인물 백과사전 지식베이스 구축을 위한 속성패턴기반 정보추출

왕지현[0] 김현진 장명길
한국전자통신연구원 미래기술연구본부 음성/언어정보연구부
{jhwang[0], jini, mgjang}@etri.re.kr

Information Extraction Based on Property Patterns to Construct a Knowledgebase for Encyclopedia Person Domain

JiHyun Wang[0] Hyun-Jin Kim Myung-Gil Jang
Dept. of Speech/Language Technology Research, ETRI

요 약

본 논문은 인물 도메인의 백과사전 지식베이스를 구축하기 위하여 백과사전 본문의 자연어 문장으로부터 인물 표제어의 특징을 잘 나타내는 속성값을 인식하여 추출하는 방법에 관하여 기술한다. 속성은 인물 공통 및 세부 분야별로 총 52개의 속성을 정의하였고 이를 태그셋으로 정의하여 1천 문서의 백과사전 인물 속성태깅 코퍼스를 구축하였다. 속성태깅코퍼스로부터 반자동으로 약 1천 8백여 개의 속성패턴을 추출하였고 백과사전 인물 표제어 24,848개에 대해 속성패턴을 적용하여 지식베이스를 구축하였다. 추출성능은 f-score 0.68의 결과를 나타내었다.

1. 서 론

백과사전(encyclopedia)이란 학문, 예술, 사회, 경제 따위의 과학과 자연 및 인간의 활동에 관련된 모든 지식을 압축하여 부문별 또는 자모순으로 배열하고 풀이한 책을 말한다. 백과사전은 우리말 단어를 풀이해 놓은 국어사전과는 달리, 실세계(real world)의 지식을 일정한 기준에 의해 집대성해 놓은 일반적인 데이터 또는 언어자원을 백과사전이라 할 수 있다.[1]

백과사전은 실세계의 객관적이고 보편적인 지식을 체계적으로 제공한다는 점에서 질의응답 시스템과 같은 지식기반 시스템을 위한 지식베이스 구축의 대상으로 활용되기도 하며, 지식베이스 자체만으로도 백과사전의 요약된 사실정보(fact database)를 제공한다는 점에서 중요한 가치를 갖는다.[2]

백과사전 지식베이스는 표제어의 요약정보를 나타내기 위하여 표제어의 ' 속성(property)' 을 정의하여 지식베이스의 지식단위(unit)로 사용한다. 지식베이스의 속성은 추출출처(source)에 따라 개요정보에 기술되어 있는 속성을 ' 개요속성' 이라고 하고 본문에서 추출된 속성을 ' 본문속성' 이라고 한다.

표제어 문서의 개요정보는 정형화된 형식으로 기술되어 있는 표제어의 요약 정보이다. 미리 정한 형식(pre-defined format)으로 기술되어 있으므로 지식베이스에 저장할 대상이 무엇인 지가 명확하여 속성을 추출하기가 비교적 용이하다. 그러나 본문속성의 경우 문장을 분석하고 문맥을 파악해야만 문장내의 속성을 인식하고 추출할 수 있다.

문장 내에서의 속성은 개체명(Name Entity)과 유사하거나 동일하게 인식될 수 있으나 문맥에 따라 다르게 해석되어 추출된다.

예를 들면, 다음과 같다.

문장 1 : 1970년 6월 *<중국:LOCATION>*에서 사망하였다.

문장 2 : 특산종으로 *<함경북도의 명천:LOCATION>*에서 자생한다.

문장1과 2의 ' 중국' 과 ' 함경북도의 명천' 은 모두 개체명 ' LOCATION' 이다. 하지만 속성은 문장 1에서는 표제어의 ' 사망장소' 를 나타내고, 문장 2에서는 표제어의 ' 자생지' 를 나타낸다. 다시 말해, 개체명은 같은 LOCATION이지만 속성은 서로 다르며, 문맥에 의해 LOCATION의 의미가 결정된다.

