KAISER: 워드 임베딩 기반 개체명 어휘 자가 학습 방법을 적용한 개체명 인식기

함영균○, 최동호, 최기선
한국과학기술원
hahmyg@kaist.ac.kr, zmal0103@kaist.ac.kr, kschoi@kaist.ac.kr

KAISER: Named Entity Recognizer using Word Embedding-based Self-learning of Gazettes

Younggyun Hahm○, Dongho Choi, Key-Sun Choi
KAIST

요 약

본 논문에서는 한국어 개체명 인식의 성능 향상을 위하여 워드 임베딩을 활용할 수 있는 방법에 대하여 기술한다. 워드 임베딩이란 문장의 단어의 공기정보를 바탕으로 그 단어의 의미를 벡터로 표현하는 분산 표현이다. 이러한 분산 표현은 단어 간의 유의미한 정도를 계산하는데 유용하다. 본 논문에서는 이러한 워드 임베딩을 통하여 단어 벡터들의 코사인 유사도를 통한 개체명 사전 자가 학습 및 매칭 방법을 적용하고, 그 실험 결과를 보고한다.

주제어: 개체명 인식, 워드 임베딩

1. 서론

개체명 인식이란 텍스트로부터 개체명(예: 사람, 장소, 조직 등)을 인식하고 분류하는, 자연언어처리의 한 분야이다. 현존하는 개체명 인식 시스템들은 전문가들의 수작업 주석을 통하여 구축된 지도학습 데이터를 학습데이터로 사용한다. 이러한 개체명 인식기의 성능은 학습데이터의 양 이외에도 언어적 특질, 그리고 개체명 사전에 의해 크게 결정되는 것으로 알려져 있다[1]. 영어권 언어의 경우에는 오랜 시간 다양한 학술대회 등을 통하여 양질의 학습데이터가 공개되어 있으며[2][3], 또한 영어의 언어적 특질, 즉 개체명은 대문자로 시작하거나 관사 등의 정보를 활용하여 상당히 많은 성능향상이 이루어지고 있다.

한국어의 경우 언어와는 다른 언어적 특질을 갖고 있어 동일한 방법론을 적용하기 어려우며, 다른 언어권에 비해 상대적으로 공개되어 있는 개체명 학습데이터가 부족한 것이 사실이다. 특히 형태소적 특질이 풍부하여 어휘의 변형이 자유롭게 이루어지기 때문에 개체명 사전의 적용의 어려움이 있다. 이는 한국어 개체명 인식에서 어휘 자질의 부족함을 의미한다.

본 논문에서는 이러한 문제를 해결하기 위하여 워드 임베딩을 사용하였다. 워드 임베딩은 문장의 단어의 공기정보를 바탕으로 그 단어의 의미를 벡터로 표현하는 분산 표현이다. 이러한 분산 표현은 단어 간의 유의미한 정도를 계산하는데 유용하다. 예를 들어, 어휘 “대한민국”의 경우, 어휘 “한국”과 동일한 의미를 갖고 있지만, 학습 데이터 및 개체명 사전에서 “한국”어휘를 포함하고 있지 않다면, 어휘 자질을 사용하였을 때 “한국”을 개체명으로 인식하지 못하게 된다. 그러나 워드 임베딩에서는 “대한민국”과 “한국”이 유사한 벡터공간에 위치함으로서, 두 어휘의 표면형이 다르다 하더라도 유사한 어휘라고 간주할 수 있다.

본 논문에서는 이러한 워드 임베딩을 활용하여, 크게 다음의 두 가지 방법을 적용하였다. 첫째로는 개체명 어휘와의 코사인 유사도 기반 매칭 방법이다. 입력 텍스트에서 개체명 어휘와 동일한 표면형이 아니라 하더라도, 워드 임베딩의 단어 벡터가 유사하다면, 해당 개체명 어휘를 사용하는 방법이다. 둘째로는 개체명 어휘의 자가 학습 방법이다. 이는 입력 텍스트 문서에서 발견된 개체명 어휘를 또 하나의 새롭게 구축된 개체명 사전으로 간주하는 방법이다. 본 논문에서 제안된 방법의 효용성을 실험하기 위하여, 한국어 개체명 인식의 기본 기계학습 모델로 CRF 모델을 사용하였고, 워드 임베딩을 활용한 경우의 성능 향상을 측정하였다. 본 논문에서 사용한 CRF 모델은 워드 임베딩 이외의 자질을 최소화하여 +-2의 창 크기(window size)에서의 형태소의 어휘와 품사를 자질로 사용하였다.

