The Automated Estimation of Content-terms for Query-focused Multi-document Summarization

Tingting He[1 2] Wei Shao [1 2] Fang Li [12] Zongkai Yang[2] Liang Ma[1 2]
[1]*Department of Computer Science, Huazhong Normal University, 430079, Wuhan*
[2]*Engineering Research Center of Education Information Technology Ministry of Education, 430079, Wuhan, China*
E-mail: shaowei0910@163.com tthe@mail.ccnu.edu.cn LLIIFF84@yahoo.com.cn

Abstract

Query-focused multi-document summarization aims to produce a summary in response to a user query. We present an approach based on estimation of content-terms to address this task. In the process of estimating content-terms, we make full use of the relevant feature and the information richness feature for assigning importance to each of them. With summary content-terms being identified correctly, the candidate sentences are ranked and best sentences are selected to form the summary. Experiments on DUC 2005 and 2006 are performed and the ROUGH evaluation results show that the proposed approach is encouraging.

1. Introduction

As the number of documents available on the internet increases, the task of query-focused multi-document summarization, who could help the user to quickly locate the desired information to a large extent, has received considerable attention. In generally speaking, the multi-document summarization aims at delivering the majority of information content from a set of documents about an explicit or implicit main topic. By contrast, query-focused Multi-document Summarization has to produce a brief, well-organized, fluent description according to the given query from a set of multiple documents. That is to say the query determines what information is appropriate for inclusion in the summary, and it makes the task potentially more challenging.

To attempt such a complex task, one needs to deal with the following steps: mine the enriched information that query-focused because the summary must be biased to the query, and this process needs significant amount of natural language process on relevant document set; merge information stored in different documents, while keeping the query-focused information as novel as possible and with the less redundancy; at last organize the units to form a fluent summary. As a hot research topic, a lot of advanced technology and research have been appeared.

In this paper, we discuss an approach based on automated estimation of content-terms to address this task. After assigning importance to each term, we score the candidate sentences based on the importance of content-terms they own, at last best sentences are selected to form the summary. In order to mine the most reasonable content-terms and get the best summary sentences, we distill two kind of key features for assigning importance to each of them: the relevance feature and the information richness feature. While the front represents the degree of relevance to the particular information need, and it is obtained with the help of the relevance based language model. The other one represents the importance of a term in the relevant document set, which is measured by the log-likelihood ratio statistic. Experimental results on DUC 2005 and 2006 tasks show that the proposed approach is encouraging.

The remainder of the paper is organized as follows: Section 2 discusses previous work; The proposed approach for estimation of content-terms is described in Section 3; Section 4 introduces the experiment; Section 5 gives the evaluation results; Section 6 presents our conclusion and future work.

978-0-7695-3305-6/08 $25.00 © 2008 IEEE
DOI 10.1109/FSKD.2008.260

2. Related work

As the automatic document summary can help concisely describe the information and facilitate the users to understand the document cluster, especially provide personalized services. It will make a great significance of research and bright prospect of vast application, when put in use in many fields such as acquisition of massive information, the figure analysis of commerce intelligence, the electron's administration and the calculation moving, email threads and dealing with the research results.

In the past few years, a series of international workshops and conferences on automatic text summarization such as NTCIR, DUC, special topic sessions in ACL, COLING have brought advanced technology and produced a great deal of experimental systems.

Among these existing systems, large corpus of them has focused on generating query-independent summary. For example: Radev et al [1] propose a centroid-based method that scores sentences based on cluster centroids, position and TF*IDF features. By investigating five different topic representations, Harabagiu and Lacatusu [2] introduce a novel representation of topics based on topic themes. With the help of affinity graphs, Wan and Yang [3] identify semantic relationships between sentences and used a graph rank algorithm to compute the amount of information a subset of sentences contain. At last they get the summary with the best subset of sentences using a greedy algorithm.

