KIPS Tr. Software and Data Eng.
Vol.4, No.12 pp.577~584 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2015.4.12.577

A Topic Classification System Based on Clue Expressions for Person-Related Questions and Passages

Gyoung Ho Lee† · Kong Joo Lee††

ABSTRACT

In general, Q&A system retrieves passages by matching terms of a question in order to find an answer to the question. However it is difficult for Q&A system to find a correct answer because too many passages are retrieved and matching using terms is not enough to rank them according to their relevancy to a question. To alleviate this problem, we introduce a topic for a sentence, and adopt it for ranking in Q&A system. We define a set of person-related topic class and a clue expression which can indicate a topic of a sentence. A topic classification system proposed in this paper can determine a target topic for an input sentence by using clue expressions, which are manually collected from a corpus. We explain an architecture of the topic classification system and evaluate the performance of the components of this system.

Keywords : Topic Classification, Clue Expression, Person-Related Topic Class

단서표현 기반의 인물관련 질의-응답문 문장 주제 분류 시스템

이 경 호† · 이 공 주††

요 약

일반적으로 질의응답 시스템은 입력된 질문에 대한 정답을 찾기 위해 질문과 관련된 문서 또는 단락 단위의 검색을 수행한다. 그렇지만 단어 기반의 검색만으로는 정답을 포함하는 단락을 찾기 어려운 경우가 있다. 본 논문에서는 이러한 문제를 각 문장이 가지고 있는 주제를 통해 해결할 수 있다고 판단하고 이를 위한 질의-응답문의 주제 분류 시스템에 대해 연구하였다. 이러한 시스템을 위해 필요한 인물과 관련한 주제 유형을 소개하고, 주제를 찾기 위한 단서표현을 정의하였다. 또한 단서표현기반으로 문장의 주제를 파악하는 시스템의 구성에 대해 소개하고, 이 시스템의 구성요소들에 대한 성능 평가를 수행하였다.

키워드 : 주제 분류, 단서표현, 인물관련 주제

1. 서 론

일반적으로 질의응답 시스템은 입력된 질문에 대한 정답을 찾기 위해 질문과 관련된 문서 또는 단락 단위의 검색을 수행한다. 검색을 통해 얻은 후보 단락들은 각 단락과 질의와의 관계를 분석하여 그 우선순위가 결정된다. 이러한 검색에서 단어 기반의 검색만으로는 정답이 포함된 단락을 찾기 어려운 경우가 있다. 아래 Table 1은 질의응답 시스템을 위한 문서 검색 시스템[1]의 검색 결과 예시이다. 질의와 응답 1, 2 모두 '허난설헌'과 '허균'이라는 인물과 관련된 단락들이다. 응답 1은 그들과 가족 관계인 '홍우정'에 대해 서술하고 있고, 응답 2는 사제 관계인 '이달'에 대해 서술하고 있다. 이 예제의 질의에 대한 정답과 관련된 단락은 응답 2이다. 하지만 검색기는 질의와 공통된 단어 수가 더 많은 단락 1이 단락 2보다 더 높은 연관성을 가지는 것으로 응답하였다.

이러한 경우 응답 1에서 잘못된 정답을 찾거나 응답 1을 분석한 후 응답 2를 분석하기 때문에 정답에 대한 반응이 늦어질 수 있다.

예제에 나타난 질의와 응답들을 글의 주제 측면에서 살펴보면, Table 1의 질의문은 '제자로 받아들여'와 같은 표현을 통해 이 단락이 스승-제자 관계를 주제로 작성한 단락임을 알 수 있다. 정답 문장인 응답 2에서도 '스승'이라는 표현을 통해 이 단락이 스승-제자 관계를 표현하고 있다. 반면, 응답 1은 '아버지', '외할아버지', '손자'와 같은 어휘들을 통해 가족관계를 주제로 이야기하고 있는 것을 알 수 있다. 이와 같이 각 단락들이 글을 통해 말하고자 하는 주제를 파악할 수 있다면, 이를 통해 질의와 응답 단락이 글을 통해 의도

※ 이 논문은 2014년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No: 1101-2014-0061, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발).
† 준 회 원 : 충남대학교 전자전파정보통신공학과 박사과정
†† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : July 28, 2015
First Revision : August 26, 2015
Second Revision : October 6, 2015
Accepted : October 16, 2015
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

하고자 하는 바를 찾아낼 수 있다. 또한 각 단락의 주제를 검색의 색인어로 활용하여 정답 문장을 찾는 검색의 성능을 높이는 데 활용할 수 있다.

Table 1. Example of Search Results of Q&A System

질의 이 사람은 서자라는 신분 때문에 벼슬길이 막히자 시를 지으며 자신의 처지를 달랬는데, 말년에 허균과 허난설헌을 제자로 받아들여 허균의 사상에 큰 영향을 미쳤다. 이 사람은 소동파 등의 시풍을 따랐던 대중적인 흐름과 달리 당나라 시에 심취하여 조선 선조 때의 '3당 시인' 중 하나로 불리는데, '홍길동전의 모델'로 추측되는 이 사람은 누구일까?
응답 1 아버지는 한성부 서윤을 역임하고 사후 이조판서에 추증된 홍영이며, 어머니는 이조판서를 지낸 허성의 딸이다. 외할아버지 허성은 허난설헌의 오빠이며, 홍길동전을 지은 허균의 형이다. 할아버지 홍가신이 하루는 꿈을 꾸는데 길몽을 꾸고, 손자 중 탁월한 문장가가 날 것이라는 예견을 했는데 이 이후 홍우정이 태어났다.
응답 2 호 손곡(蓀谷)은 강원도 원주 손곡리에 정착하면서 만들었다. 허균과 허난설헌의 스승이다. 오늘날 홍주 군청 앞에 시비가 있다.

