Word Embeddings 자질을 이용한 한국어 개체명 인식 및 분류

최윤수○ 차정원
창원대학교
cyunsu77@changwon.ac.kr, jcha@changwon.ac.kr

Korean Named Entity Recognition and Classification using Word Embedding Features

Yun-Su Choi○ Jeong-Won Cha
Changwon National University

요 약

한국어 개체명 인식 및 분류에 다양한 연구가 있었지만 영어권 개체명 인식 및 분류에 비해 자질 부족 문제를 가지고 있다. 본 논문에서는 한국어 개체명 인식 및 분류에서 자질 부족의 문제를 해결하기 위해 word embeddings으로 생성한 자질을 사용하는 방법을 제안한다. Word embeddings으로 생성한 자질을 개체명 인식 및 분류에 사용하였을 때 기존 시스템보다 성능이 향상되는 것을 보여 그 효용성을 보였다.

1. 서 론

개체명(Named Entity)이란 문서에서 특정한 의미를 가지고 있는 단어 또는 어구를 말한다. 정보 검색에서 개체명은 주요 검색 대상이 된다. 이러한 개체명을 추출하기 위해 자연어 처리 분야에서 개체명 인식 및 분류(Named Entity Recognition and Classification) 연구가 활발하게 진행되었다.

한국어 개체명 인식 및 분류를 위한 방법으로 [2]에서는 CRFs를 이용하였고 [3]에서는 2단계 최대 엔트로피 모델을 이용한 개체명 인식 방법이 있었다. [4]은 Structural SVMs 및 Pegasos 알고리즘을 이용하여 기존 CRFs를 이용한 시스템보다 높은 성능을 유지하면서 학습 시간을 4% 줄였다. [5]는 딥 러닝을 이용한 개체명 인식 방법을 연구 하였는데 영어에 비해 자질이 부족한 한국어에 자질 튜닝 작업에 들어가는 시간과 노력을 줄이면서 기존의 개체명 인식기 성능과 큰 차이가 없음을 보였다.

Word embeddings이란 문장속의 word의 관계를 비지도 학습(Unsupervised learning)방식으로 분석하여 특징화 하는 것이다. [6]에서는 다양한 word embeddings 방법을 이용하여 chunking과 개체명 인식을 수행하고 그 성능을 비교하였다. [7]에서 제안된 CBOW 모델은 현재 word의 문맥을 이루는 벡터들의 합으로 현재 word의 벡터를 결정하는 모델이다. NNLM(Neural net Language Model)의 구조를 변경해 은닉층(Hidden Layer) 대신 투영층(Projection Layer)을 사용해 학습시간을 100배 이상 단축시키고 기존 방법보다 높은 성능을 보였다.

본 논문에서는 한국어 개체명 인식 및 분류에서 자질 부족 문제를 극복하기 위해 word embeddings을 개체명 인식 및 분류를 위한 자질로 사용하는 방법을 제안한다. CBOW 모델과 K-means를 이용하여 각각 형태소 단위의 word vector와 word cluster symbol을 생성하고, 이를 CRFs의 자질로 사용하여 개체명 인식 및 분류를 실험하였다. 실험 결과 word embedding 자질을 개체명 인식 및 분류에 사용할 경우 의미 있는 성능 향상이 있었다.

2. Word Embeddings을 이용한 한국어 개체명 인식 및 분류

본 연구에서는 형태소 단위의 개체명 인식 및 분류 방법을 사용하였다. 형태소 단위로 개체명을 인식 및 분류할 때, 개체명의 형태소 경계를 구분해야하는 문제가 발생한다. 형태소 경계를 표현하기 위해 BIO 형태의 개체명 태그를 사용하였다.

또한 기계학습 기반 개체명 인식 및 분류 방법으로 CRFs를 사용하여 개체명 인식 및 분류를 수행하였다. 한국어 개체명 인식 및 분류를 위한 CRFs의 자질은 표 1과 같다. 이 중 일반 명사 사전이랑 국립국어원 표준국어대사전과 세종 말뭉치에서 명사들을 모은 후, 개체명이 될 수 있는 명사를 제외시키고 남은 명사로 생성한 사전이다.

표 1 한국어 개체명 인식을 위한 자질 정보

자질 정보
형태소 어휘 정보 (-2 ~ +2)
형태소 품사 정보 (-2 ~ +2)
형태소 길이
형태소의 어절 내 위치
어절의 마지막 형태소가 조사일 경우
개체명 사전 존재 유무
일반 명사 사전 존재 유무

Word embeddings을 개체명 인식 및 분류에 이용한 방법은 다음과 같다. 우선 word embeddings을 수행하고자 하는 단위로 구성된 학습 말뭉치를 준비한다. 본 연구에서는 형태소 단위의 word embeddings을 수행하였기 때문에 형태소 분석이 된 대량의 학습 말뭉치를 생성하였다. 대량의 학습 말뭉치가 준비되면 CBOW 모델을 사용하여 word embedding을 수행한다. Word embeddings을 수행하면 d차원의 실수 값으로 이루어진 word vector가 생성되고, 이 word vector의 d개의 실수 값을 각각 CRFs의 자질로 사용하였다.

