CRF를 이용한 백과사전 도메인의 템플릿 기반 지식베이스 설계 및 구축

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

Template-based Knowledgebase Design and Construction using Conditional Random Fields in Encyclopedia Domain

Ji-Hyun Wang[O], Chang-ki Lee, Hyeon-Jin Kim, Myung-Gil Jang
Speech/Language Information Research Department ETRI

요 약

본 논문은 백과사전 도메인의 지식베이스 설계 및 통계기반 정보추출 방법을 이용한 속성정보 인식에 대하여 기술한다. 총 13개 카테고리로 구성된 백과사전에 대해 99개의 템플릿과 285개의 속성을 정의하였으며, 각 표제어의 추출 대상인 속성정보는 표제어를 설명하는 본문에서 통계기반 기계학습모델인 CRF(Conditional Random Fields)를 적용하여 추출하였다. 백과사전 카테고리 별로 균일하게 선정된 4천 5백 문서를 학습에 사용하였고, 테스트 문서셋 500문서에 대해 속성인식률을 측정하였다. 성능 평가한 결과, F1 55.76% (P 74.89%, R 44.42%)의 성능을 나타내었다.

1. 서 론

백과사전은 실세계(Real World)의 객관적이고 보편적인 지식을 체계적으로 기술하고 있기 때문에 템플릿 기반의 질의응답 및 정보추출 시스템의 응용 도메인으로 적합하다. 또한, 백과사전 도메인의 지식베이스는 지식베이스 자체만으로도 다양한 지식기반 응용 시스템에 활용될 수 있는 사실정보(Fact Database)를 제공하기 때문에 매우 중요한 가치를 갖는다.

템플릿 기반의 정보추출 기술은 자연어로 작성된 문서에서 의미있는 정보를 추출하여 테이블 형태의 구조정보로 저장하는 기술이다. 이 분야에 대한 연구는 1950년대에도 있었으나 연구분야로 자리 잡은 것은 1991년의 MUC(Message Understanding Coreference)-3 이후의 일이다. MUC-3부터 MUC-7에 참여한 시스템들이 사용한 접근방법에는 규칙에 기반한 방법[6,7], 통계 기법을 이용한 방법[1,3], 그리고 모든 자연어 처리 기법[2,4]을 이용한 방법이 있다.

기존의 정보추출 시스템들이 사건, 사고, 테러 및 기업정보 등 한정된 도메인에 대하여 시스템을 구축한 반면, 본 논문의 시스템에서는 보다 더 넓은 분야의 문서들에 대한 실험을 수행하기 위해 추출 대상 도메인을 백과사전으로 선정하였다. 백과사전은 정치, 경제, 사회 등 인간생활 전반에 걸친 정보뿐만 아니라 동식물, 지리, 과학 등 다양한 분야로 구성되어 있기 때문에 백과사전 지식베이스를 위한 템플릿의 개수가 훨씬 많이 요구되며 본문의 추출 대상 문장의 표현이 보다 다양하다. 본 논문에서는 백과사전의 분야별로 템플릿 및 템플릿을 구성하는 속성들을 수작업으로 정의하였으며, 백과사전 각 표제어의 본문에 통계기반 교사학습방법을 적용하여 표제어에 대한 사실 정보(Fact)를 자동으로 추출하여 템플릿을 생성하였다. 표제어 단위로 생성된 템플릿은 백과사전 지식베이스를 구성한다.

2장은 백과사전 지식베이스의 구조와 템플릿 설계에 대해 기술하며, 3장은 통계기반 정보추출 방법으로 널리 사용되는 CRF(Conditional Random Fields) 와 ME(Maximum Entropy Model) 학습을 위한 자질 구성 요소 및 설계에 대해 설명한다. 4장은 전체 성능 및 분야별 속성의 성능에 대해 분석하고 결론을 맺는다.

