Description of the Google update summarizer at TAC-2011

Jean-Yves Delort
Google Research
Brandschenkestrasse 110
8002 Zurich, Switzerland
jydelort@google.com

Enrique Alfonseca
Google Research
Brandschenkestrasse 110
8002 Zurich, Switzerland
ealfonseca@google.com

Abstract

In this paper we describe the system with which we have participated at TAC-2011 in the task of update summarization. This is our first participation in TAC, and we have started exploring the use of topic models for summarization. We have participated with a lightweight system that is an extension of TOPICSUM (Haghighi and Vanderwende, 2009), which we are currently extending for update summarization. The system had almost no pre-processing nor post-processing. The resulting scores rank as average across all participants. In this paper we analyze the results and outline some ideas for improvement.

1 Introduction

Web users have shown their interest for tools giving them the possibility to follow information, as witnessed by the popularity of email alerting systems, RSS readers and other social platforms. One of the main challenges for such systems is to identify new information that is not already known by the users and to synthesize it.

This is the scenario addressed by the **update summarization** task organized in the Text Analysis Conference (TAC): let us suppose that a user is interested in how some news story develops and wants to track it over time. When she has missed the news for a time interval, she would like to read a summary that highlights what is new during the time she was not reading the news, to get up to date again. An update summarizer is a system that addresses this information need. It receives as input two different sets of documents: one ($\mathcal{A}$) containing news that the user has already read, and a second set ($\mathcal{B}$) that is new to her. The task is to generate a short, fluent, multi-document summary of the second set that does not contain information present in the first set (as it would be redundant).

There has been an update summarization task in the past five DUC and TAC competitions, between 2007 and 2011[1]. To ensure that the summary sentences are coherent, many systems choose a sentence extraction procedure[2].

A usual approach to multi-document summarization is to start by ranking sentences in order of relevance or informativeness, and then selecting the top ranked sentences. A redundancy removal mechanism such as MMR (Carbonell and Goldstein, 1998) is usually added to ensure that no sentence is selected if there is a higher-ranking similar sentence. This simple procedure can also be applied to generate *update summaries*: given the two document collections $\mathcal{A}$ and $\mathcal{B}$, it is possible to generate first a summary for $\mathcal{A}$. Next, when generating a summary for $\mathcal{B}$, the redundancy removal algorithm is applied not only using the higher-ranking sentences from $\mathcal{B}$, but also using the sentences from $\mathcal{A}$'s summary. In this way, no sentence from $\mathcal{B}$'s collection that is very similar to the central topic of $\mathcal{A}$ will be

[1] http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html#pilot
http://www.nist.gov/tac/2008/summarization/update.summ.08.guidelines.html
http://www.nist.gov/tac/2009/Summarization/update.summ.09.guidelines.html
http://www.nist.gov/tac/2010/Summarization/Guided-Summ.2010.guidelines.html
http://www.nist.gov/tac/2011/Summarization/

[2] Note that the evaluation metrics used in summarization, such as ROUGE, try to maximize the information contents of the summary but do not address the problem of properly reordering the extracted sentences in a coherent discourse. In this work we will also ignore the sentence ordering.

selected (Kennedy et al., 2010). A more generic approach may add some sophisticated penalty for sentences from set $\mathcal{B}$ according to their content overlap with any part of document set $\mathcal{A}$ (Li et al., 2010).

There is a second family of generic multi-document summarizers that has also shown success by encoding content as sets of n-grams and selecting the summary that overall maximizes some metric over them (Nenkova and Vanderwende, 2005; Haghighi and Vanderwende, 2009; Gillick and Favre, 2009; Berg-Kirkpatrick et al., 2011). Redundancy is implicitly encoded as redundant summaries will cover fewer n-grams. One example is the characterization of a collection as a multinomial distribution, so the summary that better approximates the target distribution will be considered to be the best one. However, in contrast to the previous family, that can be easily modified to generate update summaries, as far as we know no attempt has been done to generalize these approaches to generate update summaries.

