QA를 위한 백과사전 사건 추적 시스템의 설계

장혜주○ 정유철 강보영[1] 송사광[2] 김운[3] 송훈 맹성현
한국정보통신대학교, 서울대학교[1], 한국전자통신연구원[2] , 충남대학교[3]
{hjjang○, enthusia77, shoon, myaeng}@icu.ac.kr, comeng99@snu.ac.kr[1], smallj@icu.ac.kr[2], wkim@cs.cnu.ac.kr[3]

Design of Topic Detection and Tracking System for QA in Encyclopedia

Hyeju Jang○, Yuchel Jung, Bo-Yeong Kang[1], Sa Kwang Song[2], Jin Un[3], Hoon Song, Sung Hyon Myaeng
Information and Communications University
Seoul National University[1]
Electronics and Telecommunications Research Institute[2]
Chungnam National University[3]

요 약

본 논문은 백과사전 QA에서 여러 문서에서 정답을 추출한 후 종합하여 답을 출력하여야 하는 질의를 위한 백과사전 사건 추적 시스템을 제안한다. 본 시스템은 사건 관련 질문과 문서의 속성을 반영할 수 있는 템플릿을 정의하여 문서를 추적하며, 하나의 사건은 '제목', '시간', '장소', '주체', '범주'의 5가지 속성을 가진다. 이러한 방법론을 통하여 기존 QA 시스템의 정답 추출 성능 향상에 도움을 주고, 정보 구성(organizing)과 TDT(Topic Detection and Tracking) 연구에서의 새로운 관점과 방향을 제시하고자 한다.

1. 서 론

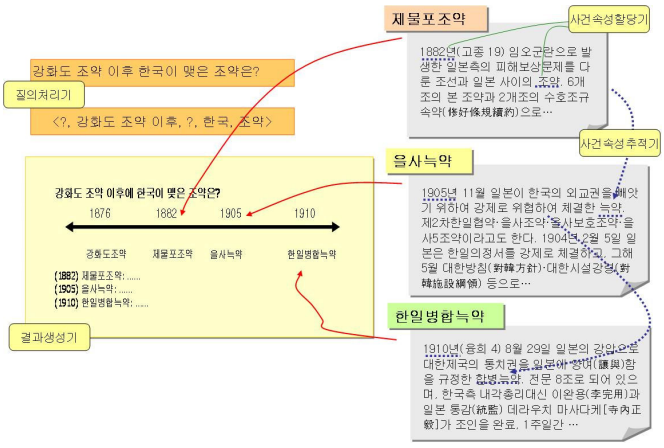
기존의 QA시스템에서는 대부분 한 문서 내에서 답을 찾을 수 있는 단답형의 질의 처리가 주류를 이루고 있으며, 여러 문서에서 정답을 추출한 후 종합하여 답을 출력하여야 하는 서술형 질의 해결에는 한계가 있다. 예를 들어, 기존의 QA시스템에서는 '임진왜란 이후에 조선에서 맺은 조약은?'과 같은 질의는 해결하기 힘들었다. 왜냐하면 임진왜란 이후에 조선에서 맺었던 각각의 조약에 대한 문서를 찾은 후 그것을 시간별/장소별로 종합하여 답을 제시하여야 하기 때문이다. 또한 기존 시스템의 경우 질의에 사용된 용어를 정확히 내포하는 문장이 문서 내에 존재하지 않을 경우 정답 추출에 실패할 수 있다.

본 논문은 이와 같이 QA시스템의 질의로 사용될 수 있으나 기존의 QA기술로는 해결하기 힘든 일부 질의 - 여러 문서에 걸쳐 있는 답을 종합하여야 하는 사건 관련 질의- 에 대하여 기존의 TDT기술을 응용 확장 적용하여 기존 QA시스템의 정답 추출 성능 향상에 도움을 주는 시스템을 제안하고자 한다.