본 논문에서는 이와 같은 점에 착안하여, 입력된 문장에 대한 주제 분류 시스템을 연구하였다. 구체적으로, 인물과 관련된 질의와 응답으로 구성된 문장들에서 각 문장이 가질 수 있는 주제들을 찾는 시스템에 대하여 연구하였다. 이를 위해 '제자로 받아들여'나 '스승'과 같이 글에 나타난 주제를 대표할 수 있는 표현을 '단서표현'으로 정의하고, 이를 기반으로 문장의 주제 인식 방법을 연구하였다. 또한 문장이 가질 수 있는 인물과 관련된 주제 목록과, 이러한 표현을 수집하기 위해 개발된 도구에 대하여 설명한다. 개발된 문장 주제 분류 시스템은 질의응답 시스템의 정답 후보 문장 검색 결과 정렬에 활용할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 단락단위 문장의 주제인식과 관련된 최근의 연구 동향에 대해 살펴본다. 3장에서는 본 논문에서 제안하는 시스템 개발을 위해 필요한 전반적인 구성요소에 대해 설명하고 이들 각각에 대해 설명하였다. 마지막으로 4장에서 본 논문의 결론에 대해 설명한다.

2. 관련 연구

주제 분류(text classification, text categorization)는 미리 정의되어있는 주제를 특정 문서에 할당하는 것으로 자연언어처리 분야의 오랜 관심사였다. 주제 분류는 질의응답 시스템(QA System)에서 질의의 주제를 분류하는 연구[2], E-mail spam 필터링[3], 뉴스 기사 주제 분류[4], 웹문서 주제 분류와 검색[5, 6] 등 다양한 응용 분야에서 사용되어왔다. 주제 분류를 분류 기법 측면에서 보면 단순히 문서를 이루는 단어를 기반으로 하는 방법[7]과 온톨로지 기반의 분류[8], LDA와 같은 Topic Modeling 기법과 기계학습 알고리즘을 이용하는 방법[9, 10] 등 다양한 접근 방법이 연구되고 있다.

본 논문은 짧은 문서(short text)에 대한 주제 분류를 다루고 있다. 이와 관련된 연구로는 [11, 12]가 있다. [11]의 연구는 검색 snippets, 채팅 메시지, 상품 리뷰 등과 같은 짧은 글들에 대한 분류가 어려운 이유 중 하나로 희소성(sparseness)을 꼽았다. 이러한 희소성 문제를 완화하기 위해 도메인과 장르에 구애받지 않는 광범위한 "Universal dataset"을 수집하고 이를 문서 분류에 활용하였다. 이 연구에서는 LDA알고리즘을 Universal dataset에 적용하여 토픽 모델을 생성하였다. 문서분류 모델학습을 위한 학습 문서들에 이 토픽 모델을 적용하여 학습 문서들을 hidden topic으로 표현하고, 이를 자질로 하여 분류 모델을 학습시킨다. 주제 분류가 필요한 새로운 문서가 입력되면, 앞서 사용한 토픽 모델을 적용하여 문서를 표현하고, 문서 분류 모델에 적용하여 새 문서의 주제를 분류한다. 이와 같은 모델을 통해 웹 문서의 도메인 분류 문제에서 16.27%의 분류 오류(Classification error)를 나타내었다. [12]의 연구에서는 문서를 구성하는 단어들을 자질로 사용하면서, chi-square 자질 선택 알고리즘을 이용하여 자질을 선택하였다. 이 연구에서 웹 문서를 6개 주제로 분류하는 문제에 대해 Naive Bayesian 알고리즘을 적용한 결과 65.6%의 precision과 66.8%의 recall을 나타내었다. 성능 향상을 위하여, 각 문서 주제에 특화된 자질을 자동으로 추출하는 알고리즘을 개발하고, 이를 적용하였을 때, 특정 주제에 대해 최고 89.3%의 precision과 65.8%의 recall을 나타내었다. 본 논문은 이러한 연구들이 다루고 있는 짧은 문서의 특징을 "단서표현"을 통해 해소하고자 하였다.

본 논문의 시스템의 응용 목적은 질의응답 시스템의 정답 문서 검색 성능 향상이다. 이러한 분야의 선행 연구로는 [13]이 있다. [13]의 연구에서는 질의문과 응답문 사이의 관계정보(relation)와 관계정보의 주제(topic) 추출을 통해 이들의 구문적, 의미적 유사성 비교를 시도하였다. 이 연구에서 문장으로부터의 관계추출을 위해 2가지 방법을 시도하였다[14]. 첫 번째는 구문 분석을 통해 입력문장의 PAS(Predicate Argument Structure)를 추출하고, 이로부터 수동으로 작성된 규칙을 통해 관계정보를 추출하는 규칙 기반(rule based approach)방법이다. 두 번째는 위키피디아 문서의 인포박스에 기록된 프로퍼티(property)정보와 해당 프로퍼티를 설명하는 문장을 추출하고, 이를 학습데이터로 하여 기계학습 알고리즘을 사용한 관계분류기를 사용하는 통계적 방법(statistical approach)이다. 이러한 2가지 방법을 이용하였을 때, 규칙 기반 방법에서는 30개의 관계 정보를, 통계적 기법에서는 7000개 이상의 관계정보를 설정하고 이에 대한 추출 정확도의 F1 score는 각각 0.553과 0.342를 기록하였다. 본 논문에서는 이 연구와 유사하게 질의/응답 문장의 관계를 나타내기 위해 주제(topic)를 추출한다. 또한 이를 추출하기 위한 방법으로 통계적 기법을 활용한다는 점에서 유사점이 있다. 반면 [13]의 연구에서 통계적 기법을 사용할 때 관계를 자동으로 추출하여 매우 많은 수의 관계를 정의했지만,

