질문 특성을 고려한 커뮤니티 질의응답 시스템(cQA) 자질 추출 방법

박용민[O*], 김보겸[†], 이재성[‡]
한국과학기술원 전산학과[O*], 충북대학교 디지털정보융합학과[†], 충북대학교 소프트웨어학과[‡]
bluem31@kaist.ac.kr, bogyum@cbnu.ac.kr, jasonlee@cbnu.ac.kr

Feature Extraction for Community Question Answering System(cQA) considering Question Characteristic

Yongmin Park[O*], Bogyum Kim[†], Jae Sung Lee[‡]
Dept. of Computer Science, Korea Advanced Institute of Science and Technology[O*]
Dept. of Digital Informatics and Convergence, Chungbuk National University[†]
Dept. of Software Engineering, Chungbuk National University[‡]

요 약

커뮤니티 질의응답 시스템(cQA)은 기존에 구축된 '질문-답' 쌍에서 사용자의 질문과 비교하여 유사도 순으로 결과를 보여주는 시스템이다. 본 논문에서는 '국립국어원'의 질의응답 게시판에 적용 가능한 '커뮤니티 질의응답 시스템'을 소개하고, 국립국어원 질의응답 게시판의 질문 특성을 분석하여 cQA의 성능 향상을 위한 자질 추출 방법을 제시한다.

주제어: 질의응답, 커뮤니티 질의응답, 자질 추출

1. 서론

온라인으로 서비스를 제공하는 수많은 사이트에서는 보통 '질의응답(QA) 게시판'을 운영하고 있다. 일반적으로 사용자가 질의응답 게시판에 질의문을 등록하면 게시판 관리자가 이에 대한 답변을 달아주는 형식이다. 하지만 사용자의 질의문이 등록되면 관리자가 답변을 해줄 때까지 사용자는 기다려야 하며, 관리자는 기존의 유사한 질의문에 대한 답이 있음에도 불구하고 각 게시글에 대한 답변을 직접 달아주어야 하기 때문에 매우 비효율적이다. 이러한 비효율적인 게시판 운영을 개선하기 위하여 '자주 묻는 질문'에 대한 답변을 따로 모아서 제공하기도 하고, 각 게시판에서는 기존의 질의문에 대한 답변을 찾아볼 수 있도록 게시판 검색 기능을 제공하지만 게시판 검색 기능은 대부분 '제목', '본문'에 등장하는 단어들을 이용한 단순 매칭 기법을 사용하기 때문에 검색어로 입력한 단어가 들어 있는 문서를 정렬해서 보여주는 형식에 불과하다.

질의응답 게시판을 효율적으로 활용하기 위해서는 사용자의 질문 의도를 올바르게 파악하고, 해당 질의문을 분석해 기존의 질의응답 문서 중 가장 적절한 '질문-답' 쌍을 유사도 순으로 보여 줄 필요가 있다. 따라서 본 논문에서는 '질문-답' 쌍에서 질문 특성을 분석하여 효율적인 '커뮤니티 질의응답 시스템(cQA)'을 구축하는 방법을 제안한다.

2. 관련 연구

커뮤니티 질의응답 시스템[1,2,3]은 일반적인 질의응답 시스템[4,5,6]과는 차이가 있다. 일반 질의응답 시스템은 다양한 데이터를 기반으로 사용자의 질문에 대한 정확한 답변을 찾아내는 작업을 수행하지만, 커뮤니티 질의응답 시스템은 커뮤니티별로 이미 구축되어 있는 '질문-답' 쌍을 이용하여 사용자 질문과 가장 유사도가 높은 순으로 '질문-답' 쌍을 제시해 주는 작업을 수행한다.

커뮤니티 질의응답 시스템에 관한 연구로는 확장된 나이브 베이즈 분류기를 이용하여 질문의 목적에 따라 정보형, 제안형, 의견형으로 자동 분류하는 기법을 제안한 연구[1]가 있으며, 사용자의 검색어나 질의어에서 추출한 키워드의 의미를 확장하여 검색 도메인을 선정한 후, 해당 도메인에 맞는 전문영역을 검색하는 연구[2]도 진행되었다. 또한 커뮤니티 질의응답 시스템의 전체 성능을 높이기 위하여 사용자의 질문을 몇 개의 주제로 분류하는 방법에 관한 연구[3]도 진행되었다.

