인지과학, 제26권 제3호
Korean Journal of Cognitive Science
2015, Vol. 26, No. 3, 263~278

온라인가나다를 위한 주제 분류 기반 유사 질문 검색 시스템*

문 정 민 송 영 호 진 지 환 이 현 섭 이 현 아†
금오공과대학교 컴퓨터소프트웨어공학과

국립국어원의 온라인가나다 서비스는 한국어에 대한 질문을 등록하면 전문가가 답변을 작성하는 인터넷 서비스이다. 이러한 서비스는 유사한 질문이 자주 등록되는 문제점이 있다, 만일 새롭게 등록되는 질문과 유사한 질문을 자동으로 찾아 그 질문에 대한 답변을 등록 즉시 제공한다면, 질문자는 빠른 시간에 답변을 얻을 수 있고 서비스 관리자는 수동 답변 작성의 부담을 덜 수 있다. 본 논문에서는 온라인가나다의 특성을 분석하여 자주 질문되는 다섯 개의 주제 분류를 설정하고, 주제 분류 유사도와 함께 음소와 음절단위 수열유사도와 벡터 유사도를 결합하여 유사한 질문을 검색하는 시스템을 제안한다. 평가에서는 본 논문에서 제시한 주제 분류 정보를 활용하여 검색 정확률이 향상되는 결과를 얻었다. 최종 실험에서는 Mean Reciprocal Rank(MRR)가 0.756, 정답이 1위와 5위내에 검색될 확률은 각각 68.31%, 87.32%를 보였다.

주제어 : 질의응답시스템, 유사 질문 검색, 질문 주제 분류, 국립국어원 온라인가나다

---

* 이 논문은 문정민 외(2014)의 논문을 확장한 것임.

† 교신저자: 이현아, 금오공과대학교 컴퓨터소프트웨어공학과, 연구 분야: 자연언어처리
Tel: 054-478-7546, E-mail: halee@kumoh.ac.kr

서 론

국립국어원의 온라인가나다 서비스[1]는 한국어 어문 규범, 어법, 표준국어대사전 내용 등에 대하여 문의하는 인터넷 서비스이다. 이 서비스는 2000년 8월 경 시작하여, 현재까지 약 14만 개의 한국어 관련 지식정보 데이터를 사용자에게 제공한다. 서비스는 사용자가 게시판에 질문을 올리면 전문성을 가진 관리자가 답변을 등록하는 방식으로 운영되어 한국어에 대한 정확한 정보를 제공한다. 이와 같이 방대한 전문 데이터에 대하여 편리한 검색 시스템이 제공된다면, 사용자는 관리자의 답변 작성을 기다리지 않고 즉시 정보를 얻을 수 있고, 관리자는 유사한 질문들에 대해 동일한 답변을 반복적으로 작성하지 않게 되어 시스템 효율을 높일 수 있다.

기존의 질의응답시스템에 대한 접근은 정제되지 않은 문서를 대상으로 하거나 (Hirschman and Gaizauskas, 2001) 통계적인 기법에 지나치게 의존하여(Ittycheriah et. al., 2001; 유동현과 이현아, 2014) 국립국어원 게시판과 같이 잘 정제된 문서에 대한 정확한 답변 추천에 적합하지 않다. 이러한 문제를 극복하기 위해 개최된 국어정보 처리 시스템 경진대회[2]에서는 국립국어원 자료 질의 응답 시스템 개발 및 적용을 위한 cQA(Community Question Answering) 시스템 개발을 목적으로 '질문-답' 쌍의 학습 말뭉치를 제공하였다. 도수종 외(2014)의 연구에서는 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. 박용민 외(2014)의 연구에서는 온라인가나다의 특수성에 맞추어 명사와 동사뿐만 아니라 다양한 형태소열을 자질로 사용할 것을 제안하였다.

본 논문에서는 국립국어원 게시판의 특징을 분석하여 자주 발생하는 다섯 가지의 질문 유형을 설정하고, 입력된 질문과 축적된 질문 문서와의 벡터 유사도 점수와 수열유사도 점수와 함께, 주제 분류 점수를 고려하여 유사 질문을 찾는 방법을 제안한다. 얻어진 유사 질문에 등록된 답변을 사용자에게 제시하여 사용자가 빠른 시간 내에 원하는 답변을 얻을 수 있도록 지원한다.