사건 추적 시스템은 백과사전 문서를 대상으로 사건 문서에 나타난 사건의 속성을 정의하여 질의 템플릿과 사건 문서 템플릿을 만든 후, 사건 간의 관계를 사건 추적(Topic Tracking)의 새로운 방향으로 봄으로써 관련 문서를 추적한다. 이를 통하여, 기존 QA시스템의 질의응답에 도움을 줄 뿐만 아니라 TDT의 새로운 적용 가능성을 시도하였다고도 할 수 있다.

2. 사건 추적 시스템의 역할

사건 추적 시스템은 사건에 관한 질의를 입력으로 받아 그 질의와 관계 있는 여러 문서를 추적하여 답을 생성해 주는 시스템이다. [그림 1]은 사건 추적 시스템이 수행하는 일을 묘사하고 있다. 사건 추적기는 자연어로 된 사용자 질의를 받아 들여 본 연구에서 정의한 사건의 속성을 추출한 후, 질의 템플릿으로 변환시켜 주는 질의 처리기, 백과사전 문서에서 사건의 속성을 추출하여 색인하는 사건 속성 할당기, 질의에 따라 사건의 속성 별로 문서를 추적해 주는 사건 속성 추적기, 결과를 취합하고 보여주는 결과 생성기로 구성되어 있다. 본 논문의 다음 절에서는 사건을 어떠한 관점으로 바라보고 정보를 모델링하였는지를 질의와 백과사전 문서로 나누어 설명하고자 한다.

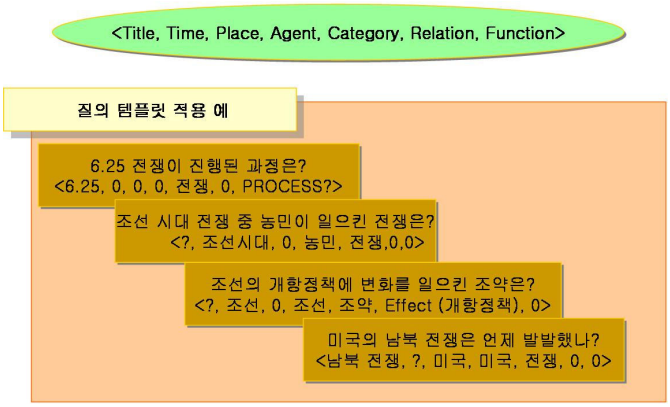


[그림 1] 사건 추적 시스템의 역할

3. 질의 모델

본 연구의 질의 모델은 한국전자통신연구원에서 제공한 전쟁 및 조약 분야의 사용자 질의 모음과 본 연구 그룹에서 생성한 사건 중심 질의 총 40개를 분석하여 설계되었다. 실제 질의의 예를 통하여 사건의 속성을 추출하는 상향식 방식을 통한 질의 분석을 수행하였다.

사건 중심 질의로부터 우리는 [그림 2]에서 볼 수 있듯이 시간, 장소, 범주, 제목, 주체, 사건 간의 관계, 질의가 요구하는 답의 방식 등의 표현을 볼 수 있다. 예를 들어, '조선 시대 전쟁 중 농민이 일으킨 전쟁은?'이라는 질의에서 '조선 시대'는 시간, '농민'은 주체, '전쟁'은 범주의 개념에 속하게 된다. 본 연구에서는 이러한 분류 중 사건 그 자체와 관련된 제목, 시간, 장소, 주체, 범주의 5가지를 사건의 속성이라 정의한다. 그리고, 질의 템플릿에는 5가지의 사건 속성과 더불어 질의에서 요구하는 바를 표현하는 관계와 방식의 항목을 포함시켰다.



