PolyU at TAC 2009

You Ouyang, Wenji Li
Department of Computing
The Hong Kong Polytechnic University
{csyouyang,cswjli}@comp.polyu.edu.hk

1 Introduction

PolyU has participated in both the two tasks in the TAC 2009 summarization track, including the update summarization task and the automatically evaluating summaries of peers (AESOP) task. The update summarization task is to generate short fluent multi-document summaries of news articles. For each topic, a topic statement and two chronologically ordered newswire document sets are given. The task requires generating a 100-word summary for each document set. The purpose of the AESOP task is to promote research and development of systems that automatically evaluate the quality of summaries. The automatic metrics are run on the data and submitted summarization systems from the update summarization task and compared to manual evaluations.

In this year, we mainly study the word-based approaches for both tasks. Simple word-based approaches are first proposed as basic solutions. More sophisticated approaches are then studied by considering the factors beyond words. The systems are detailed in following sections.

2 Update Summarization

2.1 Word-based summarization approach

Word-based summarization systems mainly study the importance of the words to the topic to generate the summaries. Various estimates of the word importance are proposed in previous work. In our system, we use a simple frequency-based estimate for the word importance. Based on the assumption that the words appearing more frequently in the documents of a topic are more likely to represent that topic, we scale the importance of a word w by log $freq_T(w)$, where $freq_T(w)$ is the frequency of w appearing in the documents of topic T. The score of a sentence is then estimated by accumulating the scores of all the words in it as $score(s) = \sum_{w_i \in s} \log freq_T(w_i)$. For the purpose of maximizing the total amount of required information, subject to the given length of summaries, the score is normalized by the length of the sentence, denoted by $len(s)$, i.e., sentences are actually ranked by $score(s)/len(s)$.

The same as most typical extractive summarization systems, the sentences in the topic are first ranked according to the estimated scores and then selected into the summary by descending order. Post-processing approaches are also applied to improve the readability of the summary. For the redundancy removal issue, we adopted the famous Maximum Marginal Relevance approach (Carbonell and Goldstein, 1998). The sentences are selected iteratively that each round the candidate sentence will be selected only when it is not too similar to any sentences already in the summary. On the other hand, sentence re-ordering technique (Barzilay et al., 2002) is used to make the summary more fluent. In our approach, the selected sentences are re-ordered chronologically. Here we simply use the date of the newswire document as the temporal information of the sentence. For two sentences in the same document, they are ordered by their original order in the document. To the requirement of the task, the length of the generated summary is strictly controlled to 100 words.

2.2 Summarizing beyond Words: Concept Hierarchical

The initial system introduced above is a typical word-based extractive summarization system. As a further study, we also submitted a hierarchical system based on studying the relations between the words. A word hierarchical is used to represent the relations. The main idea of the approach is to employ a hierarchical summarization process which is motivated by the behavior of a human summarizer. While the document set may be very large in multi-document summarization, the length of the summary to be generated is usually limited. So there are always some concepts that can not be

included in the summary. A natural thought is that more general concepts should be considered first. So, when a human summarizer faces a set of many documents, he may follow a general-specific principle to write the summary. The human summarizer may start with finding the core topics in a document set and write some sentences to describe this core topic. Next he may go to find the important sub-topics and cover the subtopics one by one in the summary, then the sub-sub-topics, sub-sub-sub-topics and so on. By this process, the written summary can convey the most salient concepts. Motivated by this experience, we propose a hierarchical summarization approach which attempts to mimic the behavior of a human summarizer. The approach includes two phases. In the first phase, a hierarchical tree is constructed to organize the important concepts in a document set following the general-to-specific order.

To construct a hierarchical representation for the words in a given document set, we follow the idea introduced by Lawrie et al. (2001) who use the subsuming relation to express the general-to-specific structure of a document set. A subsumption is defined as an association of two words if one word can be regarded as a sub-concept of the other one. In our approach, the pointwise mutual information (PMI) is used to identify the subsumption between words instead of the probability used in (Lawrie et al., 2001). Generally, two words with a high PMI is regarded as related. Using the identified relations, the word hierarchical tree is constructed in a top-bottom manner. Two constraints are used in the tree construction process:

(1) For two words related by a subsumption relation, the one which appears more frequently in the document set serves as the parent node in the tree and the other one serves as the child.

(2) For a word, its parent node in the hierarchical tree is defined as the most related word, which is identified by PMI.

An example of a tree fragment is demonstrated below by Figure 1. The tree is constructed on the document set D0701A from DUC 2007, the query of this document set is "Describe the activities of Morris Dees and the Southern Poverty Law Center".

