퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 6, pp. 813-818

Definition Sentences Recognition Based on Definition Centroid

김권양
Kweon-yang Kim

School of Computer Engineering, Kyungil University

Abstract

This paper is concerned with the problem of recognizing definition sentences. Given a definition question like "Who is the person X?", we are to retrieve the definition sentences which capture descriptive information correspond variously to a person's age, occupation, or some role a person played in an event from the collection of news articles. In order to retrieve as many relevant sentences for the definition question as possible, we adopt a centroid based statistical approach which has been applied in summarization of multiple documents. To improve the precision and 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. We see some improvements obtained by our approach over the baseline for 20 IT persons who have high document frequency.

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

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.

In spite of the great success of web search engines, such as Google(*http://www.google.com*) and Yahoo!(*http://yahoo.com*), which help people explore the web to find useful information, people still need the retrieval of exact answers in response to a query instead of returning just a list of relevant web pages.

Recent interest in question answering has focused on answering definition questions such as "Who is X?" or "What is Y?", which are different from factoid questions such as "When was X born?" or "Where was X born?" [1]. 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. On the other hand, definition questions 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 X?" are the form of following:

- X(born January 9, 1942) is the current chairman of Samsung Group.
- X has an Economics degree from Waseda University in Tokyo and an MBA from George Washington University in the United States.
- In 1996, X became a member of the International Olympic Committee.
- With an estimated net worth of $3.4 billion, X rank among the Forbes richest people in the world.
- 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.
- X has four children, one son and three daughters.
- In September 2006, X received the James A. Van Fleet Award from The Korea Society.

These definition sentences capture descriptive information which corresponds variously to a person's age, origin, education background, occupation, or some role a person played in an event.

As users are gradually coming to play a central role in the web contents, modeling personal information about person and relations among them will increasingly be important. Our research is concerned with the problem of identifying definition sentences relevant to a given person, specially with definition question about IT person, from the relevant news articles in vast text collections. To retrieve definition sentences about IT person de-

접수일자 : 2007년 9월 18일
완료일자 : 2007년 10월 20일

scribed in the news articles, 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. Related Work

A number of techniques have been proposed based on extracted terms and textual information. A set of words extracted from web documents are used as features for extracting definition sentences[2].

In typical information extraction system, definition patterns from the free text are manually constructed in a labor intensive manner, and are usually represented in the form of regular expressions. These trained patterns are matched against test sentences through the strict slot by slot matching for each position. Although manually constructed definition patterns provide high precision, they suffer from considerable labor and lack flexibility to obtain and maintain them.

Supervised learning for inducing definition sentence patterns can compensate this weakness to some extent, but it is limited by availability of annotated corpora, which also requires intensive labor[3,4,5].

The problem of discovering knowledge in the textual data is a new area in data mining. Existing text mining systems discover rules that require exactly matching strings; however, due to variability and diversity in natural language data, some form of soft matching based on textual similarity is needed.

Recently event-based approaches which treat a news topic as a series of sub-events, attempt to select and organize sentences in a summary with respect to events that the sentences describe, but not benefit much than traditional approaches because it is difficult to automatically break a news topic into sub-events[6,7].

The vector space model has been used in information retrieval to determine the similarity of two documents. We use the vector space model from information retrieval to provide an appropriate similarity measure[8]. The reason of using vector space model is that it is suitable for a web site with thousands of web pages like a newspaper web site.

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.

In this paper, we concentrate on a method for improving the recall of recognizing definition sentences.

3. Ranking Definition Sentences

The news articles, we are considering in this paper, contain descriptions of the salient attributes and activities of IT persons, along with lists of their associates such as person or organization. Our approach is more data driven, relying on discovering how IT persons are actually described in the collection of news articles retrieved from the general newspaper search engine.

3.1. System description

We present an overview of our definition sentence extraction system, which uses multiple components. Figure 1 shows high level view of our system which recognizes definition sentences.

