Query-focused supervised sentence ranking for update summaries

Seeger Fisher and Brian Roark
Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
{fishers,roark}@cslu.ogi.edu

Abstract

We present a supervised sentence ranking approach for use in extractive update summarization. We use the same general machine learning approach described in earlier DUC papers, and adapt it to the update summarization task. The system proves adaptable enough to be effective at query-focused update summaries.

1 Introduction

Our approach to the update summarization task is similar to our approach to the query-focused multi-document summarization task, with some changes to how we filter redundant sentences within a summary. The approach is a form of extractive summarization. Sentence extraction summarization systems take as input a collection of sentences (one or more documents) and select some subset for output into a summary. This is best treated as a sentence ranking problem, which allows for varying thresholds to meet varying summary length requirements. Most commonly, such ranking approaches use some kind of similarity or centrality metric to rank sentences for inclusion in the summary – see, for example, Lin and Hovy (2002); Erkan and Radev (2004); Radev et al. (2004); Blair-Goldensohn (2005); Biryukov et al. (2005); Mihalcea and Tarau (2005) and the references therein. Such an approach is typically preferred over supervised ranking approaches for reasons of domain independence.

We present an alternative approach, whereby a number of similarity/centrality metrics are used, not directly to rank the sentences, but rather as features within a supervised machine learning paradigm. Since the features themselves are not domain-specific, the benefit of domain generality is retained, while still accruing the benefits of supervised learning.

We examine this approach within the context of query-focused multi-document summarization, for which there is much less training data for supervised approaches than query-neutral multi-document summarization. We address this through the use of two separate ranking models: one trained on a large collection of document clusters and associated (query-neutral) manual summaries; the other trained on a smaller data set from the 2005 and 2006 DUC query-focused multi-document summarization task, which includes document clusters, queries, and the associated (query-focused) manual summaries. The scores from the first ranker are used as features in the second ranker. In addition to the use of two ranking models, we achieve query responsiveness by skewing the word distributions, which make up the features of our models, towards the query. All of this is achieved within a very general supervised ranking paradigm, which is robust and domain independent.

We broke the query-directed summarization problem down into three tasks:

1. Text normalization and sentence segmentation
2. Sentence ranking
   a. query-neutral ranking
   b. query-focused ranking
3. Sentence selection from a ranked list

In previous papers we have detailed the architecture and training of our query-focused multi-document summarization system (Fisher and Roark, 2006; Fisher and Roark, 2007). In this paper we review the previous non-update summary system, and then show how we modified our approach to effectively handle update summaries. The only changes we made for updating were to the third part, sentence selection from a ranked list.

2 Sentence Extraction System

The several stages of our sentence extraction system are detailed in Fisher and Roark (2006). We give just a brief review of the stages here.

2.1 Text normalization

In the multi-document summarization data[1] made available for the Document Understanding Conferences (DUC), each document set is a collection of individual articles, each article in its own file. We created one large text file for each document set by concatenating the raw content text from each article, discarding the meta-data. We then used a simple algorithm to perform sentence segmentation, making use of a list of common abbreviations extracted from the Penn Treebank.

2.2 Supervised sentence ranking

For sentence ranking, we implemented a perceptron ranker (Crammer and Singer, 2001). The objective we used for our supervised ranking is the ROUGE-2 score as configured for the DUC-06 evaluation. For a 250 word summary we are typically only interested in the top 15 or so sentences in a document set (while allowing for redundancy). As a result, we configured the perceptron ranking algorithm to produce models with only 3 ranks. Within each document cluster, feature values were normalized.

Using a limited feature set, the algorithm cannot converge to perfect ranking performance on the training set. We experimented with n-gram features, but although this allowed the perceptron to converge to the training data very accurately, it did not improve ranking performance against our held-out training data. We also experimented with a second order polynomial kernel for the perceptron. This also helped the perceptron to converge, but it did not significantly help with accuracy on the heldout data. See Fisher and Roark (2006) for further details.

2.2.1 Query-neutral sentence ranking

The base feature set that we use is the same as was used in our baseline system from DUC 2005 and DUC 2006 (Fisher and Roark, 2006). For every cluster of documents c in the set of clusters $\mathcal{C}$ comprising the training set, let $\mathcal{Z}_c$ be the collection of manual summaries for that cluster. Let $s \in c$ be the sentences in cluster c and $z \in Z_c$ be the sentences in the summaries of cluster c. For every cluster $c \in \mathcal{C}$ we scored each sentence $s \in c$ as follows

$$\rho(s) \quad = \quad \operatorname*{average}_{z \in Z_c} (\text{rouge}(s, z))$$

where rouge(s, z) is the ROUGE score (Lin, 2004) of sentence s with z as the reference summary. We calculated this value for all sentences in each cluster of the DUC 2001-2003 training data for summaries of size 100, 200 and 400 words, giving us our "gold standard" ranking for use in training the base system.

