Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

중심 벡터에 기반한 신문 기사 요약

Summarization of News Articles Based on Centroid Vector

김권양

경일대학교 컴퓨터공학부

E-mail: kykim@kiu.ac.kr

요 약

본 논문은 "X라는 인물은 누구인가?"와 같은 질의어가 주어질 때, X라는 인물에 대한 나이, 직업, 학력 또는 특정 사건에서 X라는 인물의 역할에 대한 정보를 기술하는 문장을 인식하고 추출함으로써 해당 인물에 대한 신문 기사 내용을 요약하는 방법을 제시한다. 질의어 용어에 대해 가능한 많은 관련 문장을 추출하기 위하여 중심 벡터에 기반한 통계적 방법을 적용하였으며, 정확도와 재현율 성능을 개선하기 위해 위키피디어 같은 외부 지식을 사용한 중심 단어의 개선된 가중치 측도를 적용하였다. 실험 대상인 전자신문 말뭉치 상에서 출현 빈도수가 큰 20 인의 IT 인물에 대해 제안한 방법이 개선된 성능을 보임을 알 수 있었다.

Key Words : Summarization, Centroid vector, Centroid word, Definition sentence

1. Introduction

With the rapid growth of online information, it is more important to obtain information effectively and efficiently. To find out the exact information he needs, the user often has to read most of too many web pages and give up his search while examining a few documents.

Factoid questions accept simple, factual answers and answering factoid questions need strong predictions about the type of expected answer such as date, person name, place, organization name, and etc. Recent interest in question answering has focused on answering definition questions that need a different approach because a definition can be a sentence for which global characteristics about the person or organization hold over multiple documents. Thus, answers to definition questions are usually longer, and more complex.

Definition questions might be viewed as simultaneously asking a series of factoid questions about same named-entity. Definition sentences contain the most descriptive information about the query term from multiple relevant documents as opposed to whole documents or web pages which general web search engine retrieves.

For example, the definition sentences for the definition question "Who is the person X?" capture descriptive information which corresponds variously to X's age, origin, education background, occupation, or some role a person played in an event as follow:

- X was born January 9, 1942.
- X is the current chairman of Samsung Group.
- X has an Economics degree from Waseda University in Tokyo.
- X received an MBA from George Washington University in the United States.
- X became a member of the International Olympic Committee in 1996.
- X is the third son of Samsung Group founder Lee Byung-chul.
- X is married to Hong Ra Hee who is also the Executive Director of the Hoam Foundation.
- In September 2006, X received the James A. Van Fleet Award from The Korea Society.

Our research is concerned with the biographical multi-document summarizer that summarizes information about IT person described in the news articles by identifying definition sentences relevant to a given person. To retrieve definition sentences

which contain descriptions of the salient attributes and activities of person in the corpus, along with lists of their associates, we employ a centroid based statistical approach.

To improve the recall performance for the recognizing the definition sentences, the weight measure of centroid words is supplemented by using external knowledge resource such as Wikipedia and redundant candidate sentences are removed from candidate definitions.

2. Summarization of news articles

For a given query term(person's name), our system retrieves the relevant news articles returned by a general newspaper search engine. The sentences which include the query term from the returned collection of relevant articles are definition candidates. These definition candidates are used for constructing a centroid vector which consists of centroid words. All of sentences from news articles are scored by measuring similarity with a centroid vector to select the definition sentences for a given query term.

2.1. Acquiring relevant news articles

For a given query term, our system retrieves a set of relevant news articles, which have a query term, returned by a general newspaper search engine using a query expression such as "person name && organization name". These retrieved collection of articles are split into sentences that are basis to get the definition sentences for the query term. Next, our system selects candidate sentences for definition sentences, discarding sentences that do not contain the search term to construct the centroid vector.

In general, the plain text is obtained by removing all HTML tags from the collection of news articles. Then, the stop words are usually removed from the extracted word lists. After that, we transfer each word into its stem by using morphological analyzer.

