휴리스틱을 이용한 개체명 인식 학습 말뭉치 품질 향상

이성희○, 송영길, 김학수
강원대학교 IT대학 컴퓨터정보통신공학전공
nlpfclee@kangwon.ac.kr, nlpyksong@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

Improving Quality of Training Corpus for Named Entity Recognition Using Heuristic Rules

Seong-Hee Lee○, Yeong-Kil Song, Hark-Soo Kim
Department of Computer and Communications Engineering College of Information Technology, Kangwon National University

요 약

개체명 인식은 문서에서 개체명을 추출하고 추출된 개체명의 범주를 결정하는 작업이다. 기존의 지도 학습 기법을 이용한 개체명 인식을 위해서는 개체명 범주가 수동으로 부착된 대용량의 학습 말뭉치가 필요하며, 대용량의 말뭉치 구축은 인력과 시간이 많이 들어가는 일이다. 본 논문에서는 학습 말뭉치 구축비용을 최소화하고 초기 학습 말뭉치의 노이즈를 제거하여 말뭉치의 품질을 향상시키는 방법을 제안한다. 제안 방법은 반자동 개체명 사전 구축 방법으로 구축한 개체명 사전과 원거리 감독법을 사용하여 초기 개체명 범주 부착 말뭉치를 구축한다. 그리고 휴리스틱을 이용하여 초기 말뭉치의 노이즈를 제거하여 학습 말뭉치의 품질을 향상시키고 개체명 인식의 성능을 향상시킨다. 실험 결과 휴리스틱 적용을 통해 개체명 인식의 F1-점수를 67.36%에서 73.17%로 향상시켰다.

주제어: 개체명 인식, 원거리 감독법, 휴리스틱

1. 서론

개체명(named entity)이란 고유한 의미를 지니는 단어 또는 단어열을 뜻하고 인명(person), 지명(location), 기관명(organization) 등과 같은 범주(categories)를 가지고 있다. 개체명 인식(named entity recognition)은 문서에서 개체명을 추출하고 추출된 개체명의 범주를 결정하는 작업이다. 개체명을 인식하기 위한 연구로는 통계 기반의 기계 학습 방법이 주를 이루고 있다[1]. 기계 학습 기반의 지도 학습(supervised learning)을 이용한 개체명 인식은 개체명 범주가 부착된 대용량의 말뭉치가 필요하며 말뭉치 구축을 위해서는 인적, 시간적 비용이 매우 많이 든다. 최근에는 이러한 문제를 해결하기 위해 상대적으로 비용이 적은 준지도 학습법(semi-supervised learning)[2]이 많이 사용되고 있다. 준지도 학습법의 일종인 원거리 감독법(distant supervision)[3]을 이용한 개체명 인식은 개체명 사전을 이용하여 문장에 자동으로 개체명 범주를 부착한다. 하지만 개체명 사전을 구축하기 위해서도 많은 비용이 필요하다는 단점이 있다. 본 논문에서는 말뭉치 구축비용을 최소화하기 위해 반자동 개체명 사전 구축 방법[4]으로 구축한 개체명 사전과 원거리 감독법을 이용하여 초기 개체명 범주 부착 말뭉치를 생성한다. 그리고 자동화 구축 방법으로 인해 발생한 초기 말뭉치의 노이즈(noise)를 2단계 휴리스틱을 통해 효율적으로 제거함으로써 개체명 인식 성능을 향상시키는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 기술하고, 3장에서는 본 논문의 제안 방법, 4장에서는 실험 및 평가, 5장에서는 결론에 대해서 기술한다.

* 본 연구는 엔씨소프트 산학연구용역 과제의 지원을 받아 수행되었음. 또한 우수기술연구센터 사업 중 “링크드데이터 기반 대화형 질의응답 검색 프레임워크 개발 (과제번호 : 10048448)” 과제의 지원을 받아 수행되었습니다.

2. 관련 연구

기존의 개체명 인식에 대한 연구로는 기계 학습 기반의 지도 학습이 많이 사용되어 왔다. 지도 학습을 이용한 개체명 인식은 높은 성능을 보이지만 대용량의 학습 말뭉치가 필요하다는 단점이 있다. 최근 이를 해결하기 위해서 원거리 감독법을 이용한 준지도 학습의 개체명 인식 연구[5]가 진행된 바 있다. 하지만 준지도 학습 기반의 개체명 인식은 자동으로 학습 말뭉치를 구축하기 때문에 많은 노이즈가 발생한다. 본 논문에서는 적은 비용으로 말뭉치의 노이즈를 효과적으로 제거하는 방법으로 2단계 휴리스틱을 제안한다.