---

1) http://www.korean.go.kr/front/onlineQna/onlineQnaList.do?mn_id=61

2) https://ithub.korean.go.kr/user/contest/contestIntroLastView.do

주제 분류를 활용한 질의응답 시스템

그림 1은 제안하는 시스템의 개요를 보인다. 본 논문에서는 온라인가나다의 질문들을 분석하여 얻은 다섯 분류의 질문 주제를 설정하고, 이를 유사 질문 검색에서 사용한다. 아래에서는 다섯 개의 주제 분류와 새로운 질문의 주제를 결정하는 방법, 얻어진 주제 분류와 벡터 유사도, 수열유사도를 결합하여 유사 질문을 검색하는 방법을 설명한다.

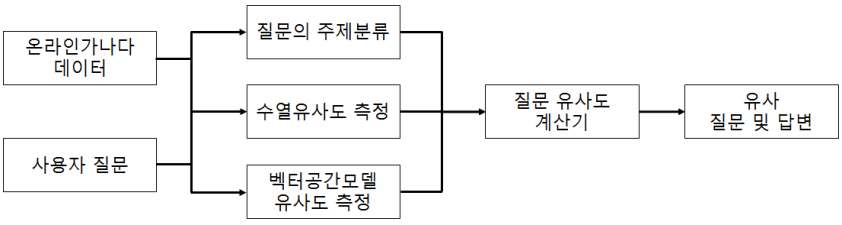


그림 1. 전체 시스템 흐름도

주제 분류를 통한 주제간 유사도 측정

아래는 온라인가나다의 질문 중 일부를 보인다. '밖에', '알다', '틀리다' 등이 세 질문에 중복적으로 나타나, 일반적인 단어 가중치에 기반한 유사 문서 검색 방식을 이용하면 세 질문이 서로 높은 유사도를 가지는 것으로 분석된다. 하지만, 이 질문들은 문법과 띄어쓰기와 같이 서로 다른 주제의 내용을 다루고 있다.

예제에서 볼 수 있듯이 온라인가나다의 질문들은 그 내용에 기반하여 표의 오른쪽과 같이 분류할 수 있으며, 500개의 질문 문서를 수동으로 분류한 결과 '문법', '띄어쓰기', '의미', '외래어', '발음'의 의미있는 다섯 분류를 얻을 수 있었다. 분류는 중복이 가능하며, 문법 78개(15.6%), 띄어쓰기 76개(15.2%), 의미 65개(13.0%), 외래어 45개(9%), 발음 38개(7.6%), 기타 217개(43.4%)이며 분류 중복된 질문은 19개(3.8%)로 나타났다.

아래에서는 본 논문에서 사용하는 각 분류와 함께 분류를 결정하기 위한 방법을 설명한다. 분류 결정에서는 키워드에 대한 가중치를 사용한다. 하나의 질문 안

질문	분류
이 **밖에**를 부사로, 이 밖의를 형용사로 보고 사용해야 할까요? 아래 예문 좀 봐 주세요 개정안에 따르면 과밀억제지역 산업단지에서는 공장 신설을 허용하고 이 밖의 지역에서는 첨단업종을 포함한 기존 공장의 증설 범위를 확대하기로 했다. 이 **밖에**로 사용하면 틀린가요?	문법
돈이 천 원**밖에** 없다. 의 **밖에**와 대문 **밖에** 누가 왔다. 의 **밖에**의 띄어쓰기에 대해 알고 싶습니다.	띄어쓰기
다음 예문이 맞는지 틀린지 알고 싶습니다. 1) 나를 알아주는 사람은 형**밖에** 없다. 2) 합격자는 나 **밖에** 없다. 제 생각으로는 1)은 맞고, 2)는 틀리지 않나요?	기타

에서 각기 다른 분류 키워드가 여러 개 발생하면 키워드별 가중치를 합하여 해당 분류에 대한 가중치로 계산한다.

