실시간 검색어를 이용한 주제어 기반의 질의응답시스템

송일현○, 강상우, 서정연*
서강대학교 컴퓨터공학과, *서강대학교 컴퓨터공학과/바이오융합기술협동과정
{ihsong, swkang, seojy}@sogang.ac.kr

Topic based Question-Answering System using Real-Time Search Terms

Il-Hyeon Song○, Sang-woo Kang, Jung-Yun Seo*
Department of Computer Science and Engineering, Sogang University
*Department of Computer Science and Engineering, and Interdisciplinary Program of Integrated Biotechnology, Sogang University

요 약

본 논문에서는 실시간 검색어를 이용한 주제어 기반의 질의응답 시스템을 제안한다. 제안 시스템은 주제어로 사용자의 질의 범위를 제한함으로써 질의과정에서 발생할 수 있는 오류의 감소를 기대할 수 있다. 제안 시스템은 주제어 기반의 질의응답을 수행하기 위해 검색대상문서 색인, 질의유형결정, 검색결과의 순위화 과정을 거친다. 제안한 방법으로 기준시스템에 비해 P@5에서 질의유형별 평균 69%의 성능향상을 얻었다.

주제어 : 질의응답 시스템, 질의유형 분류

1. 서론

질의응답 시스템(Question-Answering System)은 대상 문서집합 내에서 주어진 질의에 대한 정답을 찾아내는 시스템이다[1]. 기존의 질의 응답시스템은 TREC (Text REtrieval Conference)을 중심으로 많은 연구가 진행되어 왔으며, 개방영역에서 정답을 찾기 위해 대량의 문서집합을 사용한다[2].

최근 스마트폰과 같은 모바일 환경에서 텍스트 입력의 한계를 보완하기 위해 음성검색시스템과 같은 음성입력에 기반을 둔 응용들이 확대되고 있으며 질의 응답시스템에도 음성기반의 질의 입력방식이 적용될 것으로 기대된다. 하지만 개방영역의 질의응답은 질의의 어휘 수에 제한이 없기 때문에 음성인식에 한계가 있어 실용적 성능을 기대하기 힘들다.

실시간 검색어는 포털사이트에서 현재 가장 많이 입력되고 있는 검색어로 다수의 인터넷 사용자들의 관심 주제를 반영한다. 따라서 질의응답시스템이 실시간 검색어에 대한 질의응답을 수행한다면, 응답 성능을 유지하면서도 다수의 사용자의 관심 분야에 대응할 수 있다.

이를 위해 본 연구는 주제어를 이용한 질의응답시스템을 제안한다. 제안 시스템은 포털사이트의 실시간 검색어로 주제어를 결정하고 이를 이용하여 질의응답 시스템의 성능을 향상한다.

2. 주제어 기반의 질의응답 시스템

그림 1은 제안하는 시스템의 흐름을 보여준다. 시스템의 흐름은 크게 두 가지 단계로 나누어지는데, 전처리 과정에서는 시스템이 웹에서 주제어 목록을 가져오고 주제어와 관련된 기사를 수집 후 색인화 하여 저장한다. 질의응답 과정에서 사용자는 시스템이 제공하는 주제어 목록을 통해 주제어를 선택한다. 시스템은 사용자가 선택한 주제어와 관련된 머리기사의 목록을 제공하고, 사용자는 주제어와 관련된 질의를 입력한다. 제공된 머리기사 목록은 사용자의 질의내용을 구체화 하는데 도움을 줄 수 있으며 사용자가 시스템 내에 정답이 존재하는 문서와 관련된 질의를 하도록 유도하여 질의응답의 정확도를 높인다.

