질의 응답 시스템에서 질의 카테고리별 개념리스트 구축에 기반한 의미적 질의 확장

김혜정° 강보영 박성배 이상조
경북대학교 컴퓨터공학과
{hjkim325°, comeng}@sejong.knu.ac.kr, {seongbae, sjlee}@knu.ac.kr

Semantic Query Expansion based on a Question Category Concept List in QA system

Hae-Jung Kim° Bo-Yeong Kang Seong-Bae Park Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook National University, Korea

요 약

질의 응답(Question Answering) 시스템은 질의에서 요구하는 정답 유형(Answer tyype) 및 질의에 사용된 용어를 적용하여 보다 정확한 답을 추출하고자 한다. 그러나 질의에 사용된 용어들이 문서의 정답문장에 그대로 사용되지 않고 같은 의미의 다른 어휘로 출현하기도 하며, 혹은 다른 문법적 정보를 가진 카테고리로 등장하여 정답 추출에 어려움이 따른다. 따라서, 본 논문은 질의별 카테고리 개념 리스트를 구축하여 효과적인 의미적 질의 확장 방법론을 제안한다. 제안된 방법은 먼저 질문 문장의 패턴 및 질의 정보 유형을 파악하여 질의 카테고리 및 카테고리별 개념 리스트를 구축한다. 그런 후 구축된 질의 개념 카테고리 및 리스트를 활용하여 질의 유형을 학습하고, 새로운 질의가 입력되면 해당 개념 카테고리로 분류한 후, 개념 리스트를 기반으로 개념별 질의 확장을 수행한다. 제안된 시스템의 성능 평가를 위하여, TREC-9의 질의와 TREC 문서 중 1991년도 WSJ(Wall Street Journal) 42,654건을 대상으로 실험한 결과 질의 확장을 수행하지 않는 시스템의 경우 MRR(Mean reciprocal ratio) 측정에서 0.223의 결과를 보인 반면 제안된 시스템의 경우 0.50의 향상된 결과를 보였다.

1. 서 론

정보 검색(Information retrieval) 시스템에서는 주어진 사용자의 질의에 대해 가장 관련이 있을 정답이 포함된 문서들이 추출 되지만, 좀 더 사용자 의도를 잘 파악하고 주어진 질의에 명확한 대답을 줄 수 있는 시스템의 필요성이 대두되었다. 따라서, 이러한 요구를 만족시키기 위하여 질의와 밀접한 연관성을 갖는 단어, 문단, 절을 추출하고 순위화한 후 가장 연관성이 있는 정답을 추출하는 질의 응답 시스템에 관한 연구가 활발히 진행되고 있다[1]. 그러나 질의와 정답에 사용된 어휘들이 문서의 정답 문장에 그대로 사용되지 않고 같은 의미의 다른 어휘로 출현하기도 하며, 혹은 다른 문법적 정보를 가짐으로써 정확한 정답 추출에 어려움이 따른다. 예제 질의와 정답 문장을 살펴보자.

◆ 질의 문장 : Who is the inventor a paper?
◆ 정답 문장 : A devised paper from china.....

예제 질의를 처리를 할 경우 먼저 사용자 질문을 분석하여 질의를 의미 분류 체계인 카테고리별로 분류하여야 한다. 이때 단순히 근접성(proximity)에 기반한 질의 응답 시스템으로 처리할 경우 Wh-term의 "Who"를 보고 사람의 이름을 얻기 위한 "PERSON"이란 카테고리로 분석하게 되고, 질의 속에 포함된 어휘 중 주요한 키워드로서 "NAME", "inventor", "paper"가 추출되어 정답 분석에 사용 된다. 그러나 단순한 질의 응답의 경우 키워드만으로 다른 문법적 정보를 가진 어휘 "devised"로 구성된 정답 문장을 찾아내기가 사실상 어렵다. 즉, 질의 속의 키워드인 "inventor"를 보고 질의 확장을 하더라도 어휘의 문법적 카테고리가 다르므로 "inventor"의 동사형인 "invent"를 찾을 수 없을 뿐만 아니라, 동사 "devise"의 동의어가 "invent"라는 정보까지 알아야 정확한 정답 문장을 추출할 수가 있다. 만약 "inventor"라는 용어가 문법 카테고리 정보와 관계없이 "discoverer, make, create, devise, invent, develop, creator"와 같은 의미적으로 밀접한 어휘로 확장될 수 있다면 보다 정확한 정답 추출이 가능할 것이다.

