품사 부착 실험을 통한 Bags-of-Features 방법의 정량적 평가

이찬희○, 이설화, 임희석

고려대학교 정보대학 컴퓨터학과
chanhee0222@korea.ac.kr, whiteldark@korea.ac.kr, limhseok@korea.ac.kr

Quantitative Evaluation of Bags-of-Features Method Using Part-of-Speech Tagging

Chanhee Lee○, Seolhwa Lee, Heuiseok Lim
Department of Computer Science and Engineering, College of Informatics, Korea University

요 약

본 논문에서는 단순하지만 효과적인 단어 표현 방법인 Bags of Features에 대한 비교 실험을 수행한다. Bags of Features는 어휘집의 크기에 제한이 없으며, 문자 단위의 정보를 반영하고, 벡터화 과정에서 신경망 구조에 의존하지 않는 단어 표현 방법이다. 영어 품사 부착 실험을 사용하여 실험한 결과, one-hot 인코딩을 사용한 모델과 대비하여 학습 데이터에 존재하지 않는 단어의 경우 49.68%, 전체 부착 정확도는 0.96% 향상이 관찰되었다. 또한, Bags of Features를 사용한 모델은 기존의 영어 품사 부착 분야의 최첨단 모델들 중 학습 데이터 외의 추가적인 데이터를 활용하지 않는 모델들과 비견할 만한 성능을 보였다.

주제어: 자연어처리, 품사 부착

1. 서론

현재 단어를 입력으로 사용하는 많은 자연어처리 시스템(품사 부착, 단어 임베딩, 개체명 인식 등)은 단어를 독립적인 단위로 취급하고 one-hot 인코딩이나 lookup table을 이용하여 단어를 벡터로 변환한다[1,2]. 그러나 이러한 방법은 고정된 크기의 어휘집을 사전에 정의해야 하며, 모델의 파라미터 수가 어휘집의 크기에 따라 선형적으로 증가한다. 어휘집에 포함되지 못한 (Out-Of-Vocabulary, OOV) 단어도 추가적인 문제를 발생시킨다. 회귀 신경망이나 합성곱 신경망을 이용하여 문자 수준에서 단어를 처리함으로써 이러한 문제를 극복하는 방법도 있지만, 이는 추가적인 신경망 구조가 필요하므로 모델의 복잡도를 증가시킨다.

이찬희(2017)[3]의 연구에서는 bag of characters를 응용하여 어휘집에 제한이 없으며 신경망 구조에 의존하지 않고 문자 단위의 정보를 반영하는 단어 표현 방법인 Bags of Features(BOF)를 제안하였다. 본 연구에서는 BOF 방법을 이용하여 영어 품사 부착 실험을 수행한다.

2. 모델

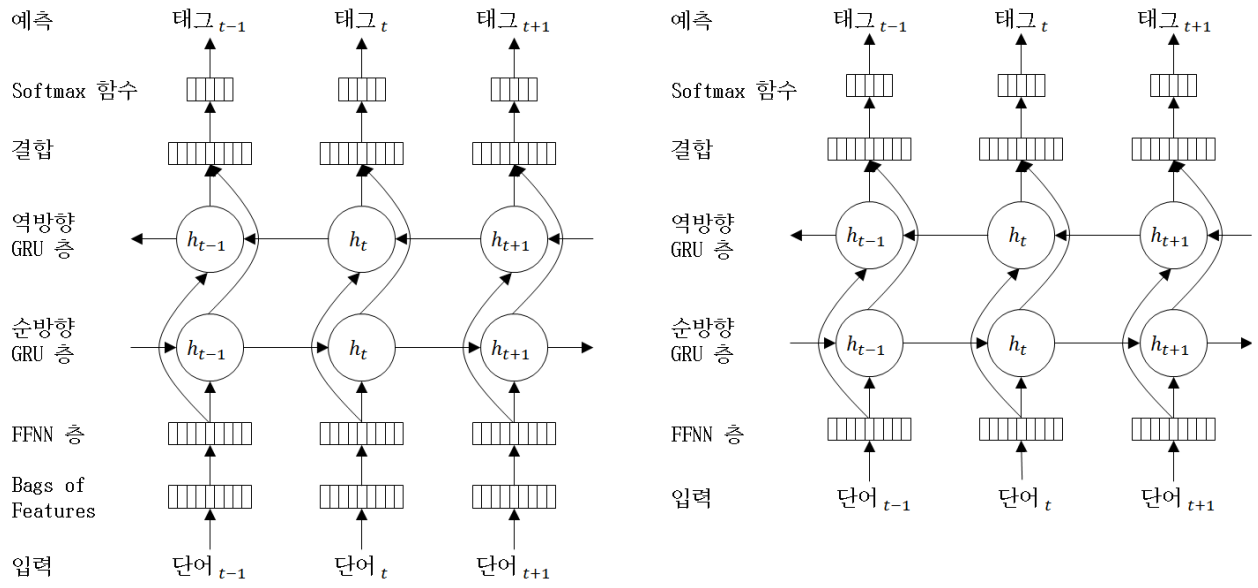
2.1. Bags of Features

이찬희(2017)[3]는 한국어를 기준으로 단어 표현 방법을 제안하였다. 한국어는 교착어이므로 품사 부착 이전에 형태소 분석이 우선적으로 수행되어야 하지만, 영어는 교착어에 속하지 않으므로 이러한 추가적인 과정이 필요하지 않다. 따라서 본 연구에서는 사전 실험으로써 영어를 대상으로 BOF 방법의 성능을 평가한다.

