템플릿에 기반한 기록정보 QA

이충희° 오효정 김현진 장명길
한국전자통신연구원
{forever°, ohj, jini, mgjang}@etri.re.kr

Record Information Retrieval based on Template

ChungHee Lee° HyoJung Oh HyeonJin Kim MyungGil Jang
Electronics and Telecommunications Research Institute(ETRI)

요 약

기네스 기록과 같은 기록정보는 사용자가 질의응답 시스템에 자주 질문할 수 있는 내용이지만, 구성단어의 수가 적고 일반적인 단어로 구성되는 기록정보 문장의 특성으로 인해 전통적인 질의응답 시스템에서는 정답을 제시하기 힘든 정보이다. 그러므로 기록정보만을 위한 접근방법이 필요하다. 우리는 기록정보는 특정 문맥에 의해 쓰여지는 경우가 많다는 가정 하에, 문맥 정보를 반영할 수 있는 템플릿을 정의하고, 이 템플릿에 의해서 기록정보를 색인하여 정답을 제시하는 시스템을 제안한다. 템플릿은 거리, 형태소, 형태소품사, 정답유형, 구문 정보의 5가지 제약정보를 나타낼 수 있게 구성된다. 전통적인 백과사전 QA 시스템과 제안 시스템을 비교하여 평가한 결과, 제안한 방법이 기록정보 QA 시스템에 효과적임을 알 수 있었다.

1. 서 론

본 논문에서는 "세계에서 가장 높은 산은 에베레스트산이다.", "피어리가 세계 최초로 북극점을 정복하였다."와 같은 특정 분야에서의 기록적 사실을 '기록정보'로 정의하였고 기록정보에 대해서 정확한 답을 줄 수 있는 방법을 제안하다.

대부분의 질의응답 시스템은 질문에 대한 정답을 찾기 위해서, 질문에 있는 키워드를 이용해서 정보검색을 하고 검색된 문서 또는 단락으로부터 정답을 추출한다[1,2,3]. 기록정보를 묻는 질문의 질의어는 개수가 적고 일반적인 단어인 경우가 많다는 특징이 있는데, 이런 특징 때문에 질의응답 시스템의 첫 번째 정보검색 단계에서의 검색정확도가 떨어지고 결국 전체 QA 정확도가 낮아진다. 그러므로 기록정보 QA를 위한 새로운 접근방법이 필요하다.

현재 서비스되고 있는 QA 시스템 중에는 기록정보에 대해서 문서검색 후에 정답을 추출하는 일반 QA 방법을 사용하는 시스템[4]과 기록정보에 대한 지식을 수작업으로 구축하여 정답을 제시하는 시스템[5]이 있다. [4]의 경우 실시간으로 문서를 검색하고 정답을 제시하므로 정답을 제시하는 시간이 느리고, 기록정보만을 위한 별도의 처리방법을 사용하지 않았기 때문에 기록정보에 대한 QA 정확률이 낮다. [5]는 기록정보에 대한 지식을 수작업으로 구축하였으므로, 시스템 속도와 정답 정확도가 매우 높다는 장점이 있지만 지식 구축에 너무 많은 시간과 인력이 소요되고 새로운 지식에 대한 업데이트 속도와 옛날 지식의 변화에 대처하기 힘들다는 단점이 있다.

본 논문에서는 기록정보에 대한 문장은 특정 문맥 형태로 구성된다는 가정 하에, 문장의 문맥 정보를 표현할 수 있는 템플릿을 정의하고 템플릿에 의해 기록정보를 색인하고 정답으로 제시하고자 한다.

2. 기록정보 색인 및 검색

기록정보 색인 및 검색을 위해서 우리는 백과사전[1] 본문을 사용하였다. 백과사전은 역사적이고 사실적인 내용들로 구성되기 때문에 기록정보를 포함하는 문장이 상당히 많다. 편의를 위해 기록정보를 포함하는 문장을 '기록문장'으로 정의하고 앞으로 사용한다.

2.1 기록정보 색인

우리는 기록문장의 문맥 정보를 표현하고 제약하기 위해 정답 색인 템플릿을 정의하였다. 색인 템플릿을 이용해 추출된 기록정보 색인결과는 RIU(Record Index Unit) 단위로 저장된다. RIU에 대해서는 2.1.2절에서 자세히 설명한다. 색인은 다음 세 단계를 거친다.

2.1.1 중심어휘 정의 및 용례 추출

기록문장은 대부분 해당 문장이 기록정보에 대한 것임을 알 수 있는 중심어휘가 존재한다.

예제1: "영국의 피시하우스는 세계 최초의 수족관이다."

