2002년 한국정보처리학회 춘계학술발표논문집 제9권 제1호

이벤트 템플릿을 이용한 정보 추출에 관한 연구

임수종, 정의석, 황이규, 윤보현
한국전자통신연구원 지식처리 연구팀
e-mail : {isj, eschung, hik63265,ybh}@etri.re.kr

A Study on Information Extraction Using Event Template

Soojong Lim, Euisok Chung, Yi-Gyu Hwang, Bo-Hyun Yoon
Knowledge Processing Team, ETRI

요 약

본 논문에서는 개체형 인식이 된 일반 문서에서 정보 추출을 하기 위하여 이벤트 템플릿 구조를 사용하는 방법을 제안한다. 제한된 도메인 지식을 주로 사용하는 기존의 정보 추출 방법과 달리 predicate-argument 구조를 갖는 이벤트 템플릿은 일반적인 지식을 주로 사용하여 정보 추출을 한다. 이벤트 템플릿을 추출하기 위해서는 형태소 분석 결과 용언의 하위범주 정보를 이용하고 이벤트 템플릿의 논항 구조를 이용하여 필요시 이벤트 템플릿을 통합한다. 문서에서 생성된 일반적인 이벤트 템플릿은 정보 수용자의 요구에 맞는 도메인 지식을 사용하여 최종적인 결과를 생성한다. 이벤트 템플릿을 사용하는 정보 추출 실험 결과는 제한된 도메인 정보를 사용하는 시스템에 비해 정확율은 떨어지지만 기존 정보 추출시스템의 문제인 이식성을 높일 수 있다.

1. 서론

최근 컴퓨터를 이용한 지식 관리가 일반화 되면서 인터넷등을 이용하여 접근할 수 있는 지식의 양이 늘어나고 있다. 그러나 이런 정보가 늘어날수록 사용자가 특정 정보에 접근하기 위해서 많은 노력을 필요로 한다. 정보 추출(Information Extraction)이란 자연어로 작성된 문서 집합에서 원하는 정보를 선택하여 구조화된 표현으로 생성하는 것을 말한다[5]. 추출된 정보는 일회성의 자료가 될 수도 있지만 데이터 베이스에 저장하여 반영구적으로 활용될 수도 있다. 공연 분야를 예를 들면, 공연에 관한 문서 안에서 정보 수요자는 공연자(주체), 공연명, 공연날짜와 시간, 공연 장소만을 알고 싶을 경우 공연에 대한 줄거리 혹은 공연 작품의 원작자등에 대한 정보를 무시하고 정보 수요자가 요구한 정보만을 추출하여 준다. 정보 추출은 정보검색, 에이전트, 정보 여과등의 다른 분야와 결합하여 시너지 효과를 유발할 수 있다.

문서에 있는 정보를 테이블 형태로 바꾸려는 시도는 1950년대에도 있었으나 연구 분야로 자리 잡은 것은 1991년의 MUC(Message Understanding Conference)-3 이후의 일이다[8]. 정보추출이 자연어로 작성된 문서를 다루기 때문에 자연어 처리 기술을 이용하지만, 추출을 목적으로 하는 문서의 일부분만을 이해하면 되기 때문에 기존의 자연언어 처리와는 다르게 문서의 모든 부분을 이해할 필요는 없다. 이러한 점에서 자연언어 처리 기술의 응용 분야들과 접근 방법이 다르다.

그러나 문서의 일부부만을 이해한다는 특성과 특정한 정보만을 관심 영역으로 삼기 때문에 전반적으로 정보 추출 기술은 특정한 도메인에 종속되어 이식성이 떨어지는 단점이 있다.

본 연구에서는 도메인 종속적인 기술이나 언어자원을 최소화하여 일반적으로 접근할 수 있는 이벤트 템플릿을 이용한 정보추출 방법을 제안한다.

