KACTEIL-NER: 딥러닝과 앙상블 기법을 이용한 개체명 인식기

박건우○, 박성식, 장영진, 최기현, 김학수
강원대학교 컴퓨터정보통신공학과
parkku01@kangwon.ac.kr, a163912@kangwon.ac.kr, buwak07@kangwon.ac.kr, pluto32@kangwon.ac.kr,
nlpdrkim@kangwon.ac.kr

KACTEIL-NER: Named Entity Recognizer Using Deep Learning and Ensemble Technique

Geonwoo Park○, Seongsik Park, Yoengjin Jang, Kihyoen Choi, Harksoo Kim
Kangwon National University Computer and Communication Engineering

요 약

개체명 인식은 입력 문장에서 인명, 지명, 기관명, 날짜, 시간 등과 같은 고유한 의미를 갖는 단어 열을 찾아 범주를 부착하는 기술이다. 기존의 연구에서는 단어 단위나 음절 단위를 입력으로 사용하였다. 하지만 단어 단위의 경우 미등록어 처리가 어려우며 음절 단위의 경우 단어 고유의 의미가 희석되는 문제가 발생한다. 이러한 문제들을 해결하기 위해 본 논문에서는 형태소 단위 개체명 인식기와 음절 단위 개체명 인식기를 앙상블하여 보정된 결과를 예측하는 개체명 인식기를 제안한다. 제안된 모델은 각각의 단일 입력 모델보다 향상된 F1-점수(0.8049)를 보였다.

주제어: 개체명 인식기, 인공 신경망, 앙상블, 형태소 확률 자질, 지명 사전 자질

1. 서론

개체명 인식은 입력 문장에서 고유한 의미를 갖는 단어 열을 찾아 인명, 지명, 기관명, 날짜, 시간과 같은 범주들을 부착하는 기술로, 정보 추출의 핵심 요소이다. 최근 자연어처리 분야에서 효율적인 입력 단위에 대해 많은 연구가 진행되고 있다[1-5]. 많은 개체명 인식 연구들은 단어(형태소) 단위나 음절 단위로 입력된다. 기존의 연구에서 단어 단위 개체명 인식기의 경우 미등록어 개체명 처리가 어려운 문제가 있고[2,3], 음절 단위 개체명 인식기의 경우 단어 고유의 의미가 희석되는 문제가 발생한다[4,5]. 이러한 문제들을 해결하기 위해 본 논문에서는 형태소 단위 개체명 인식기와 음절 단위 개체명 인식기의 결과를 이용해 상호 보완하는 모델을 제안한다.

2. 관련 연구

딥러닝(Deep learning)을 이용한 인공 신경망 기반 개체명 인식에서 성능을 향상시키는 방법 중 하나는 입력되는 단어 표상을 확장시키는 방법이 있다. 단어 표상을 확장시키는 방법 중에는 음절 단위 임베딩을 사용하는 방법, 단어 단위 임베딩을 사용하는 방법, 음절 단위 임베딩으로부터 단어 단위 임베딩 벡터를 유도하는 방법 등 여러 방법들이 존재한다[2-5].

앙상블 모델은 하나 이상의 단일 모델들을 통해 나온 다중 결과 값들을 이용해 단일 결과 값을 도출하는 모델이다[6]. 앙상블 모델에는 결과 값들 중 가장 많이 나온 결과 값을 선택하는 Voting 방법[2]과, 기계학습을 통해 최적의 결과 값을 선택하는 방법[7] 등이 있다.

본 논문에서는 다양한 입력 단위에 존재하는 문제점을 해결하기 위해 형태소 단위 개체명 인식기와 음절 단위 개체명 인식기를 이용한 앙상블 모델을 제안한다.

