국어정보 질의응답을 위한 키워드 추출

전석종O, 이수인, 이현아
금오공과 대학교 컴퓨터소프트웨어공학과
wjstjrwhd123@nate.comO, rubyou123@naver.com, halee@kumoh.ac.kr

Keyword Extraction for Korean Language Q&A

Jong-Seok JongO, Su-In Lee, Hyun-A Lee
Kumoh National Institute of Technology

요 약

국립국어원 온라인가나다에서 제공되는 질의응답 문서를 이용한 국어정보에 대한 Q&A시스템은 언어 자체에 대한 질문과 답변의 특성으로 조사나 어미로 끝나는 표현이 주어로 등장하는 등의 특이한 문장이 자주 나타난다. 이러한 이유로 형태소 분석을 거쳐 명사를 키워드로 추출하는 일반적인 키워드 추출 방식은 좋은 성능을 얻기 어렵다. 본 논문에서는 국어정보 질의응답 문서의 특징에 맞는 키워드 추출 방법을 제안한다. 제안하는 방식에서는 문장 단위로 분할된 결과에서 연결어미로 문장을 추가로 분할한 뒤에 조사 앞에 나타나는 단어열을 키워드로 추출한다. 덧붙여 다자비교형 질의에서의 키워드 추출을 위해 편집거리를 이용한 키워드 추출 방법을 제안한다.

주제어: 국어정보 질의응답, 키워드 추출

1. 서론

질의응답 시스템이란 주어진 질의에 대해 응답을 구하는 시스템을 총칭하는 용어로, 사용자가 제시하는 질의를 분석하여 문서 집합으로부터 정답을 추출하는 것을 목적으로 한다. 기존의 질의응답 시스템에 대한 접근은 정제되지 않은 대량의 문서로부터 한정되지 않은 분야의 질의에 대한 답변을 찾는 것을 주목적으로 하고 있다[1,2]. 대량의 질의응답 문서로부터 새로운 질문에 대한 정답을 찾고자 하는 연구가 국내에서 이루어졌으나[3] 단순한 통계적 기법에 의존하고 있다.

국어 정보 처리 시스템 경진대회[4]에서는 국립국어원 온라인 가나다[5]의 한국어에 대한 '질문-답' 쌍의 학습 말뭉치를 제공하고 이에 대한 cQA(Community Question Answering) 시스템 개발을 목표로 진행되었다. 국어정보 질의응답은 한정된 분야에 대한 잘 정제된 QA 문서로 인하여 일반적인 QA에 비해 높은 정확도를 보일 수 있으나, 언어 자체에 대한 질문과 답변의 특성으로 인한 문제점을 가진다. 예를 들어 조사나 어미로 끝나는 표현이 문장의 주어로 등장하거나 한 문장 전체가 주어로 오는 문장이 흔하게 발생한다. 이런 특성으로 띄어쓰기 기준으로 어절을 구분하고 명사를 키워드로 판단하는 일반적인 접근 방식으로는 올바른 답변을 찾기 어렵다.

이러한 문제를 해결하기 위해 [5]에서는 질문 분류를 13개로 구분하고 게시판의 특성상 어법에 맞지 않는 문장을 처리하기 위해 분석된 주부와 술부, 주제어 등을 자질로 사용한 벡터 유사도를 계산하여 유사 답변을 추천하였다. [6]에서는 온라인가나다의 특수성에 맞추어 명사가 동사 뿐만 아니라 다양한 형태소열을 자질로 사용할 것을 제안하였다. 하지만 두 방식 모두 지정된 패턴에 기반한 방식을 활용하는 한계가 있다.

본 논문에서는 수동으로 구축한 패턴이나 말뭉치에서 추출한 패턴 없이 국어질의응답 문서의 특성에 맞는 키워드 추출 방법을 제안한다.

2. 국어정보 질의응답을 위한 키워드 추출

아래 예문은 국어정보 질의응답에서 발생하는 실제 질문이다. 예문 [가]와 [나]에서는 '것입니다'와 '겁니다'가 질문의 핵심에 해당하는 단어이지만, 일반적인 정보검색이나 질의응답에서와 같이 형태소 분석을 거쳐 명사만을 키워드로 추출하면 명사 '것'만 추출되어 정확한 유사 질문을 찾기 어렵다. 예문 [다]에서는 핵심 단어인 '돼'와 '되'보다는 '답변', '문장', '종결' 등의 가중치가 높아 적합하지 않은 질문이 유사 질문으로 선택될 수 있다.

> [가] '-겁니다'는 '-것입니다'의 구어체이며 옳은 말인가요?
> [나] 올 것입니다의 준말이라면 당연히 올 겁니다가 맞는 것이라고 생각했는데…
> [다] 문장을 종결할 때 '-되'로 쓰이는 경우는 없나요? 항상 '-돼'로만 쓰이는 것 같아서요. 답변 부탁드립니다.

