질의 응답 시스템에서 심층적 질의 카테고리의 개념 커버리지에 기반한 의미적 질의 확장

(Semantic Query Expansion based on Concept Coverage of a Deep Question Category in QA systems)

김 혜 정 † 강 보 영 †† 이 상 조 †††
(Hae-Jung Kim) (Bo-Yeong Kang) (Sang-Jo Lee)

요 약 질의응답(Question Answering) 시스템은 질의에서 요구하는 정답 유형(Answer type) 및 질의에 사용된 용어를 적용하여 보다 정확한 답을 추출하고자 한다. 그러나 질의에 사용된 용어들이 문서에 그대로 사용되지 않고 같은 의미의 다른 어휘로 출현하기도 하며, 혹은 다른 문법적 정보를 가진 카테고리에 등장하여 정답 추출에 어려움이 따른다. 만약, 질의에서 요구하는 정보유형을 보다 깊게 세분화하고, 세분화된 질의 유형과 개념적으로 유사한 문장을 대상으로 정답 추출을 수행할 수 있다면 보다 정확한 정답을 추출할 수 있을 것이다. 따라서, 본 논문은 심층 질의 카테고리의 개념 커버리지에 기반한 효과적인 의미적 질의 확장 방법론을 제안한다. 질의에서 요구하는 정보 유형을 보다 세분화된 심층 질의 카테고리로 나누고, 이러한 심층 질의 카테고리를 표현하기 위해 동원되는 어휘 집합에 질의 확장을 적용함으로써 정답 추출의 성능을 향상시키고자 하였다. 제안된 시스템의 성능 평가를 위하여, TREC 문서 중 1991년도 WSJ(Wall Street Journal) 42,654건과 TREC-9의 질의를 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.

키워드 : 질의 응답 시스템, 질의 카테고리 개념 리스트, 질의 확장

Abstract When confronted with a query, question answering systems endeavor to extract the most exact answers possible by determining the answer type that fits with the key terms used in the query. However, the efficacy of such systems is limited by the fact that the terms used in a query may be in a syntactic form different to that of the same words in a document. In this paper, we present an efficient semantic query expansion methodology based on a question category concept list comprised of terms that are semantically close to terms used in a query. The semantically close terms of a term in a query may be hypernyms, synonyms, or terms in a different syntactic category. The proposed system constructs a concept list for each question type and then builds the concept list for each question category using a learning algorithm. In the question answering experiments on 42,654 Wall Street Journal documents of the TREC collection, the traditional system showed in 0.223 in MRR and the proposed system showed 0.50 superior to the traditional question answering system. The results of the present experiments suggest the promise of the proposed method.

Key words : Question Answering System, Question Category Concept List, Query Expansion

1. 서 론

자연어는 사용자가 필요한 정보를 가장 잘 표현할 수 있는 방법이다. 현대의 자연어 처리 시스템인 정보 검색(Information retrieval) 시스템에서는 주어진 사용자의 질의에 대해 가장 관련이 있을 정답이 포함된 문서들이 추출 되지만, 사용자 의도를 잘 파악하고 좀 더 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

† 정 회 원 : 경북대학교 컴퓨터공학과
hjkim325@hanmail.net
†† 정 회 원 : 한국정보통신대학교 컴퓨터공학과
kby@icu.ac.kr
††† 종신회원 : 경북대학교 컴퓨터공학과 교수
sjlee@knu.ac.kr
논문접수 : 2004년 10월 1일
심사완료 : 2005년 1월 31일

질의응답 시스템에 관한 연구가 활발히 진행되었다[1-3]. 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. 아래 예제 질의와 정답 문장을 살펴보자.

• 질의문장 : Who is the inventor of a paper?

• 정답문장 : A devised paper from China in A.D.

예제 질의를 질의응답 시스템으로 처리를 할 경우 먼저 사용자 질의를 분석하여 의미 분류 체계인 질의 카테고리별로 분류하여야 한다. 이때 기존의 질의응답 시스템으로 처리할 경우 질의 분석을 위한 의미 분류 체계 가운데 사람을 나타내는 "PERSON"으로 분류되고, "PERSON"의 하위 범주 가운데 사람의 이름을 나타내는 "NAME"의 질의 유형에 속한다. 즉, Wh-term의 "Who"를 보고 사람의 이름을 얻기 위한 "PERSON"이란 카테고리로 분석하게 되고, 정확한 카테고리 분석을 위한 주요한 키워드로서 질의 속에 포함된 어휘 중 "NAME", "inventor", "paper"가 추출되어 정답 분석에 사용 된다. 그러나 기존의 질의응답 시스템의 경우 키워드인 inventor 만으로 같은 의미 이지만 문법적으로 다른 어휘인 "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 수있다. 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.

