# 주제 분류를 활용한 국립국어원 질의응답 게시판 유사 질문 검색 시스템

문정민○, 송영호, 진지환, 이현섭, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
praisehim.jm@gmail.com, {keahs90, jwlghks90, leehs1017}@naver.com, halee@kumoh.ac.kr

# Similar Question Search System for Q&A board of The National Institute of the Korean Language using Topic Classification

Jung-Min Mun○, Yeong-Ho Song, Ji-Hwan Jin, Hyun-Seob Lee, Hyun-Ah Lee
Dept. of Computer Software Engineering, Kumoh National Institute of Technology.

**요 약**

국립국어원의 온라인 가나다 서비스는 한국어에 대한 다양한 질문과 정확한 답변을 제공한다. 만일 새롭게 등록되는 질문에 대해 유사한 질문을 자동으로 찾을 수 있다면, 질문자는 빠른 시간에 답변을 얻을 수 있고 서비스 관리자는 수동 답변 작성의 부담을 덜 수 있다. 본 논문에서는 국립국어원 질의응답게시판의 특성을 분석하여 질문의 주제를 6가지로 분류하고, 주제 분류 정보와 벡터 유사도, 수열 유사도를 결합하여 유사한 질문을 검색하는 시스템을 제안한다. 평가에서는 본 논문에서 제시한 주제 분류 정보를 활용한 결과 1위 정답 검색 정확률이 향상되는 결과를 얻었다. 최종 실험에서는 MRR이 0.62, 정답이 1위, 5위내에 검색될 확률은 각각 54.2%, 78.2%를 보였다.

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

## 1. 서론

국립국어원의 온라인 가나다 서비스는 한국어 어문 규범, 어법, 표준국어대사전 내용 등에 대하여 문의하는 인터넷 서비스이다. 이 서비스는 2000년 8월 경 시작하여, 현재까지 약 12만 개의 한국어 관련 지식정보 데이터를 사용자에게 제공한다. 서비스는 사용자가 게시판에 질문을 올리면 전문성을 가진 관리자가 답변을 등록하는 방식으로 운영되고 있어 한국어에 대한 정확한 정보를 보장한다. 이와 같이 방대한 전문 데이터에 대한 편리한 검색 시스템이 제공된다면, 사용자는 관리자의 답변 작성을 기다리지 않고 즉시 정보를 얻을 수 있고, 관리자는 유사한 질문들에 대해 동일한 답변을 반복 작성해야 하는 문제를 해결하여 시스템 효율을 높일 수 있다.

기존의 질의응답시스템에 대한 다양한 접근은 정제되지 않은 문서를 대상으로 하거나[1] 통계적인 기법에 지나치게 의존하여[2,3] 국립국어원 게시판과 같이 잘 정제된 문서에서의 정확한 답변 추천에 적합하지 않다. 본 논문에서는 잘 정제된 신뢰도 높은 답변 문서가 제공되는 국립국어원의 온라인 가나다 서비스의 특징을 고려한 답변 추천 시스템을 제안한다. 이 시스템에서는 기등록된 질문들을 분석하여 사용자의 질문에 적합한 답변을 추천함으로써 빠른 시간 내에 원하는 답변을 제공한다. 본 논문에서는 국립국어원 게시판의 특징을 분석하여 질문의 유형을 크게 다섯 분류로 나누고, 입력된 질문와 축적된 질문 문서와의 벡터 유사도 점수와 수열 유사도 점수와 함께, 주제 분류 점수를 고려하여 유사 질문을 찾는 방법을 제안한다.

## 2. 주제 분류를 활용한 질의응답 시스템

온라인 가나다는 관리자가 직접 답변을 등록하는 형태이기 때문에, 집단 지성의 형태로 구성되는 사용자 중심 Q&A시스템의 답변의 정보보다 신뢰성이 높아, 유사 질문 검색이 효과적으로 적용될 수 있는 분야이다.