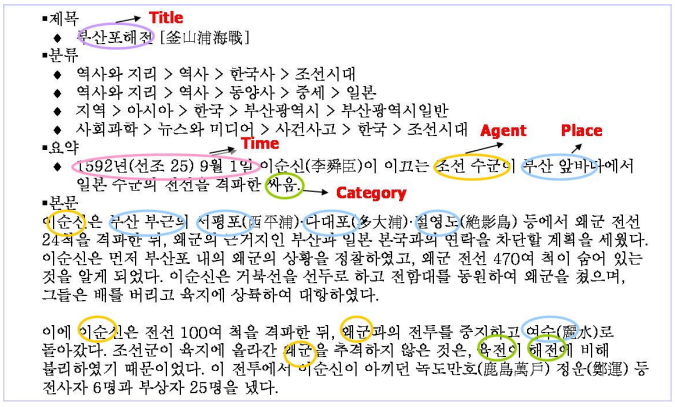
[그림 2] 질의 템플릿과 적용 예

사건의 속성 중, 제목은 그 사건 고유의 이름을 의미하고 장소는 사건이 발생한 지역을 칭한다. 또한 시간은 사건이 발생한 특정 시점 혹은 기간을 의미하여, 사용자의 표현에 따라 다르게 나타날 수 있다. 예를 들어, '6.25 휴전 이후에 맺어진 조약은?'이라는 질의와 '1953년 이후에 맺어진 조약은?'이라는 질의는 같은 답을 찾고자 한다. 장소 또한 시간과 마찬가지로, 같은 정보가 질의에 따라 다르게 표현될 수 있다. '대전 유성구'와 '대전 유성구 문지동' 그리고 '한국정보통신대학교'는 범위가 다르지만 같은 지역을 표현하기 위한 다른 표현이 될 수 있다. 이러한 이유로 시간과 장소 정보는 정규화 과정이 필요하다. 주체는 전쟁과 조약 도메인에서 전쟁과 조약이라는 사건에 참가하고 있는 대상을 칭한다. 이 때 주체는 개인이 될 수 있으며 국가, 혹은 단체 또한 주체가 될 수 있다. 주체 정보의 경우 주체의 레벨이 모호할 수 있으므로 시간이나 장소처럼 정규화 과정이 필요할 수 있다. 예를 들어, 세계 2차 대전의 주체는 전쟁에 참가하고 있는 국가도 될 수 있지만 군인이라고 할 수도 있다. 마지막으로 범주는 그 사건이 어떤 도메인에 속하는가를 말해준다. 즉, 그 사건이 전쟁인지 조약인지를 나타내 주는 항목으로 다양한 범주 계층이 존재할 수 있다. 한산도 대첩을 예로 들어 보면, 한산도 대첩은 '전쟁', '전투', '해전' 등의 범주에 속할 수 있다.

관계(relation)는 질의 안에서의 사건과 답이 되야 하는 사건의 관계를 나타낸다. 예를 들어, '조선의 개항정책에 변화를 일으킨 조약은?'이라는 질의에서 질의 안에서의 사건은 '개항정책'이고 답이 되야 하는 사건은 '개항정책'에 영향을 미친 사건이라고 할 수 있다. 이 관계를 Effect라고 명명하였다. 방식(function)은 사용자가 요구하는 답이 어떻게 보여져야 하는지를 나타낸다.

4. 사건 관련 백과 사전 문서 모델

본 연구에서 사건 관련 문서 탐색은 사건 중심의 질의 모델을 활용하여 백과사전 문서를 탐색하는 과정이므로 질의 모델에 포함되는 속성을 백과사전 문서로부터 추출하여 사건 관련 문서 모델을 설계하였다. 하나의 백과사전 문서는 하나의 사건에 대해 묘사하고 있다. 그리하여 하나의 백과사전 문서는 질의 모델에서와 같이 제목, 시간, 장소, 주체, 범주로 표현될 수 있다. [그림 3]은 하나의 문서가 나타내는 사건의 5가지 속성을 보여주고 있다.



