중간언어 기반의 Word2Vec와 CCA를 이용한 이중언어 사전 추출

김정태[O†], 김창현[‡], 천민아[†], 김재훈[†], 김재환[†]
[†]한국해양대학교, [‡]한국전자통신연구원
jtkim@kmou.ac.kr, chkim@etri.re.kr, minah0218@kmou.ac.kr, jhoon@kmou.ac.kr, jhkim@kmou.ac.kr

Pivot-based Bilingual Lexicon Extraction Using Word2Vec and CCA

Jeong-Tae Kim[O†], Chang-Hyun Kim[‡], Min-Ah Cheon[†], Jae-Hoon Kim[†], Jae-Hwan Kim[†]
[†]Korea Maritime and Ocean University, [‡]Electronics and Telecommunications Research Institute

요 약

이중언어 사전은 자연어처리 분야에서 매우 유용한 자원으로 사용되고 있다. 그러나 초기사전이나 병렬말뭉치 등 자원이 부족한 언어 쌍에 대해서 이중언어 사전을 추출하는 것은 쉽지 않다. 이러한 문제를 해결하기 위해 본 논문에서는 중간 언어 기반으로 Word2Vec와 CCA를 이용하여 이중언어 사전을 추출하는 방법을 제안한다. 본 논문에서 제안하는 방법의 성능을 평가하기 위해서 중간언어로 영어를 사용하여 스페인어-한국어에 대한 이중언어 사전을 추출하는 실험을 하였다. 무작위로 뽑은 200개의 단어에 대한 번역 정확도를 구하였다. 그 결과 최상위에서 37.5%, 상위 10위에서 63%, 그리고 상위 20위에서는 69.5%의 정확도를 얻을 수 있었다.

주제어: Word2Vec, CCA, 이중언어, 대역어 사전

1. 서론

이중언어 사전은 기계번역, 교차언어 정보 검색 등 자연어 처리 분야에서 매우 유용한 자원으로 사용되고 있다. 이중언어 사전을 수동으로 추출하는 것은 많은 시간과 비용이 든다. 따라서 자동으로 이중언어 사전을 추출하는 방법들이 연구되어 왔다. 가장 단순한 방법으로는 병렬말뭉치(parallel corpus)로부터 이중언어 사전을 추출하는 방법이 있다[1]. 그러나 잘 알려지지 않은 언어 쌍에 대해서 병렬말뭉치를 구하는 것은 쉽지 않으며 특정 도메인에 제한되어 있다. 이러한 이유 때문에 비교말뭉치(comparable corpus)로부터 이중언어 사전을 추출하는 연구들이 진행되었다[2, 3]. 하지만 잘 알려지지 않은 언어 쌍에 대해서 비교말뭉치를 구하는 것 또한 어렵고, 초기사전(seed dictionary)도 부족하다. 이와 같은 문제를 해결하기 위해 영어와 같은 잘 알려진 언어를 중간언어를 기반으로 이중언어 사전을 추출하는 방법들이 연구되었다[4, 5].

한편 두 언어의 말뭉치의 단어들을 Word2Vec를 이용하여 다차원의 단어벡터로 표현한 후 선형변환(linear transformation)을 통해서 이중언어 사전을 추출하는 방법이 제안되었다[6]. 그리고 정준상관분석(canonical correlation analysis; CCA)을 이용한 선형변환을 통해 이중언어 사전을 추출하는 방법도 제안되었다[7]. 이 방법들은 기존의 방법들 보다 높은 정확도를 보여주었다.

본 논문에서는 초기사전이나 병렬말뭉치 등 자원이 부족한 언어 쌍에 대해서 중간언어를 기반으로 Word2Vec와 정준상관분석을 이용하여 이중언어 사전을 추출하는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 방법에 대해 기술한다. 4장에서는 실험 및 결과를 제시하며, 5장에서는 결론 및 향후 연구에 대해서 논의한다.

2. 관련 연구

2.1 Word2Vec

Word2Vec는 한 단어에 대해 주변에 있는 단어들을 인공 신경망(neural network)으로 학습시켜 의미적으로 유사한 단어들을 벡터공간상에 가깝게 표현해주는 단어표현방법(word embedding method)이다[8]. 이 연구에서는 CBOW(continuous bag-of-words)모델과 skip-gram모델을 제시하였다. CBOW모델은 skip-gram모델에 비해 빠른 속도로 학습한다는 장점을 가지고 있고, skip-gram모델은 말뭉치에서 빈도수가 낮은 단어에 대해서 더 나은 벡터로 표현해준다는 장점을 가지고 있다.