1.	average tf.idf	6.	average logodds
2.	sum tf.idf	7.	sum logodds
3.	average loglike	8.	sum (max 3) logodds
4.	sum loglike	9.	Sentence position
5.	sum (max 3) loglike	10.	centrality

Table 1: Base feature set

For each sentence in a cluster, we extracted a small number of features for ranking. Most of these features are aggregated from word-based features. Word-based features were of three varieties: TF*IDF, log likelihood ratio, and log odds ratio statistics. The feature set is summarized in Table 1. See Fisher and Roark (2006; Fisher and Roark (2007) for details on calculation of the features.

Beyond these base features, we added the features from Table 1 for both the immediately previous and immediately following sentences as features for the current sentence, effectively tripling the number of features.

Using multiple similarity metrics as features is useful because all of these features score co-occurrence dependencies differently.

2.2.2 Query-focused sentence ranking

Skewing word distributions

To achieve query-sensitivity within the context of a single supervised ranking system, we examined skewing word distributions towards the query for purposes of calculating distribution sensitive features. Recall that we have a number of features (see table 1) that rely on the distribution of a word in the

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

document set relative to its distribution in the corpus. We skew the word distributions towards the query in a document set by adding the counts of each of the non-stop query words, multiplied by an empirically determined factor, to the counts of words in the document set. In effect, non-stop query words have their counts increased in the document set for purposes of calculating the word-distribution sensitive features. The result is that when extracting features from a sentence, words that are in the query will have relatively larger feature values, by virtue of having higher document set counts. When the individual words have larger values, the feature values for sentences containing those words will also be higher.

Note that this approach allows us to train the models on non-skewed training data, with the query-focused skewing happening at test time. Hence, large amounts of query-neutral multi-document summarization training data can be exploited. With this approach, we can get query sensitivity within a very simple ranking approach. This has the additional benefit of being able to convert the ranking score to a normalized probability (via softmax), thus allowing the use of these scores as features in another stage of ranking.

Re-ranking

The first-pass ranking model in our approach is trained on query-neutral summarization data. Given that we now have query-sensitive training data from the DUC-2005 and 2006 evaluation set, we can build a specifically query-focused reranker from this data. As with the query-neutral ranking, we used the perceptron ranking algorithm.

The sentences are first ranked using the skewing approach described above, and the output from this step (the softmax normalized perceptron score) is one of the features input to the reranker. In addition to this feature, which has its weight empirically fixed, the reranker has two other sets of features for which it learns parameter weights. These are features characterizing the number of non-stop query words in the sentence. We first partition the set of non-stop query words into two subsets: those with log likelihoods higher than a fixed threshold and those with log likelihoods lower than the threshold. The log likelihood is calculated for each query word for that cluster, using unskewed counts. Then, for each subset s, there are five indicator features: 0 words in the sentence from s; at least 1 word in the sentence from s; at least 2 words from s; at least 3 words; and at least 4 words. For the trials reported here, the partitioning threshold was set empirically at 10. See Fisher and Roark (2006) for further details on this approach.

For training the reranker, we used the DUC-2005 document sets as training data, and the DUC-2006 document sets as development data for testing different features. We fixed the weight of the baseline ranker at 1000.

Query expansion

Besides skewing word distributions towards the query, and then performing re-ranking with query-based features, we also performed query expansion to make our sentence ranking more sensitive to the query. We used a 300 million word corpus to estimate the probability that two words will occur in adjacent sentences. We picked the 100 non-stop words with the highest log-likelihood as expansion terms for each query term. These expansion terms were included in the re-ranking step described above, but as separate features from the actual query terms. For more details see Fisher and Roark (2007).

2.3 Sentence selection

At the sentence selection stage, we removed any sentence less than 5 words or greater than 50 words in length. The restriction on being too short is based on the intuition that in an extraction system, anything too short will be meaningless out of context. The restriction on being too long is a simple way to keep the system from extracting long lists, which generally do not make a good summary. In addition, any sentence that begins or ends with a quotation mark was also filtered out. Finally, sentences beginning with a pronoun were removed, to avoid the most obvious cases of poor anaphora resolution.

