Language Model Passage Retrieval for Question-Oriented Multi Document Summarization

Jia-Ching Ying, Show-Jane Yen, Yue-Shi Lee
Dept. of Computer Science and Information Engineering
Ming Chuan University
Taoyuan, Taiwan
jashying@gmail.com
{sjyen,lees}@mcu.edu.tw

Yu-Chieh Wu
Dept. of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
bcbb@db.csie.ncu.edu.tw

Jie-Chi Yang
Graduate Institute of Network Learning Technology
National Central University
Taoyuan, Taiwan
yang@cl.ncu.edu.tw

Abstract

The goal of question-oriented text summarization aims at producing the informative short description according to the given queries. This is somewhat similar to the target of question answering which retrieves exact answers from large raw text collections. In this paper, we present a resource, and training data-free summarization model for DUC multi-document summarization task. Similar as last year, our method simplified the two-pass retrieval as a passage retrieval task. At first the top-down clustering algorithm is used to merge similar passages into a set of groups. Then the passage retriever extracts relevant groups in response to the given query. Finally a maximizing scorer is used to re-form the sentences into the final summary. This the second time to participate in DUC. Although the result of our system is not comparable with most top-performed methods, the light-weight and rule free techniques still encourage us to further improve via integrating rich sources.

1 Introduction

In recent years, there has been an accumulation of vast amounts electronic text articles and web pages. To effective and efficient access important information, there have been several on-going research domains of natural language processing for this task such as information retrieval (IR), information extraction (IE), and automatic text summarization (ATS).

This year's document understanding conference (DUC-2007) task is the same as past two year (Dang, 2006, 2005). The target is to generate 250 words summaries from multi-documents according to the given subject or question, i.e. question-focused text summarization. This task is quite different from traditional summarization tasks that only focus on extracting important sentences without regarding the main relevance to users. The question-focused text summarization is very similar to the traditional question answering (Q/A) task (Voorhees, 2001) that pinpoints answers from huge document collections. But the difference relies on the granularity of questions and retrieved answers. Traditional Q/A put emphasis on asking the factoid questions, however, answers should be short and exact to answer the question. In contrast, in the question-focused summarization task, the question describes an event, a comparison, or changes whereas the returned summary is like a story to response the requirement. For example, the question of topic 614 in DUC-2006 is "Describe developments in the movement for the independence of Quebec from Canada.".

In this paper, we describe the overview of our resource-free and rule-free automatic text summarizer at DUC this year. Unlike previous studies (D'Avanzo, and Magnini, 2005; Ye et al., 2005; Li et al., 2005), additional training and external resources such as named entity taggers are excluded in our summarization system. Similar as last year, our method simplified the two-pass retrieval as a

