제한된 언어 자원 환경에서의 다국어 개체명 인식

천민아[o†], 김창현[‡], 박호민[†], 노경목[†], 김재훈[†]
한국해양대학교[†], 한국전자통신연구원[‡]
minah0218@kmou.ac.kr[†], chkim@etri.re.kr[‡], homin2006@daum.net[†],
kmq7542@gmail.com[†], jhoon@kmou.ac.kr[†]

Multilingual Named Entity Recognition with Limited Language Resources

Min-Ah Cheon[†], Chang-Hyun Kim[‡], Ho-min Park[†], Kyung-Mok Noh[†], Jae-Hoon Kim[†]
Korea Maritime and Ocean University[†], Electronics and Telecommunications Research Institute[‡]

요 약

심층학습 모델 중 LSTM-CRF는 개체명 인식, 품사 태깅과 같은 sequence labeling에서 우수한 성능을 보이고 있다. 한국어 개체명 인식에 대해서도 LSTM-CRF 모델을 기본 골격으로 단어, 형태소, 자모음, 품사, 기구축 사전 정보 등 다양한 정보와 외부 자원을 활용하여 성능을 높이는 연구가 진행되고 있다. 그러나 이런 방법은 언어 자원과 성능이 좋은 자연어 처리 모듈(형태소 세그먼트, 품사 태거 등)이 없으면 사용할 수 없다. 본 논문에서는 LSTM-CRF와 최소한의 언어 자원을 사용하여 다국어에 대한 개체명 인식에 대한 성능을 평가한다. LSTM-CRF의 입력은 문자 기반의 n-gram 표상으로, 성능 평가에는 unigram 표상과 bigram 표상을 사용했다. 한국어, 일본어, 중국어에 대해 개체명 인식 성능 평가를 한 결과 한국어의 경우 bigram을 사용했을 때 78.54%의 성능을, 일본어와 중국어는 unigram을 사용했을 때 각 63.2%, 26.65%의 성능을 보였다.

주제어: 개체명 인식, 다국어, limited language resources, sequence to sequence labeling

1. 서론

심층학습(deep learning)은 입력 데이터들에 대해 높은 수준의 추상화된 정보를 추출할 수 있다. 이로 인해 자연어처리, 영상처리 등과 같은 분야에서는 최적의 자질 조합을 찾기 위해 많은 시간과 노력이 필요했던 기계학습 알고리즘 대신 심층학습을 이용한 연구가 활발히 진행되고 있다[1]. 심층학습 모델 중 순차 데이터(sequential data)를 모델링 하는 방법인 LSTM과 출력 데이터(output data) 간의 전이 확률을 추가시킨 LSTM-CRF 방식이 개체명 인식 및 품사 태깅 문제에서 높은 성능을 보이고 있다[1-3].

개체명(named entity)은 문서에서 나타나는 고유한 의미를 가지는 명사이다. 개체명은 크게 인명(Person), 지명(Location), 기관명(Organization)으로 나눌 수 있다. 개체명 인식(named entity recognition)은 문서에서 개체명을 추출하고, 추출된 개체명의 종류를 결정하는 작업이다[4]. 한국어 개체명 인식의 성능을 향상을 위해 LSTM-CRF 모델과 사전 등의 외부 자원을 이용한 연구가 진행되고 있다[5-8]. 그러나 공개된 자원이 풍부한 영어에 비해 한국어와 일본어와 같은 언어들은 자연어처리에서 사용할 수 있는 자원과 공개된 자연어 처리 모듈이 한정적이다. 본 논문은 LSTM-CRF와 최소한의 언어 자원만을 사용하여 다국어에 적용 가능한 개체명 인식 방법을 찾는 데 초점을 맞춘다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 다국어 개체명 인식에 관한 LSTM-CRF 모델을 소개한다. 4장에서는 한국어, 일본어, 중국어에 대해 3장의 모델을 적용한 결과를 분석하고, 마지막으로 5장에서 결론 및 향후 연구에 관해 기술한다.

2. 관련 연구

개체명 인식은 문서에서 인명, 지명, 기관명 등의 고유한 의미를 나타내는 단위인 개체명을 추출하고, 추출된 개체명의 종류를 결정하는 작업이다[4]. 개체명 인식의 성능 향상은 정보검색 분야에서 활발하게 연구되고 있는 질의응답이나 기계번역 시스템의 성능 개선을 위해 필수적이다[4].