예제1의 문장에서 '최초의'라는 단어를 통해서 기록문장임을 알 수 있다. 우리는 백과사전 본문을 참고해서 55개의 중심어휘를 정의했고, 중심어휘를 포함한 문장을 기록정보 색인을 위한 대상 용례로 추출하였다. 표1은 중심어휘 및 관련 용례에 대한 예를 보인다. 두산 백과사전으로부터 74,203개의 용례 문장을 추출할 수 있었다.

[표1] 기록정보 중심어휘 및 용례

중심어휘	용 례
가장	주피터는 태양계에서 <u>가장</u> 큰 행성인 목성을 가리키는 말이다.
처음으로	가발은 BC 30세기경 고대 이집트에서 <u>처음으로</u> 사용되었다.
최고의	동양 <u>최고의</u> 미인으로 알려진 양귀비는 군인들에게 살해되었다.
제일의	개성은 전국 <u>제일의</u> 상업도시이다.

2.1.2 정답 색인 템플릿

위에서 추출된 모든 용례 문장이 유용한 기록정보를 포함하고

1) (주)두산에서 제공하는 백과사전으로 본 논문에서 사용할 당시에 164,509개의 표제어로 구성되어 있었다.

있지는 않고, 특정 문맥 제약을 만족하는 문장에서만 기록적 가치가 있는 정보를 추출할 수 있다. 예를 들어, "이 가곡은 감수성이 가장 강한 시절에 만들어졌다."의 문장은 '가장'이라는 중심어휘를 가지고 있지만 기록적 가치가 있는 정보를 가지고 있지는 않다. 즉, 일정 조건을 만족하는 문장에서만 기록정보를 추출할 수 있다. 예제1의 문장은 "A의 B가 C 최초의 D이다." 형태의 구조 및 제약 정보(A:AT[2]_국가|속격[3], B:TAG_NN|주어[4], C:MORP_세계[5], D:TAG_NN+CP[6])를 가지고 있다(문맥제약1). 이런 문맥으로 구성된 문장으로부터 기록적 가치가 있는 정보를 추출하여 색인할 수 있다.

문맥제약1과 같이 기록정보를 추출하기 위한 제약 정보를 우리는 템플릿 형태로 표현하였다. 정답 색인 템플릿은 크게 문맥제약 정보와 색인 정보의 두 가지 정보로 구성된다. 문맥제약 정보는 대상 문장이 기록정보를 포함하였는지를 판단하기 위해 1단계로 사용되고, 색인 정보(Record Index Unit)는 1단계를 통과한 문장으로부터 QA에서 사용할 정답 지식을 구축하기 위해 사용된다.

문맥제약정보: 제약정보는 어절을 대상으로 하고 다음과 같은 5가지 정보로 이루어진다.

- 거리제약: 중심어휘로부터의 어절 거리정보
- 형태소제약: 어절에 나타나는 형태소 정보
- 태그[7]제약: 어절에 나타나는 형태소 태그 정보
- 정답유형 제약: 어절에 나타나는 정답유형 정보
- 구문제약: 어절의 격 구조에 대한 정보

위의 제약 중에서 거리제약은 반듯이 있어야 하는 필수정보이고 나머지는 있거나 없을 수 있는 선택정보이다. 하나의 어절에 대해서 2개 이상의 제약이 중복해서 사용될 수 있다.

색인정보: 기록문장으로부터 정답 제시에 필요한 정보를 추출하기 위해 사용된다. 추출되는 정보는 다음과 같다.

- 정답: 해당 기록정보의 주체(질문에 대한 정답으로 사용)
- 용언: 정답과 의존관계에 있는 용언 정보
- 지역(분야): 해당 기록정보의 지역이나 분야에 대한 정보
- 정답종류(상위어): 정답의 종류로 상위어와 비슷한 개념
- 주어: 용언의 가주어
- 목적어: 용언의 목적어

정답과 지역(분야) 정보는 필수정보이고, 나머지는 선택정보이다.

정답 색인 템플릿(183개)은 수작업으로 생성되었고, 중심어휘에 따라서 문장의 문맥이 결정되므로 중심어휘별로 만들어졌다.

2.1.3 기록정보 색인

본 단계는 2.1.2에서 만들어진 정답 색인 템플릿을 이용해서 RIU 정보를 구축하는 단계로 그림1을 통해 구축과정과 결과를 알 수 있다. 현재 4,118개의 RIU 정보를 구축하였다.