2. 백과사전 지식베이스 구조

2.1 속성 및 인스턴스

백과사전은 신문기사나 웹 문서와는 달리, 국어사전과 같이 표제어 중심으로 기술된다. 따라서 백과사전 지식베이스는 표제어에 대한 정보를 구축하는 것이 가장 적합하다 할 것이다. 백과사전 지식베이스는 ' 표제어(Title)' 와 표제어의 ' 속성(Property)' 및 ' 속성제약(Subtype of Property)' 그리고 속성에 대한 속성값 즉, ' 인스턴스(Instance)' 로 구성된다. 예를 들어, 미국의 컴퓨터 소프트웨어 회사인 표제어 ' 마이크로소프트사' 는 기업체에 해당하는 표제어로서, ' 설립자' , ' 설립일' , ' 본사소재지' , ' 총자산' , ' 매출액' 등이 표제어의 속성이 될 수 있다. 이들은 템플릿 ' 기업체' 를 구성하는 속성이 되며 모든 기업체 표제어가 공통으로 갖는 속성들이다. 그리고 속성 ' 설립자' 에 대한 값으로서 ' 폴 앨런' , ' 빌게이츠' 가 인스턴스가 된다.

다음 그림 1은 기업체 표제어인 ' 마이크로소프트사' 의 속성 및 인스턴스를 나타낸다.

기업체(표제어 : □마이크로소프트사□)
- 국적 : 미국
- 업종 : 컴퓨터 소프트웨어
- 설립자 : 폴 앨런, 빌 게이츠
- 본사소재지 : 워싱턴주 레드먼드
- 총자산 : 371억 56백만 달러(1999)
- 매출액 : 197억 47백만 달러(1999)

그림 1. 템플릿 □기업체□의 속성과 인스턴스

총 13개의 Top카테고리로 구성된 백과사전에서 수작업으로 정의된 전체 템플릿 개수는 99개이며 속성의 개수는 285개이다.

다음은 템플릿 설계에 대한 가이드라인이다.

1) 객체(Object)를 중심으로 템플릿을 생성한다.
   예) 인물, 학교, 국제회의, 휴양지, 경기장 등
2) 하나의 템플릿에 속성이 너무 많으면 별도의 템플릿으로 나눈다. 예를 들어, 템플릿 ' 기업체' 내의 속성인 ' 총자산' 과 ' 매출액' 을 템플릿 ' 기업자산' 이란 이름의 별도의 템플릿으로 정의한다.
3) 추상적인 개념의 템플릿을 정의한다.
   예) 설립, 소장, 개최 등
4) 모호하거나 의미가 유사한 속성들은 통합한다.
   예) 속성 ' 창설일' , ' 준공일' , ' 개관일' 을 ' 설립일' 로 통합
5) 템플릿 간의 상속 개념을 둔다.
   예)
   부모 템플릿 ' 작품' (작품명, 제작자, 제작일, 대본)
   자식 템플릿 ' 영화' (감독, 상영시간, 제작사, 배우)

2.2 속성제약

속성제약은 속성의 의미를 제약함으로써 보다 정확한 의미를 부여하기 위한 것이다. 예를 들어, 속성으로 템플릿 ' 크기' 의 속성 ' 길이' 를 정의하였다면 표제어의 본문에서는 다음 문장에서 중괄호로 태깅된 부분이 인스턴스가 될 수 있다. 다음 문장은 동식물 표제어인 ' 가창오리' 의 한 문장이다.

문장 1 : 몸길이 {약 40cm@크기.길이}, 날개길이 {약 21cm@크기.길이}이다 .

즉, 하나의 문장에서 ' 약 40cm' 와 ' 약 21cm' 가 속성 ' 길이' 로 인식될 수 있다. 그러나 이 수치정보만 추출해서는 템플릿을 생성할 때 각 길이가 표제어의 어느 부위의 길이인지 구별할 수가 없게 된다. 따라서 ' 약 40cm' 와 ' 몸길이' 를, ' 약 21cm' 와 ' 날개길이' 를 각각 한 쌍으로 추출해야 만 한다.

전체 속성 수 285개 중, 17개의 속성이 속성제약을 필요로 하였다.

3. 지식베이스 구축

3.1 속성정보 인식

표제어 본문으로부터 해당 표제어의 속성에 대한 인스턴스를 자동으로 추출하기 위하여 통계기반 기계학습 방법인 CRF(Conditional Random Fields)와 ME(Maximum Entropy Model)를 각각 적용하여 비교 실험하였다.