2.2. Ranking definition candidates

Ranking candidate definitions determines the goodness of a candidate as a definition sentence. The goodness of a definition sentence is determined by measuring how likely a candidate definition is a definition sentence for the query term. We use a statistical approach based on the centroid vector to address the ranking problem.

In order to retrieve as many relevant sentences for the query term as possible, we adopt centroid based statistical approach which has been applied in summarization of multiple documents[1]. A key feature of our ranking system is its use of centroid vector, which consists of words which are central not only to on news article, but to all the relevant news articles for the query term. A definition centroid, called pseudo sentence, is computed by creating a centroid vector which consists of centroid words. Centroid words are highly relevant topical words for the given query term and central to the definition sentences.

We hypothesize that candidate sentences contain the centroid words in the centroid vector are more informative or indicative of the definition sentences similar to [2]. Centroid words are selected from the candidate sentences which are extracted from search engine by measuring their co-occurrence with the query term.

The weight of centrality for the surrounding words of the query term is calculated by the following formula:

$$weight(W_s) = \frac{f(T_q, W_s)}{f(T_q) + f(W_s) - f(T_q, W_s)} \log_2^{\frac{N}{n}}$$

where W_s is the surrounding word and T_q is the given query term. $f(W_s)$ denotes the number of sentences containing the word W_s and $f(W_s, T_q)$ is the number of sentences where word W_s co-occurs with the query term T_q. $\log_2^{N/n}$ is the measure of inverse document frequency(IDF) of word W_s, n is the number of articles where word W_s occurs at least once, and N is the total number of articles in the collection.

This weighting formula makes two assumptions about the centrality of a surrounding word. First, the more frequently a surrounding word co-occurs with the

query term, the more important it is as a centroid word. Second, the more a surrounding word appears through the entire collection of articles, the less important it is since its global importance is low[3,4].

All extracted sentences are stemmed. And then centrality weights for the stemmed words are calculated using the centrality weighting formula. We select the surrounding words which have weight beyond a predefined threshold in the collection of relevant articles as centroid words, meaning only those weight(w_i^j) > threshold are kept in the vector representation for removing the stop words from the centroid representation. For each query term, the system constructs a definition centroid vector which consists of centroid words.

Centroid(query $term_i$)
= (weight(w_i^0), weight(w_i^1),⋯,weight(w_i^n))

where weight(w_i^j) means centrality weight of a centroid word w_i^j for a query $term_i$. After creating centroid words for each query term, the similarity between each candidate sentence and the definition centroid vector is calculated by using the cosine of angle between two vectors. Our system decides which sentences to include in the definition sentences by measuring the similarity between the definition centroid vector and input sentences from the relevant news articles. Candidate sentences that have highly ranked similarity with the definition centroid vector are more likely definition sentences.

In addition to corpus statistics, we make use of online encyclopedia Wikipedia (http://ko.wikipedia.org [5]) as an external source for the query term to supplement the selection of centroid words. The description about the query term that is retrieved from Wikipedia provides a much larger and more task specific resources for the definition sentences.

The TF·IDF(Term Frequency, Inverse Document Frequency) weighting scheme is used to assign higher weights to distinguished terms in the descriptions returned from Wikipedia. After selecting the highly topical words which have TF·IDF score above a predefined threshold, we re-rank the weight of centroid words which overlap with the topical words by multiplying 1.5.

2.3. Selecting non-redundant candidates

We improve the ranking measure by selecting non-redundant sentences from the top ranked list of definition candidates. If two definition candidates are too similar, we remove the one whose score is lower. Thus we use the following SCORE measure:

$$SCORE\ (S_i) = SIM\ (S_i, S_c)(1 - R(S_i))$$

where $R(S_i)$, *d*enotes redundancy, is computed by counting number of overlap words between sentences S_i and a candidate sentence which has redundancy with the highly ranked sentence. $SIM(S_i, S_c)$ is the similarity score between sentence S_i and centroid vector S_c. $R(S_i)$ is 1 when the candidate sentences consist of only same words, $R(S_i)$ is 0 when they have no common words.

