기네스 기록 부사와 정답 유형을 이용한 기록문장에서의 정답 추출

오수현[O] 안영민 이충희[¶] 서영훈

충북대학교 컴퓨터공학과
한국전자통신연구원 지식마이닝연구팀[¶]
{ceany[O], maniac}@nlp.chungbuk.ac.kr, forever@etri.re.kr[¶], yhseo@chungbuk.ac.kr

Answer Extraction in Record Sentence using Guinness Record Adverb and Answer-Type

Su-Hyun Oh[O], YoungMin Ahn, Chung-Hee Lee[¶], Young-Hoon Seo
Chungbuk National University
Electronics and Telecommunications Research Institute(ETRI)[¶]

요 약

본 논문에서는 기네스 기록과 같은 기록정보 즉, 기록적 가치가 있는 문장에 대한 질의가 들어왔을 경우 기록 부사와 정답 유형을 이용하여 정답을 추출하는 시스템에 대해 기술한다. 기록정보는 역사적이고 사실적인 내용으로, 기록부사를 포함하는 문장을 말한다. 기록부사는 기록정보 내에서 쓰이며 어떤 사실의 기록에 대해 뜻을 명확하게 나타내어주는 한 요소이고, 이것은 해당문장이 기록문장임을 나타내준다. 이는 질의-응답 시스템에서 정답 추출의 중요한 단서로 사용될 수 있다. 질의-응답 시스템은 크게 질의를 분석하는 부분과 정답 문서를 찾는 부분으로 나뉘며, 질의 분석을 통하여 기록부사 및 지역정보 그리고 정답유형을 결정한 후 이를 이용하여 후보 문서를 검색, 추출하고 정의문 규칙과 개체명 태깅에 의하여 정답을 추출하게 된다.

1. 서론

정보검색(IR) 시스템은 사용자가 입력한 키워드나 자연어 질의를 이용, 해당 단어나 질의와 관련된 문서를 순위화하여 보여주는 시스템이고, 질의-응답 시스템은 사용자에게 자연어 질의를 입력받아 사용자가 요구하는 정답을 제시해주는 시스템이다. 정보검색 시스템은 사용자가 결과로 제시된 문서를 보고 자신이 원하는 정답의 포함 유무를 판단해야하나 질의-응답 시스템은 사용자에게 정답을 제시하여 사용자에게 편의를 제공하기 때문에 이에 대한 요구가 증가하고 있는 추세이다[1].

질의-응답 시스템에 대한 연구는 TREC(Text REtrieval Conference)[2]을 비롯한 여러 연구기관에서 진행되고 있다. 질의-응답 시스템은 크게 질의를 분석하는 부분과 정답을 추출하는 부분으로 나눌 수 있는데 질의를 분석하는 부분은 입력된 질문에 대한 정답유형을 결정하고, 정답 추출 부분에서 질문에 사용된 단어나 질의어와 관련된 어휘를 이용하여 정답이 포함되어 있을 만한 문장이나 문서를 검색한 후 정답 추출 규칙이나 가중치 부여 등의 작업을 수행하여 정답을 추출하게 된다.

따라서 기록정보를 질의-응답 시스템에 적용하여 정답을 추출하는 방법을 생각해볼 수 있다. 기록정보에 대한 기존의 QA 시스템으로는 문장의 문맥정보를 표현할 수 있는 템플릿을 이용하여 사용자의 질의가 들어왔을 경우 질의에 적합한 템플릿을 적용하여 색인한 후, 그 결과로 정답을 제시해주는 한국전자통신연구원의 AnyQ 시스템[3,4]이 있다.

템플릿을 사용한 시스템의 경우 템플릿 정의가 잘 되어 있을 경우 시스템은 높은 성능을 낼 수 있으나, 템플릿 작성에 많은 시간을 할애해야 한다는 단점이 있다.

본 논문에서는 기록정보에 관련된 질의문을 처리할 수 있는 일반화된 기록정보 QA 시스템을 제안한다.

2. 기록정보 및 기록부사

기록정보란 특정 분야에서의 기록적 가치가 있는 문장을 뜻하는 말로 문장 내에 기록 부사라 불리는 부사를 포함하고, 지역, 그리고 정답유형에 대한 명사가 있는 문장을 말한다. 기록 부사는 기록문장의 뜻을 명확히 나타내주어 해당 문장이 기록 문장

임을 나타내주며, 그 예로 "가장, 제일, 최초로, 처음, 처음으로, 최초의, 최대의, 최소의, 최고의, 제일의" 등의 단어를 들 수 있다. 기록부사의 포함여부가 기록문장인지 아닌지를 결정하는 첫 번째 요소가 된다. 예를 들어 "기독교 방송은 1954년 12월 15일 창립된 한국 최초의 민간방송이다." 의 문장의 경우에 "최초의"라는 기록부사가 들어 있고 이것을 질문 형태로 나타낼 경우 "한국 최초의 민간방송은?" 의 형식으로 표현할 수 있다[3].

