자막 병렬 코퍼스를 이용한 이중 언어 워드 임베딩*

이설화† · 이찬희† · 임희석†
† 고려대학교 컴퓨터학과

Bilingual Word Embedding using Subtitle Parallel Corpus

Seolhwa Lee† · Chanhee Lee† · Heuiseok Lim†
† Dept. of Computer Science and Engineering, Korea University

요 약

최근 자연 언어 처리 분야에서는 단어를 실수벡터로 임베딩하는 워드 임베딩(Word embedding) 기술이 많은 각광을 받고 있다. 최근에는 서로 다른 두 언어를 이용한 이중 언어 워드 임베딩(Bilingual word embedding) 방법을 사용하는 연구가 많이 이루어지고 있는데, 이중 언어 워드 임베딩에서 임베딩 결과의 질은 학습하는 코퍼스의 정렬방식에 따라 많은 영향을 받는다. 본 논문은 자막 병렬 코퍼스를 이용하여 밑바탕 어휘집(Seed lexicon)을 구축하여 번역 연결 강도를 향상시키고, 이중 언어 워드 임베딩의 사전(Vocabulary) 확장을 위한 언어별 연결 함수(Language-specific mapping function)를 학습하는 새로운 방식의 모델을 제안한다. 제안한 모델은 기존 모델과의 성능비교에서 비교할만한 수준의 결과를 얻었다.

1. 서 론

자연 언어 처리 분야에서 많은 각광을 받고 있는 워드 임베딩(Word embedding)은 단어를 실수벡터로 표현하여 벡터공간으로 임베딩하는 언어 모델링 기술이다. 특히, 워드 임베딩은 기계 번역, 자동 음성 인식 등의 많은 자연 언어 처리 태스크에서 활용되며, 대표적으로 Skip-gram 모델과 CBOW(Continuous Bag of Words) 모델이 최근 가장 많이 사용하는 대표적인 워드 임베딩 모델이다[1].

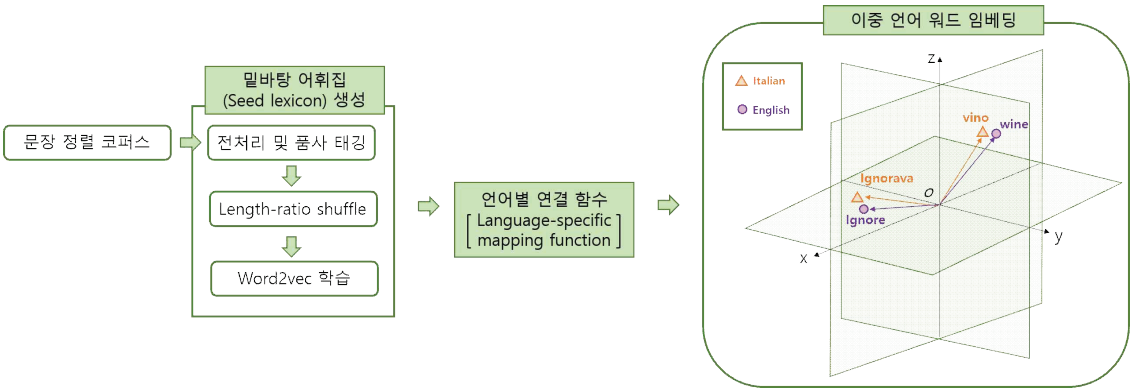
최근에는 이중 언어 워드 임베딩(Bilingual word embedding)방법을 사용하는 연구가 기계 번역 분야에서 많이 이루어지고 있다. 이중 언어 워드 임베딩은 두 개의 다른 언어로부터 하나의 공간으로 단어를 임베딩하는 방법으로, 서로 다른 두 언어에서 유사한 의미를 가지는 단어가 유사한 공간에 매핑(Mapping)되도록 하는 것을 목표로 한다.