Word vector를 생성하고 형태소 단위로 생성된 word vector의 실수 값을 K-means를 사용하여 clustering한다. Clustering 수행 후에는 각각의 형태소는 class 번호를 가진다. 이 class 번호가 word cluster symbol이 되고 각 형태소마다 하나의 CRFs 자질로 사용하였다.

3. 실험 및 토의

3.1. 실험 환경

제안된 방법의 효용성을 보이기 위해서 다양한 실험을 진행하였다. 우선 모든 실험에서 CRFs를 이용한 개체명 인식 및 분류를 위해 CRF++[1]을 이용하였다. 본 실험에서는 인명, 학술분야 및 이론, 인공물, 기관, 지역, 문명/문화 관련 명칭, 날짜, 시간, 수량 표현, 이벤트, 동물, 식물, 물질, 용어로 총 14개의 개체명 범주를 사용하였다.

한국어 개체명 인식 시스템의 성능을 측정하기 위해서 TV 도메인과 스포츠 도메인, 그리고 IT 도메인 문서를 사용하였다. TV 도메인에서는 104,759문장을 학습 데이터로 사용하고 3,896문장을 테스트 데이터로 사용하였다. 스포츠 도메인에서는 42,809문장을 학습 데이터로 사용하고 4,000문장을 테스트 데이터로 사용하였다. 마지막으로 IT 도메인에서는 14,075문장을 학습 데이터로 사용하고 1,000문장을 테스트 데이터로 사용하였다.

Word embeddings은 개체명 인식 및 분류와 동일하게 형태소 단위로 수행하였다. CBOW 모델을 사용하여 총 2억 8천만 개 형태소를 학습에 사용하고 50차원의 실수로 이루어진 569,589개 word vector를 생성하였다. Word cluster 자질은 앞서 생성된 word vector와 K-means를 이용하여 각각 200, 300, 400, 500개의 class로 clustering을 수행하였다.

개체명 인식 및 분류와 word embeddings을 수행하기 위해 형태소 분석이 필요하다. 형태소 분석은 창원대학교 적응지능연구실에서 공개한 Espresso를 사용하여 수행하였다[8].

3.2. 실험 결과

표 2와 표 3, 표 4은 TV 도메인과 Sports 도메인 그리고 IT 도메인에서의 한국어 개체명 인식 및 분류 실험 결과이다. 우선 CRFs를 사용한 기본 시스템에 word vector 자질을 추가로 사용하고 실험을 수행하였다.[2] Word vector자질을 추가로 사용하였을 때 TV 도메인과 Sports 도메인, IT 도메인에서 기본 시스템보다 각각 성능이 0.4%, 0.47%, 0.01%가 향상되었다.

다음으로 word cluster자질을 사용했을 때, class 개수별 성능을 보여준다. 세 개의 도메인에서 기본 시스템에 word cluster 자질을 추가로 사용하였을 때 기본 시스템보다 모두 성능이 향상 되었다. TV 도메인에서는 class 200개, Sports 도메인에서는 class 300개로 생성한 word cluster 자질을 사용하였을 때 class 개수 별 성능 중 가장 높은 성능을 보였다. TV 도메인에서는 88.74%로 기본 시스템보다 0.23% 성능이 향상되었고, Sports 도메인에서는 89.93%로 0.48% 성능이 향상 되었다. IT 도메인에서는 class 400개로 생성한 word cluster 자질을 사용하였을 때, 성능이 0.82% 향상되어 81.32%의 가장 높은 성능을 보였다.

마지막으로 기본 시스템에 word vector 자질과 word cluster 자질을 모두 사용하여 실험을 수행하였다. TV 도메인과 Sports 도메인, IT 도메인에서 word vector 자질은 모두 동일하게 사용하였다. Word cluster 자질은 4.2.2장의 각 도메인별 실험에서 가장 성능이 높게 나온 자질을 사용하였다. TV 도메인은 300개, Sports 도메인은 200개 그리고 IT 도메인은 400개로 clustering을 수행한 word cluster 자질을 사용하였다.

표 2과 표 3, 표 4에서 기본 시스템에 추가로 word vector 자질과 word cluster 자질을 모두 사용하였을 때, TV 도메인에서는 89.03%로 기본 시스템보다 0.52% 성능이 향상 되었다. Sports 도메인에서는 89.98%로 0.53% 성능이 향상 되었으며, IT 도메인에서는 80.69%로 0.19% 성능이 향상 되었다.

3.3. 토의

실험에서 TV 도메인에서는 word vector 자질을 추가로 사용했을 때 word cluster 자질을 추가로 사용한 것보다 성능이 좋았지만, Sports 도메인과 IT 도메인에서는 word cluster 자질을 추가로 사용한 것이 word vector를 추가로 사용한 것보다 성능이 좋았다. 또한 TV 도메인과 Sports 도메인에서는 word vector 자질과 word cluster 자질을 모두 사용하였을 때 성능이 가장 좋았지만 IT 도메인에서는 400개의 class로 생성한 word cluster 자질만을 사용하였을 때 성능이 가장 좋았다. 이는 word embeddings을 이용하여 생성한 자질을 모두 사용하는 것보다 선택적으로 사용하는 것이 성능을 더 높을 수도 있다는 것을 보여준다.