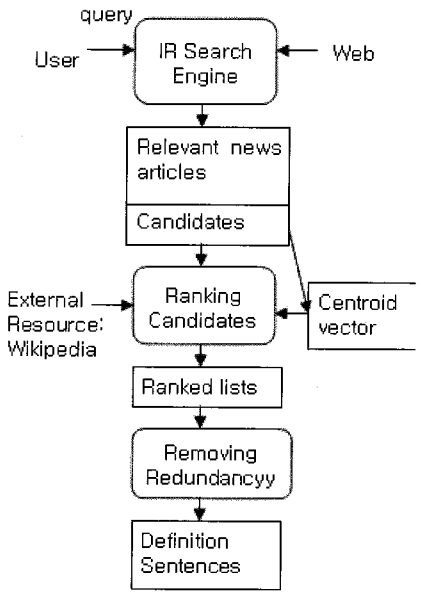

Fig. 1. The architecture of our extraction system for definition sentences.

The input to the our system is a query term, specially IT person name. We then apply the query term to search engine, the relevant news articles which contain the query term are retrieved. Candidate sentences which contain the query term are used for constructing a definition centroid for each query term. All candidate sentences are ranked by using a measure of similarity based on the centroid vector with supplement of external knowledge resource Wikipedia. After removing the redundant candidate sentences, the definition sentences are retrieved.

3.2. Acquire relevant 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", because of ambiguity problem that multiple person have same name. For example, 41 persons have same person name "Lee, Jae Yong" who is working for different organization.

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.

3.3. Ranking candidate definitions

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[9]. 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 [3]. 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 Ws is the surrounding word and Tq is the given query term. $f(Ws)$ denotes the number of sentences containing the word Ws and $f(Ws,Tq)$ is the number of sentences where word Ws co-occurs with the query term Tq. $\log_2^{N/n}$ is the measure of inverse document frequency(IDF) of word Ws, n is the number of articles including word Ws and N is the number of all 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[10,11].

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.

$$\text{Centroid(query termi)} = (\text{weight}(w_i^0),\ \text{weight}(w_i^1),\cdots,\text{weight}(w_i^n))$$

where weight(w_i^j) means centrality weight of a centroid word w_i^j for a query termi.

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 [12]) 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.

3.4. Removing redundant candidates

Candidate sentences tend to repeat some of information present in other candidate sentences. Therefore, these redundant sentences should be removed during the ranking candidate definition sentences.

After ranking the candidate sentences, we obtain a ranked list of definition candidates for each query term.

In the collection of news articles from several newspaper agency, there are duplicate sentences. 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 $R(S_i)$, denotes redundancy, is computed by counting number of overlap words between sentences Si 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:

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

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 Si. Function $f(w,S_i)$ denotes the number of common word w occurs in the sentence Si 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. Suppose we have a ranked list (S_1, S_2, S_3, ···, S_n) of candidate sentences for definition sentences, where S_1 has the highest similarity score with the definition centroid and S_n has the lowest score. If sentences S_3 and S_{10} highly overlap with sentence S_1 which has the highest score in the ranked list, sentence S_1 will inhibit S_3 and S_{10} by having their scores penalized.

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

4. Experiments

In this paper we only consider definition of person specially IT persons who are working for IT organization, we do not consider definitions of technical terms. To acquire the relevant collection of documents, we used the Electronic Times Internet[13] 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 who are working for IT-related organizations[14].

We supplement the person name with the name of organization for the works in order to solve same-name problem that multiple person have 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 2. Ranked list of centroid words for person #1(professor)

Organization #1: ICU
Centroid words(Noun)
ASEAN, Postec, ICU, ITRC, AMP, CMU, chair-professor, Malaysia, research fund, Heo unna, graduate, participant, reputation, summit meeting, research center, degree, president, Linux, KAIST
Centroid words(Verb)
establish, promise, propose, select, accompany, open, conclude, attend, support, magnify, develop

Table 3. Ranked list of centroid words for person #1(CEO).

Organization #2: Skylake incuvest
Centroid words(Noun)
Robotever, Olaworks, PEF, Jindaeje Fund, skylake incuvest, representative, financial supervisory service, venture fund, Kim Chang Geun, robot enterprise, SIC, Private Equity Fund, examiner, venture capital industry, investment, robot industry, fund, venture capital, skylake, Hynix
Centroid words(Verb)
establish, invest, induce, make, point out, accompany, succeed, be important, accomplish, attend, open, explain, need, support, use, be most difficult

As an example of generated centroid words, Table 2 and Table 3 show the higher ranked centroid words for his two recent occupations(professor, CEO) of person #1.