2. 자연어 기반 정보 추출

정보추출의 큰 부류는 자연어처리 기반 정보추출과 Wrapper 기반 정보추출로 나뉘며, 문서구조와 언어구조를 이용하여 문서로부터 적합한 컨텐츠를 결정하는 것이다[8].자연어 처리 기반 정보 추출 시스템 구조는 몇 년간의 다양한 실험과 연구를 거쳐 현재는 어느 정도 정보 추출을 위한 기본적인 시스템의 구조가 확립되고 있다. 정보 추출의 목적이나 방법 에 따라 부분적인 차이는 있지만 대체적으로 연구자들이 동의하는 정보추출 시스템의 개략적인 구조는 그림 1 과

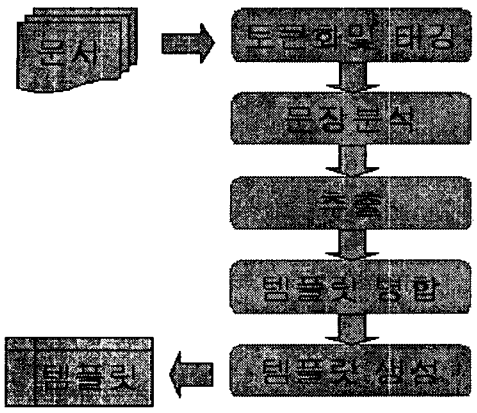

그림 1 정보추출 시스템의 구조

같다. [1]

정보 추출 분야에서 템플릿 생성 및 통합을 위한 기존의 연구로 규칙에 기반한 방법, 통계 기법을 이용하는 방법 있다. 규칙에 기반한 방법은 Complex words, Basic Phrases, Complex Phrases 로 문장을 분석하기 위하여 각각에 맞는 FSA(Finite State Automaton)를 이용하여 문장을 분석한 후 각각의 구문 정보와 어휘로 이루어진 규칙에 기반하여 TE(Template Element), TR(Template Relation), ST(Scenario Template)을 생성하는 FASTUS 시스템[4], 범용으로 사용될 수 있는 규칙 패턴 라이브러리를 구축한 후에 정보를 추출하고자 하는 대상 도메인에 학습을 통하여 해당 패턴을 바꾸는 Proteus/PET 시스템이 있다[6]. Proteus/PET 시스템은 구축된 규칙 패턴 라이브러리를 다른 도메인에 쉽게 적용할 수 있도록 GUI 기반의 tool 을 제공한다. 템플릿 생성 단계에 유용하게 쓰일 수 있도록 태그를 세분하고, 특정 태그가 어떤 세부 범주, 어떤 특징을 갖게 되는지를 정의한 규칙을 학습에 기반하여 반자동 추출할 수 있는 IE2(Information Extraction Engine) 시스템[2]은 규칙을 반자동 추출하기 위하여 여러가지 종류의 tool 을 개발하였다.

통계 기법을 이용한 방법은 공통적으로 학습 말뭉치를 필요로 한다. SIFT 시스템[7]은 엔티티와 그들의 관계를 학습하기 위한 의미 부착 말뭉치와 일반적인 문장 구조를 학습하기 위한 구문 부착 말뭉치 두가지 형태의 학습 말뭉치를 이용하여 학습한다. 학습 알고리즘은 구문 및 의미를 고려하는 통합된 통계 모델의 매개변수를 추정한다. 새로운 문장이 입력되면, 템플릿 생성 알고리즘은 통계모델을 이용하여 가장 가능성이 있는 구문 및 의미 해석 결과를 바탕으로 템플릿을 생성한다. 또한 언어 독립적인 특성을 지닌 방법으로 의존 그래프 (dependency graph)를 기반으로 (단어, 구문 범주, 수식어-피수식어 관계) 쌍의 공기 정를 이용한 방법[3]도 있다.

그림 1 과 같은 정보 추출 시스템의 구조 중에서 추출 단계는 첫번째로 도메인에 종속되는 부분이다. 기존의 정보 추출 시스템은 정보 추출의 정확성을 높이기 위해서 과도한 도메인 지식을 사용하여 이식성이 떨어지는 단점을 갖고 있다. 이러한 단점을 극복하기 위해서는 최대한 일반적인 지식을 이용하려고 시도한다.

본 논문에서는 현 정보 추출 시스템의 가장 큰 걸림돌인 이식성 문제를 해결하기 위해서 추출의 단계에서 이벤트 템플릿이라는 일반적인 논항 구조를 갖는 틀을 이용하여 정보 추출 기술의 이식성과 활용성을 높이도록 하였다.

3. 한국어 문서의 정보 추출

3.1 이벤트 템플릿

이벤트 템플릿은 '누가 언제 어디서 무엇을 어떻게' 라는 사건을 서술하는 정보가 만족되어야 한다. 문서에서 추출된 이벤트는 정보 추출을 위한 시나리오 템플릿(Scenario Template) 생성 뿐 아니라 질의/응답, 문서 요약, 텍스트 마이닝의 기초 자료로 활용될 수 있다.