4. 결론 및 향후 과제

본 논문에서는 한국어 개체명 인식 및 분류에 word embeddings을 이용하는 방법을 제시하였다. Word embeddings을 이용하는 방법으로 CRFs의 자질로써 형태소 단위의 word vector와 word cluster symbol을 사용하였다. TV 도메인과 Sports 도메인, IT 도메인에서 실험을 수행한 결과 기본 시스템 성능보다 각각 0.52%, 0.53%, 0.82%의 성능이 향상되어 그 효용성을 입증했다.

향후에는 word embeddings을 형태소 단위가 아닌 개체명 단위로 수행하고, 그 word vector를 사용하여 개체

1) https://taku910.github.io/crfpp/

2) Word vector는 실수 값으로 소수점 여섯째자리에서 반올림하여 소수점 다섯째자리로 사용하였다.

표 2 Word embeddings을 이용한 한국어 개체명 인식 및 분류 실험 결과 (TV 도메인)

사용 자질	Precision(%)	Recall(%)	F1 Score(%)
CRF Base	89.15	87.88	88.51
CRF Base + Word Vector	89.24	88.59	88.91
CRF Base + Word Cluster(Class 200)	89.08	88.01	88.54
CRF Base + Word Cluster(Class 300)	89.19	88.29	88.74
CRF Base + Word Cluster(Class 400)	89.24	87.86	88.54
CRF Base + Word Cluster(Class 500)	89.19	88.16	88.67
CRF Base + Word Vector + Word Cluster(Class 300)	89.33	88.73	**89.03**

표 3 Word embeddings을 이용한 한국어 개체명 인식 및 분류 실험 결과 (Sports 도메인)

사용 자질	Precision(%)	Recall(%)	F1 Score(%)
CRF Base	90.42	88.51	89.45
CRF Base + Word Vector	91.00	88.85	89.92
CRF Base + Word Cluster(Class 200)	90.97	88.92	89.93
CRF Base + Word Cluster(Class 300)	90.96	88.85	89.90
CRF Base + Word Cluster(Class 400)	90.81	88.77	89.78
CRF Base + Word Cluster(Class 500)	90.54	88.56	89.54
CRF Base + Word Vector + Word Cluster(Class 300)	91.10	88.89	**89.98**

표 4 Word embeddings을 이용한 한국어 개체명 인식 및 분류 실험 결과 (IT 도메인)

사용 자질	Precision(%)	Recall(%)	F1 Score(%)
CRF Base	82.78	78.34	80.50
CRF Base + Word Vector	82.86	78.30	80.51
CRF Base + Word Cluster(Class 200)	83.33	79.16	81.19
CRF Base + Word Cluster(Class 300)	83.18	78.79	80.92
CRF Base + Word Cluster(Class 400)	83.39	79.36	**81.32**
CRF Base + Word Cluster(Class 500)	83.48	79.18	81.27
CRF Base + Word Vector + Word Cluster(Class 400)	82.91	78.58	80.69

명 인식 및 분류를 수행하는 실험을 진행할 것이다. 그리고 word embeddings에 더 많은 학습 말뭉치를 사용하면 개체명 인식 성능 향상에도 도움이 될 것으로 기대된다.

사 사

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임.
(No. R0101-15-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

이 연구에 참여한 연구자(의 일부)는 「BK21플러스 사업」의 지원비를 받았음

참고문헌

[1] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, “NYU: Description of the MENE Named Entity System as Used in MUC-7”, In Proceedings of the Seventh message Understanding Conference(MUC-7) , 1998.

[2] C. Lee, et al., “Fine-Grained Named Entity Recognition using Conditional random Fields for Question Answering“, The 18th Annual conference on Human & Cognitive Language Technology, 2006.10, 268-272, 2006.

[3] S. Kim and D. Ra, “Korean Named Entity Recognition Using Two-level Maximum Entropy Model”, Korea Information Science Society, 2008.6, 81-86, 2008.

[4] C. Lee and M. Jang, “Named Entity Recognition with Structural SVMs and Pegasos algorithm”, Korean Journal of Cognitive Science, Vol. 21, No. 4, 655～667. 2010.

[5] C. Lee, J. Kim, J. Kim and H. Kim, “Named Entity Recognition using Deep Learning”, Korea Information Science Society, 2014.12, 423-425, 2014.

[6] J. Turian, L. Ratinov and Y. Bengio, “Word representations: A simple and general method for semi-supervised learning”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.07, 384-394, 2010.

[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient Estimation of Word Representations in Vector Space“, ICLR Workshop, 2013

[8] J. Hong and J. Cha “A New Korean Morphological Analyzer using Eojeol Pattern dictionary”, Korea Information Science Society, Vol. 35, No. 1, pp. 279-284, Jun, 2008.