[그림 3] 사건 관련 백과사전 문서의 사건 속성

제목은 문서가 기술하고 있는 사건의 명칭을 나타낸다. 문서의 제목과 일치한다. 시간과 장소는 각각 사건이 발생한 시점의 시간과 사건이 발생한 장소를 나타낸다. 이는 백과사전 문서의 특성 상 대개 요약 부분에 나타나며 개체명 인식 정보 및 장소 태깅 정보를 바탕으로 추출할 수 있다. 주체는 사건에 참여하고 있는 해당 대상을 의미한다. 요약 정보에 드러나는 주체를 중심으로 사건과 관련 있는 주체를 추출할 수 있으나 본문에서 추출해야 하는 경우도 있다. 범주는 문서가 기술하고 있는 사건이 속하는 범주를 나타낸다. 본 연구에서 다루고 있는 사건의 경우, 사건의 범주가 '전쟁', '조약'으로 한정되나 일반적인 모델이 고려되어야 한다. 백과사전 자체에서 문서가 속한 범주 정보를 제공

하는데 이러한 범주는 사건 중심의 분류가 아니므로 본 연구의 범주 정보로 활용하는 데에는 한계가 있다. 따라서 본 연구에서 활용하고자 하는 범주 정보는 요약 정보 및 사건 제목, 혹은 키워드 정보로부터 추출되어야 한다. 질의 모델에서와 같이, 사건의 양상에 따라 같은 전쟁이라도 '전투', '민란', '해전' 등의 하위 범주로 분류될 수 있으므로 문서 범주에 대한 계층 또한 고려된다.

질의를 해결하기 위해서는 질의에서 사용자가 제시한 사건과 특정한 관계에 있는 사건을 추적해야 하게 된다. 즉, 사건 간의 관계가 사건 탐색에 중요한 요소로 사용될 수 있다. 예를 들어, '임진왜란 이후에 한국에서 발생한 전쟁은?'은 임진왜란이라는 사건이 주어져 있고 그 사건과 시간적으로 '후(after)'에 한국이라는 장소에서 발생한 전쟁이라는 범주에 속하는 사건을 탐색하고자 하는 질의이다. 이런 경우처럼 두 문서가 기술하고 있는 사건 간의 의미적 관계를 모델링 할 필요가 있다.

따라서 본 시스템에서는 사건 간의 관계를 같은 사건, 배경 사건, 인과 관계 사건, 시간 관계 사건, 장소 관계 사건, 같은 주체의 사건, 같은 범주의 사건의 일곱 가지 유형으로 정의한다. 이와 같이, 사건 간의 관계 정보를 이용하여 문서를 추적하는 것은 같은 사건에 대해 존재하는 문서를 추적하는 기존의 TDT시스템에서의 역할과 대비되기 때문에 TDT시스템의 새로운 방향을 제시하는 중요한 요소가 된다.