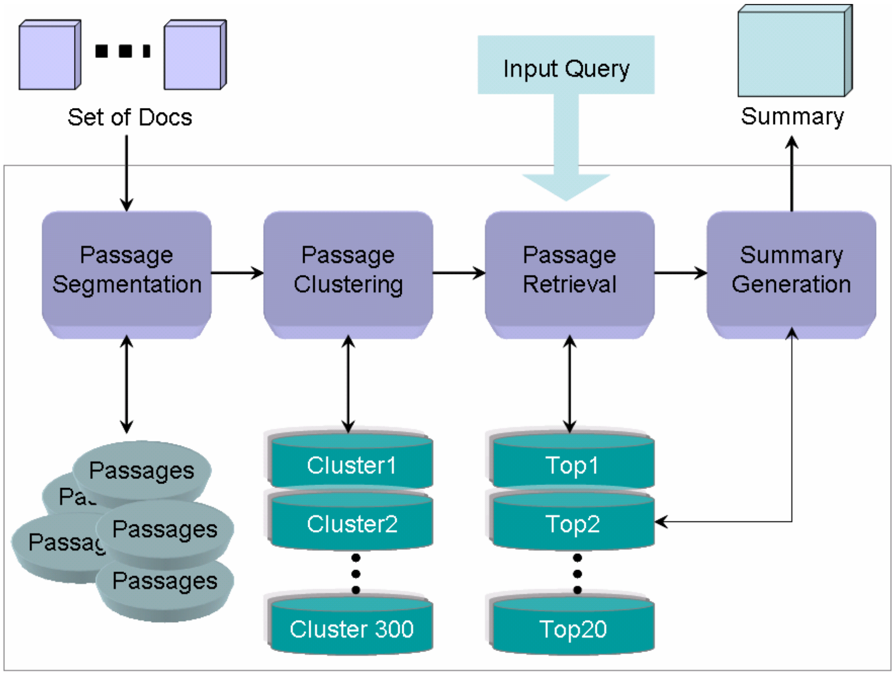

Figure1: System architecture

passage retrieval task. At first the top-down clustering algorithm is used to merge similar passages into a set of groups. Then the passage retriever extracts relevant groups in response to the given query. Finally a maximizing scorer is used to reform the sentences into the final summary.

This paper is organized as following, section 2 describes overview of our system, and section 3 describes the language model-based scoring function. In section 4, we present the evaluations and experimental results. At section 5, we draw the future direction and conclusion.

2 System Description

The target of multi-document text summarization is to extract or refine important sentences from different documents that belong to the same topic. As described above the goal is quite similar to the Q/A task. However, we tries to combine some advanced techniques in Q/A research domain, like the powerful language model-based passage retrieval algorithm (Zhai and Lafferty, 2001). Figure 1 shows the overall architecture of our Q/A model.

There are four main components within our model, namely passage segmentation, passage clustering, passage retrieval, and summary generation. For the given question, we first segment each sentences and passages in the given document set respective to the question. Some of the passages might describe the same topic, therefore we perform a clustering algorithm to group similar passages into the same paragraph. Finally the two-pass re-ranking models are used to retrieved the useful and informative passages at first pass, for each retrieved passage, the second pass sentence retrieval component selects the most important sentence for a passage and add it to be the summary. To make the summary more readable, the added sentences are re-ordered according to their time-stamps.

In the following subsections, we will introduce the first component in section 2.1. For the two retrievers and the clustering method are introduced in section 2.2, 2.3, and 2.4.

2.1 Passage Segmentation

In this step, the sentences are first segmented, the words are not stemmed and tokenized. We do not perform the word-stemmer to represent the root of words, instead, this will be done in the clustering and ranking steps. The sentences segmentation is carried out with a tool[1]. This tool can successfully identify boundaries between sentences without tokenizing words.

The documents in the DUC-provided set had been annotated with passage boundaries. Without employing additional passage segmentation tool, we directly use the tag to split paragraphs.

[1] http://l2r.cs.uiuc.edu/~cogcomp/tools.php

2.2 Passage/Sentence Clustering

In multi-document summarization scenario, multiple passages or sentences may describe the same concept meanings. To reduce the redundancy, a conventional clustering technique is applied to group similar passages into the same group. Varying from previous studies, we use the top-down bisecting *K*-means algorithm (Zhao and Karypis, 2002) for clustering. The bisecting *K*-means algorithm is a top-down step-by-step clustering method that incrementally performs *K*(=2)-means to split the largest group into two sub-clusters. In the Zhou's study, they had showed that the bisecting *K*-means outperformed the traditional *K*-means in document clustering task. Nevertheless the passage clustering is very similar to the document clustering. Hence, we select the bisecting *K*-means algorithm to avoid the risk of randomly initialization of the traditional *K*-means.

We slightly modify the bisecting *K*-means and set the number of clusters as 300. There are several criteria functions to evaluate the quality of a clustering result. We use the internal criteria functions to measure the similarity inside the cluster. This function was demonstrated as a very effect method to determine the clustering result at each splitting step for bisecting *K*-means. In addition, the settings of the algorithm are almost the same as the literature (Zhao and Karypis, 2002), except for the example representation. In order to capture more accurate meanings in passages, we do not only use the traditional bag of words model (with Porter stemming), but also include the bag of bigrams. Bigrams are meaningful than unigram.