For query-focused multi-document summarization, many competitive methods have been developed recently also, while most of which are sentence-based. They use different kinds of the sentence's features and incorporate the information of the query to select sentences. White et al. [4] and Goldstein et al. [5] create query-dependent summaries with a sentence extraction model. After cutting the documents into their component sentences, they order all the sentences according to features such as Term overlap feature between them and the query, their originality position in the passages and relevant paragraphs, and the core terms they contain. And Hovy et al. [6] select the important sentences based on the scores of basic elements (BE). Farzindar et [7] first performs a thematic analysis of the documents, and then matches these themes with the ones identified in the topic. In contrast, researches focused on scoring individual terms are comparatively less. After computing the probability of each content-term by simply counting its frequency in the document set, Nenkova *et al.*[8] socre each sentence as the average of the probabilities of the terms in it, their system performs the second best result of MSE-2005. in [9] important terms are identified through graph-based analysis where the nodes in the graph represent terms.

Compared with these work, first our approach is focused on mining the key information based on term, not sentence. And in this process, our method is different from the methods mentioned above, while we combine the relevant feature and the information richness feature to assign importance to each term. Both feature enable give a reasonable estimation of content-terms towards the information need, because the relevant feature ensures that the query-focused content-term have advantage in the process of estimation. The information richness feature can benefit the content-term who can indicate the corresponding content that contained in the relevant document set. With these kinds of information, we could naturally identify the content-terms that are suitable to including in the summary.

3. Estimation of content-terms

We use the relevance and the information richness features to estimate each terms. While the relevance feature benefits query terms and other terms which occur frequently with query terms; and the information richness feature benefits the terms who can indicate the content of relevant document set. We define a "term" to be any "non-stop word". The above step is merely for the terms in the relevant sentences, and in this paper we define the relevant sentence as the one who has more than one overlap term with the query after the process of pretreatment. The following subsections discuss in detail how these features can be used to estimating the content-terms reasonably.

3.1. Relevance Feature

The relevance based language mode [10], which is a relevance model when no training data is available in the form of relevance judgments, has been widely used in information retrieval applications. As semantically related terms tend to-occur usually, here we extend the mode to the term level and use it for observing the degree of a term's relevance to the particular information need Q. The relevance of a term w_i towards the information need is calculated as:

$$R(w_i) = P(w_i / Q) = \frac{P(w_i, q_1, \dots q_n)}{P(Q)}$$

$$\approx \frac{P(w_i)}{P(Q)} \prod_{q_i} P(q_i / w_i) \approx P(w_i) \prod_{q_i} P(q_i / w_i) \qquad (1)$$

while w is the term in the relevant document set; Q = $q_1, q_{2,...} q_n$ the information need expressed in the form of query terms; the joint probability $P(w, q_1, ...q_n)$ is computed by the Conditional sampling that assumes the query terms $q_1, q_{2,...} q_n$ to be independent of each other while keeping their dependencies on w intact.

As the meaning of the new concept can be learnt from its usage with other concepts in the same context, the required term dependencies P(q_i /w) can be expressed based on their occurrence in the document set. Here we adopt probabilistic Hyperspace Analogue to Language [11] model to define the conditional probabilities while just considering the co-occurrence. Given a term w, the probability of associating q_i with w in a window of size K, can be expressed in terms of probability of observing q_i at a distance of k < K from w, as:

$$pHAL(q_i / w) = \frac{\sum_{k=0}^{K} P(k) n(w, k, q_i)}{\sum_{k=0}^{K} \sum_{w'} n(w, k, w')} \qquad (2)$$

$\sum_{w} n(w, k, w\ )$ is the number of times some term w' followed the term w at a distance of k. if the corpus is sufficiently large, $\sum_{k=0}^{K} \sum_{w'} n(w, k, w')$ equals n(w)*K, while n(w) means the unigram frequency of w. $P(k) = K - k + 1$ is the prior probability that assumed the priors are proportional to the co-occurrence strength.

After acquiring the relevant value of a term w_i towards the information need, we normalized all the values by the maximum one as its relevance feature.

3. 2 Information Richness Feature

As query-focused multi-document summarization is quite different from QA, the summary should can be both a “compressed version” of the document cluster and satisfy the user’s need. So the standard summary should also contain the signature terms, who are generally indicative of the document cluster’s content.

In order to identify the capability of indicating the document set, we use the information richness feature to measure each term. And this is executed by computed the log-likelihood rate statistic λ suggested by Dunning [12], which is widely used in summarization work [13][14]. The statistic λ is equivalent to a mutual information statistic and is based on a 2-by-2 contingency table of counts for each term. With a background corpus, λ (w) is defined as the ratio between the probability of observing content-term w in the input and the background corpus .