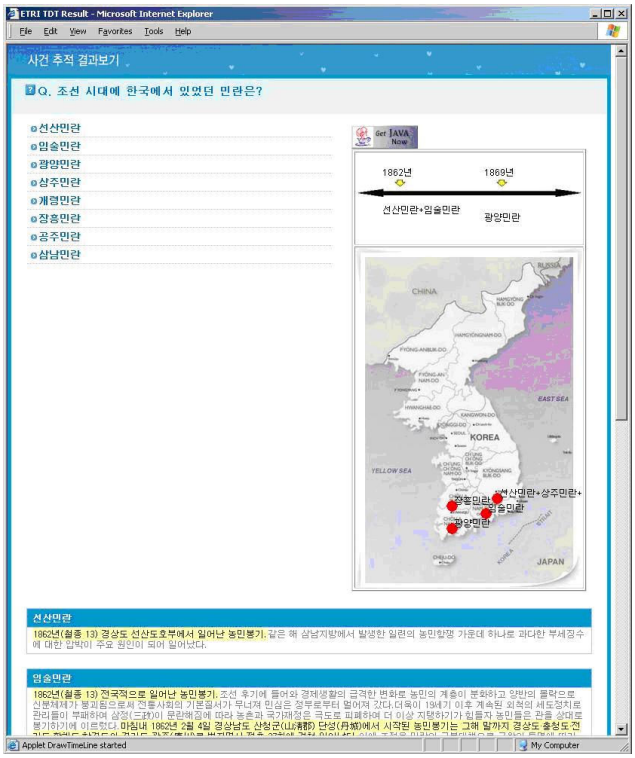
5. 사건 모델에 기반한 TDT

전통적인 TDT에서는 본 연구에서 정의한 사건 간의 관계 중 같은 사건에 관한 사건 추적을 주로 다루어 왔다. 그러나 본 연구에서 제안한 모델은 사건 추적을 같은 사건에 관한 사건을 포함하여 일곱 가지 유형의 관계성을 가지는 사건의 추적으로 바라보고 있다. 이러한 시도는 전통적인 TDT연구에 새로운 방향을 제시한다.

즉, TDT를 단순히 '유사한 사건'을 찾는 작업으로만이 아니라 보다 다양한 조건을 만족하는 문서를 찾아내는 작업으로 보는 것이다. 예를 들어, '조선시대 농민이 일으킨 반란은?'과 같은 질의에 답을 하기 위하여, 유사 사건을 추적하되 사건의 주체 정보에 초점을 맞추어 동일 주체가 일으킨 특정 시대의 같은 범주의 사건을 추적하는 과정을 거친다. 또한 한 장소에서 일어난 다양한 사건을 추적할 수 있으며, 그 장소의 남쪽에 있는 지역에서 일어난 고려 시대의 전쟁도 추적할 수 있는 것이다. 따라서 제안된 사건 관련 질의와 문서 모델이 기존 TDT에서 활용되던 사건 추적 기법을 보다 확장할 수 있으며 QA시스템에 적용되어 고난이도의 질의 처리 수행에 도움을 줄 것으로 기대된다.

5. QA를 위한 백과사전 사건 추적 시스템 구현

한국전자통신연구원으로부터 제공받은 파스칼 백과사전의 전쟁과 조약 관련 1461 문서를 대상으로 위에서 서술한 시스템을 구현하였다. [그림 4]는 '조선시대 한국에서 있었던 민란은?' 이라는 질의에 대한 시스템의 실제 사건 추적 결과 화면을 보여주고 있다. 답은 사용자의 가독성을 위하여 시간과 장소 별로 그래프와 지도를 이용하여 나타내고 있다. 이러한 Visualization 또한 여러 문서에 걸친 답을 원하는 질의에 대하여 시스템이 정보를 종합하여 보여주는 방식의 실례를 보여 준다고 할 수 있다.



[그림 4] QA를 위한 사건 추적 시스템의 결과화면

6. 결론 및 향후 연구 방향

본 논문에서는 백과사전 QA시스템을 위한 사건 추적 시스템 및 관련 질의/문서 모델을 제안하고 시스템을 통해 그 활용 가능성을 보였다. 제안된 모델과 기술은 질의응답 시스템에 활용되어 고난이도에 해당하는 추적성 나열형 정답의 처리를 개선하고 TDT연구의 새로운 지평을 열었다. 향후 연구로는 각 속성의 추출 및 활용 방법을 개선하여 시스템의 성능을 높여 완성도를 기하고자 한다.

<Acknowledgement>

본 연구는 한국전자통신연구원 질의 응답형 정보검색 기술 개발 과제 지원으로 수행되었음.

7. 참고 문헌

[1] J. Allen et al. Topic Detection and Tracking Pilot Study Final Report, In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Feb. 1998.

[2] 김현진, 이충희, 오효정, 왕지현, 장명길, "백과사전 질의응답을 위한 구문정보기반 정답색인방법", 한국컴퓨터종합학술대회 논문집 Vol. 32, No. 1(B) pp. 511, 2005