2. 관련연구

국내외의 정보추출분야에 대한 연구는 꾸준히 진행되어 왔으나 지식자원의 부족과 언어분석의 한계로 인해 제한된 영역에서의 연구, 개발만이 이루어지고 있다.

국외의 AutoSlog(Riloff, 1993)나 WHISK(Soderland, 1999)의 경우 테러(terror) 영역의 문서에 대한 패턴기반의 정보추출을 시도하고 있으며[3], 국내에서는 POSIE(포항공대)[4]의 교육 및 구인, 구직 분야의 정보추출 시스템이 개발되고 있는 것 이외에도 여러 국내연구기관에서 정보추출 기술을 연구하고 있다.

본 논문은 백과사전 인물분야의 지식베이스를 구축하기 위하여 표제어의 인물 속성을 정의하여 속성패턴을 구축하였으며, 백과사전 본문에서 인물 속성을 추출하여 지식베이스를 구축하는 방법에 대해 기술한다.

3. 백과사전 태깅 코퍼스 구축

3.1 백과사전 원시 코퍼스

지식베이스 구축에 사용한 백과사전은 전체 표제어가 약 16만 개의 표제어로 구성되어 있으며, 동식물, 의학, 철학, 인물 등 총 13개의 카테고리로 분류되어 있다.[5] 이들 중 인물 지식베이스 구축을 위하여 24,848개의 인물 표제어만을 대상으로 지식베이스를 구축하였다.

백과사전의 각 표제어 문서는 그림 1과 같이, 표제어 정보(!ID ~ !DF)와 표제어의 요약정보를 제공하기 위한 개요정보(!SU), 그리고 표제어를 설명하는 본문(!CO)로 구성된다. 이외에도 표제어의 분류체계(category)정보가 함께 기술되어 있다.

```
!ID 69928                          // # 표제어 ID
!TI 박정희                          // # 표제어 이름
!EN                                // # 표제어 영문이름
!BD 1917.11.14                     // # 출생일
!DD 1979.10.26                     // # 사망일
!CN 한국                            // # 국적
!DF 정치가, 군인.                    // # 표제어의 정의
!SU                                // # 개요정보
호 : 중수
활동분야 : 군사, 정치
출생지 : 경북 선산
주요저서 : 《우리 민족이 나아갈 길》, 《민족의 저력》
!CO                                // # 표제어 본문
경북 선산(善山) 출생. 가난한 농부인 박성빈(朴成彬)과 백남의(白南義) 사이에서 5남 2녀 중 막내로 태어났다. 1937년 대구사범학교를 졸업하고, ……
```

그림 1. 백과사전 원시 코퍼스

속성패턴을 구축하기 위한 속성태깅은 인물 전체 24,848개의 문서 중 1,000개를 대상으로 하였으며 개체명 태깅 후에 속성태깅을 하였고 본문(!CO)만을 대상으로 태깅을 하였다.

3.2 속성 태그셋

지식베이스의 속성은 인물의 의미적인 특성에 따라, 인물의 세부 분류체계에 상관없이 모든 인물의 공통적인 특징을 나타내는 속성을 ' 공통속성' , 분류체계별로 나타나는 특징적인 속성을 ' 개별속성' 으로 구분한다. 속성은 공통속성 21개, 개별속성 31개로서 총 52개의 속성을 정의하였다. 다음 표1과 표2는 지식베이스 내의 공통속성과 개별속성의 태그셋 샘플 목록이다.

표 1. 공통속성의 샘플

템플릿 이름	*속성태그*
출생	출생일, 출생장소, 본관, 국적
명칭	별칭
사망	사망일, 사망장소, 사망원인
활동	활동분야
수상	수상명, 수상일
저서	저서명, 출판일

표 2. 개별속성의 샘플

템플릿 이름	*속성태그*
구조물	건축물, 건축일
개발	개발품, 개발일
군대	부대명, 군계급
교육	교육자, 교육물
예술사조	작품
발견	발견물, 발견일
사상	철학사상

4. 패턴기반 속성 추출

4.1 속성별 개체명 분석

백과사전 인물분야의 ETRI 개체명 인식을 위한 태그는 총 76개로 구성되어 있다. 전체 52개의 속성들을 이들 개체명 태그와 비교해본 결과 41개의 속성이 12개의 개체명 태그와 1:1로 대응됨을 알 수 있었다. 41개의 속성태그를 개체명 태그와 대응한 결과는 표3과 같다.