본 논문에서는 '국립국어원 질의응답 게시판'을 대상으로, 한국어의 특성을 고려한 커뮤니티 질의응답 시스템에 관한 연구를 수행하였다.

3. 커뮤니티 질의응답 시스템(cQA)

국립국어원에서 제공한 '질문-답' 쌍을 이용하여 커

뮤니티 질의응답 시스템을 개발하였다. 전체적인 시스템 구성은 그림 1과 같다.

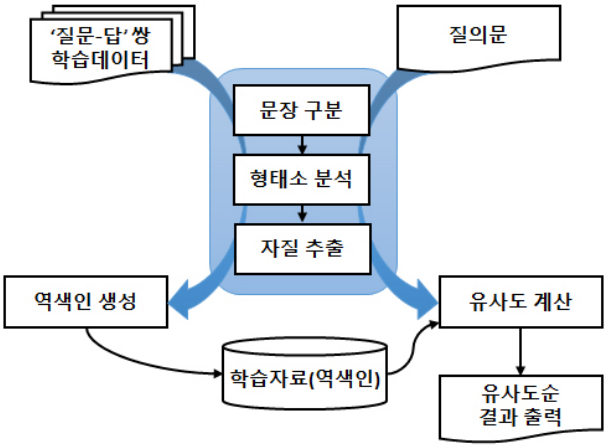

그림 1. 커뮤니티 질의응답 시스템 개요도

3.1 자질 추출

cQA의 성능은 색인과 검색을 위한 자질에 큰 영향을 받는다. 일반적으로는 명사 또는 동사를 자질로 추출하는 경우가 많지만 커뮤니티에 특화된 cQA 시스템의 경우, 해당 커뮤니티의 특성을 적극 반영하여 자질을 추출할 필요가 있다. 특히, '국립국어원 질의응답 게시판'에 등록되어 있는 질문과 답변은 특정 형태소 또는 예제에 관한 것이 많기 때문에 자질 추출 시 이러한 특성을 고려하여야 한다.

표 1. 국립국어원 '질문' 예시

질문1	잔디와 잔듸 중에 어느 것이 맞는지요?
질문2	미국산 소고기와 미국산 쇠고기 중에서 맞는 표현은 무엇입니까?
질문3	-로서와 -로써의 차이점이 궁금합니다.
질문4	~안되요가 맞나요, 아니면 ~안돼요가 맞나요?

표 1은 국립국어원 질의응답 게시판의 '질문-답' 쌍에 포함된 질문 중 일부이다. 일반적으로 질문1, 질문2와 같이 명사와 명사의 비교로 이루어진 질문이 있으며, 질문3, 질문4와 같이 부사격조사라든지 서술어의 비교를 다루는 질문도 있다. 여기서 비교의 대상이 되는 '로서', '로써', '안돼요', '안되요'는 부사격조사, 일반부사, 동사파생접미사 등 다양한 품사로 구성된다. 따라서 질문의 핵심어를 자질로 추출하기 위해서는 해당 형태소를 추출할 필요가 있다. 하지만 이러한 모든 품사의 형태소를 cQA 자질로 사용할 경우 불용어로 인하여 검색 성능 및 속도가 저하되는 결과를 초래한다. 이를 최소화하기 위하여 표 2와 같은 조건을 기준으로 자질을 추출하였다. 국립국어원 질의게시판의 특성상 형용사와 부사도 자질에 포함시켰으며, 명사류로 볼 수 있는 단어도 자질로 추출하였다.