3. 개체명 인식 학습 말뭉치 품질 향상

본 논문에서 제안하는 개체명 인식 시스템의 전체 구조도는 [그림 1]과 같다.

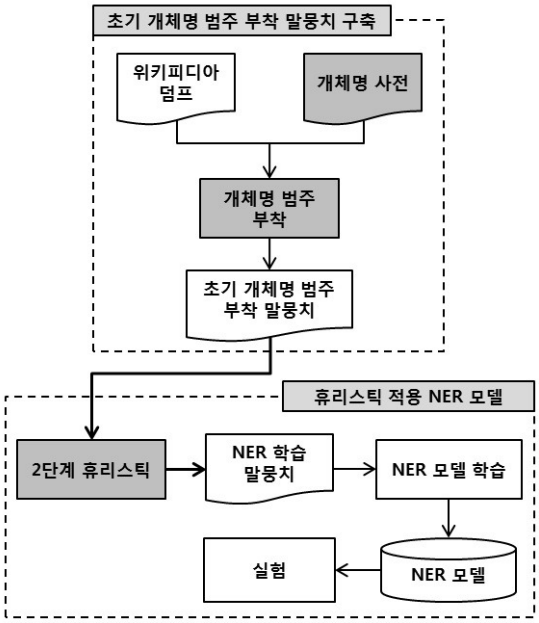

그림 1 시스템 구성도

[그림 1]에서 보는 것과 같이 제안 시스템은 2단계로 구성된다. 첫 번째 단계에서는 개체명 사전을 이용한 원거리 감독법을 이용하여 위키피디아 문장에 개체명 범주를 자동으로 부착하고 초기 개체명 범주 부착 말뭉치를 생성한다. 두 번째 단계에서는 초기 개체명 범주 부착 말뭉치에 2단계 휴리스틱을 적용하여 노이즈를 제거하고 NER(Named Entity Recognition) 학습 말뭉치를 생성한다. 그리고 CRFs(Conditional Random Fields)[6]를 이용하여 NER 모델을 학습하고 실험을 진행한다.

3.1. 초기 학습 말뭉치 구축

위키피디아 문서에는 다른 위키피디아 문서로 연결되는 링크(link)가 존재하며 링크의 어휘는 개체명일 가능성이 높다. 위키피디아 문서에서 링크의 어휘와 구축되어 있는 개체명 사전[4]을 최장 일치로 매칭(matching)하는 원거리 감독법을 이용하여 초기 개체명 범주 부착 말뭉치를 생성한다. 링크의 어휘가 아닐지라도 위키피디아 문서에서 출현한 개체명에 대해서는 사전의 각 엔트리(entry)와 최장 일치 매칭을 통해서 범주를 부착한다. [그림 2]는 위키피디아 문장에 개체명 범주를 부착한 예이다.

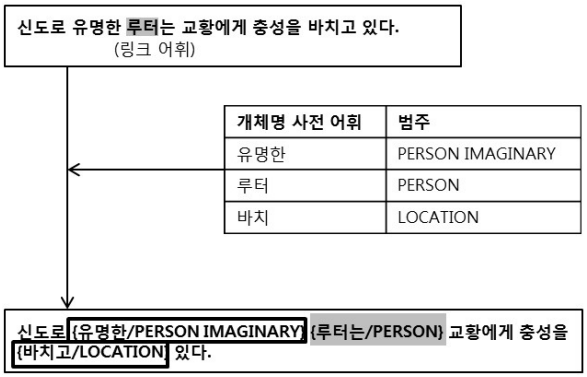

그림 2 위키피디아 문서 개체명 범주 부착 예

3.2. 2단계 휴리스틱을 통한 말뭉치 품질 향상

3.1절에서 개체명 범주 부착이 완료된 초기 말뭉치는 원거리 감독법으로 개체명 범주를 자동으로 부착했기 때문에 많은 노이즈를 포함하고 있다. 노이즈를 효과적으로 제거하기 위해서 2단계의 휴리스틱을 이용하여 잘못 부착된 범주를 제거한다. 첫 번째 단계로 개체명은 체언을 대상으로 하기 때문에 형태소 분석기를 이용하여 개체명 범주가 용언 부분에 부착된 것들을 제거한다. [그림 3]은 용언 부분의 개체명 범주 제거 예이다.