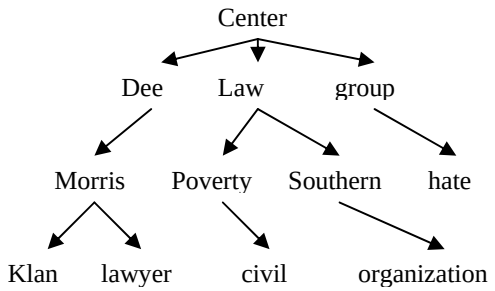

Figure 1. Example of word hierarchical

Based on the word hierarchical tree and the estimated word significance, we propose an iterative algorithm to select sentences which is able to integrate the multiple objectives for composing a relevant, concise and fluent summary. The algorithm follows a general-to-specific order to select sentences into the summary. In the implementation, the idea is carried out by following a top-down order to cover the words in the hierarchical tree. In the beginning, we consider several "seed" words which are in the top-level of the tree (these words are regarded as the core concepts in the document set). Once some sentences have been extracted according to these "seed" words, the algorithm moves to down-level words through the subsumption relations between the words. Then new sentences are added according to the down-level words and the algorithm continues moving to lower levels of the tree until the whole summary is generated. For the purpose of reducing redundancy, the words already covered by the extracted sentences will be ignored while selecting new sentences. To improve the fluency of the generated summary, after a sentence is selected, it is inserted to the position according to the subsumption relation between the words of this sentence and the sentences which are already in the summary. The detailed process of the sentence selection algorithm is described below.

Algorithm 1: Summary Generation

1: For the words in the hierarchical tree, set the initial states of the top n words as "activated" and the states of other words as "inactivated".

2: For all the sentences in the document set, select the sentence with the largest score according to the "activated" word set. The score of a sentence s is defined as $score(s) = \frac{1}{|s|}\sum score(t_i)$ where t_i is a word

belongs to s and the state of t_i should be "activated". | s | is the number of words in s.
3: For the selected sentence s_k, the subsumption relations between it and the existing sentences in the current summary are calculated and the most related sentence s_l is selected. s_k is then inserted to the position right behind s_l.
4: For each word t_i belongs to the selected sentence s_k, set its state to "inactivated"; for each word t_j which is subsumed by t_i, set its state to "activated".
5: Repeat step 2-4 until the length limit of the summary is exceeded.

3 Automatically Evaluating Summaries of Peers (AESOP)

The basic idea of the proposed evaluation schemes is similar to famous ROUGE and Pyramid, matching the concepts of the systems summaries and the human summaries.

3.1 Word-based Evaluation Theme

In our opinion, simpler evaluation criteria are preferred because they may be more adaptive to various evaluation environments. Considering the evaluation task given the human summaries, we believe that the matching scheme between system summaries and human summaries is essential for defining a evaluation criterion.

We still start with a word-based framework. If a system summary contains a word which also appears in human summaries, we regard the word as a "hit". When a summary gets a hit, its relevance to the humans summaries will increase. In the implement of the evaluation scheme, we first calculate the frequency of the words in all the human summaries (4 human summaries for NoModels track and 3 for AllPeers track), then a scoring scheme is used to calculated the significance of a hit by defining the significance as $freq_H(w)$, where $freq_H(w)$ is the number of human summaries containing the word w. The score of a system summary is then estimated by accumulating the scores of all the words in it, formulated as $score(S) = \sum_{w_i \in S} f(w_i)$. f is a function for measuring the importance of a hit by the frequency. We adopted three different functions in our implementations, including the original frequency, an exponential function and a coverage-based function. The formulas are given below as
(1) $f_1(w_i) = freq_H(w_i)$;
(2) $f_2(w_i) = 2^{freq_H}(w_i)$;
(3) $f_3(w_i) = freq_H(w_i)/N$, here N is the total number of words in the human summaries which have the same frequency with w_i.

3.2 Sentence-based Evaluation Theme

Besides words, we also considered the idea of using sentences as the matching units in evaluations. Unlike documents which may contain many concepts and topics, a sentence usually focuses on only one topic. The assumption of the sentence-based evaluation scheme is regarding the sentences in the human summaries as concepts, then examines the ability of a system summary on covering these sentences.

In the implementation of the evaluation scheme, the pairwise similarities between the sentences of a system summary and a human summary are first calculated by the following formula, i.e.

$$sim(s_a, s_h) = \frac{|s_a \cap s_h|}{|s_h|}$$

where s_a or s_h indicates a sentence in a system summary or a human summary respectively.

By the calculated similarities, a matching theme between the sentences can be established as illustrated in Figure 2.

Based on the matching theme, a criterion for evaluating the system summary can be given. In the evaluation theme, to each sentence s_h in the human summary, we find a sentence s_a from the sentence set of the system summary which can the best cover it (the one with the largest similarity). The score of a system summary S given a human summary H is then calculated by summing the matching degrees of all the sentences in H as

$$score(S \mid H) = \sum_{s_h \in H} \max_{s_a} (sim(s_a, s_h)).$$

Then the overall score of S under all the human summaries is calculated by averaging the pairwise scores.

The upper formula reflects the ability of the system summary in covering the sentences of the human summary. Since we regard sentences as concepts, it is indeed a recall-based measure of the coverage of the concepts of the human summary.

As a further consideration, the order of the sentences in the system summary can also be

evaluated by the sentence matching scheme. Regarding the sentences in the human summary are perfectly ordered, the order of sentences in the system summary can be evaluated according to the order of the corresponding sentences in human summary. However, since the matching between the sentences is not very accurate, we did not implement this issue in the submitted evaluation criterion.

