어휘정보와 명사의미정보를 이용한 사용자 질의문장 분석

Question Analysis using Lexico Information and Noun Semantic Information

정규철, 서영훈
충북대학교

Choung Kyu-Chu, Seo Young-Hoon
Chungbuk National Univ.

요약

고성능의 질의 응답 시스템을 구현하기 위해서는 질의 유형 분류기의 성능이 중요하다. 본 논문에서는 복잡한 분류 규칙이나 대용량의 사전 정보를 이용하지 않고 질의문에서 의문사에 해당하는 어휘들을 이용하여 질의 유형을 결정하고, 의문사 주변에 출현하는 명사들의 의미 정보를 이용하여 세부적인 정답유형을 결정할 수 있는 질의 유형 분류기를 제안한다. 의문사에 해당하는 어휘가 생략된 경우는 질의문의 마지막 어절의 의미 정보를 이용하여 질의 유형을 분류한다. 의문사 주변의 명사들이 마지막 어절에 출현하는 명사들에 대해 동의어 정보와 접미사 정보를 이용하여 질의 유형 분류의 성능을 향상시킨다. 본 논문에서 제안한 시스템은 질의 유형에 대한 분류는 97.4%의 정확도를 보였다.

1. 서론

인터넷의 등장으로 정보 자료의 양이 증가함에 따라 수많은 문서들 중에서 사용자가 필요로 하는 문서만을 제공하는 요구가 증가하고 있다. 정보 검색(information retrieval)이란 수집된 정보 또는 자료의 내용을 분석한 뒤 적절히 가공하여 축적해 놓은 데이터베이스로부터 사용자의 정보 요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 의미한다. 일반적으로 정보검색 시스템은 사용자의 질의에 대해 정보가 포함되어 있을 가능성이 높은 문서들을 찾아 주는 것이다. 이러한 정보 검색 시스템에서는 사용자들이 다시 한번 검색된 문서에서 정답을 찾아야 하는 불편이 있다. 반면, 질의-응답 시스템(Question Answering System)은 방대한 문서 집합에서 사용자의 요구에 대한 정확한 답을 찾는 시스템이다. 즉, 질의에 대한 결과로 문서를 제시해 주는 것이 아니라 사용자가 원하는 질문에 대해 정답을 제시해 주기 때문에 사용자의 부담을 줄여 주는 시스템이다. 이로 인해 질의 응답은 자연언어처리 분야에서 매우 관심이 집중되는 연구주제로 부각되었다. 질의-응답 시스템은 관계형 데이터베이스 시스템이나 계층형 데이터베이스 시스템에 대한 자연언어 인터페이스를 제공해 주기 때문에 자연언어로 된 질의어를 처리할 수 있다[1]. 자연언어로 된 질의어를 처리할 수 있는 질의-응답 시스템에서 중요한 것은 사용자의 질문을 이해하고 해석하는 것이다. 사용자가 찾고자 하는 것이 인물에 대한 것인지, 지명에 관한 것인지, 시간에 관련된 것인지, 조직에 관련된 것인지, 아니면 그 밖에 다른 것인지를 파악하는 것이다. 이것을 질의 처리 과정 중에서 질의 유형 분류 단계라고 한다. 이 단계는 사용자의 의도를 정확히 파악하고 정답을 구하는 것이 무엇인지 파악하는 단계이다. 질의 유형 분류 과정이 정확하게 이루어진다면 질의-응답 시스템의 성능을 향상시킬 수 있을 것이다. 본 논문에서는 질의-응답시스템의 성능향상을 위해 사용자의 질의 의도를 파악하기 위한 방법을 제안한다. 사용자의 질의 의도를 파악하기 위해서 사용자 질의 문장에서 접미사정보, 어휘정보와 명사의미정보를 이용하여 질의 유형과 정답유형을 파악하여 대량의 코퍼스를 이용하거나 복잡한 규칙을 작성하지 않고 질의문 분석을 하였다. 한국어 의문문에서 자주 발생하는 어휘들의 출현 방법을 고려하여 규칙을 작성하고 규칙에 적용되지 않는 경우는 접미사 정보, 명사 의미 정보 사전을 구축하고 동의어 사전을 이용하여 질의 유형 분류

와 정답 타입을 결정하였다.

