위키피디아 QA를 위한 질의문의 정답제약 추출

왕지현○, 허정, 이형직, 배용진, 김현기

한국전자통신연구원
{jhwang, jeonghur, leehj, yongjin, hkk}@etri.re.kr

Answer Constraints Extraction on User Question for Wikipedia QA

JiHyun Wang○, Jeong Heo, Hyungjik Lee, Yongjin Bae, Hyunki Kim
Electronics and Telecommunications Research Institute

요 약

질의응답 시스템에서 정답을 제약하기 위한 위키피디아 영역의 정답제약 9개를 정의하고 질문 문장에서 제약표현을 추출하는 방법을 제안한다. 다어절의 정답제약 표현을 추출하기 위해서 언어분석 결과를 활용하여 정답제약 후보를 생성하며 후보단위로 정답제약 표현을 학습하기 위한 자질을 제시한다. 기계학습 방법을 이용하여 정답제약 후보 별로 정답제약 태그를 분류하여 정답제약 표현을 추출한다. 성능 실험은 각 정답제약 태그 별로 F1-Score 평가를 수행하였다.

주제어: 질의응답, 질문분석, 정답제약

1. 서론

질의응답(QA) 시스템은 자연어 질문에 대해 정답을 찾는 시스템이다. 사용자 질의문으로부터 정확한 검색 의도를 알기 위해서는 질문에 포함된 질문중심어휘(Question Centric Words; 이하 QCW로 표기)와 QCW를 수식하는 질문 내의 표현들이 정답을 찾는 주요 근거가 된다. QCW는 질문에서 정답의 유형을 표현한 어휘를 말하며 이를 정답으로 대치하는 경우 정답가설(Answer Hypothesis)이 된다. QCW를 수식하여 정답을 제약하는 표현들을 '정답제약' 이라고 한다.

(1) Q : "세계에서 가장 높은 빌딩은?"
A : "부르즈할리파"
X='부르즈할리파' <- QCW='빌딩'
<- 정답제약='가장 높다','세계'

(2) Q : "'10일간의 이야기'라는 뜻의 소설집은?"
A : "데카메론"
X='데카메론' <- QCW='소설집'
<- 정답제약='10일간의 이야기'

본 논문은 위키피디아 영역의 자연어 질문에서 출현하는 정답제약 유형을 정의하였으며, 규칙과 기계학습을 사용하여 정답제약 표현을 추출하고 정답제약 유형을 결정하는 방법을 제안한다.

2. 정답제약 유형

위키피디아 영역의 자연어 질문에 대해 고빈도로 출현하는 9개의 정답제약 유형을 정의하였다. 제약마다 1개 이상의 서브필드명(sub-fieldname)이 있다. 예를 들어, 정의제약은 mean과 origin으로 구성되어 있다.

(1) 시간제약 : 정답과 관련된 시간표현
예) "조선 시대 안정복이 지은 역사책은?"
=> time='조선 시대', (QCW='역사책')

(2) 공간제약 : 정답과 관련된 구체적인 장소를 나타내는 공간표현
예) "전한을 세운 사람은?"
=> loc=전한, (QCW=사람)

(3) 별칭제약 : 정답을 달리 부르는 이름
예) "호가 '고산자'인 이 사람은 누구인가?"
=> alias=고산자 (QCW=사람)

(4) 정의제약 : 정답의 의미와 기원
예) "아랍어로 메뚜기를 뜻하며 아랍 에미리트 연방에 속한 도시"
=> mean=메뚜기
=> origin=아랍어, (QCW=도시)

(5) 언어제약 : 정답의 언어(language)
예) "'금성'을 가리키는 순우리말은?"
=> lang=순우리말, (QCW=순우리말)

(6) 부정제약 : 부정하는 대상
예) "독도를 가리키는 말이 아닌 것은 무엇일까?"
1) 우산도, 2) 삼봉도, 3) 가지도, 4) 관음도
=> neg=독도, (QCW=말)

(7) 공칭제약 : 부여된 공식명칭 및 번호. '국보', '보물', '중요무형문화재' 등
예) "이것은 천연기념물 218호로 지정된 곤충이다."
=> type=천연기념물, => number=218호, (QCW=곤충)

(8) 순서제약 : 정답과 관련된 순서 및 최상급 표현

예) “학급에서 끝에서 두 번째로 큰 학생은?”
=> domain=학급, startingPoint=끝, type=두 번째, predicate=큰, target=학생, (QCW=학생)
예) “세계 최초의 인공위성은?”
=> domain=세계, type=최초, target=인공위성, (QCW=인공위성)

(9) 차이제약 : 정답과의 비교대상
예) “사자성어 중에 ‘양’의 의미가 다른 것은 무엇일까?”
=> domain=사자성어, comp=‘양’의 의미, (QCW=것)

3. 정답제약 추출

정답제약을 추출하는 접근방법은 처리 대상 질문 문장에서 정답제약 후보들을 생성하고 각 후보 별로 자질들을 추출하여 기계학습 모델로 학습한 후, 학습된 모델을 사용하여 각 후보가 정답제약인지 여부를 판별하는 것이다(그림1). Sequence Labeling방식을 사용하지 않고 정답제약 후보를 생성하는 이유는 정답제약 추출값의 경계가 구(phrase) 단위를 넘어서는 긴 표현의 경우에 학습량이 많아지게 되고 추출 대상 표현의 중간에서 끊겨서 온전한 추출이 실패할 가능성이 높기 때문이다.