기존의 질의응답 연구들은 질의 확장을 위하여 주로 시소러스의 동의어 및 상하위어를 사용하는 방법론을 제안하였다[4-6]. 그러나 이러한 기존 연구들은 질의어의 명사 "inventor"를 동사 "devise"로 문법 카테고리를 넘어 확장하는데 있어 어려움이 있다. 또한, 상하위어로의 확장에서 확장 범위에 대한 경계가 불확실하며, 질의어에 동사 "invent"가 사용되었을 경우, 시소러스의 사용자가 같은 개념을 표현하기 위해 주로 사용하는 어휘인 동사 "make"로의 확장은 거의 불가능하다.

만약 질의에서 요구하는 정답유형과 개념적으로 유사한 문장을 대상으로 정답추출을 수행할 수 있다면 보다 정확한 정답을 추출 할 수 있을 것이다. 기존의 질의응답 시스템도 이러한 가정하에 질의에서 요구하는 정보 유형, 혹은 질의 유형을 의미체계별로 나누어 정답 추출을 수행하고 있다. 그러나 기존의 질의응답 시스템에서 정의하여 사용하고 있는 질의 유형은 "person, Location, Organization" 등 상대적으로 상위의 개념으로 체계화 되어있다. 질의 유형이 상위 개념일수록 해당 질의 유형을 표현하기 위해 사용되는 어휘의 범위도 넓게 분포된다. 그러나 질의 유형이 하위의 개념 즉, 보다 깊게 표현될수록 해당 질의 유형을 표현하기 위해 사용되는 어휘 집합 또한 좁게 분포될 수 있다.

본 논문은 이렇게 같은 질의 개념을 표현하기 위해 동원되는 어휘 집합을 질의 확장을 위해 응용하고자 하며, 따라서 심층 질의 카테고리의 개념 커버리지에 기반한 효과적인 의미적 방법론을 제안한다. 제안된 방법은 먼저 질의에서 찾고자 하는 정보의 유형을 보다 세부적으로 그룹화 할 수 있는 특성을 활용하여, 질의 문장 패턴에 기반하여 심층 질의 카테고리 및 카테고리별 리스트를 구축한다. 획득된 각 카테고리별 리스트는 해당 질의 카테고리의 개념을 표현하기 위해 동원되는 어휘 집합을 의미한다. 그런 후 심층 질의 개념 카테고리에서 획득한 카테고리별 리스트를 활용하여 각 어휘 집합에 대한 질의 확장을 수행한다. 제안된 시스템은 질의를 주요 개념에 따라 보다 깊게 세분화하고 해당 개념을 표현하는 다른 어휘들로 의미 확장을 수행할 수 있도록 함으로써 보다 정확한 정답을 추출할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 먼저, 2절에서는 질의응답 시스템에 대한 관련 연구들을 설명하고, 3절에서는 제안한 시스템의 전체적인 구성을 살펴보고 심층 질의 카테고리 및 질의 카테고리 개념 리스트 구축 방법과 질의 유형 학습 방법에 대하여 자세하게 설명한다. 4절에서는 제안된 시스템과 질의 확장을 수행하지 않은 시스템간의 비교 분석을 수행하고, 마지막으로 5절에서 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

질의응답 시스템에서는 정답 추출과 성능향상을 위해 정확한 정답 유형의 분류와 불일치 되는 단어 문제를 해결하기 위한 질의 확장이 필요하다. 또한, 질의 확장을 위해서는 시소러스가 많이 이용되는데 Voorhees는 워드넷(WordNet)[7]을 사용하여 질의 내의 모든 어휘들에 대해 동의어, 반의어, 상위어 등을 확장하고, 비교적 질의의 길이가 짧은 경우에 대한 성능 향상을 보였다[8]. Moldovan 등은 워드넷 용어 풀이에 포함된 정보가 문맥 해석에 도움이 된다는 것에 착안하여 워드넷에서 원 단어 정의문에 기술된 단어들에 대해 상위어, 하위어, 유사어 등을 확장하고 이를 위해 유사도 리스트(similarity list)를 구축하였다[9]. 이 연구에서는 원래의 질의와 비슷한 단어들로 확장을 하여 검색 대상 문서 집합의 범위를 확대하였으며, 완전히 해로운 단어들을