2. 관련 연구

질의 응답 시스템이 인터넷과 같은 실용적 환경에서 사용될 경우, 실제 사용자의 질의는 다양한 유형으로 나타나게 된다. 따라서 다양한 유형의 질의에서 사용자가 의도하는 질의의 방향을 결정할 수 있다면 질의 응답 시스템의 성능 향상에 도움이 될 것이다.

기존의 연구에서 질의 유형 분류 방법은 통계 모델에 기반한 학습을 이용하여 인식하는 통계에 기반한 방법과 수동으로 패턴을 구축하고 이를 이용하는 규칙에 기반한 방법으로 나눌 수 있다.

통계적 방법에 기반한 질의 유형 분류는 수동으로 분류된 대량의 학습 데이터를 이용하여 추출한 통계정보를 사용한다. 통계정보는 학습 코퍼스의 유무에 따라 교사학습기반[2,3,4]와 비교사 학습기반[5]으로 나눌 수 있으며, 학습 방법에 따라 Hidden Markov Model에 기반한 방법[2, 4], 결정 트리 기반[3], 최대 엔트로피 모델에 기반한 방법[5] 등이 있다.

이러한 통계적 기반의 방법은 대량의 학습 데이터를 이용한 통계 모델을 사용하기 때문에 안정적으로 질의의 유형을 분류할 수 있으며, 응용 영역의 변화에 대해 크게 영향을 받지 않는다. 그 반면, 사용자의 의도와는 상관없는 결과를 제공할 가능성이 있으며, 대량의 학습 데이터를 이용하여 추출해야 하는 어려움이 있다.

규칙에 기반한 방법은 정교한 규칙을 수동으로 작성하거나, 수동으로 작성된 규칙을 학습 코퍼스를 이용하여 수정하는 방법이다. 규칙 기반의 방법은 적용할 분야의 특성을 활용한 방법[6], 문장에 자주 발생하는 문맥을 활용한 방법[7], 접사 사전과 결합 규칙을 이용한 방법[8], 규칙과 문맥을 다단계로 적용한 방법[9] 등이 있다.

규칙에 기반한 방법은 질의 유형 분류 과정이 유한 상태 오토마타로 구현되므로 사용자의 질의에 대해서 즉각적으로 질의 유형을 분류해 낼 수 있고 수동으로 기술된 규칙에 따라 질의 유형을 분류하므로 질의 유형을 잘못 분류해서 전혀 관계없는 엉뚱한 대답을 하는 경우를 방지할 수 있다. 또한 응용 영역이 정해져 있을 경우 간단한 튜닝으로 성능을 향상시킬 수 있다. 그러나 응용 영역이 확대, 변경 되었을 때 규칙을 수정하기 위해서는 전문적인 지식을 가진 사람들의 노력이 필요하며, 규칙과 일치되지 않는 질의가 들어올 경우에는 분류가 되지 않는 어려움이 있다. 이러한 경우는 수동으로 규칙을 변경시켜 수정할 수 있다. 그리고 규칙이 많아질수록 좋은 성능을 내기 위한 튜닝이 점점 더 어려워지게 된다. 또한 시스템이 다른 응용 영역에서 사용될 경우에는 기존의 규칙들을 모두 수정하거나 재작성해야 하는 문제점이 있다.

Support Vector Machine을 이용한 질의 유형 분류[10]은 형태소 분석기를 이용해서 질의에서 형태소를 추출하고 품사를 결정한다. 개체명 인식기를 사용하여 각 형태소에 의미 표지를 부여하여 분류 알고리즘인 SVM을 이용해서 입력된 질의의 범주를 구분한다. 사용자 질의문 전체를 이용하기 때문에 실제 질의 의도와 관련 없는 수식어 등이 많이 포함된 질의문의 경우는 분류 성능이 저하되는 원인이 된다. 질의문의 어휘 정보만을 이용한 질의 분석[11]은 질의문에 나타나는 특정 어휘 정보를 이용하여 분류를 하기 때문에 어휘가 생략된 경우는 분류를 제대로 하지 못하는 문제점이 있다.