따라서 본 논문은, 사용자가 정보 요구를 위해 적용하는 질의의 경우 문장 패턴이 비교적 고정된 형식이고, 질의에서 찾고자 하는 정보의 유형 또한 그룹화 할 수 있는 특성을 활용하여, 질의 카테고리별 개념 리스트 구축에 기반한 효과적인 의미적 질의 확장 방법론을 제안한다.

2. 관련 연구

질의 응답 시스템에서는 정답 추출과 성능향상을 위해 정확한 정답 유형의 분류와 불일치 되는 단어 문제를 해결하기 위한 질의 확장이 필요하다. 질의 확장을 위해서는 대표적으로 시소러스가 많이 이용되는데 Voorhees 는 워드넷(WordNet)을 사용하여 질의 내의 모든 어휘들에 대해 동의어, 반의어, 상위어등을 확장해서 비교적 질의의 길이가 짧은 경우에 대한 성능 향상을 보였다[2]. 또한 Moldovan은 워드넷에서 단어 정의문에 기술된 단어들과, 상위어, 하위어, 유사어에 가중치를 주어 관련된 단어들의 사슬을 형성하고, 이 단어 사슬을 이용하여 관련된 문헌을 찾을 수 있음을 보였다[3]. Mandala et al.[4]는 여러개의 이질적(heterogeneous) 시소러스를 사용하여 용어들의 가중치를 결합 평균값을 계산함으로써 가장 높은 확률을 가진 용어에 대해 확장을 하였다. 그러나 이러한 기존 연구들은 문법 카테고리를 넘어선 질의 확장에는 어려움이 있을 뿐만 아니라, 시소러스의 동의어나 상하위어 정보에서는 찾을 수 없지만 사용자가 같은 개념을 표현하기 위해 주로 사용하는 어휘로의 질의 확장은 거의 불가능하다. 따라서, 본 논문에서는 주요 개념을 표현하는 어휘를 위주로 의미적 개념 리스트를 구성하고, 워드넷의 동의어 정보와 상위어 정보를 위주로 자동적인 학습을 통해 개념 리스트를 확장한다. 새로운 질의에 대해서도 정답 카테고리로 분류하고 질의 확장을 통한 정답 추출이 가능하도록 하였다.

3. 제안된 시스템의 구조

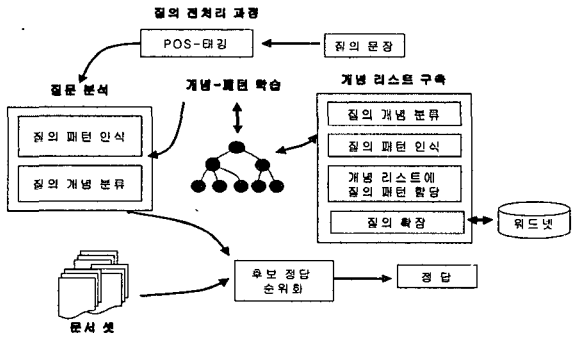

그림 1 제안된 시스템의 전체적인 구성

제안된 시스템의 구조는 크게 개념 리스트 구축 모듈과 개념 학습 모듈, 질의 확장 모듈의 세 가지로 나누어진다. 개념 리스트 구축 모듈 단계에서는 의미적 질의 확장을 위한 질의 카테고리별 개념 리스트를 구성한다. 구성된 카테고리별 개념 리스트는 워드넷을 이용하여 확장된다. 개념 학습 모듈은 확장된 개념 리스트를 활용하여 해당 개념을 표현하는 어휘를 학습한다. 마지막으로 질의 확장 모듈에서는 주어진 질문에 대해 학습된 지식을 이용하여 질문 카테고리로 분류하고 이미 구축된 개념 리스트를 참고하여 의미적 질의 확장을 수행한다.

