Bi-directional LSTM-CNN-CRF를 이용한 한국어 개체명 인식 시스템

이동엽○, 임희석

고려대학교 컴퓨터학과
judelee93@korea.ac.kr, limhseok@korea.ac.kr

Korean Entity Recognition System using Bi-directional LSTM-CNN-CRF

Dong-Yub Lee○, Heui-Seok Lim
Dept. of Computer Science and Engineering, Korea University

요 약

개체명 인식(Named Entity Recognition) 시스템은 문서에서 인명(PS), 지명(LC), 단체명(OG)과 같은 개체명을 가지는 단어나 어구를 해당 개체명으로 인식하는 시스템이다. 개체명 인식 시스템을 개발하기 위해 딥러닝 기반의 워드 임베딩(word embedding) 자질과 문장의 형태적 특징 및 기구축 사전(lexicon) 기반의 자질 구성 방법을 제안하고, bi-directional LSTM, CNN, CRF과 같은 모델을 이용하여 구성된 자질을 학습하는 방법을 제안한다. 실험 데이터는 2017 국어 정보시스템 경진대회에서 제공한 2016klpNER 데이터를 이용하였다. 실험은 전체 4258 문장 중 학습 데이터 3406 문장, 검증 데이터 426 문장, 테스트 데이터 426 문장으로 데이터를 나누어 실험을 진행하였다. 실험 결과 본 연구에서 제안하는 모델은 BIO 태깅 방식의 개체 청크 단위 성능 평가 결과 98.9%의 테스트 정확도(test accuracy)와 89.4%의 f1-score를 나타냈다

주제어: NER, sequence labelling, deep learning

1. 서론

개체명 인식(Named Entity Recognition) 시스템은 문서에서 인명(PS), 지명(LC), 단체명(OG)과 같은 개체명을 가지는 단어나 어구를 해당 개체명으로 인식하는 시스템이다. 개체명 인식을 위한 전통적인 방법으로는 주로 hand-craft된 자질(feature)을 기반으로 학습하는 HMM(Hidden Markov Models), CRF(Conditional Random Fields)와 같은 통계 기반의 모델이 있다[1, 2]. 또한 개체명 인식이나 품사 태깅(Part-of-speech Tagging)과 같이 순서 라벨링(sequence labeling) 문제를 해결하기 위해 자질을 보강(augment) 하기 위한 방법으로 RNN(Recurrent Neural Networks)와 LSTM(Long-short Term Memory)를 활용한 연구가 있다[3, 4]. 최근에는 Bi-directional LSTM와 CNN(Convolutional Neural Network) 그리고 CRF 모델들을 함께 이용하여 end-to-end learning 방식으로 개체명 인식이나 품사 태깅 모델을 학습 할 수 있도록 딥러닝 기반의 모델을 활용한 연구가 있다[5].

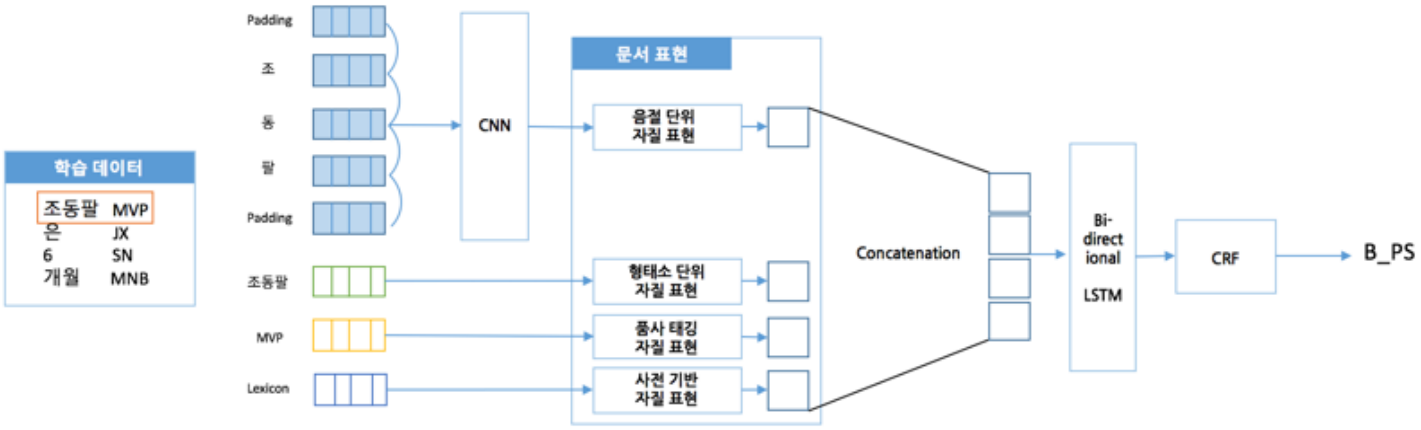
본 연구에서는 개체명 인식 시스템을 개발하기 위해 딥러닝 기반의 워드 임베딩(word embedding) 자질과 문장의 형태적 특징 및 기구축 사전(lexicon) 기반의 자질 구성 방법을 제안하고, bi-directional LSTM, CNN, CRF과 같은 모델을 통해 구성된 자질을 학습하는 방법을 제안한다. 실험은 2017 국어 정보 시스템 경진대회에서 제공한 2016klpNER 데이터를 이용하여 진행하였다. 전체 4258 문장 중 학습 데이터 3406 문장, 검증 데이터 426 문장, 테스트 데이터 426 문장으로 데이터를 나누어 실험을 진행하였다. 실험 결과 본 연구에서 제안하는 모델은 BIO 태깅 방식의 개체 청크 단위 성능 평가 결과 98.9%의 테스트 정확도(test accuracy)와 89.4%의 f1-score를 나타냈다.