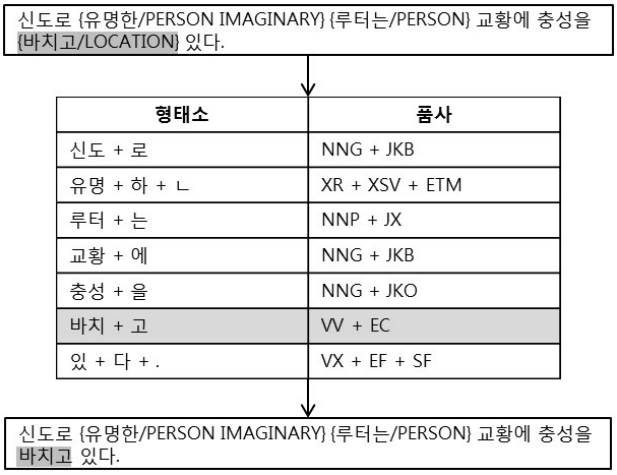

그림 3 용언 부분 개체명 범주 제거 예

[그림 3]에서 사용한 품사는 세종 형태소 품사이다. 세종 형태소 품사의 용언(VV, VA, VX, VCP, VCN) 뒤에 명사형 전성 어미(ETN)가 붙지 않는 경우에 개체명 범주를 제거한다.

두 번째 단계로는 고유한 의미를 가지는 개체명은 일반적인 단어보다는 출현할 확률이 낮을 것이므로 말뭉치에서 개체명의 출현 빈도수를 계산하고 10회 이상 출현한 개체명을 선별하여 해당 개체명이 부착된 문장 중에 잘못 부착된 것들을 수동으로 제거한다. [표 1]은 말뭉치에서 10회 이상 출현한 개체명 중 수동으로 제거한 개체명의 목록 일부이다.

표 1 개체명 수동 제거 목록

개체명 어휘	빈도
이후	2392
주도	705
동안	628
이전	408
유명한	399
승리	319
달리	198
. . .	. .

[표 1]의 개체명 어휘 목록의 범주들을 [그림 3]의 용언 부분의 개체명 범주 제거 문장에서 추가로 제거한 문장의 경우 “신도로 유명한 {루터는/PERSON} 교황에 충성을 바치고 있다.”이다. 초기 학습 말뭉치에서 2단계 휴리스틱을 모두 적용한 문장들을 대상으로 NER 학습 말뭉치를 생성한다.

3.3. 개체명 인식 모델 학습

기존의 개체명 인식 연구들은 개체명의 경계를 형태소 단위나 음절 단위로 인식하는 경우가 많았다. 형태소 단위 개체명 인식[7,8]은 형태소 분석 결과를 이용하여 개체명의 경계를 인식한다. 하지만 형태소 분석에서 오류가 발생할 경우 개체명 인식에도 오류가 전파되며, 음절 단위 개체명 인식은 많은 경계 후보가 발생하고 언어학적 정보가 부족하다는 단점이 있다. 본 논문에서는 이러한 문제를 피하기 위해 어절 단위의 개체명 인식을 수행한다. 먼저 어절 단위로 개체명을 인식하고 조사나 어미같은 형식어를 휴리스틱으로 제거한다. 어절 단위 개체명 범주 태그(tag)는 개체명의 범위를 지정해주는 BIO 태그와 B에 부착되는 개체명 범주 태그로 구성된다. BIO 태그는 개체명의 시작에 해당하는 B, 개체명의 중간 또는 끝에 해당하는 I와 비개체명에 해당하는 O로 구성된다. 예를 들어 3.2절에서 2단계 휴리스틱 적용이 완료된 문장의 경우 "신도로/O 유명한/O 루터는/B_PER 교황에게/O 충성을/O 바치고/O 있다./O"와 같이 태그가 부착된다. 기계 학습을 위해서 각 어절별로 사용된 자질은 [표 2]와 같다.

표 2 입력 자질 목록

자질 이름	설명
LEX	현재 어절의 어휘
FW_2_Lex, BW_2_Lex	현재 어절과 ±1 어절의 앞 2음절 어휘, 뒤 2음절 어휘
FW_2Tags, BW_2_Tags	현재 어절과 ±1 어절의 앞 2음절 어휘, 뒤 2음절 어휘가 일치하는 개체명들의 범주명
FW_3_Lex, BW_3_Lex	현재 어절과 ±1 어절의 앞 3음절 어휘, 뒤 3음절 어휘
FW_3Tags, BW_3_Tags	현재 어절과 ±1 어절의 앞 3음절 어휘, 뒤 3음절 어휘가 일치하는 개체명들의 범주명
BIEF (BE, BF, IE, IF)	BE : 현재 어절이 개체명의 시작이고 어절 전체가 개체명 BF : 현재 어절이 개체명의 시작이고 어절 일부가 개체명 IE : 현재 어절이 개체명의 중간 또는 끝이고 어절 전체가 개체명 IF : 현재 어절이 개체명의 중간 또는 끝이고 어절 일부가 개체명
POS_Bigram	현재 어절과 ±1 어절의 품사 바이그램
LEX-POS_Unigram	현재 어절과 ±1 어절의 '형태소-품사' 유니그램