3. 형태소 음절 상호 보완 개체명 인식기

3.1 모델 구조도

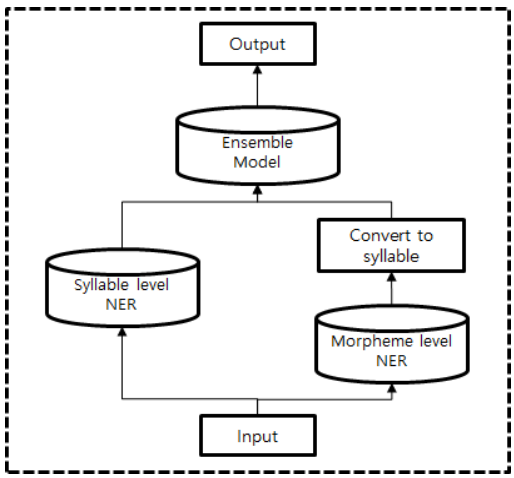

그림 1 형태소 음절 상호 보완 개체명 인식기 구조도

그림 1은 입력 문장을 각각 형태소 단위 개체명 인식기와 음절 단위 개체명 인식기에 입력하여 나온 값들을 이용해 앙상블 모델에 적용하여 최적의 출력 값을 도출

하는 모델의 구조도이다. 입력 문장은 형태소 단위, 음절 단위로 사용하여 형태소 단위 개체명 인식기와 음절 단위 개체명 인식기로 입력된다. 각 모델이 결과를 예측한 후 두 모델의 정보 결합을 위해 형태소 단위 개체명 인식기의 예측 값을 음절 단위로 변환하여 음절 단위 개체명 인식기의 예측 값과 함께 앙상블 모델의 입력이 된다. 마지막으로 두 입력을 통해 상호 보완된 개체명 예측 값을 최종 결과로 출력한다.

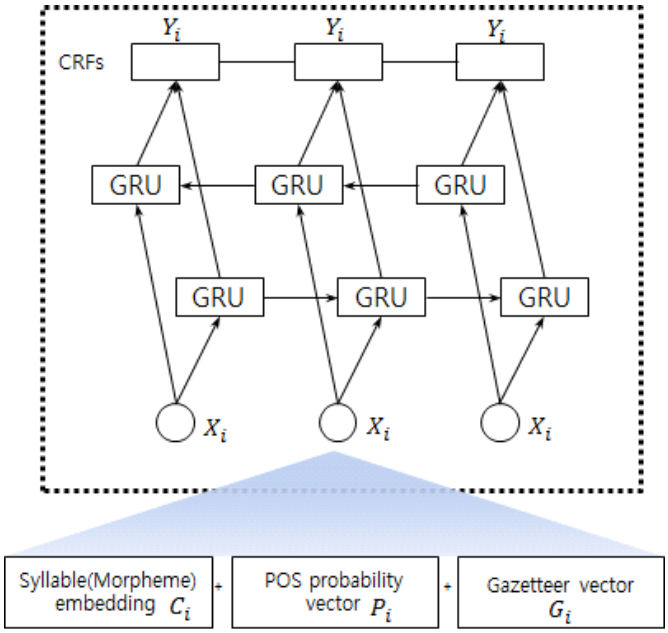

그림 2 개체명 인식기 구조도

그림 2는 음절 단위 개체명과 형태소 단위 개체명에 사용된 Gate Recurrent Unit Recurrent Neural Network Conditional Random Fields(GRU-CRF)[8]의 구조도 이다. 입력 단위의 임베딩과 함께 자질로 사용되는 형태소의 품사 확률 벡터, 지명 사전(gazetteer) 벡터를 결합하여 GRU-CRF의 입력으로 사용한다.

3.2 형태소 단위 개체명 인식기

그림 2와 같이, 형태소 단위 개체명 인식기의 입력은 세 개의 벡터로 구성된다. 형태소 임베딩(morpheme embedding: C_i)은 입력 형태소와 태그 쌍의 임베딩 벡터이고 형태소의 품사 확률 벡터(POS probability vector: P_i)는 형태소가 주어졌을 때 해당 형태소가 어떤 품사를 가지는지 출현 확률을 표현한 벡터이다. 지명 사전 벡터(gazetteer vector: G_i)는 개체명 사전 벡터이다. 개체명 사전 벡터는 개체명 사전의 형태소 분석된 개체명이 발생 하는지를 나타내는 벡터이다.