이중 언어 워드 임베딩에서 임베딩이 잘 되도록 하는 가장 중요한 요소는 서로 다른 언어 간의 연결 고리(Bilingual signal) 강도에 따라 임베딩 결과에 영향을 미칠 수 있다. 서로 다른 언어 간의 연결 고리란 단어 번역 쌍을 의미하는데, 연결 고리 강도는 어떤 정렬방식을 사용한 코퍼스를 밑바탕 어휘집(Seed lexicon)으로 쓸지에 따라 달라진다. 또한 밑바탕 어휘집은 서로 다른 언어 코퍼스를 이용해 모델을 학습시킴으로써 생성될 수 있다. 코퍼스의 정렬방식은 크게 세 가지로, 문서 정렬, 문장 정렬, 단어 정렬로 구성된다. 문서 정렬 코퍼스는 위키피디아와 같이 비교적 얻기 쉬운 코퍼스라는 장점이 있지만 문서 단위의 코퍼스이기 때문에 연결 고리강도가 약할 수 있다는 단점이 있다. 단어 정렬 코퍼스는 워드 임베딩의 특성상 많은 학습데이터를 필요로 하는데, 질 좋은 단어 단위의 번역 데이터를 얻는데 한계가 있다. 문장 정렬 코퍼스 또한 서로 다른 언어의 병렬 코퍼스로써 질 좋은 많은 양의 데이터를 구하는데 어려움이 있다.

본 논문은 문서 정렬 코퍼스보다는 언어 간의 연결 고리 강도가 강한 문장 정렬 코퍼스를 이용한 이중 언어 워드 임베딩 모델을 제안한다. 제안하는 모델은 자막 코퍼스를 강력한 언어 간의 연결고리로써 밑바탕 어휘집으로 사용하고, 서로 다른 두 언어를 동일한 공간의 벡터 공간으로 매핑한다. 이 과정에 대한 상세한 내용은 3장에서 논하기로 한다.

본 논문의 구성은 다음과 같다. 2장에서 이중 언어 워드 임베딩 모델에 관한 기존 연구, 3장에서 본 논문에서 제안하는 이중 언어 워드 임베딩 모델 및 실험 결과에 대해 서술하고, 마지막으로 4장에서 결론 및 향후 연구로 구성된다.

* 이 논문은 2017년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임(No. 10079423).



[그림 1] 이중 언어 워드 임베딩 개요

2. 관련 연구

이중 언어 워드 임베딩(Bilingual word embedding) 모델은 학습 코퍼스의 정렬 방식에 따라서 크게 세 가지로 분류할 수 있다.

첫 번째로 기계 번역 도구를 이용한 자동 번역을 통해서 단어 정렬 코퍼스를 획득하여 모델을 학습하거나 정렬된 단어 코퍼스로 이중 언어 사전 형식[2]으로 구성된 코퍼스를 획득하여 모델을 학습한다. 그러나, 단어 정렬 코퍼스는 양질의 코퍼스를 획득하기 어려운 한계점이 있고, 특히 좋은 학습 결과를 얻기에는 데이터의 양이 적다. 이중 언어 사전 형식의 코퍼스의 경우는 WMT 데이터셋[3]을 많이 사용하지만, 마찬가지의 이유 때문에 모델의 학습용으로는 적절하지 않으며, 테스트 셋으로써의 기능에 충실하다고 할 수 있다.

두 번째는 서로 다른 언어의 문서 번역 쌍을 이용한 모델로, 비교 가능한 텍스트(Comparable text)로써 위키피디아 또는 뉴스 기사를 코퍼스로 사용하여 모델의 학습을 진행한다[4]. 문서 정렬 코퍼스는 비교적 얻기 쉬운 장점이 있지만, 단순히 문서 쌍의 텍스트 위치에 기반한 문서 번역 쌍을 기준으로 학습을 진행하기 때문에 언어 간의 연결고리가 비교적 약할 수 있다.

세 번째는 단어 정렬 코퍼스와 문서 정렬 코퍼스의 한계점을 어느 정도 보완된 문장 정렬 코퍼스를 학습 코퍼스로 사용하는 것이다. 문장 정렬 코퍼스 또한 양질의 코퍼스를 얻는데 한계점이 있지만 단어 정렬 코퍼스 보다는 비교적 얻기 쉬우며, 문장 단위로 병렬 처리된 코퍼스이기 때문에 문서 정렬 코퍼스보다 언어 간의 연결고리가 강한 장점이 있다.