2.2 정준상관분석

정준상관분석은 두 데이터 집합 사이에 존재하는 상관관계(correlation)를 파악하기 위한 통계적 분석기법이다. 정준상관분석을 통해 두 데이터 집합의 벡터들을 같은 벡터공간으로 투영(projection)시키는 두 투영행렬(projection matrix)을 구할 수 있다. 두 투영행렬을 찾는 방법은 다음과 같다. 두 데이터 집합을 각각 X, Y라고하자. X의 변수들의 선형결합 v_1과 Y의 변수들의 선형결합 w_1간의 상관계수(coefficient of correlation)가 최대가 되도록 하는 X의 계수벡터 a_1과

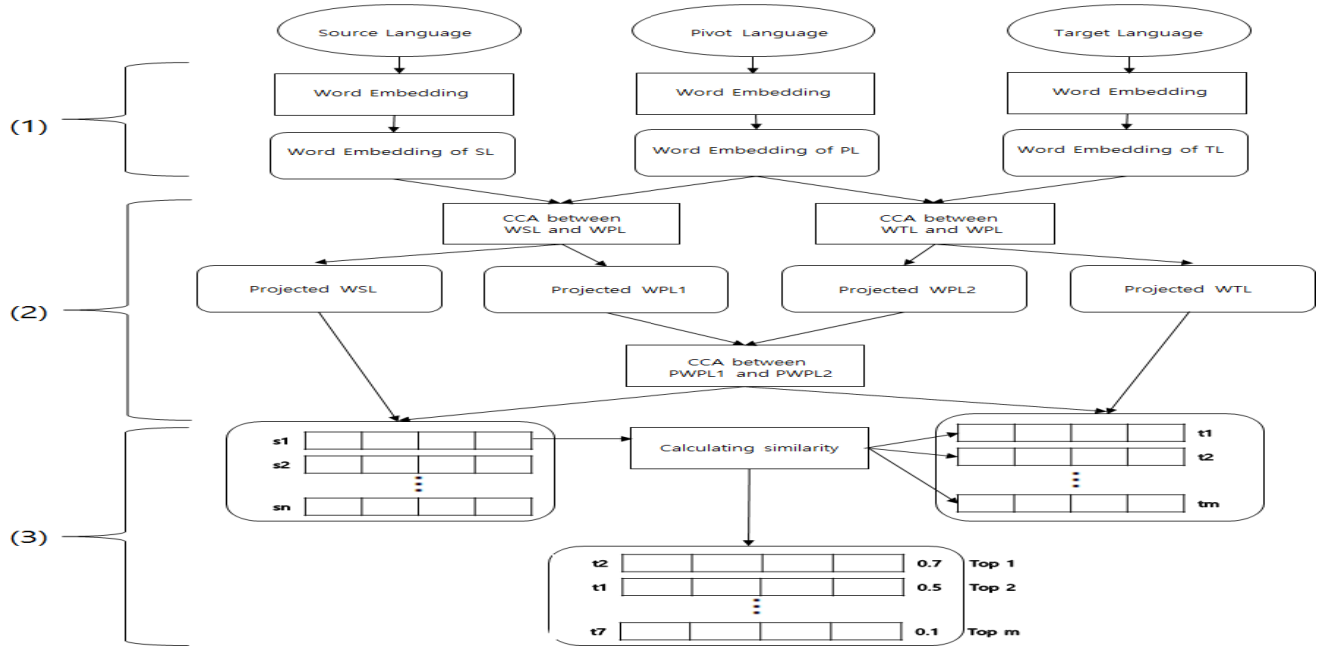

그림 1 중간언어 기반의 Word2Vec와 CCA를 이용한 중간언어 사전 추출방법의 개념도

Y의 계수벡터 b_1을 구한다. 다음으로 또 다른 X의 변수들의 선형결합 v_2과 Y의 변수들의 선형결합w_2 추정하는 일로, 이때 v_1과 w_1과 각각 서로 독립이면서 v_2와 w_2의 상관계수가 최대가 되도록 하는 X의 계수벡터 a_2와 Y의 계수벡터 b_2를 구한다. 이러한 과정을 통해 X와Y 의 계수벡터를 d=min{$rank(X)$, $rank(X)$}개 구할 수 있다. $a_1, a_2, \cdots, a_d$를 열벡터로 갖는 행렬A와 b_1, $b_2, \cdots, b_d$를 열벡터로 갖는 행렬B는 각각 X와 Y의 벡터들을 같은 벡터공간으로 투영시키는 두 투영행렬이 된다.