3.3 음절 단위 개체명 인식기

형태소 단위 개체명 인식기와 마찬가지로, 그림 2와 같이, 음절 단위 개체명 인식기의 입력은 세 개의 벡터로 구성된다. 음절 임베딩(Syllable embedding: C_i)은 입력 음절의 임베딩 벡터이고 음절 품사 확률 벡터(POS probability vector: P_i)는 n-gram 음절의 품사 확률 벡터이고 지명 사전 벡터(gazetteer vector: G_i)는 개체명 사전 벡터이다. 개체명 사전 벡터는 n-gram으로 이루어진 개체명 사전에 개체명이 발생 하는지를 나타내는 벡터이다.

음절의 품사 등장 확률 벡터 P_i를 얻기 위해 본 논문에서는 세종 형태소 말뭉치에서 내용어(고유 명사, 명사, 동사 등)로부터 n-gram 음절을 추출한다. 그 후, 각 n-gram 음절의 빈도를 계산하고 빈도를 확률로 변환한다. 표 1은 '대한민국'의 n-gram 음절을 보여준다. 표 2에서 보이듯이, '대한민국'은 고유 명사의 빈도가 다른 품사 태그의 빈도 보다 높다. 이 사실은 확률 벡터가 고유 명사와 일반 명사를 구별하는 단서가 될 수 있으며 대부분의 개체명이 고유 명사로 이루어져있기에 효과적인 자질이 될 수 있다.

표 1 n-gram 음절의 빈도

Bi-gram	고유명사	명사	동사	형용사	...
대한	1,191	511	0	7	0
한민	454	67	0	0	0
민국	396	29	0	0	0
Tri-gram	**고유명사**	**명사**	**동사**	**형용사**	**...**
대한민	344	2	0	0	0
한민국	344	2	0	0	0

지명 사전 벡터 G_i를 얻기 위해 우리는 개체명 사전에서 n-gram 음절을 추출 한다. 그 후, n-gram 음절 개체명의을 카이 제곱 분포의 확률 밀도 함수[9]를 사용하여 상위 N개의 n-gram 음절을 선택한다. 마지막으로, 입력된 n-gram 음절을 상위 N개의 n-gram 음절 개체명과 일치하는지 여부를 나타내는 이진 벡터로 변환한다. 표 2은 선택된 n-gram 음절의 일부를 보여준다.

표 2 범주별 상위 4개의 n-gram 음절

Bi-PER	Tri-PER	Bi-LOC	Tri-LOC	Bi-ORG	Tri-ORG
_김	_남궁	리_	1동_	사_	식회사
_박	레스_	시_	2동_	주식	주식회
_최	_황보	동_	1리_	식회	학교_
_이		면_	2리_	업_	_한국

표 2에서, 'Bi'와 'Tri'는 각각 bi-gram과 tri-gram을 의미하고 'PER', 'LOC', 'ORG'는 각각 인명, 지명, 기관명을 의미한다. 그리고 '_' 기호는 어절간 띄어쓰기를 의미한다. 'PER'의 bi-gram과 tri-gram은 사람의 이름이고 'LOC'의 bi-gram과 tri-gram은 행정 구역의 부분을 의미하고 'ORG'의 bi-gram과 tri-gram은 회사를 의미하는 단어이다. 이러한 사질은 지명 사전 벡터가 개체명 인식기에 대한 단서가 될 수 있음을 보여준다.

3.4 앙상블 모델

그림 1과 같이 형태소 단위 개체명 인식기와 음절 단위 개체명 인식기의 결과들을 앙상블 모델의 입력으로 이용한다. 앙상블 모델은 두 쌍의 연속된 개체명 태그들을 GRU-CRF의 입력으로 사용하고 한 개의 연속된 개체명 태그를 출력한다. 입력된 개체명 태그들은 각각 50차원의 임베딩 벡터로 구성되어있고 하나의 입력으로 사용하기 위해 100차원의 벡터로 결합한다.

4. 실험 및 결과

4.1 실험 환경

본 논문의 개체명 인식기에서는 2017 국어 정보 처리 시스템 경진대회의 말뭉치를 실험에 사용한다. 전체 말뭉치 3,660개에 Active Bagging[10,11]을 이용하여 추가적인 말뭉치를 생성하였고 평가에 사용된 데이터는 366개를 사용하였다. 개체명 태그는 인명, 지명, 기관명, 날짜, 시간 5개 이며 개체명 경계는 잘 알려진 음절 단위 BIO 태그를 이용하였다.