확장함으로써 검색 정확률을 향상 시켰다. Mandala 등은 일반적인 연구에서처럼 역시 시소러스를 사용하였으나 하나의 시소러스만으로는 좋은 성능을 얻기가 어렵기 때문에 특징이 다른 여러 개의 이질적(heterogeneous) 시소러스를 사용하여 용어들의 가중치를 결합하고 평균값을 계산함으로써 가장 높은 확률을 가진 용어에 대한 확장 방법을 제안하고 한 가지 타입의 시소러스만을 사용하는 것 보다 성능이 우수함을 입증하였다[10].

Prager 등은 질의 확장을 위해 워드넷(WordNet) 시소러스를 사용하였는데 대상 문서에서 명사절, 전치사구, 부사구등을 이용하여 질의 유형별로 20개의 카테고리로 분류하였다. 정답 추출을 위해 미리 정답 후보들을 찾아 색인하는 방법(predictive annotation)을 적용하고 단순한 패턴 매칭 기법에 의해 식별할 수 있도록 하였다. 그러나 간혹 어떤 용어에 대해서는 워드넷의 상하위 관계를 이용하여도 상위어를 찾을 수 없을 경우가 생기고 이때는 다른 외부적인 시소러스의 사용이 필요하다고 하였다[11,12].

Kiyota 등은 비슷한 어휘 또는 구 표현에 대한 사전을 구축하여 질의의 단어와 문서의 단어 사이의 대응에 이용하였고[13], Hovy 등은 질의 형태에 가능한 다른 정답 유형을 분류한 분류표와 워드넷의 용어 설명, 약어, 동의어 등의 정보를 이용하여 문서로부터 관련된 정답을 찾아 기술하였다[14].

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 어휘의 질의 확장은 거의 불가능하다. 따라서 본 논문은 질의에서 검색하고자 하는 정보의 유형을 표현하기 위해 동원되는 어휘 집합이 있다고 간주하고, 이러한 정보 유형을 표현하는 어휘 집합을 질의 확장에 적용하여 정답 추출의 효율을 높이고자 한다.

3. 심층 질의 카테고리의 개념 커버리지에 기반한 의미적 질의 확장

사용자가 정보 요구를 위해 적용하는 질의의 경우 사용되는 문장 패턴이 비교적 고정된 형식이고, 질의에서 찾고자 하는 정보의 유형 또한 그룹화 할 수 있다. 따라서 질의응답 시스템에서 질의에서 요구하는 정답 유형과 개념적으로 유사한 문장을 대상으로 정답 추출을 수행할 수 있다면 보다 정확한 정답을 추출할 수 있을 것이다. 기존의 질의응답 시스템 또한 질의에서 요구하는 정보의 유형을 의미체계별로 나누어 정답추출을 수행하고는 있다. 질의 유형은 보통 상위 레벨(upper level), 하위 레벨(lower level)로 나뉘고 각자의 연구 필요성에 따라 상위 레벨을 세분화 하는 심층 레벨(deep level) 등의 세 가지로 나뉠 수 있다. Who의 경우 상위 레벨 질의 유형은 "Person, Location, Organization" 등으로 체계화 되어있다. 질의 유형이 상위 레벨일수록 해당 질의 유형을 표현하기 위해 사용되는 어휘의 범위도 넓게 분포되며, 질의 유형이 하위의 개념 즉, 보다 깊게 표현될수록 해당 질의 유형을 표현하기 위해 사용되는 어휘 집합 또한 좁게 분포될 수 있다. 하나의 깊은 질의 유형을 표현하기 위해 동원되는 어휘 집합은 하위 레벨로서 해당 개념을 표현하는 다양한 의미 관계의 어휘들로 구성된다.

