler divergence between the outputs of the prior model and the learnt function with $P(t) = (1 + e^{-t})^{-1}$ a sigmoid function, used to transform the score $f(s)$ into a posterior probability estimation of relevance, λ is a discount factor used to balance the relative influence of the prior model and $\mathcal{D}$ is the set of documents related to q. Initially, S^+ and S^- are the generated training set. Then, a function is learned, and some unlabeled examples are added to S^+ or S^- using the predicted score. The process is repeated until S^+ achieves a sufficient size.

3. EXPERIMENTAL RESULTS

We conducted our experiments on DUC 2005[1] data set. Documents are news articles collected from the AQUAINT corpus. For a given question, a summary is to be formed on the basis of a subset of documents to its corresponding topic. Each question comes also with a set of keywords that we used for probability estimation in our prior model. It is to be noted that for each topic, we dispose of 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 purposes. For the evaluation criteria we used the ROUGE toolkit (version 1.5.5) applied by NIST for performance evaluation in DUC competitions. This program measures the quality of a produced summary by counting the relative number of its unit overlaps with a set of reference summaries - produced by three human assessors in these competitions. The following table provides the comparison results on DUC 2005. We compared our approach with two base-level summarizers, namely lead and random, and the top two performing

systems on DUC 2005. The latter are those which achieved the highest ROUGE scores in that competition. 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. To see the effect of the prior knowledge model in learning the ranking function, we report experimental results obtained by TranSumm in the cases where the prior knowledge model is not used ($\lambda = 0$) or that summaries are exclusively made using the latter (λ_∞) and finally for the best value of lambda (λ^*).

Summarizer	ROUGE-2	ROUGE-L	ROUGE-SU4
Lead	0.04320	0.27089	0.09303
Random	0.04143	0.26395	0.09066
System 5	0.06975	0.34094	0.12767
System 8	0.07132	0.33869	0.13065
TranSumm ($\lambda = 0$)	0.06842	0.32945	0.12594
TranSumm (λ_∞)	0.07012	0.33876	0.13108
TranSumm (λ^*)	**0.07546**	**0.35042**	**0.13657**

We observe that on DUC 2005 the proposed algorithm achieves the best results over other systems for the optimal value of the discount factor. We believe that this improvement is due to two conjugated factors. First, expanded keywords and question terms on this collection contain summary terms. We have further seen that questions from DUC 2005 are almost long, containing 12.42 words in average. On the other hand, as TranSumm initializes the set of relevant sentences first by the expanded question and that it increases the score of sentences containing expanded keywords terms via the prior knowledge model. It turns out that summary-like sentences in DUC 2005 have potentially high scores.

4. CONCLUSION

We proposed a learning to rank approach for extractive summarization based on a transductive setting. Our approach allows to extract sentences having similar words than questions, their topically related terms and the initial keywords. Our experiments on DUC 2005 show that our algorithm achieves the best results compared to state-of-the-art.

Acknowledgments

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] I. Mani and E. Bloedorn, Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2):35-67, 1999.

[3] R. Reichart and A. Rappoport, Self-Training for Enhancement and Domain Adaptation of Statistical Parsers Trained on Small Datasets. In *Proceedings of ACL*, pages 616-623, 2007.

[4] Robert E. Schapire and Marie Rochery and Mazin Rahim and Narendra Gupta, Incorporating Prior Knowledge into Boosting. In *Proceedings of ICML*, pages 538-545, 2002.