Table 2. A Topic Class Used in This Paper

기본		인물					
				객체		활동	
명칭 시기 위치 원인 상세 기호 정의 별칭	약어 배경 원이름 기능 최초 순위 구성 상징 평가	국적 직업/직책 가족 관계 출생 출생-시기 출생-위치 사망 사망-시기	사망-원인 사망-위치 대표 인물 관련 인물 활동 인물 인물-평가	작품 작품-대표작 작품-관련 인물 작품-활동 인물 작품-상세 작품-시기 단체	단체-대표 인물 단체-활동 인물 상 상-활동 인물 상-시기 상-작품 이론 이론-활동 인물 사건 사건-시기 질병	발견 발견-활동 인물 발견-시기 발견-위치 제작 제작-활동 인물 제작-시기 제작-상세	설립 설립-활동 인물 설립-시기 발표 발표-활동 인물 발표-시기 결성 결성-시기 명명 명명-활동 인물

본 논문에서는 인물과 관련된 질의응답 시스템에서 필요로 하는 주제를 미리 정의하고 이에 대한 분류시스템을 연구하였다. 또한 문장의 처리 속도 및 주제 분류의 적용 단계 등을 고려하여 문장의 형태소 분석 결과와 표층정보만을 이용하여 주제를 분류하는 것에서 기존의 연구와 차별점이 있다.

3. 본 론

3.1 질의-응답문 유형

본 논문에서는 주제 분류에 필요한 단서표현 추출과 주제 분류 시스템 구축 및 평가를 위한 문서집합으로 기존의 텔레비전 퀴즈쇼에서 출제된 질문들로 구성된 질의문과 이에 대한 [1]의 응답 결과로 구성된 문서집합을 사용하였다. 이 문서집합의 특징은 다음과 같다.

1) 1개의 질의문과 15개의 응답 문장으로 구성
2) 질의문은 1개 이상의 문장으로 구성
3) 응답문들은 3개 이하의 문장으로 구성

3.2 주제 분류체계

주제 분류 시스템 개발을 위해, 먼저 주제 분류체계를 정의하였다. 주제 분류체계는 수집된 질의-응답 문서집합에서 질의에 대한 정답 인물과 관련된 문장을 이용하여 생성하였다. 문장들을 [15]의 언어분석 결과를 통해 분석하고 정답 또는 정답과 밀접한 관계가 있는 대상의 개체명 유형(Named entity type)을 기반으로 주제 분류체계 목록을 정의하였다. 주제 분류체계는 Table 2와 같다.

주제 분류체계는 크게 4개의 범주를 가지고 있다.

1) 기본정보 범주: 시간, 위치, 정의와 같이 질의-응답에서 주제로 사용할 수 있는 일반적인 주제 태그를 포함하고 있다.

2) 인물정보 범주: 국적, 출생, 사망과 같이 질의-응답에서 많이 나오는 인물과 관련된 주제태그를 포함하고 있다. 이와 함께, 출생-시기, 사망-원인과 같이 인물과 관련된 주제태그에 기본 주제태그를 더하여 그 주제를 더 명확히 할 수 있도록 확장된 태그를 포함하고 있다.

3) 객체정보 범주: 작품, 단체, 상과 같이 주로 인물과 관련하여 질의-응답에서 나올 수 있는 객체를 표현하는 주제태그를 담고 있다. 인물 정보와 마찬가지로 객체정보 태그도 상-시기, 사건-시기와 같이 객체범주의 태그와 기본범주의 태그를 결합하여 그 의미를 명확히 하고 있다. 또한, 단체-대표인물, 단체-활동인물과 같이 객체범주 태그와 인물범주 태그를 결합하여 명확성과 활용성을 높일 수 있도록 하였다.

4) 활동정보 범주: 발견, 제작, 설립과 같이 주로 인물과 관련하여 질의-응답에서 나올 수 있는 활동을 주제로 한 태그를 담고 있다. 객체정보 범주와 같이 활동범주의 태그와 함께 기본범주의 태그와 인물범주의 태그를 조합하여 함께 활용하였다.

3.3 단서표현

1) 단서표현 및 주제 태깅

본 논문에서는 입력된 문장의 주제를 문장에 포함된 단어나 구절과 같은 표현을 기반으로 찾는다. 이처럼 문장에서 그 문장의 주제를 명확히 나타낼 수 있도록 도와주는 표현을 단서표현(Clue Expression)이라 정의하였다. 그리고 학습 데이터로부터 이러한 표현의 수집을 돕는 태깅 도구를 제작하였다. 태깅 도구는 전문가가 문장에서 나타난 주제를 선정하고 그 주제를 나타내는 표현들을 표시할 수 있도록 설계하였다. 응답문장에서는 질의에 대한 정답과 관련이 있는 문장에 주제를 태깅하였다. 한 문장에 나타난 주제가 여럿일 경우, 전문가의 판단에 따라 문장과 가장 관련이 있다고 판단되는 주제와 핵심단어를 최고 3개까지 태깅하였다. 주제 태깅 도구의 사용 예는 Fig. 1과 같다.

전문가는 이 도구을 통해 입력된 문장에 주제문장과 핵심단어를 표시한다. 전문가가 입력문장을 읽고, 문장에서 주제를 가진 내용에 대해 그 범위와 주제를 태깅한다(__B:주제__ ~ __E__주제__). 그리고 그 범위 내에서 주제를 나타내는 핵심 단어를 << >>를 이용하여 표시하여준다. 태깅 방법의 예는 Table 3과 같다.

