An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation

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

Abstract

This paper introduces a novel hierarchical summarization approach for automatic multi-document summarization. By creating a hierarchical representation of the words in the input document set, the proposed approach is able to incorporate various objectives of multi-document summarization through an integrated framework. The evaluation is conducted on the DUC 2007 data set.

1 Introduction and Background

Multi-document summarization requires creating a short summary from a set of documents which concentrate on the same topic. Sometimes an additional query is also given to specify the information need of the summary. Generally, an effective summary should be relevant, concise and fluent. It means that the summary should cover the most important concepts in the original document set, contain less redundant information and should be well-organized.

Currently, most successful multi-document summarization systems follow the extractive summarization framework. These systems first rank all the sentences in the original document set and then select the most salient sentences to compose summaries for a good coverage of the concepts. For the purpose of creating more concise and fluent summaries, some intensive post-processing approaches are also appended on the extracted sentences. For example, redundancy removal (Carbonell and Goldstein, 1998) and sentence compression (Knight and Marcu, 2000) approaches are used to make the summary more concise. Sentence re-ordering approaches (Barzilay et al., 2002) are used to make the summary more fluent. In most systems, these approaches are treated as independent steps. A sequential process is usually adopted in their implementation, applying the various approaches one after another.

In this paper, we suggest a new summarization framework aiming at integrating multiple objectives of multi-document summarization. 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 topic 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. Also, the general-specific relation can be used to serve other objectives, i.e. diversity, coherence and etc.

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. In the second phase, an iterative algorithm is proposed to select the sentences based on the constructed hierarchical tree with consideration of the various objectives of multi-document summarization.

2 Word Hierarchical Representation

2.1 Candidate Word Identification

As a matter of fact, the concepts in the original document set are not all necessary to be included in the summary. Therefore, before constructing the hierarchical representation, we first conduct a

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 113–116,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

filtering process to remove the unnecessary concepts in the document set in order to improve the accuracy of the hierarchical representation. In this study, concepts are represented in terms of words. Two types of unnecessary words are considered. One is irrelevant words that are not related to the given query. The other is general words that are not significant for the specified document set. The two types of words are filtered through two features, i.e. *query-relevance* and *topic-specificity*.

The *query-relevance* of a word is defined as the proportion of the number of sentences that contains both the word and at least one query word to the number of sentences that contains the word. If a feature value is large, it means that the co-occurrence rate of the word and the query is high, thus it is more related to the query. The *topic-specificity* of a word is defined as the entropy of its frequencies in different document sets. If the feature value is large, it means that the word appears uniformly in document sets, so its significance to a specified document set is low. Thus, the words with very low query-relevance or with very high topic-specificity are filtered out[1].

2.2 Word Relation Identification and Hierarchical Representation

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. 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:

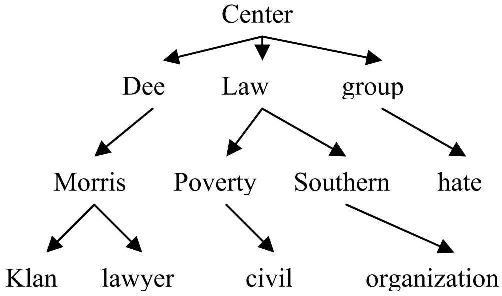
(1) For two words related by a subsumption relation, the one which appears more frequently in the document set serves as the parent node in the tree and the other one serves as the child node.

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

The construction algorithm is detailed below.

```
Algorithm 1: Hierarchical Tree Construction
1: Sort the identified key words by their frequency in the document set in descending order, denoted as $T = \{t_1, t_2, \dots, t_n\}$
2: For each $t_i$, $i$ from $1$ to $n$, find the most relevant word $t_j$ from all the words before $t_i$ in $T$, as $T_i = \{t_1, t_2, \dots, t_{i-1}\}$. Here the relevance of two words is calculated by their PMI, i.e.
   $PMI(t_i, t_j) = \log \frac{freq(t_i, t_j) * N}{freq(t_i) freq(t_j)}$
   If the coverage rate of word $t_i$ by word $t_j$ $P(t_i \mid t_j) = \frac{freq(t_i, t_j)}{freq(t_i)} \geq 0.2$, $t_i$ is regarded as being subsumed by $t_j$. Here $freq(t_i)$ is the frequency of $t_i$ in the document set and $freq(t_i, t_i)$ is the co-occurrence of $t_i$ and $t_j$ in the same sentences of the document set. $N$ is the total number of tokens in the document set.
4: After all the subsumption relations are found, the tree is constructed by connecting the related words from the first word $t_1$.
```

An example of a tree fragment is demonstrated below. The tree is constructed on the document set D0701A from DUC 2007[2], the query of this document set is “Describe the activities of Morris Dees and the Southern Poverty Law Center”.



3 Summarization based on Word Hierarchical Representation

3.1 Word Significance Estimation

In order to include the most significant concepts into the summary, before using the hierarchical tree to create an extract, we need to estimate the significance of the words on the tree first. Initially, a rough estimation of the significance of a word is given by its frequency in the document set. However, this simple frequency-based measure is obviously not accurate. One thing we observe from the constructed hierarchical tree is that a word which subsumes many other words is usually very important, though it may not appear

[1] Experimental thresholds are used on the evaluated data.

