자기조직화 지도를 이용한 이중언어사전 자동 구축

서형원, 천민아, 김재훈
한국해양대학교 컴퓨터 공학과
wonn24@gmail.com
minah-cheon@naver.com
jhoon@kmou.ac.kr

Bilingual Lexicon Extraction Using Self-Organizing Maps

Hyeong-Won Seo, Minah Cheon, Jae-Hoon Kim
Department of Computer Engineering, Korea Maritime and Ocean University

요 약

본 논문은 인공신경망(artificial neural network)의 한 종류인 자기조직화 지도(self-organizing map)를 이용하여 비교말뭉치(comparable corpora)로부터 이중언어사전(bilingual lexicon)을 자동으로 구축하는 방법에 대하여 기술한다. 일반적으로 우리가 대상으로 하는 언어 쌍마다 말뭉치 혹은 초기사전과 같은 언어 자원을 수집하고 그것을 필요에 맞게 가공하는 것은 매우 어려운 일이다. 이런 관점에서 볼 때, 비지도학습(unsupervised learning) 방법 중 하나인 자기조직화 지도를 이용하여 사전을 구축하면 다른 방법에 비해 적은 노력으로도 더 높은 성능을 얻을 수 있다. 본 논문에서는 한국어와 불어에 대하여 실험을 하였고, 그 결과 적은 양의 초기사전으로도 주목할 만한 정확도를 얻을 수 있었다. 향후 연구로는 학습 파라미터에 대해 좀 더 다양한 실험을 하고, 다른 언어 쌍으로의 적용 및 기존의 평가사전을 확장하여 더 많은 경우에 대해 실험하는 것을 들 수 있다.

Keyword: 인공신경망, 자기조직화 지도, 이중언어사전, 비교말뭉치, 초기사전

1. 서론

최근에 심층학습(deep learning) 및 인공신경망(artificial neural network)에 대한 관심이 매우 높아지고, 이를 이용하면 기계학습(machine learning) 및 패턴인식(pattern recognition) 등의 분야에서 기존의 방법을 뛰어 넘는 성능을 보인다는 연구 사례들이 차례차례 발표되고 있다[1].

이중언어사전(bilingual lexicon)을 구축하는 연구에도 다양한 기계학습 방법을 적용할 수 있으며, 이미 적용한 사례[2]도 찾아볼 수 있다. 물론, 이중언어사전을 구축하는 방법에 여러가지가 있지만, 그 중, 비교말뭉치(comparable corpora)를 이용하는 방법은 초기사전(seed dictionary)이 반드시 필요하다. 초기사전은 한 쪽 언어를 다른 쪽 언어로 번역하는 곳에 사용되고 이것의 크기가 클 수록 더 좋은 성능을 얻을 수 있다는 점이 큰 특징이자 맹점이다.

본 논문에서는 비교말뭉치로부터 이중언어사전을 자동으로 구축함에 있어서 적은 양의 초기사전을 사용해도 높은 성능을 이끌어내는 것을 목표로 한다. 이를 위해, 비지도학습(unsupervised Learning) 중의 하나인 자기조직화 지도(self-organizing map: SOM)를 이용하는 방법을 제안한다.

논문의 구성은 다음과 같다. 2 장에서 관련된 연구들을 소개하고 3 장에서는 SOM 을 이용한 이중언어사전 구축 방법에 대해 자세히 기술한다. 그리고 4 장에서는 실험 결과를, 마지막 5 장에서는 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

2.1 이중언어사전 구축 방법

이중언어사전을 구축하는 방법은 다양하지만 주로 Rapp [3]이 제안한 방법 (standard approach)을 기초로 한다. Rapp 의 방법은 두 개의 서로 다른 비교말뭉치(원시 언어와 목적 언어)로부터 공기빈도(co-occurrence frequency)와 연관척도(association measure)를 이용하여 벡터를 만들고 초기사전으로 한 쪽 언어를 다른 쪽 언어로 번역한다. 이 때, 초기사전이 얼마나 많은 양의 단어를 포함하는지에 따라 성능이 좌우 된다는 특징이 있다. 이렇게 구축된 벡터는 차원이 같기 때문에 서로 비교가 가능하며, 이를 이용하여 벡터 간 유사도를 계산하고 원시(source) 단어의 번역이 될 목적(target) 언어의 상위 k 개 후보들을 추출하게 된다. 본 논문에서는 새롭게 제안된 방법과 이 Rapp 의 방법을 비교하여 분석한다.

