Fudan University at DUC 2005

Lin Zhao, Xuanjing Huang, Lide Wu
Dept. of Computer Science and Engineering
Fudan University
{linzhao, xjhuang, ldwu}@fudan.edu.cn

Abstract

In this paper, we described our participation in the question-focused multi-document summarization task of DUC 2005. Our system was based on a supervised machine learning method in which feature extraction was an important issue. We present the whole procedure of our system, focusing on the features we used. We also analyze the results of manual and automatic evaluations.

1. Introduction

This year was the third time that Fudan University participated in the DUC evaluation. We developed a new system which is based on supervised machine learning technique. Specifically, we trained a conditional maximum entropy model to rank the sentences in input documents, and then extract some sentences into the final summary.

In this report, we first describe in detail our system procedure in section 2, and then provide official evaluation results in section 3 and section 4 concludes.

2. System Description

Our system consists of four main components: (1) document preprocessing, including sentence boundary detection, part-of-speech tagging, named entity tagging and etc; (2) extracting features that are regarded useful in determining whether a sentence should be included in the summary or not; (3) training a conditional maximum entropy model to rank the sentences in the documents; (4) redundancy reduction and summary generation. The overall architecture of the system is shown in Figure 1. We will describe some of the main steps in detail.

2.1 Feature extraction

The criterion that we used to judge whether a sentence should be extracted is the informativeness and relevance to the question. Therefore we extract five features from each sentence: sentence position, sentence length, number of name entities in a sentence, similarity with document cluster, and similarity with the question. These features are combined together to decide how possible a sentence belongs to the summary.

2.1.1 Sentence position, length, number of NEs

These three features can be easily extracted from a sentence. Generally speaking, a sentence at the beginning of a document is more likely to be included in the summary, so we used a binary-valued feature which is set to 1 for the first sentence in the document and 0 otherwise.

The second feature is relevant to sentence length since a very short sentence is less possible to be included in the summary. This feature is set to 1 if the sentence is longer than a predefined threshold after removing the non-content words in it, 0 otherwise. In our system, this threshold is set to 5.

It is believed that name entities often contain important information, so we used an automatic

NE tagger to extract them and took the number of NEs in a sentence as the third feature.

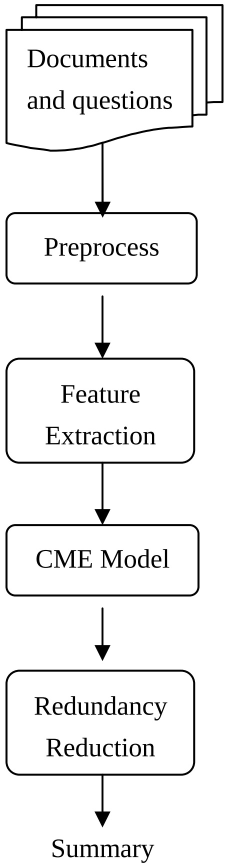

Figure 1. System Architecture

2.1.2 Similarity with cluster and question

The two most important features we used are sentence's similarity with the document cluster and with the question. Since similarity calculation is a key step, we designed three algorithms for it, trying to find the best.

(1) Baseline algorithm: similarity between two sentences is calculated as cosine value of TF*IDF-weighted vectors.

(2) Word similarity based algorithm.

The first algorithm is totally based on bag_of_word method, without considering semantic information, so it cannot deal with such cases that two sentences use completely different words while still have similar meaning. In order to solve this problem, we introduced word similarity based on WordNet into sentence similarity calculation. For details of word similarity calculation, readers can refer to [Budanitsky and Hirst 2001, Budanitsky 1999]. In this way, two arbitrary words that have same part-of-speech can be related together based on their relations described in WordNet. The larger the word similarity is, the more similar the two words are. Thus similarity between two sentences S_1 and S_2 can be calculated as:

$$sim(S_1,S_2) = \frac{\sum_{1\le p\le n} \max_{1\le i\le m} sim(w_{1p},w_{2i}) + \sum_{1\le q\le m} \max_{1\le j\le n} sim(w_{2q},w_{1j})}{m+n}$$

Where $S_1=\{w_{11}, w_{12}, \ldots, w_{1n}\}$, $S_2=\{w_{21}, w_{22}, \ldots, w_{2m}\}$.

(3) Synset based algorithm.

This algorithm is similar to the first algorithm because it is also based on cosine calculation between TF*IDF weighted vectors. The main difference is that for each dimension of a vector, a word synset [Miller et al. 1993] is used instead of the word itself. The concept of “synset” is defined in WordNet as a group of synonymous words. For example, {car, auto, automobile …} is used in place of “car”. In this way, similar words can be put together to represent one concept, and similarity is calculated between concepts instead of words. The limitation at present is that for each word, we only consider its first sense in WordNet. Other processes are the same as in the first algorithm.

We did experiments with the above three algorithms respectively, and from the results, we found that the second algorithm performed worse than the other two, which was unexpected by us. A reasonable explanation may be that the statistic information is totally neglected during semantic analysis. Probably a feasible way is to combine both statistic and semantic methods, such as using TF*IDF as weight of words when calculating the word similarities. As for the third algorithm, we found that it sometimes performed better than the first one, while sometimes worse, the reason for its unsteady performance is still not clear. Finally we chose to use the first algorithm in the final system.

Therefore sentence similarities with the cluster and question are calculated in this stage. And centroid [Radev, et al. 2000] is used as a representative of the document cluster.

2.2 Conditional Maximum Entropy(CME) model

After feature extraction, we trained a model to do the supervised learning. Here we selected CME because of its good performance on other NLP tasks. For details of maximum entropy principle, please refer to [Berger, et al. 1996].