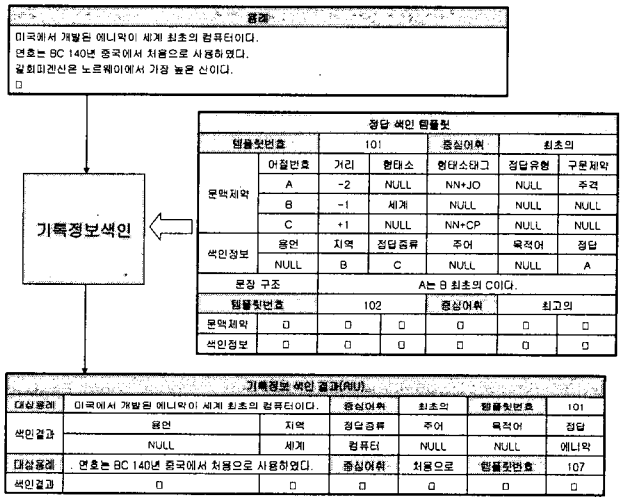

그림 1 템플릿에 기반한 정답 색인

그림1의 정답색인템플릿을 간단히 설명하면,
101 템플릿은 중심어휘 '최초의'에 대한 템플릿으로, 문맥제약정보를 분석하면 다음과 같다. A 어절은 중심어휘로부터 왼쪽 두 번째 어절이고, 어절의 형태소는 '명사+조사'로 구성되어 있고, 주격이어야 한다. B 어절은 중심어휘의 왼쪽 첫 번째 어절이고, 형태소가 '세계'로 구성되어 있어야 한다. C 어절은 중심어휘의 오른쪽 첫 번째 어절이고, 어절이 명사+지정사 형태이어야 한다. 또한 색인 정보로부터 B 어절이 지역, C 어절이 정답종류, A 어절이 정답임을 알 수 있다.
기록정보 색인결과를 보면, "미국에서 개발된 에니악이 세계 최초의 컴퓨터이다."라는 문장이 101 템플릿의 문맥제약정보를 만족하므로 RIU로 [지역:세계][정답종류:컴퓨터][정답:에니악]가 색인되어 저장되었다.

2.2 기록정보 검색

기록정보 검색에서는 정답색인단계에서 구축된 ARIU(Answer Record Index Unit)와 질문분석결과인 QRIU(Question Record Index Unit)을 비교해서 정답을 제시한다.

2.2.1 질문 색인 템플릿 정의

기록정보를 묻는 질문 또한 특정 형태의 문맥을 가지고 있으므로 정답색인 템플릿과 유사한 템플릿에 의해 정답검색에 필요

2) 정답유형(AT): 정답유형은 QA를 위한 정답의 유형들로 본 논문에서는 178개를 정의해서 사용한다(예: 사람, 경제기관, 도시, 수도,...)
3) AT_국가|속격: 해당 어절이 AT가 '국가'이고 속격이어야 한다는 제약정보
4) TAG_NN|주어: 해당 어절이 주어이고 주어의 형태소가 명사
5) MORP_세계: 해당 어절의 형태소가 '세계'
6) TAG_NN+CP: 해당 어절이 '명사+지정사' 형태의 용언
7) 본 논문의 형태소분석기는 26개 태그셋을 사용한다. 주요태그로는 NN(명사), JO(조사), VV(동사), AJ(형용사), CP(지정사), EM(어미) 등이 있다.

한 정보를 추출할 수 있다. 정답색인단계의 6가지 RIU 중에서 정답을 제외한 5가지 정보(QRIU)를 질문분석에서 추출해야 한다. 질문 색인 템플릿은 색인정보로 5가지를 추출한다는 것을 제외하고 정답 색인 템플릿과 동일한 구조와 제약정보를 가지고 있다. 질문 분석 템플릿은 기록정보에 대한 질문셋을 분석해서 수작업으로 구축되었고, 63개가 만들어졌다.

2.2.2. 기록정보 검색

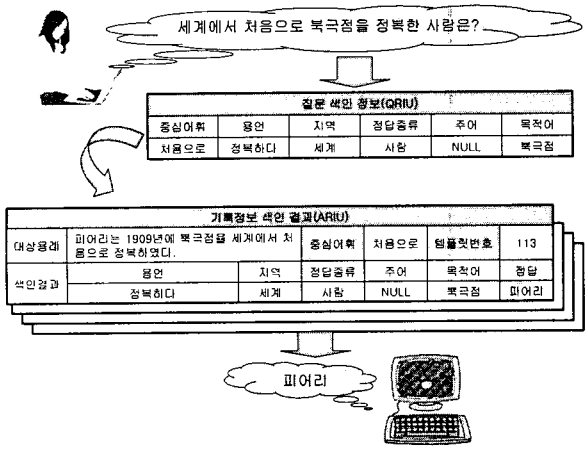

그림 2 기록정보에 대한 질의응답 예제

그림2는 질문색인 정보와 정답색인 정보를 비교해서 피어리를 정답으로 제시하는 과정을 보여준다.