3.1 질의 카테고리 분류

질의 응답 시스템에서 정답 유형의 판단은 정답을 찾을 검색 공간과 시간을 상당히 줄일 수 있기 때문에 중요하다. 본 논문은 질의에서 자주 사용된 어휘들이 질의의 주요 개념을 표현할 수 있다고 가정하고, 용어 빈도에 기초하여 질의 개념을 분류한다. TREC-9 질의 201~893 중 Who 질문 117개의 전체 용어 빈도에 대한 상위 30%를 주요 개념으로 간주하고 일반화 및 구체화를 통하여 질문 카테고리를 분류한다. 세분화된 질문 카테고리는 표 1과 같다. 또한 용어 빈도에 기초하여 Who에 이어지는 중요한 용어들을 고려하여 보면 "INVENTOR", "KILLER", "WRITER", "LEADER", "PLAYER", "FOUNDER", "OWNER", "OTHERS" 등의 하위 카테고리로 세분화 할 수 있다.

표 1 Who 용어들에 대한 세부적 질의 카테고리

질문 카테고리		예 제
PERSON_NAME	INVENTOR	Who invented television?
	KILLER	Who killed Martin Luther King?
	WRITER	Who wrote the Farmer's Almanac?
	LEADER	Who is the prime minister of Australia?
	PLAYER	Who is the fastest swimmer in the world?
	FOUNDER	Who is the founder of the Wal-Mart stores?
	OWNER	Who is the owner of CNN?
	OTHERS	Who is Coronado?

3.2 질의 패턴 추출과 의미적 질의 확장

본 절은 구축된 질의 카테고리 개념을 표현하기 위해 자주 사용되는 어휘들을 추출하는 방법을 소개한다.

[질의 패턴 정의]

문장 패턴은 의문사를 중심으로 주변 명사(N), Be동사(BE_V) 및 일반 동사(V) 태그 셋을 가진 형태로 아래와 같이 두 가지 유형으로 정의된다.

- 질의 패턴 = [Wh_term, N1, BE_V, N2]
- 질의 패턴 = [Wh_term, null, V]

예문 "Who is an inventor of a paper?" 에서 추출되는 문장 패턴은 <Who, null, is_BE, inventor_NN>이다.

정의된 형태로 질의로부터 추출된 패턴은 해당 개념 카테고리로 할당된다. 할당된 패턴에서 카테고리 개념을 보다 풍부하게 표현하기 위해 워드넷을 이용하여 패턴들을 확장

한다. 이때 Be_V가 포함된 패턴은 명사들만 일반 동사가 포함된 패턴은 해당 동사만 확장한다. 그림 2는 추출된 패턴의 해당 카테고리로의 할당 및 패턴 확장을 설명한다.