따라서, 본 논문은 심층 질의 개념을 표현하기 위해 동원되는 어휘 집합 즉 심층 카테고리의 개념 커버리지 리스트를 질의 확장을 위해 응용하여 심층 카테고리의 개념 커버리지에 기반한 효과적인 의미적 질의 확장 방법론을 제안한다. 개념 커버리지 리스트란 같은 개념을 표현하는 문법적 정보가 다른 어휘들로 구성된 어휘집합을 나타낸다. 먼저 3.1절에서 전체적인 시스템 구성을 설명하고, 3.2절에서는 질의 카테고리 개념 리스트 구축에 대해서 자세히 설명한다. 다음 3.3 절에서 구축된 개념 리스트를 활용하여 질의 카테고리를 학습하는 방법론을 기술하고, 마지막으로 3.4절에서 의미적 질의 확장 과정을 설명한다.

3.1 제안된 시스템 구조

제안된 방법은 먼저 질의 문장 패턴 및 질의 정보 유형을 파악하여 심층 질의 카테고리 및 카테고리별 개념 리스트를 구축한다. 그런 후 구축된 심층 질의 개념 카테고리 및 리스트를 활용하여 질의 유형을 학습하고, 새로운 질의가 입력되면 해당 심층 질의 카테고리로 분류한 후 개념별 질의 확장을 수행한다.

전체적인 시스템의 구조는 그림 1과 같다. 시스템은 크게 개념 리스트 구축 모듈과 개념 학습 모듈, 질의 확장 모듈의 세 가지로 나누어진다. 개념 리스트 구축 모듈 단계에서는 의미적 질의 확장을 위한 심층 질의 카테고리별 개념 리스트를 구성한다. 이때 질의에 대해 사용자가 자주 사용하는 어휘를 대상으로 빈도수가 높은 어휘가 주요 개념을 나타내는 것으로 보고, 주요 개념별로 심층 질의 카테고리를 분류하고, 해당 질의 개념을 표현하기 위해 사용되는 어휘들로 각 카테고리별 리스트를 구성한다. 개념 학습 모듈은 개념 리스트를 활용하여 해당 개념을 표현하는 어휘를 학습한다. 마지막으로 질의 확장 모듈에서는 주어진 질의에 대해 학습된 지식을 이용하여 심층 질의 카테고리로 분류한 다음 이미 구축된 심층 질의 카테고리 개념 리스트를 참고하여 의미적 질의 확장을 수행한다.

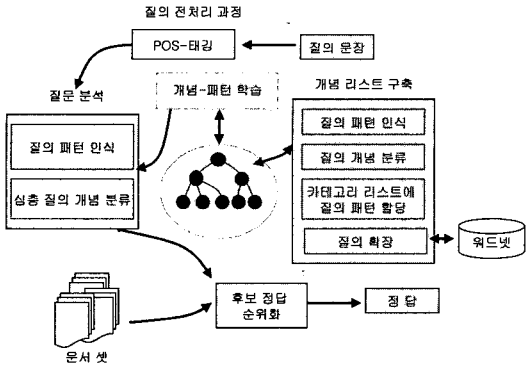

그림 1 제안된 시스템의 전체적인 구성

3.2 심층 질의 카테고리 및 카테고리 개념 리스트 구축

질의에 대해 사용자가 자주 사용하는 빈도수가 높은 어휘를 대상으로 주요 개념별로 심층 질의 카테고리를 분류한다. 또한, 해당 질의 개념을 표현하기 위해 사용되는 같은 의미의 문법적 정보가 다른 어휘들로 각 심층 카테고리의 개념을 공유하는 개념 리스트를 구성한다.

3.2.1 심층 질의 카테고리 분류

질의로부터 정답 유형을 판단하는 것은 질의응답 시스템의 기본적인 구성 요소이다. 정답 유형을 판단함으로써, 정답을 찾을 검색 공간을 상당히 줄일 수 있기 때문이다. 정답 유형은 워드넷과 같은 온톨로지에 정의된 개념 분류나 MUC에서 사용되었던 개체명 분류를 기반으로 정의 된다. 본 논문은 질의에서 자주 사용된 어휘들이 질의의 주요 개념을 표현할 수 있다고 가정하고, 온톨로지 구축시에 주로 활용되는 방법론을 적용하여 용어 빈도에 기초한 심층 질의 카테고리 분류를 수행하였다. 즉 질의에 자주 등장하는 어휘들이 질의의 정보 요구 유형을 잘 반영할 수 있다고 간주하고, 용어 빈도에 기초하여 일정 기준치 이상의 발생 빈도를 보이는 어휘를 질의에서 요구하는 주요 정보 요구 유형, 즉 심층 카테고리 후보 어휘로 추출하는 것이다. 질의 카테고리 분류에 사용되는 질의들을 대상으로 설명하면 다음과 같다. 예를 들어 TREC-9 질의 201~893 중 Who 질의 117개에 대하여 동사 및 명사의 용어 빈도수는 표 1과 같이 나타난다. 이때 전체 용어 빈도에 대한 상위 30%의 용어에 대해 질문을 구성하는 주요 개념으로 간주하고 일반화 및 구체화를 통하여 질의 카테고리를 분류한다. 계산된 용어 빈도수는 표 1과 같다.

