KIPS Tr. Software and Data Eng.
Vol.4, No.7 pp.283~290 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2015.4.7.283

Recognition of Answer Type for WiseQA

Heo Jeong† · Ryu Pum Mo†† · Kim Hyun Ki††† · Ock Cheol Young††††

ABSTRACT

In this paper, we propose a hybrid method for the recognition of answer types in the WiseQA system. The answer types are classified into two categories: the lexical answer type (LAT) and the semantic answer type (SAT). This paper proposes two models for the LAT detection. One is a rule-based model using question focuses. The other is a machine learning model based on sequence labeling. We also propose two models for the SAT classification. They are a machine learning model based on multiclass classification and a filtering-rule model based on the lexical answer type. The performance of the LAT detection and the SAT classification shows F1-score of 82.47% and precision of 77.13%, respectively. Compared with IBM Watson for the performance of the LAT, the precision is 1.0% lower and the recall is 7.4% higher.

Keywords : Question Answering, Answer Type, Question Analysis, WiseQA

WiseQA를 위한 정답유형 인식

허 정† · 류 법 모†† · 김 현 기††† · 옥 철 영††††

요 약

본 논문에서는 WiseQA 시스템에서 정답유형을 인식하기 위한 하이브리드 방법을 제안한다. 정답유형은 어휘정답유형과 의미정답유형으로 구분된다. 본 논문은 어휘정답유형 인식을 위해서 질문초점에 기반한 규칙모델과 순차적 레이블링에 기반한 기계학습모델을 제안한다. 의미정답유형 인식을 위해 다중클래스 분류에 기반한 기계학습모델과 어휘정답유형을 이용한 필터링 규칙을 소개한다. 어휘정답유형 인식성능은 F1-score 82.47%이고, 의미정답유형 인식성능은 정확률 77.13%이다. 어휘정답유형 인식성능은 IBM 왓슨과 비교하여, 정확률은 1.0% 저조하고, 재현율은 7.4% 높다.

키워드 : 질의응답, 정답유형, 질문분석, 와이즈QA

1. 서 론

범람하는 텍스트 콘텐츠(contents)와 모바일(mobile) 환경의 도래가 정보검색(information retrieval)의 한 분야인 질의응답(question answering) 기술의 중요성을 부각시키고 있다. 키워드(keywords)에 기반한 질의(query) 중심의 기존 정보검색은 정제되지 않은 콘텐츠를 단순 순위화(ranking)하여 사용자에게 제공하기 때문에 사용자는 제시된 콘텐츠를 검토하여 원하는 정보를 확인하는 번거로움이 있다. 또한 모바일 기기의 제한된 화면에 기존 정보검색 방식과 같이 정제되지 않은 콘텐츠를 제시하는 것은 비효율적인 방식이다.

질의응답은 사용자의 질문(question)에 적합한 정답(answer)과 함께 근거(evidence) 문장이나 문서를 제시하는 기술이다. 정보검색 기반 질의응답시스템(IR based Q&A system)은 일반적으로 질문분석(question analysis), 문서검색(document retrieval), 정답후보추출(answer candidate extraction), 정답순위화(answer ranking)의 4가지 컴포넌트들로 구성된다. 질문분석은 사용자의 질문을 이해하여 사용자가 요구하는 정답의 유형을 파악하고, 중요한 키워드를 인식하는 것이 핵심 기능이다. 문서검색은 정답을 포함하고 있는 문서나 문장을 검색하는 것으로 질문에서 추출된 키워드에 기반하여 검색엔진에 맞는 질의를 생성(query reformulation)하는 것이 중요하다. 정답후보추출은 검색된 문서나 문장에서 정답후보(answer candidate)에 해당하는 개체(entity), 어휘(word) 또는 구(phrase)를 추출하는 것으로 질문분석에서 인식된 정답유

※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.R0101-15-0062, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발).
† 정 회 원 : 울산대학교 정보통신공학 박사과정, 한국전자통신연구원 선임연구원
†† 비 회 원 : 한국전자통신연구원 책임연구원
††† 정 회 원 : 한국전자통신연구원 책임연구원
†††† 종신회원 : 울산대학교 전기공학부 IT융합전공 교수
Manuscript Received : February 13, 2015
First Revision : May 7, 2015
Second Revision : May 19, 2015
Accepted : May 21, 2015
* Corresponding Author : Heo Jeong(jeonghur@etri.re.kr)

형(AT; Answer Type)이 중요한 단서가 된다. 정답순위화는 질문과 정답후보가 포함된 문장의 의미적 유사도(semantic similarity)와 정답유형과 정답후보의 의미적 연관성(semantic relation)에 기반한다.

