음절 기반의 CNN를 이용한 개체명 인식

박혜웅, 송영숙
아이리마인즈, 경희대학교
bage79@gmail.com, klanguage1004@gmail.com

Named Entity Recognition using CNN for Korean syllabic character.

Hye-woong Park, Young-Sook Song
Iriminds, KyungHee University

요 약

개체명 인식(Named Entity Recognition, 이하 NER)은 인명(PS), 기관명(OG), 장소(LC), 날짜(DT), 시간(TI) 등에 해당하는 개체명에 일정한 태깅 값을 주어 그 정보를 가시화하는 작업이다. 한국어 개체명 인식은 아직 그 자질이 충분히 밝혀져 있지 않아 자연어 처리 분야의 발전을 더디게 하는 한 요소로 작용하고 있다.

한국어가 음절 기반으로 단어를 형성하고 비교적 어순이 자유롭다는 특성이 있기에, 이런 특징을 잘 포착할 수 있는 "음절 기반의 Convolutional Neural Network(CNN)"의 아키텍쳐를 제안하여 66.80%의 성능을 보였다. 이 방법을 사용하면 형태소 분석등 개체명 이전 단계에서 발생하는 오류에 의해 개체명 인식(NER)의 성능이 떨어지는 문제를 해결할 수 있고, 조사나 어미 등을 제거하기 위한 후처리를 생략할 수 있다.

Convolutional Neural Network, Named Entity Recognition, 음절 기반

1. 서론

본고에서는 다른 전처리를 하지 않은 자연어 코퍼스에서 국어의 음절 임베딩을 통해 인명, 기관명, 장소 날짜, 시간의 다섯 가지 개체명에 속하는 단어를 추출하여 자동 태깅하려고 한다. 성능 평가를 위한 개체명 말뭉치로는 2017 국어 정보 처리 시스템 경진대회에서 배포한 6259개의 문장을 사용하였다. 본고의 구성은 2장에서 개체명 분야에서의 관련 연구를 소개하고 3장에서는 CNN 모델을 바탕으로 음절 단위의 임베딩 벡터를 통해 개체명을 포함하고 있는 단어를 인식할 수 아키텍쳐를 제안하고자 한다. 4장에서는 모델링의 구체적 특징과 테스트 결과를 분석한다.

2. 관련 연구

CoNLL(2003) shared task에서 사람, 위치, 조직, 기타의 개체명에 대한 데이터셋을 구축한 이후 개체명에 대한 연구는 크게 두 가지 방향으로 발전해 가고 있다고 할 수 있다. 먼저 Collobert et al. (2011)에서 지명을 추가하는 등 꾸준히 그 개체명의 범주를 늘리는 작업이 이어졌고 국내에서도 조은경(2014)에서 정보와 요구 기능을 하는 개체명과 그 동의어를 찾아서 개체명의 범주가 확장 될 수 있음을 밝히기도 했다.

또 한편에서는 RNN, CRF 등의 다양한 방법론(4~6)이 시도되고 있다. 국내에서는 아직 CNN을 활용한 개체명 연구가 활발하지 않지만 J. P.C. Chiu(2015)나 Emma Strubell(2017) 등의 연구를 통해 CNN도 개체명 인식에서 충분히 좋은 성능과 속도를 낼 수 있음을 확인할 수 있었다. 본 논문에서는 태깅되지 않은 원시 말뭉치를 직접 입력으로 사용하여, 음절 자체 정보만으로 음절 단위의 개체명 인식을 수행하는 있는 새로운 신경망 아키텍처를 제시한다.

3. 특징

CNN이 이미지 처리분야에서 높은 성능을 낼 수 있는 것은 이미지 전체 영역에 대하여 필터를 이용해 패턴을 스스로 학습하는 능력이 뛰어 나기 때문이다. CNN은 패턴이 이미지 안의 어떤 위치에 있는지 상관 없고, 회전, 대칭, 확대 등 어떠한 변형에도 강건한 학습이 가능하다.

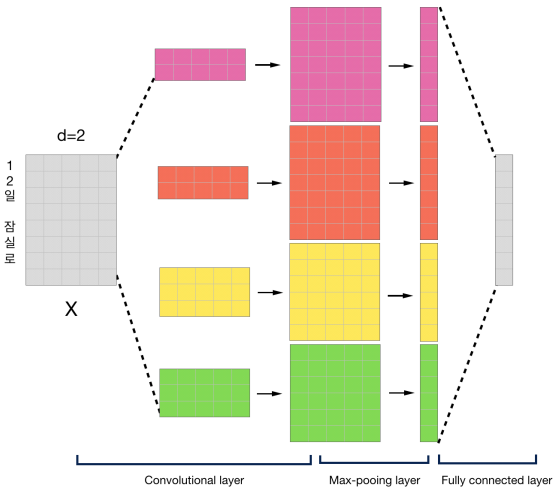
CNN의 이런 특성을 응용하면 문장안의 어떤 위치에 있든지, 다른 위치의 개체명에 영향을 주는 패턴을 인식할 수 있다고 가정하였다. 특히 한국어는 어순이 자유로워서 의존소와 지배소의 위치 관계가 일정치 않다. 또한 어미와 조사에 의한 음절의 변형이 많아 어절 단위의 임베딩을 이용하는 경우 저빈도 어절의 특징을 놓치는 일이 발생할 수 있다.

