개체명 비율 사전을 결합한 Bidirectional LSTM-CRF 기반 개체명 인식

박동주○, 안창욱*
광주과학기술원 전기전자컴퓨터공학부
{toriving, cwan}@gist.ac.kr

Named Entity Recognition using Bidirectional LSTM-CRF Combining Named Entity Ratio Dictionary

Dongju Park○, Chang Wook Ahn*
School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology

요 약

머신러닝 및 딥러닝의 급속한 발전은 다양한 자연어 처리 분야에 많은 영향을 미치고 있다. 특히, 주어진 문장 내에서 미등록 단어로 나타나는 인명, 장소 그리고 기관명과 같은 고유한 의미를 갖는 단어들을 분류하는 개체명 인식에서도 많이 활용되고 있다. 개체명인식의 전통적인 방법인 언어 문법 기반 시스템은 구축하는데 많은 자원과 시간, 그리고 도메인 지식이 필요하지만 머신러닝 및 딥러닝 방법은 보다 적은 시간으로 좋은 성능을 얻을 수 있다. 머신러닝 방법 중 하나인 Conditional Random Field (CRF)는 개체명 인식과 같이 연속적인 레이블링에 효과적이다. 하지만, CRF를 학습하기 위한 자질 선택에는 많은 시간과 자원이 소모된다. 이러한 단점을 보완하기 위해 Long Short-Term Memory를 활용하여 자질을 선별해 줄 수 있다. 본 연구에서는 개체명 인식을 위한 개체명 비율 사전을 결합한 Bidirectional LSTM-CRF 모델을 제안한다. 제안된 모델에서는 Bidirectional LSTM이 적합한 자질 선택을 돕기 위해 Bidirectional LSTM의 입력으로 개체명 비율을 함께 사용한다. 제안한 모델을 학습 및 검증을 위해 NAVER NLP Challenge의 개체명 인식 분야의 데이터를 사용하였으며, 검증 결과 다른 모델들에 비해 뛰어난 성능을 얻었다. 제안 모델에 포함된 개체명 비율을 자질로 사용하는 방법은 모델의 복잡도가 적게 증가하면서도 좋은 성능을 보여주었다.

1. 서 론

효과적이고 효율적인 머신러닝 및 딥러닝의 빠른 발전은 자연어 처리 분야의 패러다임을 바꾸었다. 특히, 주어진 문장 내에서 미등록어로 자주 나타나는 인명이나 장소, 기관명과 같은 고유한 의미를 갖는 단어들을 미리 정의된 개체 범주로 분류하는 작업인 개체명 인식에서도 좋은 성능을 보이고 있다. 이러한 개체명 인식은 기계 번역과 정보 검색, 그리고 대화 시스템과 같은 다양한 자연어 처리 분야에서 전처리의 한 부분으로 사용되어 성능을 향상 시킬 수 있어서 중요한 분야이다.

전통적인 개체명 인식 방법으로는 통계 기반 방법과 수작업으로 만들어진 언어 문법 기반 시스템 존재하며, 그 중 언어 문법 기반 시스템이 더 나은 성능을 보였다. 하지만, 언어 문법 기반 시스템은 구축하는데 시간이 많이 걸릴 뿐 더러 많은 도메인 지식을 요구한다. 반면, 통계 기반 방법은 문법 기반 방법에 비해 시간이 적게 들었지만 성능이 좋지 않았다. 최근에는 Long Short-Term Memory (LSTM) [1]와 Conditional Random Field (CRF) [2]를 활용한 딥러닝 및 머신러닝 기반 방법들이 연구되면서 시간 단축과 성능 향상을 동시에 이루고 있다.

본 논문에서는 NAVER NLP Challenge[1]의 개체명 인식 분야에서 좋은 성능을 보인 개체명 사전 자질 결합 양방향 LSTM-CRF 모델을 제안한다. 제안된 모델은 NAVER NLP Challenge에서 제공한 개체명 인식 데이터 셋을 이용하여 NLP Challenge에 제출된 모델들과 성능을 비교 및 검증한다.

