개체명 인식을 위한 개체명 사전 자동 구축

전원표○, 송영길, 최맹식, 김학수
강원대학교, IT대학, 컴퓨터정보통신공학전공
{nlpjwp, nlpyksong, nlpmschoi, nlpdrkim}@kangwon.ac.kr

Automatic Construction of a Named Entity Dictionary for Named Entity Recognition

Wonpyo Jeon○, Yeongkil Song, Maengsik Choi, Harksoo Kim
Program of Computer and Communication Engineering,
College of Information Technology,
Kangwon National University

요 약

개체명 인식기에 대한 연구에서 개체명 사전은 필수적으로 필요하다. 그러나 공개된 개체명 사전은 거의 없기 때문에, 본 논문에서는 디비피디아의 데이터로부터 개체명을 효과적으로 추출하여 자동으로 구축할 수 있는 방법을 제안한다. 제안 방법은 엔트리의 '이름'과 '분류' 정보를 사용한다. 엔트리의 '이름'은 개체명으로 사용하고, 엔트리의 '분류'는 각 개체명 클래스와의 상호정보량을 계산하여 엔트리와 개체명 클래스 사이의 점수를 계산한다. 이렇게 계산된 점수를 이용하여 개체명과 개체명 클래스를 매핑한다. 그 결과 76.7%의 평균 정확률을 보였다.

주제어: 개체명 사전, 디비피디아, 상호정보량

1. 서론

개체명 인식은 자연어에서 사용된 인명, 지명, 조직명 등의 정보를 식별하여 해당 클래스로 분류하는 작업을 말한다. 개체명은 대부분의 문서에서 중요한 핵심어 역할을 하기 때문에 자연어처리 연구에 다양하게 응용된다. 특히 도메인에 독립적인 질의응답 시스템(Open Domain Question Answering)에서는 많은 종류의 개체명을 인식할 수 있는 개체명 인식이 필수적이다[1]. 이러한 개체명 인식을 위해서는 개체명 사전이 필요하다. 그러나 개체명 사전을 구축하기 위해서는 많은 인력과 시간이 소비된다. 또한 어렵게 개체명 사전을 구축하더라도 지속적으로 생성되는 개체들을 추가하고, 관리하는 것은 현실적으로 많은 어려움이 있다. 따라서 본 논문은 디비피디아(DBpedia) 데이터를 이용하여 개체명을 효과적으로 추출하여 자동으로 구축할 수 있는 방법을 제안한다.

2. 관련 연구

개체명 사전을 구축하기 위해서는 개체명 클래스 분류를 정의해야 한다. 개체명 클래스의 분류는 연구 분야에 따라 서로 다른 정의를 사용하고 있으며, 그에 대한 많은 연구가 있었다. MUC-6(Message Understanding Conference)에서는 'PERSON', 'LOCATION', 'ORGANIZATION'의 3개 분류를 사용하고 있으며, 이것은 여러 연구에서 대분류 개체명 인식의 기본 클래스로 사용되고 있다[2]. BBN(Bolt Beranek and Newman)은 QA(Question Answering)를 위한 가이드 라인에서 29개의 클래스 계층 구조를 정의하였다[3]. 또한 Sekine는 박물관, 강 등 많은 세밀한 하위 클래스를 포함하고 있는 개체명 클래스 계층 구조를 정의했다[4]. Tkachenko는 위키피디아(WikiPedia) 데이터를 사용한 개체명 인식을 위해 BNN과 Sekine의 클래스 계층 중에서 15개의 메인 클래스와 3개의 보조 클래스를 정의하여 사용했다[5].

본 논문의 내용과 유사한 연구로 배상준[6]의 연구가 있다. 배상준[6]은 위키피디아 데이터를 이용하여 개체명 사전을 자동 구축하는 연구를 하였다. 위키피디아 엔트리(entry) 내에 있는 '분류' 정보를 이용하여 개체명 클래스의 분류체계를 구성하고, 엔트리를 매핑(mapping)시키는 방법이다. 이때, 양질의 분류체계를 얻기 위해 불확실성 측정기법을 사용하여 불확실성이 높은 분류체계를 제거하는 방법을 사용하였다. 그러나 이 방법은 새로운 개체명 분류체계를 만들어 개체명 사전을 만드는 방법이기 때문에 기존의 개체명 분류를 이용하는 연구에는 그래도 사용하기에 어려움이 있다. 그래서 본 논문에서는 정의된 개체명 분류체계에 개체명을 효과적으로 매핑하는 방법을 제안한다.

