질의응답 시스템에서 의미 연관성 참조를 위한 온톨로지의 자동 구축

김혜정° 강보영 황선욱 이상조
경북대학교 컴퓨터공학과
{hjkim325° comeng hodduk}@sejong.knu.ac.kr, silee@knu.ac.kr

Automatic Ontology Construction for Semantic Relevance in Question Answering System

Hae-jung Kim° Bo-Yeong Kang Sun-wook Hwang Sang-Jo Lee
Dept. Computer Engineering, Kyungpook National University, Korea

요 약

본 논문에서는 질의응답 시스템에서 질의에 포함된 언어 정보와 검색 대상 문장 사이의 의미 연관성을 참조하여 정확한 결과를 추출 가능하도록 하는 온톨로지의 자동 구축 방법을 제시한다. 검색 대상 문장은 웹에서의 활용과 표준화를 위하여 단어 태그, 품사 정보 및 파싱 구조를 갖는 XML 문서로 변환하고, 이 구조를 이용한 연관성 분석을 위해 의미망을 갖는 온톨로지를 자동으로 생성할 수 있도록 하였다.

온톨로지에서 의미 연관성을 결정하는데 중요하게 활용되는 개념으로써는 동사의 행위, 명사절 그룹 매치, 복합명사 선별, 고유명사 매치, 품사 태깅 등이 있다. 제안한 방법의 성능은 NIST TREC-10의 질의 응답문을 사용해서 단어 패턴 매치 방법과 비교 분석하였으며, 본 논문에서 제안한 방식이 재현율과 정확율 측면에서 우수한 성능을 나타냄을 입증하였다.

1. 서 론

정보검색 시스템은 사용자가 요구하는 특정 정보를 포함하고 있는 문서를 효율적으로 찾아내는 것을 목표로 한다. 반면에 질의응답 시스템은 질의수행 결과로서 문서 전체가 아니고 필요한 정보를 포함하고 있는 단어절, 혹은 문장과 같은 간략하고 잘 정의된 응답을 출력해야한다[1]. 특히 질의로부터 응답을 추출하는 과정(answer extraction) 에서는 대상 문장과 질의 사이에서 유사도(similarity)를 비교해서 순위화(ranking)할 수 있는 방법이 중요하게 다루어지고 있다[1,2].

본 논문에서는 자연어 질의응답 시스템에서 질의에 포함된 언어 정보와 검색 대상 문장 사이의 의미 연관성(semantic relevance)을 참조하여 정확한 결과를 추론 가능하도록 하는 온톨로지의 자동 구축 방법을 제시한다. 제안한 온톨로지는 동사의 행위, 명사절 그룹 매치, 복합명사 선별, 고유명사 매치, 품사 태깅 등을 의미 연관성을 결정하는 개념으로 활용하고 있다. 이러한 개념은 질의문과 대상문 사이에 유사성을 계산하는 중요한 척도로 사용되고 있다. 제안한 방법의 성능은 대상 문장과 질의문에 포함된 단어를 패턴 매치하는 방법과 비교하여 분석하였으며, 본 논문에서 제안한 방식이 재현율(recall)과 정확율(precision) 측면에서 우수한 성능을 나타냄을 확인하였다.

2. 관련 연구

최근 정보검색 분야에서는 사용자의 의도에 맞는 효율적인 검색을 위해 온톨로지를 구축하여 활용하고자 하는 다양한 시도가 이루어져 왔다. 그 중에서 Jose[3]등은 NLP 기술을 이용하여 동사의 행위를 중심으로 온톨로지의 자동 구축 방안을 제시한 바가 있다. 그러나 이 연구는 온톨로지에 내포된 의미 관계가 동사의 행위만을 중심으로 구축되었기 때문에 적용에 한계가 있었다. 또한 Amalia[4]등은 시소러스나 사전을 참고하여 반자동으로 온톨로지를 구축하는 방법을 제안하였으며, Sabrina[5]등은 온톨로지를 이용해서 문서의 주제를 식별할 수 있는 방법을 제시하였다.

질의응답 시스템에서는 Marco[2]가 대상 문장과 질의 사이에서 의미 연관성을 이용한 유사도 계산 척도를 제시한 바가 있으며, Robert[6] 등은 질의응답과 정보검색을 융합한 모델을 제시하였다. 본 논문은 질의응답 시스템 구현 시 문장 간의 의미 연관성을 추론하기 위해 온톨로지의 개념을 적용하되, 대상 문장으로부터 온톨로지를 자동으로 구축하여 활용하고자 하는데 목적이 있다.

3. 온톨로지의 자동 구축 방법

그림 1은 본 논문에서 제안하는 온톨로지의 자동 구축 방안과 활용 과정을 보여준다.