표 1 용어 빈도수의 예

명사/동사	빈도수	명사/동사	빈도수
Founder_NN	13	Tallest_NN	8
Invented_VBD	11	Swimmer_NN	7
Wrote_VBD	11	Thought_VBD	7
Signed_VBD	11	Coronado_NN	7
Arthitect_NN	9	Emperor_NN	7
Creator_NN	9	Author_NN	7
King_NN	9	portrayed_VBD	7
President_NN	8	Inventor_NN	6
Pyramid_NN	8	Novel_NN	6
Book_NN	8	God_NN	6

추출된 용어 빈도에 기초하여 세분화된 질의 카테고리는 표 2와 같다. 예를 들어 Who의 경우는 주로 어떤 이벤트와 관련 있는 사람의 이름을 질문하므로 "PERSON_NAME"으로 분류가 가능하다. 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.

3.2.2 질의 패턴 추출과 확장

본 절은 구축된 심층 질의 카테고리 개념을 표현하기 위해 자주 사용되는 어휘들을 추출하여 심층 질의 카테고리 개념 리스트를 구성하는 방법을 소개한다. 본 논문은 질의가 표현하는 정보 유형을 보다 쉽게 파악하기 위하여 질의로부터 패턴을 추출하여 처리한다. 즉, "Who is an inventor of a paper?" 와 같은 질의는 "Who"와 "inventor"만 보아도 질의가 요구하는 개념 유형을 알 수 있으며, "Who invented a paper?"와 같은 질의에서는 "Who"와 "invented" 만 보아도 알 수 있다. 따라서 질의 개념 파악을 위한 질의 패턴은 다음

표 2 Who 용어들에 대한 심층 질의 카테고리

질의 카테고리			예 제
Who	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?

과 같이 정의된다.

[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. 이때, 명사 N1은 동사가 발생하기 이전의 명사이고, N2는 동사 이후의 명사이며, 해당 명사가 존재하지 않으면 null 처리된다.

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

예문 "Who is an inventor of a paper?"에서 추출되는 문장 패턴은 <Who, null, is_BE, inventor_NN>이고, 예문 "who invented a paper"는 <who, null, invented_VBN>가 된다. 정의된 형태로 질의로부터 추출된 패턴은 해당 개념 카테고리로 할당된다. 할당된 패턴에서 카테고리 개념을 보다 풍부하게 표현하기 위해 워드넷을 이용하여 패턴들을 확장한다. 이때 Be_V가 포함된 패턴은 명사들만 확장하며, 일반 동사가 포함된 패턴은 해당 동사만 확장한다. 즉 <Who, null, is_BE, inventor_NN>의 경우 Be동사를 포함한 패턴이므로 명사 "inventor"만 워드넷을 참조하여 확장된다. 또한 <who, null, invented_VBN> 패턴의 경우 일반 동사 "invented" 를 포함한 패턴이므로 일반 동사만 확장된다. 그림 2는 추출된 패턴의 해당 카테고리로의 할당 및 패턴 확장을 설명한다.

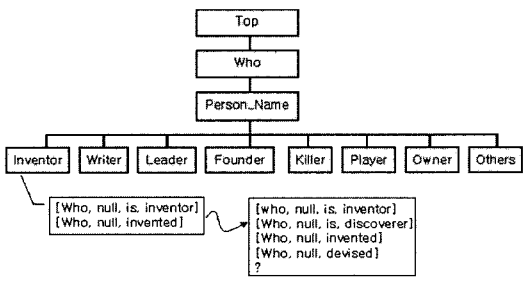

그림 2 심층 카테고리 개념별 질의 패턴 할당 및 확장

3.3 심층 질의 카테고리 학습

본 절은 심층 질의 카테고리 학습 및 분류 방법에 대해서 설명한다. 패턴 학습과 개념 분류를 위해서는 나이브 베이즈 분류자(Naive Bayes Classifier)를 사용한다. 먼저 학습할 질의에 대한 패턴을 구한 다음, 패턴 내 모든 단어들에 대해 각각의 단어가 카테고리에 속할 확률을 구한 후 학습한다. 학습 방법을 식으로 나타내면 다음과 같다.