CRF는 조건부확률계산에 기반한 단방향 그래프 모델(Unidirected graphical model)이다. CRF는 ME나 HMM(Hidden Markov Model)의 레이블 편향(Label Bias) 문제의 영향을 받지 않기 때문에 학습속도는 느리지만 인식 성능은 다른 알고리즘보다 낳다.[5]

속성은 개체명(Name Entity)과 유사하지만 같은 개체명이라 하더라도 문맥에 따라 다르게 해석될 수 있는 특징이 있다.

문장 1 : {<빈대학교:ORGANIZATION>@학력.출신학교}에서 의학을 공부하고 1891년에 졸업하였다.
문장 2 : 1926년 {<버클리신학교:ORGANIZATION>@설립.설립단체}를 설립하여 신학교수를 지냈다.

예를 들어, 문장 1과 2의 ' 빈대학교' 와 ' 버클리신학교' 는 모두 개체명 ' ORGANIZATION' 이다. 하지만 속성은 문장 1에서 표제어의 ' 출신학교' 를 나타내고, 문장 2에서는 표제어가 설립한 ' 설립단체' 를 나타낸다. 다시 말해, 개체명은 같은 ORGANIZATION이지만 속성은 서로 다르며, 보다 많은 문맥정보로부터 ORGANIZATION의 의미를 결정해야만 한다. 개체명이 지역적인 문맥(local context)에 의해 주로 결정되는 반면, 속성은 보다 먼 거리의 문맥(long distance context)에 의해 영향을 크게 받는다. 즉, 문장 1에서 ' 빈대학교' 가 개체명 ' ORGANIZATION' 으로 인식되기 위해서 단어의 suffix인 ' 대학교' 가 중요한 자질로 사용될 수 있다. 그러나 속성에서는 suffix ' 학교' 만 가지고 출신학교로 인식하면 문장 2의 경우에 잘못 인식될 확률이 높아지게 된다. 따라서 예제 문장에서 볼 수 있는 것과 같이 속성을 결정하기 위해서는 중심 용언까지의 문맥을 봐야 한다.

토큰 시퀀스를 $G^n = g_1g_2g_3\cdots g_n$이라 할 때, 속성 태깅은 가장 적합한 속성태그 시퀀스 $P^n=p_1p_2p_3\cdots p_n$을 결정하는 작업이다. 속성태깅을 위한 토큰과 자질 집합을 표현하면, 토큰은 $g_i=<F^n_i, m_i>$라 정의되며 $M^n = m_1m_2m_3\cdots m_n$은 형태소 시퀀스이고 $F^n_m = f_1f_2f_3\ldots f_n$은 자질 집합을 나타낸다. 그리고 문장에서 속성의 경계를 나타내기 위하여 각각의 속성태그를 속성의 시작(B)과 계속(I) 그리고 속성 아닝(O)을 나타내는 BIO의 클래스로 나누었다. 따라서 각 토큰에 대해 분류할 전체 속성클래스 수는 571개이다. 285(속성의 총 개수) * 2(B,I) + 1(O) = 571.

다음 표 1은 속성을 인식하기 위하여 정의된 자질표이다. (m=형태소, e=어절, p=품사, c=청크값, pt=속성값, s=stopword, sfx=suffix, pfx=prefix, +-=방향, 숫자=거리)

자질타입	자질	설명
m	m±2,m±1, m0	좌우 2거리의 형태소
m-s-	m-s-3, m-s-2, m-s-1	앞방향 3거리까지의 형태소. 불용어 제외
m-	m-	거리정보 없는 앞방향 형태소(3개까지)
m0-sfx	m0-sfx3, m0-sfx2, m0-sfx1	현재 형태소의 Suffix(3음절까지)
e+1-pfx	e+1-pfx3, e+1-pfx2, e+1-pfx1	다음 어절의 Prefix (3음절까지)
p	p-2-1,p-10,p0,p0+1,p+1+2	품사. 좌우 2거리 품사
c	c-10,c0+1	청크경계값(BIO)
verb	verb	문장 중심용언
pt	pt-2-1, pt-1, pt-	앞방향 2거리까지의 속성태그

표 1. 속성태깅을 위한 자질표

3.2 속성제약 인식

속성제약은 전체 속성들 중에서 17개에만 해당하여 개수가 적기 때문에 LSP패턴을 구축하여 추출하였다.