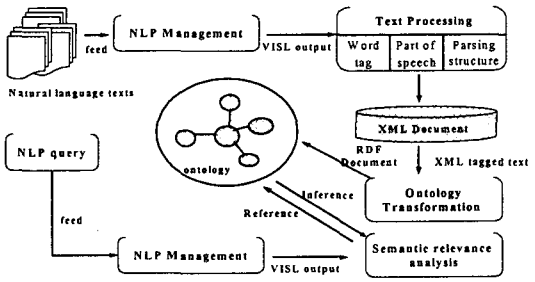


그림 1. 온톨로지의 자동구축과 활용

그림 1에서 보는 것처럼 자연어 대상문은 NLP 과정을 거쳐 파싱된 출력 구조를 생성한다. 다음 단계에서는 파싱된 출력 구조를 텍스트 처리하여 단어 태그와 품사, 파싱 구조를 갖는 XML 문서로 변환하며, 온톨로지 변환 모듈에서는 XML 문서에서 의미 연관성과 관련된 성분을 추출하여 온톨로지를 자동적으로 구축하게 된다. 구축된 온톨로지는 자연어 질의 처리 시에 의미 연관성 분석의 용도로 활용되며, 주어진 질의 역시 NLP 처리에 의해 파싱된 출력구조로 변환하고, 이 출력구조에서 추출한 중심어 성분과 온톨로지 탐색에 의해 얻어진 의미 연관성 정보를 사용하여 검색 대상문을 정확하게 식별하게 된다.

본 연구에서는 NLP 과정의 메인 파서로서 E.Bick에 의해 개발되어진 VISL 프로젝트(http://visl.hum.sdu.dk/visl [7]) 파서를 사용하였다. VISL 문법적 파서의 경우 대부분의 영어 문장을 파싱하고 출력할 수 있지만, 온톨로지의 자동 구축을 위해서는 이러한 비표준 형식의 출력을 XML로 전환하는 것이 필요하다. XML 문서는 현재 웹 응용에서 표준 형식이 되고 있으며 XSLT 등 강력한 도구에 의해 문서의 형식간 변환이 용이하다는 장점이 있다.

우선 본 연구에서는 온톨로지의 자동 구축에 필요한 구문적 성분과 의미적 성분을 XML 형식으로 표현하기 위해 JAVA API에서 제공하는 메소드들을 이용해 VISL 파싱된 출력 구조로부터 단어 태그, 품사 및 파싱 구조를 갖는 XML 문서로 자동으로 생성 하였다.

이 과정을 예를 들어 설명하면, 자연어 대상문 "Myeondong offers cultural variety."에 대한 VISL 출력 구조는 다음과 같다.

```
STA : fcl
SUBJ : np
    ⇒ N : n ( 'Myeondong' M S )     Myeondong
P : v-fin ( 'offer' PS 3S IND )     offer
ACC : np
    ⇒ N : adj ( 'cultural' F S )    cultural
    ?H : n ( 'variety' F S )        variety
```

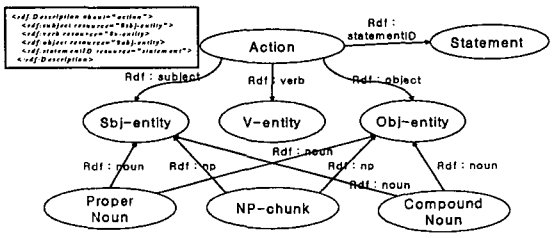
위에 기술한 VISL 출력 구조로부터 Java API 프로그래밍을 이용하여 단어 태그, 품사 정보, 파싱 구조를 갖는 XML 문서가 자동으로 생성되며, 파싱 구조에 대한 XML 문서의 생성 결과를 예로 들어보면 다음과 같다.

```
<text>
<paragraph id="paragraph_1">
<sentence id="sentence_1 span="word_1..word_5">
<chunk id="chunk_1" ext="subj" form="np" span="word_1">
</chunk>
<chunk id="chunk_2" ext="p" form="v" span="word_2">
</chunk>
<chunk id=chunk_3" ext="acc" form="np" span="word_3..word_4">
<chunk id="chunk_4 ext="n" form="adj" span="word_3">
</chunk>
<chunk id="chunk_5 ext="h" form="n" span="word_4">
</chunk>
</chunk>
</sentence>
</paragraph>
</text>
```

변환된 XML 문서로부터 XSLT를 이용해 동사의 액션을 중심으로 한 chunk들과, 명사절 chunk, 복합 명사, 고유 명사, 품사 정보들을 참조하여 RDF 인스턴스를 생성한다. 이 때 적용하는 온톨로지 변환 규칙은 다음과 같다.

```
1)If(ext="subj"or"acc") and{form="np"})
  thenADDrelation(np{np-chunk,entity})
2)If(form="np") and{span="word..word"})
  thenADDrelation(c-noun{compoundnoun,entity})
3)If(form="n")and(att="p-n")
  thenADDrelation(p-noun{propernoun,entity}
```

위 규칙을 적용하여 자동으로 생성된 온톨로지의 구성은 그림 2와 같다.