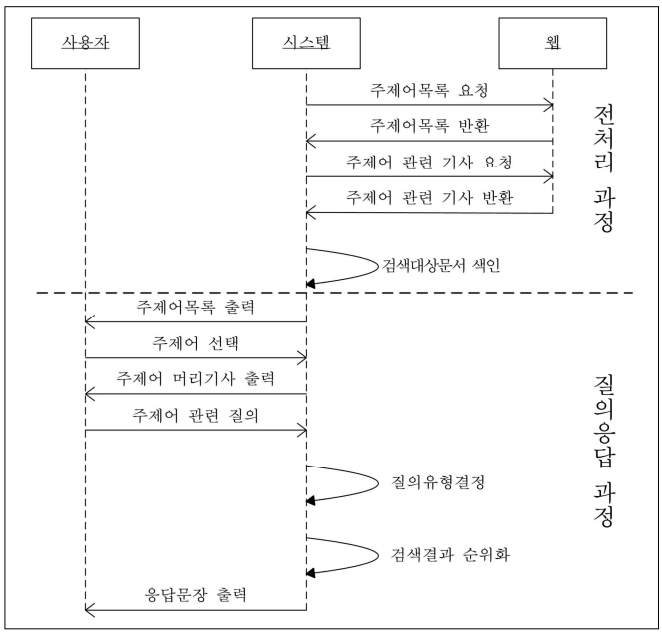

그림 1. 제안시스템의 흐름

본 논문의 2장은 제안 시스템의 주요요소를 기술한다. 2.1절 질의유형결정에서는 입력질의를 분석하여 질의의 정답이 어떤 유형인지를 파악하는 과정을 기술한다. 2.2

절 검색대상문서 색인에서는 웹에서 수집한 주제어 관련 기사 문장들이 어떤 정답유형의 정보를 갖고 있는지 결정하여 데이터베이스에 저장하는 과정을 기술한다. 2.3절 검색결과 순위화에서는 질의대상유형-정답유형의 일치여부와 질의-검색대상문장의 유사도를 계산을 통해 정답후보문장의 순위를 결정하는 과정을 다룬다.

2.1 질의유형결정

질의유형결정 단계는 주제어 관련 문서들의 검색 범위를 제한하기 위하여 사용자 질의가 어떤 유형인지를 파악하는 단계이다. 질의유형은 질의의 정답대상의 종류에 따라 정의한다.

기존 연구에서는 질의유형을 '사람', '장소', '날짜/시간', '수', '조직', '물체/사물', '이유', '방법'의 범주로 정의하였다[3]. 하지만 위 분류는 '장소', '조직', '물체/사물' 유형들을 구분하기 모호한 경우가 빈번히 발생한다. 예를 들어 "박정현의 '나 가거든' 어디서 볼 수 있지?"질의의 유형은 조직('MBC')이나 물체/사물('나는 가수다') 중에 하나로 결정하기 모호하다. 따라서 본 연구에서는 '장소', '조직', '물체/사물'과 같은 범주를 하나의 '대상(Object)'범주로 통합해 질의유형을 표 1과 같이 적용한다. '방법'유형에 대한 질의처리는 후속 연구에서 다룬다.

표 1. 질의유형

질의유형	질의의 예
사람(Person)	'나가수' 1위는 누가 했지?
날짜(Time)	'나가수' 중간 평가는 언제하지?
숫자(Number)	박정현이 중간평가에서 몇 위 했어?
대상(Object)	'나가수'가 뭐지?
이유(Reason)	박정현이 왜 검색어 1위야?

본 연구에서는 통계적 방법을 사용하여 질의유형을 결정한다. 질의유형결정은 입력된 질의를 표 1의 5가지 유형 중 하나의 유형으로 분류하는 문제로 정의되며 분류를 위한 알고리즘으로 지지벡터기계(Support Vector Machine)를 사용한다[4]. 지지벡터기계는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik에 의해 소개된 학습기법으로 두 개의 범주의 구성 데이터들을 가장 잘 분리해 낼 수 있는 결정면을 찾는 모델이다. 지지벡터기계는 이진 분류기이기 때문에 주어진 질의가 특정 유형에 속하는지 여부를 가려낼 수는 있지만, 여러 가지의 유형 중에서 적절한 질의유형이 무엇인지를 가려낼 수 없다. 따라서 본 논문에서는 지지벡터기계 이진 분류기를 이용하여 5개의 질의 유형 중에서 입력된 질의의 유형을 결정하기 위해 one-versus-all 방법으로 분류를 수행한다[5]. 분류에 사용되는 자질은 입력된 질의 문장의 형태소 uni-gram을 사용하고, 카이제곱 통계량을 이용하여 적합한 자질의 수를 결정한다[6].

2.2 검색대상문서 색인

제안 시스템의 검색대상문서를 수집하기 위해 실시간 검색어를 검색엔진의 질의로 요청하여 주제어와 관련된 기사를 수집한다. 질의에 대한 정답을 찾기 위해 시스템이 검색하는 대상은 수집된 기사의 문장들이다. 검색대상문서 색인단계에서는 주제어와 관련된 기사의 문장들의 정답유형을 결정하여 색인화한다. 정답유형은 기사의 문장이 포함하고 있는 정답의 종류에 따라 '사람', '날짜', '숫자', '대상', '이유'로 구분된다. '사람'은 사람의 이름, '날짜'는 날짜와 시간, '숫자'는 숫자와 관련된 표현, '대상'은 장소, 조직, 물체 및 사물의 이름, '이유'는 인과관계 표현을 나타낸다.