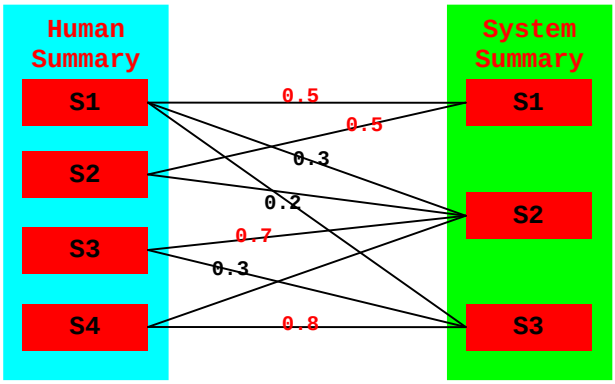

Figure 2. Example of sentence matching theme. The overall score equals to 0.5 + 0.5+ 0.7 + 0.8 = 2.5

4 TAC 2009 Results

In the update summarization track of TAC 2009, a total of 52 runs are submitted. Three NIST baseline systems are also included in the evaluation. We submitted two runs to this track, i.e. the Polyu1 system which followed the simple word-based framework and the Polyu2 system which followed the hierarchical-based framework. The results of the TAC evaluations are listed in Table 1. The "Best" in the table indicates the best result in all the submitted systems. In the results, the ranks of our systems are between 10th and 20th which is an acceptable result, given the fact that we did not include any sentence trimming strategies in our systems. As a good thing, the hierarchical-based system PolyU2 performed better than the word-based system which might have proved the benefit of using word relations.

In the update AESOP track of TAC 2009, a total of 35 runs are submitted. Two NIST baseline systems are also included in the evaluation, i.e., the ROUGE-4 and BE evaluations. We submitted four runs to this track, i.e. the Polyu1 system which followed the sentence-based framework and the other three also followed the word-based framework, with different functions for estimating the word significance. The results of the TAC evaluations are listed in Table 2. The "Best" in the table indicates the best result in all the submitted systems. In the results, our system performed very well in the NoModel track, very close to the best performing system. However, in the AllPeers track, the performance was not so well. The reason is that to enable the evaluation on human summaries, a jackknife strategy was adopted by comparing the human summary to be evaluated to only three other human summaries. In contrast, each summary is compared to four human summaries in the NoModel track. Naturally, the usage of more human summaries can benefit the accuracy of the evaluation since more relevant information is covered.

The Poly1 system which used sentence-based evaluation scheme performed worse than the other three systems which used the word-based evaluation scheme. This means that the sentence-based framework still need more work on it.

	PolyU1	PolyU2	Best
A Rouge-2	0.091	**0.095**	0.122
A ROUGE-SU4	0.129	**0.139**	0.151
A Response	3.886	**4.091**	5.159
A Pyramid	**0.274**	0.261	0.383
B Rouge-2	0.085	**0.087**	0.104
B ROUGE-SU4	0.125	**0.131**	0.140
B Response	3.659	3.591	5.023
B Pyramid	**0.200**	0.192	0.307

Table 1. Update Summarization Results.
A and B indicate the two topic sets respectively.

	PolyU1	PolyU2	PolyU3	PolyU4	Best
A AP Pyramid	0.839	0.884	0.884	**0.885**	0.982
B AP Pyramid	0.843	0.841	0.856	**0.858**	0.976
A NM Pyramid	0.887	0.965	0.962	**0.967**	0.978
B NM Pyramid	0.905	0.944	**0.968**	0.962	0.970
A AP Response	0.758	0.801	0.800	**0.802**	0.968
B AP Response	0.722	0.729	0.733	**0.739**	0.957
A NM Response	0.767	0.846	0.853	**0.856**	0.872
B NM Response	0.718	0.821	0.814	**0.825**	0.833

Table 2. AESOP Results. Pearson's R is reported.
AP and NM indicate AllPeers and NoModel respectively

5 Conclusion

We proposed several word-based summarization systems to the update summarization track and evaluation schemes to the ASEOP task. Results showed that word-based approaches are effective in both summarization systems and evaluations. Approaches beyond words are also examined, including a hierarchical system for update summarization and a sentence-matching–based evaluation for ASEOP. As a matter of fact, the sophisticated approaches still need to be further studied.

References

R. Barzilay, N. Elhadad, and K. R. McKeown. 2002. *Inferring strategies for sentence ordering in multidocument news summarization.* Journal of Artificial Intelligence Research, 17:35-55, 2002.

J. Carbonell and J. Goldstein. 1998. *The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries*. In Proceedings of ACM SIGIR 1998, pp 335-336.

K. Knight and D. Marcu. 2000. *Statistics-based summarization --- step one: Sentence compression.* In Proceeding of The American Association for Artificial Intelligence Conference (AAAI-2000), pp 703-710.

D. Lawrie, W. B. Croft and A. Rosenberg. 2001. *Finding topic words for hierarchical summarization.* In Proceedings of ACM SIGIR 2001, pp 349-357.

C. Lin and E. Hovy. 2003. *Automatic evaluation of summaries using n-gram co-occurance statistics*. In Proc. of HLT-NAACL 2003, pp 71-78.