최상급 단서 어휘를 이용한 질의-응답시스템

Question-Answering System using the Superlative Words

박희근, 오수현, 안영민, 서영훈
충북대학교 컴퓨터공학과

Park Hee-Geun, Oh Su-Hyun, Ahn Young-Min,
Seo Young-Hoon
Chungbuk National Univ.

요약

본 논문에서는 최상급 질의에 대한 정답을 추출하는 질의-응답시스템에 대해 기술한다. 최상급 질의란 “가장”, “제일”, “처음”, “최고의”, “최대의”, “최소의”, “최초로”, “최초의” 등의 최상급 단서 어휘를 포함하고 있는 질의를 말한다. 최상급 질의는 4가지 주요 성분-최상급 단서 어휘, 정답유형, 지역정보, 용언-과 기타 문장 성분으로 구성된다. 이 중 최상급 단서 어휘는 자신이 수식하는 용언을 반드시 필요로 하느냐에 따라 두 가지 유형으로 나뉘며, 이는 정답 추출을 위한 필수요소를 결정하는 기준이 된다. 모든 최상급 질의에 대해 최상급 단서 어휘, 정답유형, 지역정보는 정답을 추출하기 위한 필수요소이지만, 용언은 최상급 단서 어휘의 유형에 따라 필수요소로 결정된다. 본 논문의 시스템은 최상급 질의 분석을 통하여 정답 추출을 위한 필수요소를 찾고, 이를 이용하여 후보 문서와 후보 문장을 검색한 후, 정답을 추출한다. 실험 결과 최상급 질의에 대한 높은 정확률과 재현율을 보였다.

Abstract

In this paper, we describe a question-answering system which extracts answers for the superlative questions which include the superlative words such as "the most", "the best", "the first", "the largest", "the least", and so on. The superlative questions are composed of four main components and others. Four main components are the superlative word, answer type, regional information, and a verb modified by the superlative word. We classify the superlative words into two types as to whether the verb has to be needed to be a question or not. The superlative word, answer type and regional information are essential elements to extract answer for all superlative questions. But the verb may be an essential element by the type of superlative word. Our system analyzes input question, and finds four main components of the superlative question. Also, our system searches relative documents and candidate sentences using them, and extracts answers from candidate sentences. Empirical result shows that our system has high precision and high recall for the superlative questions.

I. 서 론

정보검색(IR) 시스템은 사용자가 입력한 키워드나 자연어 질의를 분석하여, 분석된 결과와 관련된 문서를 순위화하여 제공하는 시스템이고, 질의-응답(QA) 시스템은 사용자로부터 다양한 자연어 질의를 입력으로 받아 분석하여 사용자 요구에 적합한 정답을 문서가 아닌 단어나 구 혹은 문장의 형태로 제공하는 시스템이다. 정보검색 시스템은 사용자가 결과로 제시된 문서에서 자신의 요구에 적합한 정답의 포함 유무를 판단해야 하지만, 질의-응답 시스템은 사용자 요구에 적합한 정답을 제시하여 사용자의 편의를 제공하기 때문에 이에 대한 요구가 증가하고 있는 추세이다[1].

일반적인 질의-응답 시스템은 질의문으로부터 질의 유형이나 정답 유형, 키워드 등을 추출한다. 그 다음으로 기존의 정보검색 방법을 이용하여 질의문과 유사한 문서를 검색하고, 검색된 문서 내에서 다시 정답을 포함할 가능성이 높은 문단을 추출한다. 마지막으로 정답 유형과 관련이 있는 개체를 정답으로 추출한다[2][5].

이러한 일반적인 질의-응답 시스템은 질의문의 키워드로 어떠한 개체를 추출하느냐에 따라 정답 유형과 후보들이 달라지고 결국에는 사용자가 원하지 않는 정답이 제시될 수도 있다[1][4].