본 논문에서는 질의응답 기술 중 첫 번째 컴포넌트인 질문분석에서 정답유형을 인식하기 위한 방법과 기술에 대해서 소개하고자 한다. 정답유형은 크게 두 유형으로 구분될 수 있다. 첫째, 기정의된(predefined) 의미범주(semantic category)를 대상으로 질문에서 요구하는 정답의 의미범주를 분류하는 의미정답유형(SAT; Semantic Answer Type)이 있다. 일반적으로 PLO를 중심으로 하는 개체명 범주를 많이 사용한다. 둘째, 질문 내에서 정답의 유형을 제약하는 어휘인 어휘정답유형(LAT; Lexical Answer Type)이 있다. 질문 1의 정답은 '내비게이션'으로, 질문 내에 정답을 제약하는 어휘로 '장치'가 있다. 본 논문은 앞서 언급된 의미정답유형과 어휘정답유형을 인식 방법과 기술에 대해서 제안할 것이다.

질문 1. 지도를 보여줘 자동차 운전을 도와주는 길 안내 장치

본 논문의 구성은 2절에서 관련 연구를 소개하고, 3절에서 정답유형인식의 상세 방법론에 대해서 제안한다. 4절에서는 평가 데이터에 대해서 기술하고, 5절에서 평가방법과 실험결과에 대해서 토의한다. 6절에서는 본 논문에 대한 결론과 향후 연구방향에 대해서 제안한다.

2. 관련 연구

질의응답 기술의 발전에서는 1999년부터 시작된 TREC(Text Retrieval Conference)의 질의응답 경연대회(competition)가 큰 역할을 하였다. TREC에서는 사실기반 질의응답(factoid Q&A)부터 시작하여 리스트형 질의응답(list Q&A), 정의형 질의응답(definition Q&A), 대화형 질의응답(interactive Q&A)을 거쳐 실시간 질의응답(live Q&A)으로 기술적 난이도를 높이고 있다[1].

IBM은 1997년 'Deep Blue' 이후 인공지능(artificial intelligence)의 새로운 도전 분야로 질의응답을 선택하였고 2011년에 DeepQA인 왓슨(Watson)을 야심차게 발표하였다. 왓슨은 'Jeopardy! 퀴즈쇼'에서 인간챔피언을 이기면서 크게 주목을 받았다. 이는 인공지능 연구 부흥에 큰 역할을 하였다[2].

왓슨에 대한 기술 개발은 1999년 IBM의 질의응답 시스템인 PIQUANT를 기반으로 2006년부터 2011년까지 5년간 개발되었다. PIQUANT는 1999년부터 2005년까지 TREC의 QA track에서 상위의 성적을 보였다. 그러나 PIQUANT도 고정된 정답유형이 미리 결정되어있었고, 이는 'Jeopardy!' 퀴즈의 폭넓은 도메인과 정답을 충족하기에는 한계가 있었다[2-4].

LASSO 시스템의 질문분석모듈은 크게 3단계로 구분된다. 첫째, 질문의 유형(question type)을 결정한다. 질문유형은 5W1H에 기반한다. 둘째, 정답유형을 결정한다. 정답유형은 정답의 의미범주로 정의한다. 이때, Who형 질문의 경우 정답유형이 PERSON으로 명확하지만, What형 질문의 경우 모호성 문제(ambiguous problem)가 발생한다. 이 문제를 해소하기 위해서, 질문초점(QF; Question Focus)의 개념을 정의하고 있다. 질문초점은 질문에서 찾고자 하는 것을 지칭하여 질문의 모호성을 해소하는 질문 내의 단어나 구로 정의된다[5].

[6]에서는 오픈 도메인(open-domain) 질의응답을 위해서 정답유형 텍소노미(answer type taxonomy)를 정의하고 있다. 정답유형 텍소노미는 3단계로 구성된다. 첫째, 가장 대표적인 개념노드(conceptual node)를 텍소노미의 최상위 의미범주(semantic category)에 추가한다. 둘째, 텍소노미의 최상위 의미범주와 개체범주(named entity category)를 다대다로 매핑(many-to-many mapping)한다. 셋째, 최상위 의미범주에 워드넷(WordNet)의 하위계층들(sub-hierarchies)을 연결한다. [6]에서는 질문의 구문분석 트리(syntactic parsing tree) 중 핵심 의존 수식관계(head-modifier dependency)를 파악하고 정답유형 텍소노미를 이용하여 정답유형을 결정한다. 정답유형은 일부 예외적인 경우를 제외하고는 기본적으로 개념노드로 인식된다.

왓슨에서는 정답유형으로 기정의된 의미범주를 사용하지 않는다. LASSO 시스템과 같이 질문초점을 정의하고 패턴(pattern)을 이용하여 인식한다. 인식된 질문초점의 중심어를 어휘정답유형으로 인식한다. 어휘정답유형을 인식할 때 대용어 인식(anaphora resolution)을 활용한다. 패턴에 의해 인식된 어휘정답유형은 로지스틱 회귀분석(logistic regression)에 기반한 분류기(classifier)로 신뢰도(confidence)를 계산하여 필터링(filtering)을 수행한다[7].