다음은 신문의 공연 기사에 출연한 문장으로 일반적인 predicate-argument 구조를 갖는 Event Template 을 생성한 예이다.

<TITLE>명성황후</TITLE> 제작사인 <ORGANIZATION>㈜에이콤인터내셔날</ORGANIZATION> (대표 <PERSON>윤호진</PERSON>)은 이 작품을 <DATE> 내년 2 월 1-16 일 </DATE> <LOCATION>런던</LOCATION>의 <LOCATION>아폴로 헤머스미스 극장</LOCATION> 무대에 올린다.

```
<Event_Template>
<PREDICATE = "올리다"> 무대</PREDICATE>
  <ARG1 type ='ORGANIZATION'>
  ㈜에이콤인터내셔날 </ARG1>
  <ARG2 type ='TITLE'>명성황후</ARG2><TIME
  type ='DATE'> 내년 2 월 1-16 일 </TIME>
  <LOCATION type ='LOCATION'> 런던, 아폴로
  헤머스미스 극장 </LOCATION>
</Event_Template>
```

위의 예에서 시간과 장소가 아닌 ARG1, ARG2 로 표시된 부분은 문장에서 그에 해당하는 개체형이 출현하는 개수에 따라 가변적이다.

3.2 이벤트 템플릿 추출

본 연구에서 대상으로 삼고 있는 이벤트의 정의는 다음과 같다.

- 사건을 서술하는 동사
- Argument 로 2 개 이상의 개체명을 갖고 적어도 한 개는 필수 개체형
- 필수 개체형 : 인명, 조직명
- 필수 개체형 이외는 보조 개체형

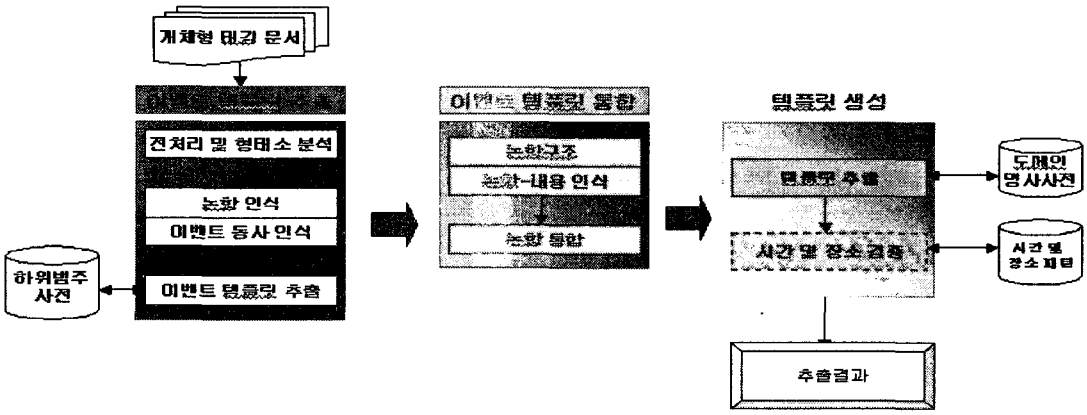

그림 2 이벤트 템플릿을 이용한 정보추출시스템

'누가 언제 어디서 무엇을 어떻게' 라는 사건을 기술하는 문장을 선택하여 이벤트 템플릿을 생성하기 때문에 대상 문서에서 이벤트 템플릿을 생성할 수 있는 문장을 선택하여야 한다.

본 논문에서 정의한 이벤트에 적합한 문장 중에서 용언의 하위범주 정보를 이용하여 이벤트 템플릿을 추출하게 된다.

이러한 이벤트 템플릿의 추출 방법은 사건을 기술하는 문장을 대상으로 이벤트 템플릿을 표현할 수 있다는 장점이 있지만 predicate 으로 사용된 '올리다'의 경우 '공연하다' 라는 의미 외에도 여러가지 뜻으로 쓰일 수 없기 때문에 특정 사건을 인지하는데 모호성을 발생시킨다.

predicate 을 추출할 때 한가지 고려할 점은 동사의 경우 여러가지 형태가 존재할 수 있다는 것이다.

첫째로 보조 용언이 올 수 있다.

이와 같은 경우 형태소 분석기의 결과가 2 개 이상의 동사이어도 하나의 본동사를 찾아서 predicate 으로 결정을 하여야 한다.