문법

‘문법’ 주제의 질문을 분석해보면, 약 90%의 질문이 문법에 관련된 용어들을 포함한다. 어느 질문에서나 자주 나올 수 있는 키워드인 ‘주어’, ‘동사’, ‘형용사’, ‘부사’에는 0.3점의 가중치를 부여한다. 위의 키워드를 제외한 ‘문법’, ‘어법’, ‘관형사’, ‘파생어’, ‘조어법’, ‘접미사’, ‘의태어’, ‘양성모음’, ‘과거형’, ‘진행형’에는 1.0의 가중치를 부여한다. 이러한 문법에 관련된 키워드를 통하여 문법 질문들을 판정한다.

띄어쓰기

‘띄어쓰기’ 주제의 질문에서는 아래의 예문 [가]와 [나]와 같이 ‘띄어쓰기’, ‘띄어쓰는’, ‘붙여 쓰는’과 같은 형태소 ‘띄’, ‘붙’이 포함된 질문이 약 70%를 차지한다. 해당 단어 ‘띄어쓰기’, ‘붙여쓰기’에는 0.8점을, ‘띄어 쓰는’, ‘붙여 쓰는’에는 0.7점의 가중치를 부여한다. 또한 해당 주제에서는 문장 길이가 350자를 넘는 질문이 없어 문장의 길이가 짧다는 것을 확인할 수 있었다. 문장 길이가 350자 이하인 질문에 대해서는 가중치의 0.4점을 곱한다.

예문 [다]와 [라]를 살펴보면 키워드에 해당하는 단어를 포함하지 않지만, 띄어

쓰기에 대한 질문이다. 동일한 글자열이 각기 다른 공백을 가지는 형태(예를 들어, ‘신 빈곤층’, ‘신빈곤층’)로 하나의 질문 안에서 발생하면, 이러한 사실을 활용하여 분류를 결정할 수 있지만, 질문의 주제가 되는 내용이 아닌 부분(예를 들어 ‘써야 하나요’, ‘써야하나요’)에서의 띄어쓰기 차이로 인하여 띄어쓰기 분류가 아닌 질문이 띄어쓰기 분류로 판별되는 경우가 많아, 이에 대한 가중치는 사용하지 않는다.

[가] 부지런한지의 띄어쓰기가 궁금합니다.
[나] 승리를 위해 한 잔. 위의 예문에서 한 잔은 띄어 쓰는 게 맞나요, 붙여 쓰는 게 맞나요?
[다] 앞논, 뒤논/윗논, 아랫논 이렇게 쓰는 게 맞나요?
아니면 앞 논, 뒤 논/위 논, 아래 논 이것이 맞나요?
[라] 신 빈곤층으로 써야 하나요, 신빈곤층으로 써야하나요?

의미

아래 예문 [마]와 [바]는 단어의 뜻이나 실생활에서 쓰이는 의미를 물어보는 질문이다. 수작업으로 분류된 질문들에서 의미 분류의 문서는 ‘의미’, ‘차이’, ‘뜻’의 키워드를 약 62% 포함하고 있어, 각 키워드에 대해 0.3의 가중치를 부여한다. [사]의 예문에서 ‘뜻인가요’와 같이 명확히 의미에 대한 질문에 대해서는 1.0의 가중치를 부여한다.

[마] 미쁘다의 뜻을 알려 주세요.
[바] 이번 방학 때 뭐 할 거예요?와 이번 방학 동안 뭐 할 거예요?, 이번 휴가 때 뭐 할 거예요?와 이번 휴가 동안 뭐 할 거예요? 여기서 때와 동안의 의미에 차이가 있나요?
[사] 슈퍼마켓과 편의점은 같은 뜻인가요?

외래어

외래어 분류는 외래어 표기에 대한 질문들을 포함한다. 예문 [아]~[차]에서는 ‘로마자’, ‘외래어’, ‘표기’가 외래어 질문을 파악하는 키워드임을 알 수 있다.

[아] 맞히다를 로마자로 표기하면 음운 변화를 고려해서 machida가 되는 건가요?
[자] 미용을 목적으로 머리를 자르는 것을 외래어로 커트(cut)라고 합니까, 컷이라고 합니까?
[차] 외래어 표기 중에서, 캘린더, 카렌다 둘 다 맞나요?
[카] snow는 스노, 스노우 중 어느 걸로 쓰나요?