LASSO 시스템은 5W1H의 질문유형, 의미정답유형과 질문초점을 분석하고 있다. 반면, 왓슨에서는 질문초점을 패턴으로 인식하고 질문초점의 중심어를 어휘정답유형으로 인식하고 있다. 본 논문에서는 두 시스템의 장점을 모두 적용한다. 5W1H를 중심으로 10개의 질문유형을 분류하고, 질문초점과 어휘정답유형을 인식한다. 또한 정답유형의 의미범주인 의미정답유형을 정의하고 인식한다. 의미정답 유형은 정답후보의 개체명 정보와의 연관성을 기반으로 정답을 제약할 때 사용된다. 그리고 어휘정답유형은 정답후보와의 의미적 연관성을 텍소노미나 온톨로지(ontology)와 같은 어휘지식자원을 이용하여 계산하고 정답을 순위화하는 자질(feature)로 활용한다.

LASSO, 왓슨과 [6]에서의 정답유형은 구문분석의 의존관계나 의문사와의 구문적 관계(syntactic relation)에 기반한 패턴으로 인식된다. 반면, 본 논문에서는 기계학습 방법과 패턴을 하이브리드한 방식으로 정답유형인식의 재현율(recall)을 크게 향상시켰다.

3. WiseQA를 위한 질문분석기

Fig. 1은 본 논문에서 소개하는 질문분석의 구성도이다.

질문분석은 질문초점 인식, 어휘정답유형 인식과 의미정답유

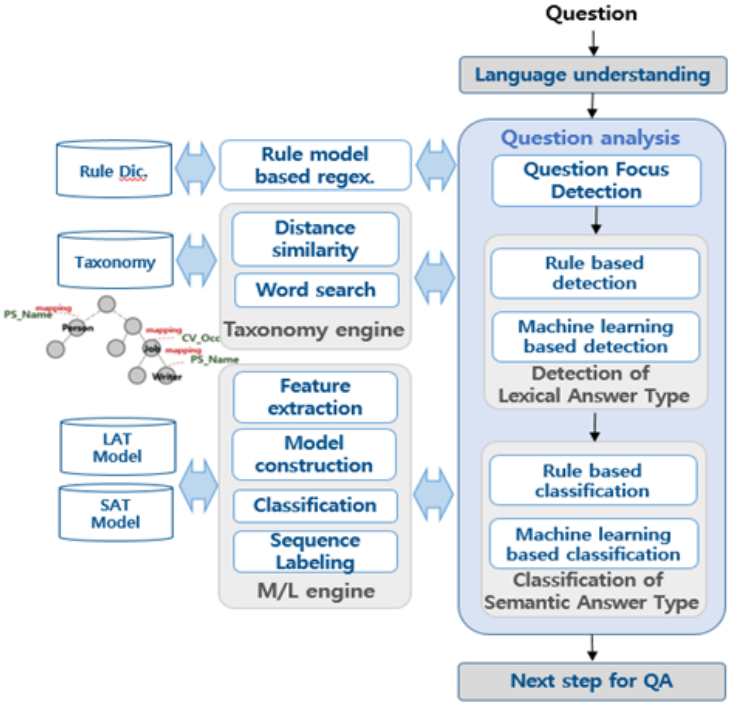

Fig. 1. The Architecture of Question Analysis for WiseQA

형 인식 모듈로 구성된다. 질문분석을 위한 세부모듈로는 규칙사전(rule dictionary)을 이용하는 정규표현식(regular expression) 기반 규칙모델(rule model), sSVM을 이용하는 기계학습모델(machine learning model)과 텍소노미를 대상으로 검색(search)과 거리유사도(distance similarity)를 측정하는 텍소노미 엔진으로 구성된다. 텍소노미는 울산대의 UWordMap, 부산대의 Korlex와 한국전자통신연구원(ETRI)의 개체명 태그셋(NE tag set)이 이용되었다[8-10].

3.1 질문초점 인식

LASSO와 왓슨에서의 질문초점은 다음과 같이 정의되고 있다.

- LASSO : A focus is a word or a sequence of words which define the question and disambiguate it in the sense that it indicates what the question is looking for, or what the question is all about.
- 왓슨 : The focus is the part of the question that is a reference to the answer.

결국, “질문에서 찾고자 하는 정답을 지칭하고 있는 질문 내의 특정 부분”을 의미한다. 본 논문에서의 질문초점은 정답후보가 대치될 수 있는 질문 내의 특정 부분이다. 이와 같은 정의로 인해, 질문에 포함되는 의문사도 질문초점으로 고려될 수 있다. 이는 정답후보의 근거점수 계산(evidence scoring)을 위한 재검색에서 질문을 재생성(question rewriting)할 때 이용된다. Fig. 2는 질문초점을 이용한 질문재생성의 예시를 보여주고 있다. 의문형을 평서형으로 변환하고 질문초점의 위치에 정답후보를 대치함으로써 새로운 문장을 생성할 수 있다. 재생성된 질문은 검색모듈의 질의로 변환되어 검색에 이용된다. 검색된 결과의 점수는 정답후보의 근거점수 계산을 위한 자질로 이용된다.

질문초점 인식은 어휘구문패턴(lexico-syntactic pattern)을 이용한다. 질문초점 인식을 위한 기본 패턴은 다음과 같다.