3. 제안하는 방법

그림 1은 본 논문에서 제안하는 이중언어 사전 추출 방법의 전체적인 개념도이며, 전체적인 과정은 크게 3단계로 이루어져 있다. 3.1절에서는 각 언어의 단어벡터를 구하는 방법에 대해 설명하고, 3.2절에서는 원시언어의 단어벡터와 대상언어의 단어벡터를 같은 벡터공간에 투영하는 방법에 대해 설명한다. 마지막으로 3.3절에서는 번역 후보 순위 결정방법에 대해 설명한다.

3.1 전처리 및 Word2Vec를 이용한 단어임베딩

원시언어, 중간언어 그리고 목표언어의 말뭉치는 다음과 같은 전처리과정을 거친다.

- 형태소분석기를 사용하여 토큰을 분리하고 품사 태깅한다.
- 문자기호(!?<...)는 제거한다.
- 숫자들은 모두 @Number@로 바꾼다.

전처리과정을 거친 각 언어들의 단어에 대해 Word2Vec의 skip-gram모델을 이용하여 단어벡터를 각각 생성한다. skip-gram모델을 이용한 이유는 2.1절에서 언급한 바와 같이 빈도수가 낮은 단어에 대해 더 나은 벡터로 표현 해주기 때문이다. 다음 단계부터는 성능의 개선을 위해 명사들만 뽑아 사용한다. 이 과정은 그림 1에서 (1)에 해당한다.

3.2 학습벡터 구성 및 CCA를 이용한 투영

본 본문에서는 원시언어의 단어벡터와 대상언어의 단어벡터를 같은 벡터공간에 투영하기 위해서 중간언어의 단어벡터를 이용한다. 방법은 다음과 같다. 우선 정준상관분석을 이용하여 원시언어의 단어벡터와 중간언어의 단어벡터를 같은 벡터공간으로 투영한다. 두 투영행렬을 구하기 위한 학습벡터는 원시언어-중간언어의 초기사전을 이용하여 구성한다. 원시언어의 한 단어의 단어벡터 k와 그 단어의 번역단어들의 단어벡터들 $v_1, \cdots, v_p$의 중심벡터(centroid vector)를 학습벡터로 사용한다. 같은 방법으로 중간언어의 단어벡터와 대상언어의 단어벡터를 같은 벡터공간으로 투영한다. 앞의 방법을 통해 구한 두 투영된 중간언어의 단어벡터들을 학습벡터로 사용하여 정준석관분석을 통해 투영된 원시언어의 단어벡터와 투영된 대상언어의 단어벡터를 같은 벡터공간에 투영한다. 이 과정은 그림 1에서 (2)에 해당한다.

3.4 유사도 계산 및 번역 후보 순위 결정

코사인 유사도(cosine similarity)를 이용하여 원시언어의 단어벡터와 대상언어의 단어벡터 사이의 유사도를 계산한다. 계산된 코사인 유사도의 값에 따라 순위를 매긴 뒤 번역 후보를 추출한다. 이 과정은 그림 1에서 (3)에 해당한다.

4. 실험 및 결과

본 논문에서는 원시언어는 스페인어, 중간언어는 영어 그리고 대상언어로는 한국어를 사용하여 이중언어 사전을 추출하는 실험을 해보았다.

본 논문에서 사용된 말뭉치는 다음과 같다. 스페인어 말뭉치는 웹사이트 www.statmt.org/europarl에 있는 스페인어-영어 병렬말뭉치를 이용하였고, 한국어 말뭉치는 뉴스기사로부터 수집된 한국어-영어 병렬말뭉치와 독립적인 뉴스기사 말뭉치를 사용하였다. 그리고 영어 말뭉치로는 스페인어-영어 병렬말뭉치와 한국어-영어 병렬말뭉치를 사용하였다.