본 논문에서는 유사 질문 검색에서 널리 사용되는 벡터 유사도와 수열 유사도와 함께, 질문의 주제 분류를 활용할 것을 제안한다. 그림 1은 제안하는 시스템의 개요를 보인다. 예를 들어 국립국어원의 질의응답 게시판에는 띄어쓰기에 대한 질문이 자주 발생하는데, 사용자의 질문이 띄어쓰기에 대한 질문이라면 띄어쓰기에 대한 다양한 답변들이 질문자에게 큰 도움이 될 수 있다. 본 논문에서는 국립국어원 질의응답게시판의 질문들에 대한 분석을 통해 다섯 분류의 질문 주제를 설정하고 유사 질문 검색에서 주제 분류를 사용하고자 한다. 아래에서는 국립국어원 질의응답의 다섯 주제와 새로운 질문의 주제를 결정하는 방법을 소개하고, 얻어진 주제 분류와 벡터 유사도, 수열 유사도를 결합하여 유사 질문을 검색하는 방법을 설명한다.

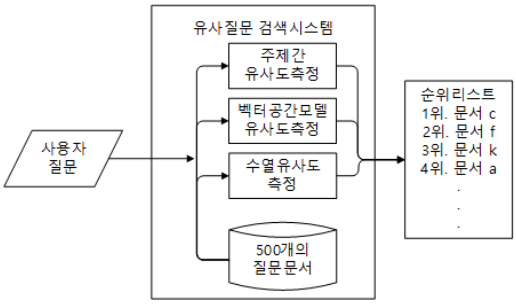


그림1. 전체 시스템 개요도

### 2.1 주제 분류를 통한 주제간 유사도 측정

국립국어원 게시판의 경우 많은 수의 질문이 띄어쓰기 등에 대해 집중되어 있는 양상을 보인다. 게시판의 특성을 분석하기 위해 500개의 질문 문서를 수동으로 분류한 결과로 '의미', '띄어쓰기', '외래어', '발음', '문법' 과 다섯 분류 어디에도 속하지 않는 '기타' 분류를 얻을 수 있었다. 주제 분류의 필요성에 대해 다음 예제에서 살펴보자.

| | |
|---|---|
| 질문 1)<br>[문법] | 이 **밖에**를 부사로, 이 밖의를 형용사로 보고 사용 해야 할까요? 아래 예문 좀 봐 주세요<br>개정안에 따르면 과밀억제지역 산업단지에서는 공장 신설을 허용하고 이 밖의 지역에서는 첨단업종을 포함한 기존 공장의 증설 범위를 확대하기로 했다. 이 **밖에**로 사용하면 틀린가요? |
| 질문 2)<br>[띄어쓰기] | 돈이 천 원**밖에** 없다.의 **밖에**와 대문 **밖에** 누가 왔다.의 **밖에**의 띄어쓰기에 대해 알고 싶습니다. |
| 질문 3)<br>[기타] | 다음 예문이 맞는지 틀린지 알고 싶습니다.<br>1) 나를 알아주는 사람은 형**밖에** 없다.<br>2) 합격자는 나 **밖에** 없다. 제 생각으로는 1)은 맞고, 2)는 틀리지 않나요? |

위 예제에 대해서 어절 '밖에'나 '알다', '틀리다' 등이 세 문서에 중복적으로 나타나므로, 일반적인 가중치에 기반한 유사 문서 검색 방식을 이용하면 세 문서가 서로 높은 유사도를 가지는 것으로 분석될 수 있다. 하지만, 이 질문들은 문법과 띄어쓰기와 같이 서로 다른 주제에 대한 내용을 다루고 있어, 유사 질문 검색에서 주제 분류가 필요함을 알 수 있다.

본 시스템에서는 입력되는 질문을 여섯가지 주제로 분류하여 유사도 계산의 정확도를 높이고자 한다. 시스템에서는 주제 적합도를 계산하여, 유사 주제의 질문이 높은 점수를 가질 수 있게 한다.