수동 분류된 외래어 분류의 질문 중에서 53.3%의 질문이 '로마자'와 '외래어'를 포함하고 있었으며, 이는 다른 주제에서는 잘 나오지 않는 표현들이었다. 시스템에서는 외래어 주제 가중치로 '로마자'와 '외래어'라는 단어를 포함한 질문에는 0.7점을 부여한다. 이에 반해 '표기'는 다른 주제에서도 발생 가능한 단어이다. 수동 분류된 전체 질문 중에서 10.8%의 질문이 '표기'를 포함하는 질문이었으며, '표기'가 포함된 질문 중 50%는 외래어 분류에 해당하는 질문이다. 시스템에서는 '표기'를 포함한 질문에 0.3의 가중치를 부여한다.

질문 [아]와 [자], [카]의 'machida'와 'cut', 'snow'와 같은 알파벳열이 발생하는 질문도 외래어 주제로 볼 수 있다. 시스템에서는 알파벳을 동반한 질문에는 가중치 0.4점을 부여한다. 수동 분류된 외래어 분류 질문 중에서 13.15%의 질문이 알파벳열과 함께 '쓰다'와 '적다'를 포함하고 있어, 알파벳 배열과 함께 동사 '적다'나 '쓰다'가 포함된 질문에 대해서는 외래어 주제 가중치로 0.3점을 추가 부여한다.

수동 분류된 질문 중에서 위에서 나열한 키워드를 포함하지 않은 질문이 22.22% 발생하였다. 분류되지 않은 "슈퍼마켓, 슈퍼마켓, 슈퍼마켙, 수퍼마킷 어느게 맞나요?"나 "커튼이에요, 커텐이에요?"와 같은 질문은 벡터공간모델과 수열유사도로 유사 질문을 추천할 수 있다.

발음

'발음' 주제의 질문 중 84.21%가 '발음'이라는 키워드를 포함하고 있어, '발음'을 포함한 질문은 발음 분류에 대해 0.7점의 가중치를 부여한다. 또한 '[ ]' 형태의 발음 기호 역시 84.21%의 질문이 포함하고 있어, []의 안에 2글자 이상의 한글이 존재할 경우에 가중치 0.7을 적용한다.

발음이라는 단어와 []를 모두 사용하면 발음 질문을 전부 찾을 수 있었지만, 불필요한 질문이 추가되는 결과도 발생하였다. 발음 질문의 경우 질문의 길이가 130글자 미만으로 이루어진 질문이 73.68% 존재한다. 시스템에서는 130글자 초과의 질문에 -0.3의 가중치를 적용하여 38개중 36개의 질문을 정확히 찾아 낼 수 있었으며, 발음 질문 특성상 [타]와 같이 '[]'의 출연이 다수 발생한다. 다수의 '[]' 가 존재한다면 가중치 0.3을 적용해 38개중 37개의 발음 질문을 찾아 낼 수 있다.

[타] 임진왜란 때 활약한 신립 장군이요, 읽을 때 [신닙] / [실립] 어떻게 읽나요?

하나의 질문에 대하여, 위에 제시한 각 분류의 키워드 가중치를 [0,1]의 범위로 정규화한 뒤 합산하여 분류별 가중치를 계산한다. 한 질문에 대한 각 주제별 가중치 다섯 개의 값으로 적합도 벡터를 구성하고, 입력된 질문과 질문 집합의 문서의 적합도 벡터의 유사도를 계산한다. 벡터 유사도는 항목별 차이값의 합을 최대값 5에서 뺀 뒤 다시 5로 나누어 유사도가 낮을수록 0에 가까운 값을 가지고 높을수록 1에 가까운 값을 가지도록 정규화한다.

아래는 다섯 분류에 해당하지 않은 기타 분류에 포함되는 질문의 예를 보인다. 예문 [하]의 경우 '의미'라는 단어가 나타나지만 의미 분류의 질문이 아닌 유래에 대한 질문으로 보는 것이 맞다. 이와 같은 질문들은 분류를 지정하기에는 그 빈도가 높지 않지 않고 정확하게 판별하기 어렵다. 본 논문에서는 이러한 질문들은 다섯 개의 어느 분류에도 포함되지 않는 것으로 판단하고, 벡터 유사도와 수열유사도로 유사 질문을 결정하고자 한다.

[파] 남편에 대한 호칭어, 지칭어는 무엇인가요?
[하] 짜다라는 의미에 대해서 알고 싶어서 이렇게 질문을 드리는데요, 언제부터 짜다가 인색하다와 동일시되어 사용되었는지 알고 싶습니다. 자료가 있으면 알려주시면 감사하겠습니다.