따라서, 본 연구는 문장 수준의 정렬 코퍼스를 사용한 모델을 제안한다.

3. 이중 언어 워드 임베딩 모델

본 연구에서 제안하는 이중 언어 워드 임베딩 모델은 문장 정렬 코퍼스로 자막 코퍼스를 이용하여 밑바탕 어휘집을 생성하여 두 언어를 하나의 공간으로 매핑하고, 언어별 연결 함수(Language-specific mapping function)를 통하여 번역 쌍을 확장하는 모델이다.

3.1 문장 정렬 코퍼스

본 연구에서 사용한 문장 정렬 코퍼스는 OPUS 코퍼스(영화 자막 코퍼스)[5]를 사용하였고, 이탈리아어-영어 쌍에 대해 학습을 진행하였다. 데이터의 병렬 문장 쌍은 2634M이며, 영화 자막 데이터는 서로 다른 언어가 문장 단위로 병렬 처리되어 있어서 밑바탕 어휘집을 만드는데 양질의 데이터가 될 수 있으며 강력한 연결고리로써의 작용을 할 수 있다.

English	Italian
1 care for the earth	1 cura della terra
2 care for people	2 cura delle persone
3 share the surplus	3 condividi il superfluo

[그림 2] 이탈리아어-영어 자막 병렬 코퍼스 예시

3.2 밑바탕 어휘집 (Seed lexicon)

[그림 1]의 모델 개요에서 문장 정렬 코퍼스를 학습 데이터로 사용하고, 기호 등을 제거하는 전처리 작업을 거친다. 한국어의 경우에는 품사 태깅을 통하여 다시 전처리 과정을 거칠 필요성이 있으나 본 논문에서는 이탈리아어-영어에 대해 학습을 수행하였으므로 별도의 품사 태깅이 필요하지는 않는다.

그 후에 서로 다른 두 언어쌍의 병렬 문장들을 길이-비율 섞음(Length-ratio shuffle)[4]을 통하여 서로 다른 언어의 각 문장 토큰단위 비율로 섞는 작업을 수행한다.

마지막으로 워드 임베딩 기법 중 하나인 Word2vec 학습을 하게 되는데, Skip-gram 모델 [1]을 이용하여

단어로부터 문맥을 학습하는 방식으로 학습이 이루어진다. 300차원을 기준으로 학습을 진행하였고, 학습 후, 이탈리아어(IT)-영어(EN)와 영어(EN)-이탈리아어(IT) 단어 쌍에서 출력되는 단어 벡터 값의 코사인 유사도 값을 계산하여 최종 단어 쌍을 밑바탕 어휘집으로 사용하게 된다. 이렇게 생성된 최종 단어 쌍은 작은 규모의 이중 언어 워드 임베딩 모델로써 밑바탕 어휘집으로 사용된다.

3.3 언어별 연결 함수 (Language-specific mapping function)

연결 함수는 밑바탕 어휘집 쌍의 수가 적은 한계점을 보완하고자 사전(Vocabulary)을 확장 및 강력한 연결 강도를 밑바탕으로 모델을 학습하는 것에 목적이 있다. 여기서 확장으로 쓰이는 코퍼스는 방대한 양과 얻기 쉬운 특성을 가지고 있는 위키피디아 코퍼스를 사용한다.

Vulic [5]의 연구에서는 서로 다른 언어에 대한 단어 쌍 X와 Y를 단일 연결 함수를 사용하여 동시에 학습하였다.

이는 번역 단어 쌍을 하나의 임베딩 공간에 같은 인스턴스로 보기 때문에 각 단어에 대한 임베딩 정보 손실을 야기할 수도 있다. 본 논문은 위의 가정에서 출발하여 각 언어에 대해 [수식 1]과 [수식 2]와 같이 연결함수를 각각 학습하는 모델을 제안한다. 다음 수식은 L2-regularization 문제로 간주하여 해결할 수 있다.

$$\min_{W \in R^{d_{seed} \times d_{wiki}}} ||XW_{Italian} - Y||_F^2 + \lambda ||W_{Italian}||_F^2$$

수식 1

$$\min_{W \in R^{d_{seed} \times d_{wiki}}} ||XW_{english} - Y||_F^2 + \lambda ||W_{english}||_F^2$$

수식 2

수식 1은 이탈리아어에 대한 연결함수이고, $R^{d_{seed}}$는 밑바탕 어휘집에 대한 임베딩 공간 ,$R^{d_{wiki}}$는 위키피디아에 대한 임베딩 공간을 나타낸다. X는 밑바탕 어휘집에서 이탈리아어에 대한 단어 벡터이고, Y는 이탈리아어 위키피디아에서의 단어 벡터이다.