3. 질의 유형 분류

질의문에 대한 질의 유형을 분석하고 결정하는 데는 의문사가 중요한 역할을 한다. 그러나 대부분의 한국어 문장에서는 생략이 빈번하게 이루어지므로 의문사 정보만을 가지고 질의 유형을 분석하는 데는 한계가 있다. 이러한 경우는 질의문에 나타나는 어휘들의 의미와 이들의 출현 규칙을 살피고 추출한 어휘들의 정보를 담고 있는 사전을 이용하여 질의 유형을 분류할 수 있다. 또한, 질의 유형에 대한 자세한 분류를 하여 정답후보들의 적합성을 판단하고, 정답을 추출하는데 중요한 역할을 한다.

질의 유형 분류는 자연언어로 된 질문에 대하여 7개의 질의 유형과 각각의 질의 유형에 대해 세분화된 70개의 하위 의미 범주로 분류한다. 사람에 관해 나타내는

(HUMAN), 어떠한 장소를 나타내는 (LOCATION), 날짜나 시간은 (TIME), 수를 나타내는 (NUMBER), 조직을 나타내는 (ORGANIZATION), 물체나 사물 등의 개체를 나타내는 (OBJECT), 기타 분류되지 않는 (UNKNOWN)으로 나눌 수 있다. 각각의 질의 유형에 대해 세분화 된 하위 의미 범주는 정답 유형을 결정하고 정답 후보를 결정하는데 유용하게 사용될 수 있다.

3.1 어휘 정보를 이용한 질의 유형 결정

질의 유형을 분석하기 위해서는 질의의 초점이 무엇인지를 파악하는 것이 중요하다. 한국어 의문문은 대부분의 경우 문장의 마지막에 의문의 초점을 나타내는 중요한 정보를 가지고 있다. 각각의 질의 유형을 의문의 초점을 나타내는 어휘가 존재한다면 어휘 정보만 이용해서 질의 유형을 분류할 수 있다.

(사람)에 관한 질의 유형의 경우 '누구인가', '누구입니까', '사람은 누구입니까' 등의 의문의 초점을 나타내는 어휘들이 문장의 마지막에 존재하게 된다. 그러나 '누가', '누구의' 등의 의문의 초점을 나타내는 어휘들이 나타나면 의문의 초점이 마지막에 나타나지는 않는다.