2. 제안하는 과정

본 연구에서 제안하는 한국어 개체명 시스템의 전체 구조도는 [그림 1]과 같다. 2017 국어 정보 시스템 경진대회에서 제공하는 학습데이터는 문장이 형태소 단위로 나누어져 있다. “조동팔” 이라는 단어와 그에 해당하는 품사 태깅 결과가 주어졌을 때 제안하는 시스템은 4가지의 방법으로 문서를 표현한다. 음절 단위의 자질 표현을 구성하기 위해 단어를 이루는 음절 단위로 임베딩을 구성한 뒤 CNN을 통하여 음절의 자질을 추출 한 후 이를 음절 단위 자질 표현으로 활용한다. 형태소 단위로 나눠진 단어에 대해 glove vector를 이용한 워드 임베딩을 구성하여 이를 형태소 단위의 자질 표현으로 활용한다. 또한 학습 데이터에 포함되어 있는 품사 태깅 정보를 기반으로 품사 태깅에 대한 임베딩을 구성하여 이를 자질로 활용할 수 있다. 마지막으로 형태소 단위로 나눠진 단어를 대상으로 미리 구축된 기구축 사전을 이용하여 사전 기반의 자질을 표현할 수 있다. 학습 데이터를 이용하여 표현된 각각의 자질들을 연결(concatenation)한 뒤 이를 bi-directional LSTM의 입력으로 사용한다. 그 결과 LSTM은 은닉 상태(hidden states)를 계산하여 출력하고, 이 은닉 상태를 CRF의 입력으로 사용하여 최종적으로 형태소에 대응하는 개체명을 예측한다.

2.1 문서 표현(Document Representation)

2.1.1 CNN을 이용한 한국어 음절 단위의 자질 표현



[그림 1] 본 연구에서 제안하는 한국어 개체명 시스템의 전체 구조도

글자(character) 임베딩을 기반으로 자질을 추출하기 위해 CNN을 이용할 수 있다[6]. 본 실험에서는 CNN의 필터 크기(filter size)와 필터 개수를 다양하게 설정하여 성능 비교를 진행하였다. 최종적으로 2, 3, 4, 5 만큼의 필터 크기와 128 개의 필터 개수를 사용하였을 때, 가장 좋은 성능을 보였다. 또한 글자의 자질 표현 방법에 따른 성능 비교를 진행하기 위해 자소 단위와 음절 단위의 자질 표현 방법을 비교하였다. 그 결과 음절 단위로 글자를 구성하여 자질을 표현할 경우 자소 단위로 글자를 구성하여 자질을 표현하는 경우보다 f1-score가 약 2% 정도 만큼 높은 성능을 보였다.

2.1.2 Glove vector를 이용한 형태소 단위의 자질 표현

형태소 단위의 자질을 표현하기 위해 glove vector를 이용한 임베딩 공간을 구성하였다. Glove vector를 학습하기 위해 한국어로 구성된 위키피디아 데이터 약 345만건을 이용하였다.

