Extractive Multi-Document Summaries Should Explicitly Not Contain Document-Specific Content

Rebecca Mason and Eugene Charniak
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI 02912
{rebecca,ec}@cs.brown.edu

Abstract

Unsupervised approaches to multi-document summarization consist of two steps: finding a content model of the documents to be summarized, and then generating a summary that best represents the most salient information of the documents. In this paper, we present a sentence selection objective for extractive summarization in which sentences are penalized for containing content that is specific to the documents they were extracted from. We modify an existing system, HIERSUM (Haghighi & Vanderwende, 2009), to use our objective, which significantly outperforms the original HIERSUM in pairwise user evaluation. Additionally, our ROUGE scores advance the current state-of-the-art for both supervised and unsupervised systems with statistical significance.

1 Introduction

Multi-document summarization is the task of generating a single summary from a set of documents that are related to a single topic. Summaries should contain information that is relevant to the main ideas of the entire document set, and should not contain information that is too specific to any one document. For example, a summary of multiple news articles about the *Star Wars* movies could contain the words "Lucas "and "Jedi", but should not contain the name of a fan who was interviewed in one article. Most approaches to this problem generate summaries *extractively*, selecting whole or partial sentences from the original text, then attempting to piece them together in a coherent manner. Extracted text is selected based on its relevance to the main ideas of the document set. Summaries can be evaluated manually, or with automatic metrics such as ROUGE (Lin, 2004).

The use of structured probabilistic topic models has made it possible to represent document set content with increasing complexity (Daumé & Marcu, 2006; Tang et al., 2009; Celikyilmaz & Hakkani-Tur, 2010). Haghighi and Vanderwende (2009) demonstrated that these models can improve the quality of generic multi-document summaries over simpler surface models. Their most complex hierarchial model improves summary content by teasing out the words that are not general enough to represent the document set as a whole. Once those words are no longer included in the content word distribution, they are *implicitly* less likely to appear in the extracted summary as well. But this objective does not sufficiently keep document-specific content from appearing in multi-document summaries.

In this paper, we present a selection objective that *explicitly* excludes document-specific content. We re-implement the HIERSUM system from Haghighi and Vanderwende (2009), and show that using our objective dramatically improves the content of extracted summaries.

2 Modeling Content

The easiest way to model document content is to find a probability distribution of all unigrams that appear in the original documents. The highest frequency words (after removing stop words) have a high likelihood of appearing in human-authored summaries (Nenkova & Vanderwende, 2005). However, the raw

Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pages 49–54,
Portland, Oregon, June 23, 2011.

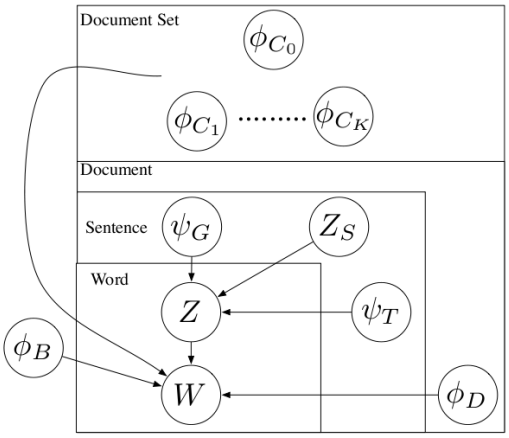

Figure 1: The graphical model for HIERSUM (Haghighi & Vanderwende, 2009).

unigram distribution may contain words that appear frequently in one document, but do not reflect the content of the document set as a whole.

Probabilistic topic models provide a more principled approach to finding a distribution of content words. This idea was first presented by Daumé and Marcu (2006) for their BAYESUM system for query-focused summarization, and later adapted for non-query summarization in the TOPICSUM system by Haghighi and Vanderwende (2009). [1] In these systems, each word from the original documents is drawn from one of three vocabulary distributions. The first, ϕ_b, is the background distribution of general English words. The second, ϕ_d, contains vocabulary that is specific to that one document. And the third, ϕ_c, is the distribution of content words for that document set, and contains relevant words that should appear in the generated summary.