제안한 방법의 성능을 평가하기 위해서 무작위로 200개의 평가 단어를 뽑아 실험하였고, 평가척도로는 식 (1)과 같은 정확도(accuracy)를 사용하였다.

$$ACC_k = \frac{1}{N}\sum_{i=1}^{N}\max_{1 \le j \le k} a_{ij}, \quad a_{ij} = \begin{cases} 1 \, if \, t_{ij} \in A_i \\ 0 \, otherwise \end{cases} \quad (1)$$

정확도는 모든 평가 단어에 대하여 평가기준 상위 k개 이내에 정답이 한 개 있는 평가 단어들의 수의 평균이다. N은 평가 단어의 수이고, A_i는 i번째 평가단어의 정답 집합이고 t_{ij}는 i번째 평가단어에 대해서 시스템이 제시한 j번째 후보대역어이다. 즉, a_{ij}는 상위 k개 이내에 적어도 한 개 이상 올바르게 번역되면 1이고 그렇지 않으면 0 이다.

그림 2는 스페인어-한국어의 이중언어 사전추출 정확도에 대한 막대그래프이다. 최상위에서 37.5%, 상위 10위에서 63%, 그리고 상위 20위에서 69.5%의 정확도를 확인할 수 있었다.

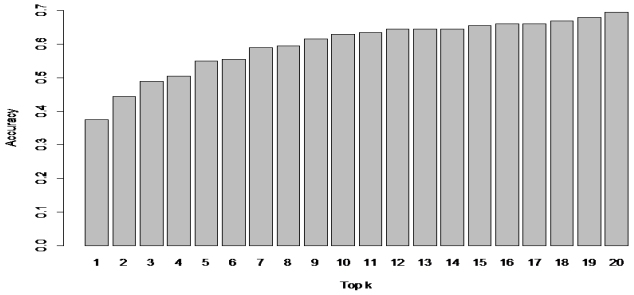

그림 2 스페인어-한국어의 사전추출 정확률

5. 결론 및 향후 연구

본 논문에서는 중간언어 기반으로 Word2Vec와 정준상관분석을 이용하여 이중언어 사전 추출 방법을 제안하였다. 제안한 방법을 성능을 확인하기 위해서 원시언어로 스페인어, 중간언어로 영어 그리고 대상언어로는 한국어를 사용하여 이중언어 사전을 추출하는 실험을 하여 좋은 성능을 확인할 수 있었다.

정준상관분석은 선형성을 가정한다는 문제점을 가지고 있다. 이러한 문제점을 극복하기 위해 딥러닝(deep learning)을 이용한 정준상관분석이 제안되었다[9]. 향후 연구로는 이 방법을 이용하여 이중언어 사전을 추출해 볼 계획이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R7119-16-1001, 지식증강형 실시간 동시통역 원천기술개발]

참고문헌

[1] D. Wu and X. Xia, "Learning an English-Chinese lexicon from a parallel corpus", Proceedings of the First Conference of the Association for Machine Translation in the Americas, pp. 206-213, 1994.

[2] K. Yu and J. Tsujii, "Bilingual dictionary extraction from Wikipedia", Proceedings of the 12th Machine Translation Summit, pp. 379-386, 2009.

[3] A. Tamura, T. Watanabe and E. Sumita, "Bilingual lexicon extraction from comparable corpora using label propagation", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pp. 24-36, 2012.

[4] R. Rapp, "Automatic identification of word translation from unrelated English and German corpora", Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 519-526, 1999.

[5] H. Kwon, H. Seo, and J. Kim, "Bilingual lexicon extraction via pivot language and word alignment tool", Proceedings of the 6th Building and Using Comparable Corpora, pp. 14-19, 2012.

[6] T. Mikolov, Q. V. Le and I. Sutskever, "Exploiting Similarities among languages for machine translation", arXiv preprint arXiv:1309.4168, 2013.

[7] C. Zhang and T. Zhao, "Bilingual lexicon extraction with forced correlation from comparable corpora", Proceedings of the International Conference on Neural Information Processing. Springer International Publishing, pp. 528-535, 2015.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013.

[9] G. Andrew, R. Arora, J.A. Bilmes and K. Livescu, "Deep canonical correlation analysis", International Conference on Machine Learning, pp. 1247-1255. 2013.