현재 서비스되고 있는 QA 시스템 중에는 “가장”, “제일”, “처음”, “최고의”, “최대의”, “최소의”, “최초로”, “최초의” 등의 최상급 단서 어휘를 포함하는 질의인 최상급 질의에 대한 정답을 추출하는 일반적인 QA 방법을 사용하는 시스템[6]과 최상급 질의에 대한 정답 정보를 가진 문서들을 수작업으로 구축하

여 정답을 제시하는 시스템[7]이 있다. [6]의 경우에는 최상급 질의에 대한 별도의 처리방법을 사용하지 않기 때문에 제시된 정답의 정확도가 낮아지게 된다. [7]은 최상급 질의에 대한 정답 정보를 문서들을 수작업으로 구축하였으므로, 정답의 정확도가 높다는 장점이 있지만 정보 구축에 너무 많은 시간과 인력이 소요되고 새로운 정보에 대한 변화에 빠르게 대처하지 못한다는 단점이 있다.

이에 본 논문에서는 최상급 질의의 형태가 크게 두 그룹으로 정형화 되어 있다는 것을 연구를 통하여 알게 되었고, 그 특성을 이용하여 최상급 단서 어휘, 정답유형, 지역정보 등을 추출하여 사용자의 요구에 적합한 정답을 추출하는 시스템에 대해서 기술하였다. 본 논문에서 제안하는 시스템은 [6]과 비교하였을 때 일반적인 질의-응답 시스템 내부에 최상급 질의에 대한 별도처리 모듈이 되기 때문에 정답의 정확도가 높아지고, [7]과 비교하였을 때 정답 정보들을 수작업으로 구축하지 않고 일반적인 질의-응답 시스템의 방식을 확장하기 때문에 시간과 인력의 소요를 줄일 수 있다. 실제로 실험을 통하여 본 논문에서 제안하는 시스템을 일반적인 질의-응답 시스템내의 한 모듈로 활용할 경우에 전체적인 성능이 향상되었음을 보여준다.

Ⅱ. 최상급 질의와 최상급 단서 어휘

일반적으로 질의-응답 시스템은 사용자의 자연언어 질의를 입력으로 받는다. 이러한 자연언어 질의가 최상급 단서 어휘를 포함하고 있을 때의 질의를 최상급 질의라고 한다. 그리고 여기서 말하고 있는 최상급 단서 어휘는 “가장”, “제일”, “처음”, “최고의”, “최대의”, “최소의”, “최초로”, “최초의” 등의 어휘를 말한다.

[표 1] 최상급 단서 어휘 및 최상급 질의 예문

최상급단서어휘		예문
그룹 A	최고의	중국 ***최고의*** 부자는?
	최대의	세계 ***최대의*** 자동차회사는?
	최소의	세계 ***최소의*** 책은?
	최초의	세계 ***최초의*** 대통령은?
그룹 B	가장	세계에서 ***가장*** 오래된 학교는?
	제일	세계에서 ***제일*** 큰 나무는?
	최초로	세계 ***최초로*** 실용화된 계산기는?

최상급 질의는 일반적으로 표 1과 같이 두 그룹으로 나눌 수 있다. 그룹 A는 최상급 단서 어휘 자체에 서술적인 의미를 지니고 있어 별도의 용언이 필요 없는 어휘이다. 그룹 B는 그 외의 경우나 용언을 동반하는 어휘이다. 그룹 A의 경우는 “한국 최초의 동물원은?”과 같이 최상급 질의가 ‘지역 | 최상급 단서 어휘 | 정답유형’의 정형적인 형태를 보이고, 그룹 B의 경우는 “세계에서 가장 큰 폭포는?”과 같이 ‘지역 | 최상급 단서 어휘 | 용언 | 정답유형’의 정형적인 형태를 보인다.

일반적인 질의-응답 시스템의 질의가 최상급 질의일 경우 사용자의 질의가 본 논문에서 제안하는 시스템으로 입력되어 처리하게 된다. 이 때, 최상급 단서 어휘를 중심 키워드로 하여 ‘지역’, ‘정답유형’, ‘용언’ 등의 정답 추출에 필요한 요소들을 추출한다. 이 요소들을 추출할 때 최상급 질의의 정형적인 문장 형태를 이용하게 된다.