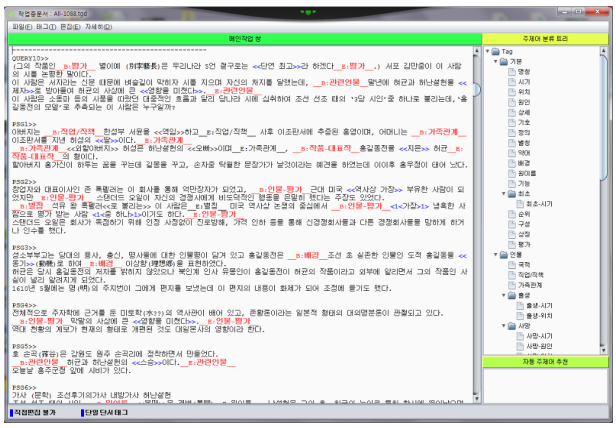

Fig. 1. Example of a Topic Tagging Tool

Table 3. Example of Tagging Results

Original	아버지는 한성부 서윤을 역임하고 사후 이조판서에 추증된 홍영이며, 어머니는 이조판서를 지낸 허성의 딸이다.
Tagging	아버지는 __B:직업/직책__한성부 서윤을 <<역임>>하고__E:직업직책__ 사후 이조판서에 추증된 홍영이며, 어머니는 __B:가족 관계__이조판서를 지낸 허성의 <<딸>>이다.__E:가족 관계__

2) 단서표현

본 논문의 시스템은 단서표현을 이용하여 문장의 주제를 파악한다. 1) 항의 태깅 도구를 통해 주제와 핵심 단어가 태깅된 학습데이터 문장(4.1절에서 설명)으로부터 총 6,218개 단서표현을 추출하였다. 이러한 단서표현과 이에 해당하는 주제의 예는 Table 4와 같다.

단서표현은 문장에서 주제를 나타낼 수 있는 단어나 어구를 형태소 단위의 패턴으로 나타낸 것이다. 입력된 문장의 형태소 분석 결과에서, “역할/NNG을/JKO 하/VV”의 패턴이 발견되면 그 문장이 “기능”과 관련된 주제를 포함하고 있는 것으로 생각할 수 있다. 또한, 문장에 “태생/NNP”라는 표현이 나온다면, 이러한 문장은 인물의 출생지와 관련된 주제를 가지고 있는 것을 알 수 있다.

문장에서 단서표현과 함께 나타나는 구체적인 대상이 주제 결정에 중요한 요소로 사용될 수 있다. Table 4의 예제 문장에서 인물의 출생 위치를 나타내는 “태생”이라는 단어와 함께 국가를 나타내는 단어가 함께 나왔으면 문장의 주제는 국적으로 더 구체화될 수 있다. 그렇기 때문에 본 논문의 시스템에서는 문장에서 함께 사용된 개체명 유형을 주제 분류의 자질로 함께 사용한다.

부사의 위치가 비교적 자유로운 한국어의 특성에 따라 [16], 단서표현의 패턴을 문장에 적용하기 유연하게 할 필요가 있다. 이러한 패턴을 입력문장에서 매칭시키기 위하여 각 형태소들 사이에 다른 형태소가 들어올 수 있는 최대 거리를 2로 설정하여 단서표현의 적용에 유연성을 두었다. 특별한 경우, 패턴을 구성하는 앞, 뒤 형태소 사이의 거리가 2보다 커야 할 필요가 있다. 이러한 경우 “@@/@@”패턴을

Table 4. Examples of Clue Expressions

Clue expression	Topic	Example
역할/NNG 을/JKO 하/VV	기능	조선 시대에 각각 지방 대학 역할을 하였으며…
영감/NNP 을/JKO 얻/VV @@/@@ 작곡/NNG	배경	리스트는 이탈리아 여행 도중 사랑의 장편 서사시를 읽고 영감을 얻어 소나타를 작곡했다
에서/JKB 태어나/VV	출생-위치	그녀는 비엔나에서 태어났으나…
출신/NNG	출생-위치	스탈린은 조지아 출신이었지만
태생/NNP	출생-위치	호로비치는 키에프 태생으로…
태생/NNP	국적	… 작품으로 유명한 루마니아 태생의 소설가이자 시인…

정의하여 2보다 먼 거리의 패턴 매칭이 필요함을 나타내었다. *“리스트는 이탈리아 여행 도중 사랑의 장편 서사시를 읽고 영감을 얻어 ‘이 사람의 소나타’를 작곡했다.”*라는 문장에서 “영감을 얻어 ~ 작곡”이라는 표현을 통해 문장이 어떤 작품에 대한 탄생 배경을 나타냄을 알 수 있다. 이러한 패턴 사이에 작품 이름과 같이 여러 형태소로 구성된 구절이 들어올 수 있다. 단서표현에서 “@@/@@”패턴은 이러한 구를 대체하여 표현한다.

3) 단서표현의 주제활성도

1) 항과 2) 항을 통해 단서표현 목록을 수집한다. 단서표현은 태깅된 학습데이터를 기반으로 생성된다. 본 논문에서는 단서표현이 임의의 주제를 얼마나 활성화시킬 수 있는지를 나타내도록 **주제활성도(topic activeness)**를 정의하였다(여기서 주목해야 하는 부분은 특정 주제가 아닌 임의의 주제여도 상관없다). 주제활성도는 다음과 같이 계산된다.

Te: 단서표현 e가 나타난 문장 중, 주제가 태깅된 문장의 수
Tn: 단서표현 e가 나타난 문장 중, 주제가 태깅되지 않은 문장의 수
주제활성도(e): Te/(Te+Tn)

학습데이터의 문장에 주제를 태깅할 때, 입력문장과 답안과의 관계, 문장의 주제로서의 명확성 등을 고려하여 주제와 단서표현을 태깅하였다. 그렇기 때문에 1) 단서표현이 나타나지만, 답안과 관계가 없는 문장이거나 단서표현들이 문장에서 주요한 역할을 하지 못하여 문장에 주제가 태깅되지 않는 경우, 2) 단서표현이 나타나지만 다른 단서표현이 발생시킨 주제가 더 적합하여 해당 단서표현과 관련된 주제가 태깅되지 않은 경우, 3) 단서표현이 나타나고 해당 단서표현이 발생시킨 주제가 태깅된 경우와 같이 3가지 상황이 가능하다. Te는 2)와 3)과 같은 경우이고 Tn은 1)과 같은 경우이다. 단서표현이 여러 문장에 나타났지만 그중 주제가 태깅되지 않은 문장이 많다면, 해당 단서표현은 주제와의 연

관성이 약한 것을 의미한다. 반대로 단서표현이 나타냈을 때 주제를 가진 문장이 많다면, 해당 단서표현이 주제를 발생시킨 것이 아니라고 하더라도, 그 단서표현이 문장의 주제와 연관이 많다는 것을 의미한다. 이렇게 계산된 주제활성도는 인물관련 주제 포함 판별 모듈에서의 자질과 인물관련 주제 분류 모듈에서 필터링 기준으로 사용된다.