[표 1] 기록부사 및 예문

기록부사		예 문
그룹1	최초의	한국 **최초의** 신부인 김대건은 ...
	최대의	1733년에는 세계 **최대의** 섬인 그린란드에 ...
그룹2	가장	현재 전하고 있는 한국의 **가장** 오래된 가계부는 조선시대 영조때 암행어사로 유명한 박문수(朴文秀) 종가(宗家)의 양입제출이다.
	최초로	중국에서 **최초로** 통일국가가 성립된 것은 진나라 때이지만

"최초의, 최대의, 최소의, 최고의, 제일의, 가장, 제일, 최초로, 처음, 처음으로" 등의 기록부사 중, 본 논문에서는 기록부사 "최초의, 최대의, 최소의, 최고의, 제일의"를 기록부사 그룹 1로 정의하고 이를 이용한 정답 추출에 대해 기술한다. 기록부사 그룹 1을 제외한 나머지 기록부사는 그룹 2로 정의하였다. 기록부사 그룹1의 경우 용언이 필요 없는 질문을 생성할 수 있고 그룹2의 경우는 질문 생성 시 용언이나 목적어 등이 필요하다.

기록문장은 파스칼 백과사전[1][5]에서 기록 부사를 가지고 있는 문장 중 기록적 가치가 있을 것으로 판단되는 16000여 문장에 대한 분석결과 기록부사 "가장, 제일"을 포함하는 문장은 5383개, "최초로, 처음, 처음으로"를 포함하는 기록문장은 4472개, 그리고 "최초의, 최대의, 최소의, 최고의, 제일의"를 포함하는 문장은 4502개였다.

기록부사 그룹 1을 포함하는 4502문장을 분석한 결과 지역|기록부사|정답유형 의 순서로 나오는 문장이 1000여개로 상당히 많은 수의 문장이 추출되었다.

3. 질의 분석 및 정답 추출

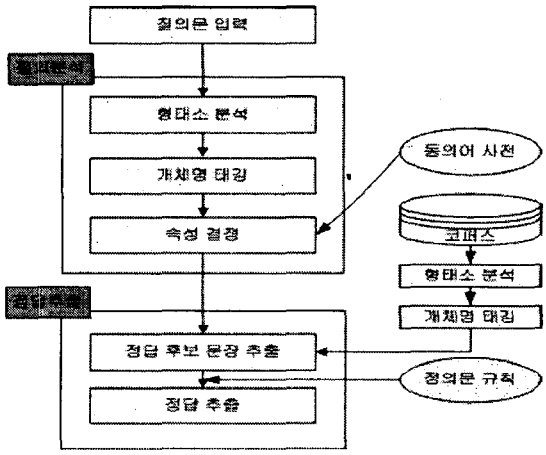
3.1 질의분석

본 논문에서 제안한 시스템은 일반 질의-응답 시스템의 서브 모듈로 기록정보 관련 질의를 처리한다.

1) 동서문화사에서 제공하는 백과사전으로 한국전자통신연구원의 지원으로 사용되었다.

질의분석 순서는 질문이 입력되면 우선 형태소 분석기를 통해 형태소 분석을 하며, 형태소 분석 과정에서 기록 부사가 발견되면 기록정보 추출 모듈에서 처리를 하게 된다. 기록정보 추출 모듈은 개체명 인식기를 통해 지역과 정답유형을 결정하며, 만일 이러한 성분이 발견되지 않으면 일반 질의-응답 모듈에서 해당 질의를 처리하게 된다. 기록부사가 형태소 분석 과정에서 발견되면 개체명 태깅 모듈을 통해 지역과 정답유형을 개체명 태깅하게 되며, 이 때 동의어 사전을 통해 정답유형의 어휘를 확장한 후 정답 추출 시 적용한다.

본 논문의 시스템 구성도는 그림 1에 예시하였다.



[그림 1] 시스템 구성도

그리고 다음 표 2에 질문 구조를 예시하였다.

[표 2] 질문 구조

한국 최초의 박물관은?		
기록부사	지역	정답유형
최초의	한국	박물관

이렇게 질문에 대한 분석이 완료되면 각 정보를 이용하여 정답 후보문서를 검색한다.

3.2 정답추출

질의분석이 끝나게 되면 추출된 기록부사와 지역정보 그리고 정답유형을 이용하여 정답후보문서를 검색한 후 정답을 추출한다.