① 지시관형사 + 명사
② 지시대명사
③ 의문사
(ㄱ) 몇+명사
(ㄴ) 어떤+명사

대상 의문사는 ‘누구’, ‘어떤’, ‘무엇’, ‘어디’, ‘어느’, ‘언제’, ‘얼마’, ‘몇’이다. 의문사들 중, ‘몇’과 ‘어떤’은 뒤따르는 명사에 따라 정답의 대상이 결정된다. 이로 인해 뒤따르는 명사까지 질문초점의 경계를 확장하였다.

질문 2. 석회암 지대에서 주성분인 탄산칼슘이 물에 녹으면 깔때기 모양으로 파인 웅덩이가 생긴다. 이 웅덩이 안에 경작할 수 있는 지형을 ‘돌리네’라고 하는데, 돌리네 여러 개가 연결돼 긴 와지를 이루는 이것은 무엇일까?

‘지시관형사+명사’와 ‘지시대명사’로 구성된 질문초점은 상

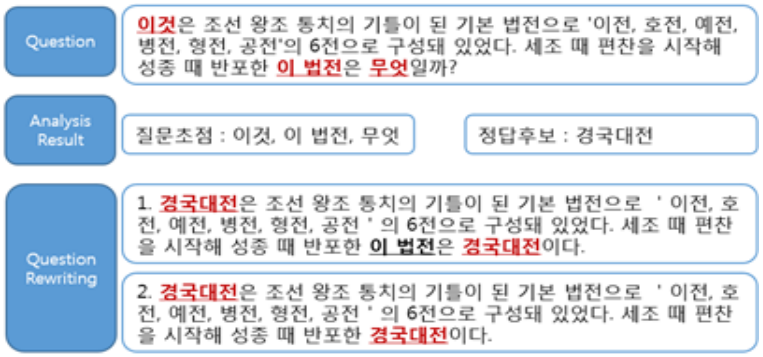

Fig. 2. An Example of Question Focus and Question Rewriting for Evidence Retrieval

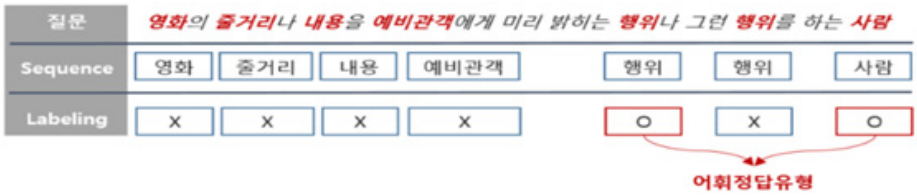

Fig. 3. An Example of Noun-based Sequence Labeling for LAT Detection

호참조(co-reference resolution)를 이용하여 필터링이 수행된다. 질문 2를 대상으로 질문초점 인식 규칙을 적용하면, '이 웅덩이', '이것', '무엇'이 질문초점으로 인식된다. 그러나 '이 웅덩이'는 앞 문장의 '웅덩이'를 지칭하는 것으로 정답을 지칭하는 것이 아니다. 이를 필터링하기 위해서 상호참조기술을 이용한다. 필터링 알고리즘은 다음과 같다.

```
For i = 0, 1, ..., QF size
    Find Co-reference List (CR) of QF[i]
    For j = 0, 1, ..., CR size
        If CR[j] is not question focus,
            QF[i] is filtered out
```

규칙에 의해서 인식된 질문초점 QF에 대해서 상호참조 정보인 CR의 리스트를 찾고, 해당 CR 리스트에서 QF에 포함되지 않는 멘션(mention)이 하나라도 있으면, 해당 질문초점은 필터링 된다.

3.2 어휘정답유형(LAT) 인식

어휘정답유형은 정답과 의미적인 연관성을 가지는 질문 내의 어휘를 의미한다. 의미적인 연관성은 일반적으로 Instance-of, isa, part-of 등이 있다. 따라서 어휘정답유형은 정답을 제약하는 중요한 자질로 활용될 수 있다.

질문초점의 중심어에 해당하는 명사는 일반적으로 어휘정답유형이 된다. Fig. 2의 질문초점, '이 법전'에서 중심어인 '법전'은 어휘정답유형이 된다. 정답후보인 '경국대전'과는 instance-of의 관계로 연결된다. 의미적 연관성은 워드넷과 같은 어휘지식자원(lexical knowledge resource)을 기반으로 파악할 수 있다. 본 논문에서는 Fig. 1에서 언급한 텍소노미를 이용하였다.

어휘정답유형 인식을 위해서 휴리스틱한 규칙과 기계학습을 이용하였다.

규칙은 앞서 언급한 바와 같이 질문초점의 중심어를 어휘정답유형으로 인식하는 기본적인 규칙과 어휘구문패턴(lexico-syntactic pattern)에 기반한 규칙으로 나뉜다. 어휘구문패턴은 형태소 분석과 의존구문 분석 결과에 기반하여 정규표현식으로 패턴을 정규화한 것이다.