표 3. 개체명 태그와 속성태그의 대응

개체명 태그	속성 태그
DATE (날짜)	출생일, 사망일, 수상일, 출판일, 제작일, 졸업일, 소속일, 건축일, 개발일, 발견일, 사건일, 전쟁일, 설립일, 등단일, 당선일, 대회일, 지정일 (총 17개)
ORGANIZATION (기관, 단체)	졸업학교, 단체(경력), 설립단체, 정당(당선), 부대명(군대), 등단공간 (총 6개)
LOCATION (지역, 국가)	출생장소, 본관, 국적, 사망장소 (총 4개)
POSITION (직위, 직책)	직위(경력), 지위(당선), 군계급 (총 3개)
WORKS (작품명)	저서명, 작품명, 등단작 (총 3개)
EVENT (사건, 전쟁, 대회)	소송사건, 전쟁명, 대회(성적) (총 3개)
PERSON (사람)	교육자(교육) (총 1개)
PRIZE (상 이름)	수상명(수상) (총 1개)
THEORY (이론)	철학사상 (총 1개)
QUANTITY or TIME (양, 시간)	경기성적 (총 1개)
CULTURALASSET (문화재)	문화재명 (총 1개)

대응된 속성의 개수가 가장 많은 개체명 태그는 DATE 이었고, 그 다음이 ORGANIZATION, LOCATION 순이었다. 표3의 이외의 나머지 11개의 속성은 해당 개체명이 없거나 정의된 속성의 개념이 다수의 개념에 속하거나 아예 개체명이 아닌 경우이다. 예를 들어, 속성 ' 주요업적' 이나 ' 주요평가' 는 개체명에 대응되지 않는 속성에 해당한다.

4.2 속성 정보추출

속성패턴기반의 속성정보추출은 크게 다음의 세가지 과정으로 이루어 진다.

① 후보패턴 수집. 속성패턴을 구축하기 위하여 먼저 속성태깅코퍼스에 패턴수집을 위한 템플릿을 적용하여 후보패턴을 수집하였다.

그림2의 패턴수집 템플릿은 속성태깅코퍼스에서 태깅된 속성값을 중심으로 좌우의 문맥을 패턴화하여 구축한다. 패턴을 구성하는 토큰은 NE(개체명), 품사 또는 단어이며 추출 대상인 속성값은 반드시 NE이어야 한다. 패턴을 구성하는 좌우 문맥은 거리 2까지의 토큰을 보도록 하였으며 패턴 내에 반드시 하나의 단어 토큰을 보도록 하였다. 단어는 형식형태소가 아닌 실질형태소만을 보았다.

	t_{-2}	t_{-1}	t_0	t_1	t_2
템플릿 1 :	단어		NE		
템플릿 2 :	NE		NE		
템플릿 3 :		단어	NE		
템플릿 4 :		NE	NE		
템플릿 5 :		품사	NE		
템플릿 6 :			NE	단어	
템플릿 7 :			NE	NE	
템플릿 8 :			NE	품사	
템플릿 9 :			NE		단어
템플릿 10 :	단어	NE	NE		
템플릿 11 :			NE	NE	단어
템플릿 12 :		NE	NE	단어	
템플릿 13 :		단어	NE	NE	
템플릿 14 :		NE	NE	NE	

그림 2. 패턴수집 템플릿

② 최종 속성패턴 구축. 자동으로 수집한 후보패턴에서 수작업을 하여 약 1천 8백 여 개의 최종 속성패턴을 걸러내었다. 최종 속성패턴을 결정하는 조건으로는 패턴 출현 빈도수를 3개 이상으로 하였고, 주변 토큰 중 단서가 되는 단어나 NE토큰이 있는 패턴을 속성패턴으로 선정하였다.