These words as shown in Table 2 and Table 3 are taken as centroid words that characterize the person #1 for the respective occupations. Since these centroid words receive higher weights, candidate sentences containing these words are likely to be ranked higher as a definition sentence.

As we do not have standard list for the definition questions as answers, three different assessors were asked to select definition sentences from the candidate sentences. In the decision process, assessors perform matching at the conceptual level, abstracting away from issues such as vocabulary differences, syntactic divergences, paraphrases, etc. The baseline system uses only the centroid based approach to rank candidate sentences. Table 4 shows the performance of our system for the definition question extraction.

Table 4. The performance of definition sentence extraction.

	F measure	Improvement (over baseline)
Centroid(baseline)	30%	-
Centroid + Wikipedia	35%	5%
Centroid + Redundancy	33%	3%
Centroid+Wikipedia +Redundancy	38%	8%

F-measure is defined as the harmonic mean for precision and recall as follows:

$$F-measure=\frac{2\times \mathrm{Precision}\times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$$

where Precision = the number of correct sentences marked by the system/the total number of correct sentences marked humans and Recall = the number of correct sentences marked by the system/the total number of sentences marked by the system.

From the Table 4, 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.

5. Conclusion

In this paper, we have proposed a method of extracting definition sentences about IT person from Korean news articles. Although the nature of descriptions in the news articles can very, we focus that facts about a person's life are extracted. In order to retrieve as many relevant sentences for the query term as possible, we adopt a centroid based statistical approach.

For the given definition question, specially IT person, we defined the definition centroid consists of centroid words which are more informative or indicative of the definition sentences. 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[15], we expect some improvement of performance will be attained.

In future work, we plan to construct automatically IT person event ontology from the definition sentences which our system extracted.

References

[1] E. M. Voorhees, "Evaluating Answer to Definition Questions", *Proceeding of HLT-NAACL*, pp. 109-111, 2003.

[2] A. K. McCallum, "BOW: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow, 1996.

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

[4] H. Cui, M. Y. Kan and T.S. Chua, "Unsupervised learning of soft patterns for generating definitions

from online news", *Proceedings of the 13th World Wide Web conference, New York*, pp. 90-99, 2004.
[5] J. Xu, R. M. Weischedel and A. Licuanan, "Evaluation of an extraction-based approach to answering definitional questions", *Proceedings of SIGIR'04, Sheffield, UK*, pp. 418-424, 2004.
[6] N. Daniel, D. Radev and T. Allison, "Sub-event based Multi-document Summarization", *Proceedings of the HLT-NAACL 2003 Workshop on Text Summarization*, pp. 9-16, 2003.
[7] E. Filatova and V. Hatzivassiloglou, "Event-based Extractive summarization", *Proceedings of ACL 2004 Workshop on Summarization*, pp. 104-111, 2004.
[8] G. Salton, *Automatic text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, 1989.
[9] D. Radev, H. Jing and M. Budzikowska, "Centroid based Summarization of Multiple Documents", *Proceeding of ANLP/NAACL'00 Workshop on Automatic Summarization, Seattle, WA*, pp. 21-29, 2000.
[10] 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.
[11] 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.
[12] http://ko.wikipedia.org
[13] http://www.etnews.co.kr/
[14] http://people.joins.com/
[15] 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.

저 자 소 개

김권양(Kweon Yang Kim)
1983년 : 경북대학교 전자공학과(학사)
1990년 : 경북대학교 전자공학과(석사)
1998년 : 경북대학교 컴퓨터공학과(박사)
1983~1988년 : ETRI 연구원
1999년~2000년 : University of Central Florida 방문교수
1991년~현재 : 경일대학교 컴퓨터공학부 교수

관심분야 : 시멘틱웹, 한글공학
Phone : 053-850-7287
Fax : 053-850-7609
E-mail : kykim@kiu.ac.kr