기계학습에서의 어휘정답유형인식은 sSVM을 이용한 명사기반 순차적 레이블링 문제(noun based sequence labeling problem)로 해결한다. 어휘정답유형은 텍소노미에 등록된 어휘만을 대상으로 한다. 이는 어휘정답유형과 정답후보의 의미적 연관성 파악을 위해 텍소노미를 이용하기 때문이다. 이로 인해, 어휘정답유형을 명사로 제한하였다. 또한, 복합명사로 인한 많은 문제를 해결하기 위해서 어절 단위로 복합명사의 범위를 제한하였다. 명사기반 순차적 레이블링의 예는 Fig. 3과 같다. Table 1은 기계학습에 사용되는 자질목록이다.

3.3 의미정답유형(SAT) 인식

의미정답유형은 질문에서 요구하는 정답의 의미범주를 의미한다. 본 논문에서의 의미범주는 [11]을 참조하여 180여 개의 개체명 태그로 정의된다.

의미정답유형 인식은 어휘의미패턴(lexico-semantic pattern)에 기반한 규칙과 sSVM을 이용한 기계학습모델을 사용하였다. 어휘의미패턴은 형태소 분석, 개체명 분석과 의존구문 분석 결과에 기반하여 정규표현식으로 정규화한 규칙이다. 의미정답유형 인식은 기계학습의 다중클래스 분류문제(multiclass classification)와 동일하다. 기계학습의 단점은 짧은 시간에 개별 질문에 대한 튜닝이 어렵다는 것이다. 본

Table 1. Features of LAT for Machine Learning

Categories	Features	Description
위치(Position)	Word index	단어가 포함된 어절 번호
형태소정보(Morpheme Info.)	Word	단어와 단어의 bi-gram 문자열
	Morpheme tag	형태소 태그
	Morpheme info. of ±N position	지정된 앞뒤 어절의 형태소 정보
개체명정보(NE Info.)	NE tag	단어의 개체명 태그
구묶음정보(Chunking Info.)	Chunking label	단어가 포함된 구묶음의 레이블
구문정보(Parsing Info.)	Parsing tag	단어가 포함된 어절의 구문 태그
	Parsing tag of modifier	수식 어절의 구문 태그
	Modifier word	수식 어절의 단어 문자열
	Morpheme tag of modifier	수식 어절의 형태소 태그
질문유형(Question Type)	Interrogative type	의문사의 유형
질문초점(Question Focus)	True/False	해당 단어가 질문초점에 포함되었는지 여부

Table 2. SAT Features which are Added to [11]

Categories	Features	Description
질문유형(Question Type)	Interrogative type	의문사의 유형
질문초점(Question Focus)	Word in question focus	질문초점에 포함되어있는 단어와 단어의 bi-gram 문자열
	Word of ±1 position	질문초점 앞뒤 어절의 단어와 단어의 bi-gram 문자열
어휘정답유형(Lexical Answer Type)	LAT words	LAT 단어와 bi-gram 문자열

논문에서는 이 문제를 해결하기 위해서 규칙모델과 기계학습모델을 하이브리드 하였다.

의미정답유형 인식에서 기계학습모델의 자질은 [11]의 언어분석 자질에 Table 2의 자질을 추가하여 사용되었다.

질문 3. 영화의 줄거리나 내용을 예비관객에게 미리 밝히는 행위나 그런 행위를 하는 사람

본 논문에서는 어휘정답유형에 기반한 의미정답유형 필터링 규칙을 추가하였다. 인식된 어휘정답유형과 의미정답유형 간의 의미적 연관성을 중심으로 의미정답유형을 필터링할 수 있다. 질문 3에서 인식된 어휘정답유형은 '행위'와 '사람'이다. 의미정답유형으로 PS_NAME이 인식된다면, '사람'과 PS_NAME은 텍소노미 상에서 밀접하게 매핑(mapping)되어있을 것이다. 반면, OGG_BUSINESS의 경우, '사람'과 연관성이 적을 것이다. 이를 이용하여 기계학습으로 인식된 의미정답유형 후보를 필터링할 수 있다.

4. 평가데이터

기계학습을 위해서는 학습데이터(training data)를 구축하여야 한다. 본 논문에서는 정답유형인식을 위한 학습데이터로 약 81,500개의 질문을 수집하고 어휘정답유형과 의미정답유형을 부착하였다. 학습데이터의 구성은 Table 3과 같다. WiseQA는 장학퀴즈를 대상으로 하는 질의응답 시스템이다. 그러나 수집된 퀴즈질문의 개수가 부족하여 네이버 지식IN의 타이틀 중 질문에 해당하는 것을 수집하여 보완하였다.

Table 3. Statistics of Training Data

Sources	The number of questions	Description
Naver	64,237(78.82%)	네이버 지식IN의 Title
Quiz	17,259(21.18%)	장학퀴즈, 도전골든벨, …
Total	81,496	

평가데이터(evaluation data)는 '장학퀴즈'의 질문으로 586개를 선정하였다. '장학퀴즈'에는 다양한 유형의 질문이 있다. '장학퀴즈'는 기본적으로 3라운드로 구성되고, 객관식과 주관식으로 질문이 구분된다. 객관식 문제 중, 보기(examples)를 제거하면 문제가 성립되지 않는 질문도 있다.