4. 실험 및 평가

본 논문에서는 3.1절의 방법으로 구축한 초기 학습 말뭉치로부터 학습 데이터 55,000 문장과 테스트 데이터 1,000 문장을 추출하였다. 테스트 데이터에 대해서는 수작업을 통해서 정답 개체명 범주를 부착하고 개체명 사전은 위키피디아 표제어로부터 반자동화 개체명 사전 구축 방법[4]을 이용하여 구축한 것을 사용하였다. 실험에 사용된 개체명 범주는 11종류(person, location, organization, celestial body, event, facility, game, language, law, person imaginary, study field)이다. 원거리 학습법에 기반 한 기존 방법[5]과 기존 방법에 3.2절의 휴리스틱을 적용한 방법의 성능을 측정하였다. [표 3]은 성능 측정 결과이다.

표 3 성능 비교

모델	정확률 (%)	재현율 (%)	F1-점수 (%)
원거리 감독	72.63	62.80	67.36
원거리 감독 + 휴리스틱	85.87	63.74	73.17

[표 3]의 '원거리 감독' 은 기존의 원거리 감독법을 기반으로 한 개체명 인식 모델이다. '원거리 감독 + 휴리스틱' 은 기존 방법에 3.2절에서 제안한 휴리스틱을 통해 초기 학습 말뭉치의 노이즈를 제거한 개체명 인식 모델을 의미한다. [표 3]에서 보는 것과 같이 원거리 감독법에 간단한 휴리스틱을 적용했음에도 불구하고 정확률을 13.24%, 재현율을 0.94%, F1-점수는 5.81% 향상시킬 수 있었다.

5. 결론

본 논문에서는 개체명 범주 부착 말뭉치 구축비용을 최소화하는 방법을 제안하였다. 제안 방법은 반자동으로 구축한 개체명 사전을 이용한 원거리 감독법으로 개체명 범주를 자동으로 부착하고, 개체명 범주 부착 말뭉치의 노이즈를 줄여 2단계 휴리스틱을 통해 효과적으로 제거하고 말뭉치의 품질을 향상시키고 개체명 인식의 성능 향상을 꾀한다. 실험 결과 2단계 휴리스틱을 적용하여 정확률 13.24%, 재현율 0.94%, F1-점수 5.81%의 성능 향상을 보였다. 실험을 통해 제안 휴리스틱이 초기에 오부착된 개체명 범주를 효과적으로 제거함으로써 정확률을 크게 향상시킬 수 있었다. 향후 연구로는 2단계 휴리스틱을 적용한 후에도 남아있는 노이즈의 유형을 분석하고 최소한의 비용으로 노이즈를 제거하기 위한 방법에 대해 연구할 예정이다.

참고문헌

[1] K. Uchimoto, Q. Ma, M. Murata, H. Ozakum, and H. Isahara, "Named Entity Extraction Based on A ME Model and Transformation Rules," Proc. of the ACL, 2000.

[2] A. Blum, Semi-supervised Learning, Encyclopedia of Algorithms, pp. 1-7, Jan, New York, 2015.

[3] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol.2, pp. 1003-1011, 2009.

[4] Y. Song, H. Kim, "Semi-automatic Construction of a Named Entity dictionary Based on Active Learning," Proc. of the Computer Science and its Applications Lecture Notes in Electrical Engineering, Vol.330, pp. 65-70, 2015.

[5] Y. Kim, “Automatic training corpus generation method of Named Entity Recognition using Big data,” M.S. Thesis, Sogang University, 2015.
[6] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Proc. of the ICML, pp. 282-289, 2001.
[7] C. Lee, M. Jang, “Named Entity Recognition with Structural SVMs and Pegasos algorithm,” Journal of Cognitive Science, Vol.21, No.4, pp. 655-667, 2010.
[8] Y. Park, S. Kang, B. Kyu, and J. Seo, “Title Named Entity Recognition using Wikipedia and Making Acronym,” Proc. of the KIISE Korea Computer Congress 2013, pp. 637-639, 2013.