3.4 주제 분류 시스템

전체적인 주제 분류 시스템의 구성은 Fig. 2와 같다.

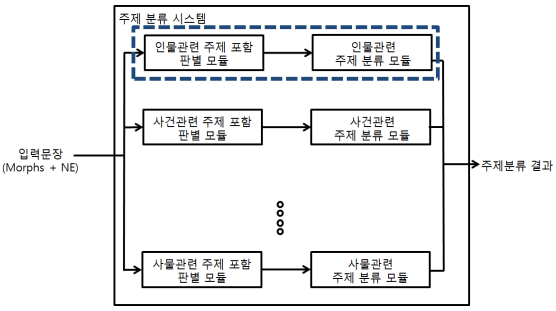

Fig. 2. Overview of the Topic Classification System

이 시스템은 입력문장의 언어 분석 결과(형태소 분석 결과, 개체명 인식 결과)를 입력받고, 해당 문장이 어떤 주제들을 가지고 있는지 찾는다. 전체 시스템은 인물, 사건, 사물 등, 질의와 응답 문장이 가질 수 있는 여러 주제들에 대해 검출할 수 있는 내부 시스템들로 구성된다. 본 논문에서는 이 중 인물과 관련된 주제를 찾는 시스템의 개발에 대해 다루고 있다.

본 논문의 시스템은 크게 인물관련 주제 포함 판별 모듈과 인물관련 주제 분류 모듈로 구성된다. 인물관련 주제 포함 판별 모듈은 입력문장이 인물과 관련된 주제를 포함하고 있는 문장인지 여부를 판단한다. 인물관련 주제 분류 모듈은 인물과 관련된 주제가 있다고 판별된 문장에 대해 앞서 정의한 주제 목록 중, 해당 문장에 포함된 주제를 찾는 역할을 한다. 인물관련 주제 분류 시스템의 구체적인 구성은 Fig. 3과 같다.

1) 인물관련 주제 포함 판별 모듈

인물관련 주제 포함 판별 모듈은 입력문장의 언어 분석 결과를 입력받아 해당 문장이 인물과 관련된 주제를 포함하고 있는지 여부를 결정한다.

이 모듈의 **자질 추출기**는 입력된 문장의 언어 분석 결과와 단서표현 목록을 이용하여 문장에서 출현한 단서표현 목록과 문장에 나타난 개체명 유형을 추출하여 자질 벡터를 생성한다.

인물관련 주제 포함 판별기는 기계학습 알고리즘을 이용하여 입력문장이 인물과 관련된 주제를 담고 있는 문장인지 여부를 판단한다. 태깅 도구를 이용하여 작업한 문장 중, 주제가 태깅된 문장과 그렇지 않은 문장을 주제 유무 판별을 위한 학습데이터로 이용한다.

이들 학습 문장의 자질을 기계학습 알고리즘에 적용하여 학습 모델을 생성한다. 인물관련 주제 유무 판별을 위해 사용하는 주요 자질은 단서표현 포함여부이다. 단서표현은 인물과 관련된 주제를 나타낼 수 있는 표현들을 통해 생성되었다. 그렇기 때문에 이 단서표현이 문장에 나타났다면, 이 문장이 인물과 관련된 주제를 이야기하고 있을 것이다. 이를 기반으로 새로운 입력문장이 들어왔을 때 학습 모델과 입력 문장의 자질을 이용하여 해당 문장의 인물관련 주제 포함 여부를 판별한다.

2) 인물관련 주제 분류 모듈

인물관련 주제 분류 모듈은 앞선 단계에서 입력문장이 인물과 관련된 주제가 있다고 판별하였을 때, 해당 문장이 가지고 있는 주제들을 찾는 모듈이다. 주제가 없다고 판별된 문장에 대해서는 아무런 출력 결과를 내주지 않음으로써 문장이 주제를 담고 있지 않다는 것을 나타낸다. 인물관련 주제 분류 모듈은 자질추출기, 주제 분류기, 주제 순위화의 3개의 모듈로 구성된다.

이 모듈의 **자질추출기**는 앞선 단계의 자질추출기와 같이 입력 문장의 언어 분석 결과와 단서표현 목록을 이용하여 자질을 생성한다. 앞선 단계의 자질추출기에서는 한 문장에서 여러 단서표현이 발견되면 이를 하나의 자질벡터에 각 단서표현의 출연 여부를 나타내었다. 하지만 이 단계의 자질추출기는 한 문장에서 여러 개의 단서표현이 발견되면 각각의 단서표현별로 해당 표현의 출현 여부를 나타내는 자질벡터를 만들고 각 자질 벡터별로 주제를 추출한다. 예를 들어, 전체 단서표현의 목록이 {ㄱ,ㄴ,ㄷ,ㄹ}이고 문장에서 ㄱ, ㄷ 두 개의 단서표현이 발견되었다면, 앞선 단계에서는 [1,0,1,0] 형태의 하나의 자질 벡터를 만들지만, 이 단계의 자질추출기는 [1,0,0,0], [0,0,1,0] 두 개의 자질 벡터를 생성한