다음은 속성제약을 추출하는 예제문장과 패턴이다.

(^:문자열, !:POS, &:개체명, []:Option, %:부분매치, {}:추출위치)

-예제문장 : " 몸길이 10cm." , " 다리길이 5mm이고" , " 꼬리길이가 5m나 되며"
-패턴 : 크기.길이 = {%길이} [!jx|!jc] &LENGTH [^.]

중괄호 ' {' 와 ' }' 사이의 문자열이 추출되어 템플릿 ' 크기' 의 속성' 길이' 에 대한 속성제약이 추출된다. 예제문장에 있어서, ' 몸길이' , ' 다리길이' , ' 꼬리길이' 가 속성제약이 된다.

4. 성능평가 및 결론

추출된 속성정보에 대한 성능 평가를 위해 백과사전 카테고리에서 균일하게 선정된 5천 문서를 수작업으로 속성태깅하였다. 속성태깅 5천문서 중, 임의의 500문서를 테스트 문서로 실험하였고 나머지 문서들을 학습에 이용하였다. 표 2에서 볼 수 있는 바와 같이 CRF는 ME보다 성능이 약 5% 정도 높은 것을 알 수 있다.

학습알고리즘	F1	정확률	재현율
CRF	55.76	74.89	44.42
ME	50.59	71.74	39.07

표 2. 전체 성능결과

표 3은 카테고리별 성능결과이다.

카테고리	F1	정확률	재현율
동식물	78.89	71.70	75.13
서적	84.75	52.63	64.94
문화예술	32.10	61.9	21.67

표 3. 카테고리별 성능

전체 성능결과가 높지 못한 이유는 카테고리별 성능결과에 의해 파악될 수 있다. 동식물과 서적 카테고리의 인스턴스들은 대부분 단일 명사구 내에서 쉽게 인식되기 쉬운 유형들이 많았고 주로 크기, 먹이, 색, 모양 등 단편적인 정보들이 많이 기술되어 있는 반면, 문화예술 카테고리의 인스턴스들은 관형구를 포함한 2개 이상의 명사구로 이루어진 경우가 많으며, 표제어를 설명하기 위한 배경 설명으로 인해 문장이 비교적 긴 경향이 많았다.

지식베이스의 특성상 F-score를 올리기 위해 재현율을 높이기 보다는 재현율을 낮추더라도 정확률을 높은 수치로 유지해야 할 필요가 있다.

본 논문은 백과사전 도메인의 지식베이스를 구축하기 위해 템플릿 기반의 지식베이스를 설계하였고 CRF를 이용하여 자동으로 지식베이스를 구축하였다. 그리고, 백과사전의 카테고리별 성능분석을 통해서 템플릿 기반의 정보추출 접근방법이 적합한 카테고리도 있지만 그렇지 않은 카테고리도 있다는 점을 알 수 있었다. 향후에는 전체 성능을 향상시키기 위하여 통계자질의 보강을 좀더 진행할 계획이다.

참고문헌

[1] A. Borthwick and J. Sterling, Eugene Agichtein and Ralph Grishman, *Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition*, In Proceedings of the Sixth Workshop on Very Large Corpora, pp.152~160. 1998.

[2] C. Huyck, *Description of the American University in Cairos System Used for MUC-7*, In Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, MD, April 1998.

[3] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, *Nymble : a high-performance learning named-finder*, In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp194~201, 1997.

[4] J. Fukumoto, F. Masui, M. Shimohata, and M. Sasaki, *Oki Electric Industry: Description of the Oki System as Used for MUC-7*, In Proceedings of the Seventh Message Understanding Conference(MUC-7), Columbia, MD, April 1998.

[5] J. Lafferty, A. McCallum, and F. Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. ICML, 2001.

[6] 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.

[7] 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.