정답을 추출하는 방법은 정답후보문서의 형태소 분석 후 기록

부사와 지역정보 그리고 정답유형을 포함한 문장을 검색하여 정의문 규칙을 적용하여 해당 규칙에 맞는지를 검토하고 개체명 태깅을 통하여 정답유형과 같은 유형의 개체명 태그를 가지는 단어를 정답으로 제시한다.

정의문이란 "X는 Y이다." 형식의 문장을 말하는 것으로 기록정보의 문장 형식을 정의문 규칙에 적용해보면 X는 질문을 의미하고, Y는 정답을 의미한다[6].

기록문장을 분석한 결과 기록문장에 적용할 수 있는 일반화된 정의문 규칙을 찾아냈고, 다음은 "최초의"에 대한 정의문 규칙의 일부이다.

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

기록문장 "**국내 최초의 박물관**은 1908년 9월에 건설된 **이왕가 박물관**이다."의 문장에 첫 번째 규칙을 적용할 수 있고, 또 다른 기록문장 "1883년(고종 20) 박영선(朴永善)과 함께 박문국(博文局)을 설치하고 일본인 이노우에 가쿠고로[井上角五郎(정상각오랑)]를 초빙하여 **한국 최초의 신문**인 **한성순보**를 간행하였으며, ..." 에는 두 번째 정의문 규칙을 적용할 수 있다. 정의문 규칙 중 '*' 기호는 한 문장 안에 임의의 단어가 올 수 있다는 의미이며, ":!etm"은 관형형 어미가 붙을 수 없다는 제약조건이다.

정의문 규칙을 적용할 때에는 규칙 구성시 출현빈도에 따라 가중치를 부여 우선순위를 부여하였고, 우선순위별로 차례대로 각 규칙에 맞는지의 여부를 검사하여 정답을 제시하게 된다.

4. 실험 및 결과

실험은 기록정보 QA용으로 작성된 800여개의 질문셋 내의 기록부사 그룹1에 해당하는 질문 250여개 중 80개를 이용하였다.

선택된 80개의 질문 중 50개의 질문과 각 질문의 정답을 가지고 있는 1~5개의 해답문서는 학습에 이용되었고, 나머지 30개의 질문은 각 질문 당 1~3개의 정답후보문서와 함께 실험용 데이터로 이용하였다. 실험에 사용한 정답 후보 문서는 백과사전 기록문장과 인터넷 웹 검색을 통하여 수집하였다.

[표 3] 제안한 시스템의 실험 결과

코퍼스(질문수)	정답제시	정답	재현율	정확률
trained (50)	43	40	0.86	0.93
untrained (30)	16	7	0.53	0.44

학습된 코퍼스를 이용한 실험의 경우 재현율과 정확률 면에서 높은 성능을 내고 있으나 실험용 코퍼스를 이용하였을 경우에는 재현율 및 정확률이 좋지 못하다.

오류의 예로는 "(간사이국제공항) 일본 최초의 24시간 하이테크 공항으로 개항과 함께..." 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 정답후보문서 검색에서 각 속성이 하나의 문장 내에 함께 출현하지 않고 문장 또는 단락을 달리하여 나타나는 경우와 정의문 규칙이 없을 경우 정답을 제시하지 못하거나 오답을 출력하였다.

또한 다른 이유로는 정의문 규칙의 적용 우선순위로 인해 적합한 정의문 규칙이 아닌 다른 정의문 규칙이 적용되어 정답이 추출되지 않았다.

5. 결론

본 논문에서는 기록정보 질의에 대해 일반화된 질의-응답 시스템에 대하여 기술하였다. 질의 분석을 통해 기록부사, 지역정보, 정답유형을 결정하고, 정답 추출 모듈에서 질의분석모듈에서 결정된 속성을 사용하여 정답 후보 문서 검색을 실행하고, 템플릿이 아닌 개체명 태깅과 정의문 규칙을 이용하여 정답을 찾아낼 수 있었고, 기록부사를 정답 추출의 중요한 단서로 사용함으로써 일반 질의-응답 시스템보다 나은 결과를 얻을 수 있었다.

향후 연구방향으로는 정의문 규칙의 확장 및 제약, 기록부사 그룹 2에 해당하는 질문의 처리를 위한 용언과 목적어의 분석, 그리고 복합명사나 구의 형태로 나타나는 정답유형의 처리에 대한 연구가 진행되어야할 것이다.

6. 참고문헌

[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] TREC(Text REtrieval Conference), http://trec.nist.gov

[3] 이충희, 오효정, 김현진, 장명길, 템플릿에 기반한 기록정보 QA, 2005 한국컴퓨터종합학술대회 발표논문집, 2005

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

[5] 파스칼 백과사전 : http://www.epascal.com http://kr.dic.yahoo.com/search/enc/search.html?prop=enc&p=&a=c&x=32&y=8·

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