4.2 실험 결과

표 3은 제안한 모델의 성능을 보여준다. 표 3에서, 'M_NER' 은 각각 초기 값이 다른 3개의 형태소 단위 개체명 인식기이고 'S_NER' 은 각각 초기 값이 다른 3개의 음절 단위 개체명 인식기이다. Ensemble NER' 은 형태소 단위 개체명인식기와 음절 단위 개체명 인식기의 결과를 이용한 앙상블 모델이다.

표 3 성능 비교

	recall	precision	f1-measure
M_NER	0.6736	0.7511	0.7511
S_NER	0.6890	0.8191	0.7484
Ensemble NER	0.7683	0.8452	0.8049

표 3에서 보이듯이, Ensemble 모델이 M_NER 보다 F-1 점수 5.38%p 향상이 있었고, S_NER 보다 F-1 점수 5.65%p 향상이 있었다.

5. 결론

본 논문에서는 형태소 단위 개체명 인식기의 단점인 미등록어 처리 문제와 음절 단위 개체명 인식기의 단점인 단어 고유의 의미를 희석시키는 문제를 상호 보완하여 성능을 향상시키는 형태소 음절 상호 보완 개체명 인식기를 제안하였다. 실험 결과 형태소 및 음절을 입력하는 개체명 인식기보다 0.0538, 0.0565의 성능 향상을 보였다.

감사의 글

이 논문은 2016년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.R-20160906-004163, 빅데이터 자동 태깅 및 태그 기반 DaaS 시스템 개발) 또한, 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2016R1A2B4007732)

참고문헌

[1] 심광섭. "기분석 어절 사전과 음절 단위의 확률 모델을 이용한 한국어 형태소 분석기 복제", 정보과학회 컴퓨팅의 실제 논문지, 제22권, 제3호, pp. 119-126, 2016.

[2] 이성희, 송영길, 김학수. "원거리 감독과 능동 배깅을 이용한 개체명 인식", 정보과학회논문지, 제43권, 제2호, pp. 269-274, 2016.

[3] 최윤수, 차정원. "Word Embedding 자질을 이용한 한국어 개체명 인식 및 분류", 정보과학회논문지, 제43권, 제6호, pp. 678-685, 2016.

[4] 나승훈, 민진우. "문자 기반 LSTM CRF를 이용한 개체명 인식", 한국정보과학회 학술발표논문집, pp. 729-731, 2016.

[5] 유홍연, 고영중. "품사 임베딩과 음절 단위 개체명 분포 기반의 Bidirectional LSTM CRFs를 이용한 개체명 인식", *한글 및 한국어 정보처리 학술대회 논문집*, pp. 105-110, 2016.

[6] 배지윤, 이민혁, 김유중, 태동현, 석준희. "재표집을 활용한 앙상블 인공 신경망 모델", 한국정보과학회 학술발표논문집, pp. 669-671, 2016.

[7] J. Feng, T. Zahavy, B. Kang, H. Xu and S. Mannor, Ensemble Robustness of Deep Learning Algorithms, *arXiv preprint arXiv:1602.02389*, 2016.

[8] C. Lee, LSTM-CRF Models for Named Entity Recognition. *IEICE Transactions on Information and Systems* 100(4), pp. 882-887.2017.

[9] A. D. Ball and G. D. Buckwell, in Statistics A Level, *Edited Macmillan Education*, UK, pp. 186-201. 1991.

[10] S. Lee, Y. Song, M. Choi and H. Kim. Bagging-based active learning model for named entity recognition with distant supervision. *Big Data and Smart Computing (BigComp)*, International Conference on. IEEE, 2016.

[11] 박건우, 이성희, 김학수. "개체명 사전과 원시 말뭉치를 이용한 준지도 학습 기반 개체명 인식 모델", 한국정보과학회 학술발표논문집, pp. 1757-1759. 2016.