개체명 인식이 어려운 이유의 핵심 키워드는 미등록어(unknown word)와 중의성(ambiguity)이다. 언어의 특성상 시간이 흐름에 따라 새로운 단어가 계속 생겨나고, 해당 단어들을 모두 개체명 사전에 등록할 수 없기 때문에 사전으로 개체명을 처리하는 데는 한계가 있다. 또 문맥에 따라 같은 단어가 개체명이 될 수도, 되지 않을 수도 있으며 시간의 흐름에 따라 과거 개체명이 아니었던 단어가 개체명이 되거나 그 반대의 경우도 있다. 과거에는 규칙 기반이나 통계 기반의 기계학습을 이용하여 개체명 인식을 처리했으나[9-10], 최근 심층학습 모델 중 하나인 LSTM-CRF 모델이 좋은 성능을 보이고 있다[1-3,5-8]. [5-8]은 LSTM-CRF를 한국어 개체명 인식에 적용한 논문이다. [5-6]은 입력된 문자열의 각 문자를 양방향 LSTM을 적용하여 문자 임베딩을 얻은 후, CNN으로 합성하여 단어의 임베딩 벡터를 얻는다. 단어 임베딩 벡터에 품사 정보, 띄어쓰기 정보, 사전 자질 등의 외부 자원을 추가하여 양방향 LSTM-CRF의 입력이 되는 확장단어 임베딩을 구성하여 86.53%의 성능을 얻었다. [7]은

단어/품사 임베딩을 기본으로 음절 정보, 각 음절의 개체명 품사 분포 정보, 사전 자질 벡터를 결합하여 80.68%의 성능을 보였다. [8]의 연구 역시 형태소/품사를 입력으로 하여 형태소 임베딩, 자음/모음 자질, 품사, 기구축 사전 정보를 결합하여 85.71%의 성능을 보였다. 이처럼 기존의 개체명 인식 연구에서는 다양한 자질의 조합과 외부 자원을 통한 성능 향상에 초점을 맞추고 있다. 이러한 방법은 학습에 사용할 수 있는 자원이 상대적으로 부족한 언어에 대해서는 성능 향상이 쉽지 않다는 문제점이 존재한다.

3. 다국어 개체명 인식을 위한 LSTM-CRF

공개된 자원이 풍부한 영어에 비해 한국어와 일본어와 같은 언어들은 자연어처리에서 사용할 수 있는 자원이 한정적이다. 특히 모국어가 아닌 외국어에 대한 자연어처리에서는 해당 언어에 대한 깊은 이해가 필요하므로 필요한 데이터를 수집 및 가공하는 일에 어려움이 따른다. 본 장에서는 최소한의 학습 자원을 사용한 다국어 개체명 인식에 관해 설명한다.

3.1 입력 형식 : 문자 단위의 n-gram 기반의 임베딩

기존 연구는 단어, 형태소, 품사 임베딩 벡터를 입력 노드의 기본 입력 단위로 삼는다. 해당 임베딩 벡터에 문자(음절)에 대한 정보를 결합하는 형태로 확장하는데, 본 논문에서는 문자의 n-gram을 입력 단위로 사용하여 gensim[11]의 word2vec 기능을 이용하여 임베딩 시킨다. n-gram을 임베딩 시키기 전, 아래의 규칙을 적용한다.

1. **공백 문자**는 **@sp@**로 치환
2. 학습할 말뭉치에서 나타난 **n-gram의 빈도수가 5 이하**인 경우, 해당 n-gram은 **@unk@**로 치환
3. n-gram을 위한 **패딩 문자**는 **@pad@**로 치환
4. 이 외의 n-gram은 그대로 사용

n-gram 임베딩 학습 parameter에 관한 정보는 다음과 같다. 학습 단위는 unigram, bigram이며, 임베딩 차원의 크기는 32, 64, 128, 256이다. window size는 10으로 정했다. window size가 10인 이유는 단어 임베딩의 경우 window size가 기본 4로 고정되어 있기 때문이다. 한국어, 중국어, 일본어 단어의 평균 문자수가 2~3글자이므로 2.5글자×4(단어의 window)=10이라는 숫자를 도출했다. 그 외의 parameter는 gensim의 기본값으로 설정했다.