#### (가) 외래어

'외래어' 주제의 질문을 분석해보면 외래어 질문에만 나타나는 몇 가지 키워드를 발견할 수 있다. 아래의 예문 [가]~[다]에서는 '로마자', '외래어', '표기'가 외래어 질문을 파악하는 키워드임을 알 수 있다. 이는 다른 주제에서는 잘 나오지 않는 표현들이므로, 이러한 키워드에 가중치를 부여하여 외래어 주제의 질문을 분류한다. 시스템에서는 외래어 주제 가중치로 '로마자'와 '외래어'를 포함한 질문에는 0.7를, 다른 주제에서도 발생 가능한 표현인 '표기'를 포함한 질문에는 0.3의 가중치를 부여한다.

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

질문 [가]와 [나], [라]에서는 알파벳열 'machida'와 'cut', 'snow'이 발생하며, 이러한 알바벳열이 발생하는 질문도 외래어 주제로 볼 수 있다. 알파벳을 동반한 외래어 관련 질문들은 대부분 동사 '적다' 또는 '쓰다'가 포함하여, 알파벳 배열과 함께 동사 '적다'나 '쓰다'가 포함된 질문에 대해서는 외래어 주제 가중치로 0.3를 부여한다.

#### (나) 의미

아래 예문 마와 바는 단어의 뜻이나 실생활에서 쓰이는 의미를 물어보는 질문이다. 본 논문에서는 이러한 주제를 의미 주제로 본다. 수작업으로 분류된 질문들에서 의미 분류의 문서는 '의미', '차이', '뜻'의 키워드를 주로 포함하고 있어, 각 키워드에 0.3의 가중치를 준다.

[마] 미쁘다의 뜻을 알려 주세요.'
[바] 기술과 설명이 서로 반대말임을 알았습니다. 두 낱말의 개념과 차이에 관한 좀 더 쉬운 풀이와 예를 알려주세요.

#### (다) 띄어쓰기

아래의 예문에서 볼 수 있듯이, '띄어쓰기' 주제의 질문에서는 '띄어 써야', '띄어 쓰는', '붙여쓰는', '띄어쓰기'와 같이 형태소 '띄', '붙'이 포함된 질문이 많다는 것을 발견할 수 있어, 해당 단어에 0.6점의 가중치를 부여한다.

예문 [차]를 살펴보면 키워드에 해당하는 단어는 포함하지 않지만, 다양한 띄어쓰기의 경우로 질문을 구성한다. 공백문자를 이용하여 띄어쓰기 질문을 표현한 경우에는 앞에서 나온 형태소가 뒤이어 같은 패턴으로 반복되어, 이러한 사실을 활용하여 띄어쓰기 주제를 구분한다. 이 방식에는 0.2점의 가중치를 부여한다.

[사] 체육 교수라고 할 때 띄어 써야 하는지 문의드립니다.
[아] 승리를 위해 한 잔. 위의 예문에서 한 잔은 띄어 쓰는 게 맞나요, 붙여 쓰는 게 맞나요?
[자] 돈이 천 원밖에 없다.의 밖에와 대문 밖에 누가 왔다.의 밖에의 띄어쓰기에 대해 알고 싶습니다.
[차] 앞논, 뒤논/윗논, 아랫논 이렇게 쓰는 게 맞나요? 아니면 앞 논, 뒤 논/위 논, 아래 논 이것이 맞나요?

#### (라) 발음

'발음' 주제의 질문을 분석해보면, 약 90%가 '발음'이라는 키워드를 포함하고 있다. 따라서, 이 주제에서는 키워드를 '발음'으로 정하였다. 또한 발음 질문 중 '[ ]'와 같은 발음 기호를 사용한 질문도 있다. 이를 토대로 문장에 발음기호가 포함되었을 경우에 가중치를 적용하는 방법을 적용하였다. 본 시스템에서는 '발음' 키워드는 가중치 0.6, '[ ]' 발음 기호의 경우 가중치 0.5 '을 적용하였다.