'사람', '대상'유형은 사전기반으로, 정형화된 규칙을 도출하기 용이한 '날짜'와 '숫자'유형은 정규표현식으로 패턴일치방식을 적용하여 정답 유형을 추출한다. 또한 인과관계를 나타내는 문장 유형과 관련된 황화상의 연구를 참고하여 인과관계를 추출하기 위한 규칙집합을 생성한다[7]. 규칙집합의 예로 "-어서", "-니까"와 같은 연결어미 사용, "때문"과 같은 의존명사의 사용을 들 수 있다. 표 2는 주제어 "박정현"에 관련된 정답유형의 예를 보여준다.

표 2. 검색대상문장 색인 예

검색대상 문장	한편, 이날 *박정현*이 부른 '*나 가거든*'은 *2001*년 방영된 드라마 <*명성황후*>의 OST로 세계적 소프라노 *조수미*가 불러 큰 이슈를 모았으며 당시 *30만장*의 판매고를 기록한 히트곡이다.
정답유형 정보	사람 ('박정현', '조수미'), 대상 ('명성황후', '나 가거든'), 날짜 ('2001년'), 숫자 ('30만장')

2.3 검색결과 순위화

정답후보문장은 색인되어 저장된 문장들 중에 사용자가 선택한 주제어와 동일한 주제어를 가진 문장들이다. 정답 후보문장들 간의 순위화를 위해 식 (1)을 적용한다.

$$score(d_i) = \alpha \cdot sim(d_i,q) + (1-\alpha) \cdot f(d_i,q) \quad (1)$$

식 (1)에서 q는 질의 문장, d_i는 정답후보문장, α는 가중치를 나타낸다. 이진함수 f는 질의문장의 질의유형과, 정답후보문장의 정답유형이 일치할 경우 1, 그 외의 경우 0이 되며, 유사도 함수 $sim(d_i,q)$는 질의 문장과, 정답후보 문장 간의 유사도를 계산한다.

$$sim(A,B) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}} \quad (2)$$

식 (2)에서 문장 A, B에 대한 유사도 계산 함수는 문장에 각각 출연한 실질 형태소들을 이진가중치(binary weighting)기법에 의하여 벡터로 표현한다[8].

α는 질의문장-정답후보문장 유사도와 질의유형-정답유형 일치 중 어디에 비중을 두는지를 결정하는 가중치이다. α를 0.5 이하로 설정할 경우, 질의유형-정답유형이 일치하는 문서들이 그렇지 않은 문서보다 항상 높은 점수를 갖는다. 따라서 정답후보문장에 질의유형과 일치하는 정답유형이 없다면 질의에 대한 정답이 될 수 없다.

하지만 정답유형 중'대상', '이유'유형의 경우 정답후보문장에 질의유형과 일치하는 유형이 없어도 질의에 대한 정답이 될 수 있다. 예를 들어 "박정현의 인기 비결은 뭐야?"질의는 '대상'질의유형으로 분류되지만 질의의 정답이 되는 "박정현의 인기비결은 뛰어난 가창력이다"정답후보문장은 '대상'을 정답유형으로 포함하지 않는다. 질의유형-정답유형간의 불일치를 해결하기 위해 '대상', '이유' 질의유형에는 α를 0.5이상으로 설정하여 정답유형이 존재하지 않는 문서에서 유사도가 높은 문장이 높은 점수를 갖도록 한다. 성능이 최대가 되는 α는 실험을 통해 결정한다.

3. 실험 및 평가

3.1 말뭉치

질의유형분류 학습을 위한 질의 수집을 위해 주제어를 인물의 범주로 설정하고, 온라인 지식 검색 서비스에서 인물과 관련된 800개의 질의를 수집하였다. 또한 주제어 관련 기사 수집을 위해 3명의 인물을 주제어로 결정하고 주제어 당 20개의 기사를 수집하였다. 수집된 기사는 총 495문장이다. 평가를 위한 질의 수집을 위해 일반인을 대상으로 주제어 관련 기사의 머리글을 보여주고 주제어와 관련된 질의를 190개 수집하였다. 또한 주제어 관련 기사 중 수집한 질문에 답이 될 수 있는 문장을 함께 수집하였다. 190개의 질의 문장 중 수집된 기사 내에 답이 존재하는 질의는 158개이며, 질의유형분포는 그림 2와 같다.