2.2 자기조직화 지도

일반적으로 SOM 은 자연언어처리 뿐만 아니라, 매우 다양한 분야(기후연구, 시장조사 등)에서 사용되고 있는 기술 중의 하나이다. SOM 은 비지도학습 방법으로 주로 저차원(2 차원 이하)의 지도를 생성하여 주어진 입력 예제(sample, instance)에 대한 위상(topological) 속성을 보존하려는 성질을 가진다. 잘 학습된 SOM 을 이용하면 수집된 자료의 군집화(clustering) [4], 자질추출(feature extraction) [5] 및 분류(classification) [6] 등 여러 곳에 사용할 수 있다.

본 논문에서는 SOM 을 학습하여 SOM 벡터를 만들고 이것을 서로 비교하여 원시 언어와 목적 언어간의 번역 사전을 추출한다. SOM 의 입력은 각 언어의 의미 벡터가 되고 이것을 가지고 SOM 을 학습하게 된다. 학습이 무사히 끝나면 2 차원 공간에 입력 예제(혹은 단어)에 대한 위상이 표현되며, 이 지도로 인해 비슷한 입력은 비슷한 입력들끼리 승자를 공유하게 된다. 따라서, 본 논문에서는 입력 벡터가 잘 만들어지고 SOM 의 학습이 잘 된다면 유의어들은 하나의 승자 혹은 그것과 가까운 주변의 승자를 갖게 된다고 가정한다. 이 가정을 전제로, 원시 언어의 SOM 학습이 끝난 후에 목적 언어의 SOM 학습 시 초기사전을 이용하여 각 번역 단어에게 원시 단어의 승자를 알려주어 똑같은 승자 노드를 갖도록 하는 것이 본 논문에서 제안하는 방법의 핵심이다. 자세한 학습 방법은 다음 장에서 설명한다.

3. 이중언어사전 추출 방법

본 논문에서 제안하는 이국어 추출 방법을 개념적으로 설명하면 다음과 같다. (1) 먼저, 두 개의 비교말뭉치로부터 가능한 모든 단어에 대응하는 의미 벡터들을 만든다. (2) 그 다음, 각 언어에 대하여 SOM 학습을 통해 SOM 벡터를 생성한다. (3) 구축된 SOM 벡터들간의 거리 유사도를 계산하고 그것을 기준으로 정렬한다.

3.1 의미 벡터 구축

이중언어사전을 추출하기 위한 가장 첫 번째 단계는 단어를 벡터 공간에 표현하는 것이다. 단어를 벡터 공간에 표현하기 위해서는 여러가지 방법이 있지만 일반적으로는 두 단어 간에 연관척도(point-wise mutual information, log-likelihood, chi-square 등)를 이용하여 의미 벡터를 구성한다. 본 논문에서는 의미적으로 비슷한 단어들(즉, 유의어)은 벡터 공간 안에서 서로 가까운 곳에 위치한다고 가정한다. 가령, '학교'에 대한 의미 벡터와 '대학교'에 대한 의미 벡터 간의 거리 내적(Euclidean distance)을 계산하면 '학교'와 '부동산'의 경우보다 가깝다는 것이다. 이것은 잘 학습된 SOM(다음 절에서 상세히 설명함)을 이용하면 설사 학습에 참여하지 않은 단어라 할지라도 그것과 비슷한 의미를 지닌 다른 단어가 이미 학습되었기 때문에 일련의 반복적 학습 후에 도출된 그들의 SOM 벡터들은 서로 유사할 것이라는 전제에서 출발한 것이다.

이렇게 구축된 의미 벡터들은 문서에 따라 독립적으로 생성되었기 때문에, 각기 다른 문서에서 생성된 벡터들끼리는 그 요소 하나하나가 의미하는 바가 서로 다르다는 특징이 있다. 이 단계에서 생성된 의미 벡터들은 다음 단계의 입력으로 사용된다.

3.2 SOM 의 학습 및 SOM 벡터 생성

본 절에서는 SOM 의 학습 과정과 SOM 벡터 생성에 대하여 설명한다. 본 논문에서 가정한 환경에서는 두 개의 서로 다른 언어(원시와 목적)에 대한 의미 벡터들이 존재한다. 따라서 이들에 대한 독립적인 SOM 학습은 의미가 없다. 다시 말해, 서로 번역이 될 두 단어의 승자가 달라지는 것은 의미가 없기 때문에, 두 SOM 이 서로 관여하면서 학습되는 것이 중요하다. 예를 들어,'학교'에 대한 SOM 의 승자와 'school'에 대한 승자를 같게 하는 것이 매우 중요하다.