이찬희(2017)[3]의 연구에서 제안된 방법에 추가적으로, 영어의 특성을 반영한 대소문자 정보를 포함시켜 실험을 수행하였다. Collobert(2011)[4]에서 제안된 방법에 따라, 단어를 [*모두 대문자, 모두 소문자, 첫 글자 대문자, 기타*] 중 하나로 분류하여 이를 one-hot 인코딩을 사용하여 벡터로 변환 후 BOF 벡터에 결합하는 방법으로 이를 구현하였다.

2.2. 양방향 회귀 신경망

품사 부착과 같은 순열 분류 작업에서는 중심 단어의 앞 뒤 단어 정보에 접근 가능하다. 양방향 회귀 신경망[5]은 특정 입력에 대하여 이전 및 이후 단어 정보를 효과적으로 활용할 수 있다. 순수한 회귀 신경망은 장거리 의존성 문제에 취약하므로, 본 연구에서는 이 문제를 개선시킨 Gated Recurrent Unit(GRU)[6]을 사용한다. 모델의 뼈대는 GRU를 이용한 다층 회귀 신경망이며, 각 층 사이에는 Dropout[7]을 적용하였다. 양방향 회귀 신경망의 출력에 Feed-Forward Neural Network(FFNN)을 적용한 후, softmax 함수를 이용하여 품사들에 대한 확률 분포를 얻는다. 입력과 회귀 신경망 사이에도 FFNN 층들을 추가하여 더 높은 차원의 자질이 추출될 수 있도록 하였으며, 이 FFNN 층에는 Dropout을 적용하지 않는다. 이러한 모델의 구조는 [그림 1](a)에 나타나 있다.



(a) Bags of Features 사용 모델

(b) One-hot 인코딩 사용 모델

그림 1: 품사 부착 실험에 사용된 모델의 구조도. 점선은 Dropout의 적용을 나타냄. 도식의 단순화를 위하여 GRU 및 FFNN 층들은 다수의 층을 하나로 축약하여 표시함. (a) Bags of Features를 이용한 품사 부착 모델. (b) One-hot 인코딩을 이용한 품사 부착 모델.

2.3. One-hot 인코딩

BOF를 이용한 모델과의 성능 비교를 위하여 단어의 벡터화에 one-hot 인코딩을 사용하는 모델을 제작하였다. 이 모델의 학습 시에는 학습 데이터에 등장한 모든 어휘가 어휘집에 포함되도록 하였으며, 개발 또는 평가 데이터에만 등장하는 단어는 OOV를 나타내는 특수 단어로 치환하였다. 이 모델의 구조는 [그림 1](b)에 나타나 있다.

3. 실험 설계

3.1. Penn TreeBank 말뭉치

본 실험에서는 영어 품사 부착 실험에서 가장 널리 쓰이는 말뭉치인 Penn TreeBank(PTB)의 Wall Street Journal(WSJ)부분을 사용하였으며, 표준에 따라 0-18항은 학습, 19-21항은 개발, 22-24항은 평가 데이터로 활용하였다. 해당 말뭉치는 총 45 종류의 품사 중 하나로 단어들이 구분되어 있다.

3.2. 실험 방법

실험에 사용된 모델의 학습에는 경사 하강법을 바탕으로 한 최적화 알고리즘인 Adam[8]을 사용하였으며, 이를 이용하여 cross-entropy 손실 함수를 최소화 시켰다. 모든 모델의 구현에는 TensorFlow 라이브러리[9]를 사용하였다.

모든 FFNN 층 및 GRU 층에는 512개의 은닉 노드를 사용하였다. BOF를 사용한 모델에는 입력과 회귀 신경망 사이에 3개의 FFNN 층을 추가하였다. One-hot 인코딩을 사용한 모델의 단어 임베딩은 무작위로 초기화된 512차원 벡터가 사용되었다. 모든 모델은 20 에포크 동안 학습되었으며, 배치 크기는 64로 실험하였다.

4. 실험 결과

[표 1]은 BOF를 사용한 모델과 one-hot 인코딩을 사용한 모델의 품사 부착 정확도를 정리한 것이다(PTB WSJ 말뭉치의 평가 데이터 기준). One-hot 인코딩을 이용한 모델은 OOV 단어에 대해 매우 낮은 성능을 보인다. 이는 모델이 OOV 단어 자체에 대한 정보 없이 주변 단어를 기반으로 품사를 추정해야 함에 따른 것으로 사료된다. 반면 BOF를 이용한 모델은 단어를 구성하는 문자 정보를 활용할 수 있으며, 이는 정량적 실험 결과로도 확인할 수 있다. BOF를 이용한 모델은 one-hot 인코딩 이용 모델과 비교하여 OOV 단어는 49.68%, 전체 단어는 0.96% 향상된 성능을 기록하였다.

