문서 임베딩을 이용한 소셜 미디어 문장의 개체 연결

박영민[1], 정소윤[2], 이정엄[3], 신동수[3], 김선아[3], 서정연[1]

서강대학교 컴퓨터공학과[1], LG전자 소프트웨어센터[2], 현대자동차 융합기술개발팀[3]

pymnlp@gmail.com, soyun.jeong@lge.com, {lee.jeongeom, dshin, seona}@hyundai.com, seojy@sogang.ac.kr

Document Embedding for Entity Linking in Social Media

Youngmin Park[1], Soyun Jeong[2], Jeong-Eom Lee[3], Dongsoo Shin[3], Seona Kim[3], Junyun Seo[1]

Department of Computer Science and Engineering, Sogang University[1],

LG Electronics Software Center[2],

Convergence Technology Development Team, Hyundai Motor Company[3]

요 약

기존의 단어 기반 접근법을 이용한 개체 연결은 단어의 변형, 신조어 등이 빈번하게 나타나는 비정형 문장에 대해서는 좋은 성능을 기대하기 어렵다. 본 논문에서는 문서 임베딩과 선형 변환을 이용하여 단어 기반 접근법의 단점을 해소하는 개체 연결을 제안한다. 문서 임베딩은 하나의 문서 전체를 벡터 공간에 표현하여 문서 간 의미적 유사도를 계산할 수 있다. 본 논문에서는 또한 비교적 정형 문장인 위키백과 문장과 비정형 문장인 소셜 미디어 문장 사이에 선형 변환을 수행하여 두 문형 사이의 표현 격차를 해소하였다. 제안하는 개체 연결 방법은 대표적인 소셜 미디어인 트위터 환경 문장에서 단어 기반 접근법과 비교하여 높은 성능 향상을 보였다.

주제어: 개체 연결, 개체명 인식, 위키백과, 문서 임베딩

1. 서론

자연어 문장의 개체명 인식(Named Entity Recognition) 결과는 인명, 지명, 조직명 등의 태그를 갖지만 각 개체명이 구체적으로 어떤 개체를 의미하는지는 불분명 하다. 예를 들어 '*김기범은 반올림에 출연했다.*' 라는 문장에서 '*김기범*'을 인명으로 인식하였더라도 다양한 개체(야구선수, 배우, 가수 등…) 중 어느 개체에 해당하는지 모호하다. (그림 1)과 같이 이러한 모호성을 해결하는 작업을 개체 연결(Entity Linking)이라 한다.

기존의 개체 연구는 주로 입력 문장과 지식 베이스 문서에 동시에 출현하는 단어를 기반으로 한 단어 기반 접근법(Word-based Approach)을 사용하기 때문에 비정형 문장에서는 높은 성능을 기대하기 어렵다. 본 논문에서는 문서 임베딩(Document Embedding)과 선형 변환(Linear Transformation)을 이용하여 이러한 단점을 극복하고자 한다.

2. 관련 연구

기존의 개체명 연결은 개체와 지식베이스에 출현하는 단어, 하이퍼링크 등을 이용하여 두 대상의 문맥 유사도를 계산하는 방법들이 연구되었다[1-2]. 소셜 미디어의 문장은 길이가 짧아 문맥정보가 충분하지 않은 문제가 있다. 이러한 문제를 극복하기 위해 Shen의 연구[3]에서는 사용자의 성향을 모델링하였고, 정소윤의 연구[4]에서는 사용자 과거 문장과 최근 발생한 뉴스 주제를 모델링 하여 성능 향상을 이룬바 있다.