$$P(S_k) = \frac{C(S_k)}{C(w)}, \quad P(v_j \mid S_k) = \frac{C(v_j, S_k)}{C(v_j)} \quad then\ k = 1\ to\ 8$$

어휘 w 는 주제를 분류하고자 하는 패턴에서의 키워드 단어이고 어휘 w와 함께 나타나는 주변 각 단어를 v_j, 어휘 w가 속할 수 있는 모든 질의 카테고리를 $S_k\,(k = 1\ to 8)$라고 하자. 이때, 어휘 w가 카테고리 S_k에 속할 확률 $P(S_k)$과 주변단어 v_j가 속할 수 있는 모든 질의 카테고리 S_k에 대한 확률을 계산한다. 또한, 새로운 질의가 주어졌을 경우 학습된 지식을 기반으로 질의가 속할 수 있는 심층 카테고리 S^* ($then\ *=1\ to\ 8$)를 분류하는 방법은 다음과 같다.

$$Decide\ S^*\ \text{if}\ S^* = \arg\max_{S_k}[\log P(S_k) + \sum_{v_j\ in\ C} \log P(v_j \mid S_k)]$$

즉, 새로운 질의가 주어졌을 경우 해당 질의 패턴이 어떤 카테고리인지의 확률 S^*은 학습된 지식 즉 모든 단어에 대한 $P(v_j \mid S_k)$ 및 $P(S_k)$를 적용하여 계산된 값 가운데 가장 큰값을 취해 해당 어휘에 대한 심층 카테고리로 정한다.

3.4 의미적 질의 확장

본 절에서는 새로운 질의가 입력되면 학습된 내용을 기초로 의미적으로 질의를 확장하는 부분에 대해서 기술한다. 제안된 시스템은 먼저 입력 질의의 패턴을 추출하고, 학습 지식을 기반으로 추출 패턴의 심층 질의 카테고리를 분류한다. 분류된 심층 질의 카테고리로부터 해당 개념을 공유하는 확장 어휘들을 획득하여 각 심층 카테고리별 개념 커버리지 리스트를 구성한다. 질의 확장 과정은 그림 3에 자세히 설명되어 있다.

예를 들어, 질의 "Who is an inventor of a paper?"와 같은 질의가 입력되면 질의의 개념을 파악하기 위해 먼저 질의 패턴을 추출한다. 그런 후 학습된 패턴 지식을 기초로 추출된 질의 패턴의 심층 개념 카테고리를 분류한다. 입력 질의의 개념은 심층 질의 카테고리 중 "INVENTOR"에 해당하고, "INVENTOR" 카테고리로부터 해당 개념을 공유하는 확장 어휘들을 사용하여 질의를 확장한다. 따라서 입력 질의는 "inventor, discoverer, invent, devise" 등으로 확장될 수 있다.

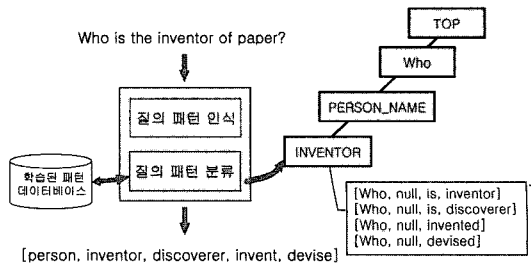

그림 3 개념별 질의 패턴 할당 및 확장

4. 실험 및 평가

본 절에서 제안된 시스템의 성능을 검증하기 위해 TREC-9[15]의 질의 중 Who 질의를 대상으로 다음의 두 가지 방법으로 평가되었다. 먼저 제안한 시스템의 심층 질의 카테고리 학습 및 분류 성능에 대한 평가를 수행하였고, 그런 후 제안한 질의 확장 방법을 적용한 질의응답 시스템과 질의 확장을 수행하지 않는 시스템의 성능을 비교 분석 하였다.

