문자 기반 LSTM-CRF 한국어 개체명 인식을 위한 사전 자질 활용

민진우○, 나승훈
전북대학교
jinwoomin4488@gmail.com, nash@jbnu.ac.kr

Lexicon Feature Infused Character-Based LSTM CRFs for Korean Named Entity Recognition

Jin-Woo Min○, Seung-Hoon Na
Chonbuk National University, Chonbuk National University

요 약

문자 기반 LSTM CRF는 개체명 인식에서 높은 인식을 보여주고 있는 LSTM-CRF 방식에서 미등록어 문제를 해결하기 위해 단어 단위의 임베딩 뿐만 아니라 단어를 구성하는 문자로부터 단어 임베딩을 합성해 내는 방식으로 기존의 LSTM CRF에서의 성능 향상을 가져왔다. 한편, 개체명 인식에서 어휘 사전은 성능 향상을 위한 외부 리소스원으로 활용하고 있는데 다양한 사전 매칭 방법이 파생될 수 있음에도 이들 자질들에 대한 비교 연구가 이루어지지 않았다. 본 논문에서는 개체명 인식을 위해 다양한 사전 매칭 자질들을 정의하고 이들을 LSTM-CRF의 입력 자질로 활용했을 때의 성능 비교 결과를 제시한다. 실험 결과 사전 자질이 추가된 LSTM-CRF는 ETRI 개체명 말뭉치의 학습데이터에서 F1 measure 기준 최대 89.34%의 성능까지 달성할 수 있었다.

1. 서론

LSTM CRF는 순차 입력열 태깅을 위한 딥러닝 모델로, 최근 품사 태깅 및 개체명 인식 문제에서 높은 성능을 보여주고 있다 [1,3-6]. LSTM CRF는 입력 문자열을 LSTM을 이용하여 양방향으로 은닉벡터를 얻고 출력 태그간의 의존성을 CRF로 모델링하여 품사 태깅 및 개체명 인식에서 높은 성능을 보이고 있다 [1,3].

새롭게 생성되고 사라지는 개체명의 특성 때문에 발생하는 미등록어 문제를 해결하기 위하여 단어를 구성하고 있는 문자들을 딥러닝을 활용하여 해당 하는 단어의 임베딩을 CNN, LSTM 네트워크를 통해 합성하는 방식들이 제안되었다 [2-4,6-7].

한편 개체명 인식에서의 성능 향상을 위해 사전 등 외부 리소스를 활용하는 방안들이 연구되었다[2,8]. 본 논문에서는 다수의 리소스로부터 개체명 사전을 구축하고 사전의 매칭 방법을 다양화하여 각 매칭법의 장단점을 구분하고 매칭법을 조합하여 한국어 개체명 인식에 활용하고자 하였다. 최적의 사전 매칭조합을 자질로 사용한 결과 ETRI 개체명 말뭉치의 학습데이터에서 F1 measure 기준 [4]에 제시된 86.53%의 더욱 개선시켜 89.34%의 성능을 얻었다.

2. 관련 연구

개체명 인식을 위한 딥러닝 연구로는 순차 입력열 태깅에 높은 성능을 보이고 있는 LSTM에 출력 노드 간의 의존성을 모델링하는 CRF를 결합한 방식이 많은 연구에서 보여지고 있다 [1,3-6].

이에 미등록어 문제를 해결하기 위하여 입력 단어의 표상을 구성하기 위해 [2-4,6-7] 단어 자체의 임베딩 벡터 뿐만 아니라 단어를 구성하고 있는 문자들을 LSTM 혹은 CNN 네트워크로 합성하여 임베딩 벡터를 생성하는 방법들이 연구되었다. [4]에서는 LSTM, CNN을 합성 방식을 결합한 LSTM-CNN 하이브리드 합성 방식을 제안하고 LSTM, CNN을 개별로 문자를 합성한 결과보다 나은 성능을 보여주고 있다.

또한 최근 개체명 인식 시스템에서 외부 자질의 한 형태로 사전 자질을 활용하고 있는데 [2,8]에서는 영문 개체명 인식에서 사전 자질을 활용한 LSTM-CRF의 성능 향상을 보여주고 있다.

3. 사전 자질을 이용한 문자 기반 LSTM CRF

본 논문에서는 한국어 개체명 인식에서 영문에서와 같이 사전 자질을 활용하여 개체명 인식의 성능을 높이고자 하였다. 기존 문자 LSTM-CNN 기반한 단어 표상 방식에 품사 정보, 띄어쓰기 정보에 사전 자질을 추가하여 확장 단어 표상을 구성하였다.