#### (마) 문법

'문법' 주제의 질문을 분석해보면, 약 70%의 질문이 문법에 관련된 용어들을 포함한다. 어느 질문에나 자주 나올 수 있는 '주어'를 제외한 키워드로는 '동사', '형

용사', '부사', '조사', '어간', '어미', '능동', '주동', '사동' 등이 있다. 이러한 문법에 관련된 키워드를 통하여 문법 질문을 판정한다. 문법의 경우 키워드 별로 가중치를 부여하지 않고, 키워드의 포함 여부만 판별한 후 키워드를 포함하면 1.0의 가중치를 부여한다.

**(바) 기타**

아래의 예문과 같이 위의 모든 주제가 아닌 모든 질문은 기타 분류로 처리한다.

> [카] 짝퉁은 사전에 등재되지 않은 말로 알고 있었는데, 오늘 뉴스를 보니 계속 사용을 하더라고요.
> [타] 명확한 거짓말을 흔히 새빨간 거짓말이라고 하는데요, 거짓말과 빨간색이 어떻게 같이 붙어 쓰이게 되었는지 궁금합니다. 일본어 관용구에서 따온 말이라는 얘기도 있고, 인터넷에는 여러 설만 난무할 뿐 정확한 유래를 알기 어렵더라고요.

위의 각 방법으로 기타를 제외한 다섯 주제에 대한 주제 적합도를 계산할 수 있다. 주제별 적합도는 가중치의 합으로 계산하되 최대 1의 값을 가질 수 있게 한다. 만일 주제 적합도가 0.5보다 작은 값을 가지는 경우는 비적합 주제로 취급하여 적합도를 0으로 변경한다. 이 방식을 통해 한 문서에 대한 각 주제별 적합도 다섯 개의 값으로 적합도 배열을 구성한다. 최종적으로 입력된 질문과 질문 집합의 문서의 적합도 배열의 유사도를 계산하여, 주제 분류 유사도를 계산한다. 배열 유사도는 항목별 차이값을 활용하여 구한다. 이와 같은 방식으로 하나의 문서에 하나의 주제 분류를 결정하는 것이 아니라, 비슷한 주제 분류로 볼 수 있는 질문에 높은 점수를 주어, 주제가 유사한 질문에 높은 점수를 부여할 수 있다.

## 2.2 벡터 공간모델을 이용한 유사도 측정

질문간의 유사도를 측정하기 위해 질문을 단어의 벡터로 표현한다. 질문 문서 $d$에 포함된 단어 $t$의 가중치를 $w_{t,d}$로 표현했을 때, $n$개의 단어를 포함한 질문 문서의 벡터는 $V_d = \{w_{1,d}, w_{2,d}, \cdots, w_{n,d}\}$로 표현할 수 있다. 단어 $t$는 질문 문장의 어절 또는 형태소가 될 수 있다. 어절을 기준으로 벡터를 생성하면 다른 활용형으로 쓰인 단어를 유사하게 판별할 수 없으므로, 본 논문에서는 문장을 형태소 분석하여 얻어진 어근을 기준으로 벡터를 구성한다.

일반적인 유사 문서 검색에서는 명사만을 추출하여 문서의 유사성을 판별한다. 이에 반하여 국립국어원의 질문에 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, 명사이외의 단어를 중요 단어로 포함한다. 따라서, 시스템에서는 모든 품사를 대상으로 질문 문장의 벡터를 구성한다.

단어가중치는 기본적으로 문서 $d$에서의 단어 $t$의 빈도 $tf_{t,d}$와 단어 $t$의 역문서빈도 $idf_t$의 곱 $tf_{t,d} \cdot idf_t$를 사용한다. 단어 빈도를 나타내는 $tf_{t,d}$는 정수 빈도 $tf_{t,d}$ 또는 단어 빈도간 차이를 줄인 $\log(1+tf_{t,d})$, 문서 최대 빈도로 정규화한 $tf_{t,d}/\max(tf)$ 등을 사용하여 다양하게 구할 수 있다. 역문서빈도 값인 $idf_{t,d}$는 전체 질문 문서를 $Q$로, 전체 문서 개수를 $|Q|$로, 단어 $t$의 문서 빈도를 $df_t$이라 할 때 $\log(|Q|/df_t)$ 를 이용하여 구할 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?