We are currently working on extending topic model-based summarizers for the particular case of update summarization. We have participated in TAC-2011 with a preliminary system that is deeply inspired by the TOPICSUM algorithm (Haghighi and Vanderwende, 2009). This is a preliminary version of DUALSUM, which is described by Delort and Alfonseca (2012) and performs much better.

2 System Description

Given a target distribution T, there are a number of metrics that can be used to rank candidate summaries based on their similarity to this distribution. The Kullback-Leibler divergence (KL) is one of the most widely used, especially in a probabilistic context. In summarization, KL can be used to find the set of extracted sentences that has the minimal divergence with the target distribution:

$$S^* = \underset{S}{\operatorname{argmin}} KL(T, S) = \sum_{w \in \mathbf{V}} p_T(w) \log \frac{p_T(w)}{p_S(w)}$$

where w is a word from the vocabulary $\mathbf{V}$.

Although some variations of KL can be considered, such as penalizing sentences that contain document-specific topics (Mason and Charniak, 2011), here we follow the same approach as Haghighi and Vanderwende (2009). Since finding this subset of sentences is a NP-complete set covering problem, a greedy algorithm is often used in practice:

1. Initialize $S^* = \emptyset$.
2. While $length(S^*) < limit$,
   (a) Let $s_i = \underset{i}{\operatorname{argmin}}\, KL(T, S^* \cup s_i)$
   (b) $S^* = S^* \cup s_i$

An important part of KL-based summarizers is the selection of the target distribution to optimize. Daumé and Marcu (2006), Chemudugunta et al. (2006) and Haghighi and Vanderwende (2009) proposed the use of LDA-like algorithms to automatically discern between words that belong to a generic background distribution, words that are mostly document-specific, and words that refer to the central topic of each news collection. We have reimplemented KLSUM and TOPICSUM (Haghighi and Vanderwende, 2009) as baselines to use as the starting point of our work. We have adapted them in a very trivial way to the generation of update summaries by running them only on the set of new collections ($\mathcal{B}$), ignoring the original collections. This will produce summaries that are close to the central topic of $\mathcal{B}$, but it is not guaranteed that they will not repeat the original information already present in $\mathcal{A}$.

The system with which we have participated in TAC-2011 is an extension of TOPICSUM specifically adapted to generate update summaries (manuscript unpublished). We shall call this algorithm AGGSUM.

3 Evaluation

One of the criteria used at TAC to evaluate and compare the informativeness of generated summaries is the ROUGE score. ROUGE is a recall-oriented measure which measures the overlap between n-grams in a generated summaries with n-grams found in gold-standard human-produced summaries.

Figure 1 shows ROUGE scores for various summarizers during the last TAC competitions with respect to summary size. Shown summarizers are the top-5 performers for 100-words summaries as well as KLSUM, TOPICSUM and AGGSUM. Figure 2