The extracted features were used as input to CME model, in the form of feature vectors with discrete values. CME can output a score which represents the probability that a sentence should be included in the summary, i.e., p(y=1|x), where x stands for input feature vector, and y stands for whether a sentence belongs to the summary or not. y=1 means that a sentence belongs to the summary; y=0, otherwise. We applied CME on each sentence and got a score to use as a measure for sentence extraction in the next step.

2.3 Summary generation

This is the final stage of the system. In this stage, the module of redundancy reduction was implemented to generate the final summary that is no longer than 250 words.

According to the scores that we computed in the previous stages, all the sentences in the document cluster were ranked in descending order. Since two sentences may have redundant information, it is not appropriate to extract both sentences into the summary. The method we used is as follows. When sentence X is extracted, the weight of the remaining sentence Y is recalculated as:

$$Y^{'} = \sqrt{Y^2 - (Y * sim(X,Y))^2}$$

In each iteration, we extract the sentence with the highest score, and then adjust scores of the remaining sentences using the above formula. Scores of sentences that are very similar with the extracted sentence are adjusted downwards in this way. This process is repeated until we reach the length restriction of the summary.

2.4 Issue on granularity

Another issue that should be mentioned is about the granularity specified in the user profile, which is either ‘general’ or ‘specific’. Summaries with different granularities are required. In order to measure the granularity, we designed two methods.

One is to compute the value of information content of each word [Seco, et al., 2004] to measure how general a word is, which is based on WordNet, and then the values of all words in a sentence are averaged to get the degree of generalization of this sentence.

Another way is to use the number of name entities in a sentence as a measurement. Generally, the more NEs a sentence contains, the more specific it is.

We tested both ways but according to manual checking, we did not find their significance in distinguishing the sentences, so finally we chose not to consider the granularity when generating a summary.

3. Experimental Results

3.1 Experiment data

The training data we used to train the CME model are from the task 4 of DUC 2003. There are 30 documents clusters, totally 675 documents. Each cluster is provided with topic (question) and relevant sentences extracted from the documents. The set of relevant sentences are used as summary for training.

Evaluation data of DUC 2005 contains 50 topics. For 30 topics, 4 human summaries are provided and for the other 20, 9 or 10 human summaries are provided for evaluation.

3.2 Evaluation metrics

Three kinds of evaluation methods are used in this year. Besides the automated ROUGE evaluation and manual evaluation on summary quality and fluency, a semi-automatic evaluation method called Pyramid evaluation is developed. Its basic idea is to identify summarization content units (SCUs) that are used for comparison of information in summaries. SCUs that appear in more manual summaries will get higher weights, so a pyramid will be formed after SCU annotation of manual summaries. The SCUs in peer summary are then compared against an existing pyramid to evaluate how much information is agreed between peer summary and manual summary.

3.3 Experimental results

There are 31 groups participated in DUC2005, and each group submitted one run. DUC also created a baseline system that took the first 250 words of the most recent document for each topic.

Table 1 shows the ROUGE evaluation results of our system, with our rank, score, and median score of all systems under different ROUGE metrics.

Metrics	Rank	Score	Median
Rouge-1	6	0.3609	0.3469
Rouge-2	14	0.0609	0.0597
Rouge-3	15	0.0162	0.0158
Rouge-4	15	0.0070	0.0068
Rouge-SU4	11	0.1188	0.1167
Rouge-L	6	0.3320	0.3141
Rouge-W	6	0.0953	0.0911

Table 1. Rouge evaluation results

Our system performed well on Rouge-1, Rouge-L and Rouge-W, but dropped on other Rouge metrics such as Rouge-N where N is bigger than 1. The reason may be that our algorithm did not take N-gram into consideration.

Actually the ROUGE scores for most submitted systems are very close, and therefore seem not very effective in differentiating the system performances.

Table 2 shows manual evaluation results of our system on responsiveness. And in additional pyramid evaluation, we ranked 7th when using modified pyramid scores.

Metrics	Rank	Score	Median
Responsiveness	11	18.65	17.16

Table 2. Evaluation results on responsiveness

4. Conclusion and Future Developments

In this paper, we described the architecture and evaluation results of our system in DUC 2005. When designing the system, we conducted lots of experiments trying to find some new ways in text summarization, especially in deeper understanding of the texts, such as syntactic and semantic analysis of the texts. Although they did not show the expected effect at present, we still believe they are worth further exploring.

Another possible and interesting research direction is to combine question answering with question-focused text summarization because of some common characteristics of these two problems. Any improvement on one is probably useful to the other.

5. Acknowledgments

This research was partly supported by NSF of China under contract of 60435020. We are very thankful to Zhushuo Zhang, Xian Qian, Zhongqin Wu, Fei Liu, Yaqian Zhou, Ningyu Chen, Qi Zhang and Yingyi Qi for their contributions in our work.

References

A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *In Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics,* Pittsburgh. June 2001.

A. Budanitsky. Lexical semantic relatedness and its application in natural language processing. *Technical Report CSRG390*. University of Toronto. 1999.

Nuno Seco, Tony Veale, Jer Hayes. An intrinsic information content metric for semantic similarity in WordNet. *In Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence.* Valencia, Spain, 2004.

D. Radev, H. Jing, M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *In ANLP/NAACL workshop on Summarization*, Seattle, WA, 2000.

Adam L. Berger, et al. A maximum entropy approach to natural language processing. *Computational Lingustics*, 22(1), 39-71. 1996.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five Papers on WordNet. *Technical report,* Princeton University, 1993.