3. 개체명 인식기

본 연구는 개체명 인식기 연구의 선행 연구로 수행하

*이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2013R1A1A4A01005074) 또한 본 논문은 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발]

였다. [그림 1]은 본 연구팀에서 연구하고 있는 개체명 인식기의 구조를 보여준다.

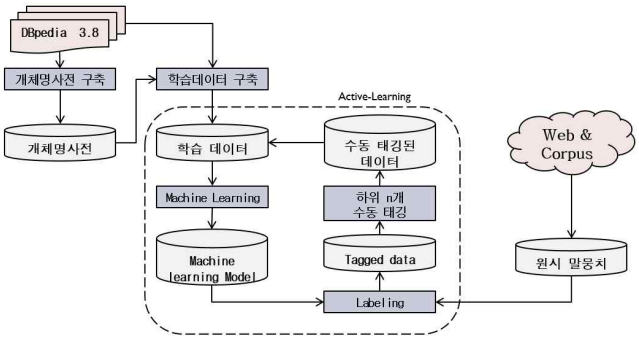

그림 1. 개체명 인식기 구조도

기계학습 기반의 개체명 인식기의 연구를 위해서는 개체명이 태깅된 대용량의 말뭉치가 필요하다. 그러나 국내에는 공개된 개체명 태깅 말뭉치가 거의 없기 때문에, 직접 말뭉치를 구축해야 한다. 말뭉치 구축을 위해 수작업으로 개체명 태깅을 하는데 시간적 인적 자원의 소모가 크므로 개체명 사전이 필수적이다.

4. 개체명 사전 자동 구축

4.1. 개체명 클래스 정의

<표 1>은 본 논문에서 정의한 개체명 클래스이다.

표1. 개체명 클래스 정의

번호	Tkachenko	본 연구	비고
1	PERSON	PERSON	
2	GEOLOGICAL_REGION	LOCATION	의미 확장
3	FACILITY	FACILITY	
4	AUTO	VEHICLE	의미 확장
5	GAME	GAME	
6	PLANT	PLANT	
7	PERIOD	PERIOD	
8	COMPUTER	COMPUTER	
9	DISAMBIGUATION_PAGE	-	삭제
10	GPE	GPE	
11	ASTRAL_BODY	-	삭제
12	ORGANIZATION	ORGANIZATION	
13	WORK_OF_ART	WORK_OF_ART	
14	ANIMAL	ANIMAL	
15	INSECT	INSECT	
16	LANGUAGE	LANGUAGE	
17	LIST_OF_PAGE	-	삭제
18	OTHER_PAGE	OTHER	이름 변경
19	-	EVENT	추가
20	-	THEORY	추가
21	-	FOOD	추가

<표 1>의 개체명 클래스는 Tkachenko가 정의한 개체명 클래스 일부를 수정하여 정의한 것으로, 향후 진행할 개체 관계추출 연구를 위해 기존의 개체명 클래스에서 추가 및 삭제, 수정을 하였다. <표 1>의 의미 확장은 기존 개체명 클래스의 한정적인 의미 때문에 디비피디아의 엔트리를 커버하기 어려운 문제를 보완하기 위한 것이다. 예를 들어 'AUTO'의 경우 자동차를 의미하는데 'VEHICLE'로 확장함으로써 탈 것에 대한 엔트리를 커버할 수 있다.

4.2. 개체명 사전 자동 구축

개체명 사전 자동 구축을 위해 디비피디아 데이터를 사용하였다. 디비피디아는 위키피디아의 데이터를 RDF(Resource Description Framework) 형태의 파일로 저장한 것으로 위키피디아의 거의 대부분의 데이터를 저장하고 있다. 본 논문에서는 개체명 사전 자동 구축을 위해 디비피디아의 데이터 중 엔트리의 '제목'과 '분류'를 사용하였다.