2. 워드 임베딩을 통한 개체명 인식 성능의 향상

본 논문에서 사용한 개체명 사전과 워드 임베딩 데이터는 2016 국어 정보 처리 경진대회[1]에서 공개된 데이터 셋을 사용하였다. 워드 임베딩 데이터는 한국어 위키피디아를 대상으로 skip-gram 모델, 창 크기 5, 50차원의 벡터 값으로 구축되었다.

1) https://sites.google.com/site/2016hclt

2.1 코사인 유사도를 통한 개체명 사전 매칭

어휘 w_1과 어휘 w_2에 대한 워드 임베딩 벡터 값을 각각 $\overrightarrow{w_1}$, $\overrightarrow{w_2}$라고 할 경우, 두 벡터간의 코사인 유사도는 다음과 같이 계산된다.

$$cosine\ similarity = \frac{\overrightarrow{w_1} \cdot \overrightarrow{w_2}}{\|\overrightarrow{w_1}\|\|\overrightarrow{w_2}\|}$$

본 시스템은 CRF 모델에 의해 발견된 개체명 어휘를 제외하고, 발견되지 않은 어휘들 중 특정 품사태그를 갖는 어휘들에 대하여 개체명 사전들과의 코사인 유사도를 측정, 특정 임계값(threshold)를 넘을 경우, 해당 개체명 어휘의 개체명 태그를 부여한다. 예를 들어, 입력 텍스트의 "한국"이 CRF 모델에 의해 개체명으로 발견되지 않았으나, "대한민국" 어휘가 개체명 사전에서 "LC"로 태깅되어 있을 경우, 임계값 0.8에서 코사인 유사도가 0.9로 나타났다면, 어휘 "한국"에 대하여 "LC"로 태깅하여 준다. 이때 사용한 품사태그는 공개된 개발 데이터 셋에서 개체명으로 태깅된 어휘의 품사태그에 한정하여 다음을 사용하였다: {NNP, SN, NNB, NNG, VV, ETM, SL, MM, SO, NR, JX, VX, XSN, JKG, SS, XPN, SF, SH, MAG, JC, JKS, SW, VA.}

2.2 개체명 어휘 자가 학습 기반 개체명 태깅

본 시스템은, 개체명 사전과의 코사인 유사도를 적용할 수 있다는 점에 착안하여, 입력 텍스트 문서에 대하여 CRF 모델에 의해 개체명으로 분류된 어휘를 일종의 개체명 사전으로 간주하는 방식을 적용해 보았다. 예를 들어, 입력 텍스트의 "한국"이 CRF 모델에 의해 개체명으로 발견되지 않은 경우를 생각해 보자. 이 때, "대한민국" 어휘가 CRF 모델에 의해 개체명으로 "B-LC"로 태깅되었다면 이를 하나의 개체명 사전으로 간주한다. 이후 "한국"과 "대한민국"간의 워드 임베딩 코사인 유사도를 측정하여 임계값를 넘을 경우 "B-LC"로 태깅 해 주는 방식이다. 이때, 본 논문에서는 명사에 한정하여서만 본 방식을 적용하였다: {NNG, NNB, NNP}

3. 실험 및 토의

2.1 개체명 어휘 자가 학습 기반 개체명 태깅

본 실험은 다음과 같이 진행하였다. 기본 베이스라인으로 CRF 모델을 사용하였고, 이 모델에서 개체명으로 태깅된 어휘들을 자가 학습된 개체명 사전으로 간주한 뒤, 개체명으로 태깅되지 않은 어휘 중 명사들에 대하여 자가 학습된 개체명 사전 어휘들과의 코사인 유사도를 측정하였다. 이 유사도가 임계값을 넘을 경우, 해당 개체명 어휘의 태그를 부여하는 방식으로 진행하였다. 이 때에는 미리 제공된 개체명 사전 데이터는 사용하지 않았다. 이에 대한 결과는 표 1과 같다.

사용 자질	정밀도	재현율	F1
CRF	91.89	79.81	85.42
CRF + 자가 학습 (유사도 >0.70)	79.77	83.17	81.44
CRF + 자가 학습 (유사도 >0.75)	83.07	82.94	83.00
CRF + 자가 학습 (유사도 >0.80)	83.92	82.90	83.40
CRF + 자가 학습 (유사도 >0.85)	84.58	82.96	83.71
CRF + 자가 학습 (유사도 >0.90)	84.95	82.79	83.85
CRF + 자가 학습 (유사도 >0.95)	84.97	82.75	83.85

표 1 개체명 어휘 자가 학습 기반 개체명 태깅 결과

위 실험 결과에서 볼 수 있듯, CRF 모델만 사용하였을 경우, 정밀도는 매우 높으나 재현율이 상대적으로 낮게 나오는 것을 확인할 수 있다. 이러한 재현율을 향상시키기 위한 방법으로서 본 논문에서는 개체명 사전을 사용하였다.