둘째로 명사가 동사 파생 접미사(-하-, -되-, -시키-)와 결합하여 동사의 역할을 하는 경우이다. 이 경우도 대부분의 형태소 분석기가 동사로 결과를 내지 않기 때문에 이러한 종류의 파생 동사를 인식하여 predicate 으로 결정을 하여야 한다.

3.3 이벤트 템플릿 통합

하나의 문장에서 생성된 이벤트 템플릿은 하나의 이벤트를 완전하게 표현하지 못 하여 다른 문장에서 생성된 이벤트 템플릿과 통합하여야 하거나 혹은 다른 이벤트와 내용이 중복되어 제거되어야 하는 경우가 있다.

입력된 이벤트 템플릿에 대해서 통합 여부를 결정하기 위해서 먼저 논항의 개수와 분류된 논항의 구조를 인식한다. 인식된 논항 중에서 '누가' '언제' '어디서' '무엇을' '어떻게' 라는 정보를 모두 담고 있다면 중복 관계의 후보로 분류하고 그렇지 않은 경우에는 보완관계의 후보로 분류한다.

인식된 내용을 바탕으로 하여 보완 관계의 경우에는 논항의 타입과 논항의 내용이 일치하는지 여부를 판별하여 최종적으로 보완 관계 여부를 결정하고 중복 관계의 경우에는 이벤트 템플릿이 소유하고 있는 논항의 타입과 논항의 내용이 모두 일치하는지 여부를 판별하여 최종적으로 중복 관계 여부를 결정한다.

보완 관계에 있는 템플릿은 두 개 이상의 템플릿이 통합되어 하나의 이벤트를 기술하는 경우를 말하는데 3 개 이하 논항을 가지면서 하나 이상의 논항이 중복된 형태를 말한다. 중복 관계에 있는 템플릿은 두 이벤트 템플릿의 구조와 컨텐츠가 일치하는 경우를 말하는데 '누가(arg type = 인명 or 조직명)"무엇을(arg type = 제목)"언제(arg type = 시간 or 날짜)' '어디서(arg type = 장소)'가 일치하는 경우이다.

3.4 결과 생성 모듈

생성된 일반적인 이벤트 템플릿에서 필요로 하는 특정 영역의 정보를 추출하여야 한다. 한 문서에서 추출된 이벤트 템플릿은 다양한 정보를 포함할 수 있기 때문에 원하는 특정 영역의 정보만을 선택하는 과정이 필요하다.

본 연구에서는 공연 정보에 대한 정보를 추출하는 것을 최종 목표로 하기 때문에 표 1 과 같은 정보를 갖는 이벤트 템플릿을 선택하여야 한다.

표 1 정보추출 명세 및 내용

공연명	명성황후
공연자	㈜에이콤인터내셔날
공연장소	런던 아폴로헤머스미스극장
공연날짜	내년 2 월 1~16 일

먼저 일반적인 이벤트 템플릿에서 특정 영역에 해당하는 템플릿만을 선택하기 위해서는 학습 데이터를 이용하여 해당 영역의 문서에 자주 출현하는 도메인 명사 사전을 구성한다. predicate 의 argument 인 일반명사 중에서 해당 영역의 명사 사전에 포함되어 있는 이벤트 템플릿은 특정 영역의 이벤트에 속한다고 판

단을 한다. 다음은 공연 도메인 명사 사전의 엔트리이다.

그리고 공연 도메인에서 공연 장소로 쓰일 수 있는 패턴을 조사하여 이벤트 템플릿 중에서 특정 영역에 해당하는 템플릿이라고 판정을 한 경우에 적용하여 공연 장소로 사용될 수 있는지를 검사한다. 공연 장소와 더불어 공연 날짜에 대해서 똑 같은 검사를 반복한다.

장소의 경우 서울이나 미국과 같은 애매한 장소가 공연 장소로 선택되는 것을 피하기 위함이고 날짜의 경우 1980 년대, 일제 강점기 같은 날짜 개체형이 공연 날짜로 선택되는 것을 피하기 위함이다.

본 연구에서 제안하는 이벤트 템플릿을 이용한 정보추출시스템의 구조는 그림 2 와 같다.

4. 실험 및 평가

정보 추출 시스템의 실험을 위해서 한국전자통신연구소에서 구축하고 있는 개체형 태깅 문서 중에서 공연 분야의 문서를 이용하였다. 개체형 태깅 문서는 2001 년 9 월부터 10 월까지 일간지 기사를 사람, 조직명, 제목, 날짜, 시간, 장소등의 범주로 개체형 태깅하였다[9].