For each content-term, we compute the $-2\log\lambda$ value marked as $\lambda(w_i)$. And then the information richness feature is defined as:

$$IR(w_i) = \frac{\lambda(w_i)}{\underset{j \in \{1,2...n\}}{Max} \lambda(w_j)} \qquad (3)$$

3.3 scoring terms

With the relevance and location information for each term, the next step is how to incorporate them together to create an overall score. Here we use a linear combination of the above two features to score each of them. While α, β are the experience weight assigned by human.

$$W(w_i) = \alpha * R(w_i) + \beta * IR(w_i) \qquad (4)$$

4. Experiment

4.1 Date

In order to evaluate the performance of our approach, we used the data from the DUC 2005 and 2006 respectively for our experiments. Both tasks are to produce a 250-term summary from relevant documents in response to the need of information that defined in each topic. As there are 50 topics each year, for each given topic, we can use the remaining 49 topics as a background corpus in the process of estimating content-terms for information richness.

After assigning importance to individual terms in the input, we score each relevant sentence based on the weight of the content-terms it has.

$$Weight_{R(s)} = \sum_{w_i \in s} W(w_i) \qquad (5)$$

4.2 Evaluation Metric

We used the ROUGE for evaluation, which measures summary quality by counting overlapping units such as the n-gram, term sequences and term pairs between the candidate summary and the reference summary.

ROUGE-N is an n-gram recall measure computed as follows:

$$Rough-N=\frac{\sum_{S\in\{\mathrm{Re}fSum\}}\sum_{n-gram\in S}Count_{match}(n-gram)}{\sum_{S\in\{\mathrm{Re}fSum\}}\sum_{n-gram\in S}Count(n-gram)} \quad (6)$$

Where n stands for the length of the n-gram, and $Count_{match}(n-gram)$ is the maximum number of n-grams co-occurring in a candidate summary and the reference summaries. *Count(n-gram)* is the number of n-grams in the reference summaries. Here we use ROUGE-1, ROUGE-2 and ROUGE-SU4 metrics.

5 Result

In order to analyse the feasibility of our approach, we compare with the systems S15 and baseline that described in DUC 2005 publications. They are the best performing system and baseline system respectively. At the same time, the methods that estimation of content-terms only based on relevant feature or information richness feature are also experimented, which are marked as based-R and base-IR. The parameters in the process of estimation of content-terms are empirically set as: K =8, $\alpha=\beta=0.5$. Table 1 gives the performances, while the system of median is the average performance of the systems in DUC2005.

Table 1 System comparison on the test of DUC 2005

System	ROUGE-1	ROUGE-2	R-SU4
S15	0.3752	0.0725	0.1316
media	0.3469	0.0597	0.1167
The proposed approach	0.3701	0.0718	0.1338
Base-R	0.3660	0.0678	0.1320
Base-IR	0.3422	0.0462	0.0916
BaseLine	0.2752	0.0402	0.0872

From table 1 we can see our approach could get competitive result in ROUGE score. This may be benefited by the sentences in the summary that always are the ones who contain content-terms expressed in the query or high related to. Therefore our approach holds a good chance of satisfying the information need, but this is not always the case. In the process of estimating content-terms, our approach naturally proportions both query-focused and information richness feature, so the result is much higher than Base-R or Base-IR respectively. The following list gives parts of the top most highly weighted words for the topic set of D332h in DUC2005.

Parts of the highly weighted words: *investigate, tax, theft, money, charge, federal, prosecution, last, official, crime, trade, government, case, guilty, allege, evasion, tell, state, attorney, report, business, involve, sentence, income, court, company, depart, count, depute, distribute, lawyer, member, story, polite, work, prison, narcotic, plead, preside, dollar, judge, record, office, agent, month, drug, team, indict, bank, major, time, firm, convict, pay, baseball ,file, scheme, ask, investor, home, campus, carter, include, client, come, finance, during, continue, fraud, account, call, world, people, justice, affidavit, trial, sever, bet, right, administrate, sheriff, newspaper, record, hear, exchange, organ, college, law, weapon, manage, move, student, jury, believe, board, place, want, angel, immigrant*

As we havn't done any analyse on the query, all the information of the content-words are based on surface character. The result would be worse when the task is for special summary and there are same modificatory information in the topic. In this condition, the summary may still contain same content-words that matched with model summary, but the capacity of response is poor.