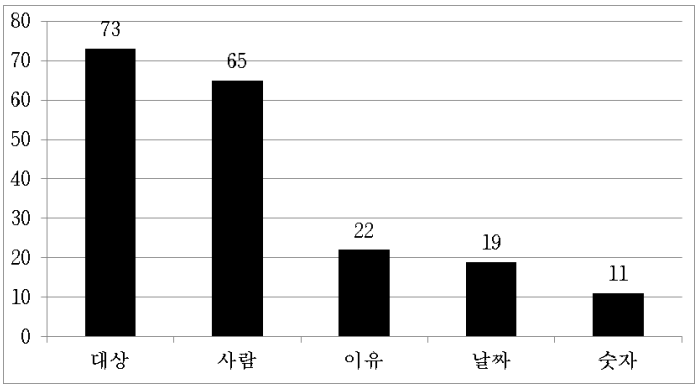

그림 2. 질의유형분포

3.2 평가방법

시스템의 질의유형분류와 정답유형 정보추출의 성능을 평가하기 위해 정확률, 재현율, F1-평가치를 사용하였고, 검색결과의 순위화 성능을 평가하기위해 MRR(Mean Reciprocal Rank), P@n(Precision at n)을 이용하였다.

표 3. 시스템 평가 방법

평가치	설명
정확률(P)	$\frac{\text{시스템이 정확하게 인식한 범주의 개수}}{\text{시스템이 인식한 범주의 개수}}$
재현율(R)	$\frac{\text{시스템이 정확하게 인식한 범주의 개수}}{\text{전체 범주의 개수}}$
$F1-$평가치	$\frac{2PR}{P+R}$
MRR	$\frac{1}{\lvert Q\rvert}\sum_{i=1}^{\lvert Q\rvert}\frac{1}{rank_i}$
$P@n$	$\frac{\text{후보 } n\text{위 안에 정답이 있는 질의 개수}}{\text{전체 질의 개수}}$

3.3 실험결과

- 질의유형분류

시스템의 질의 처리 모듈은 사용자에 의해 입력된 질의를 표 1의 범주 중 하나로 분류한다. 지지벡터기계[9]에 의해 질의문장이 각 정답유형으로 결정될 확률을 계산하여 가장 높은 확률 값을 갖는 유형을 질의문의 정답유형으로 결정하였다. 실험결과 카이제곱 통계량 상위 200개의 형태소 uni-gram을 자질로 사용하였을 때 정확도는 96.3% (183/190)이었고, 질의유형별 정확률과 재현율, F1-측정치는 표 4와 같다.

표 4. 질의유형분류 결과

질의유형	정확률	재현율	F-1
사람	0.984	0.953	0.968
날짜	0.944	0.895	0.918
숫자	0.917	0.917	0.916
대상	0.947	0.986	0.966
이유	1.000	1.000	1.000

실험 결과 대부분의 질의유형에서 90%이상의 높은 분류정확도를 갖는데, 이는 수집된 대부분의 질의문장이 의문사만으로도 질의유형판별이 가능한 단순한 질의였기 때문이다. 오분류된 질문은 "박정현이 중간점검 때 같이 즉흥곡을 부른 가수들을 알려주세요"와 같이 의문사를 포함하지 않는 문장들이었다. 이와 같은 문제를 해결하기 위해서는 질의문에 포함된 단어인 "가수"의 의미정보를 이용하여 질의유형이 '사람'으로 분류되도록 하는 것과 같이 단어의 의미정보를 질의유형분류의 추가자질로 사용하는 접근방법이 필요하다.

- 정답유형 추출

수집된 495개의 문장에 대해 정답유형 정보를 추출하는 실험결과는 표 5와 같다.

표 5. 정답유형 추출 결과

정답유형	정확률	재현율	F-1
날짜	0.928	1.000.	0.963
숫자	0.867	0.949	0.906
이유	0.523	0.719	0.605

'사람', '대상'유형은 수집된 기사의 고유명사를 추출하여 사전을 작성하였다. '날짜'와 '숫자'유형은 신문기사에서 정형화 되어있기 때문에 높은 정확도를 갖지만, '이유'유형은 패턴을 정형화하기 어렵기 때문에 추후 문장의 인과관계 추출을 위한 연구가 필요하다.