벡터 공간모델을 이용한 유사도 측정

질문간의 유사도를 측정하기 위해 질문을 단어의 벡터로 표현한다. 질문 문서 d에 포함된 단어 t의 가중치를 $w_{t,d}$로 표현했을 때, n개의 단어를 포함한 질문 문서의 벡터는 $V_d = \{w_{1,d}, w_{2,d}, \cdots, w_{n,d}\}$로 표현할 수 있다. 단어 t는 질문 문장의 어절 또는 형태소가 될 수 있다. 어절을 기준으로 벡터를 생성하면 다른 활용형으로 쓰인 단어를 유사하게 판별할 수 없다. 또한 형태소 분석기는 어절을 구분하는 기준이 띄어쓰기이다. 그래서 같은 의미의 어절이라 할지라도 띄어쓰기가 다르면 다른 것으로 판단한다. 결국 게시판에서 사용된 문장에 어절 단위로 벡터를 적용하는 것은 적합하지 않다. 본 논문에서는 문장을 형태소 분석하여 얻어진 어근을 기준으로 벡터를 구성한다.

일반적인 유사 문서 검색에서는 명사만을 추출하여 문서의 유사성을 판별하지만, 온라인가나다에 대해서는 모든 품사를 기준으로 유사성을 판별하는 것이 적합하다. 아래 보기의 질문을 보면 중요 단어는 각각 '반드시', '반듯이', '에서부터', '부터', '그때', '그 때', '이때', '이 때', '초가집', '초가'로, 명사 이외의 단어를 중요 단어로 포함하는 것을 알 수 있다.

[ㄱ] 반드시와 반듯이 중 어느 것이 맞습니까?
[ㄴ] 에서부터와 부터의 쓰임이 혼동되어 질문 올립니다.
[ㄷ] 그때와 그 때는 쓰임새가 어떻게 다른가요? 이때와 이 때는요?
[ㄹ] 초가집이라고 하지 않고, 초가라고만 해야 하지 않습니까?

단어가중치는 기본적으로 문서 d에서의 단어 t의 빈도 $tf_{t,d}$와 단어 t의 역문서빈도 idf_t의 곱 $tf_{t,d} \bullet idf_t$를 사용한다. 단어 빈도를 나타내는 $tf_{t,d}$는 정수빈도 $tf_{t,d}$ 또는 단어 빈도간 차이를 줄인 $\log(1+tf_{t,d})$, 전체 문서에서의 최대빈도로 정규화한 $tf_{t,d}/\max(tf)$ 등을 사용하여 다양하게 구할 수 있다. 역문서빈도 값인 $idf_{t,d}$는 전체 질문 문서를 Q로, 전체 문서 개수를 $|Q|$, 단어 t의 문서 빈도를 df_t이라 할 때 $\log(|Q|/df_t)$ 를 이용하여 구할 수 있다.

제안하는 시스템에서는 단어 빈도를 계산하는 3가지 방법에 역문서빈도값을 곱하여 단어 t에 대한 가중치 $w_{t,d}$를 다음과 같이 계산한다.

(1) $w_{t,d_i} = tf_{t,d_i} \cdot \log\frac{|Q|}{df_t}$ (2) $w_{t,d_i} = \log(1+tf_{t,d_i}) \cdot \log\frac{|Q|}{df_t}$

(3) $w_{t,d_i} = \frac{tf_{t,d_i}}{\max_{d_j \in Q} tf_{t,d_j}} \cdot \log\frac{|Q|}{df_t}$

사용자가 입력한 질문 문서 q와 주어진 질문 문서 d_i간의 유사도 $sim(d_j, q)$는 q의 벡터 V_q와 질문 문서 V_{d_i}에 대한 코사인유사도를 이용하여 계산한다.

$$sim(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{N} w_{i,w} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

수열유사도를 이용한 유사도 측정

온라인가나다의 질문은 다양한 형태의 띄어쓰기와 인용 등으로 인해 정확한 형태소 분석 결과를 얻기 어려운 경우가 있다. 또한 벡터 공간 모델을 이용한 유사도 측정에서는 단어의 발생 순서를 고려할 수 없다. 아래 예문에서 두 예문은 음절이나 화소의 구성이 매우 유사하다.