2. 관련 연구

기존의 개체명 인식은 문장 내에서 연속적인 레이블링을 해야하기 때문에 CRF와 같은 머신러닝 기법을 주로 사용했다[3]. 하지만, CRF에서는 학습을 위해 다양한 자질을 선별하는 과정에서 비용이 많이 발생하게 된다. 이러한 문제점은 Recurrent Neural Network(RNN) 계열 딥러닝 네트워크를 사용함으로써 극복할 수 있게 되었고, 그에 따라 RNN 계열 네트워크와 CRF를 결합하는 방법들이 많이 연구되고 있다.

RNN 계열 네트워크와 CRF를 결합한 방법 중 하나인 Bidirectional LSTM-CRF[4]는 주어진 문장을 양방향으로 처리하여 자질을

[1] https://github.com/naver/nlp-challenge

만들어 낸 후 CRF를 적용시켰다. 이러한 방법은 기존에 사용하던 단방향 방식의 단점을 보완함으로써 좋은 성능을 달성하였다.

Bidirectional LSTM-CNN[5]에서는 단어 단위에 LSTM을 사용하고, 문자 단위에서는 Convolutional Neural Network (CNN)를 사용함으로써 자질 추출에서 단어와 문자 모두 중요함을 보여주었다.

최근에는 다양한 문서에 대해 언어 모델을 사전 학습 시킨 후 개체명 인식 작업에 대해 전이 학습을 통해 좋은 성능을 보여주고 있다[6, 7].

3. 제안 모델

본 연구에서는 NAVER NLP Challenge의 개체명 인식 분야에 사용된 모델을 제안한다. 제안된 모델의 구조는 크게 Embedding block, Bidirectional LSTM block, CRF block으로 나누어 진다. 전체적인 모델은 그림 1과 같다.