첫 번째 성능 평가를 위하여 TREC-9의 201~893번 질의 692개 중에서 Who 질의 117개를 사용하여 질의의 개념을 표현 할 수 있는 심층 질의 카테고리 및 심층 질의 카테고리별 개념 리스트를 구축하였다. 구성된 개념 리스트는 나이브 베이즈 분류자를 이용하여 각 질의 카테고리별로 학습하였고, 정확한 학습 성능을 검증하기 위해 10-fold 교차 검증(cross validation)을 수행하고 정확율(precision)을 기준으로 성능을 실험하였다. 즉, 주어진 패턴의 90%를 학습하고, 10%는 validation set으로 테스트하는 방법을 10번 반복하여 평균값을 계산하였다. 이때 각 패턴에 대한 정확률은 표 3과 같고, 전체적인 정확도의 경우 학습된 패턴에 대해서는 94.44%, 학습되지 않은 패턴에 대해서도 78.70%의 정확도를 얻을 수 있었다.

표 3 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

두 번째 성능 평가로 제안된 질의 확장의 시스템 성능 테스트를 위해서는 TREC에서 1991년 WSJ(Wall Street Journal) 42,654 건의 문서 및 해당 문서에 대한 TREC-9 의 6개 who 질의를 사용하였다. 실험에 사용된 모든 질의 및 문서는 POS태깅, 스태밍, 불용어 제거 등의 전처리 과정을 거친다. 질의와 정답 문서 사이의 유사도는 아래 수식과 같이 정의하여 사용하였다[12].

$$Sim(Q, D) = \alpha \cdot d(q_i, d_j), where \ \delta(q_i, d_j)$$

$$= 1 \quad if \quad q_i = d_j \quad otherwise \quad 0$$

위 수식에서 질의(Q)와 대상문서(D) 사이의 유사도(δ)는 질의와 대상문서에서 단어가 일치하는 경우는 1 그렇지 않은 경우에는 0값으로 나타내었다. α 는 가중치로서 질의어에 존재하는 단어와 일치하는 경우에 대해서는 1.0 을, 워드넷을 이용하여 확장된 단어와 일치하는 경우에는 가중치를 0.5 로 설정하여 실험하였다.

표 4는 대상문서에서 정답을 포함한 문장의 최대 크기를 세 문장으로 보았을 경우 각 질의에 대한 질의응답 시스템의 결과이며, 평가 척도로는 TREC에서 사용하는 것과 마찬가지로 각 질의에 대해 정답을 포함하는 첫 번째 문서의 순위를 역순한 MRR(Mean Reciprocal Rank)의 평균을 사용하였다.

표 4 기존 시스템과 제안된 시스템과의 MRR

	기존 시스템	제안된 시스템
세 문장	0.223	0.500

표 4는 대상문서에서 정답을 포함한 문장의 최대 크기를 세 문장으로 보았을 경우 각 질의에 대한 질의응답 시스템의 결과이다. 표 4에서 알 수 있듯이 제안된 시스템이 질의 확장을 수행하지 않은 단순한 질의응답 시스템에 비하여 좋은 성능을 보임을 알 수 있다.

5. 결론 및 향후 연구

현대의 자연어 처리 시스템인 정보 검색 시스템에서는 주어진 사용자의 질의에 대해 가장 관련이 있을 정답이 포함된 문서들이 추출 된다. 그러나 좀더 정확하고 명확한 정답 구나 절에 대한 정보를 줄 수 있는 시스템의 필요성이 있다. 따라서 자연어로 주어진 질의와 밀접한 연관성을 갖는 단어, 문단, 절을 추출한 후 순위화하여 가장 정확한 정답을 추출할 수가 있는 질의응답 시스템에 관한 연구가 활발히 진행 중이다.

본 논문에서는 질의응답 시스템의 질의에서 질의 유형의 특성을 의미적으로 분석하여 심층 질의 카테고리 및 카테고리의 개념별 리스트를 구성하고, 심층 질의 카테고리가 제공하는 카테고리 리스트를 기반으로 효과적인 의미적 질의 확장 방법론을 제안하였다.