개체형 태깅이 된 공연 관련 신문 기사 205 개중에서 직접적으로 공연자, 공연제목, 공연장소, 공연 날짜 및 시간에 대해 언급하지 않았거나 두개 이상의 공연을 병렬적으로 소개한 문서를 제외하고 한 문서에 하나의 공연만을 소개하는 문서 59 개를 선택하여 실험을 진행하였다.

실험을 평가하기 위해서는 공연에 관계된 공연의 주체, 공연제목, 공연장소, 공연시작일, 공연 종료일, 공연 시간에 대해서 정답을 제시한 경우(Cor), 오답을 제시한 경우(Inc), 답을 제시하지 못한 경우(Mis), 존재하지 않는 답을 제시한 경우(Spu)로 나누어 재현률(R)과 정확률(P)을 다음 식과 같이 계산하여 F-Measure 로 평가하였다.

$$R = \frac{Cor}{Cor + Inc + Mis}$$

$$P = \frac{Cor}{Cor + Inc + Spu}$$

위와 같은 식을 사용하여 본 논문의 실험 결과는 재현률 73.9%, 정확률 76.6%이며 F-Measure 는 75.2% 이다.

이러한 실험의 결과는 MUC-7 의 가장 좋은 결과[2]보다는 우수하지만 추출하고자 하는 대상 영역과 정보, 그리고 복잡도가 다르기 때문에 단순 비교를 하는 것은 무리가 있다. 그러나 특정 영역에서 문서의 구조에 기반하지 않고 자연어처리 기술을 사용하였으며 이식성을 고려하여 설계되었다는 측면에서 의의가 있다.

5. 결론

정보 추출은 범람하는 전자문서 중에서 정보 수요자가 원하는 정보만을 추출하여 제시하는 것이다. 이러한 정보 추출은 인터넷의 확장에 비례하여 늘어나는 전자 문서 안에 내재하여 있는 정보에 손쉽게 접근할 수 있다는 장점이 있다.

본 연구는 일반적인 텍스트 문서에서 특정한 정보만을 추출하기 위해서 이벤트 템플릿을 사용하는 방법을 제안한다. 이벤트 템플릿 구조는 일반적인 정보추출 시스템의 단점 중의 하나인 이식성의 문제를 최소화할 수 있는 구조이다. 문서에 있는 문장 중에서 이벤트에 해당하는 문장을 대상으로 형태소 분석 정보와 용언의 하위범주 정보를 이용하여 이벤트 템플릿을 추출하고 추출된 이벤트 템플릿을 검사하여 통합 여부를 결정하여 통합을 한 후에 일반적인 목적으로 추출한 이벤트 템플릿 중에서 정보 수요자가 원하는 정보를 담고 있는 이벤트 템플릿을 도메인 명사 사전과 도메인 제한적인 지식을 사용하여 선택하였다.

향후 정교한 하위범주 사전의 구축과 이를 이용하기 위한 구문 분석 및 의미 해석등을 거쳐 정확한 이벤트 템플릿의 추출 및 이벤트의 흐름에 대한 이해를 통한 정보 추출이 요구되며 이러한 이벤트 템플릿의 구조가 질의/응답, 문서 요약 등의 분야에 성공적으로 적용할 수 있는지 여부도 검증을 해야 할 것이다.

참고문헌

[1] Cardie, C., "Empirical Methods in Information Extraction", AAAI-97, pp.65-79, 1997

[2] C. Aone, L. Halverson, T. Hampton, and M. Ramos-Santacruz, "SRA: Description of the IE2 System used for MUC-7", In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998.

[3] D. Lin, "Using collocation statistics in information extraction", In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998

[4] J. R. Hobbs, D.Appelt, J.Bear, D. Israel, M.Kameyama, M.Stickel, and M. Tyson, "FASTUS: Extracting Information from Natural Language Texts", In Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, MD, November 1995

[5] Ralph Grishman, "Information Extraction: Techniques and Challenges", In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998.

[6] R. Yangarber and R. Grishman. "NYU: Description of the Proteus /PET system as used for MUC-7 ST", In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998

[7] S. Miller, "Algorithm that learn to extract information BBN: Description of the SIFT System as Used For MUC-7", In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998

[8]윤보현, 황이규, 정의석, 임수종, 왕지현, 임명은, "웹 정보 추출의 동향", 인터넷 정보학회지, 2001

[9]황이규,윤보현, "NE/CO 태깅 지침서", ETRI 내부 기술 보고서, 2001