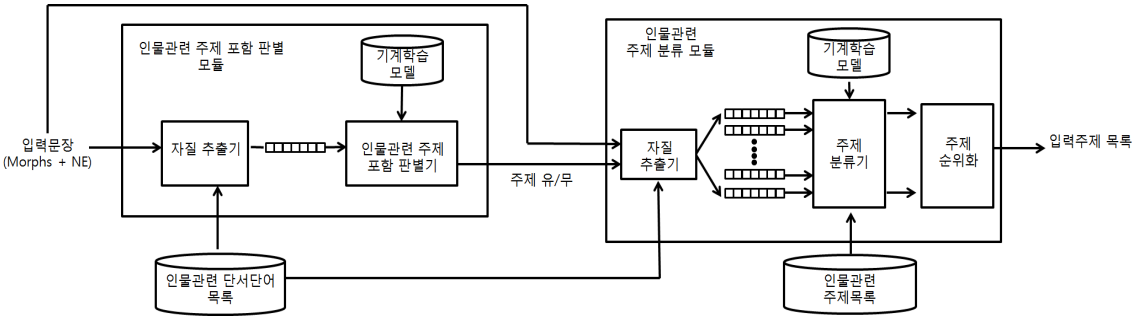

Fig. 3 Architecture of the Topic Classification System for Person-Related Sentences

다. 이렇게 생성된 각 자질 벡터들은 주제 분류기에 입력되어 각 단서표현들이 나타내는 주제를 찾는다.

인물관련 주제 분류 모듈의 **주제 분류기**는 기계학습 알고리즘을 이용하여 주제를 분류한다. 주제가 태깅된 문장을 이용하여 학습모델을 생성하고, 새로운 입력문장에서 추출된 자질을 모델에 적용하여 입력문장의 주제를 판별한다. 앞선 단계의 자질추출기에서 문장에서 인식된 단서단어의 수만큼 자질을 생성하고, 주제 분류기는 이를 입력받아 각 자질을 학습모델에 적용하여 해당 단서단어별로 주제를 판별한다.

한 문장에서 추출될 수 있는 주제가 여러 가지일 수 있다. **주제 순위화** 프로그램은 한 문장에서 나타난 여러 주제들의 우선순위를 결정한다.

예1) ***친구인 슈타틀리를 위해 작곡한 이 사람의 유일한 클라리넷 협주곡은 당대 최고의 사랑을 받았다.***

예 1)의 문장에서 "~를 위해 작곡"이라는 표현을 통해 이 문장이 어떤 인물의 작품활동과 관련된 글임을 알 수 있고 이를 통해 문장의 주제를 "작품-관련 인물"로 결정할 수 있다. 또한, "당대 최고의 사랑"이라는 문장을 통해 이 문장이 작품에 대한 평가를 담고 있으며 이에 대해 "평가"라는 주제를 할당할 수 있다. 이와 같이 질의와 응답문의 특성상 하나의 문장에도 여러 가지 주제가 담겨있는 경우가 있으므로, 추출된 주제들의 우선순위를 정할 필요가 있다. 이러한 작업을 인물관련 주제 분류 모듈의 주제 순위화 프로그램이 수행한다. 각각의 주제와 그 주제를 발생시킨 단서표현의 상호정보량(Mutual information)을 학습데이터를 통해 계산하고 이를 이용하여 순위를 결정한다. 이렇게 순위가 결정된 주제들 중 순위가 높은 N개의 주제를 해당 문장의 최종 주제로 결정한다.

4. 실 험

4.1 학습데이터 분석

실험에 사용한 질의-응답문은 3.1에서 밝힌 종류의 질의-응답문이며 전문가에 의해 태깅된 질의문은 총 993개이다. 이 중 893개는 단서표현 추출과 학습모델 생성을 위한 학습 데이터로, 나머지 100개의 질의문은 모델 평가를 위한 평가 데이터로 사용하였다.

학습데이터에서 893개에 대한 응답문은 13,395개이다. 질의문과 응답문은 각각 2,004개의 문장과 39,308개의 문장으로 구성되고 학습데이터의 총 문장 수는 41,312개이다. 평가 데이터는 100개의 질의문과 1,500개의 응답문으로 구성되며 각각 219개의 문장과 4,357개의 문장을 포함하여 총 4,576개의 문장으로 구성된다.

학습데이터의 문장 중 주제가 태깅된 문장의 수는 15,006개이고 이들 문장에 총 18,730개의 주제가 태깅되어있다. 평가 데이터에는 1,572개 문장에 주제가 태깅되어있고 이들 문장에 2,008개의 주제가 태깅되었다.

학습데이터에서 태깅된 주제 중 발생빈도 상위 5개와 하위 5개를 Table 5에 나타내었다.

태깅된 학습데이터로부터 추출한 단서표현들의 주제활성도 일부를 Table 6에 나타내었다. 주제활성도가 높은 단서표현은 질의-응답문에서 주제와 동반되어 사용되는 경우가 높다는 것이고, 반대로 주제활성도가 낮은 단서표현은 그렇지 않음을 나타낸다.

Table 5. Topics and Their Frequencies (5 Most Frequently Occurring and 5 Least Frequently Occurring Topics)

	Topic	Count
5-most	직업/직책	4382
	가족 관계	2005
	별칭	1003
	정의	933
	작품-활동 인물	910
5-least	발견-위치	34
	사망-위치	33
	약어	32
	발견	28
	제작-상세	20

Table 6. Example of Topic Activeness of Clue Expression

Topic	Topic Activeness
본명	0.950
뜻에서 유래	0.937
는 @@ 의미 하ㄴ다	0.562
중요 하ㄴ 역할	0.25

4.2 인물관련 주제 포함 판별 모듈 실험

인물관련 주제 포함 판별 모듈을 위한 자질 조합과 기계학습 알고리즘을 선택하고 이를 검증하기 위하여 실험을 수행하였다. 인물관련 주제 포함 판별 모듈의 자질추출기에서 추출하는 자질의 유형은 다음과 같다.