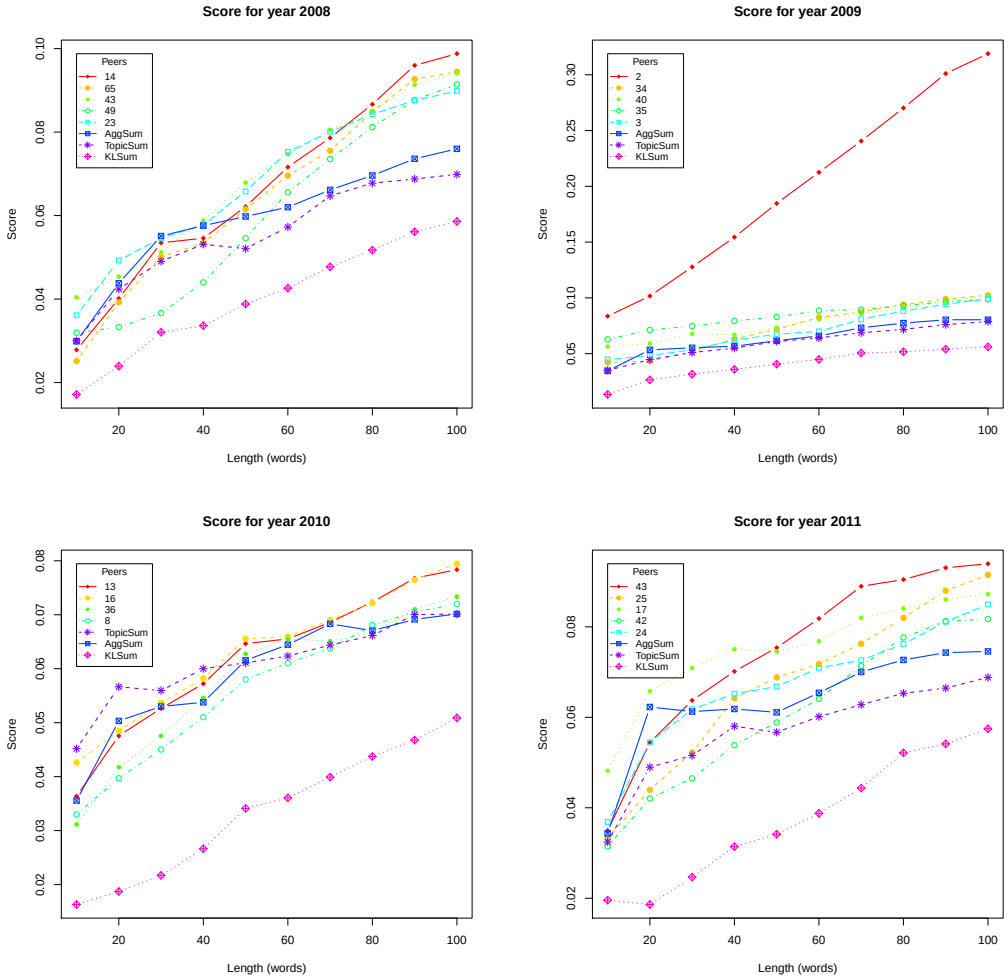

Figure 1: ROUGE score for different summary lengths of the top five systems and the three tested benchmarks

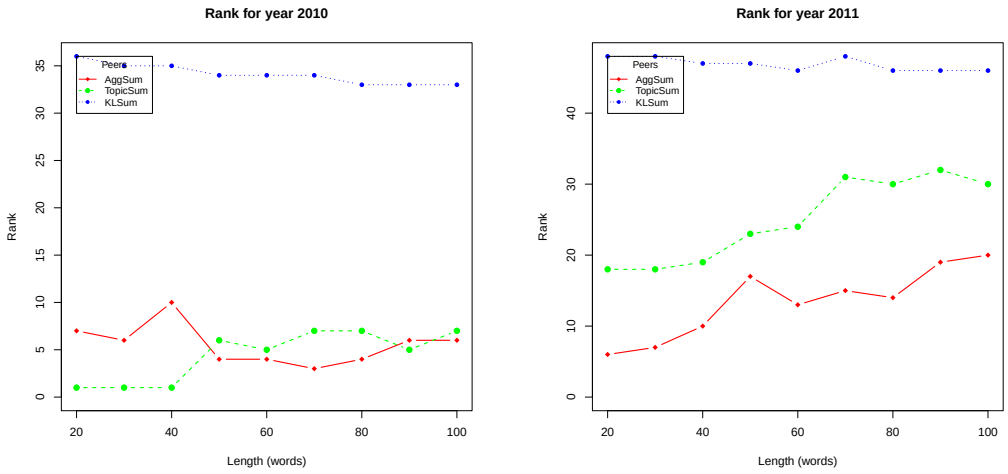

Figure 2: Rank of the three benchmarks for different summary lengths.

shows the rank of the three benchmark systems for different summary lengths.