그룹 B의 경우에는 그룹 A에 비해 용언, 목적어 등 많은 처리가 필요하여 향후 연구로 남기고, 본 논문에서는 용언을 동반하지 않는 최상급 단서 어휘 그룹 A인 “최고의”, “최대의”, “최소의”, “최초의”에 대한 질의-응답 시스템에 대해서 기술한다.

본 논문에서 제안하는 질의-응답 시스템의 실험을 위하여 파스칼 전자 백과사전[3]에서 최상급 질의에 대하여 응답이 가능한 16,000여 문장을 분석한 결과 “가장, 제일”을 포함하는 문장은 5383개, “최초로, 처음, 처음으로”를 포함하는 문장은 4472개, 그리고 그룹 A인 “최고의, 최대의, 최소의, 최초의, 제일의”를 포함하는 문장은 4502개였다. 이 중 ‘지역 | 최상급 단서 어휘 | 정답유형’의 순서로 나타나는 문장이 거의 대부분이었으며, 일부는 ‘최상급 단서 어휘 | 지역 | 정답유형’의 형태로 나타나기도 하였다. 또한 ‘지역 | 최상급 단서 어휘 | 정답유형’의 순서로 나타나는 경우에 약 22%는 지역, 최상급 단서 어휘, 정답유형이 각각 단일어로 구성되어 있었다.

Ⅲ. 질의 분석 및 정답 추출

1. 질의 분석

본 논문에서 제안하는 시스템은 일반 질의-응답 시스템 내의 한 모듈로, 사용자의 질의가 최상급 단서 어휘를 포함한 질의인 최상급 질의에 대한 처리를 한다.

일반 질의-응답 시스템에 사용자의 자연언어 질의가 입력되면 우선 형태소 분석을 수행한다. 질의문장 내에서 최상급 단서 어휘가 발견되면 최상급 질의에 대한 정답 추출 모듈로 넘겨져 처리를 하게 된다. 형태소 분석이 되어 최상급 질의-응답 모듈로 전달된 질의는 개체명 인식 시스템을 이용하여 최상급 단서 어휘를 중심으로 지역과 정답유형을 추출한다.

[표 2] 최상급 질의 분석 예문

질의 : 한국 최초의 동물원은?		
최상급 단서 어휘	지역	정답유형
최초의	한국	동물원

2. 정답 추출

질의 분석이 완료되면 추출된 지역 정보 어휘를 확장하게 된다. 지역 정보 어휘는 동의어 사전을 이용하며, 표 2의 경우 지역에서 '한국'은 '대한민국', '국내', '우리나라' 등으로 확장된다.

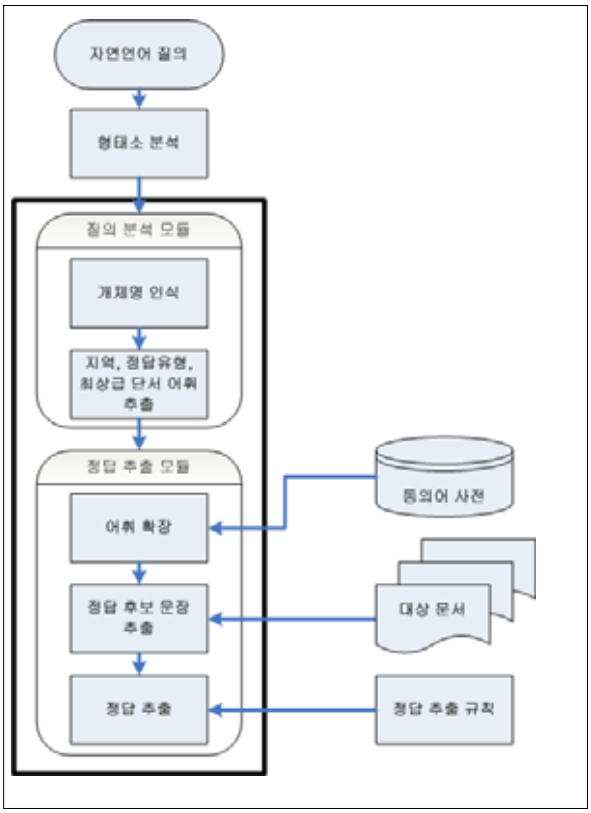
기본적으로 추출된 정보와 어휘 확장을 통해 확장된 정보를 이용하여 대상 문서들에서 정답 후보 문서를 추출한다. 정답 후보 문장을 추출할 때에는 '최상급 단서 어휘 | 지역 | 정답유형'의 순서를 가지는 문장을 우선적으로 추출한 뒤, 순서에 관계없이 '최상급 단서 어휘', '지역', '정답유형'의 출현 여부로 추출하게 된다.

