Peking University at DUC 2006

Sujian Li, You Ouyang, Bin Sun,
Inst. of Computational Linguistics, Peking University
{lisujian, oyangu, bswen}@pku.edu.cn

Zhili Guo
IBM China Research Lab.
guozhili@cn.ibm.com

Abstract

This paper has described a summary extraction system implemented by Peking University at DUC 2006. The system follows the assumption that there must exist some sentences which can summarize the topic briefly. Then our system depends on various features to judge whether a sentence is appropriately included in the summary. Our results are satisfying in the evaluations including linguistic quality, responsiveness, Rouge and pyramid evaluations.

1. Introduction

The system task for DUC 2006 models real-world complex question answering [1]. Focused on information users are most interested in, a brief summary is generated from a set of relevant documents. That is, given a topic and a set of 25 relevant documents, a fluent, well-organized 250-word summary will be produced. Thus, QA and Multi-document summary (**MDS** for short) tasks are combined here. QA task requires that the summary can answer the questions included in the topic. MDS task requires that the summary must summarize the contents of all documents as comprehensively and fluently as possible.

It is our first try to participate in the DUC evaluation. To build our summarization system, we follow the assumption that there must exist some sentences in the 25 files, which can summarize the topic briefly. Due to immaturity of text generation techniques and the assumption above, our system is designed with a summary extraction framework. Sentences in the documents are picked out to compose a summary. Various features in the sentence are used to judge whether the sentence should be appropriately included in the summary.

The rest of the paper is organized as follows. Section 2 describes our system design. Section 3 emphasizes on the feature calculation and sentence scoring which decide the importance of each sentence. Section 4 presents the evaluation results of our system. Section 5 shows the future work and concludes the paper.

2. System overview

Our summarization system is designed with a summary extraction framework. Important sentences are extracted and re-organized to form a summary. Thus, the whole system is divided into three modules: text preprocessing, sentence extraction, post-processing. The flowchart is as figure 1.

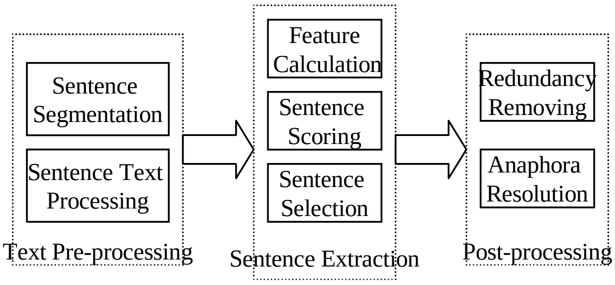


Figure 1: System Architecture

In the module of text preprocessing, the main task is sentence segmentation for the query and 25 relevant files. At the same time, some mature text processing tools, such as word stemming, POS tagging and entity recognition, are used for processing each sentence.

The module of sentence extraction is the kernel of the system. Important sentences are extracted from multiple documents to serve in the summary. In order to quantify the possibility of each sentence being in the summary, appropriate features are extracted and assigned values to score each sentence. Then sentences with higher scores are extracted. Thus, which features to select and how to conduct feature calculation is the key of the module, which will be introduced in section 3.

In the module of post-processing, in order to avoid information redundancy, we have conducted some remedy. Some redundant phrases, for example some parenthesized phrases function as explanatory roles which seldom appear in the summary, are removed for efficiency. Now, the extracted sentences are listed according to their information scores. In order to assure that the text of a summary is readable, these sentences should be organized with some reasonable order. Here, we adopt the simple time order to reorganize all the extracted sentences. In addition, anaphora resolution, especially for time and persons, are conducted.

3. Feature Calculation and Sentence Scoring

To evaluate whether a sentence is appropriately included in the summary, two factors are considered. One is the association between a sentence and the query, and the other consideration is the information density of a sentence compared to other sentences in the documents. More responsive a sentence is to the query, more possible the sentence is to be included in the summary. Due to the length limit of 250 words, a sentence which will serve in the summary should contain as much as information with the same length. More information density the sentence contains, more important it is than other sentences. For each sentence, the measure of information density depends on the quantification values of features.

All features are divided into three kinds: word based features, chunk based features, and global features. Each feature is assigned a normalized value. Then bonus and penalty are conducted on the features correspondingly. It is worthy noted that feature selection is prudent. The features involved in semantic meaning are usually thought as good selection. However, semantic parsing technique is still immature. Features of lexical and syntactic levels are still our favorites. The features are listed as follows.

