BLLIP at TAC 2011: A General Summarization System for a Guided Summarization Task

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

1 Introduction

The TAC[1] update summarization task is a shared task in which peer systems summarize a collection of documents, and then summarize a second set of documents assuming the user has already read the first set of documents. In the 2010 and 2011 versions of this task, the summaries additionally are "guided" to include certain aspects for pre-defined categories. For example, if the category of the document collection is a natural disaster, the aspects might include the date and location of the event, and the number of people killed.

For our submission, we run our *general* extractive multi-document summarization system (Mason & Charniak, 2011) on the TAC 2011 data. This system was developed for the earlier DUC-style summarization tasks[2], and has outperformed other state-of-the-art systems in both automatic and manual evaluation. The purpose of this submission is to measure how well a general word frequency-based system does on the guided summarization task, and whether the generated summaries cover the required aspects.

2 Methodology

For the non-update guided summaries, we make no changes from our system as presented in (Mason & Charniak, 2011). Our system is based on the document model of (Haghighi & Vanderwende, 2009), in which each word from the original documents is drawn from one of three vocabulary distributions: general English vocabulary (ϕ_b), vocabulary specific to one document (ϕ_d), and content vocabulary for the document set (ϕ_c). Their system then finds an extractive summary that minimizes $KL(\phi_c||\mathbf{S})$, the KL divergence between the content vocabulary distribution, and the distribution of all words in the extracted summary. The extracted sentences are ordered according to their relative positions in their source documents.

We improve upon that objective by explicitly penalizing summaries that contain content that is specific to individual documents. These measures are linearly combined to get the objective $\min KL(\phi_c||\mathbf{S}) - KL(\phi_d||\mathbf{S})$. For further details about the implementation of our system and its performance on the DUC 2008 dataset, see (Mason & Charniak, 2011).

For the update summaries, we train a new document model on the second set of documents, and find updated topics ϕ_{bu}, ϕ_{du}, and ϕ_{cu}. In order to not contain earlier information, we include sentences from the earlier summary, $\mathbf{S}$ in the objective along with the candidate update summaries $\mathbf{Su}$. We then find the extractive summary from the updated documents that minimizes $KL(\phi_{cu}||\mathbf{S}+\mathbf{Su}) - KL(\phi_{du}||\mathbf{S}+\mathbf{Su})$.

3 Results

The TAC dataset provided contains 44 document collections, with 20 documents in each. The documents are split equally into two sets, where all the documents in the first set chronologically precede the documents in the second. Submissions to the guided summarization task contain for each document collection, a summary 'A' on the first docu-

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

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

Metric	Run 28-A	Run 28-B
Pyramid		
Average Modified Pyramid Score (Rank)	0.446 (5)	0.321 (14)
Average numSCUs (Rank)	5.682 (9)	3.545 (15)
Average numrepetitions (Rank)	1.136 (32)	0.523 (30)
s Macroaverage modified score w/ 3 models (Rank)	0.441 (5)	0.317 (14)
Assessor Scores		
Average Overall Response (Rank)	3.000 (10)	2.341 (23)
Average Linguistic Quality (Rank)	2.955 (20)	2.864 (20)

Table 1: Manual evaluation results for our submission, with ranks relative to 50 submitted runs. Note that for average numrepetitions, lower scores are better (eg, Run 28-A has the 18th highest number of repetitions, so its rank is 32).

ment set, and an update summary 'B' on the second set (assuming the user has already read the documents in the first set).

There were 50 submissions in total. The ID for our system is 28. Table 1 shows the results of manual evaluation for the summaries.

For the 'A' summaries, our system ranks competitively among its peers at covering required aspects, as measured by the Pyramid scores. The Pyramid scores measure coverage of manually defined summary content units (SCUs) (Nenkova et al., 2007). We rank 5th out of 50 submissions for average modified Pyramid score, and 9th for the average number of SCUs. However, our system's summaries are relatively weak in linguistic quality, and have more redundancies than the average submitted summaries. Due to these readability issues, our summaries are only ranked 10th overall.

Our 'B' summaries (update summaries) did not score as highly as the 'A' summaries for content. This is partially due to a mistake in our update objective. The update summary should be written assuming the user has already read the previous *document set* – while our update objective only assumes the user has read the previous *summary* based on those documents.

4 Conclusion

We participated in the TAC 2011 guided summarization task using a state-of-the-art general summarization system. Although our system does not explicitly model categories and aspects, its performance for content is still competitive with the other peer systems. The weakest quality of the summaries is the amount of redundancy. For future work, the summaries could be improved by compressing the extracted sentences to remove redundant information, and reordering the sentences in order to make the summaries easier to read.

References

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.

Mason, R., & Charniak, E. (2011). Extractive multi-document summaries should explicitly not contain document-specific content. *ACL-HLT 2011 Workshop: Automatic Summarization for Different Genres, Media, and Languages*.

Nenkova, A., Passonneau, R., & McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, *4*.