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

frequently in the document set. The reason is that the word covers many key concepts so it is dominant in the document set. Motivated by this, we develop a bottom-up algorithm which propagates the significance of the child nodes in the hierarchical tree backward to their parent nodes to boost the significance of nodes with many descendants.

```
Algorithm 2: Word Scoring Theme
1: Set the initial score of each word in $T$ as its log-frequency, i.e. $score(t_i)$ =log $freq(t_i)$.
2: For $t_i$ from $n$ to $1$, propagate an importance score to its parent node $par(t_i)$ (if exists) according to their relevance, i.e. $score(par(t_i)) = score(par(t_i)) + \log freq(t_i, par(t_i))$.
```

3.2 Sentence Selection

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 3: Summary Generation
1: For the words in the hierarchical tree, set the initial states of the top $n$ words[3] 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.
```

4 Experiment

Experiments are conducted on the DUC 2007 data set which contains 45 document sets. Each document set consists of 25 documents and a topic description as the query. In the task definition, the length of the summary is limited to 250 words. In our summarization system, pre-processing includes stop-word removal and word stemming (conducted by GATE[4]).

One of the DUC evaluation methods, ROUGE (Lin and Hovy, 2003), is used to evaluate the content of the generated summaries. ROUGE is a state-of-the-art automatic evaluation method based on N-gram matching between system summaries and human summaries. In the experiment, our system is compared to the top systems in DUC 2007. Moreover, a baseline system which considers only the frequencies of words but ignores the relations between words is included for comparison. Table 1 below shows the average recalls of ROUGE-1, ROUGE-2 and ROUGE-SU4 over the 45 DUC 2007document sets. In the experiment, the proposed summarization system outperforms the baseline system, which proves the benefit of considering the relations between words. Also, the system ranks the 6th among the 32 submitted systems in DUC 2007. This shows that the proposed approach is competitive.

	ROUGE-1	ROUGE-2	ROUGE-SU4
S15	0.4451	0.1245	0.1771
S29	0.4325	0.1203	0.1707
S4	0.4342	0.1189	0.1699
S24	0.4526	0.1179	0.1759

[3] n is set to 3 experimentally on the evaluation data set.

[4] http://gate.ac.uk/

S13	0.4218	0.1117	0.1644
Ours	0.4257	0.1110	0.1608
Baseline	0.4088	0.1040	0.1542

Table 1. ROUGE Evaluation Results

To demonstrate the advantage of the proposed approach, i.e. its ability to incorporate multiple summarization objectives, the fragments of the generated summaries on the data set D0701A are also provided below as a case study.

The summary produced by our system
The Southern Poverty Law Center tracks hate groups, and Intelligence Report covers right-wing extremists. Morris Dees, co-founder of the Southern Poverty Law Center in Montgomery, Ala. Dees, founder of the Southern Poverty Law Center, has won a series of civil right suits against the Ku Klux Klan and other racist organizations in a campaign to drive them out of business. In 1987, Dees won a $7 million verdict against a Ku Klux Klan organization over the slaying of a 19-year-old black man in Mobile, Ala.
The summary produced by the baseline system
Morris Dees, co-founder of the Southern Poverty Law Center in Montgomery, Ala. The Southern Poverty Law Center tracks hate groups, and Intelligence Report covers right-wing extremists. The Southern Poverty Law Center previously recorded a 20-percent increase in hate groups from 1996 to 1997. The verdict was obtained by lawyers for the Southern Poverty Law Center, a nonprofit organization in Birmingham, Ala.

Comparing the generated summaries of the two systems, we can see that the summary generated by the proposed approach is better in coherence and fluency since these factors are considered in the integrated summarization framework. Various summarization approaches, i.e. sentence ranking, redundancy removal and sentence re-ordering, are all implemented in the sentence selection algorithm based on the word hierarchical tree. However, we also observe that the proposed approach fails to generate better summaries on some document sets. The main problem is that the quality of the constructed hierarchical tree is not always satisfied. In the proposed summarization approach, we mainly rely on the PMI between the words to construct the hierarchical tree. However, a single PMI-based measure is not enough to characterize the word relation. Consequently the constructed tree can not always well represent the concepts for some document sets. Another problem is that the two constraints used in the tree construction algorithm are not always right in real data. So we regard developing better tree construction approaches as of primary importance. Also, there are other places which can be improved in the future, such as the word significance estimation and sentence inserting algorithms. Nevertheless, we believe that the idea of incorporating the multiple summarization objectives into one integrated framework is meaningful and worth further study.

5 Conclusion

We introduced a summarization framework which aims at integrating various summarization objectives. By constructing a hierarchical tree representation for the words in the original document set, we proposed a summarization approach for the purpose of generating a relevant, concise and fluent summary. Experiments on DUC 2007 showed the advantages of the integrated framework.

Acknowledgments

The work described in this paper was partially supported by Hong Kong RGC Projects (No. PolyU 5217/07E) and partially supported by The Hong Kong Polytechnic University internal grants (A-PA6L and G-YG80).

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.