학습 순서를 설명하면 다음과 같다. (1) 먼저, 원시 언어에 대하여 SOM 을 학습한다. 이 때에 승자 선택은 사람이 직접 관여하지 않도록 한다. 이것은 학습이 무사히 잘 되었다면 학습에 참여한 '학교'에 대한 승자와 학습에 참여하지 않은 '대학교'에 대한 승자는 서로 같거나 그 주위에 있을 것이라는 가정을 전제로 한다. (2) 그 다음, 목적 언어에 대해서도 SOM 을 학습한다. 이 때, 원시 언어에 대한 승자를 목적 언어에 해당하는 번역 단어들에게 알려준다. 즉, '학교'에 대한 승자 노드가 3 번이라면 'school'에 대해 학습할 때에는 따로 승자를 계산하는 것이 아니라 원시 단어의 승자인 3 번을 이어 받아서 가중치를 조정하는 것이다. 따라서, 목적 언어의 학습이 끝나면 'school'에 대한 승자도 3 번이 되어야 한다.

두 언어에 대하여 SOM 학습이 모두 끝나면 각 단어에 대한 SOM 벡터를 생성해야 한다. 다시 말해, 각 단어에 대해 의미 벡터가 아닌 SOM 벡터를 생성해야 하는데, 각 노드에 해당하는 벡터들과 입력 벡터 간의 내적 값이 입력 단어에 대한 SOM 벡터의 각 요소 값이 되는 것이다. 이렇게 되면, 원시 언어와 목적 언어 간에 SOM 벡터들이 서로 비교가 가능해진다. 주의해야 할 점은 SOM 을 학습할 때 필요한 학습 파라미터와 SOM 의 크기는 실험하는 문서마다 또 입력 벡터의 크기에 따라 달라질 수 있다는 점이다.

3.3 SOM 벡터 간의 유사도 계산 및 정렬

SOM 벡터가 생성되었다면 이제 벡터 간 유사도를 계산하고 그에 따라 정렬을 한다. 이 때, 유사도는 코사인 유사도(cosine similarity)를 이용하여 계산한다. 예를 들어, '학교'에 대한 SOM 벡터와 목적 언어의 모든 단어에 대한 SOM 벡터 간의 유사도를 계산한 후, 그에 따라 상위 k 개의 단어를 추출하면 '학교'의 번역 후보 중 가장 실제 정답과 가까운 후보들이 추출되는 것이다.

4. 실험

본 논문에서 사용된 비교말뭉치는 [2]에서 사용된 한국어와 불어 비교말뭉치를 사용하였다. 각 말뭉치

는 U-Tagger (한국어) [7]와 Tree-Tagger (불어) [8]를 이용하여 모든 형태소들의 품사를 부착하였고 불용어와 기호 등을 제거하였다. 이렇게 전처리가 된 말뭉치를 이용하여 3.1 절에서 언급한 의미 벡터를 구축해야 하는데 벡터의 차원이 굉장히 고차원이기 때문에 학습의 시간이 매우 오래 걸리는 단점이 있다. 따라서, 본 논문에서는 실험의 단순화를 위해 Word2vec[1] [9]를 이용하여 의미 벡터를 만들었다. Word2vec 를 이용하여 단어를 벡터로 표현하면 의미적으로 같은 단어들이 비슷한 성질을 띄는 경우보다는 서로 동시에 발생하는 공기 단어들이 주로 비슷한 값을 갖도록 되는 특징이 있다. 가령, '학교'의 벡터와 유사한 값을 가지는 벡터는 '대학교'나 '고등학교'가 아니라 '선생님', '교실', '급식' 등과 같은 단어도 포함되게 되는 것이다. 엄밀히 말하면, 이런 단어들은 유의어가 아니지만 본 논문에서는 이렇게 Word2vec 로 만들어진 벡터가 번역 단어를 찾는 작업에 충분히 사용될 수 있다고 가정한다. 이렇게 만들어진 벡터는 100 차원의 크기를 가지도록 고정하였고, SOM 학습에 필요한 파라미터들은 여러 번의 실험을 통해 학습 비율(learning rate)은 0.1, 그리고 SOM 의 크기는 900(30×30)의 공간으로 고정하여 실험하였다.