3.1 Word based features

For word-based features, we mainly consider the word overlap, cosine similarity, centroid and semantic similarity between each sentence and the corresponding query.

Word overlap feature

The words in a sentence can be of three cases: Stop words, Non-stopwords[1] also appearing in the query, Non-stopwords not in the query. Obviously, stop words have a negative effect on the information density of a sentence if its length is fixed, while non-stopwords have a positive effect. Here we make an assumption that the non-stopwords overlap of the sentence and the query somewhat reflects the degree of their association. Then, a negative value is assigned according to the number of stop words in the sentence, while a positive value is assigned according to the non-stopwords overlap. In addition, considering the sentence expected in the summary should carry information that is needed but not contained in

[1] A non-stopword means a word which is not a stop word.

the query, we assumed that non-stopwords in the sentence but not in the query usually play such a role. Thus, a positive value is also assigned according to the non-stopwords not in query. It is difficult for the system to have a uniform measure among sentences of different lengths. That is, there often exists a long sentence bias for the non-normalized values. The longer the sentence is, the higher the values usually are. Thus, we conduct normalization for the three values mentioned above.

Given N_1, N_2, N_3 as the word number of the three cases respectively and N_{All} as the total word number of the sentence, the value of word overlap feature is calculated as:

$$V_{WO} = (\lambda_1 N_1 + \lambda_2 N_2 + \lambda_3 N_3) / N_{All}$$

where λ_i $(1 \le \mathrm{i} \le 3)$ is the weight of N_i. Since $(N_1 + N_2 + N_3) / N_{All} = 1$, V_{wo} can be seen as a linear function of $N_1 / N_{All}, N_2 / N_{All}$. After the constant removed, the formula is as follows.

$$V_{WO} = (\lambda_1' N_1 + \lambda_2' N_2) / N_{All}$$

Then the computation of V_{wo} only considers the number of stop words in the sentence and the non-stopword overlap between the sentence and the query.

Cosine feature

The cosine similarity of the sentence and the query is also used as a feature. It computes the cosine value between the sentence vector and the query vector. Each dimension of the vector is the TFIDF value of the words.

Given Q, S as the TFIDF vector of the query and a sentence, the cosine feature is calculated as:

$$V_{Cos} = (Q, S) / |Q| * |S|.$$

Semantic feature

The WordNet framework [2] has provided a function for computing the semantic distance of two words. With this function, we can compute the semantic distance between a sentence and the query through accumulating the distance values of all word pairs. A normalization factor – the word number of the sentence, is also chosen for overcoming the length bias. For a specific query and a sentence, the semantic feature is calculated as follows.

$$V_S = (\sum_{\mathrm{i}} \sum_{\mathrm{j}} \mathrm{Sim}(\mathrm{s_i}, \mathrm{q_j})) / N_{All}$$

where $\mathrm{s}_i, \mathrm{q}_j$ are words from the sentence and query respectively.

Centroid feature

Centroid feature focuses on describing the importance of a sentence compared to other sentences in the documents. The documents' centroid is selected by computing words' Count*IDF scores. Then the hypothesis is that sentences that contain words from the centroid are more indicative of the topic of the documents, which is introduced detailedly in [3]. Our system directly makes use of the centroid feature values produced by the MEAD system [4].

3.2 Chunk based features

Here, features involved in named entities are mainly considered as chunk-based features.

Entity feature

Named entities are the important part of the sentences. Generally, the more named entities a sentence has, more important the sentence is. Also, a normalization factor of sentence length is assigned. Four types of named entities are mainly considered: Person, Localization, Date, and Organization. GATE [5] is used to extract these types of entities.

Given $NE_i \ , i = 1, 2, 3, 4$ as the number of the four types of named entities in the sentence, the value of the entity feature is calculated as:

$$V_E = (\sum_i \lambda_i \ NE_i \ ) / N_{All}$$

Where λ_i is the weight for each entity type, and depends on the question types in the query.

Entity overlap feature

Entity overlap feature is designed to count the overlap of named entities. Different from word overlap feature, we use the total number of named entities in the query as a normalization factor, but not the sentence length.

Given NE_{Same} as the number of overlapped named entities between the sentence and the query, NE_{All} as the total number of named entities in the sentence, the entity overlap feature is calculated as follows.

$$V_{EO} = NE_{Same} / NE_{All}$$

3.3 Global features

Here, global features mainly consider the length and some text patterns of the sentence.