CPA: 전체 단서표현의 길이를 가지는 벡터를 정의하고, 문서에서 단서표현이 발견되면 벡터의 해당 위치에 1, 아니면 0으로 표현한 벡터

CPB: CPA에 각 단서표현의 주제활성도를 곱한 벡터

NE: 개체명 유형의 개수를 길이로 가지는 벡터에 문장에서 개체명이 출현하면 해당 개체명 유형의 위치에 1이 아닌 경우를 0으로 표현한 벡터

성능 실험에 이들 3가지 자질 벡터의 조합을 이용하였다. 실험에 사용한 자질의 조합은 Table 7과 같다.

Table 7. Combination of Features for a Predictor of Person-Related Topic Presence

Type	Combination
(A)	CPA
(B)	CPB
(C)	CPA+NE
(D)	CPB+NE
(E)	CPA+CPB+NE

추가적으로 분류기의 기본 성능(Baseline)을, 모든 문장을 주제가 없는 것으로 분류할 경우의 성능으로 하였다. 기계학습 모델 생성을 위한 알고리즘으로 SVM[17]과 Naive bayesian 분류기[18]를 사용하였다.

각 자질과 알고리즘을 사용한 인물관련 주제 포함 판별 모듈의 실험 결과는 Table 8에 나타내었다.

Table 8. Experimental Results for a Predictor of Person-Related Topic Presence

Features	SVM	Naive bayesian
Baseline	65.647	65.647
(A)	77.469	67.767
(B)	77.797	67.767
(C)	78.238	68.466
(D)	78.409	68.659
(E)	*78.628	68.291

4,576개의 평가 문장은 1,572개 문장에 주제가 태깅되어 있고 3,004개의 문장이 주제가 없는 것으로 태깅되어있다. 인물관련 주제 포함 판별 모듈의 성능은 분류기가 이들 문장의 주제 유/무를 분류하였을 때 얼마나 정확하게 맞추었는지를 평가하였다.

전체적으로 SVM알고리즘을 사용한 경우가 Naive Bays 알고리즘을 사용한 경우보다 더 성능이 나았다. SVM을 사용하였을 경우, (E)의 성능이 78.628%로 가장 높았고, Naive Bays알고리즘의 경우 (D)조합을 자질을 사용했을 때 68.659%로 가장 높은 성능을 나타내었다. 모든 문장을 주제가 없는 것으로 분류할 때의 성능인 65.647%의 정확도보다 각각 12.781%, 3.012% 더 나은 성능을 나타낸다. 이 실험을 통해 인물관련 주제 포함 판별 모듈은 SVM을 (E)자질조합과 함께 사용할 때 가장 좋은 성능을 나타냄을 알 수 있다.

4.3 인물관련 주제 분류 모듈 실험

이 장에서는 인물관련 주제 분류 모듈의 성능을 측정하기 위한 실험 결과를 나타내었다. 앞선 실험에서 인물관련 주제 포함 여부를 가장 잘 판별한 자질 조합인 SVM알고리즘과 (E)조합을 인물관련 주제 분류 모듈 실험의 자질 조합으로 사용하였다. 이 실험에서 입력문장은 앞선 단계인 인물관련 주제 포함 판별 모듈에서 주제가 있다고 판별한 1,294개 문장을 이용하였다.

Table 9는 수집된 모든 단서단어를 자질로 사용한 결과이다. 이 표에서 정확률(precision)은 입력문장의 주제들 중 주제 분류기가 정확히 예측한 주제의 개수를 주제 분류기가 추출한 주제의 개수로 나눈 값이다. 재현율(recall)은 정확률과 같은 분자를 입력문장들의 주제의 개수로 나눈 값이다.

Table 9. Results of Experiment for Topic Classification System

Top N	Precision	Recall	F1
3	0.317	0.754	0.446
2	0.369	0.676	0.477
1	0.454	0.522	0.486

F1은 정확률과 재현율을 통해 계산된 F1 점수를 나타낸다.

Table 9의 실험에서 입력된 문장들에 들어있는 주제의 개수는 총 1,114개이다. Top 1은 입력 문장에 대해 최대 1개의 주제를 추출하는 것이고 Top 2는 최대 2개, Top 3는 최대 3개까지의 주제를 추출한 결과이다. 주제 분류기가 Top 1의 주제를 찾았을 때 1,281개의 주제가 있다고 판별하였고 이 중 실제 입력문장의 정답에 있었던 경우는 581개였다. Top 2에서는 2,042개의 주제를 찾고 이 중 752개를 맞추었다. Top 3에서는 2,648개의 주제를 찾고 839개를 맞추었다. 이를 기반으로 정확률과 재현율, F1값을 계산한 것이 Table 9의 값이다.

Table 10은 주제활성도 0.8 이상인 단서표현들, 즉 단서표현들 중 주제를 발생시킬 가능성이 높은 표현들을 자질로 사용했을 때의 실험 결과이다. 여기서 주제의 개수는 1,114개이고 Top 1~3까지 추출한 주제의 개수와 올바르게 찾아진 주제의 개수는 각각 953개와 540개, 1,227개와 620개, 1,338개와 651개이다.

Table 10. Results of Experiment for Topic Classification System Filtering by Topic Activity

Top N	Precision	Recall	F1
3	0.487	0.584	0.531
2	0.505	0.557	0.530
1	0.567	0.485	0.523

Table 9와 Table 10의 실험결과에서 더 적게 주제를 추출할수록 추출한 주제에서 정답이 뽑힌 비율이 높아지므로 정확률이 높아지는 것과 반대로, 더 적게 뽑으면 정답에 포함되어있는 주제들에서 찾아진 정답의 비율이 낮아지므로 재현율이 낮아지는 것을 알 수 있다. 또한 주제활성도가 높아질수록, 정확률은 더 적게 뽑지만 더 적합한 주제를 찾으므로 정확률이 높아지고 재현율이 낮아지는 것을 볼 수 있다.

5. 결 론

본 논문에서는 질의응답 시스템의 정답문장 검색 결과의 성능 향상을 위해 사용될 수 있는 입력문장의 주제 분류기 시스템 개발에 대해 소개하였다. 이를 위한 주제 정의와 태깅 방법, 그리고 전체적인 주제 분류 시스템의 구성을 설명하고 문장에서 나타난 주제를 찾기 위한 단서표현의 유형을 정의하였다.