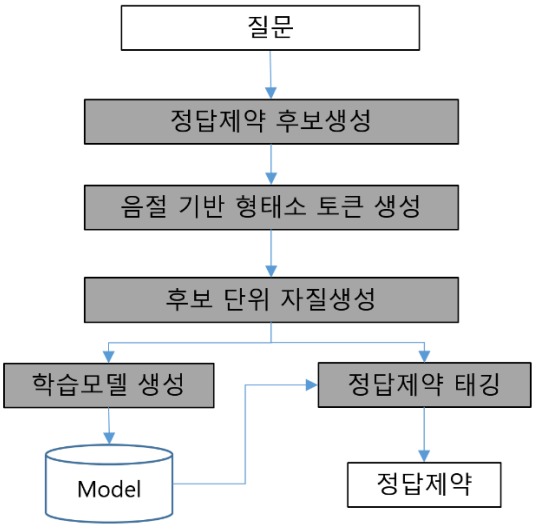

그림 1. 정답제약 추출 흐름도

2.1 정답제약 후보생성

정답제약 후보는 자연어 질문을 언어분석하여 어절, 개체명, 단일명사, 명사구 청크, 절의 수식을 받는 명사구, 기호문자로 둘러싼 인용구문을 대상으로 후보를 생성하였다[1][2]. 예를 들어, 질문 “흰 사슴이 물을 마시던 연못이라는 뜻을 가진 한라산 정상의 화구호 이름은?”이라는 문장에서 정답제약 후보를 생성하면 다음과 같이 생성하게 된다.

(1) 어절 : C1=(흰), C2=(사슴이), C3=(물을),..C11=(이름은)
(2) 개체명 : C12=(한라산)
(3) 단일명사 : C13=(사슴), C14=(정상), C15=(화구호), C16=(이름)
(4) 명사구 청크 : C17=(흰 사슴), C18=(물), C19=(연못), C20=(뜻), C21=(한라산 정상), C22=(화구호 이름)
(5) 절의 수식을 받는 명사구 : C23=(흰 사슴이 물을 마시던 연못), C24=(흰 사슴이 물을 마시던 연못이라는 뜻), C25=(흰 사슴이 물을 마시던 연못이라는 뜻을 가진 한라산), C26=(흰 사슴이 물을 마시던 연못이라는 뜻을 가진 한라산 정상), C27=(흰 사슴이 물을 마시던 연못이라는 뜻을 가진 한라산 정상 화구호 이름은)

위의 후보 생성 규칙은 4,358개 질문 문장으로 구성된 정답제약 태그드코퍼스에 대해서 약 70.83%의 커버리지를 보였다. 후보 유형별 비율을 보면, 명사구 청크가 48.72%로 가장 많은 비중을 차지하였고 개체명이 37.75%, 절의 수식을 받는 명사구가 13.51%의 순으로 커버하였다.

2.2 음절 기반 형태소 토큰 생성

질문 문장에 정답제약을 태깅할 때, 음절 경계에 태깅을 한다. 예를 들어, (흰) (사슴)(이) (물)(을) (마시)(던) (연못)(이)(라는) (뜻)(을) (가진) (한라산) (정상)(의) (화구)(호) (이름)과 같이 최소 분해 단위인 형태소 단위의 토큰을 생성하되 어미활용을 복원하지 않고 음절 단위의 태깅을 한다. 이와 같이 하는 이유는 생성한 정답제약 후보의 경계와 형태소 단위 토큰의 경계를 쉽게 정렬하기 위한 것으로, 수작업한 태그드코퍼스의 정답제약 태깅 위치를 형태소 토큰의 위치에 쉽게 매핑할 수 있다.

2.3 정답제약 자질생성

생성된 각각의 정답제약 후보에 대해서 자질을 생성한다. 후보별로 생성된 자질들은, 학습단계에서 정답제약 레이블과 함께 학습하여 기계학습 모델을 생성한다. 태깅단계에서는 기계학습 분류기의 입력이 되어 정답제약 레이블을 출력하게 된다.

생성하는 자질셋은 2가지가 있으며, 학습단계와 태깅단계에서 (2)의 정답제약 후보 단위 자질을 사용한다.