Length feature

A simple fact is that short sentences cannot carry enough information corresponding for the query. Thus, too short sentences are not appropriate candidates of summary sentences and will not be considered. And due to the 250-word limit, too long sentences are not appropriate too. We use 20 as the lower bound of sentence length and 60 as the upper bound. A relatively strict threshold is chosen because much more worse sentences are removed though some better sentences are filtered. Even with this striction we can still extract enough good candidate sentences for composing the summary.

Pattern penalty feature

There are some patterns which are unsuitable for being in the summary. The sentences which have these patterns will be discounted for being summary sentence. We list some example patterns as follows.

(1) Somebody’s saying, e.g.: some body say/said/says, ”…”.

(2) Internet address appearing in the text.

(3) Capitalized text, e.g.: MAKAH HUNT FOR WHALING HERITAGE.

In case that there may be some special sentences which contain these patterns but correspond closely to the query, here we use a penalty for each possible pattern, instead of banning them. The pattern penalty feature is calculated as:

$$V_{BP} = \begin{cases} -1 & \text{if the sentence matches some penalty pattern,} \\ 0 & \text{otherwise} \end{cases}$$

3.4 Sentence scoring

After all the features are calculated, we use those feature values to score for the sentence. A linear function is used to combine all the feature values, except the length feature which decides whether the sentence is a candidate. The sentence score is calculated as:

$$S = \sum_{\mathrm{i}} \lambda_i V_i$$

Where V_i represents each feature value, λ_i is the experience weight assigned by human.

With the sentence scores, we can sort sentences from high to low, here MMR (Maximal Marginal Relevance) technique is adopted to reduce redundancy and improve summary universality.

4 Evaluations

4.1 Test Data and Metrics

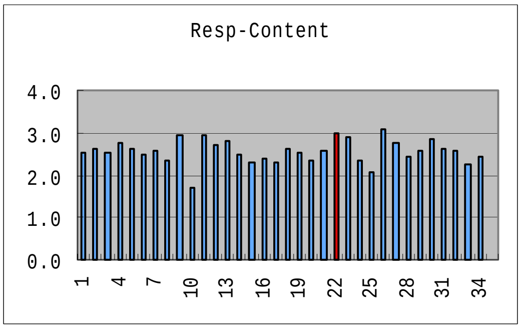
Same with DUC 2005, DUC 2006 provides fifty document sets for test evaluation. However, each document set includes a fixed number – 25 documents and its query. Each query contains a query title and a query narrative. A query title is usually a phrase which describes briefly the topic. A query narrative is usually composed of several factoid or definition questions, which need answers given in the summary. NIST assessors created 4 reference summary for each topic. There are 34 participants in DUC2006, each participant submit one summary. All submitted systems are either manually or automatically evaluated, including linguistic quality, responsiveness, ROUGE-2, ROUGE-SU4 [6], and Pyramid [7].

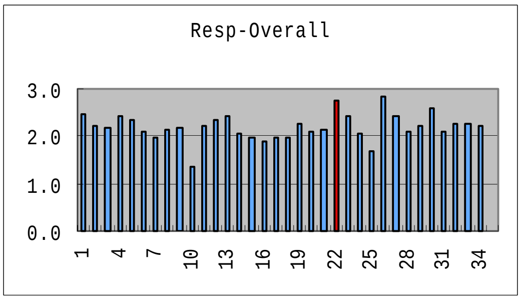
4.2 Results

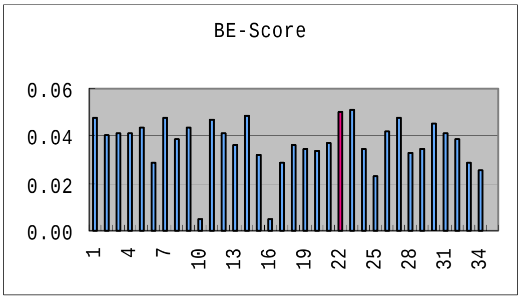
Among the 34 systems, our system ranks 2nd in the responsiveness evaluations of both content and overall, 4th in the linguistic quality evalution, 5th and 6th respectively in the automatic Rouge-2 and Rouge-SU4 evaluations, 2nd in the BE evaluation, and 2nd in an extra pyramid evaluation. Table 1 and figure 2 show the details of the evaluation results.