정답 후보 문장을 추출한 후에는 각각의 정답 후보 문장에 정답 추출 규칙을 적용하여 정답 후보 문장으로부터 정답을 추출한다.

표 3은 최상급 단서 어휘 '최초의'에 대한 정답 추출 규칙의 예이다.

[표 3] 최상급 단서 어휘 '최초의'에 대한 정답 추출 규칙

순번	정답 추출 규칙
1	[지역] 최초의 [정답유형]+은/는/jx * <정답>+이/co:!etm
2	[지역] 최초의 [정답유형]+이/co+ㄴ/etm <정답>+은/는/jx
3	<정답>+이/가/jc [지역] 최초의 [정답유형]+이/co:!etm
4	...



▶▶ 그림 1. 시스템 구성도

최상급 단서 어휘를 포함한 정답 후보 문장 "한국 최초의 동물원은 1909년에 개원한 창경원동물원이다."는 최상급 단서 어휘 '최초의'에 대한 첫 번째 정답 추출 규칙을 적용할 수 있고, 정답 후보 문장 "1932년에 한국 최초의 창작동요집인 윤석중(尹石重)의 |_윤석중동요집_|이 나왔으며, ..."에는 두 번째 정답 추출 규칙을 적용할 수 있으며, 정답 후보 문장 "벽골제는 한국 최초의 저수지라는데 의미가 있을 뿐 아니라, ..." 세 번째 정답 추출 규칙을 적용할 수 있다. 정답 추출 규칙 중 '*' 기호는 한 문장 안에서 임의의 단어들이 올 수 있다는 의미이며, ':!etm'은 관형형 어미가 붙을 수 없다는 제약조건이다.

Ⅳ. 실험 및 결과

최상급 질의-응답용으로 작성된 800여개의 질문 셋에서 그룹 A에 해당되는 최상급 단서 어휘를 포함하는 질문 250여개 중 80개의 질문을 이용하여 실험하였다.

무작위로 선별된 80개의 질문에 대하여, 각각의 질문마다 정답을 포함하는 문서 5개와 정답을 포함하지 않는 문서 3개를 해당 질문에 대한 말뭉치로 구축하였다. 이 중 50개의 질문과 각 질문의 정답을 포함하는 문서들은 시스템 구축과 정답 추출 규칙을 작성하는데 사용된 학습 말뭉치이고, 나머지 30개의 질문과 각 질문의 정답 후보 문서들은 시스템의 성능을 측정하는데 사용된 비학습 말뭉치이다. 실험에 사용된 정답 후보 문서들은 파스칼 전자 백과사전에서 최상급 질의에 응답이 가능한 문장과 인터넷 웹 검색을 통하여 수집하였다.

[표 4] 제안한 시스템의 실험 결과

말뭉치	질문수	정답제시	정답	재현율(%)	정확률(%)
학습	50	44	40	88.0	90.9
비학습	30	18	12	60.0	66.7

실험 결과 학습 말뭉치는 최상급 질의-응답 시스템을 구축하고 정답 추출 규칙을 작성하는데 쓰였기 때문에 표 4에서 보는 바와 같이 높은 재현율과 정확률을 보였다.

