의미검색을 위한 지식표현 연구

김명관° 박영택
서울보건대학, 숭실대학교
binsum°@shjc.ac.kr

A Study on Knowledge Representation for Semantic Search

Myung Gwan Kim° Young Tack Park
Seoul Health College, Soongsil University

요 약

웹은 사람만이 읽을 수 있는 자연언어 문장들로 구성되어있다. 웹을 기계가 이해할 수 있게 하기 위해 의미적 표기로 구성되어야 한다. 광대한 웹의 성격상 수작업으로 이를 해결할 수는 없다. 따라서 본 연구에서는 링크 파서 및 개념그래프를 사용하여 자연어 문장을 지식표현으로 변환하고 이에 대한 검색을 다룬다. 기존의 연구에서는 3쌍으로 이루어진 지식표현과 검색으로 접근하고 있다. 그러나 이 경우 각 구(Phrase) 사이에 관계를 표현할 수가 없다. 또한 동의어 및 다의어에 대한 문제가 발생한다. 본 연구에서는 이 문제를 해결하기 위해 개념그래프를 사용하여 단어 사이의 의미를 표현하며 동의어 및 다의어 문제를 해결하기 위해 다중 단어로 된 동의어 즉 동일구(Paraphrase)를 사용한다. 이 경우 의미검색에서 다의어 및 동의어 문제가 개선됨을 보였다.

1. 서 론

웹은 사람만이 읽을 수 있는 자연언어 문장들로 구성되어있다. 웹을 기계가 이해할 수 있게 하기 위해 의미적 표기로 구성되어야 한다. 광대한 규모인 웹의 성격상 수작업으로 이를 해결할 수는 없다. 따라서 본 연구에서는 링크 파서 및 개념그래프(Conceptual Graph)를 사용하여 자연어 문장을 의미 표현으로 변환하고 이에 대한 검색을 다룬다. 기존의 연구에서는 3쌍으로 이루어진 지식표현과 검색으로 접근하고 있다.[1] 그러나 이 경우 각 어절 사이에 관계를 표현할 수가 없다. 이를 해결하기 위해서 개념그래프를 사용한다. 또한 동의어 및 다의어에 대한 문제가 발생한다. 예로 "light"는 "heavy"의 반대와 "illumination" 등 여러 가지의 의미를 갖는다. 이런 애매한 의미가 의미검색[4] 작업에서 정확도(Precision)를 떨어뜨린다. 동의어(Synonymy)는 서로 교환해서 사용할 수 있는 단어들을 말한다. 그러나 많은 단어들이 다의어이기 때문에 동의어에 대한 정확한 정의는 **특정 문맥에서만 교환될 수 있는 단어**라고 할 수 있다. 이렇게 정확히 교환 가능한 동의어를 발견 할 수 있을 경우에 만 회상도(Recall) 및 정확도를 개선할 수 있다. 대부분의 의미표현 자동 생성에 사용한 온톨로지 도구들이 워드넷과 같이 단일 동의어를 사용하는 경우이므로 다의어를 정확하게 구분하여 처리할 수가 없다.

이를 해결하기 위해 복수 단어로 이루어진 동의어를 구성하여야 한다. 즉 문장에 같은 의미를 지닌 다른 표현들을 가진 온톨로지가 필요하다. 이 복수 단어 동의어를 동일구(Paraphrase)라고 부른다. 이 경우 단점과 장점을 갖는다. 단점은 단일 동의어에 비해 처리해야할 엄청나게 커다란 규모의 대상 도메인을 갖는다는 것이다. 장점은 복수 단어의 동의어를 사용할 경우 **다의어 문제가 거의 발생하지 않는다는 것**이다[5].

본 논문에서는 이를 위해 2장에서 자연어 지식표현 자동생성 연구와 링크 문법, 의미그래프(Conceptual Dependency), 워드넷, 동일구(Paraphrase) 관련연구를 설명하고 3장에서는 지식표현 자동생성방법을 제시한다. 또한 이를 사용한 의미검색 시스템 구축을 4장에서 논의한다.. 결과로 기존에 비해 정확도와 회상도가 얼마나 개선되는지에 대한 실험과 결과를 5장에서 보여준다.

2. 관련 연구

웹의 지식표현자동생성 및 검색에 대한 연구는 대표적으로 3쌍의 단어 표현으로 지식표현을 하고 이를 SQL 방식으로 검색해주는 MIT의 Sapere[1]의 연구와 각 웹 사이트의 자연어 내용을 링크 문법과 워드넷을 이용 웹 페이지의 주석을 세만틱웹의 RDF로 작성해 주는 Li[12]의 연구 등이 있다. 본 논문에서는 좀더 일반화된 자연어 웹 페이지의 지식표현 구조를 생성해 주는 후자의 방법을 채택하였다. 그러나 이 경우 지식표현을 구성한 후 의미검색시스템과 연계할 때 매칭과정이 필요하다[6]. 대표적인 의미검색시스템은 질의 그래프와 리소스의 부분 그래프 사이를 유질동형(Isomorphism) 기반의 의미검색을 하는 OntoSeek[7]와 이웃한 노드들의 유사성을

기반으로 한 Similarity Flooding[8], 온톨로지를 이용하여 그래프의 패스를 따라 같은 위치의 같은 의미의 짝들에 점수를 누적하여 유사값을 구하는 Anchor-PROMPT[9] 등이 있다.