이 예제들에서는 질문의 핵심 단어가 모두 조사 앞에 나타나는 공통점을 찾을 수 있다. 또한 형태소 분석을 통한 명사 추출이 적합하지 않음도 알 수 있다. 본 논문에서는 이 점에 착안한 키워드 추출 방법을 제안한다.

2.1 어미 기준 문장 분리와 조사 기반 키워드 추출

예문 [가]~[다]에서 살펴본 바와 같이 정상적인 문장을 대상으로 한 형태소 분석과 명사 키워드의 사용은 국어정보 질의응답에 적합하지 않다. 이 문제를 해결하기 위해 조사 어휘만을 활용하여 키워드를 추출할 수 있다. 아래는 예문 [라]와 이에 대한 형태소 분석 결과를 보인다. 형태소 분석에서는 종결어미를 기준으로 문장을 분리하여 [라-1]과 [라-2]의 분석 결과를 낸다. 결과에서는 신청일로부터’가 올바른 결과인 ‘신청일+로부터’가 아닌 ‘신청+일로+부터’로 분석는 형태소 분석 오류를 발견할 수 있다. 이로 인하여 조사를 기준으로 키워드를 추출하면 ‘일로’가 키워드로 추출되는 문제가 발생한다. 마찬가지로 예문 [나]에 대해서도 조사만을 활용하면 ‘올 것입니다’가 아닌 ‘것입니다’만을 키워드로 추출하여, 적합한 유사 질의를 검색할 수 없게 된다.

[라] 신청일로부터 5일 후라고 하면 언제를 얘기하는 걸까요? 오늘을 포함해서 계산해야 되는 건 알겠는데, 5일 후라는 말의 뜻은?
[라-1] 신청(NNG) 일로(NNG)부터(JX) 5(NR)일(NNM) 후(NNG)라고(JX) 하(VV)면(ECE) 언제(NP)를(JKO) 얘기(NNG)하(XSV)는(ETD) 걸(VV) ㄹ까요(EFQ)?(SF)
[라-2] 오늘(NNG)을(JKO) 포함(NNG)하(XSV)어서(ECD) 계산(NNG)하(XSV)어야(ECD) 되(VV)는(ETD) 건(NNM) 알(VV)겠(EPT)는데(ECD) 5(NR)일(NNM) 후(NNG)이(VCP)라는(ETD) 말(NNG)의(JKG) 뜻(NNG)은(JX)?(SK)

이러한 문제는 조사 앞에 나타나는 단어열 ‘신청일로’와 ‘올 것입니다’를 연결하여 키워드로 추출하는 방식으로 해결할 수 있다. ‘신청일로’는 정확한 결과는 아니지만, 입력 질의와 질의응답 문서의 질의가 동일하게 분석된다면, 적합한 유사질의 검색의 결과를 얻게 해 준다.

하지만 이 방식을 적용하면 [라-1] ‘하면 언제’가 키워드로 추출되는 문제가 발생한다. 이러한 문제를 해결하기 위해 본 논문에서는 연결어미를 기준으로 문장을 술부 단위로 추가 분리한다. [라-1]에 대해 문장 분리를 시행하면 아래와 같은 결과를 얻을 수 있다. 분리된 문장에서는 ‘신청일로’, ‘5일후’, ‘언제’가 키워드로 추출되어 보다 정확한 결과를 얻을 수 있다.

[라-1-1] 신청(NNG) 일로(NNG)부터(JX) 5(NR)일(NNM) 후(NNG)라고(JX) 하(VV)면(ECE)
[라-1-2] 언제(NP)를(JKO) 얘기(NNG)하(XSV)는(ETD) 걸(VV) ㄹ까요(EFQ)?(SF)

2.2 수열 유사도에 기반한 유사 형태 키워드 추출

위의 예문 [가]~[다]와 아래 예문 [마], [바]에서 볼 수 있듯이 국어정보 질의응답 문서에서는 두 가지 표현이나 그 이상의 표현 중에서 어떤 것이 적합한지를 묻는 질의가 다수 발생한다. 이러한 다자비교 질의는 맞춤법이나 발음, 띄어쓰기, 표기 등의 대부분의 질의 분류에서 발생한다.

[마] ~의 윤곽이 가시화되다가 맞나요, ~의 윤곽이 가시화하다가 맞나요?
[바] 부딪히다와 부딪치다의 차이점이 무엇인가요?

예문 [마]의 경우 기본 문장 분석을 위해 형태소 분석을 시행하면, 어말어미 기준으로 문장이 분리되어 ‘~의 윤곽이 가시화되다’, ‘가 맞나요’, ‘의 윤곽이 가시화하다’, ‘가 맞나요’로 분석되어 2.1의 조사 기반 키워드 추출 방식으로는 ‘가시화하다’를 키워드로 추출할 수 없다. 이처럼 다자선택형 질의는 형태소 분석 결과를 사용하지 않고 질의에 발생하는 유사형태의 어구(예를 들어 ‘가시화되다’와 ‘가시화하다’)를 키워드로 추출하는 것이 적합하다. 이 점에 착안하여 본 논문에서는 수열유사도를 활용하여 키워드를 추출하고자 한다.