수작업으로 필터링하여 최종적으로 구축한 속성패턴의 샘플은 다음의 그림3과 같다.

```
출생.출생장소   :=   <DATE> @<LOCATION>출생/NN
출생.출생장소   :=   <DATE> @<LOCATION>태어나/VV
출생.출생장소   :=   <DATE> @<LOCATION>출생/NN하/SV
출생.출생장소   :=          @<LOCATION>출생/NN
출생.출생장소   :=          @<LOCATION>태생/NN
출생.출생장소   :=          @<LOCATION>태어나/VV
출생.출생장소   :=          @<LOCATION>출생/NN하/SV
출생.출생장소   :=          @<LOCATION>출신/NN
출생.출생장소   :=          @<LOCATION>$x     출생/NN
출생.출생장소   :=          @<LOCATION>$x     태어나/VV
출생.출생장소   :=          @<LOCATION>$x     출생/NN하/SV
출생.출생장소   :=          @<LOCATION>$x     출신/NN
```

그림 3. 속성패턴의 샘플

@다음의 개체명 <LOCATION>은 현재 토큰으로 추출될 속성값을 나타내며, 좌우의 토큰들이 속성값을 잡아내기 위한 문맥정보가 된다.

③ 속성값 추출. 속성패턴은 백과사전의 입력문장을 언어분석한 결과와 비교하여 매칭된 결과를 속성값으로 추출하였다.

5. 성능평가

인물 도메인 문서 24,848개 중 1,000문서를 속성태깅하여 패턴을 추출하였고 1,000 문서들 중 100문서를 임의로 선택하여 평가하는 작업을 10번을 시도하여 평균을 내었다. 그 결과 f-score 0.68(R 0.73, P 0.64)의 평가 결과를 내었다.

6. 결론 및 고찰

본 논문은 인물 도메인의 백과사전 지식베이스 구축을 위하여 인물 표제어의 속성을 정의하였고 정의된 속성의 패턴기반 정보추출 방법을 구현하였다.

실험결과에 따르면 질의응답시스템 등의 지식기반 응용시스템에 제대로 활용하기 위한 충분한 성능에는 미치지 못하지만 지속적인 보완에 의해 성능을 향상시킬 수 있을 것이다.

향후 연구에서 보완되어야 할 사항을 정리하면 다음과 같다.

- 정의된 지식베이스 속성이 백과사전 표제어들의 특징을 얼마나 잘 나타내는 가?
- 개체명에 해당하는 속성값만을 대상으로 추출하기 때문에 개체명이 아닌 속성값은 추출하지 못한다. 예를 들어, " ~ *레이더*를 개발하였다." 의 경우, ' 레이더' 는 일반 보통명사로서 개체명이 아니다.
- 속성패턴의 단어토큰으로 실질형태소외에도 형식형태가 중요한 의미를 갖는다. 예를 들어, " <~학교:ORGANIZATION>를 졸업하였다." 와 같은 문장의 경우, 격조사' 를' 이 용언 ' 졸업하다' 와 함께 속성값 ' 학교' 를 추출하기 위한 중요한 단서가 된다.
- 패턴구축의 수작업 오류 보정 및 커버리지를 위한 패턴의 일반화(generalization)가 필요함.

참고문헌

[1]최호섭, 옥철영, 김창환, 왕지현, 장명길, " 질의응답 시스템을 위한 백과사전 기반 지식베이스와 온톨로지", 제15회 한글 및 한국어 정보처리학술대회 자료집, pp.177-183, 2003

[2]애니퀘스천 백과사전 질의응답시스템, http://anyq.etri.re.kr

[3]Ion Muslea, " Extraction Patterns for Information Extraction Tasks: A Survey" , Proceedings of the American Association for Artificial Intelligence, 1999

[4]포항공대 정보추출 시스템 POSIE, http://nlp.postech.ac.kr/Research/POSIE/search.php

[5]두산세계대백과 엔싸이버, http://www.encyber.com