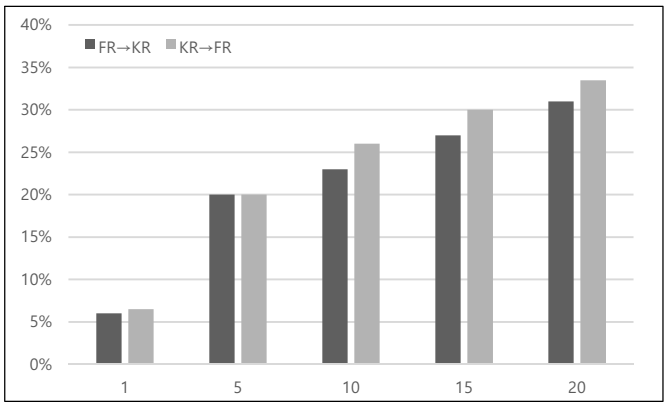
실험 평가를 위해 사용된 평가사전은 다음과 같은 방법으로 사람이 직접 구축하였다. 먼저, 각 말뭉치에서 빈도수가 높은 순으로 언어 당 200 개씩 평가 단어를 선별하였다. 본 논문에서는 실험의 단순화를 위해 명사만을 선별하였다. 이렇게 선별된 명사 단어 200 개 (한국어 200 개, 불어 200 개)에 대하여 사람이 직접 사전 [2]을 이용하여 번역 정답을 부착하였다. 사실상, 원시 단어의 번역 단어가 여러가지 [3]일 수 있지만, 본 논문에서는 하나의 번역 단어를 갖는다는 전제 하에 번역 단어를 하나씩만 부착하였다. 번역 단어를 선정하는 기준은 가능한 여러 번역 후보 중에서 실험에 사용된 문서 안에 출현한 빈도수가 가장 높은 단어를 선정하였다. 우리는 이것이 최선의 번역이라고 가정한다.

초기사전은 다음과 같이 구축하였다. 먼저 원시 단어의 경우, 문서에 출현한 모든 명사 단어 가운데 빈도수가 100 이상인 단어를 287 개씩 [4] 선택한 대신, 위에서 언급한 평가사전에 포함된 단어는 제외하였다. 이것은 정답 단어를 학습에 포함시키지 않아 실험 평가에 공정성을 기하기 위함이다. 이렇게 선정된 원시 단어의 번역 역시 평가사전을 구축한 방법과 동일하게 번역 단어를 하나씩만 부착하였다. 수집된 초기사전의 통계는 <표 1>과 같다.

<표 1> 초기사전의 통계

	한국어→ 불어	불어→ 한국어
원시 단어의 수	287	287
목적 번역의 수	265	266

<표 1>에서 알 수 있듯이 몇몇 번역 단어(불어 22 개,한국어 21 개)들은 하나 이상의 원시 단어에 대응하고 있다. 이럴 경우, 만약 해당하는 원시 단어들이 서로 비슷한 의미를 지니고 있다면 상관이 없지만, 서로 다른 의미를 가질 경우 복수의 승자로 인해 정확성이 떨어지는 단점을 가지게 된다. 물론, 우리의 가정은 하나의 평가 단어로부터 파생된, 즉 유사도가 매우 높은, 복수의 원시 단어(유의어)들이 하나의 승자 혹은 그 주변에 있는 승자를 가지게 된다면 목적 언어에 대해서도 높은 확률로 정답과 그것의 유의어들이 서로 높은 유사도를 갖게 될 것이라고 전제를 하고 있다. 하지만 그렇지 못 할 경우, 즉 어떤 원시 단어로부터 파생된 (초기사전에 포함된) 학습 단어들이 지형적으로 멀리 떨어져 있는 복수의 승자들을 가지고 그로 인해 평가 단어는 그들의 중간점을 승자로 갖게 된다면 번역의 정확도는 떨어질 수 밖에 없다.



(그림 1) 한국어-불어 쌍에 대한 평가 정확도 (가로축: 랭킹, 세로 축: 정확도)

(그림 1)은 한국어-불어 언어 쌍에 대해 200 개의 평가 단어 중 실제 정답을 얼마나 맞췄는지를 나타낸 그림이다. 보이는 바와 같이, 상위 5 위까지를 고려했을 경우 20% 정도는 제안된 방법으로 올바른 번역을 찾을 수 있었다. 이 성능은 200 개의 평가 단어 중 일부(불어: 54 개, 한국어: 49 개) [5]로부터 파생된 학습 단어가 초기사전에 없다는 점을 감안했을 때 주목할만한 결과라고 할 수 있다.