제안하는 시스템에서는 단어 빈도를 계산하는 3가지 방법에 역문서빈도값을 곱하여 단어 $t$에 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 $q$와 주어진 질문 문서 $d_i$ 간의 유사도 $sim(d_j, q)$는 $q$의 벡터 $V_q$와 질문 문서 $V_{d_i}$에 대한 코사인유사도를 이용하여 계산한다.

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

## 2.3 수열유사도를 이용한 유사도 측정

국립국어원 게시판에 등록되는 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. 이러한 유사성을 측정하기 위해 수열 유사도를 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 유사성을 유사 질문 검색에 활용한다. 유사도 계산에서는 편집거리(edit distance) 알고리즘을 사용한다.

> [A] ~안되요가 맞나요, 아니면, ~ 안돼요가 맞나요?
> [B] ~하면 안 돼.인가요, ~하면 안 되인가요?

## 2.4 질문 간 유사도 결정방법

시스템에서는 획득된 주제 분류 유사도, 벡터 유사도, 수열 유사도를 합산하여, 입력 질문과 질문 집합의 문서의 유사도를 계산하고, 유사도가 가장 높은 문서를 정답으로 사용자에게 제시한다.

## 3. 실험 및 평가

구축된 시스템에 대한 실험과 평가를 수행하였다. 실험에서는 2014년도 국어 정보 처리 시스템 경진 대회에서 제공하는 500개의 국립국어원 질의응답 문서를 사용한다.

평가를 위하여 주제 분류, 벡터공간모델, 수열유사도 평가에 맞는 실험데이터를 준비하였다. 주제 분류 알고리즘에는 500개의 질문 문서를 수작업으로 분석한 내용을 토대로 시스템의 정밀도와 재현율을 측정하였다. 벡터공간모델을 이용한 유사도측정과 수열유사도 측정에서는 순위를 매기기 위해 새로운 질문 문서들이 필요하다. 본 논문에서는 500개의 질문 문서 외에 200개의 추가적인 질문 문서를 구성하고, $MRR$(Mean Reciprocal Rank)을 이용해 성능을 평가한다.

평가를 위한 전체 질문 문서를 $Q$, 전체 문서 개수를 $|Q|$, 시스템에서 $i$번째 문서의 유사순위를 $rank_i$라고 할 때, $MRR$은 다음 식을 따른다.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

### 3.1 주제 분류 평가

시스템에서 제공하는 분서 중에서 외래어 질문은 40개, 의미 질문은 56개, 띄어쓰기 질문은 79개, 발음 질문은 35개, 문법은 90개였다. 질문 문서를 각 주제로 분류하는 알고리즘의 정밀도(precision)와 재현율(recall)은 아래와 같았다.

| | 외래어 | 의미 | 띄어쓰기 | 발음 | 문법 |
|---|---|---|---|---|---|
| 정확도 | 88.89% | 51.61% | 82.21% | 77.77% | 72.83% |
| 재현률 | 80.00% | 57.14% | 67.09% | 100.0% | 77.44% |

### 3.2 벡터 공간모델 평가

객관적 성능 평가를 위해 $MRR$(Mean reciprocal rank)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 그림은 그 결과를 보인다.