The first observation from those graphs is that ROUGE is significantly affected by the summary length. While adding more sentences increases the ROUGE score for all the systems, the speed of improvement can be quite different among summarizers. In other words, considering the ROUGE score as the gold-standard evaluation metric, the ranking between summarizers depends on the summary length. This issue is important since, in many real-world scenarios, users need very short summaries, sometimes shorter than the 100 words limit given by TAC.

The figure also confirm the effectiveness of TOPICSUM (and AGGSUM) over KLSUM as reported by (Haghighi and Vanderwende, 2009). While KLSUM uses the raw n-gram distribution as a representation of the core content of the collection, TOPICSUM and AGGSUM are based on topic-models that can learn a more representative distribution. Thus, topic-model approaches can more effectively discard unwanted words, such as stopwords and document specific words.

The results show that the TOPICSUM and AGGSUM can select very useful sentences at the first iteration of the greedy algorithm, ranking very well amongst all the other TAC participants. However, the ROUGE gains obtained by adding new sentences are comparatively very small. This may indicate that this particular sentence selection algorithm is not very good at redundancy removal, as the new added sentences are not adding much new coverage over the manual summaries. Other reason is because of the way the greedy algorithm works and the way the collection distribution is generated. If the distribution is too focused, i.e. the majority of the mass is concentrated on a few n-grams about the very central topic, then the first sentence is likely to cover most of the collection probability mass, and the algorithm tends to add the subsequent sentences with *as little information as possible*. For example, if the first selected sentence contains all of the important n-grams in the collection and in a similar proportion, then the algorithm will avoid selecting sentences with important words in order to keep the KL divergence low.

A final comment is the fact that AGGSUM outperforms TOPICSUM on the 2011 dataset for all summary lengths.

4 Conclusions

This paper describes our contribution to the task of update summarization in TAC-2011, which consists of an extension of TOPICSUM to be able to generate update summaries. By analyzing the results, we have observed that it performs comparably very well for very short summaries (20 words, usually containing just one sentence) in terms of ROUGE-2. Using TOPICSUM executed on the update set $\mathcal{B}$ as a baseline, we show that it also performs better on shorter summaries. However, as the size of the summary increases, the ROUGE scores grow with a comparatively smaller slope than the output from other participants. This is particularly true for TAC-2011, where the ROUGE score does not improve at all as the summary size increases from 20 to 50 words.

We believe that the problem is due to (a) the greedy implementation of KLSUM, selecting one sentence at a time instead of globally minimizing the divergence of the whole summary; and (b) the learned collection distributions, which may be very skewed towards the very central topic of the collection, but probably do not have enough information to cover a 100-words summary.

Future work includes further modifications to the model to better learn the distribution about the news updates, and a better sentence selection model that does not suffer from the drawbacks of the greedy algorithm. We also plan to explore the effects of pre-processing and post-processing on the results.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Program (FP7/2007-2013) under grant agreement number 257790.

References

T. Berg-Kirkpatrick, D. Gillick, and D. Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490. Association for Computational Linguistics.

J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248.

Hal Daumé, III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-2006, pages 305–312, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jean-Yves Delort and Enrique Alfonseca. 2012. DUALSUM: A topic-model based approach for update summarization. In *Proceedings of EACL-2012*.

D. Gillick and B. Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18. Association for Computational Linguistics.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

A. Kennedy, T. Copeck, D. Inkpen, and S. Szpakowicz. 2010. Entropy-based sentence selection with roget's thesaurus. In *Proceedings of the third Text Analysis Conference, TAC-2010*. NIST.

S. Li, T. Song, and X. Wang. 2010. TAC2010 update summarization and AESOP of ICL. In *Proceedings of the third Text Analysis Conference, TAC-2010*. NIST.

Rebecca Mason and Eugene Charniak. 2011. Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, WASDGML '11, pages 49–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Nenkova and L. Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.