수열 유사도에 기반한 유사 형태 키워드 추출을 위해 입력된 원 문장에서 어절 쌍을 추출한다. 예문 [마]에서는 ‘~의’부터 ‘맞나요?’까지의 8개의 어절을 추출할 수 있고, 이로부터 ${}_8C_2 = 28$개의 어절 쌍을 추출할 수 있다. 얻어진 어절 쌍 w_i와 w_j에 대해서 자소단위 수열유사도 $SIM(w_i, w_j)$를 $EditDistance$ 즉 편집거리[7]를 사용하여 아래의 식(1)으로 구한다. 식에서는 $|w_i|$는 어절 w_i의 길의를 나타낸다. 편집거리의 최대값은 두 어절 중 긴 어절의 길이이므로, 식 (1)에서는 편집거리를 긴 어절의 길이로 나누고 이를 1에서 빼서, [0,1]의 범위로 어절 유사도를 구한다. 결과로 동일한 어절이 두 번 발생하는 ‘~의’과 ‘윤곽이’에 대해서는 1, ‘가시화되다가’와 ‘가시화하다가’에 대해서는 0.94, ‘맞나요,’와 ‘맞나요?’에 대해서는 0.98의 수열유사도를 얻을 수 있다.

$$SIM(w_i, w_j) = 1 - \frac{EditDistance(w_i, w_j)}{Max(|w_i|, |w_j|)} \quad (1)$$

제안하는 시스템에서는 수열유사도가 임계치 이상인 어절 쌍을 키워드로 추출하여, 다자비교형 질의에 적합한 유사 질의를 검색에 사용한다.

2.3. 추출된 키워드를 활용한 유사 질의 추천

시스템에서는 제안한 방법으로 추출한 키워드를 이용하여 온라인 가나다를 위한 유사 질의를 추천한다. 문서 d에서 추출된 키워드 t에 대한 가중치는 식 (2)의 tf-idf를 사용한다.

$$W_{t,d} = \log(1 + tf_{t,d}) \times \log\frac{Q}{df_t} \quad (2)$$

사용자 질의 q와 시스템의 질의-응답 집합의 문서 d와 유사도는 코사인 유사도[8]을 이용하여 계산한다. 시스템에서는 코사인 유사도의 내림차순으로 유사 질의를 사용자에게 제시한다.

3. 평가 및 결론

본 논문에서는 국어정보 질의응답을 위한 키워드 추출 방법을 제안하였다. 형태소 분석과 명사 추출로 정확한 결과를 낼 수 없는 문서의 특성을 반영하여 조사와 어미를 이용한 추출 방식과 함께, 다자선택형 질의를 위한 편집거리 기반 키워드 추출 방식도 제안하였다.

제안한 키워드 추출 방식의 평가를 위해 [4]에서 사용한 20개의 질의를 사용하여 2.3에서 설명한 유사 질의 검색 방법을 사용한 실험을 수행하였다. 형태소 분석기를 통해 얻어진 어근을 키워드로 사용한 경우 10위권 이내에 정답이 포함되는 경우가 60%이며 MAP는 0.302 인 것에 반하여, 제안한 키워드 추출 방식을 사용한 경우 10위권 이내에 정답이 포함되는 경우가 80%이며 MAP는 0.4252 를 나타나 성능 향상을 얻을 수 있었다.

향후 연구로는 기존 연구에서 사용하였던 '띄어쓰기' 나 '발음' 등의 질문의 주제를 효과적으로 결합하는 방법과 답변에 존재하는 키워드들을 효율적으로 사용하는 방법을 예정하고 있다.

참고문헌

[1] Hirschman, L., Gaizauskas, R., "Natural language question answering", Cambridge University Press, 2001.

[2] Rivindu Perera, "IPedagogy: Question answering system based on web information clustering", IEEE 4th International Conference on Technology for Education, pp.245-246, 2012.

[3] 유동현, 이현아, "Q&A 문서의 검색 결과 요약을 활용한 질의응답 시스템", 정보처리학회지 3(4), 2014.

[4] 2014 국어 정보 처리 시스템 경진대회, https://ithub.korean.go.kr/user/contest/contestIntro LastView.do

[5] 도수종, 김용성, 염홍선, 정소윤, 김광준, 서정연, "주·술부 분석과 주제어 추출을 이용한 국문정보 커뮤니티 기반 질의응답 시스템", 한국정보과학회 동계학술발표회 논문집, 1290-1292, 2014.

[6] 박용민, 김보겸, 이재성, "질문 특성을 고려한 커뮤니티 질의응답 시스템(cQA) 자질 추출 방법", 제26회 한글 및 한국어 정보처리 학술대회, 119-121, 2014.

[7] https://en.wikipedia.org/wiki/Edit_distance

[8] https://github.com/need4spd/devyssim