[ㅁ] ~ 안돼요가 맞나요, 아니면 ~안되요가 맞나요?
[ㅂ] ~하면 안 돼.인가요, ~하면 안 되인가요?

이러한 유사성을 측정하기 위해 수열유사도를 사용한다면, 단어의 발생 순서를 고려할 수 있으며 형태소 분석의 오분석으로 인한 문제도 해결할 수 있다. 시스템에서는 문장을 음절 또는 음소 단위의 수열로 변환하고, 얻어진 수열간의 유사성

을 유사 질문 검색에 활용한다.

음절 또는 음소 단위 분할에는 각각의 장단점이 있다. 음절 단위 분할은 띄어쓰기 정도만 다르거나 하나의 음절만 차이나는 질문에 대해서는 매우 좋은 성능을 보인다. 그 반대로 음절 하나가 가지는 의미가 큰 경우에는 음절 단위 분할은 좋은 성능을 낼 수 없다. 음소 단위 분할은 문장에서 오타가 발견되는 경우 좋은 성능을 보인다. 한국어 문장의 경우 음소의 탈락이나 추가와 같은 오타가 많이 발생하므로, 이러한 경우는 음소단위 분할이 적절하다. 두 가지 방법의 성능 비교를 통해 음절 혹은 음소단위 분할방법 중 하나를 시스템에서 사용한다. 유사도 계산에서는 편집거리(edit distance) 알고리즘(Levenshtein, 1965)을 사용한다. 수열유사도 계산에서는 얻어진 편집거리를 더 많은 개수를 가지는 질문 쪽의 총 음절 개수 또는 음소 개수로 나눈 값을 1에서 빼서 [0,1]의 범위의 수열유사도 값을 구한다.

질문 간 유사도 결정방법

시스템에서는 획득된 주제 분류 유사도, 벡터 유사도, 수열유사도를 합산하여, 입력 질문과 질문 집합의 문서의 유사도를 계산하고, 유사도가 가장 높은 문서를 정답으로 사용자에게 제시한다. 실험에서는 벡터 유사도와 수열유사도, 주제분류 유사도에 각기 다른 가중치를 부여하여 최적의 조합을 찾고자 한다.

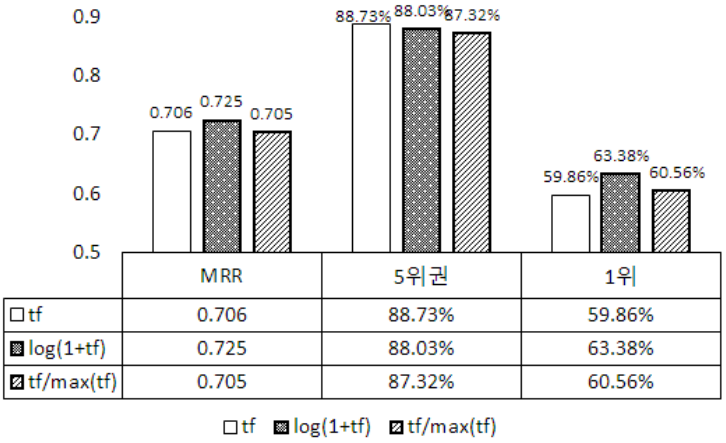
실험 및 평가

구축된 시스템에 대한 실험과 평가를 수행하였다. 주제 분류 알고리즘에는 국어정보 처리 시스템 경진대회에서 제공하는 500개의 국립국어원 질의응답 문서를 수작업으로 분석한 내용을 토대로 시스템의 정확율과 재현율을 측정하였다. 형태소 분석기로는 꼬꼬마 형태소 분석기(이동주 외, 2010)를 사용하였다. 벡터공간모델을 이용한 유사도측정과 수열유사도 측정에서는 순위를 매기기 위해 새로운 질문 문서들이 필요하다. 평가에서는 시스템 개발에 참여하지 않은 대학생 5명이 500개의 질문 문서에 유사 질문이 존재하는 실제 질문을 온라인가나다에 추출하여 142개의

테스트 질문-답변 집합을 구성하였다.