1음절 단어를 제외하면 음절 정보는 특별한 그 자체만으로 의미를 지닌 것은 아니라고 여겨서 개체명 연구에서는 음절 임베딩을 사용하는 모델이 적다. 본 논문에서 음절 임베딩을 사용한 이유는 복잡한 전처리 또는 후처리 단계를 생략하고 그 단계에서 발생하는 오류를 최소화하기 위함이다. 일반적으로 개체명 인식을 하기 위해서는 형태소 분석 결과가 필요하다. 조사, 어미 등이 제거된 명사 단위에 태깅을 해야 하기 때문이다. 하지만 음절 단위의 임베딩을 이용하면, 이러한 전처리가 생략될 수 있다.

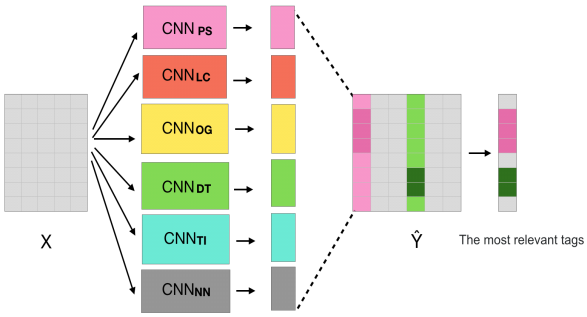
개체명을 인식하기 위해서는 품사 정보, 주위의 단어 또는 음절 정보, 사전 어휘부에 포함되어 있는지 여부 등을 특징으로 사용하기도 한다. 하지만 본 모델은 입력되는 원시 말뭉치의 순수한 음절 정보만을 학습하므로 매우 사용하기 쉬우면서 각 단계에서는 CNN 모델의 특성이 자연어 처리에서 잘 작동할 수 있도록 응용한 모델이다.

4. 모델링

클래스의 분류는 PS, LS, OG, DT, TI 를 포함하여, 개체명이 아닌 명사(NN)를 하나의 분류로 추가하여 총 6개로 정의하였다.



각각의 문자에 대한 분류 문제이기 때문에, 윈도우안의 음절의 위치를 최대한 보존하기 위함이다. Max-pooling된 결과를 모두 붙여 Fully-connected Layer의 입력으로 하였고, 최종적으로 각 윈도우 대하여, 한 개의 클래스에 대한 확률을 결과(Y^ps)로 출력한다.



한 개의 CNN을 사용하여 여러 개의 클래스로 분류하는 방법에 비하여 여러 개의 CNN이 각각의 클래스 분류 문제를 담당하게 되면 몇 가지 장점이 있다. 먼저 입력되는 클래스별로 임베딩 벡터를 각각 학습하여 성능을 향상시킬 수 있다. 마찬가지로 Convolution Layer 필터의 크기나 개수도 클래스별로 최적화 시킬 수 있다. 최종 각각의 CNN의 출력을 종합하여, 각 클래스에 대한 확률값(Y^)을 얻게 된다.

5. 평가방법

입력된 문장을 고정된 크기의 윈도우로 슬라이딩 하면서 여러 개의 입력 임베딩(X)을 얻는다.
예를 들어 윈도우의 크기가 5이고 입력 문장이 "12일 잠실로 이사간다."인 경우, 아래와 같이 첫 번째 윈도우에 대한 입력 임베딩은 "12일 잠"이고 레이블은 [TI, TI, TI, NN, LC] 이다. 마찬가지로 한 문자씩 슬라이딩하면 입력 임베딩은 "2일 잠실", "일 잠실로", " 잠실로 "의 순서로 얻게 된다.
모든 윈도우를 입력하여 모델에서 얻은 결과를 모두 합산하면, 각 문자에 대하여 가장 많이 예측된 클래스로 분류를 할 수 있다. 일부의 결과에서 예측 오류가 발생하더라도 다른 결과들에 의해서 오류를 보정해 주는 이점이 있다.

	정밀도 (Precision)	재현율 (Recall)	F-score
DT	69.43	69.43	69.43
LC	60.75	79.75	68.97
OG	58.39	66.20	62.05
PS	71.86	80.40	75.90
TI	60.00	55.56	57.69
합	64.09	70.29	66.80

6. 결론

본 논문에서는 CNN을 개체명 인식에 적용하였다. 원시 말뭉치의 음절 정보만을 이용하여, 분류외의 분야에 CNN을 활용할 수 있음을 보였다. 앞으로도 다양한 신경망 모델을 조합하여 인위적인 특징 추출 없이 성능을 개선하고자 한다.

참고문헌

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuogluand P. Kuksa. Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research (JMLR),2011.
[2] 조은경, 정보 기술 분야에서 개체명 동의어 연구, 언어과학연구 69, 2014.

[3] 나승훈, 민진우. 문자 기반 LSTM CRF를 이용한 개체명 인식. 한국정보과학회 학술발표논문집, 729-731, 2016
[4]이창기. Long Short-Term Memory 기반의 Recurrent Neural Network를 이용한 개체명 인식, 한국컴퓨터 종합학술대회 논문집, No.6, pp.645-647, 2015.
[5]이창기, 김준석, 김정회, 김현기, 딥러닝을 이용한 개체명 인식, 한국정보과학회 동계학술발표회 논문 집, No.12, pp.423-425, 2014.
[6]유홍연, 고영중, Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장. 정보과학회논문지, 44(3), 306-313, 2017.
[7] J. P. C. Chiu, E. Nichols, Named Entity Recognition with Bidirectional LSTM-CNNs, arXiv: 1511. 08308, 2015
[8]Emma Strubell, Patrick Verga, David Belanger, Andrew McCallum, Fastand Accurate Entity Recognition with Iterated Dilated Convolutions, arXiv: 1702. 02098, 2017