CRF 모델에서 개체명 어휘를 자가 학습한 뒤, 이 개체명 어휘 사전과 입력 문서의 어휘들과의 유사도를 측정하여, 임계값을 0.7 으로 두었을 때에는 매칭 할 수 있는 개체명 어휘들이 늘어나 재현율이 3.9%가량 증가하였음을 확인하였고, 상대적으로 정밀도가 낮아 전체적인 성능이 하락되었음을 확인할 수 있다. 그러나 유사도의 임계값을 점차 높여갈 경우, 재현율은 크게 손상되지 않으면서 정밀도가 높아지는 것을 확인할 수 있었다. 본 실험 결과로부터, 정밀도의 하락을 일정 부분 감수함으로서 재현율을 높일 수 있음을 확인할 수 있었다. 정밀도를 기본 모델의 수준으로 유지하면서 재현율을 높이는 것을 추후 연구로서 고려할 만 해 보인다. 이는 2.2장에서 다시 논의된다.

2.2 코사인 유사도를 통한 개체명 사전 매칭

본 실험은 2.1에서의 실험과 달리 공개된 개체명 사전을 사용하여 다음과 같이 진행하였다. 기본 베이스라인으로 CRF 모델을 사용하였고, 이 모델에서 개체명으로 태깅되지 않은 어휘들에 대하여 개체명 사전과의 코사인 유사도를 측정하여, 각 임계값 별로 정밀도와 재현율, F1 스코어를 계산하였다. 이에 대한 결과는 표 2와 같다.

사용 자질	정밀도	재현율	F1
CRF	91.89	79.81	85.42
CRF + 개체명 사전 (유사도 >0.70)	76.20	86.83	81.17
CRF + 개체명 사전 (유사도 >0.75)	82.70	85.99	84.31
CRF + 개체명 사전 (유사도 >0.80)	88.65	85.73	87.17
CRF + 개체명 사전 (유사도 >0.85)	90.26	85.40	87.76
CRF + 개체명 사전 (유사도 >0.90)	91.16	85.22	88.10
CRF + 개체명 사전 (유사도 >0.95)	91.89	85.14	88.39

표 2 코사인 유사도를 통한 개체명 사전 매칭 결과

위 결과에서 볼 때, 2.1장에서 자가 학습된 개체명 사전과 달리 보다 정교하게 구축된 개체명 사전의 경우, 정밀도를 베이스라인과 동일하게 확보하면서도, 개체명 사전에서의 어휘를 보다 많이 발견함으로서 재현율이 5.9%가량 향상되었음을 확인할 수 있었다. 이러한 결과를 볼 때, 2.1 장에서 자가 학습된 개체명 어휘들의 품질을, 수작업에 의해 구축 및 검증된 개체명 사전들의

수준으로 끌어올린다면 개체명 인식 시스템의 전반적인 정밀도와 재현율을 향상시킬 수 있을 것으로 기대할 수 있다.

4. 결론 및 향후 과제

본 논문에서는 한국어 개체명 인식의 성능 향상을 위하여 워드 임베딩을 활용할 수 있는 방법 중에 하나로, 개체명 사전을 자가 학습하고 매칭 하는 방법을 제안 하였다. 수작업을 통해 정교하게 구축된 개체명 사전에 대한 코사인 유사도 방식의 매칭 방법을 통해, 정밀도를 유지하면서도 재현율을 향상시킬 수 있음을 확인 하였다. 또한 자가 학습을 통해 생성된 개체명 사전을 사용하였을 경우에도, 정밀도의 하락을 일정 부분 감수하면서 재현율이 향상됨을 확인 하였다. 이러한 두 실험 결과를 통해, 자가 학습을 통해 생성되는 개체명 사전의 품질을 향상시키는 연구가 향후 진행된다면 자가 학습 기반의 개체명 태깅 방법 역시 효과적일 것으로 기대할 수 있다.

사사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임.
(No. R0101-16-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] D. Nadeau, S. Sekine. “A survey of named entity recognition and classification”, Linguisticae Investigationes, 30:3-26, 2007.
[2] N. Chinchor, P. Robinson. “MUC-7 named entity task definition”, In Proceedings of the 7th Message Understanding Conference. 1998.
[3] E. F. Tjong Kim Samg, F. De Meulder. “Introduction to CoNLL-2003 Shared task: Language-independent named entity recognition”, In Proceedings of the 6th Conference on Natural Language Learning, pp.142-147, 2003.