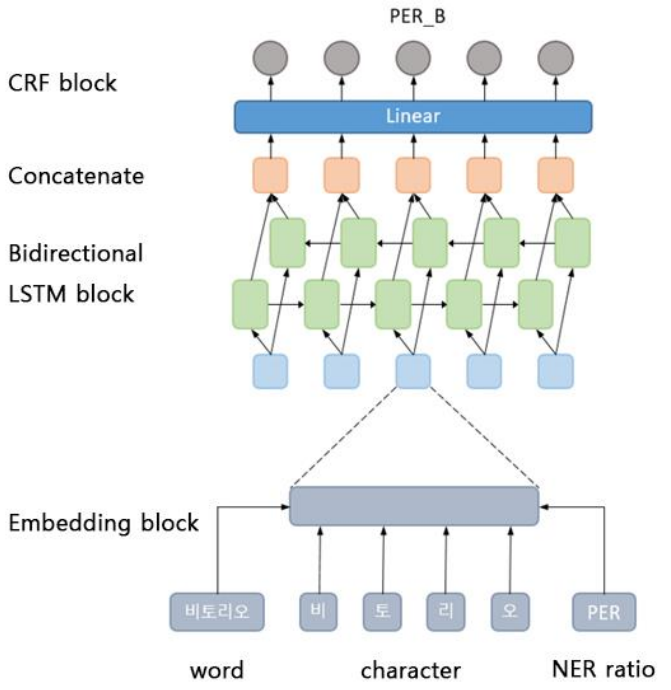

그림 1. 제안 모델의 전체적인 구조

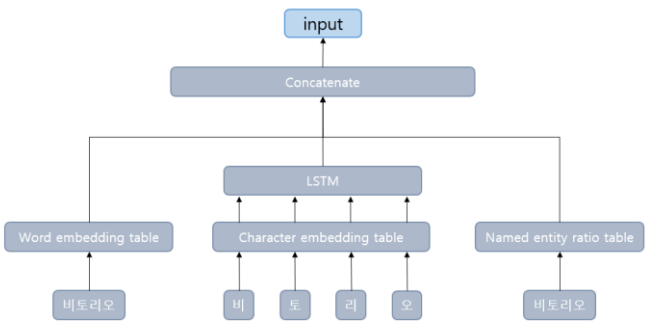

그림 2. Embedding block의 구조

Embedding block은 입력으로 들어온 데이터를 단어 단위 임베딩, 문자 단위 임베딩, 그리고 단어가 트레이닝 데이터 내에서 차지하고 있는 개체명 비율을 합쳐서 최종 입력으로 변환해 준다. 이때, 문자 단위는 각 문자를 임베딩 한 후 LSTM을 통해 최종적으로 나오는 상태를 사용한다. Embedding block의 구조는 그림 2와 같다.

Embedding block에서 최종 입력으로 들어온 값은 정방향 LSTM과 역방향 LSTM의 마지막 상태 값을 합친 Bidirectional LSTM을 거치게 된다. 이때 Bidirectional LSTM에서 사용되는 LSTM은 모두 과적합 방지를 위해 dropout과 layer normalization을 사용한다. Bidirectional LSTM의 출력은 CRF block의 입력으로 사용된다.

CRF block에서는 Bidirectional LSTM으로 만들어진 상태 값을 선형 함수를 통해 한번 변형 시킨다. 그 후 최종적으로 CRF를 사용하여 각 단어들의 개체명을 예측한다.

4. 실험 및 결과

4.1 실험 설정

제안된 모델을 학습 및 검증하기 위한 데이터로 NAVER NLP Challenge에서 사용된 개체명 인식[2] 데이터를 사용하였다. 데이터 셋의 개체명 범주는 총 14가지로 구성되어 있으며, 여러 어절에 걸쳐 표현된 개체명을 처리하기 위해 범주 뒤 'B'/'I' 의 추가 정보를 사용하거나 개체명이 부여되지 않았을 경우 '-'로 표시한다. B는 개체명의 시작 어절을 표시하며, I는 앞의 어절과 연속된 같은 개체명을 의미한다.

데이터 셋은 총 100,000문장으로 이루어져있으며, 그중 90,000문장은 학습 데이터이고 나머지 10,000문장은 테스트 데이터이다. 문장들은 어절 단위로 구분이 되어있어, 모델의 입력은 단어 단위 대신 어절 단위로 입력 된다.

모델 학습을 하기 전 훈련 데이터를 이용하여 훈련 데이터 내에서 각 어절들이 차지하는 개체명 비율을 하나의 사전으로 저장한다. 또한, 어절과 어절에 포함되어 있는 각 문자들에 대해도 임베딩 사전을 만든다. 이렇게 저장된 개체명 비율 사전과 어절 사전 및 문자 사전은 학습할 때와 테스트할 때 Embedding block에서 사용된다.

최종적인 모델은 3개의 모델을 앙상블 하여 F1-score를 계산하였다. 앙상블 방법으로는 각 모델에서 추론한 개체명을 Hard voting 하는 방식으로 사용하여 가장 많이 나오는 개체명을 최종 출력으로 사용한다.

학습 알고리즘 선택에서는, 일반적으로 사용되는 AdamOptimizer[8]는 학습데이터의 양이 많아 1~2 epoch 사이에서도 빠르게 수렴하여 과적합 되는 경향이 있어 RMSPropOptimizer[9]를 사용하였다. 본 연구에서 RMSPropOptimizer는 느리게 수렴하였지만 다른 학습 알고리즘에 비해 안정적이고 좋은 성능을 보여주었다.