Our system selects non-redundant sentences from the top list of candidate sentences ranked by the similarity measure to avoid introducing redundant sentences into the definition sentences. The overlap measure for redundancy is computed as follow:

$$R(S_i) = \arg\max_{SIM\ (S',S_c) > SIM\ (S_i,S_c)} Overlap\ (S_i, S')$$

where

$$Overlap(S_i,S') = 2 \times \frac{\sum_{w \in S_i \cap S'} MIN\ (f(w,S_i), f(w,S'))}{length\ (S_i) + length\ (S')}$$

$length(S_i)$ is the number of words in the sentence S_i. Function $f(w,S_i)$ denotes the number of common word w occurs in the sentence S_i and $f(w,S')$ is the number of common word w occurs in the other candidate sentence S'. If a common word occurs m times in the sentence S_i and n times in the sentence S', we choose the MIN of them as an overlapping count.

This redundancy measure is the harmonic mean of the percentage of each sentence

that overlaps with other sentence S'. All candidate sentences are re-ranked using the modified SCORE formula which assigns the lower score to the redundant candidate sentences.

3. Experiments and evaluation

We only consider definition of person specially IT persons who are working for IT organization. To acquire the relevant collection of documents, we used the Electronic Times Internet[6] advanced search and submitted queries of the type "person name |$| organization name" with the date filter from 1996 to 2007 for about 38,000 IT persons[7].

We supplement the person name with the name of organization for the works in order to solve ambiguity problem that multiple person has same name. These table-like information about IT person such as person name, organization name, birthday, occupation, etc are already constructed by the wrapper program from the original HTML documents through the previous our work. For our experiments, we select the top 20 IT persons who have high document frequency. Table 1 shows the test data in our work.

Table 1. Test data.

person	# of documents	person	# of documents
person #1	2201	person #11	923
person #2	2181	person #12	888
person #3	1702	person #13	822
person #4	1693	person #14	820
person #5	1386	person #15	587
person #6	1129	person #16	585
person #7	1075	person #17	584
person #8	1043	person #18	562
person #9	1001	person #19	557
person #10	994	person #20	545

The baseline system uses only the centroid based approach to rank candidate sentences. F-measure is defined as the harmonic mean for precision and recall. We see improvements obtained by Centroid + Wikipedia, Centroid + Redundancy, Centroid + Wikipedia + Redundancy over the baseline Centroid approach, with the improvement of 5%, 3% and 8%, respectively for F measure.

4. Conclusion

We have proposed a summarization method by extracting definition sentences about IT person from Korean news articles. Centroid words are selected from the candidate sentences which are extracted from newspaper search engine by measuring their co-occurrence with the query term. All candidate sentences extracted from newspaper search engine are ranked by similarity measure with the centroid vector.

To improve the recall performance, the weight measure of centroid words is supplemented by using external knowledge resource such as Wikipedia and redundant candidate sentences are removed from candidate definitions. If we employ some extra important features such as named entities, sentence position, sentence length and headline in news article which have been used in the text summarization[8], we expect some improvement of performance will be attained.

참 고 문 헌

[1] D. Radev, H. Jing and M. Budzikowska, "Centroid based Summarization of Multiple Documents", Proceeding of ANLP/NAACL 2000 Workshop on Automatic Summarization, Seattle, WA, pp. 21-29, 2000.

[2] W. Hilderbrandt, B. Katz and J. Lin, "Answering definition questions using multiple knowledge sources", Proceedings of HLT/NAACL2004, Boston, MA, pp.49-56, 2004.

[3] B. Schiffman, I. Mani and K. J. Conception, "Producing biographical summaries: Combining linguistic knowledge with corpus statistics", Proceedings of European Association for Computational Linguistics, pp. 450-457, 2001.

[4] U. Y. Nahm and R. J. Mooney, "Mining soft matching rules form textual data", Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 979-986, 2001.

[5] http://ko.wikipedia.org

[6] http://www.etnews.co.kr/

[7] http://people.joins.com/

[8] C. Nobata, S. Sekine and H. Isahara, "Evaluation of features for sentence extraction on different types of corpora", Proceedings of ACL 2003 workshop on multilingual summarization and question answering, Vol. 12, pp. 29-36, 2003.