벡터 공간모델 평가

객관적 성능 평가를 위해 MRR(Mean Reciprocal Rank)(Voorhees, 1999)을 평가 척도로 이용한다. 추가적으로 시스템의 1등 문서가 정답인 경우에 대한 정확도와, 시스템의 1~5등 문서에 정답이 포함되면 정확한 결과로 보고 측정한 정확도와 1위에 책정된 확률도 비교한다. 아래 그림에서와 같이 결과에서는 $\log(1+tf_{t,d})$값을 사용하는 것이 가장 높은 MRR을 보였다.



수열유사도 평가

편집거리 알고리즘을 음절 단위와 음소 단위로 각각 나누어 실험하였다. 실험은 수열유사도와 벡터 유사도, 주제분류 유사도를 최적의 값으로 조합한 결과를 평가하였다. 측정 결과 음절 단위를 적용한 알고리즘이 0.752의 MRR을 보여 음소 단위의 0.738보다 높은 값을 가졌다. 정답 문서가 1위와 5위에 포함될 확률은 음절 단위는 각각 68.31%, 86.62%, 음소 단위 방식은 64.79%, 88.73%를 보였다.

두 가지 측정 방법에서 성능의 차이는 있지만 음절 단위 방법은 1위권 정확도가, 음소 단위 방법은 5위권 정확도가 우세함을 알 수 있다. 따라서 시스템에서는 음절 단위 방법과 음소 단위 방법, 각각 점수의 절반을 합산하여 평가점수로 활용한다. 그 결과 두 방법의 장점을 모두 얻어 전체 MRR도 향상되는 효과를 얻을 수 있었다.

	MRR	1위 정확도	5위 정확도
음절 단위	0.752	68.31%	86.62%
음소 단위	0.738	64.79%	88.73%
음절 + 음소	0.756	68.31%	87.32%

성능 종합

아래 표는 각 자질을 결합하여 사용한 경우의 평가 결과를 보인다. 실험에서는 벡터공간모델에서 최고 성능을 보인 $\log(1+tf_{t,d})$와 음절과 음소 수열유사도 평균을 사용하였다. 결과에서는 벡터 유사도와 수열유사도, 주제 분류를 모두 결합한 경우 벡터 유사도만 사용한 경우에 비해 낮은 성능을 보였다. 각 자질의 특성상 벡터 유사도는 작은 값을 가지며, 수열유사도에 비해 주제 분류 유사도가 큰 값을 가져, 각기 다른 가중치를 곱하여 각 값을 합산하였다. 그 결과 5:2:1의 가중치 조합으로 가장 높은 성능을 보였다.

	MRR	1위 정확도	5위권 정확도
벡터	0.725	63.38%	88.03%
벡터+수열	0.727	64.79%	84.51%
벡터+수열+주제분류	0.662	58.45%	77.46%
벡터+수열+주제분류(3:1:1)	0.745	66.90%	86.62%
벡터+수열+주제분류(5:2:1)	0.756	68.31%	87.32%

평가에서는 세 가지 정보를 적절한 가중치로 결합한 경우 MRR과 1위 정확도가 가장 높은 결과를 보여, 음소 기반의 수열유사도와 시스템에서 제안한 주제 분류의 사용이 성능 향상에 영향을 미치는 것을 알 수 있었다. 이에 반하여 5위권 정확도는 벡터 유사도만으로도 상당수 정답에 근접한 문서를 찾을 수 있는 것으로 나타나, 주제분류 알고리즘이 문자기반 알고리즘으로 구분할 수 없는 1, 2위 문서간의 경계를 더욱 명확히 구분 짓는 역할을 수행함을 알 수 있었다.