또 한가지 주목할 점은 모델에 사용된 파라미터의 수이다. BOF를 이용한 모델은 성능이 우수할 뿐만 아니라 사용된 파라미터의 수도 one-hot 인코딩 이용 모델의 32.96%에 불과하다. 추가적으로, one-hot 인코딩을 사용하면 어휘집의 크기와 비례하여 파라미터의 수가 증가하지만 BOF를 사용할 경우 어휘집의 크기와 무관하게 파라미터의 수가 동일한 수준으로 유지된다.

[표 1] PTB WSJ 말뭉치의 평가 데이터를 기준으로 한 품사 부착 정확도. 전체: 모든 단어에 대한 정확도. OOV: OOV 단어에 대한 정확도. 파라미터: 모델의 파라미터 수.

모델	전체	OOV	파라미터
One-hot 인코딩	96.16	57.71	32,119K
Bags of Features	**97.08**	**86.38**	10,585K

[표 2]는 본 연구의 실험 결과와 기존의 영어 품사 부착 연구들의 결과 비교이다. '추가' 열은 해당 연구에서 PTB WSJ 학습 데이터 외의 추가적인 데이터를 사용했

[표 2] 기존 연구들의 품사 부착 정확도(PTB WSJ 말뭉치의 평가 데이터 기준). 전체: 모든 단어에 대한 정확도. OOV: OOV 단어에 대한 정확도. 추가: 모델의 학습에 PTB WSJ 말뭉치 외의 데이터를 활용했는지 여부.

모델	전체	OOV	추가
Manning (2011)	97.32	90.79	Yes
Shen (2007)	97.33	89.61	No
Sun (2014)	97.36	-	No
Moore (2015)	97.36	91.09	Yes
Hajič (2009)	97.44	-	Yes
Søgaard (2011)	97.50	-	Yes
Tsuboi (2014)	97.51	91.64	Yes
Huang (2015)	97.55	-	Yes
Choi (2016)	**97.64**	**92.03**	Yes
본 연구	97.08	86.38	No

는지 여부를 나타내는데, 최근 최첨단 자연어처리 시스템들에서 널리 사용되는 단어 임베딩이 추가 데이터 활용의 대표적인 예이다. 본 연구에서의 실험 결과는 모델의 파라미터 학습에 추가적인 데이터를 활용하지 않는 연구 방법들과 비견할 만한 성능을 보였다.

Collobert(2011)[4]에 따르면, 추가 데이터를 활용한 단어 임베딩의 적용은 모델의 성능에 큰 향상을 가져온다. 본 연구에서 제안된 모델에도 마찬가지로 단어 임베딩을 적용하여 성능을 향상시킬 수 있을 것으로 기대되며, 이는 향후 추가 연구로 수행할 수 있을 것이다.

5. 결론

본 연구에서는 단어를 고정 길이 벡터로 변환하는 단순하면서도 효과적인 방법인 Bag of Features를 이용하여 영어 품사 부착 모델을 구현하고 성능을 정량적으로 비교 평가하였다. BOF 방법은 문자 수준에서 동작하므로 사전에 정의된 한정적인 어휘집이 필요하지 않다. 또한, 문자 단위로 단어를 처리하는 기존의 방법과 달리 합성곱 신경망 혹은 회귀 신경망과 같은 추가적인 구조가 요구되지 않는다. 영어 품사 부착 실험 결과, one-hot 인코딩을 사용한 비교 모델과 대비하여 OOV 단어에 대한 품사 부착 정확도에서 49.68%의 성능 향상을 보였으며, 모든 단어에 대한 품사 부착 정확도 또한 0.96% 상승함을 관찰하였다. 기존의 영어 품사 부착 연구들과의 비교에서도 추가 데이터를 활용하지 않는 모델들과 비견할 만한 성능을 나타냄을 확인할 수 있었다.

Acknowledgement

본 연구는 문화체육관광부 및 한국콘텐츠 진흥원의 2017년도 문화기술 연구개발 지원 사업으로 수행되었음. [2017. 스마트 시니어세대의 문화향유를 위한 인지반응 맞춤형 UI /UX기술 개발]

참고문헌

[1] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. pages 3111-3119, 2013.

[2] Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

[3] 이찬희, 이설화, 임희석. "Bag of Characters를 응용한 단어의 벡터 표현 생성 방법", 한국컴퓨터교육학회 하계학술대회 학술발표 논문집, 제21권, 제2호, pp. 47-49, 2017.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. Journal of Machine Learning Research 12(Aug):2493-2537, 2011.

[5] A. Graves and J. Schmidhuber. Frame-wise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks 18(5):602-610, 2005.

[6] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[7] N. Srivastava, G. E Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929-1958, 2014.

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

[9] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S Corrado, A. Davis, J. Dean, M. Devin. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.