HIERSUM (Haghighi & Vanderwende, 2009) adds more structure to TOPICSUM by further splitting the content distribution into multiple sub-topics. The content words in each sentence can be generated by either the general content topic or the content sub-topic for that sentence, and the words from the general content distribution are considered when building the summary.

[1] The original BAYESUM can also be used without a query, in which case, BAYESUM and TOPICSUM are the exact same model.

3 KL Selection

The KL-divergence between two unigram word distributions P and Q is given by $KL(P||Q) = \sum_w P(w) \log \frac{P(w)}{Q(w)}$. This quantity is used for summary sentence selection in several systems including Lerman and McDonald (2009) and Haghighi and Vanderwende (2009), and was used as a feature in the discrimitive sentence ranking of Daumé and Marcu (2006).

TOPICSUM and HIERSUM use the following KL objective, which finds $\mathbf{S}*$, the summary that minimizes the KL-divergence between the estimated content distribution ϕ_c and the summary word distribution $P_\mathbf{S}$:

$$\mathbf{S}* = \min_{\mathbf{S}:|\mathbf{S}|\leq L} KL(\phi_c||P_\mathbf{S})$$

A greedy approximation is used to find $\mathbf{S}*$. Starting with an empty summary, sentences are greedily added to the summary one at a time until the summary has reached the maximum word limit, L. The values of $P_\mathbf{S}$ are smoothed uniformly in order to ensure finite values of $KL(\phi_c||P_\mathbf{S})$.

4 Why Document-Specific Words are a Problem

The KL selection objective effectively ensures the presence of highly weighted content words in the generated summary. But it is asymmetric in that it allows a high proportion of words in the summary to be words that appear infrequently, or not at all, in the content word distribution. This asymmetry is the reason why the KL selection metric does not sufficiently keep document-specific words out of the generated summary.

Consider what happens when a document-specific word is included in summary $\mathbf{S}$. Assume that the word w_i does not appear (has zero probability) in the content word distribution ϕ_c, but does appear in the document-specific distribution ϕ_d for document d. Then w_i appearing in $\mathbf{S}$ has very little impact on $KL(\phi_c||P_\mathbf{S}) = \sum_j \phi_c(w_j) \log \frac{\phi_c(w_j)}{P_\mathbf{S}(w_j)}$ because $\phi_c(w_i) = 0$. There will be a slight impact because the presence of the word w_i in $\mathbf{S}$ will cause the probability of other words in the summary to be slightly smaller. But in a summary of length 250 words (the

length used for the DUC summarization task) the difference is negligible.

The reason why we do not simply substitute a symmetrical metric for comparing distributions (e.g., Information Radius) is because we want the selection objective to disprefer *only* document-specific words. Specifically, the selection objective should not disprefer background English vocabulary.

5 KL(c)-KL(d) Selection

In contrast to the KL selection objective, our objective measures the similarity of both content and document-specific word distributions to the extracted summary sentences. We combine these measures linearly:

$$\mathbf{S}* = \min_{\mathbf{S}:|\mathbf{S}|\leq L} KL(\phi_c||P_{\mathbf{S}}) - KL(\phi_d||P_{\mathbf{S}})$$

Our objective can be understood in comparison to the MMR criterion by (Carbonell & Goldstein, 1998), which also utilizes a linear metric in order to maximize informativeness of summaries while minimizing some unwanted quality of the extracted sentences (in their case, redundancy). In contrast, our criterion utilizes information about what kind of information should *not* be included in the summary, which to our knowledge has not been done in previous summarization systems.[2]

For comparison to the previous KL objective, we also use a greedy approximation for $\mathbf{S}*$. However, because we are extracting sentences from many documents, the distribution ϕ_d is actually several distributions, a separate distribution for each document in the document set. The implementation we used in our experiments is that, as we consider a sentence s to be added to the previously selected sentences $\mathbf{S}$, we set ϕ_d to be the document-specific distribution of the document that s has been extracted from. So each time we add a sentence to the summary, we find the sentence that minimizes $KL(\phi_c||P_{\mathbf{S}\cup s}) - KL(\phi_{d(s)}||P_{\mathbf{S}\cup s})$. Another implementation we tried was combining all of the ϕ_d distributions into one distribution, but we did not notice any difference in the extracted summaries.