[그림 2]는 개체명 사전 자동 구축의 과정을 보여준다.

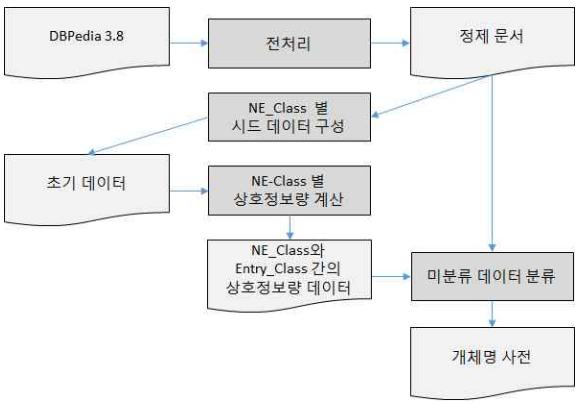

그림 2. 개체명 사전 자동 구축 과정

디비피디아의 데이터로부터 엔트리의 '제목'과 '분류'를 가져온다. 이때 '분류'에 대하여 전처리를 수행한다. '분류'는 여러 어절로 구성될 수 있는데, 본 논문에서는 '분류'의 마지막 어절만을 가져왔다. 이와 같은 방법으로 얻어진 정제 문서로부터 일정 분량의 엔트리를 각각의 개체명 분류로 수작업 분류하여 시드 데이터를 구성한다. 본 논문에서는 각 클래스별로 30개의 엔트리를 시드 데이터로 사용하였다.

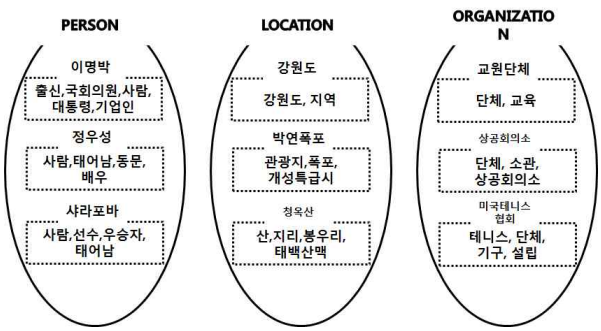

그림 3. 시드 데이터 예

[그림 3]과 같은 초기 데이터로부터 수식 (1)을 이용하여 개체명 클래스와 각 엔트리의 '분류' 정보간의 상호정보량을 계산한다.

$$I(t_i;C_j)=p(t_i,C_j)\log\frac{p(t_i\wedge C_j)}{p(t_i)\cdot p(C_j)} \quad (1)$$

수식 (1)에서 t_i는 엔트리의 '분류'로 [그림 3]에서 점선으로 표현된 "출신", "사람" 등을 의미한다. C_j는 "PERSION", "LOCATION" 등과 같은 개체명 클래스를 의미한다. 이렇게 계산된 값과 수식 (2)를 이용하여 초기 데이터 이외의 엔트리별 점수를 계산하여 개체명 클래스를 매핑한다.

$$Score(E)=\arg max_j \sum_i I(t_i;C_j) \quad (2)$$

이 방법은 엔트리를 구성하는 모든 분류들의 점수를 합산하여 계산하므로 특정 분류 때문에 잘못 분류되는 경우를 처리할 수 있다. 예를 들어 "미국 테니스 협회"의 분류 정보 중 "테니스"가 "ORGANIZATION"이 아닌 다른 클래스와 높은 상호정보량 값을 가지더라도 나머지 분류 정보 "단체", "기구", "설립" 때문에 최종적으로 "ORGANIZATION" 클래스에서 높은 점수가 나온다.

5. 구축 결과

본 논문에서는 개체명 사전 구축을 위해 한국어 디비피디아를 이용하였다. 한국어 디비피디아의 엔트리 중 '분류' 정보를 가지고 있는 엔트리는 186,520개이다. 이중 개체명 사전에 추가할 대상은 단일 '분류' 정보를 가지고 있는 엔트리는 56,031개를 제외한 130,489개로 제한한다. 그 이유는 단일 '분류' 정보를 가지고 있는 엔트리는 각 개체명 클래스와 상호정보량 점수를 구할 수 없기 때문이다. <표 3>은 각 클래스 별로 매핑된 개체명의 수이다.