표 4의 실험과 별도의 질문 50개를 무작위로 선별하여 일반적인 질의-응답 시스템과 제안한 시스템을 비교하여 표 5에 나타내었다.

[표 5] 일반적인 질의-응답 시스템과 제안한 시스템의 비교

시스템	질문수	정답제시	정답	재현율(%)	정확률(%)
일반	50	48	29	96.0	60.4
제안	50	37	35	74.0	94.6

비학습 말뭉치의 실험 결과는 학습 말뭉치에 대한 결과보다는 좋지 않은 결과를 보였지만, 일반 질의-응답 시스템과 비교하였을 때 높은 정확률의 향상을 보였다. 이는 본 논문에서 제안하는 시스템을 일반 질의-응답 시스템 내의 한 모듈로서 사용하여 최상급 단서 어휘를 포함하는 질의를 담당하여 전체적인 질의-응답 시스템의 성능을 향상시킬 수 있음을 의미한다.

정답을 제대로 제시하지 못하거나 제시된 정답이 오답일 경우를 분석해 보면, "(간사이국제공항) 일본 최초의 24시간 하이테크 공항으로 개항과 함께..."와 같은 정답 후보 문장에서 정답유형이 여러 어절 즉, 복합명사나 구의 형태로 이루어진 경우 정답유형 처리가 이루어지지 않아 올바른 정답을 추출할 수 없었다. 다른 예로, 질문 "한국 최초의 종합경기장은?"에 대한 정답 후보 문서 검색에서 각 속성이 하나의 문장 내에 함께 출현하지 않고 문장 또는 단락을 달리하여 나타나는 경우 정답 추출은 한 문장을 대상으로 이루어지는데, 정답 추출에 필요한 '최상급 단서 어휘', '지역', '정답유형' 등의 정보가 한 문장이 아닌 여러 문장에 나타나 정상적인 정답을 제시하지 못하였다. 또한 정답 추출 규칙이 없어 정답을 제시하지 못하거나, 정답 추출 규칙의 제약조건이 부족하여 오답을 추출하기도 하였다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 최상급 단서 어휘를 포함하는 질의에 대하여 최상급 단서 어휘를 중심 키워드로 하여 정답을 추출하는 질의-응답 시스템에 대하여 기술하였다. 질의-분석을 통하여 '지역', '최상급 단서 어휘', 정답유형'등의 정답 추출을 위한 필수 요소들을 추출하였고, 동의어 사전을 이용하여 어휘를 확장하고, 대상 문서로부터 정답 후보 문장을 추출한 다음 정답 추출 규칙을 적용하여 정답을 추출한다.

실험 및 결과는 제안된 최상급 질의-응답 시스템이 일반 질의-응답 시스템 내의 한 모듈로 사용됨으로써 전체 질의-응답 시스템의 성능을 향상시킬 수 있음을 보인다.

향후 연구 방향으로는 최상급 단서 어휘 그룹 B에 대한 처리를 할 수 있도록 확장하고, 올바른 정답을 제시할 수 있도록 복합명사나 구에 대한 처리가 필요하며, 적절한 정답 추출 규칙의 확장에 대한 연구가 이루어질 것이다.

▌참 고 문 헌▐

[1] 강유환, 안영민, 서영훈, "개념 기반 질의-응답 시스템에서 개념 규칙을 이용한 해답 추출", 제17회 한글, 언어, 인지 학술대회 발표집, pp.184-187, 2005.

[2] 고병일, 강유환, 신승은, 서영훈, "질의응답시스템을 위한 서술형 정답 추출", 제16회 한글, 언어, 인지 학술대회 발표집, pp.303-307, 2004.

[3] http://www.epascal.com
http://kr.dic.yahoo.com/search/enc/

[4] ETRI AnyQ QA System : http://anyQ.etri.re.kr

[5] Ellen M. Voorhees, The TREC question answering track, Natural Language Engineering, 7(4), pp.361-378, 2001.

[6] http://www.answerbus.com AnswerBus^{tm} QA system

[7] Encarta^{tm} QA system : http://www.msn.com