[2]A few anonymous reviewers asked if we tried to optimize the value of λ for $KL(\phi_c||P_{\mathbf{S}}) - \lambda KL(\phi_d||P_{\mathbf{S}})$. The answer is yes, but optimizing λ to maximize ROUGE results in summaries that are perceptibly worse, and manually tuning λ did not seem to produce any benefit.

6 Evaluation

6.1 Data

We developed our sentence selection objective using data from the Document Understanding Conference[3] (DUC) 2006 summarization task, and used data from DUC 2007 task for evaluations. In these tasks, the system is given a set of 25 news articles related to an event or topic, and needs to generate a summary of under 250 words from those documents.[4] For each document set, four human-authored summaries are provided for use with evaluations. The DUC 2006 data has 50 document sets, and the DUC 2007 data has 45 document sets.

6.2 Automatic Evaluation

Systems are automatically evalatued using ROUGE (Lin, 2004), which has good correlation with human judgments of summary content. ROUGE compares n-gram recall between system-generated summaries, and human-authored reference summaries. The first two metrics we compare are unigram and bigram recall, R-1 and R-2, respectively. The last metric, R-SU4, measures recall of skip-4 bigrams, which may skip one or two words in between the two words to be measured. We set ROUGE to stem both the system and reference summaries, scale our results by 10^2 and present scores with and without stopwords removed.

The ROUGE scores of the original HIERSUM system are given in the first row of table 1, followed by the scores of HIERSUM using our KL(c-d) selection. The KL(c-d) selection outperforms the KL selection in each of the ROUGE metrics shown. In fact, these results are statistically significant over the baseline KL selection for all but the unigram metrics (R-1 with and without stopwords). These results show that our KL(c-d) selection yields significant improvements in terms of ROUGE performance, since having fewer irrelevant words in the summaries leaves room for words that are more relevant to the content topic, and therefore more likely to appear in the reference summaries.

The last two rows of table 1 show the scores of two recent state-of-the-art multi-document sum-

[3]http://duc.nist.gov/

[4]Some DUC summarization tasks also provide a query or focus for the summary, but we ignore these in this work.

System	ROUGE w/o stopwords			ROUGE w/ stopwords		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
HIERSUM w/ KL	34.6	7.3	10.4	43.1	9.7	15.3
HIERSUM w/ KL(c)-KL(d)	35.6	**9.9**	**12.8**	43.2	**11.6**	**16.6**
PYTHY	35.7	8.9	12.1	42.6	11.9	16.8
HYBHSUM	35.1	8.3	11.8	45.6	11.4	17.2

Table 1: ROUGE scores on the DUC 2007 document sets. The first two rows compare the results of the unigram HIERSUM system with its original and our improved selection metrics. Bolded scores represent where our system has a significant improvement over the orignal HIERSUM. For further comparison, the last two rows show the ROUGE scores of two other state-of-the-art multi-document summarization systems (Toutanova et al., 2007; Celikyilmaz & Hakkani-Tur, 2010). See section 6.2 for more details.

marization systems. Both of these systems select sentences discriminatively on many features in order to maximize ROUGE scores. The first, PYTHY (Toutanova et al., 2007), trains on dozens of sentence-level features, such as n-gram and skip-gram frequency, named entities, sentence length and position, and also utilizes sentence compression. The second, HYBHSUM (Celikyilmaz & Hakkani-Tur, 2010), uses a nested Chinese restaurant process (Blei et al., 2004) to model a hierarchical content distribution with more complexity than HIERSUM, and uses a regression model to predict scores for new sentences.

For both of these systems, our summaries are significantly better for R-2 and R-SU4 without stopwords, and comparable in all other metrics.[5] These results show that our selection objective can make a simple unsupervised model competitive with more complicated supervised models.