질문 4. 날짜를 가리키는 단어 중 순우리말이 아닌 것은 무엇일까? [어제, 오늘, 내일, 모레]

질문 5. 오전 열한 시부터 오후 한 시까지를 가리키는 말로 십이시의 일곱 번째 시는 무엇일까? [인시, 묘시, 오시, 술시]

질문 6. 지금 보이는 숫자들은 일정한 규칙에 따른 것입니다. 규칙을 따를 때 ?에 들어갈 숫자는 무엇일까?

질문 4의 경우, 보기 없이는 질문이 되지 않는다. 반면 질문 5는 보기가 없어도 주관식 질문으로 정답을 찾을 수 있다.

질문 6는 그림이 제시되어야만 하는 질문이다. 즉, 텍스트만 이용해서는 질문의 의미를 확인할 수 없다. 평가데이터에서는 질문 4와 질문 6과 같이 보기나 시청각자료(음원, 그림, 동영상 등) 없이 질문이 성립되지 않는 질문은 제외하였다.

5. 실 험

실험은 기계학습모델에 대한 평가와 하이브리드 모델(hybrid model)에 대한 평가를 모두 진행하였다. 기계학습모델에 대한 평가는 학습데이터를 대상으로 10 fold cross validation을 수행하였다. 하이브리드 모델은 평가데이터를 대상으로 평가를 수행하였다. 질문초점과 어휘정답유형은 한 질문에 복수 개의 정답을 가질 수 있으므로 F1-Score를 평가척도로 이용하였고, 의미정답유형은 질문에 가장 적합한 의미범주 하나를 결정하는 문제이므로 정확률(precision)을 평가척도로 이용하였다.

$$\text{Precision} = \frac{\#Correctly\ Detected\ ATs(or\ QFs)}{\#Detected\ ATs(or\ QFs)} \quad (1)$$

$$\text{Recall} = \frac{\#Correctly\ Detected\ ATs(or\ QFs)}{\#ATs(or\ QFs)\ in\ Evaluation\ Set} \quad (2)$$

$$\text{F1\_Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

기계학습모델에 대한 평가에서는 다양한 학습자질들이 성능에 미치는 영향을 평가하였다. Table 4와 Table 5는 언어분석 정보를 기본자질(baseline)로 사용하고, 질문유형(QT), 질문초점(QF)과 어휘정답유형(LAT) 자질들을 추가하면서 성능을 평가한 결과이다. 어휘정답유형의 경우, 질문유형과 질문초점 자질이 추가됨에 따라 정확률은 상승하지만, 재현율이 저하되는 것을 알 수 있다. 그러나 재현율의 하락보다 정확률 상승의 폭이 커서 조화평균인 F1-score의 값은 상승하였다. 의미정답유형의 경우에는 [11]에서 사용된 언어분석 정보를 기본자질(baseline)로 사용하고, 추가적인 자질들을 포함시키며 성능변화를 평가하였다. 개별적인 자질들 중, 어휘정답유형 자질(LAT)이 성능향상에 가장 큰 기여를 하였다. 이는 어휘정답유형과 의미정답유형 사이의 의존성이 크다는 것을 확인할 수 있는 결과이다.

Table 4. Evaluation Results of LAT(Training Data)

Features	Precision	Recall	F1-Score
Baseline	0.8267	0.7326	0.7767
+ QT	0.8293 (+0.0026)	0.7312 (-0.0014)	0.7771 (+0.0004)
+ QF	0.8455 (+0.0188)	0.7280 (-0.0046)	0.7822 (+0.0055)
+ QT + QF	0.8466 (+0.0199)	0.7276 (-0.0050)	0.7825 (+0.0058)

Table 5. Evaluation Results of SAT(Training Data)

Features	Precision
Baseline	0.7692
+ QT	0.7725 (+0.0033)
+ QF	0.7717 (+0.0025)
+ LAT	0.7775 (+0.0083)
+ QT + QF	0.7742 (+0.0050)
+ QT + LAT	0.7802 (+0.0110)
+ QF + LAT	0.7796 (+0.0104)
+ QT + QF + LAT	0.7818 (+0.0126)

평가데이터에 대한 실험결과는 Table 6과 Table 7에 제시되어있다. 질문초점과 어휘정답유형에 대한 평가는 하이브리드 모델에 대한 평가결과이다. 의미정답유형은 하이브리드 모델과 기계학습모델에 대한 평가결과를 구분하여 제시하고 있다.

Table 6. Evaluation Results of QF and LAT(Evaluation Data)

Components	Precision	Recall	F1-Score
QF	0.9467	0.9375	0.9421
LAT	0.7430	0.9265	0.8247

Table 6에서 질문초점(QF)은 F1-score로 94.21%의 성능을 보인다. 오류의 대부분은 상호참조에 기반한 필터링 오류이다.

질문 7. [거룩한 분노는/종교보다 깊고/불붙은 정열은/사랑보다 강하다/아! 강낭콩꽃보다도 더 푸른/그 물결 위에/양귀비꽃보다도 더 붉은/그 마음 흘러라] 이 시는 임진왜란 당시 충절을 보여준 인물을 진주 남강의 강물로 형상화해 표현했다. 변영로 작품의 시적 대상이자 제목이기도 한 이 인물은 누구일까?