수식 2는 영어에 대한 연결함수로 마찬가지로, X는 영어에 대한 단어 벡터이고, Y는 영어 위키피디아에서의 단어 벡터이다. 전체 연결함수의 학습은 밑바탕 어휘집에 있는 단어를 기준으로 위키피디아 단어를 밑바탕 어휘집의 공간으로 매핑하는 최적의 W 매트릭스를 학습시키는 것이다. λ는 L2-regularization에서의 패널티를 주는 가중치를 나타낸다.

3.4 실험 결과

3.4.1 정성적 평가

<표 1>은 18M의 밑바탕 어휘집 쌍 중에서 일부만을 정성적인 평가결과로 나타낸 것이다. 정성적인 평가로 모든 내용을 넣기에는 양이 너무 방대하지만 대체적으로 결과가 좋은 것을 확인할 수 있었다.

<표 1> 밑바탕 어휘집의 정성적 평가

IT	EN
Ignorava	Ignored
Lasciarmi	Leave
Prestatore	Lender
Diurna	Daytime
Rivelero	Reveal

3.4.2 정량적 평가

본 연구에서 사용한 정량적 평가의 데이터 셋은 [6]에서 오픈한 이탈리아어-영어 테스트 데이터셋 1K를 사용하여 수행하였고, Top-1 정확도 스코어 값을 계산하였다.

<표 2> 모델의 정량적 평가

모델	IT-EN (Acc_1)
Vulic [6]	0.667
제안한 모델	0.657

<표 2>의 결과에서 기존 모델과 제안한 모델간의 성능을 비교하였을 때, 0.001 정도의 score값 차이가 발생하였지만, 기존 모델과 비교할만한 수준의 결과가 나왔다. 본 결과는 1K개의 정답 셋 중에 10개가 더 오답으로 체크된 스코어 값이므로, 연결 함수에서의 L2-regularization의 패널티 파라미터 λ에 얼마나 가중치를 두느냐에 따라 모델의 성능이 달라질 수 있다.

4. 결론 및 논의

본 연구는 이중 언어 워드 임베딩의 학습 데이터의 정렬 유형에 따라 임베딩 결과에 영향을 미칠 수 있다는 것에서 출발하여 자막 코퍼스를 이용하여 강력한 밑바탕 어휘집으로 사용하고, 서로 다른 두 언어를 동일한 공간의 벡터 공간으로 매핑하는 모델을 제안하였다.

또한 각 언어에 대해 각각 학습하는 언어별 연결 함수를 제안하여, 기존의 단일 연결 함수의 한계점을 검증할 수 있는 실험을 수행하였다. 기존 모델에 비해 스코어 값이 낮게 나왔지만, 비교할만한 정도의 유의미한 수치이다. 이는 학습시에 패널티의 가중치에 따라 모델의 성능이 개선될 수 있다는 점에서 긍정적인 결과를 기대할 수 있을 것이다.

참 고 문 헌

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

[2] Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

[3] Specia, L., Turchi, M., Cancedda, N., Dymetman, M., & Cristianini, N. (2009, May). Estimating the sentence-level quality of machine translation systems. In 13th Conference of the European Association for Machine Translation (pp. 28-37).

[4] Vulić, I., & Moens, M. F. (2016). Bilingual distributed word representations from document-aligned comparable data. Journal of Artificial Intelligence Research, 55, 953-994.

[5] Lison, P., & Tiedemann, J. (2016, May). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In LREC.

[6] Vulic, I., & Korhonen, A. (2016). On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In ACL (1).