In figure 1, it illustrates the ROUGE-2 scores for all the topics that in DUC2006. From the result, we can see that the performance of stable need to be improved as there are many parameters assigned by human. It is also one of our future work and we will try to train the parameters automatically.

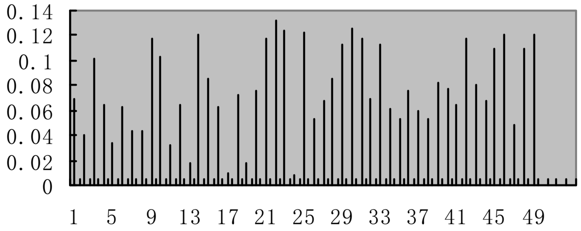

Figure 1: ROUGE-2 scores for all topic set

6. Conclusion and future work

In this paper, we have tried to use term as a basic unit, and assigned the importance to each terms based on the relevant feature and the information richness feature.

After content-terms were estimated, sentences were ranked based on the importance of terms it contains, at last summary was generated out of them. The features that used have been proved be able to identify the relevant terms correctly.

The result of evaluation shows that our method is encouraging. In the future, we will go on our work from the following aspects: As in the process, we gave any deeper analysis on how the content-terms connect with each other in the query. So it can not always satisfy the information that needed. After the content-words that estimated, how to get the best sentences and form the summary is also important. We will try new approach for this step.

Acknowledgement

The paper is supported by National Science Foundation of China, Grant No. 60773167; National Social Science Foundation of China, Grant No. 06BYY029; the Natural Science Foundation of Hubei Province, Grant NO.2006ABC011; 973 National Basic Research Program, Grant NO.2007CB310804; the Program of Introducing Talents of Discipline to Universities, Grant NO. B07042

References

[1] G.Erkan and D.Radev. LexPageRank: prestige in multi-document text summarization. In Proceedings of EMNLP'2004

[2] S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In Proceedings of SIGIR'2005.

[3] X.Wan and J.Yang. Improved affinity graph based multi-document summarization. In Proceedings of HLT-NAACL, Companion Volume: Short Papers, pages 181–184, 2006.

[4] R.W. White, I. Ruthven and J. M. Jose: Finding Relevant Documents using Top RankingSentences: An Evaluation of Two Alternative Schemes, SIGIR, 2002

[5] J.Goldstein, M.Kantrowitz, V.Mittal, J.Carbonell: Summarizing text documents: Sentenceselection and evaluation metrics. ACM SIGIR, 1999

[6] E. Hovy, C.-Y. Lin and L. Zhou. 2005. A BE-based multi-document summarizer with query interpretation. In Proceedings of DUC 2005.

[7] A.Farzindar, F.Rozon and G.Lapalme. 2005. CATS a topic-oriented multi-document summarization system at DUC 2005. In Proceedings of the 2005 Document Understanding Workshop (DUC2005).

[8] A.Nenkova, L.Vanderwende, and K. McKeown. A compositional context sensitive multidocument summarizer. In Proc. of SIGIR, 2006.

[9] I. Mani and E. Bloedorn. Multidocument summarization by graph search and matching. In Proc. of AAAI, 1997.

[10] Victor Lavrenko and W. Bruce Croft. Relevance-based language models. In International ACM SIGIR conference on Research and development in Information Retrieval, pages 120{127, 2001.

[11] Lund K and Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. In Behavior Research Methods, Instrumentation, and Computers, pages 203{208, 1996.

[12]Ted Dunning, Accurate Methods for Statistics of Surprise and Coincidence, Computational Linguistics, 19:61-74, 1993.

[13]J. Conroy, J. Schlesinger, and D. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In Proceedings of the COLING/ACL'06 (Poster Session).

[14]F. Lacatusu, A. Hickl, K. Roberts, Y. Shi, J. Bensley, B. Rink, P. Wang, and L. Taylor. 2006. Lcc's gistexter at duc 2006: Multi-strategy multi-document summarization. In Proceedings of DUC'06.