표 2. 자질 추출 조건

구 분	조 건	예 시
자질①	1.어절 끝이 조사일 경우,	
	조사를 제거한 형태소열	안돼와
자질②	2.어절 끝이 조사가 아닐 경우,	
	1) 명사 또는 복합명사	보조용언
	2) 체언접두사 + 명사/복합명사	불완전동사
	3) 명사/복합명사 + 명사형파생접미사	활용형 양도불가성
	4) 체언접두사 + 명사/복합명사 + 명사형파생접미사	비음운화 불완전성
	5) 동사 + 명사형전성어미	보기
	6) 숫자 + 수사 숫자 + 의존명사 수사 + 의존명사 숫자 + 수사 + 의존명사	2만 2014년 수백만원 5백만원
	7) 외국어, 한자	cake, 反
기 타	8) 동사, 부사, 형용사	

자질은 어절을 기본 단위로, 표 2의 조건에 해당하는 것을 추출하였으며, 자질 내에 포함되어 있는 특수기호는 모두 제거하였다. 예를 들어, 표 1의 질문3과 같이 '-로서와', '-로써의'라는 어절은 형태소 분석 결과 '와/의'가 조사이므로 해당 조사를 떼어낸 '-로서'와 '-로써'가 자질로 추출된다. 이때, 자질 내 특수기호는 모두 제거하였기 때문에 최종적으로 '로서'와 '로써'가 자질로 추출된다.

단순히 명사, 동사 등의 형태소만을 자질로 추출하였을 때 보다 해당 커뮤니티 게시판의 특성을 고려하여 자질을 추출하였을 경우, cQA의 성능 향상을 도모할 수 있다. 이에 대한 실험 결과는 표 3에서 확인할 수 있다.

3.2 유사도 비교

질의문과 기존의 질문-답 쌍과의 유사도는 각 자질별 tf-idf 가중치[7]와 코사인 유사도를 이용하였으며, 수식은 식 1, 식 2와 같다.

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad , \quad idf_i = \log\frac{N}{n_i + 1}$$

$$dw_{i,j} = f_{i,j} \times idf_i \, , \quad qw_{i,q} = \left(0.5 + \frac{0.5 \times freq_{i,q}}{\max_l freq_{l,q}}\right) \times idf_i \quad \text{(식 1)}$$

$$CosSim(d,q) = \frac{\sum_i (dw_{i,j} \times qw_{i,q})}{\sqrt{\sum_i (dw_{i,j})^2} \times \sqrt{\sum_i (qw_{i,q})^2}} \quad \text{(식 2)}$$

N : 총 문서 수
n_i : 용어 k_i가 출현한 문서 수
$freq_{i,j}$: 문서 d_j의 용어 k_i 출현 빈도 수($1 \le j \le N$)
$f_{i,j}$: 문서 d_j에서의 용어 k_i의 정규화 빈도
$qw_{i,q}$: 질의문 내 용어 가중치
$dw_{i,j}$: 문서 내 용어 가중치
$CosSim(d,q)$: 코사인 유사도

4. 실험 및 결과

실험은 국립국어원에서 제공한 '질문-답' 쌍 5,336개를 학습 자료로 이용하였으며, 기존의 질문을 변형시킨 100개의 질의문으로 실험하였다. 형태소 분석은 세종 형태소 품사 부착 말뭉치[8]를 이용한 3단계 확률기반 형태소 분석기[9,10]를 사용하였으며, 실험 성능은 식 3과 같이 재현율(Recall)로 평가하였다.

$$\text{재현율(Recall)} = \frac{\text{올바르게 검색된 문서 수}}{\text{정답 문서 수}} \quad \text{(식 3)}$$

자질 추출은 9가지 유형으로 나누어 실험을 진행하였으며, 각 유형별 재현율은 표 3과 같다.

표 3. 자질에 따른 cQA 성능(재현율)

구 분	1-best	3-best	5-best
명사	64%	76%	81%
명사, 동사	69%	86%	89%
명사, 동사, 부사, 형용사	77%	92%	93%
자질②	67%	82%	86%
자질②, 동사	77%	91%	91%
자질②, 동사, 부사, 형용사	80%	93%	94%
자질①, 자질②	80%	93%	95%
자질①, 자질②, 동사	85%	98%	99%
자질①, 자질②, 동사 부사, 형용사	89%	99%	100%

전체적으로 자질①을 사용하였을 경우가 그렇지 않은 경우보다 성능이 우수함을 알 수 있다. 즉, 국립국어원의 게시판 특성에 따라 비교의 대상이 되는 형태소열을 자질로 추출함으로써 커뮤니티 질의응답 시스템의 성능을 높일 수 있었다.