질문 7과 같이 인용문을 지칭하는 '지시관형사+명사'의 경우, 상호참조에서 인용문과 '지시관형사+명사'를 동일한 멘션으로 인식하기에는 기술적으로 한계가 있다. 이와 같은 상호참조의 기술적 한계로 인해 질문 7의 '이 시'를 질문초점에서 필터링하지 못하는 오류가 발생한다.

Table 6의 어휘정답유형(LAT) 평가결과는 학습데이터에 대한 기계학습모델 평가와 양상이 많이 다르다. 기계학습모델 평가 대비, 정확률은 많이 하락(-10.36%)한 반면, 재현율은 많이 상승(+19.89%)하였다. 이는 학습데이터의 구성비율과 관련된 것이다. 약 80%의 비율인 Naver 지식IN의 질문데이터는 상대적으로 짧은 단문이나 명사구로 구성된 질문이 대부분이다. 반면, 장학퀴즈질문은 평균 1.48문장으로 구성되는 복합문장이 많다. 이와 같은 차이로 정확률이 크게 하락한 것으로 분석된다. 반면 재현율의 상승은 질문초점에

기반한 규칙의 효과이다. 장학퀴즈질문은 정답을 지칭하기 위해서 '지시관형사+명사'의 패턴을 많이 사용한다. 질문 7의 '이 인물'과 같은 경우로, 질문초점의 중심어인 '인물'은 쉽게 어휘정답유형으로 인식될 수 있다.

Table 7. Evaluation Results of SAT(Evaluation data)

Models	Hybrid model		ML model	
	Precision	Recall	Precision	Recall
SAT	0.7713	-	0.7747	-
SAT(big class)	0.7918	-	0.7935	-
SAT@Rank3	-	0.7933	-	0.7882

Table 7과 같이 의미정답유형의 성능은 크게 3가지 방식으로 나누어 평가하였다. 의미정답유형의 의미범주는 개체명 태그이다. 개체명 태그는 계층구조로 구성되며, 최상위 계층(big class)은 15개의 범주로 구분된다. 15개의 최상위 범주에 대한 분류 성능을 평가하였다. 또한 제시되는 의미정답유형의 순위에서 3위 내에 정답이 있는 경우(Recall@Rank3)를 파악하기 위한 평가를 수행하였다. 세부 의미범주를 대상으로 한 평가결과(SAT)는 하이브리드 모델과 기계학습모델에서 각각 77.13%와 77.47%의 정확률로 0.34%의 차이를 보이고 있다. 최상위 범주에 대한 평가결과(SAT : big class)는 각각 79.18%과 79.35%의 정확률을 보이고, 순위 3위 내에 올바른 의미정답유형이 있는 경우(SAT@Rank 3)의 재현율이 각각 79.33%과 78.82%이다. 하이브리드 모델이 기계학습모델 대비 정확률은 다소 낮고, Rank3의 재현율은 다소 높다. 그러나 1% 미만의 성능차이로 큰 의미는 없다. 앞서 언급한 바와 같이 규칙모델은 기계학습모델의 단점인 개별 질문분석 결과의 튜닝이 어렵다는 문제를 해결하기 위한 목적으로 하이브리드 되었다.

6. 결론 및 향후 연구

본 논문에서는 WiseQA를 위한 정답유형인식 기술에 대해서 소개하였으며, 정답유형을 어휘정답유형과 의미정답유형으로 구분하였다. 두 정답유형을 인식하기 위해서 기계학습모델과 규칙모델을 하이브리드 한 모델을 제시하였다. 어휘정답유형의 기계학습모델은 순차적 레이블링 문제로 해결하였고, 의미정답유형은 다중클래스 분류 문제로 해결하였다. 기계학습모델에서 사용된 다양한 자질들을 소개하였고, 자질별 성능기여 정도를 평가하였다. 어휘정답유형 인식을 위한 규칙모델에서는 질문초점을 이용한 휴리스틱한 방법을 소개하였다. 의미정답유형 인식을 위한 규칙에서는 어휘정답유형과 의미정답유형의 관계에 기반한 필터링 규칙에 대해서 소개하였다.

본 논문에서 제안한 정답유형인식은 IBM 왓슨과 비교하여 다음의 장점이 있다.

첫째, IBM 왓슨은 어휘정답유형만을 인식했으나, 본 논문에서는 의미정답유형을 함께 인식하고 있다. 의미정답유형을 이용하여 정답후보에 대한 강력한 제약을 적용할 수 있기 때문에 질의응답의 정확률을 개선할 수 있다.

Table 8. Comparison of LAT Performance between IBM Watson and WiseQA

System	Precision	Recall	F1-Score
Rule (Watson)	0.753	0.853	0.800
WiseQA	0.743 (-0.010)	0.927 (+0.074)	0.825 (+0.025)

둘째, IBM 왓슨은 어휘정답유형 인식을 위해 질문초점에 기반한 규칙만을 사용하였다. 그러나 본 논문에서는 sSVM에 기반한 기계학습모델을 추가함으로써, Table 8처럼 IBM 왓슨보다 재현율을 크게 향상시켰다. IBM 왓슨과 같이 규칙만을 이용한 모델(Rule)과 비교하여 WiseQA의 재현율이 7.4% 향상되었다.