<그림 2> 온톨로지의 구성

4. 온톨로지 기반 질의 분석

의미 연관성 참조를 위한 온톨로지를 구축한 다음의 단계에서는 구축된 온톨로지를 참조하면서 사용자의 질의를 처리하게 된다. 그림 1에서처럼 질의문은 NLP 과정을 거쳐 VISL 출력 구조로 변환되고, VISL 출력구조로부터 중심 단어를 식별한 후 온톨로지 참조를 통해 의미 연관성을 분석한다. 의미 연관성은 질의의 중심어와 대상 문장 사이에 동사의 액션을 중심으로 한 chunk들과, 명사절 chunk, 복합 명사, 고유 명사, 품사 정보들을 매치하는 과정으로 볼 수 있다. 다음의 예는 온톨로지로부터 NP chunk를 식별하여 유사도가 높은 대상 문장을 가려내는 과정을 보여준다.

질의문: When did the big bad wolf eat red riding hood.
중심어 추출: q=([the, big, bad, wolf], eat, [red, riding, hood])
온톨로지 참조: NP chunk의 식별
대상문s1 : [The bad wolf] ate [red riding hood] who carried [a pink chicken].
대상문s2 : [The wolf] who wore [a pink riding hood] ate [the bad red chicken].
분석 결과: semantic-relevance(s1, q) > semantic-relevance(s2, q)

5. 평가

본 연구에서는 제안한 방법의 성능을 평가하기 위하여 NIST TREC-10[8]에서 50개의 질의를 추출하고 NIST에서 제공하는 15개의 문서에서 응답문을 추출하였다. 수작업으로 상위 3개의 응답 문장을 순위화하여 50개의 질의를 순서대로 실행한 결과를 MRR(Mean Reciprocal Rank) 측면에서 분석하였다. 본 논문에서 제안한 방법과 단어의 패턴 매치 방법을 50개의 테스트 질의문에 대하여 비교한 실험 결과는 표 1과 같다.

<표1> 질의의 실험결과 비교

	응답의 수				MRR	정답비율 (%)
	순위1	순위2	순위3	합계		
패턴매치	14	6	5	25	0.27	50.40
온톨로지	16	5	6	27	0.29	54.80

실험결과, 본 논문에서 제안한 온톨로지 기반의 의미 연관성 검사 방식이 질의문에 포함된 단어를 패턴 매치하는 방식에 비해 정확한 답을 구하는 비율이 높음을 확인할 수 있었다. 이는 본 논문에서 제안한 방식이 의미 연관성을 반영함으로써 문장 간의 유사도 측정에 효율적임을 나타내 보여준다..

6. 결론

본 논문에서는 문장 간의 유사도를 산출하기 위해 의미 연관성을 참조할 수 있는 온톨로지를 자동으로 구축하는 방법을 제안하였다. 제안한 방법은 자연어 문장을 문법적 파서를 이용하여 파싱한 후 XML 변환 및 XSLT를 이용하여 동사의 행위, 명사절 그룹 매치, 복합명사 선별, 고유명사 매치, 품사 태깅 등의 의미정보를 갖는 온톨로지를 자동으로 구축한 후, 질의 처리 시에 이 온톨로지를 참조함으로써 질의와 가장 유사도가 높은 문장을 식별할 수 있도록 하였다. 제안한 방법의 성능을 평가한 결과 온톨로지 기반의 의미 연관성 검사 방식이 질의문에 포함된 키워드를 패턴 매치하는 방식에 비해 우수함을 확인할 수 있었다.

향후 연구는 보다 다양한 의미 연관성을 추가하여 온톨로지를 확장해 나가는 문제와 질의문에 포함된 중요 성분들을 일차 논리 등 온톨로지 검사에 유리한 형식으로 자동 변환하는 방법을 개발하는 것이다.

참고 문헌

[1] Ido Milstein and Magnus Sandberg. Using Semantic Networks for Question-Answering. in CS224N/Ling237 Final Projects 2000

[2] Marco De boni and Suresh Manandhar. The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering in Proceedings of the AAAI Symposium on New Directions in Question Answering, Stanford, 2003

[3] Jose Saias and Paulo Quaresma. Using NLP techniques to create legal ontologies in a logic programming based web information retrieval system. In Proceedings of the ICAIL - International Conference on Artificial Intelligence and Law, 2003.

[4] Amalia Todirascu and Francois de Beuvron. Using Description Logics for Ontology Extraction in Proceedings of ROMAND'2000 Workshop on Robust Parsing, Lausanne, 19-20 October 2000, pp. 89 - 105.

[5] Sabrina T.A, Rosni Abdullah and Tang Enya Kong. Automatic Topic Identification Based on Extended Ontology Hierarchy. in Proc. of the 2nd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)

[6] Robert Caisauskas and Kevin Humphreys. A Combined IR/NLP Approach to Question Answering Against Large TextCollections.

[7] http://visl.hum.sdu.dk/visl