2.1.3 품사 태깅 정보를 이용한 자질 표현

학습 데이터에는 앞, 뒤 단어에 대한 연관 성을 고려 할 때 활용될 수 있는 단어에 대한 품사 태깅 정보가 존재하여 이를 자질 표현으로 활용하였다. 그 결과 품사 태깅 정보를 이용하였을 때, 품사 태깅 정보를 이용하지 않았을 때 보다 f1-score가 약 1.9% 향상되었다.

2.1.4 기구축 사전 정보를 이용한 자질 표현

기구축 사전 정보를 이용하여 사전 기반의 자질을 표현하기 위해 gazette 기구축 사전을 이용하였다. 그 결과 기구축 사전 정보를 이용하였을 경우, 기구축 사전 정보를 이용하지 않았을 때 보다 f1-score가 약 1% 향상됨을 알 수 있었다.

2.2 Bi-directional LSTM과 CRF를 이용한 구성된 자질 학습

2.1절에서 구성된 문서 표현을 bi-directional LSTM의 입력으로 사용하여 각 형태소의 정보에 대한 은닉 상태를 계산할 수 있다. Bi-directional LSTM은 주어진 문장에 대해 각각 전향(forward), 후향(backward)으로 문장의 정보를 고려하여 보다 풍부하게 문장 정보를 표현하여 은닉 상태를 계산할 수 있다. CRF는 전 단계에서 계산된 은닉 상태에 대해 조건부 확률(conditional probability)를 계산하여 주어진 형태소에 대응하는 개체명을 예측한다.

2.3 데이터 구성 정보 및 하이퍼 파라미터(Hyper-Parameter) 설정

[표 1]은 본 실험에서 사용한 데이터의 구성 정보 및 각 모델들의 하이퍼 파라미터 설정 값을 나타낸다.

[표 1] 데이터 구성 정보 및 하이퍼 파라미터 설정 값

	Hyper-parameter	Value
Training data	word vocab size	6516
	char vocab size	1899
	entity tag vocab size	7
	morphological vocab size	45
	lexicon vocab size	6
Glove	window size	20
	dimension	100
CNN	filter sizes	2,3,4,5
	number of filters	128
	dropout	0.8
LSTM	initial state	0.0
	state size	600
	dropout	0.8
	training epoch	17
	initial learning rate	0.01
	decay rate	0.9
	char dimension	100

3. 실험 결과

제공된 데이터 전체 4258 문장 중 학습데이터 3406 문장, 검증 데이터 426 문장, 테스트 데이터 426 문장으로 데이터를 나누어 실험을 진행하였다. 성능 평가는 BIO 태깅 방식의 개체 청크 단위 성능 평가를 이용하여 진행하였다. [표 2]는 문서 표현을 하기위한 자질 표현 방법에 따른 한국어 개체명 인식 시스템의 성능을 나타낸다.

[표 2] 문서 표현 방법에 따른 한국어 개체명 인식 시스템 성능

Feature Representation		Accuracy	F1-score
형태소 단위		97.4	78.4
형태소 단위 + 글자	자소 단위	97.5	84.1
	음절 단위	97.8	86.2
형태소 단위 + 음절 단위 + 품사 태깅 정보		98.3	88.1
형태소 단위 + 음절 단위 + 품사 태깅 정보 + 사전 정보		**98.9**	**89.4**

4. 결론

본 연구에서는 2017 국어 정보시스템 경진대회에서 제공한 2016klpNER 데이터를 이용하여 한국어 개체명 인식 시스템을 개발하였다. 실험은 전체 4258 문장 중 학습 데이터 3406 문장, 검증 데이터 426 문장, 테스트 데이터 426 문장으로 데이터를 나누어 진행하였다. 문서 표현을 구성하기 위해 CNN을 이용한 한국어 음절 단위의 자질 표현, glove vector를 이용한 형태소 단위의 자질 표현, 품사 태깅 정보, 기구축 사전 정보를 이용하였다. 구성된 문서 표현을 bi-directional LSTM의 입력으로 사용하여 은닉 상태를 계산한 후 CRF의 입력으로 사용하여 최종적으로 형태소에 대응하는 개체명을 예측 하였다. 실험 결과 본 연구에서 제안하는 모델은 98.9%의 테스트 정확도(test accuracy)와 89.4%의 f1-score를 나타냈다.

참고문헌

[1] Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In Proceedings of CoNLL-2009, pages 147-155.

[2] Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In Proceedings of EMNLP-2015, pages 879-888, Lisbon, Portugal, September.

[3] Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In Proceedings of ICASSP- 2013, pages 6645-6649. IEEE.

[4] Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308.

[5] Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs- CRF. In Proc. of ACL.

[6] Kim, Yoon. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 .