모델의 하이퍼파라미터는 서로 다른 값을 여러 번 시도하여 적합한 조합을 찾았으며, 하이퍼파라미터 중 어절의 임베딩 크기나 문자의 임베딩 크기, 그리고 개체명의 임베딩 크기는 모두 동일하게 설정하

[2] https://github.com/naver/nlp-challenge/tree/master/missions/ner/data/train

였으며, 모델 내에 사용되는 LSTM 상태의 크기도 동일하게 설정하였다. 자세한 하이퍼파라미터는 표 1과 같다.

표 1. 모델의 하이퍼파라미터에 따른 값

하이퍼파라미터	값
배치 사이즈	128
학습률	0.001
드롭아웃 확률	0.35
임베딩 크기	128
LSTM 상태 크기	128
최대 문장 길이	180
최대 어절 길이	8

4.2 실험 결과

실험 결과 본 논문에서 제안한 모델은 90.4219라는 높은 F1-score를 얻었다. 이 결과는 3개의 모델을 앙상블 한 결과이며, 앙상블을 하지 않은 단일 모델의 경우는 90.2499를 얻었다. 이 수치는 NAVER NLP Challenge의 개체명 인식 분야에서 가장 높은 수치이다. 다른 참가자 및 모델의 F1-score는 표 2 및 NAVER NLP Challenge 리더보드[3]에서 확인 할 수 있다.

표 2의 nlp-pln 팀의 모델은 BiLSTM+CRF+multi-head attention and separable convolution 이며, Sogang_Alzzam 팀의 모델은 BiLSTM-CRF-ELMo, bible 팀의 모델은 CNN 기반 모델이다. 그 외 나머지 팀은 공개하지 않았다.

표 2. NAVER NLP Challenge 개체명 인식 분야 F1-score

팀명	F1-score
제안 모델 (앙상블)	**90.4219**
제안 모델 (단일)	**90.2499**
cheap_leaning	90.2417
nlp_pln	89.7830
Sogang_Alzzam	88.8506
ner_master	88.5818
bible	88.3348

5. 결 론

본 논문에서는 자연어 처리 분야의 하나인 개체명 인식을 위한 학습 데이터 기반 개체명 비율 사전을 결합한 Bidirectional LSTM-CRF 모델을 제안하였다. 제안 모델은 학습 데이터에 존재하는 어절들을 바탕으로 개체명 비율 사전을 미리 만든 후, 학습과 테스트 시에 사용하여 좋은 성능을 보여주었다. 데이터에서 새로운 자질을 얻어 사용하는 이러한 방법은 모델의 구조를 개선하는 방법보다 쉽고 간단하게 성능을 향상 시킬 수 있음을 보였다.

또한, 제안된 방법은 학습 데이터가 많아 질 수록 그에 따른 개체명 비율 사전이 현실 세계에서 사용하는 비율을 잘 반영할 것이고, 비율을 잘 반영 할수록 성능이 더욱 향상 될 수 있을 것이다.

[3] http://air.changwon.ac.kr/?page_id=10

Acknowledgements

이 논문은 2019년도 광주과학기술원의 재원으로 GRI(GIST연구원) 사업의 지원을 받아 수행했으며, NAVER NLP Challenge에 참여해서 개발한 알고리즘 기반 연구임.

참 고 문 헌

[1] Hochreiter, S., & Schmidhuber, J. Long short-term memory, Neural computation, 9(8), 1735-1780. 1997.

[2] Lafferty, J., McCallum, A., & Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning (ICML), 2001.

[3] McCallum, A., & Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Conference of the North American Chapter of the Association for Computational Linguistics & Human Language Technologies (NAACL-HLT), pages 188–191, 2003.

[4] Huang, Z., Xu, W., & Yu, K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

[5] Chiu, J. P., & Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics (ACL), 4, 357-370, 2016.

[6] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. Deep Contextualized Word Representations. In North American Association for Computational Linguistics (NAACL), 2018.

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] Kingma, D. P., & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. (2014).

[9] Geoffrey Hinton, N Srivastava, and Kevin Swersky. Lecture 6a overview of mini–batch gradient descent. Coursera Lecture slides https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2012, [Online].