3.1 개체명 인식을 위한 확장 단어 표상

문자 기반으로 합성하여 얻은 임베딩 벡터와 단어 임베딩 벡터를 결합하여 기본 단어 표상을 얻은 후 형태소의 품사와 띄어쓰기 정보 그리고 새롭게 추가한 사전 자질등과 결합하여 확장 단어 표상을 얻는다. 그림 1은 LSTM CRF의 입력이 되는 확장 단어표상을 보여주고 있다.

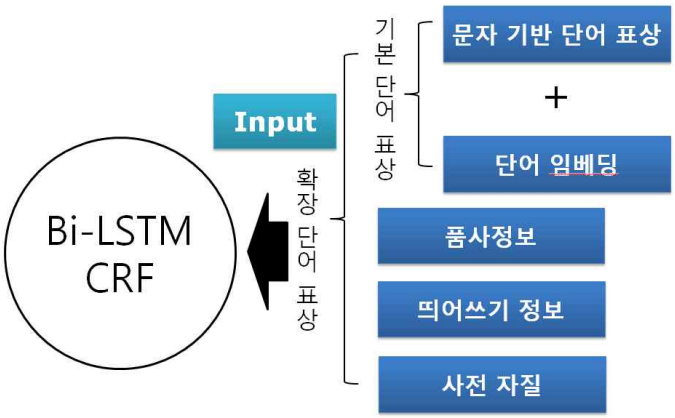

그림 1. 확장 단어표상

3.2 사전구축

ETRI 범용 개체명 인식 데이터 셋의 개체명 범주는 PS(인명), LC(지명), OG(기관), DT(날짜표현), TI(시간표현)이다. 여기서 사전 구축이 어려운 DT, TI를 제외하고 PS, LC, OG 세 범주에 대하여 한국어 위키 백과에서 추출한 위키 사전과 위키 사전과 네이버 지식백과, 세종 고유명사, 영문 위키 번역 사전을 통합한 통합사전을 구축하였다. 이후 이상 데이터들을 수작업으로 제거한 후 CRF 형태소 분석기를 적용하여 사전의 각 엔트리를 한국어 개체명 인식의 입력 단위인 형태소 단위로 분할하여 사전 매칭의 정확도를 높이고자 하였다. 구축한 사전의 엔트리 수는 표 1과 같다.

표 1. 위키 사전과 통합사전의 범주별 개체명 엔트리 수

	PS	LC	OG
위키	67401	41019	27681
통합	109488	100464	152322

3.3 사전의 매칭

사전의 매칭은 [2]의 BIEOS(Begin, Inside, End, Outside, Single) 표기법을 사용하여 매칭하였다. 이것은 매칭된 엔트리 내의 각 부분의 위치를 나타낸 것으로 엔트리가 단순 매칭되는 것인지를 표기하는 Yes/no 표기법 보다 나은 성능을 보여준다.

[2]에서는 사전을 매칭할 때 최장 매칭을 우선으로 하여 부분 매칭을 한 후 전체 매칭으로 덮어쓰는 방식으로 사전을 매칭하였다.

본 논문에서는 [2]과 달리 부분매칭은 고려하지 않고 전체(정확)매칭을 베이스로 최장, 최빈도, L2L3LL의 3가지 방식으로 매칭 한 후 3가지 방법을 조합하여 CRF로 성능을 측정하여 최고 성능을 보이는 사전 자질 조합을 몇 가지 결정 후 LSTM CRF에서 실험을 진행하였다.

자질의 조합은 하나의 자질로써 합하는 것이 아닌 각 방법들이 개별의 자질 임베딩 벡터로 사용된다. 사전의 3가지 매칭 법에 대한 예시는 표 2에서 보이는 것과 같다.

표 2. 사전이 적용되는 방법에 대한 예시

Text	우리	금융	그룹	은	,
최장	B-OG	I-OG	E-OG	O	O
최빈도	B-OG	E-OG	E-OG	O	O
L2	L2-OG	O	O	O	O
L3	L3-OG	O	O	O	O
LL	O	O	O	O	O

최장 매치는 문장을 구성하는 토큰들이 사전에 매칭이 되었을 때 BIOES 표기법에 따라 시작은 B, 끝은 E, 나머지 중간을 차지하는 부분은 I로 매치한다.

최빈도 매치는 최장 매치와 기본적으로 같지만 사전에 매치가 되었을 때 사전 자체를 구성하는 토큰들이 사전 토큰의 B,I,E,S 중 어느 부분에 가장 최빈도로 매칭되었는지의 여부를 매치한다.