표 3. 매핑된 개체명의 수

클래스	매핑된 수	클래스	매핑된 수
GPE	8,642	ANIMAL	1,524
PERSON	41,066	INSECT	443
LOCATION	13,924	PLANT	854
WORK_OF_ART	8,730	FOOD	1,445
ORGANIZATION	11,986	FACILITY	10,728
VEHICLE	1,328	LANGUAGE	1,172
THEORY	3,688	COMPUTER	1,166
GAME	3,522	PERIOD	36
EVENT	5,457	OTHER	14,779

구축된 개체명 사전의 신뢰도를 측정하기 위해서는 정답 비교를 위한 개체명 사전이 필요하다. 그러나 공개된 개체명 사전을 찾지 못하였기 때문에 정확한 신뢰도를 측정할 수 없었다. 그래서 차선책으로 분류된 개체명 중 각 클래스 별로 임의로 100개씩 샘플링(sampling)하여 정확률을 측정하였고, 총 5회 반복하여 평균 정확률을 계산하였다. 태깅된 정답이 없기 때문에 재현율은 계산하지 못했다. 매핑된 개체가 36개인 'PERIOD' 클래스의 경우 전체를 대상으로 측정한다. <표 4>는 클래스별로 각 샘플링한 개체명의 평균 매핑 정확률을 나타낸다.

표 4. 평균 매핑 정확률

개체명 클래스	평균 정확률	개체명 클래스	평균 정확률
GPE	0.615	ANIMAL	0.588
PERSON	0.988	INSECT	0.620
LOCATION	0.778	PLANT	0.840
WORK_OF_ART	0.840	FOOD	0.715
ORGANIZATION	0.808	FACILITY	0.858
VEHICLE	0.890	LANGUAGE	0.725
THEORY	0.855	COMPUTER	0.790
GAME	0.695	PERIOD	0.861
EVENT	0.573	OTHER	-

실험 결과 가장 높은 정확률을 보인 것은 PERSON으로 98.8%의 정확률을 보였고, 가장 낮은 정확률을 보인 것은 EVENT로 57.3%의 정확률을 보였다. 전체적으로 평균 76.7%의 정확률을 보였다.

6. 결론

본 논문에서는 개체명 인식기 구축에 필수적인 개체명 사전을 자동으로 구축하는 방법을 제안하였다. 개체명 클래스의 분류로 Tkachenko의 개체명 클래스를 수정하여 총 18개의 개체명 클래스 분류를 만들었다. 개체명 사전의 구축 대상으로 위키피디아의 정보를 저장하고 있는 디비피디아의 정보 중 엔트리 '이름'과 '분류' 정보를 이용하였다. 엔트리의 '이름'과 개체명 클래스를 매핑하기 위해 엔트리를 구성하는 '분류'들과 개체명 클래스 간의 상호 정보량의 합을 사용하였다. 그 결과 개체명 클래스 분류에서 평균 76.7%의 정확률을 보였다.

향후 연구로 개체명 사전의 정확률 향상과 더불어 개체명 인식기, 개체간의 관계추출에 대한 연구를 진행할 예정이다.

참고문헌

[1] Satoshi Sekine, "Named entity: history and future", Technical Report, New York University, 2004.

[2] MUC-6, http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html (2013.09.13 확인)

[3] Annotation guidelines for answer types, http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html (2013.09.13. 확인)

[4] Definition of Sekine`s Extended Named Entity, http://nlp.cs.nyu.edu/ene/version6_1_0eng.html (2013.09.13. 확인)
[5] Maksim Tkachenko, Alexander Ulanov and Andrey Simanovsky, “Fine grained classification of named entities in wikipedia”, Technical report, HP Laboratories, 2010.
[6] 배상준, 고영중, “한국어 위키피디아를 이용한 분류체계 생성과 개체명 사전 자동 구축”, 정보과학회논문지(B), 제16권, 제4호, pp. 492-496, 2010.