예를 들어, "철의 여인 마가렛 대처란 책의 저자는 누구입니까?" 라는 질의문의 경우 '누구입니까'의 어휘 정보에 의해 (사람)에 관한 질의문임을 알 수 있다. 또한, "1991년 노벨 평화상을 누가 수상했습니까?"라는 질의문의 경우 "누가"라는 어휘 정보에 의해 (사람)에 관한 질의문임을 알 수 있다. (장소)에 관한 질의 유형의 경우, '어디인가', '어디입니까' 등의 의문의 초점을 나타내는 어휘들이 존재한다. 의문의 초점을 나타내는 어휘가 마지막에 나타나지 않는 경우 예를 들어, "마케도니아는 어느 나라로부터 독립을 하여 자유를 얻었는가?"라는 질의문의 경우 '어느'라는 어휘 정보에 의해 (장소)에 관한 질의문임을 알 수 있다. 이와 같이 의문의 초점을 나타내는 어휘가 존재하는 경우 어휘의 종류에 따라 질의 유형을 판단 할 수 있다. 즉, '누구인가', '어디인가'라는 실마리 어휘에 의해 질의문이 어떠한 범주에 속해야 하는 지를 판단할 수 있다. 질의 유형 중에 (조직)과 (장소)의 경우는 의문의 초점을 나타내는 어휘가 '어디'로 중복되는데, 이러한 경우는 '어디' 앞에 나타나는 어절의 단어를 살펴서 질의 유형을 판단할 수 있다. 예를 들어, "국내 시장 점유율이 가장 높은 일본의 자동차 생산 회사는 어디인가?"와 "미로의 비너스상이 있는 곳은 어디인가?"라는 질의문의 경우 첫 번째 질의문은 '자동차 생산 회사'라는 (조직) 묻는 질의문이고 두 번째 질의문은 '비너스상이 있는 곳'이라는 (장소)를 묻는 질의문이다. 이러한 경우 '회사'라는 명사를 분석하고 '곳'이라는 명사를 분석하여 두 개의 실마리 단어가 중복되는 질의문을 분류할 수 있다. 또한 한국어는 생략이 빈번하게 발생하기 때문에 의문의 초점을 나타내는 어휘가 존재하지 않는 경우가 있다. 이러한 경우는 질의문의 마지막 어절에 위치하는 명사를 분석하여 질의 유형을 판단할 수 있다. 예를 들어, "1988년 가장 높은 살인비율을 가졌던 미국 도시는?"이라는 질의문은 '어디인가'라는 어절이 생략되었다. 이러한 질의문의 분석은 마지막 어절의 '도시'라는 명사의 의미를 담고 있는 사전을 이용하여 '도시'가 (장소)의 (CITY)라는 것을 알 수 있다. 만약, "목성탐사 '갈릴레오' 우주선을 전송한 기관은?"이라는 질의문의 경우 '기관'은 (조직)과 (신체기관)이라는 두 가지의 의미 정보를 가진다. 이러한 경우는 중의성을 해결하기 위해 {'혈관' '호르몬', '호흡', '순환', ...} 등의 어휘가 질의문에 나타난 경우에만 (신체기관)으로 분류하도록 한다.

3.2 명사 의미 정보를 이용한 질의 유형 결정

앞 절에서 예를 들어 살펴본 "알마아타가 수도인 나라는 어디인가?'"라는 질의문은 '어디인가'라는 어휘에 의해 [장소]에 관한 질의문으로 분류할 수 있다. (장소)라는 범주 중에서 (COUNTRY)라는 하위 의미 범주로 분류를 위해 '나라'라는 명사의 의미가 (LOC)(COUNTRY)라는 정보를 가지고 있는 사전을 이용한다. 이러한 의미 정보 사전은 수동으로 구축하였다. 아래의 표 2는 질의 유형에 대한 하위 의미 범주를 나타낸다. 각각의 하위 의미 범주는 TREC-10, TREC-11과 NTCIR에 참가한 질의 응답 시스템들을 참고 하였다. 하위 범주까지의 분류는 정답의 유형을 찾고 정답 후보를 생성하는데 이용될 수 있다.

[표 2] 질의 유형에 대한 하위 의미 범주

질의 유형	하위 의미 범주
HUMAN	ARTIST, POLITICIAN, ECONOMIC, SPORTS, ...
LOCATION	PLACE, PLANET, CONTINENT STATE, CAPITAL, ...
TIME	YEAR, MONTH, DAY SEASON, PERIOD, ...
NUMBER	COUNT, PRICE, PERCENT, WEIGHT, HEIGHT, ...
ORGANIZATION	SCHOOL, COMPANY GOVERNMENT, GENERAL, ...
OBJECT	PLANET ,WAR, RELIGION REASON, ORGAN, ...

3.3 질의 유형 분석의 성능 향상

앞 절에서 설명한 것과 같이, 질의 유형에 대해 하위 의미 범주로 분류하기 위해 명사 의미 정보 사전을 이용하였다. 이러한 명사의 의미정보를 모두 사전을 구축하는 것은 불가능하다. 질의문에서 다양한 형태의 출현 가능한 명사들을 분류하기 위해서 대용량의 사전을 구축 하지 않고 유의어를 이용하는 방법과 접미사 정보를 이용하는 방법을 이용한다.

3.3.1 유의어 사전 이용

만약, 추출된 명사가 의미 정보 사전에 포함되지 않은 경우 표 3과 같이 약 10만개의 고빈도 명사들로 구성된 유의어 사전을 이용하여 유의어를 명사 의미 정보 사전에서 찾아 분류한다.