그러나 커뮤니티 질의응답 시스템의 특성상 기존에 구축된 '질문-답' 쌍과 전혀 관련 없는 질의문이 들어올 경우 제대로 된 답변을 찾아줄 수 없는 문제점이 있다. 이는 학습 말뭉치를 증가시키면 해결될 것으로 보인다.

본 실험에 더불어, [1,2,3]의 실험과 유사하게 사용자 질의문 유형(발음, 띄어쓰기, 표기법 등)을 분석하고, 해당 유형에 맞는 검색 결과를 도출하기 위한 실험도 추가로 진행하였다. 실험 결과, 정답에 해당하는 문서는 잘 찾아냈지만 소규모 학습 말뭉치로 인하여 그 외의 비슷한 문서는 제대로 구분하지 못하였다. 따라서 본 논문에 해당 실험 결과는 제외하였다.

5. 결론

커뮤니티 질의응답 시스템은 사용자 질의문과 기존에 구축된 '질문-답' 쌍을 비교하여 유사도 순으로 '질문-답' 쌍을 검색해 주는 시스템이다.

본 논문에서는 '국립국어원의 질의응답 게시판'의 커뮤니티 질의응답 시스템에 적용 가능한 자질 추출 방법을 연구하여 적용시켰으며, 조사를 제외한 형태소의 묶음, 명사류, 동사, 형용사, 부사 등을 자질로 사용했을 때 검색 성능이 가장 높음을 확인하였다.

국립국어원의 질의응답 게시판에 본 논문에서 제시한 자질 추출 방법을 이용한 커뮤니티 질의응답 시스템을 적용시킬 경우, 사용자와 관리자 모두에게 편리한 검색 시스템이 될 것이다.

또한 이를 확장한다면, '자주 나오는 질문' 게시판에 검색이 많은 '질문-답' 쌍을 자동으로 등록하여 게시판 효율을 높일 수 있을 것으로 기대된다.

참고문헌

[1] 연종흠, 심준호, 이상구, "확장된 나이브 베이즈 분류기를 활용한 질문-답변 커뮤니티의 질문 분류", 정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제1호, pp.95-99, 2010.

[2] 정옥란, 오제환, 이은석, "Q&A 커뮤니티 기반 전문 영역 검색을 위한 프레임워크", 한국전자거래학회지 제16권 제2호, pp.143-158, 2011.

[3] 배경만, 고영중, 김종훈, "커뮤니티 기반의 질의 응답서비스(cQA)에서 질문-응답 쌍의 구조적 특징을 이용한 언어 모델 기반의 주제 분류 기법", 정보과학회논문지: 소프트웨어 및 응용 제39권 제8호, pp.664-671, 2012.

[4] E. M. Voorhees, "Overview of TREC 2003 QA Track", Prof. of TREC-12, NIST, 2003.

[5] 김한준, 김민경, 장재영, "문서 말뭉치 기반 질의응답 시스템", 한국디지털콘텐츠학회논문지 제11권 3호, pp.375-383, 2010.

[6] 허정, 류법모, 장명길, 김현기, "오픈 도메인 질의응답을 위한 검색문서 제약 및 정답유형 분류기술", 정보과학회논문지: 소프트웨어 및 응용 제39권 제2호, pp.118-132, 2012.

[7] Jones, Karen Sparck, "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation 28(1), pp.11-21, 1972.

[8] 국립국어원, "21세기 세종계획 최종 성과물(2011년 12월 수정판 2쇄)", 2011.

[9] 이재성, "한국어 형태소 분석을 위한 3단계 확률 모델", 정보과학회논문지: 소프트웨어 및 응용, 제38권 제5호, pp.257-268, 2011.

[10] 김보겸, 이다니엘, 이재성, "3단계 확률기반 형태소 분석기를 이용한 한국어 품사 태거 구축 방법", 한국컴퓨터교육학회, 동계학술대회 발표논문집, 제16권 제1호, pp.129-134, 2012.