(1) 토큰 단위 자질
- 형태소 : m-2, m-1, m0, m+1, m+2
- 품사 : p-2, p-1, p0, p+1, p+2
- 개체명 : n-2, n-1, n0, n+1, n+2
- 어휘의미 : s-2, s-1, s0, s+1, s+2

(2) 정답제약 후보 단위 자질
- 좌측 경계 토큰의 토큰 단위 자질
- 우측 경계 토큰의 토큰 단위 자질
- 후보 중심어의 지배소의 형태소와 품사
- 후보 중심어의 술어 지배소의 형태소와 품사
- 후보 중심어의 술어 지배소의 대상격(obj) 논항의 형태소와 품사
- 후보 중심어의 의미역(SRL)
- 좌측 경계 토큰의 피지배소 개수

정답제약 태그드코퍼스 4,358 문장에서 출현하는 숫자 표현들 중에서 순서와 순차적 표현을 나타내는 다음과

같은 숫자들의 형태소 길이 분포를 조사하였다.

예) “제201-1호”, “제11항”, “1,2호”, “제19대”, “두 번째”, “18호”, “2층” 등

코퍼스 내에서 약 95% 이상이 +-2 내의 토큰 거리 안에 포함됨을 알 수 있었다. 이와 같은 이유로 토큰 단위 자질의 거리를 +-2로 정하였다.

2.4 정답제약 학습 및 태깅

정답제약 유형은 총 9개 유형이다. 9개의 정답제약에 정의된 전체 서브필드명을 고려하면 총 16개의 태그가 만들어 지며, 정답제약이 아닌 NOT태그까지 고려하여 17개의 태그를 만들었다. 각 정답제약 후보가 생성하는 자질들에 대해서 17개의 태그를 Structural SVM [3]으로 학습하고 태깅하였다.

3. 성능평가

위키피디아 영역의 장학퀴즈형 자연어 질문 4,358개 문장(평균어절수는 약 18어절)을 개발셋으로 학습하고 별도의 657개 문장을 블라인드 평가셋으로 실험하였다. 정확한 추출경계와 정답제약의 각 서브필드명의 태그를 잘 맞췄는지를 F1-Score로 평가하였다.

표1. 장학퀴즈형 자연어 질문의 성능평가

제약명	개발셋	블라인드셋	제약명	개발셋	블라인드셋
시간	92.15%	80.10%	부정	84.68%	82.17%
공간	94.01%	65.01%	공칭	76.12%	83.87%
별칭	90.27%	65.55%	순서	84.65%	72.39%
정의	79.55%	77.70%	차이	77.97%	92.31%
언어	90.91%	76.60%	-	-	-

위키피디아 영역의 단문형 자연어 질문 2,878개 문장(평균어절수는 약 3어절)을 개발셋으로 하여 학습하고 평가를 하였다.

표2. 단문형 자연어 질문의 성능평가

제약명	개발셋	제약명	개발셋
시간	95.20%	부정	-
공간	92.99%	공칭	64.29%
별칭	96.97%	순서	80.61%
정의	85.71%	차이	-
언어	97.22%	-	

단문형 평가셋에서는 부정제약과 차이제약이 출현하지 않아서 학습과 평가를 할 수 없었다.

4. 결론 및 토의

본 논문은 질의응답 시스템을 위한 질문분석 과정에서 정답을 제약하기 위한 위키피디아 영역의 정답제약 9개를 정의하였고 질문 문장에서 제약표현을 추출하는 방법을 제안하였다. 비교적 긴 형태의 장학퀴즈 질문셋과 짧은 단문형 질문셋에 대한 성능평가를 수행했다.

의미기반의 태깅에 있어서 어렵고 중요한 이슈는 학습 코퍼스를 수작업으로 구축할 때 모호한 점을 최소화하기 위해 태그 별로 명확한 태깅 기준을 정하는 것이고 태깅 일관성을 유지하는 것이다. 정답제약 태깅도 표현에 따라 태깅 여부가 상당히 애매한 경우가 많았으며 코퍼스 구축 작업이 진행될수록 일관성을 유지하기가 쉽지 않았다.

사사문구

본 연구는 과기정통부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [2013-0-00131, (엑소브레인-1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

참고문헌

[1] 임수종, 김현기, “의미 정보를 이용한 한국어 의미역 인식 연구”, 제27회 한글 및 한국어 정보처리 학술대회 논문집, pp.18-22, 2015.

[2] 임준호, 배용진, 김현기, 김윤정, 이규철, “의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치”, 제27회 한글 및 한국어 정보처리 학술대회 논문집, pp.234-239, 2015.

[3] Y.Altun, T.Hofmann and I.Tsochantaridis, SVM Learning for Interdependent and Structured Output Spaces, Proceedings of the ICML, 2004.