오류 분석에서는 질문의 길이가 긴 경우에 의한 오류, 핵심어가 실질형태소가 아닌 경우에 의한 오류가 많이 발견되었다. 질문의 길이가 길면 분류와 상관없는 키워드가 많이 포함될 수 있어 잘못된 분류로 분석되는 경우가 많았다. 핵심어가 실질형태소가 아닌 경우에 의한 오류는 온라인가나다의 특성상 체언이나 용언이 아닌 단어가 질문의 핵심으로 사용되는 경우에 발생했다. 예를 들어 "'-겁니다'는 '-것입니다'의 구어체이며 옳은 말인가요?"와 "올 것입니다의 준말이라면 당연히 올 겁니다가 맞는 것이라고 생각했는데..."에서는 '것입니다'와 '겁니다'가 질문의 핵심에 해당하는 단어이지만, 형태소 분석을 거치면 명사 '것'만 추출되어 정확한 유사 질문을 찾지 못했다. "문장을 종결할 때 '-되'로 쓰이는 경우는 없나요? 항상 '-돼' 로만 쓰이는것 같아서요. 답변 부탁드립니다."의 경우 핵심 단어인 '돼'와 '되'보다는 '답변', '문장', '종결' 등의 가중치가 높아 적합하지 않은 질문이 유사 질문으로 선택되었다. 이러한 문제를 해결하기 위해 문장 길이에 대비하여 어떤 단어가 질문의 핵심어인지 파악하는 방법과 형식형태소 중에서 질문의 핵심어를 파악하는 방법에 대한 연구를 향후에 진행할 예정이다.

결 론

본 논문에서는 벡터공간모델과 수열유사도에 질문 주제 분류를 결합한 유사 문장 검색 방법을 제안하였다. 벡터공간모델과 수열유사도를 적용한 모델에서보다 주제 분류 알고리즘을 적용하였을 때 향상된 성능을 얻을 수 있었다. 또한 음절단위와 음소단위의 수열유사도의 적용으로 시스템의 성능을 향상시킬 수 있었다. 자

동으로 주제를 분류하기 위해서 더욱 다양한 질문의 특성을 분석하여 주제 분류 알고리즘을 개선한다면 더욱 좋은 성능을 낼 것으로 기대된다. 또한, 상호직교성의 문제를 내제한 벡터 공간모델을 보완하고 단어간 상관도 개념이 추가된 일반화 벡터공간 모델(Generalized Vector Space Model) 등의 사용을 예정하고 있다.

참고문헌

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

문정민, 송영호, 진지환, 이현섭, 이현아, "주제 분류를 활용한 국립국어원 질의응답 게시판 유사 질문 검색 시스템", **제 26회 한글 및 한국어 정보처리 학술대회 발표논문집**, 201-205, 2014.

박용민, 김보겸, 이재성, "질문 특성을 고려한 커뮤니티 질의응답 시스템(cQA) 자질 추출 방법", **제 26회 한글 및 한국어 정보처리 학술대회**, 119-121, 2014.

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

이동주, 연종흠, 황인범, 이상구, "꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구", **정보과학회논문지: 컴퓨팅의 실제 및 레터, Vol. 16**, No.11, 1046-1050, 2010.

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

Ittycheriah, A., Franz, M., Zhu, W. -J. and Ratnaparkhi, A. "IBM's statistical question answering system", Proceedings 9th Text Retrieval Conference (TREC-9), 2001.

LevenshteinV. I., "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals" Soviet Physics Doklady, Vol.10, 707-710, 1965.

Voorhees, E. M., “Proceedings of the 8th Text Retrieval Conference”. TREC-8 Question Answering Track Report. 77-82, 1999.

1차원고접수 : 2015. 04. 02
1차심사완료 : 2015. 04. 22
2차원고접수 : 2015. 07. 02
2차심사완료 : 2015. 07. 22
최종게재승인 : 2015. 08. 03

(*Abstract*)

Similar Question Search System for online Q&A for the Korean Language Based on Topic Classification

Jung-Min Mun Yeong-Ho Song Ji-Hwan Jin
Hyun-Seob Lee Hyun Ah Lee
Kumoh National Institute of Technology

Online Q&A for the National Institute of the Korean Language provides expert's answers for questions about the Korean language, in which many similar questions are repeatedly posted like other Q&A boards. So, if a system automatically finds questions that are similar to a user's question, it can immediately provide users with recommendable answers to their question and prevent experts from wasting time to answer to similar questions repeatedly. In this paper, we set 5 classes of questions based on its topic which are frequently asked, and propose to classify questions to those classes. Our system searches similar questions by combining topic similarity, vector similarity and sequence similarity. Experiment shows that our method improves search correctness with topic classification. In experiment, Mean Reciprocal Rank(MRR) of our system is 0.756, and precision for the first result is 68.31% and precision for top five results is 87.32%.

Key words : Question Answering System, Similar Question Search, Topic Classification, online Q&A of the National Institute of the Korean Language