그러나 이들 기법들의 대부분은 단일 단어로 이루어진 유사어를 사용한다. 이 때문에 다의어(Polysemy) 문제가 발생한다. 즉 문맥에 따라 다른 의미로 사용하는 경우 때문에 의미검색 시 정확도와 회상도(Recall)를 떨어뜨린다. 이를 해결할 수 있는 방법이 복수 단어로 된 유사어를 구하는 동일구(Paraphrase)를 사용하는 것이다[5].

3. 자연어 의미표현 자동생성

본 논문에서 제시하는 자연어 지식표현 자동생성은 다음과 같은 과정으로 이루어진다.

1. 웹 문서 수집
2. Tag 등 불필요한 정보 제거
3. Link Parser를 이용한 형태소 분석
4. 구문 분석 결과 중 주어-동사-목적어, 명사-명사-수식, 전치사-명사-수식 등 관계 분석
5. 관계 분석된 결과를 개념그래프로 표현
6. 개념 그래프로 표현된 의미표현을 데이터베이스 관계로 변형하여 데이터베이스 구축

예를 들어 MIT의 Sapere시스템에서 문제가 되었던[1] 다음과 같은 문장은 위와 같은 과정을 거쳐서 의미표현을 하게 된다.

- 문장 :

[the president.n of Russia visited.v to the president.n of China]

- Link Parser를 이용한 형태소 분석

(m)	LEFT-WALL	Wd	<---Wd---->	Wd	president.n
(m)	the	D	<---Ds---->	Ds	president.n
(m)	president.n	Ss	<---Ss---->	S	visited.v
(m)	president.n	M	<---Mp---->	Mp	of
(m)	of	J	<---Js---->	Js	Russia
(m)	visited.v	O	<---Os---->	Os	president.n
(m)	the	D	<---Ds---->	Ds	president.n
(m)	president.n	M	<---Mp---->	Mp	of
(m)	of	J	<---Js---->	Js	China

- 개념그래프표현(CG):

[visited] [Person: president] <- (Agnt) <- [Russia]
[Person: president] <- (Dest) <- [China] .

- 데이터베이스표현(DBF):

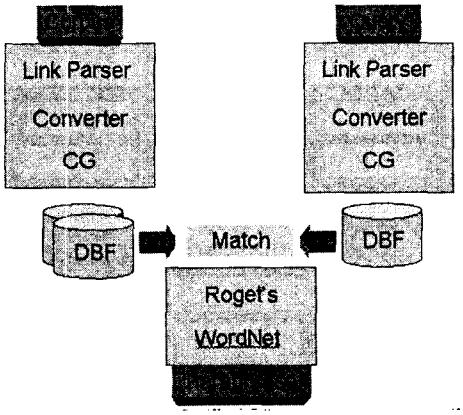
R	S	O
Visited	president	president
Agnt	president	Russia
Dest	president	China

(그림 1) 자연어문장의 지식표현 변환 과정

기존 시스템에서는 visited에서 어떤 President가 방문자인지 나타나지 않지만 우리 지식표현에서는 개념그래프를 사용하여 각 명사들의 관계가 잘 나타나게 된다.

4. 동일구(Paraphrase)를 이용한 의미검색시스템

3장에서 언급된 과정을 거쳐서 생성된 지식표현은 질의어와 데이터베이스로 나누어져서 의미검색이 이루어진다. 의미검색은 데이터베이스의 지식표현 문장과 질의어의 유사도를 구하는 과정으로서 유사도가 가장 높은 문장들이 질의어에 대한 결과로 주어진다. 이 과정에서 동일구를 사용하기 위하여 단일단어 유의어를 가진 워드넷과 다중 단어 유의어 표현을 포함하는 로젯의 시소러스[10]를 사용한다. 로젯의 시소러스는 워드넷과 함께 자연어처리 과정에 가장 많이 사용하는 유의어 시소러스이다.[11] 로젯의 시소러스는 예를 들어 "forty"-"quarter of hundred" 등의 다중 단어 유의어를 포함하고 있기에 워드넷에 보완적인 역할이 가능하다. 이밖에 MIT의 Ibrahim[5]이 구성한 동일구도 일부 사용하였다.