3.2 출력 형식 : 개체명 태그

기존의 한국어 개체명 인식기에서는 위치 표시 접두어 방법인 BIO(Begin, Inside, Outside)나 BIEOS(Begin, Inside, End, Outside, Single) 태그와 개체명 태그를 결합한 태그를 사용했다. 본 논문에서는 [3]과 같이 각 개체명 태그가 최장 일치법일 경우를 가정하여 표시 접두어를 제외한 개체명 태그 그 자체를 사용한다. 각 위치의 출력 개체명 태그는 그림 1과 같이 각 n-gram의 가장 처음에 오는 문자에 대응되는 개체명 태그를 기준으로 한다. 진한 표시의 음절이 개체명 태그와 대응된다.

<**서호프**:PES>와 <**파사노**:PES>에게 연속 안타를									
서호	**호**프	**프**와	**와**_	_파	**파**사	**사**노	**노**에	…	**를**#
PES	**PES**	**PES**	O	O	**PES**	**PES**	**PES**	…	O

그림 1. n-gram 기반 개체명 인식에서 출력 개체명 태그 형식 예시

3.3 양방향 LSTM-CRF

그림 2는 양방향 LSTM-CRF의 구조이다. 양방향 LSTM은 각 LSTM cell에 입력 문자열을 양방향으로 받아서 각 입력 단위에 대해 은닉 벡터를 얻는다. 이 결과에 전이 확률(의존성)을 추가한 것이 LSTM-CRF이다. 입력은 3.1절에서 설명한 n-gram 기반의 임베딩 벡터이고, 최종 출력은 3.2에서 설명한 개체명 태그이다. 설명한 데이터 외의 형태소, 품사 정보나 기구축 사전 등의 외부 자원은 사용하지 않는다.

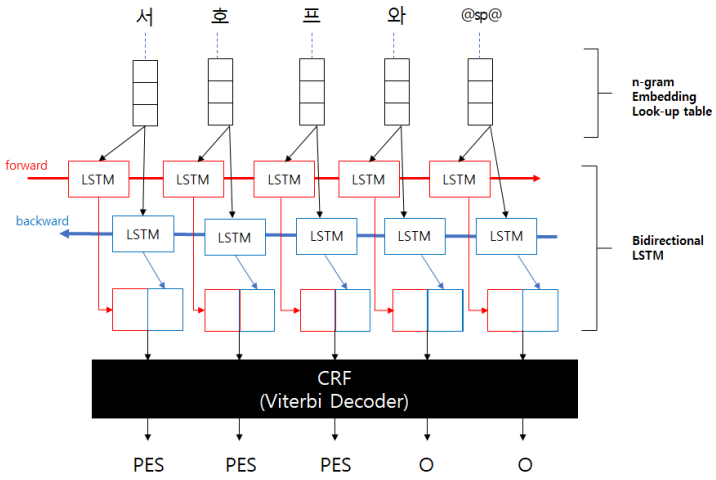

그림 2. 양방향 LSTM-CRF 모델 (unigram을 입력으로 했을 때)

4. 실험

제안한 개체명 인식에 대한 성능 평가를 위한 3.3의 전체 모델 구현은 gensim[11]과 Tensorflow[12]를 사용했다.

4.1 실험 환경

다국어 개체명 인식에 대한 대상 언어는 한국어, 일본어, 중국어이다. 개체명 인식의 범위는 인명(PES), 지명(LOC), 기관명(ORG), 날짜(DAT), 시간(TIM)의 다섯 가지다. 각 언어에 대한 개발, 학습, 평가 말뭉치에 대한 정보는 표 1과 같다. 한국어는 ETRI에서 배포한 개체명 말뭉치[13] 5,000문장을 사용했다. 일본어는 개인이 구축하여 공개한[14] 개체명 말뭉치 500문장을 표1의 일본어前의 항목처럼 나눈 후, 각 항목의 문장을 5배씩 증가시킨 상황에서 실험했다. 중국어는 CONLL2007형식으로 제

작되어 배포된 웨이보 개체명 말뭉치[15] 1,887문장을 사용했다.

표 1. 성능 평가에 사용한 각 언어 개체명 말뭉치 정보

언어	개발(dev)	학습(train)	평가(test)
한국어	**500**	**3,500**	**1,000**
일본어前	50	400	100
일본어	**250**	**2,000**	**500**
중국어	**269**	**1,349**	**269**

n-gram 임베딩 실험에서 본 논문에서 성능 평가를 할 때 사용한 단위는 unigram과 bigram이다. 한국어 n-gram 임베딩은 뉴스 말뭉치 약 2GB를 사전학습(pre-train)한 결과를 사용했다. 일본어와 중국어의 경우 대량의 뉴스 말뭉치를 모으는 도중이라 개발, 학습, 평가 말뭉치로부터 n-gram 임베딩을 사전 학습했다.