6.3 Manual Evaluation

For manual evaluation, we performed a pairwise comparison of summaries generated by HIERSUM with both the original and our modified sentence selection objective. Users were given the two summaries to compare, plus a human-generated reference summary. The order that the summaries appeared in was random. We asked users to select which summary was better for the following questions:[6]

System	Q1	Q2	Q3	Q4
HIERSUM w/ KL	29	36	31	36
. . . w/ KL(c)-KL(d)	58	51	56	51

Table 2: Results of manual evaluation. Our criterion outperforms the original HIERSUM for all attributes, and is significantly better for Q1 and Q3. See section 6.3 for details.

Q1 Which was better in terms of overall content?

Q2 Which summary had less repetition?

Q3 Which summary was more coherent?

Q4 Which summary had better focus?

We took 87 pairwise preferences from participants over Mechanical Turk.[7] The results of our evaluation are shown in table 2. For all attributes, our criterion performs better than the original HIERSUM selection criterion, and our results for Q1 and Q3 are significantly better as determined by Fisher sign test (two-tailed P value < 0.01).

These results confirm that our objective noticably improves the content of extractive summaries by selecting sentences that contain less document-specific

[5] Haghighi and Vanderwende (2009) presented a version of HIERSUM that models documents as a bag of bigrams, and provides results comparable to PYTHY. However, the bigram HIERSUM model does not find consistent bags of bigrams.

[6] These are based on the manual evaluation questions from DUC 2007, and are the same questions asked in Haghighi and Vanderwende (2009).

[7] In order to ensure quality results, we asked participants to write a sentence on why they selected their preference for each question. We also monitored the time taken to complete each comparison. Overall, we rejected about 25% of responses we received, which is similar to the percentage of responses rejected by Gillick and Liu (2010).

information. This leaves more room in the summary for content that is relevant to the main idea of the document set (Q1) and keeps out content that is not relevant (Q4). Additionally, although neither criterion explicitly addresses coherence, we found that a significant proportion of users found our summaries to be more coherent (Q3). We believe this may be the case because the presence of document-specific information can distract from the main ideas of the summary, and make it less likely that the extracted sentences will flow together.

There is no immediate explanation for why users found our our summaries less repetitive (Q2), since if anything the narrowing of topics due to the negative $KL(\phi_d||P_{\mathbf{S}})$ term should make for more repetition. We currently hypothesize that the improved score is simply a spillover from the general improvement in document quality.

7 Conclusion

We have described a new objective for sentence selection in extractive multi-document summarization, which is different in that it explicitly gives negative weight to sentences that contain document-specific words. Our objective significantly improves the performance of an existing summarization system, and improves on current best ROUGE scores with significance.

We have observed that while the content in our extracted summaries is often comparable to the content in human-written summaries, the extracted summaries are still far weaker in terms of coherence and repetition. Even though our objective significantly improves coherence, more sophisticated methods of decoding are still needed to produce readable summaries. These problems could be addressed through further refinement of the selection objective, through simplification or compression of selected sentences, and through improving the coherence of generated summaries.

References

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems*.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 335–336). New York, NY, USA: ACM.

Celikyilmaz, A., & Hakkani-Tur, D. (2010). A hybrid hierarchical model for multi-document summarization. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 815–824). Stroudsburg, PA, USA: Association for Computational Linguistics.

Daumé, III, H., & Marcu, D. (2006). Bayesian query-focused summarization. *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 305–312). Morristown, NJ, USA: Association for Computational Linguistics.

Gillick, D., & Liu, Y. (2010). Non-expert evaluation of summarization systems is risky. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 148–151). Stroudsburg, PA, USA: Association for Computational Linguistics.

Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 362–370). Boulder, Colorado: Association for Computational Linguistics.

Lerman, K., & McDonald, R. (2009). Contrastive summarization: an experiment with consumer reviews. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (pp. 113–116). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. *Proceedings of*

the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona, Spain.

Nenkova, A., & Vanderwende, L. (2005). *The impact of frequency on summarization* (Technical Report). Microsoft Research.

Tang, J., Yao, L., & Chen, D. (2009). Multi-topic based query-oriented summarization. *SDM'09* (pp. 1147–1158).

Toutanova, K., Brockett, C., Gamon, M., Jagarlamudi, J., Suzuki, H., & Vanderwende, L. (2007). The PYTHY Summarization System: Microsoft Research at DUC 2007. *Proc. of DUC*.