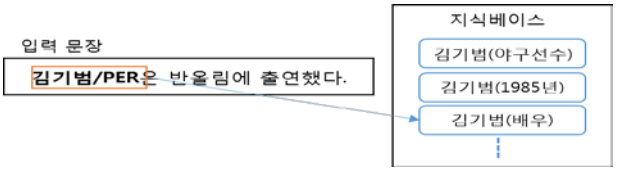

그림 1. 개체 연결의 예

3. 심층학습을 이용한 개체 연결

본 논문에서 제안하는 개체 연결 모델의 전체 구조는 (그림 2)에서 확인할 수 있다. 제안하는 개체 연결 모델은 2 개의 문서 임베딩 모델, 1 개의 선형 변환 모델 그리고 코사인 유사도로 구성된다. 개체 연결에서 연결 대

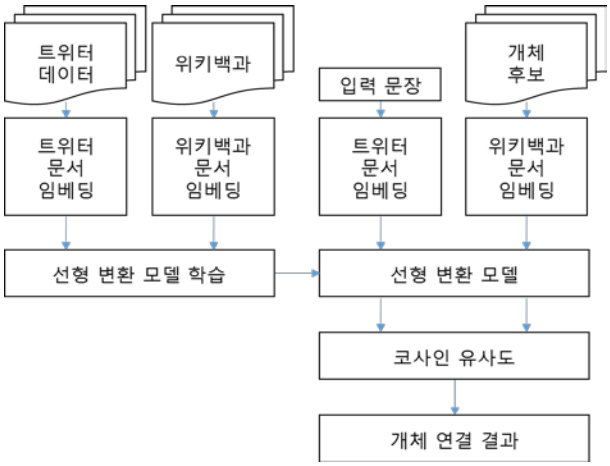

그림 2. 개체 연결 모델의 구성

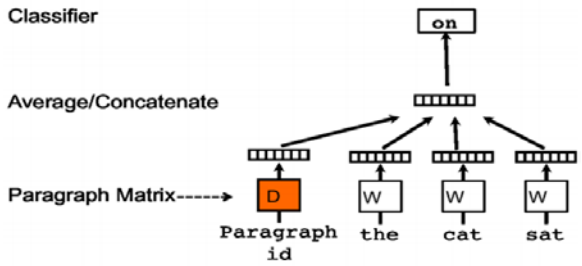

그림 3. Distributed Model의 구조

상이 되는 지식베이스의 개체로서 위키백과(wikipedia.org)의 문서를 사용하는 것을 위키화(wikification)라고 한다. 본 논문에서도 위키화를 대상으로 한다.

3.1 문서 임베딩

최근 단어 임베딩(Word Embedding)을 확장하여 자연어 문장이나 문서를 벡터 공간으로 표현하는 문서 임베딩(Document Embedding)[5]이 제안되었다. 문서 임베딩은 DM(Distributed Memory) 또는 DBOW(Distributed Bag of Word)로 구성되는데 본 논문에서는 DM 모델을 사용한다. DM 모델은 (그림 3)과 같이 각 문서에 단락 식별자(Paragraph id)와 해당 문서를 구성하는 단어 열을 입력으로 하고, 다음 단어가 Softmax 분류기에 의해 예측되는 신경망으로 구성된다. 이때 D와 W는 확률적 경사 강하법(Stochastic Gradient Descent)에 의해 학습된다. 그리고 테스트 단계에서 문서가 입력되면 Softmax 분류기와 W의 가중치(weight)를 고정한 상태에서 신경망 학습을 수행이 수행된다. 학습이 수행 된 후 단락 행렬 D가 입력 문서에 대한 문서 임베딩이 된다.

3.2 선형 변환

트위터와 같은 소셜 미디어의 문장과 위키백과 문서의 문장은 표현 방식에서 큰 차이가 있기 때문에 유사한 내용을 포함하더라도 문서 임베딩의 유사도가 낮게 나올 가능성이 높다. 본 논문에서는 위키백과 문서의 벡터 표현과 트위터 문장의 벡터 표현 사이에 선형적 사상이 가능하다는 가정을 하고 선형 변환을 수행하여 이러한 문제를 해결하고자 한다.

트위터 문장의 문서 임베딩을 T, 위키백과의 문서 임베딩 출력을 W 그리고 두 임베딩의 선형 변환 행렬을 M이라 하면 다음과 같은 식으로 표현된다.

$$TM=W \quad (1)$$

M을 학습하기 위한 T와 W를 수집하기 위해 문서 W는 위키백과 문서를 임베딩의 입력으로 사용하고, T는 해당 위키백과 문서의 제목을 포함하는 트워터 문장을 수집하여 임베딩의 입력으로 사용한다. 또한 M을 학습하기 위해 확률적 경사 강하법을 사용한다.