2.3 Passage Retrieval

Passage retrieval components adopts the developed language model to extract important passages (clusters). Last year we developed the two pass ranking models for DUC-2006 and replaced the two pass framework with one-pass sentence ranking. However, in this way, it will cause the sentences too similar to make the reader over-understanding. In section 3, we will discuss the new passage retrieval method in detail. By clustering similar passages, the passage retrieval is designed for retrieve the relevant "clusters" instead of passages.

For the passage retrieval, we simply use the given question as query to our retrieval algorithm to retrieve top 20 clusters. We then kept them for further summary generation..

2.4 Summary Generation

We treat each retrieved cluster as separable concept. Our summary is mainly derived from these concepts. Therefore for each cluster, we extract an important sentence within it to represent the concept. Unlike most approaches, which estimate the similarity between the centroid and each sentence, we seek to find the sentences for other ranks that maximize the difference between previous ranks. With except for top-1 cluster, we choose exact one sentence from a cluster that achieved the best different score than the other summary sentences. This kept the newly added sentences to be fresh and new. For example, for top-1 cluster, we select one sentence, which is closest to the centroid to be the first extracted summary sentence. Hereafter, for top-2 cluster, we choose the sentence that is "most different" from the summary as fresh and insert to the summary. Similar producing ways for the other sentences until the pre-defined summary size achieved.

To make the summary more readable, each sentence is further re-ordered according to its timestamps in the original document. If the summary contains more than 250 words, we remove words from the final part of the summary to enable the size no more than the size limitation.

3 Language Model-based Retrieval Algorithm

Searching answers in a small dataset is more efficient than in the whole corpus. To find out answers in passages is much easier than searching the whole relevant document set. In this section, we will introduce our retrieval model.

3.1 Passage Retrieval

The passage retriever segments each retrieved document into passages and retains the paragraphs that contain at least one of the query terms. We implemented the similar idea of the *n*-gram language model (Chen and Goodman 1998).The goal of language modeling is to predict the probability of natural word sequences; or more simply, to put high probability on word sequences those actually occur (and low probability on word sequences that never occur). The simplest and most successful basis for language modeling is the *n*-gram model. Note that by the chain rule of probability we can write the probability of any sequence as

$$P(w_1 w_2 ... w_T) = \prod_{i=1}^{T} P(w_i \mid w_1 ... w_{i-1})$$

An *n*-gram model approximates this probability by assuming that the only words relevant to predicting $P(w_i \mid w_1 ... w_{i-1})$ are the previous *n*-1 words; that is, it assumes the Markov *n*-gram independence assumption

$$P(w_i \mid w_1 ... w_{i-1}) = P(w_i \mid w_{1-n+1} ... w_{i-1})$$

A straightforward maximum likelihood estimate of n-gram probabilities from a corpus is given by the observed frequency

$$P(w_i \mid w_{1-n+1} ... w_{i-1}) = \frac{\#\left(w_{1-n+1} ... w_i\right)}{\#\left(w_{1-n+1} ... w_{i-1}\right)}$$

where #(.) is the number of occurrences of a specified gram in the training corpus. Because of the heavy tailed nature of language one is likely to encounter novel *n*-grams that were never witnessed during training. Therefore, some mechanism for assigning non-zero probability to novel n-grams is a central and unavoidable issue. One standard approach to smoothing probability estimates to cope with sparse data problems is to use some sort of linear interpolation estimator

$$P_{li}(w_i \mid w_{1-n+1} ... w_{i-1}) = \sum_{i=1}^{n} \lambda_i P(w_i \mid w_{1-n+1} ... w_{i-1})$$