[표 3] 유의어 사전 이용

추출된 명사	유의어
작가	글쓴이, 소설가, 문필가, 집필자, 문예가, 저작자, 제작자, 대문호, 저자, 지은이
장소	곳, 데, 처소, 지점, 부분, 점, 위치, 지역
나라	국가, 사직, 네이션

예를 들어, "텅스텐의 가장 큰 생산 국가는 어디인가?"라는 질의문에서 '국가'라는 명사가 질의 유형의 하위 의미 범주로 분류하기 위한 명사로 추출된 경우이므로 '국가'의 유의어 사전을 이용하여 '나라'라는 유의어로 변경되어 분류될 수 있다.

3.3.2 접미사 사전 이용

접미사는 접사의 하나로 낱말의 끝에 붙어 의미를 첨가하여 다른 낱말을 이르는 말이다. 단독으로는 사용할 수 없고, 항상 다른 단어의 어근 뒤에 결합되어, 여러 가지 의미를 첨가해 주는 역할을 한다.

질의문이 여러 가지 종류의 접미사가 붙어 다양한 형태로 나타나기 때문에 접미사 처리를 하지 않는다면 시스템의 성능을 저하시키는 원인이 된다. 예를 들어, "개최국은 어디인가?", "생산국은 어디인가?"와 같은 질의문의 경우 "개최국"가 "생산국"이라는 명사를 이용하여 의미 범주를 분류하려면 명사 의미 정보 사전에 "개최국"과 생산국"이라는 것의 의미 범주가 존재해야만 한다. 그러나 이러한 모든 명사를 사전으로 구축하는 것은 불가능하다.

"국가"를 나타내는 접미사 '국'이 여러 낱말의 끝에 붙어서 (COUNTRY)라는 의미를 가진다. 따라서 접미사에 대해 의미 범주를 미리 정의한다면 명사 의미 사전의 구축이나 어휘들의 정보를 이용하지 않고도 질의문를 분류할 수 있다. 그러나 중의적인 뜻을 가진 접미사가 사용된다면 질의 유형을 올바르게 분류해낼 수 없다. 예를 들어, '가'라는 접미사의 경우는 "소설가", "은행가"에 같이 사람의 뜻으로도 사용되고 노래 이름이나 노래 종류를 나타내는 "애국가", "응원가"로도 사용되며 특색을 낀 거리라는 의미로 "상점가", "환락가"라는 뜻으로도 사용이 된다. 이러한 경우는 단순한 접미사 정보만을 가지고 질의 유형을 분류할 수 없다.

4. 실험 및 평가

시스템의 성능 평가를 위해 질의응답 시스템의 성능 평가를 위한 테스트컬렉션[12]의 90개 질의문과 TREC의 성능 평가를 위한 질의문 중에서 500개를 임의로 선택하여 사용하였다. 각각의 질의문에 대해 수동으로 질의유형을 분류하고 제안된 시스템의 결과와 비교한다. 실험 결과, "최근 한국에서 구매된 잠수함의 형태는?"이라는 질의문의 경우, '형태'라는 명사와 '형태'의 동

의어들이 의미 정보사전에 존재하지 않기 때문에 분류되지 않았다. 또한 "보험 산업에서 피해가 가장 큰 항목은 무엇입니까?"라는 질의문의 경우 '무엇입니까'라는 어휘 정보에 의해 (OBJECT)와 (ORGANIZATION)의 두 가지 결과를 가지게 된다. 둘 사이의 중복결과를 피하기 위해 앞 어절의 명사를 살펴본 결과 '항목'이라는 어휘에 대한 분석에 실패하여 분류되지 않았다.