(그림 2) 동일구를 사용한 의미검색시스템 구성

5. 실험 및 평가

본 논문의 실험과 평가를 위해서 미국 CIA에서 매년 발표하는"http://www.cia.gov/cia/publications/factbook/"에 있는 2002 CIA World Fact book을 기반으로 의미변환 및 검색시스템을 구현하였다.

질의어는 다음과 같은 문장들을 선정하였다.

Q1. When did FUJIMORI's election?
Q2. When has Iraq invaded Kuwait?

Q3. When did the Kosovo war start?
Q4. When did hurricanes take place?
Q5. What is the largest country in the world?

이와 같은 질의어로 시스템을 평가하는 기본 목적은 시스템의 기본 성능을 평가하기 위한 것이 아니라 동일구를 사용한 의미변환 및 검색 시스템의 장점을 강조하기 위한 것이다. 따라서 위에 언급한 예제들을 그림 8.의 알고리즘을 사용하여 의미변환을 수행하고 다이스 계수를 통한 유사도 검색을 수행하였을 때 표 1.과 같은 결과를 얻을 수 있었다.

<표 1> 우리 시스템과 관계만을 사용한 시스템, 키워드만을 사용한 시스템의 비교

n : 검색된 문장의 수, I : 질문에 대한 바른 답변, p : 정확도

	System			Relation			Keyword		
질문	n	I	p	n	i	p	n	i	p
Q1	3	3	1	2	2	1	43	2	0.046
Q2	2	2	1	1	1	1	72	4	0.005
Q3	3	1	0.66	2	0	0	65	1	0.015
Q4	2	1	0.5	1	0	0	138	2	0.014
Q5	1	1	1	1	1	1	35	1	0.028
평균	2.2	1.6	0.72	1.4	0.8	0.57	70.6	2	0.028

6. 결 론

본 논문에서는 기존 의미검색 시스템의 문제점이었던 단위 구 간의 관계표현의 결여, 다의어 및 동의어 미처리 등을 개선한 의미변환 및 검색시스템을 제안하였다. 단위 구 간의 관계 표현은 기존 시스템에서 고려하고 않으나 의미검색의 중요한 요소이다. 이를 위해 링크 문법과 개념 그래프를 사용해서 구(Phrase)들 사이의 관계를 표현하였다. 또한 한 단어가 여러 가지 의미를 갖는 다의어(Polysemy) 문제는 전혀 다른 결과를 검색하게 한다. 동의어가 질의어에 추가되지 않으면 역시 원하는 결과의 일부를 찾지 못하게 된다. 이와 같은 문제를 해결하기 위해서 같은 의미를 갖는 다중단어와 유사어 등 동일구를 사용하였다. 이 동일구 구축을 위해서 워드넷과 로젯의 시소러스 및 MIT Ibrahim이 생성한 어휘를 사용하였다. 이렇게 구축한 동일구(Paraphrase)를 사용하여 의미를 축소(다의어처리)하고 확장(유사어처리)하여 검색의 정확도와 회상도(Recall)를 개선하였다. 이를 실험 및 평가하기 위해 미국 CIA의 World Book 2002 사이트의 자료를 기반으로 수행하였다. 실험결과, 기존 관계정보 만을 사용한 시스템에 비해 실험 데이터에서 더 좋은 성능을 보여주었다.

참고 문헌

[1] J. Lin, "Indexing and Retrieving Natural Language Using Ternary Expression", Master's Thesis, Massachusetts Institute of Technology, 2001

[2] D.D. Sleator and D. Temperly, "Parsing English with a Link Grammar", Third International Workshop on Parsing Technologies, 1993

[3] G. A. Miller, "WordNet: An On-Line Lexical Database", International Journal of Lexicography, Vol. 3, No. 4, 1990

[4] H. Zhu and J. Zhong, "An Approach for Semantic Searh by Matching RDF Graphs", American Association for Artificial Intelligence, 2002

[5] A. Ibrahim, "Extracting Paraphrases from Aligned Corpora", MIT Master degree Thesis, 2002

[6] J. Poole and J. Campbell, "A Novel Algorithm for Matching Conceptual and Related Graphs", In Proceedings of the Third International Conference on Conceptual Structures(ICCS '95), pp. 293-307, 1995

[7] N. Guarino, "OntoSeek: Content-Based Access to the Web", IEEE Intelligent System 14(3), pp. 70-80, 1999

[8] Melnik, S., "Similarity Flooding: A Versatile Graph Matching Algorithm", In Proceedings of the 18th International Conference on Data Engineering(ICDE), 2002

[9]Tomek Strzalkowski,"Natural Language Information Retrieval", In Proceedings of the 5th Text Retrieval Conference, 1996

[10] Peter M. Roget, "Roget's Thesaurus", Gramercy Books, 1979

[11] M. L. Hale, "A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity", CMP-LG, 1998