성능의 평가 단위는 각 개체명의 청크(chunk) 단위이다. 전체적인 실험 성능은 개발 데이터에서 가장 좋은 성능을 보인 15 epoch에서 평가했다. 실험에 사용한 평가방법은 precision, recall, F_1-measure이다.

$$precision = \frac{|\text{시스템(개체명)} \bigcap \text{정답(개체명)}|}{|\text{시스템(개체명)}|}$$

$$recall = \frac{|\text{시스템(개체명)} \bigcap \text{정답(개체명)}|}{|\text{정답(개체명)}|}$$

$$F_1 - measure = \frac{2 \times precision \times recall}{precision + recall}$$

양방향 LSTM-CRF에서 입력의 크기는 n-gram 임베딩 벡터의 크기와 같다. mini-batch는 50, 100, 200일 때 hidden node의 수는 128, 256개 일 경우에 대해 각각 실험했다. 최적화에는 Adam 함수를 사용했다.

4.3 실험 결과

표 2는 각 언어에 대한 개체명 인식 성능 평가 결과를 unigram, bigram에 대해 각각 정리한 것이다. 각 입력 단위에서 가장 좋은 성능을 냈을 때의 상황을 보여준다. 입력은 임베딩의 크기를 나타내고, n-gram은 입력에 사용한 n-gram의 종류를 나타낸다. mini-batch는 각 batch에서 학습한 문장의 수를 나타내고, 은닉 노드는 양방향 LSTM에서 forward와 backward를 구성하는 LSTM cell의 개수이다. 은닉 노드가 128이라면 forward와 backward에 각 128개의 노드가 부여되어 최종적으로는 256개의 은닉 노드를 사용한 것이 된다. 한국어의 경우 bigram을 사용했을 때가 unigram을 사용했을 때보다 2.5%p 정도 향상된 성능을 보였다. 일본어와 중국어의 경우에는 unigram을 사용했을 때가 bigram을 사용했을 때보다 더 높은 성능을 나타냈다.

표 2. 양방향 LSTM-CRF를 사용한 다국어 개체명 인식 성능 평가 결과

언어	입력	n-gram	mini-batch	은닉 노드	precision	recall	F_1
한국어	64	unigram	200	256	84.30	69.26	76.04
	256	bigram	200	256	**84.46**	**73.40**	**78.54**
일본어	64	unigram	100	128	**65.86**	**60.75**	**63.20**
	128	bigram	50	128	24.62	27.30	25.89
중국어	128	unibram	100	128	**31.48**	**23.10**	**26.65**
	256	bigram	100	256	7.79	16.76	10.64

(※ 모든 성능 평가를 위한 실험에서 사용한 epoch 수는 15로 고정했다.)

중국어의 경우 unigram과 bigram 모두 다른 언어에 비해 현격히 낮은 성능을 보이는 것을 볼 수 있다. 실제 같은 말뭉치를 사용하여 중국에서 진행된 중국어 개체명 인식 결과[16]는 F_1-measure가 44.09%이다. 한국어의 경우에는 충분한 양의 뉴스 말뭉치를 사용하여 유효한 n-gram 임베딩을 구축할 수 있었으나, 일본어와 중국어의 경우에는 개체명 말뭉치로부터 n-gram 임베딩을 구축했기 때문에 n-gram의 임베딩이 충분하게 학습되었다고 보기 어렵다. 일본어와 중국어 모두 한자를 포함하는 언어이다. 유니코드에 등록된 한중일 통합한자의 경우 그 수가 8만여 개에 이른다. 특히 중국어의 경우에 같은 문자를 쓰는 방법이 번자체와 간자체의 두 종류가 있다. 예를 들어 바퀴가 달린 수레를 의미하는 차(車)를 번자체로 쓸 경우 車가 되지만 간자체로 쓸 경우 车로 표기된다. 즉, n-gram을 충분히 반영하기 위해서는 더 많은 말뭉치가 필요하다는 의미이다. 일본어와 중국어에 대해 많은 양의 뉴스 말뭉치를 수집하여 n-gram 임베딩을 추가 학습한 뒤, 다시 성능평가를 해 볼 필요가 있다.