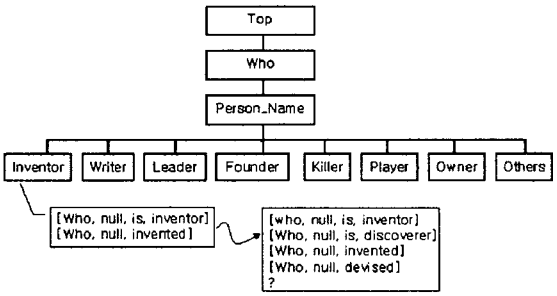

그림 2 개념별 질의 패턴 할당 및 확장

4. 실험 및 평가

제안된 시스템의 성능을 검증하기 위한 실험은 TREC-9의 질의 중 who 질의를 대상으로 두 가지 방법으로 평가되었다. 먼저 TREC-9의 201~893번 질문 692개 중에서 who 질의 117개를 사용하여 질의의 개념을 표현 할 수 있는 질의 카테고리 및 카테고리별 개념 리스트를 구축하고, 개념 리스트는 나이브 베이지안 분류 알고리즘을 이용하여 카테고리별로 학습하였고, 정확한 학습 성능을 검증하기 위해 10-fold 교차 검증(cross validation)을 수행하고 정확율을 기준으로 성능을 실험하였다. 이때 각 패턴에 대한 정확율은 표 3과 같고, 전체적인 정확도의 경우 학습된 패턴에 대해서는 94.44%, 학습되지 않은 패턴에 대해서도 78.70%의 정확도를 얻을 수 있었다.

표 2 패턴에 대한 학습 정확율

	학습된 집합	교차 검증 집합
1	0.917874	0.695652
2	0.966184	0.782608
3	0.966184	0.869565
4	0.917874	0.826086
5	0.951691	0.826086
6	0.951691	0.826086
7	0.946860	0.869565
8	0.971014	0.826086
9	0.932367	0.695652
10	0.922705	0.652173
평균	0.944444	0.786955

또한 제안된 질의 확장의 시스템 성능 테스트를 위해서는 TREC에서 1991년 WSJ(Wall Street Journal) 42,654건의 문서 및 해당 문서에 대한 TREC-9 의 6개 who 질의를 사용하였다. 실험에 사용된 모든 질의 및 문서는 POS태깅, 스태밍, 불용어 제거등의 전처리 과정을 거친다. 질의와 대상문서 사이의 유사도는 질의와 대상문서에서 단어가 일치하는 경우는 1 그렇지 않은 경우에는 0값으로 나타내었다.

표 3은 대상문서에서 정답을 포함한 문장의 최대 크기를 세 문장으로 보았을 경우 질의에 대한 결과이며, 평가 척도로는 MRR(Mean Reciprocal Rank)의 평균을 사용하였다.

표 3 기존 시스템과 제안된 시스템과의 MRR

	기존 시스템	제안된 시스템
세 문장	0.223	0.50

5. 결론 및 향후 연구

본 논문에서는 질의 응답 시스템의 질의에서 자주 사용된 어휘들이 질의의 주요 개념을 표현할 수 있다고 가정하고, 질의 유형의 특성을 의미적으로 분석하여, 질의 카테고리 개념 리스트로 구성하고, 개념 리스트를 기반으로 효과적인 의미적 질의 확장 방법론을 제안하였다.

그러나 제안된 시스템에서는 사용된 질의의 카테고리 유형이 who만을 대상으로 하였기 때문에 when, why와 같은 다른 질의 카테고리에 대한 확장이 필요하고, 학습 문장 패턴에 대해서도 광범위한 데이터에 대해 실험을 확장시켜 봄으로써 학습 성능을 높일 수 있다면 더 좋은 시스템 성능을 얻을 것으로 기대된다.

참 고 문 헌

[1] S. Na, I.Kang, O. Kwan, J. Lee, "Answer Candidate Ranking based on syntactic Proximity in Question Answering", In Proceedings of the 29th KISS sprint conference, pp.478-480, 2002.

[2] E. M. Voorhees, "Query expansion using lexical-semantic relations", In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Informatin Retrieval, pp. 61-69, 1994.

[3] Dan I. Moldovan, A. Novischi, "Lexical Chains for Question Answering" In proceedings of the COLING, pp. 325-332 , 2002.

[4] R. Mandala, T. Tokunaga, H. Tanaka, "Query expansion using heterogeneous thesauri" In Proceedings International Journal archive Volume 36, pp.361 - 378, 2000