TREC 질의 201~893 중 who 질의 117개에 대하여 온톨로지 구축 방법론중 하나인 빈도수를 이용하여, 동사 및 명사의 용어 빈도수를 구하고, 일반화 및 구체화를 통하여 세분화된 심층 질의 카테고리를 분류하고 카테고리 리스트를 구성하였다. 구성된 카테고리별 개념 리스트는 분류 학습에서 일반적으로 이용되는 나이브 베이즈 분류자를 이용하여 학습되었고, 10-fold 교차 검증을 수행한 결과 학습된 패턴에 대해서는 정확률

(precision)을 기준으로 94.44%, 학습되지 않은 패턴에 대해서도 78.70%의 정확률을 얻을 수 있었다. 또한, 질의응답 시스템에 적용하였을 경우 질의 확장을 수행하지 않은 질의응답 시스템 보다 제안된 시스템의 경우 높은 성능을 보임을 알 수 있었다. 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 데이타에 대해 실험이 필요하다.

참 고 문 헌

[1] M. Pasca and S. Harabagui, "High Performance Question/Answer," *In Proceedings of the 24rd ACM-SIGIR Conference*, pp. 366-374, 2001.

[2] J. Kupiec, "MURAX : A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia," In *Proceedings of the 16th ACM-SIGIR Conference*, pp. 181-190, 1993.

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

[4] R. Mandela, T. Tokunaga, and H. Tanaka, "Combining Multiple Evidence from Different Types of Thesaurus," In *Proceedings of the 17th ACM-SIGIR Conference*, pp. 15-19, 1999.

[5] J. Prager and J. C-Carroll, "Use of WordNet Hypernyms for Answering What-Is Questions," In *Proceedings of TREC-2001*, pp. 143-150, 2000.

[6] C. Cardie, V. Ng, D. Pierce, and C. Buckley, "Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System," In *Proceedings of the 6th ANLP*, pp. 180-187, 2000.

[7] G. Miller, "WordNet: A Lexical Database for English," In *Proceedings of the Communications of the ACM*, pp. 39-41, 1995.

[8] E. Voorhees, "Query Expansion using Lexical-Semantic Relations," In *Proceedings of the 17th ACM-SIGIR Conference*, pp. 61-69, 1994.

[9] D. Moldovan and R. Mihalcea, "Using WordNet and Lexical Operators to Improve Internet Searches," In *Proceedings of IEEE Internet Computing*, pp. 34-43, 2000.

[10] R. Mandela, T. Tokunage and H. Tanaka, "Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion," In *Proceedings of the 22nd Annual International ACM SIGIR Conference*, pp. 15-19, 1999.

[11] J. Prager, D. Radev, E. Brown, and A. Coden, "The Use of Predictive Annotation for Question Answering in TREC8," In *Proceedings of TREC-2000*, pp. 399-411, 2000.

[12] J. Prager, D. Radev, and K. Czuba "Answering What-Is Questions by Virtual Annotation," In *Proceedings of Human Language Technology Conference*, pp. 26-30, 2001.

[13] Y. Kiyota, S. Kurohashi, and F. Kido, "Dialog Navigator:A Question Answering System based on Large Text knowledge Base," In *Proceedings of COLING*, pp. 460-466. 2002,

[14] D. Ravichandran and E. Hovy "Learning Surface Text Patterns for a Question Answering system," In *Proceedings of the ACL Conference*, pp. 41-47, 2002.

[15] TREC(Text REtrieval Conference) : http://trec.nist.gov/overview.html

김 혜 정

1987년 경북대학교 수학과(이학사). 1989년 경북대학교 전자공학과 전산전공(공학석사). 2004년 12월 경북대학교 컴퓨터공학과(공학박사). 관심분야는 정보 검색, 자연어 처리, 시멘틱 웹, 기계 학습

강 보 영

1997년 경북대학교 컴퓨터공학과(공학사) 1999년 경북대학교 영어영문학과(문학석사). 2002년 경북대학교 컴퓨터공학과(공학석사). 2004년 8월 경북대학교 컴퓨터공학과(공학박사). 2004년 9월~현재 한국정보통신대학교(ICU) 박사후연구원. 관심분야는 자연어 처리, 정보 검색, 질의응답, 문서분류

이 상 조

1974년 경북대학교 수학교육과(이학사) 1976년 한국과학기술원(이학석사). 1994년 서울대학교 컴퓨터공학과(공학박사) 1976년~현재 경북대학교 컴퓨터공학과 교수. 관심분야는 자연어 처리, 정보 검색, 기계 학습, 운영체제, 프로그래밍 언어, 시멘틱 웹