At this point we also applied some simple compression to the remaining sentences. Namely, we removed any paired parentheticals, defined as stretches of text in a sentence that were delimited by parentheses, single dashes, or em-dashes.

Sentences were selected in order based on the final ranking, until the summary size limit was reached, with some sentences being removed for lack of novelty, as follows. Stop-words were removed from a candidate sentence, then the bigram overlap with non-stop words already in the summary was calculated. If the overlap amounted to 65 percent or less of the non-stop words in the candidate (determined empirically), the candidate was added to the summary, otherwise it was discarded. Finally, we ordered the extracted sentences by document-id, and then by order they occurred in the document.

3 Update Summaries

Our system for producing update summaries is very similar to the query-focused multi-document summarizer we fielded at DUC 2007, see Fisher and Roark (2007) , but with some important differences in sentence selection from the ranked list of sentences. We use the same classifier and feature set, trained in the same way as the DUC 2007 main task summarizer. For summaries of a first partition, the 2008 system was identical to the 2007 system, excepting that the summary is shorter. For the other partition, we allowed the system to rank sentences in the same way. However, in the sentence selection stage when checking for overlap between a candidate sentence and the sentences in the summary so far, we checked not only against sentences already in the new summary, but also against sentences from the summary of the first partition. Thus, there was no change to our ranking algorithm, only to the part of the system that adds already ranked sentences to the growing summary.

4 TAC 2008 Results

The OGI-08 system was fairly average in the field of participants in TAC 2008, which is unsurprising given that the system was unchanged from the system we fielded in the update summarization pilot at DUC in 2007. There are quite a few different evaluation metrics used at DUC. Our system scores ranged from somewhat better than the mean, to the bottom of the 3rd quartile of submitted systems, depending on which metric is used. Given that our system is unchanged from the pilot, where we consistently ranked in the top 3rd, this shows that a number of the other systems substantially improved.

5 Summary and future directions

We have presented the application of general supervised machine learning techniques to the problem of sentence ranking for extractive summarization. By exploiting model summaries to define a gold-standard ranking over sentences, we can use well-motivated learning approaches, which handle an arbitrary number of features. We have demonstrated that many common metrics used for sentence ranking can be combined into a single ranking model that provides better performance than any of the metrics in isolation. We straightforwardly extended the model to include features of neighboring sentences, which was demonstrated to improve performance. We have applied this approach to query-directed summarization through a number of techniques: (1) query word count inflation; (2) reranking based on query-directed training data; and (3) query expansion techniques. The resulting approach is highly competitive, and its generality and ease of extension should allow for substantial future developments.

There are a number of ways to improve the current system. The feature set for the reranker is an area we will continue to explore, since we have experimented with relatively few different features for the current system. Though including all unigrams as features led to over-fitting, we would like to find a subset of lexical n-gram features that are relevant to indicating importance and applicability to inclusion in a summary. We also want to include features that are indicative of what sort of question the query is. Also, we believe that clause segmentation prior to ranking could lead to substantially better performance. A related set of features to explore are discourse connectives, and how they relate one clause to another.

References

M. Biryukov, R. Angheluta, and M.F. Moens. 2005. Multidocument question answering text summarization using topic signatures. *Journal on Digital Information Management.*

S. Blair-Goldensohn. 2005. Columbia University at DUC 2005. In *Document Understanding Workshop (DUC) 2005.*

K. Crammer and Y. Singer. 2001. Pranking with ranking. In *Neural Information Processing Systems*. NIPS.

G. Erkan and D. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*.

Gunes Erkan. 2006. Using biased random walks for focused summarization. In *Document Understanding Workshop*.

S. Fisher and B. Roark. 2006. Query-focused summarization by supervised sentence ranking and skewed word distributions. In *Proceedings of the Document Understanding Workshop (DUC)*.

S. Fisher and B. Roark. 2007. Feature expansion for query-focused supervised sentence ranking. In *Proceedings of the Document Understanding Workshop (DUC)*.

C.Y. Lin and E. Hovy. 2002. Automated multi-document summarization in NeATS. In *Proceedings of the Human Language Technology Conference*.

C.Y. Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop in Text Summarization, ACL'04*.

R. Mihalcea and P. Tarau. 2005. An algorithm for language independent single and multiple document summarization. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Jahna Otterbacher, Gunes Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 915–922, Vancouver, British Columbia, Canada, October. ACL.

D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. 2004. MEAD - a platform for multidocument multilingual text summarization. In *LREC*, Lisbon, Portugal.