Where $0 < \lambda_i < 1$ and. $\sum_{i=1}^{n} \lambda_i = 1$

This parameter can be set automatically using the Expectation-Maximization (EM) algorithm. However, the estimation of some probability by some linear way is not perfect enough. Especially, the occurrence of some *n*-gram is a kind of response of a human been. Most statistician believe that the logistic regression can solve the problem which the estimation is a kind of response of a human been. The logistic regression assumes that there is a sigmoid function, called response function, to represent the probability of occurrence of a response as follows:

$$P_{lo}(w_i \mid w_{1-n+1} ... w_{i-1}) = \frac{1}{1 + e^{-v_i}}$$

, where $v_i = \alpha_0 + \sum_{i=1}^{n} \alpha_i P(w_i \mid w_{1-n+1} ... w_{i-1})$

This parameter can be set automatically using the MIRA algorithm for binary problem.

In the following, we take an example to describe the logistic regression modeling to estimate the conditional probability of *bi*-grams in some passage. Table 1 represents the response variable and independent variables.

Table 1: the response variable and independent variables

Response variable		independent variables	
$\#(w_{i-1}w_i)$	$\#(w_{i-1})$	$P(w_i \mid w_{i-1})$	$P(w_i)$
n_1	r_1	x_{11}	x_{21}
n_2	r_2	x_{12}	x_{22}
…	…	…	…
n_j	r_j	x_{1j}	x_{2j}

We can treat $\#(w_{i-1})$ as the frequency we observe and $\#(w_{i-1}w_i)$ as the frequency of the response which *bi*-gram is occurred. So we translate the problem become a binary classification problem and we can adopt MIRA algorithm for binary problem to estimate the parameter α_0, α_1, α_2 as follows:

```
Initialize: set $\alpha = (\alpha_0, \alpha_1, \alpha_2) \neq 0$
Loop For 1,2…,j
    $x_j = \left(1, x_{1j}, x_{2j}\right)$
    Loop For 1,2…,$n_j$
        $\alpha \leftarrow \alpha + (\frac{1}{\|x_j\|^2} \times \frac{1}{1 + e^{-\alpha \cdot x_j}}) x_j$
Output: α
```

After parameter estimation, we can estimate all the *bi*-gram for this passage. Therefore we can calculate the probability of a query $Q=w_1,w_2,...w_q$ occurred in this passage as follows:

$$P(Q) = \prod_{i=1}^{q} P_{lo}(w_i \mid w_{i-1}) = \prod_{i=1}^{q} \frac{1}{1+e^{-v_i}}$$

where $v_i = \alpha_0 + \alpha_1 P(w_i) + \alpha_2 P(w_i \mid w_{i-1})$

4 Evaluation Results

DUC-2007 has evaluated summaries in several ways: human evaluation with pyramid score, responsiveness to the topic and linguistic quality, and automatic ROUGE evaluations. The overall results are shown in Table 2.

Table 2: Overall score of our system in DUC-2007

NCU+MCU	ROUGE-2
Ling Quality Mean	2.47
Avg. Content	1.64
Basic Elements	0.023
Min	0.021
Max	0.025

4.1 Responsiveness

This evaluation gives the responsiveness score between one (lowest) to five (highest) to each automatic summaries. Responsiveness is a measurement that is supposed to contribute toward satisfying the information need expressed in the topic statement. Our summarization system achieved 2.4 and rank 24 (out of 34) on content response score. The overall responsiveness score was 1.9 and rank 31 (out of 34). We do not surprise the low score since we do not pre-define any templates or rules to pre-assume the summaries. All of the sentences are fully scored from the set of documents.

4.2 Linguistic Quality

This measurement estimates the linguistic quality of the auto-generated summaries. NIST employ several human experts who develop the given topic. They created the following judgments for evaluations.

- ♦ Grammatically
- ♦ Non-redundancy
- ♦ Referential clarity
- ♦ Focus
- ♦ Structure and coherence

Each summary is judged for each of the above factor and gave it the score from one (lowest) to five (highest). This is the second year we participate in DUC, and also join the pyramid score. As shown in Table 1, the pyramid score of our method was 0.137 and rank 28 (out of 30).