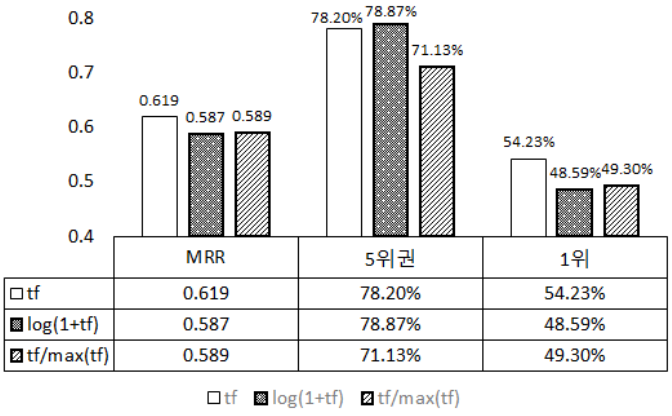


그림 2. $tf$ 값에 따른 유사도 평가

결과에서는 정수형 $tf$값을 그대로 사용하는 것이 가장 높은 $MRR$(0.619)을 보였다. 타 응용분야에서는 단어의 $tf$값이 2배 높다고 해서 해당 단어가 2배 중요한 것은 아니기 때문에, 식 (2),(3)과 같은 정규화과정을 거친다. 하지만 본 시스템의 질문 문서의 경우 문서의 길이가 짧은 것이 대부분이고 질문문서에서 단어의 $tf$값 자체가 중요도에 큰 영향을 미친다는 것을 확인하였다.

### 3.3 수열 유사도 평가

편집거리 알고리즘을 음절 단위와 음소 단위로 각각 나누어 실험하였다. 실험은 벡터공간 모델유사도와 주제분류 유사도를 모두 적용하여 시스템 전체의 $MRR$차이를 측정하였다. 측정결과 음소 단위를 적용한 알고리즘이 음절 단위보다 $MRR$이 0.01이 높아 큰 차이를 보이지 않지만, 정답 문서가 1위에 포함 될 확률은 약 5%p 높게 측정되어, 음소 단위가 보다 유용함을 확인할 수 있다.

| | MRR | 1위 정확도 |
|---|---|---|
| 음절 단위 | 0.61 | 49.3% |
| 음소 단위 | 0.62 | 54.2% |

### 3.4 성능 종합

아래 표는 각 자질을 사용한 경우의 평가 결과를 보인다. 세 가지 정보를 모두 이용한 경우 MRR과 1위 정확도가 가장 높은 결과를 보여, 음소 기반의 수열유사도와 시스템에서 제안한 주제 분류의 사용이 성능 향상에 영향을 미치는 것을 알 수 있었다.

| | MRR | 1위 정확도 | 5위권 정확도 |
|---|---|---|---|
| 벡터 | 0.60 | 49.3% | 77.5% |
| 벡터+수열 | 0.61 | 51.4% | 79.6% |
| 벡터+수열+주제분류 | 0.62 | 54.2% | 78.2% |

## 4. 결론 및 향후 연구

본 논문에서 국립국어원 질의응답게시판의 특성을 분석하여 질문의 주제를 6가지로 분류하고, 주제 분류 정보와 벡터 유사도, 수열 유사도를 결합하여 유사한 질문을 검색하는 시스템을 제안하였다. 향후 연구로는, 다양한 사용자 질문 분석을 통한 주제 분류 알고리즘을 개선을 예정하고 있다. 또한 상호직교성의 문제를 내제한 벡터 공간모델을 보완하고 단어간 상관도 개념이 추가된 일반화 벡터공간 모델(Generalized Vector Space Model) 등의 사용 등도 필요하다.

## 참고문헌

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

[2] Ittycheriah, A., Franz, M., Zhu, W.-J. and Ratnaparkhi, A. “IBM's statistical question

answering system", Proceedings 9th Text Retrieval Conference (TREC-9), 2001.
[3] 유동현, 이현아, "Q&A 문서의 검색 결과 요약을 활용한 질의응답 시스템", 정보처리학회지 3(4), 2014.
[4] 이동주, 연종흠, 황인범, 이상구, "꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구", 정보과학회논문지: 컴퓨팅의 실제 및 레터 (Journal of KIISE: Computing Practices and Letters), Volume 16, No.11, Page 1046-1050, 2010.