5. 결론 및 향후 연구

본 논문에서는 고성능의 질의 응답 시스템을 만들기 위한 필수 조건인 사용자의 질의문을 분석하는 방법에 대해 제안하였다. 의문의 초점을 나타내는 어휘가 질의문에 존재하는 경우라면 쉽게 사용자의 질의 의도를 파악할 수 있지만 한국어의 생략이 많이 발생하는 특성상 의문의 초점을 나타내는 어휘가 생략되는 경우가 빈번하게 발생한다. 이러한 경우에도 처리가 가능하도록 하여 질의 유형 분류에서 만족할 만한 성능을 보였다. 먼저 사용자 질의문에서 의문의 초점을 판단할 수 있는 어휘의 존재 여부를 파악하고 추출된 어휘로 질의 유형을 분류한다. 또한 의문의 초점을 나타내는 어휘 주변의 명사를 추출하여 명사 의미 정보 사전을 이용하여 질의문을 세부 단계까지 분류하여 질의 응답 시스템에서 정답후보 생성 시 보다 효과적으로 사용할 수 있도록 하였다. 만약 사용자 질의문에서 의문의 초점을 나타내는 어휘가 생략된 경우라면 질의문의 마지막 어절의 명사를 추출하여 명사 의미 정보 사전을 이용하여 질의 유형을 해당 범주를 분류해 낼 수 있다. 질의 유형 분석의 성능을 향상시키기 위하여 동의어나 유의어 정보를 이용하고 접미사 정보를 이용하였다. 실험결과 본 시스템은 97.4%의 높은 정확률을 보였다. 이것은 제안한 시스템이 복잡한 구문 규칙이나 대용량의 사전 정보, 통계 정보 등을 이용하지 않고도 충분히 만족할 만한 질의 유형 분류를 할 수 있다는 것을 나타낸다.

앞으로 행해질 향후 연구는 질의 유형의 하위 의미 분류의 범주를 보다 다양하고 폭넓게 적용하여 사용자 질의문에서 정답 유형을 좀 더 구체적으로 제시할 수 있도록 하는 연구이다. 또한 여러 가지 중복된 의미를 가지는 접미사를 보다 효과적으로 이용하는 방안에 대한 연구이다.

▌참고문헌▐

[1] 김영택, "자연언어처리", 생능출판사, 2001.

[2] Bikel, D. M., Miller, S., Schwartz, R. and Weischedel, R., "Nymble : A High-Performance Learning Named-finder", In-Proceeding of the Fifth Conference on Applied Natural Language Processing, pp.194-201, 1997.

[3] Sassano, M. and Utsuro, T., "Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition", In-Proceedings of the 18th International Conference on Computational Linguistics, pp.705-711, 2000.

[4] Yu, S., Bai, S. and Wu, P., "Description of the Kent Ridge Digital Labs System Used for MUC-7", In-Proceedings of 7th Message Understanding Conference, 1998.

[5] Collins, M. and Singer, Y., "Unsupervised Models for Named Entity Classification", EMNLP/VLC-99, pp.189-196, 1999.

[6] Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T., "Toward Information Extraction : Identifying protein names from biological papers", In-Proceedings of the Pacific Symposium on Biocomputing '98(PSB '98), 1998.

[7] 노태길, 이상조, "규칙기반의 기계 학습을 통한 고유 명사의 추출과 분류", 한국정보과학회 가을 학술발표논문집, Vol 27, No2, pp.170-172. 2000.

[8] Fukumoto, J., Shimohata, M., Masui, F. and Saski, M., "Description of the Oki System as Used for Met-2", In-Proceedings of 7th Message Understanding Conference, 1998.

[9] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", 제12회 한글 및 한국어 정보처리 학술대회, pp.292-299, 2000.

[10] 안영훈, 김학수, 서정연, "지지 벡터 기계를 이용한 질의 유형 분류기", 제14회 한글 및 한국어 정보처리 학술대회, pp.129-136, 2002.

[11] 이경순, 김재호, 최기선, "한국어 질의응답 시스템에세 개체인식에 기반한 대답 추출", 제12회 한글 및 한국어 정보처리 학술대회, pp184-189, 2000.

[12] 이경순, 김동완, 최기선, "질의 응답 시스템의 평가를 위한 테스트컬랙션 구축", 제12회 한글 및 한국어 정보처리 학술 대회, pp.190-197, 2000.