앞으로 진행되어야 할 연구는 다음과 같다.

첫째, 시스템에서 제시하는 정답유형의 신뢰도(confidence) 계산방법에 대한 연구가 필요하다. 정답유형의 신뢰도 값은 정답유형을 기반으로 정답후보를 제약할 때 활용될 수 있는 중요한 자질이다.

둘째, 질문분할 기술을 이용한 정답유형 검증에 대한 연구이다. 앞선 언급한 바와 같이 장학퀴즈는 평균 1.48개의 문장으로 구성된다. IBM 왓슨에서와 같이 질문분할(question decomposition)을 적용한다면[12], 동일한 정답을 요청하는 복수 개의 분할질문들(sub-questions)을 얻을 수 있다. 분할질문들에 대한 정답유형을 인식한다면 원 질문과 분할질문들 간의 정답유형을 상호 검증할 수 있다.

References

[1] John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel, "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)," *Document Understanding Conferences Roadmapping Documents*, 2001.

[2] FERRUCCI, David A., "Introduction to "this is watson"," *IBM Journal of Research and Development*, Vol.56, No.3.4, pp. 1:1-1:15, 2012.

[3] Prager, J., Chu-Carroll, J., Czuba, K., Welty, C., Ittycheriah, A., & Mahindru, R., "IBM's PIQUANT in TREC2003," pp.283-292, *TREC*, 2003.

[4] Chu-Carroll, J., Czuba, K., Prager, J. M., Ittycheriah, A., and S, "IBM's PIQUANT II in TREC 2004," in *TREC*, 2004.

[5] Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, and Vasile Rus,

"The Structure and Performance of an Open-Domain Question Answering System," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000.

[6] Pasca, Marius A., and Sandra M. Harabagiu, "High performance question/answering," *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2001.

[7] LALLY, Adam, et al. "Question analysis: How Watson reads a clue," *IBM Journal of Research and Development*, Vol.56, No.3.4, pp.2:1-2:14, 2012.

[8] Choe Ho-seop, "Construction Method of Large-scale 'Urimal(Korean)-Word Intelligent Network'," *Hangul* 273, pp.125-141, 2006(in Korean).

[9] Aesun Yoon, Soonhee Hwang, Eunryoung Lee, and Hyukchul Kwon, "Construction of Korean Wordnet 「KorLex 1.5」," *Journal of KIISE: Software and Applications*, Vol.36, No.1, pp.92-108, Korea, 2009.

[10] C. Lee and M. Jang, "Named Entity Recognition with Structural SVMs and Pegasos algorithm," *Korean Journal of Cognitive Science,* Vol.21, No.4, pp.655-667, Korea, 2010.

[11] Jeong Heo, Pum-Mo Ryu, Myung-Gil Jang, and Hyun-Ki Kim, "Search Space Reduction and Answer Type Classification for Open Domain Q&A," *Journal of KIISE: Software and Applications*, Vol.39, No.2, pp.118-132, Korea, 2012.

[12] KALYANPUR, Aditya, et al. "Fact-based question decomposition in DeepQA," *IBM Journal of Research and Development*, Vol.56, No.3.4, pp.13:1-13:11, 2012.

허 정

e-mail : jeonghur@etri.re.kr
1999년 울산대학교 전자계산학과(학사)
2001년 울산대학교 전자계산학과(석사)
2001년~현 재 한국전자통신연구원 선임연구원
2013년~현 재 울산대학교 정보통신공학전공 박사과정
관심분야 : 자연어처리, 정보검색, 텍스트마이닝, 빅데이터처리

류 법 모

e-mail : pmryu@etri.re.kr
1995년 경북대학교 컴퓨터공학과(학사)
1997년 POSTECH 컴퓨터공학과 (공학석사)
2009년 KAIST 전자전산학과(공학박사)
2009년~현 재 한국전자통신연구원 자동통역인공지능연구센터 책임연구원
관심분야 : 정보검색, 자연어처리, 온톨로지, 질의응답시스템

김 현 기

e-mail : hkk@etri.re.kr
1995년 전북대학교 컴퓨터공학부(석사)
2005년 University of Florida 전산학(박사)
1995년~현 재 한국전자통신연구원 책임연구원
관심분야 : 자연어 처리, 정보검색, 자연어 질의응답

옥 철 영

e-mail : okcy@ulsan.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 서울대학교 컴퓨터공학과(석사)
1993년 서울대학교 컴퓨터공학과(박사)
1994년 러시아 TOMSK 공과대학 교환교수
1996년 영국 GLASGOW 대학교 객원교수
2007년~2008년 한국정보과학회 언어공학연구회 위원장
2008년 국립국어원 객원교수
1984년~현 재 울산대학교 전기공학부 IT융합전공 교수
2010년~현 재 울산대학교 국어국문학부 겸직교수
관심분야 : Korean Language Processing, Korean Homograph Tagging, Ontology, Knowledge Base, Document Clustering