5 Conclusion and Future Remark

Text summarization is one of the most important issues in information retrieval and natural language processing community. This paper presents the impact of the automatic and rule-free summarization system with minimally human effort. To go state-of-the-art, our method still need to combine more rich resources as most advanced techniques. The main focus of our work coincides with the original target goal of DUC conference, i.e. to automatic summarize multi-document without human intervene. One of the future works is to integrate more and more resources such as full parsers, human-made rules and thesaurus to help to refine the text summaries. In addition, we also find that many question answering technologies can be applied to assist to retrieve important concepts in documents. We start to address the issues of combining question answering models for text summarization.

References

E. Brill. 1995. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging, Computational Linguistics, 21(4):543-565.

E. D'Avanzo, and B. Magnini. 2005. A Keyphrase-based Approach to Summarization: the LAKE System at DUC-2005. In Proceedings of the Document Understanding Conference (DUC).

S. Chen and J. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. TR-10-98, Harvard University.

H. T. Dang. 2005. Overview of DUC 2005. In Proceedings of the Document Understanding Conference (DUC).

H. T. Dang. 2006. Overview of DUC 2006. In Proceedings of the Document Understanding Conference (DUC).

J. Goldstein, M. Kantrowitz, V. Mittal, J. Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In Proceedings of the 22th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 121-128.

E. Hovy, U. Hermjakob, and C. Y. Lin. 2001. The use of external knowledge in factoid QA. In Proceedings of the 10th Text Retrieval Conference (TREC), pp. 644-652.

K. Ishikawa, S. Ando and A. Okumura. 2001. Hybrid Text Summarization Method based on the TF Method and the Lead Method. In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop.

H. Jing. 2000. Sentence Simplification in Automatic Text Summarization. In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP).

G. G. Lee, J. Y. Seo, S. W. Lee, H. M. Jung, B. H. Cho, C. K. Lee, B. K. Kwak, J. W. Cha, D. S. Kim, J. H. An, and H. S. Kim. 2001. SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP. In Proceedings of the 10th Text Retrieval Conference (TREC), pp. 437-446.

W. Li, W. Li, B. Li, Q. Chen, and M. Wu. 2005. The Hong Kong Polytechnic University at DUC2005. In Proceedings of the Document Understanding Conference (DUC).

C. Y. Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In Proceedings of the Association of Computational Linguistics Workshop, pp. 74-81.

K. Ohtake, D. Okamoto, M. Kodama, and S. Masuyama. 2001. Yet another Summarization System with Two Modules Using Empirical Knowledge. In Proceedings of the 2nd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop.

D. Radev. 2000. Text summarization tutorial. In Proceedings of the 23th ACM SIGIR Conference on Research and Development in Information Retrieval.

Y. Seki. 2002. Sentence Extraction by Tf/idf and Position Weighting from Newspaper Articles. In Proceedings of the 3rd National Institute of Informatics Test Collection Information Retrieval (NTCIR) Workshop.

S. Tellex, B. Katz, J. J. Lin, A. Fernandes, and G. Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41-47.

E. Voorhees. 2001. Overview of the TREC 2001 question answering track. In Proceedings of the 10th Text Retrieval Conference (TREC), pp. 42-52.

Y. C. Wu, K. C. Tsai, Y. S. Lee, and J. C. Yang. 2006b. Light-weight multi-document summarization based on two-pass re-ranking. In Proceedings of the Document Understanding Conference (DUC).

S. Ye, L. Qiu, T. Chua, and M. Kan. 2005. NUS at DUC 2005: Understanding Documents via Concept Links. In Proceedings of the Document Understanding Conference (DUC).

C. Zhai, and J. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval, In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 334-342.

Y. Zhao, and G. Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In Proceedings of Information and Knowledge Management, pp. 515-524.