- 검색결과 순위화

정답이 존재하는 158개의 문장에 대해 검색결과 순위화 성능을 측정하였다. 질의와 수집된 문장 간의 단순 유사도 계산(식 (2))을 했을 때를 기본 시스템으로 설정하고, '정답유형 추출'에서의 결과로 본 연구에서 제안한 방식(식 (1))으로 순위와를 했을 때의 성능을 P@5, MRR로 평가하였다. 식 (1)의 α는 순위화 성능이 최대가 되는 값을 실험을 통해 결정하였는데, '대상'유형은 0.9, '이유'유형은 0.79였다.

표 6. 검색결과 순위화 성능

유형	P@5			MRR		
	기본	제안	향상율	기본	제안	향상율
사람	0.474	0.489	3%	0.311	0.321	3%
날짜	0.466	0.932	100%	0.398	0.730	83%
숫자	0.200	0.398	99%	0.104	0.468	350%
대상	0.613	0.633	3%	0.403	0.432	7%
이유	0.264	0.632	139%	0.168	0.374	123%

표 6에서와 같이 제안 시스템은 P@5에서 유형별 평균 69%, MRR에서 유형별 평균 113%의 성능향상이 있었다. P@5 평가치가 '사람', '대상'유형에 대해서는 3%의 낮은 향상율을 보인 반면, '날짜', '숫자', '이유' 유형에서 기본시스템 대비 평균 113%의 향상율을 보였는데, '날짜', '숫자'유형의 경우 높은 정답유형 추출 성능과 함께 질의유형-정답유형 일치 문장으로 검색대상이 축소되어 큰 성능향상을 가져왔다. 하지만 '사람', '대상'유형의 경우 인명과 대상의 고유명사를 포함하고 있는 후보 문장이 전체 검색대상문서에서 차지하는 비중이 높아 검색범위를 줄이지 못했으며 질의문장과 정답후보문장의 유사도 비교만으로는 정답문서를 찾아내지 못하는 한계가 있었다.

4. 결론

본 논문에서는 실시간 검색어를 이용한 주제어 기반의 질의응답 시스템을 제안했다. 제안하는 시스템은 질의응답 대상을 주제어로 제한함으로써 사용자의 질의 범위를 제한시켜 질의응답 성능을 향상시켰다.

제안 시스템은 질의에서 질의유형을 분류하여 검색대상 범위를 제한하였고 정답 후보 순위화 과정에서 질의유형의 일치와 질의문장과 정답후보문장간의 유사도 결과를 가중합하였다. 실험을 통하여 P@5 유형별 평균 69%, MRR 유형별 평균 113%의 성능향상이 있었다. 특히 날짜, 숫자, 이유 유형에서 기본 시스템에 비해 113%의 P@5 평가치 향상이 있었다.

향후 연구로는 제안한 시스템의 성능을 향상시키기 위해서 검색대상 문서에서 '이유'와 '방법'유형을 결정하기 위한 특징 추출 방법에 대한 연구와 문장 순위화 성능을 높이기 위하여 질의에 포함된 의미정보를 활용하는 방법에 대한 연구가 필요하다.

* 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업(NIPA-2011-(C1090-1131-0008))의 연구결과로 수행되었음.

참고문헌

[1] L. Hirschman and R. Gaizauskas, "Natural Language Question Answering: the View from here," Natural Language Engineering, vol. 7, no. 4, pp. 275-300, 2002.

[2] E. M. Voorhees and D. M. Tice, "Building a Question Answering Test Collection," In Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 200-207, 2000.

[3] 윤성희, 백선욱, "단어 의미 정보를 활용하는 이용자 자연어 질의 유형의 효율적 분류," 정보관리학회지, 제21권 제4호, pp. 251-263, 2004.

[4] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[5] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition,"In Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 77-82, 1994.

[6] 김학수, 안영훈, 서정연, "한국어 질의응답시스템을 위한 지지벡터기계 기반의 질의유형분류기," 정보과학회논문지(B), 제30권, 제5·6호, pp. 466-475, 2003.

[7] 임채훈, "인과관계의 형성과정과 국어의 연결어미," 담화·인지언어학회 겨울학술대회, 151-164. 2006.

[8] 김한경, 나휘동, 이금희, 이종혁, "문장구조 유사도와

단어 유사도를 이용한 클러스터링 기반의 통계기계 번역,"정보과학회논문지(B), 제37권, 제4호, pp. 297-304, 2010.
[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/