한국어의 경우에는 unigram과 bigram에 대해 모두 76% 이상의 성능을 보였다. 이는 사전학습된 단어/품사 임베딩이나 형태소/품사 임베딩만을 사용했을 때와 비슷한 수준의 결과이다.

5. 결론 및 향후 연구

본 논문에서는 양방향 LSTM-CRF를 사용하여 한국어, 일본어, 중국어의 개체명 인식을 실험하고 그 결과를 살펴봤다. 형태소, 품사나 사전 정보 등의 자원이 충분하지 않을 경우를 고려하여 문자를 n-gram 단위로 입력했을 때의 결과를 살펴봤다. 한국어의 경우 unigram을 사용했을 때 76.04%, bigram을 사용했을 때는 2.5%p 증가한 결과인 78.54%의 성능을 보였다. 이는 문자를 n-gram 단위로 사전 임베딩 하는 것으로 단어/품사 임베딩이나 형태소/품사 임베딩만을 사용했을 때와 비슷한 수준의 결과에 도달할 수 있다는 의미이다. 일본어와 중국어의 경우에는 문자 n-gram을 사전학습하기 위한 말뭉치가 충분하지 않아 bigram을 사용했을 때가 unigram을 사용했을 때보다 현격히 낮은 성능을 보였다.

향후에는 일본어와 중국어의 뉴스 말뭉치를 한국어와 비슷한 수준까지 수집하여 문자 단위의 n-gram 임베딩을 다시 학습한 후, 성능 평가를 진행할 예정이다. 또한,

학습 말뭉치에서 얻을 수 있는 정보인 문자 단위 n-gram의 개체명 품사 분포 정보와 n-gram 임베딩을 결합한 결과를 입력으로 사용하여 다국어 개체명 인식에 대한 성능 평가할 계획이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

참고문헌

[1] Ronan Collobert, et al., Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12, 2011.
[2] Xuezhe Ma, Eduard Hovy, “End-to-end Sequence Labeling via Bidirectional LSTM-CNNs-CRF”, arXiv:1603.01354, 2016.
[3] Onur Kuru, Ozan Arkan Can, and Deniz Yuret, “CharNER: Character-Level Named Entity Recognition”, arXiv:1603.01354, 2016.
[4] David nadeau and Stasoshi Sekine, “A survey of named entity recognition and classification”, Journal of Linguisticae Investingations, 30(1): 3-26, 2007.
[5] 민진우, 오효정, 나승훈, “식품 도메인 개체명 인식을 위한 문자 기반 LSTM CRF”, 한국정보과학회 2016년 동계학술대회 논문집, pp.500-502, 2016.
[6] 나승훈, 민진우, “문자 기반 LSTM CRF를 이용한 개체명 인식”, 한국정보과학회 2016년 한국컴퓨터종합학술대회 논문집, pp.729-731, 2016.
[7] 유홍연, 고영중, “Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 임베딩의 확장”, 정보과학회논문지 44(3):306-313, 2017.
[8] 신유현, 이상구, “양방향 LSTM-RNNs-CRF를 이용한 한국어 개체명 인식기”, 제28회 한글 및 한국어 정보처리 학술발표 논문집, pp.340-341, 2016.
[9] Sergey Brin, “Extracting Patterns and Relations from the World Wide Web”, WebDB '98 Selected papers from the International Workshop on The World Wide Web and Databases, pp.172-183, 1998.
[10] Daniel M. Bikel et al., “Numble: a High-Performance Learning Named-finder”, In: Proc, The Fifth Conference on Applied Natural language Processing, 1997.
[11] gensim [Online] https://radimrehurek.com/gensim/index.html
[12] Onur Kuru, Ozan Arkan Can, and Deniz Yuret, “CharNER: Character-Level Named Entity Recogition”, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp.911-921, 2016.
[13] Tensorflow [Online] https://www.tensorflow.org/
[14] ETRI 지식마이닝 연구실, 엑소브레인 언어분석 말뭉치(ver. 1.0), 2015.
[15] 일본어 개체명 말뭉치 [Online] https://github.com/Hironsan/IOB2Corpus
[16] 중국어 개체명 말뭉치 [Online] https://github.com/hltcoe/golden-horse/tree/master/data