3.3 개체 연결을 위한 코사인 유사도

각 모델의 학습이 끝난 후 입력 문장에 대해서 개체 연결을 수행할 때 입력 문장의 문서 임베딩T_t를 선형 변환하여 위키백과 문서 임베딩 W_t로 변환한다. 그 후 각 개체 후보 문서들의 문서 임베딩 W_i에 대해서 아래의 코사인 유사도를 계산하여 가장 유사도가 높은 개체 i를 선택하게 된다.

$$\cos(W_t, W_i) = \frac{W_t \cdot W_i}{||W_t||\,||W_i||} \quad (2)$$

4. 실험

실험은 한국어 위키백과 문서와 한국어 트위터를 대상으로 수행하였다. 위키백과 문서의 문서 임베딩을 학습하기 위해 본문 내용이 200 어절 이상으로 구성된 약 32만개의 문서를 수집하였다. 트위터 문서의 문서 임베딩을 학습하기 위해 임의의 트위터 문장을 약 38만개 수집하였다. 위키백과 지식베이스는 동명이인 문서가 존재하는 인명을 대상으로 구축하였고, 트위터에서 임의의 300명 사용자가 작성한 문장 중 지식베이스의 인명이 출현한 문장 248개를 수집하여 테스트 문장으로 사용하였다. 테스트 문장에서 평균 약 3.45명의 개체 모호성이 있었다. 형태소 분석을 위해 Twitter 한국어 형태소 분석기(https://github.com/twitter/twitter-korean-text)를 사용하였다.

비교 모델은 [2]과 [3]의 모델을 재구현하여 사용하였고 실험 결과는 (표 1)과 같다. 실험 결과에서 볼 수 있듯이 기존의 모델은 비정형 문장에 대해서 성능이 크게 하락한 것을 확인할 수 있다. 반면 본 논문에서 제안하는 문서 임베딩과 선형 변환 모델은 단어 기반 모델과 비교하여 크게 개선된 성능을 보여주었다. 이러한 결과는 제안하는 문서 임베딩과 선형 변환이 트위터 문장과 위키백과 문서 사이의 표현 차이 문제를 일정 부분 해소하는 것이라 할 수 있다.

모델	정확도
링크 기반[2]	0.31
링크 + 사용자 모델[3]	0.59
문서 임베딩	0.71
문서 임베딩 + 선형 변환	0.74

표 1. 실험 결과

5. 결론

본 논문에서는 문서 임베딩과 선형 변환을 이용한 개체 연결 모델을 제안하였다. 제안하는 모델은 단어 기반 개체 연결의 단점을 보완하여 비정형 문장에 대해서도 뛰어난 성능을 보여주었다. 향후 소셜 미디어에 부착된 다양한 메타 정보(해쉬태그, 날짜와 시간, 리트윗 관계 등)의 활용 법, 인명 개체 이외의 개체에 대한 실험 등

을 수행할 계획이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [R0126-16-1112, 퍼스널 미디어가 연결공유결합하여 재구성 가능케 하는 복합 모달리티 기반 미디어 응용 프레임워크 개발]

참고문헌

[1] R. Bunescu and M. Pasca, Using Encyclopedic Knowledge for Named Entity Disambiguation, in Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 9-16, 2006.

[2] E. Charton, M. J. Meurs, L. Jean-Louis and M. Gagnon, Mutual Disambiguation for Entity Linking, in Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 476-481, 2014.

[3] W. Shen, J. Wang, P. Luo and M. Wang, Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling, in Proc. of the 19th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 68-76, 2013.

[4] 정소윤, 박영민, 강상우, 서정연, “유저 모델과 실시간 뉴스 스트림을 사용한 트윗 개체 링킹”, 인지과학, 제26권, 2제4호, pp. 435-452, 2015, 12.

[5] Q. Le and T. Mikolov, Distributed Representations of Sentences and Documents, In in Proc. of the 31st International Conference on Machine Learning, pp. 1188-1196, 2014.