L2, L3, LL매치는 사전의 단일(Single) 토큰 매치가 아닌 각각 2개 , 3개, 4개 이상의 매치 일 때 매칭 되는 토큰의 시작에 L2, L3, LL등을 표기해주는 방법이다.
또한 좀 더 정확한 매칭을 위해 단일 매칭일 때 토큰의 형태소 품사가 동사일 경우는 제외하였고 2개 이상 매치일 때 사전에 매칭되는 토큰들의 품사 중에 하나라도 동사가 포함되면 매칭에서 제외하였다.

4. 실험

4.1 실험 셋팅

한국어 개체명 인식 성능 평가를 위한 평가셋으로 ETRI 범용 개체명 인식 데이터로 사용하였다. 총 5000문장 중에서 4250, 250, 500 문장을 각각 학습, 개발, 평가 셋으로 사용하였다. 개체명 인식의 Baseline으로는 CRF를 사용하였고 표 3에서의 보듯이 CRF에서도 사전 활용의 성능 향상을 볼 수 있다. LSTM-CRF에서 사용할 사전 자질의 조합은 CRF에서 테스트하여 가장 높은 성능을 보여준 3~4가지의 방식을 선정하였다.

4.2 실험 결과

표 4에서는 한국어 개체명 인식 결과를 보여주고 있다 사전의 종류와 매칭 방법에 관계없이 사전을 사용하지 않았을 때보다 모두 1% 이상의 성능 향상을 보여주고 있다. 특히 최고 성능을 보이고 있는 위키 사전의 최빈도-

최장 매치의 결과는 89.34%로 3%에 가까운 성능 향상을 보여주고 있다.

표 3 사전 적용 CRF 개체명 인식 결과 (F1)

방법	위키		통합	
	Dev	Test	Dev	Test
Baseline	85.57%	82.65%	85.57%	82.65%
최빈도	86.07%	84.65%	86.33%	85.40%
최장	85.78%	85.06%	86.33%	85.26%
최빈도-최장	86.07%	85.16%	86.61%	85.53%
최빈도-LL	**86.32%**	84.70%	86.89%	85.96%
최빈도-최장-LL	85.31%	**85.70%**	**87.01%**	**85.99%**

표 4. 한국어 개체명 인식 실험 결과 (F1)

방법	개발셋	평가셋
CRF(baseline)	85.57%	82.65%
LSTM-CRF(문자기반)+nnlm	88.60%	86.53%
LSTM-CRF(문자기반)+nnlm (위키사전 최빈도-최장)	89.89%	**89.34%**
LSTM-CRF(문자기반)+nnlm (위키사전 최빈도-L2L3LL)	88.96%	87.51%
LSTM-CRF(문자기반)+nnlm (위키사전 최빈도-최장-L2L3LL)	**90.22%**	88.72%
LSTM-CRF(문자기반)+nnlm (통합사전 최빈도-최장)	89.37%	88.56%
LSTM-CRF(문자기반)+nnlm (통합사전 최빈도-L2L3LL)	88.31%	87.57%
LSTM-CRF(문자기반)+nnlm (통합사전 최빈도-최장-L2L3LL)	89.72%	88.80%

5. 결론

본 논문에서는 개체명 인식에서 최고의 성능을 보여주고 있는 문자 기반 LSTM-CRF에 사전 매칭의 정교화와 조합을 통하여 기존의 성능을 크게 개선시켰다. 본 논문에서 사용한 매칭방법과 조합을 영문 개체명 인식에도 적용할 예정이다.

참고문헌

[1] Z. Huang, W. Xu, K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging,” arXiv:1508.01991, 2015

[2] J. P.C. Chiu, E. Nichols, “Named Entity Recognition with Bidirectional LSTM-CNNs,” arXiv:1511.08308, 2015

[3] X. Ma, E. Hovy, “End-to-end Sequence Labeling via Bidirectional LSTM-CNNs-CRF,” arXiv:1603.01354, 2016

[4] 나승훈, 민진우, “문자 기반 LSTM CRF를 이용한 개체명 인식,” KCC 2016

[5] 이창기. “Long Short-Term Memory 기반의 Recurrent NeuralNetwork 를 이용한 개체명 인식,” KCC 2015

[6] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer. “Neural Architectures for Named Entity Recognition,” NAACL-HLT 2016

[7] C. D. Santos and B. Zadrozny. “Learning Character-level Representations for Part-of-Speech Tagging,” ICML 2014

[8] Alexandre Passos, VineetKumar, Andrew McCallum, “Lexicon Infused Phrase Embeddings for Named Entity Resolution,” 2014

[9] 이창기, 김준석, 김정회, 김현기, 딥러닝을 이용한 개체명 인식, KCC 2014