3. 실험

평가는 백과사전 QA용으로 작성된 510개의 평가셋을 이용하였다. 평가셋은 백과사전과 관련해서 질문할 수 있는 모든 질문을 포함하였는데, 510개 중 61개가 기록정보에 대한 질문이었다. 제안 시스템의 비교대상으로는 전통적 QA시스템인 [6]의 anyQ 시스템을 사용하였고, 평가 measure는 [6]에서 사용한 상위5개 안의 정답유무를 보는 Top5를 사용하였다.

본 논문은 기록정보에 대한 질문에 대해서는 전통적인 QA 시스템의 정확도가 낮을 것이라는 가정 하에 출발하였으므로, 첫 번째 실험으로 기록정보에 대한 질문과 일반 질문에 대해 anyQ 시스템의 QA 성능을 비교하였다.

[표2] anyQ 성능: 일반 질문 vs 기록정보 질문

질문유형	정답제시	정답	재현율	정확률
일반	449	336	0.75	0.75
기록정보	61	37	0.61	0.61

표2에서 보듯이 일반 질문에 비해, 기록정보 질문에 대해서 anyQ 시스템이 14.1% 낮은 성능을 보임을 알 수 있었다.

위 실험으로 기록정보를 처리할 수 있는 별도의 방법이 필요함을 알 수 있었으므로, 두 번째 실험으로 본 논문에서 제안한 방법이 기록정보에 대한 QA 시스템으로 얼마나 효과적인지를 테스트하였다. 실험으로 기록정보에 대한 61개의 질문에 대해서 제안한 시스템과 anyQ 시스템을 비교하였다.

[표3] 기록정보 QA: 제안 시스템 vs anyQ

시스템	정답제시	정답	재현율	정확률	F-score
제안	22	21	0.34	0.96	0.65
anyQ	61	37	0.61	0.61	0.61

실험결과, 제안 시스템이 anyQ 시스템에 비해서 훨씬 높은 정확률을 보이므로 기록정보 QA에 매우 효과적임을 알 수 있었다.

4. 결론

기록정보를 묻는 질문에 대해서 전통적인 QA시스템은 정답을 줄 확률이 낮으므로 기록정보를 처리할 수 있는 방법이 필요하다. 기록정보에 대한 문장은 특정 문맥으로 구성되는 특징이 있으므로 문장의 문맥 정보를 템플릿으로 표현하고, 이런 템플릿을 이용해서 기록정보를 색인하고 제시하는 기록정보 QA 시스템을 본 논문은 제안하였다. 61개의 기록정보 질문에 대한 실험에서 제안 시스템이 일반 질의응답 시스템에 비해 높은 정확률과 F-score를 보였으므로 기록정보 QA에 템플릿에 기반한 방법이 효과적임을 알 수 있었다.

하지만 템플릿에 기반한 현재의 제안 시스템의 경우, 제약 정보 부족과 고난이도 언어분석을 수행하지 않으므로 중심어휘로부터 먼 거리에 중요한 정보들이 있는 복잡한 문장을 처리하지 못하므로 재현율이 낮다는 문제점이 있다.

앞으로의 연구방향은 재현율을 높이기 위해 복잡한 문장 구조를 반영할 수 있도록 제약정보를 확장하고, 색인된 정보를 유사어를 통해 확장할 계획이다.

5. 참고

[1] Vlado Keselj, Anthony Cox, "DalTREC 2004: Question Answering using Regular Expression Rewriting", in the Text Retrieval Conference(TREC) 13, 2004

[2] Lide Wu, Xuangjing Huang, Lan You, Zhushuo Zhang, Xin Li, Yaqian Zhou, "FDUQA on TREC2004 QA Track", in the Text Retrieval Conference(TREC) 13, 2004

[3] Kyoung-Soo Han, Hoojung Chung, Sang-Bum Kim, Young-In Song, Joo-Young Lee, and Hae-Chang Rim, "Korea University Question Answering System at TREC 2004", in the Text Retrieval Conference(TREC) 13, 2004

[4] AnswerBus[tm] QA system: www.answerbus.com

[5] Encarta[tm] QA system: www.msn.com

[6] Hyeon-Jin Kim, Hyo-Jung Oh, Ji-Hyun Wang, Chung-Hee Lee, Myung-Gil Jan, "The 3-step Answer Processing Method for Encyclopedia Question-Answering System: AnyQuestion 1.0", in the Proceedings of Asia Information Retrieval Symposium(AIRS), pp309-312, 2004