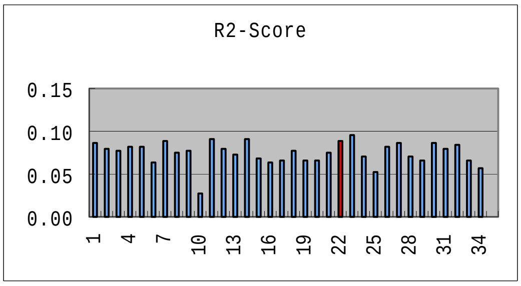
	Rank	Score	Best
Resp Content	2	3.0	3.08
Resp overall	2	2.76	2.84
BE	2	0.05049	0.05107
Rouge-2	5	0.08792	0.09558
Rouge-SU4	6	0.14486	0.15529
Pyramid	2	0.2514	0.257114
Ling quality	4	3.704	4.08

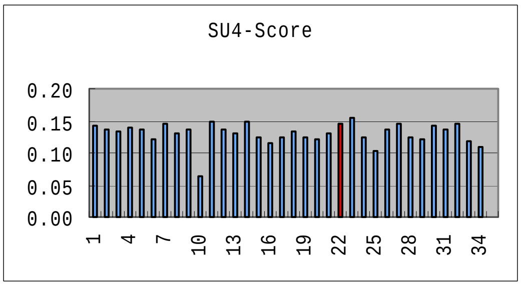
Table 1. DUC evaluation results











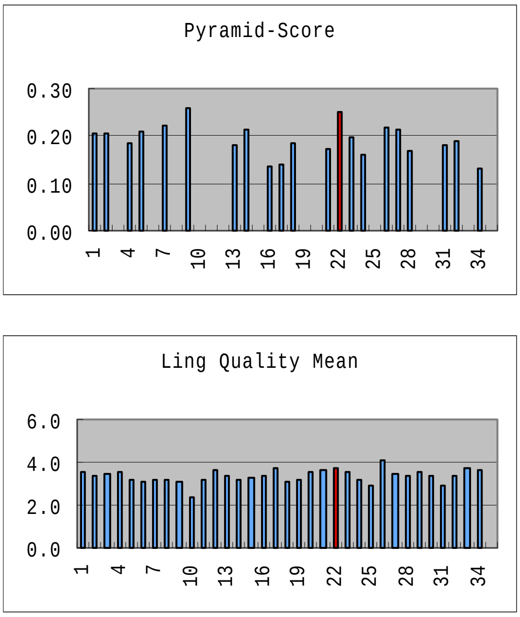


Figure 2: DUC 2006 evaluation results (the red pillars represent our system)

4.3 Analysis

Our system performed comparably in the evaluations, showing that our system is steady using different evaluation metrics. The main reason is that appropriate lexical and syntactic features are adopted and the weight parameters are assigned suitably. The performance in automatic Rouge evaluations is not so good, and the reason may be that no N-gram based features are introduced. We have introduced some chunk-based features, however, these features are only about named entities which only occupy a small percentage in the text. Other chunk based features will be considered to improve the summarization performance.

5 Conclusions and Future Work

Our system adopts the traditional methods of automatic summarization. That is, some sentences are extracted from original text and reorganized into a summary with consideration of the query. The process of sentence extraction depends on various features. Thus, feature selection and calculation is the key of our system to getting competitive results in this evaluation. It can be seen that in the environment of current techniques, simple and mature techniques still play an important role in a summarization system. Also, there are a lot of rooms for improvement. In future work, we will strengthen anaphora resolution. Complex syntactic features involving subjects, objects and so on will be considered.

Acknowledgements

This work is supported by IBM University Joint research program and 973 National Basic Research Program of China (2004CB318102).

References

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

[2] Christiane F.. 1998. WordNet: an Electronic Lexical Database. MIT Press.

[3] Radev D., Jing H., Stys M., and Tam D. Centroid-based summarization of multiple documents. Information Processing and Management, 40:919-938, December 2004.

[4] Radev D., Blitzer J., Winkel A, Topper M., Celebi A., and Lam W.. MEAD Documentation. http://www.summarization.com/mead/.

[5] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V., 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In ACL 2002.

[6] Lin.C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization, Barcelona. ACL.

[7]Pyramid,http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html