개발된 시스템의 성능 검증을 위한 실험 결과, 인물관련 주제 포함 판별 실험에서는 SVM을 이용한 실험에서 단서표현과 개체명 인식 결과, 주제활성도를 모두 자질로 사용한 경우가 78.628%로 가장 높은 정확도를 보였다. 인물관련 주제 분류 성능 검증 실험에서는 주제활성도와 추출되는 주제의 개수를 1~3개로 조절하며 실험을 하였다. 시스템이 사용될 때 이 실험 결과 정확률과 재현율을 참고하여 필요한 주제의 개수와 주제활성도의 기준을 정할 수 있다.

본 논문에서의 시스템은 질의응답 시스템에 실제로 적용하여 정답문장 검색의 성능이 얼마나 향상되는지, 그리고 이를 통해 얼마나 질의응답 시스템의 성능이 향상되는지를 평가해야 진정한 시스템의 성능이 평가된다고 할 수 있다.

하지만 이는 본 논문의 범위를 벗어나기 때문에 아쉬움이 남는 부분이다. 또한 주제 유무 판별과 주제 분류에 구문구조 등과 같이 높은 수준의 언어분석 결과를 사용하지 않고 문장의 형태소 분석과 개체명 인식 결과만을 분석에 사용하였다. 이를 통해 분석 속도에 대한 이점을 얻을 수 있었지만, 분류 성능에서는 한계점이 있다. 향후 더 높은 차원의 언어 분석 결과를 도입함으로써 더 나은 성능을 기대할 수 있을 것으로 생각한다.

본 논문에서 사용한 인물관련 주제 판별과 주제 추출은 단서표현과 기계학습 알고리즘을 이용하여 수행되었다. 좀 더 정밀한 판별 방법과 진보된 기계학습 알고리즘을 사용한다면 전체적으로 시스템의 성능을 끌어올릴 수 있을 것으로 기대되며 이는 향후 과제로 남겨둔다.

References

[1] Yongjin Bae and Hyunki Kim, "Estimating Block Weighting Scheme of Structured Text in the Information Retrieval for Question Answering," *Korea Computer Cogress*, pp.963-965, 2015.

[2] Zhang, Dell and Wee Sun Lee, "Question classification using support vector machines," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM*, 2003.

[3] Androutsopoulos, Ion, et al., "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages," *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, 2000.

[4] Antonellis, Ioannis, Christos Bouras, and Vassilis Poulopoulos, "Personalized news categorization through scalable text classification," *Frontiers of WWW Research and Development-APWeb 2006,* Springer Berlin Heidelberg, pp.391-401, 2006.

[5] McCallum, Andrew, and Kamal Nigam, "A comparison of event models for naive bayes text classification," *AAAI-98 Workshop on Learning for Text Categorization*. Vol.752. 1998.

[6] McCallumzy, Andrew, et al., "Building domain-specific search engines with machine learning techniques," *AAAI Technical Report SS-99-03*, 1999.

[7] Chen, Jingnian, et al., "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, Vol.36, No.3, pp.5432-5435, 2009.

[8] Wijewickrema, Chaaminda Manjula, and Ruwan Gamage, "An ontology based fully automatic document classification system using an existing semi-automatic system," *IFLA WLIC 2013 - Future Libraries: Infinite Possibilities*, Singapore, 2013.

[9] Morchid, Mohamed, Richard Dufour, and Georges Linares, "A LDA-based topic classification approach from highly imperfect automatic transcriptions," *LREC'14,* 2014.

[10] Quercia, Daniele, Harry Askham, and Jon Crowcroft, "TweetLDA: supervised topic classification and link prediction in Twitter," *Proceedings of the 4th Annual ACM Web Science Conference. ACM*, 2012.

[11] Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," *Proceedings of the 17th international conference on World Wide Web. ACM*, 2008.

[12] Faguo, Zhou, et al., "Research on short text classification algorithm based on statistics and rules," *Electronic Commerce and Security (ISECS), 2010 Third International Symposium on. IEEE*, 2010.

[13] Wang, Chang et al., "Relation Extraction with Relation Topics," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.1426-1436, 2011.

[14] Wang, Chang, et al., "Relation Extraction and Scoring in DeepQA," *IBM Journal of Research and Development*, Vol.56, Issue.3.4, pp.9:1-9:12, 2012.

[15] Changki Lee, Yi-Gyu Hwang, and Myung-Gil Jang, "Fine-grained named entity recognition and relation extraction for question answering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.799-800, 2007.

[16] Chae, "On the Classification and Distribution of Korean Adverbials: Focusing on the Distinction between Regular and Concord Adverbials," *Language and Linguistics*, Vol.29, pp.283-323, 2002.

[17] Cortes, Corinna and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, Vol.20, Issue.3, pp.273-297, 1995.

[18] Murphy, Kevin P., "Naive bayes classifiers," University of British Columbia, 2006.

이 경 호

e-mail : gyholee@gmail.com
2011년 충남대학교 정보통신공학과(학사)
2013년 충남대학교 정보통신공학과(석사)
2013년~현 재 충남대학교 전자전파정보통신공학과 박사과정
관심분야 : 자연언어처리, 기계학습, 인공지능

이 공 주

e-mail : kjoolee@cnu.ac.kr
1992년 서강대학교 전자계산학과(학사)
1994년 한국과학기술원 전산학과(공학석사)
1998년 한국과학기술원 전산학과(공학박사)
1998년~2003년 한국마이크로소프트(유) 연구원
2003년 이화여자대학교 컴퓨터학과 대우전임강사
2004년 경인여자대학 전산정보과 전임강사
2005년~현 재 충남대학교 전파정보통신공학과 교수
관심분야 : 자연언어처리, 기계번역, 정보검색, 정보추출