제안된 방법이 어느 정도의 성과를 얻은 것인지 비

[1] http://code.google.com/p/word2vec

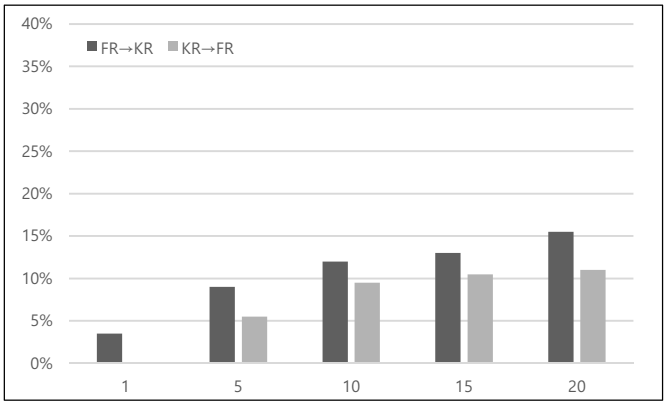
[2] http://dic.naver.com

[3] 예를 들어, 'bank'의 번역이 '은행', '강둑', '경사' 등이 될수 있다. 뿐만 아니라, 각 번역 단어의 유의어 역시 번역 단어로 선정될 자격이 있다.

[4] 실험의 공정성을 위해 동일한 수(287 개)의 소스 단어를 빈도수에 따라 높은 것부터 선택하였다.

[5] 평가사전의 단어와 관련된 학습 단어(유의어)가 없는 경우는 해당 유의어(명사)가 원시말뭉치에 없거나, 그 유의어의 번역이 목적말뭉치에 존재하지 않기 때문에 학습에 포함되지 않은 것이다. 이럴 경우, 원래의 평가 단어는 학습된 다른 평가 단어와 유사하지 않다면 그 성능을 기대할 수 없다.

교 분석하기 위해서 기존에 연구된 Rapp 의 방법 [3]과 비교하였다. 최대한 공정한 비교를 위해 초기사전과 평가사전 그리고 말뭉치 모두 동일한 환경에서 실험하였다. 대신에, SOM 의 입력이 될 의미 벡터는 3.1 절에서 언급한 과정을 거쳐서 구축하였다. 또한, 명사를 제외한 단어들은 무시하였다. 이것은 최대한 의미 벡터의 차원을 줄여서 학습에 걸리는 시간과 부하를 줄이기 위함이다.



(그림 2) 'Rapp 의 방법' 대한 평가 정확도 (가로 축: 랭킹, 세로 축: 정확도)

(그림 2)에서 보이는 바와 같이, 상위 5 위까지를 고려했을 경우 10%에 못 미치는 정확도를 보였다. 하지만 이것은 실험에 사용한 초기사전의 단어 수(200 개)가 [3]에서 사용된 초기사전의 단어 수(16,380 개)에 비해 턱없이 부족한 수이기 때문이다. 하지만, 같은 양의 초기사전을 이용함에도 불구하고 제안된 방법으로는 더 나은 성능을 보이는 것을 확인할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 SOM 을 이용한 이중언어사전 구축 방법에 대하여 제안하였다. 동일한 평가사전과 초기사전으로 한국어와 불어 쌍에 대해 실험한 결과, SOM 은 비지도학습 방법임에도 불구하고 Rapp 의 방법에 비해 주목할만한 성능을 보인다는 것을 확인할 수 있었다. 다만, 초기사전의 양이 적은 것을 감안한다면 앞으로도 충분히 성능 향상의 가능성이 열려 있다고 할 수 있다.

향후 연구로는 더욱 더 다양한 학습 파라미터와 언어 쌍에 대해서 실험하고, 기존의 평가사전과 초기사전을 확장하여 실험해볼 수 있을 것이다. 또한 단일 단어(single-word) 뿐 아니라 다중단어(multi-word)에 대해서도 실험해 볼 수 있겠다.

[감사의 글]

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10041807 , 지식학습 기반의 다국어 확장이 용이한 관광/국제행사 통역률 90%급 자동 통번역 소프트웨어 원천 기술 개발]

참고문헌

[1] 김인중. 2014. "Deep Learning: 기계학습의 새로운 트랜드", 한국통신학회지 (정보와 통신) 31(11): 52-57.

[2] H. Kwon, H.-W. Seo, M. Cheon, and J.-H. Kim. 2014. "Iterative Bilingual Lexicon Extraction from Comparable Corpora Using a Modified Perceptron Algorithm", Journal of Contemporary Engineering Sciences, 7(24): 1335-1343.

[3] R. Rapp. 1999. "Automatic Identification of Word Translations from Unrelated English and German Corpora", In Proceedings of the ACL 37, pp. 519-526.

[4] B. Curry, F. Davis, M. Evans, L. Moutinho, and P. Phillips. 2003. "The Kohonen Self-organizing Map as an Alternative to Cluster Analysis: An Application to Direct Marketing", The Market Research Society, 45(2): 191-211.

[5] Y. Liu, R. H. Weisberg, and C. N. K. Mooers. 2006. "Performance Evaluation of the Self-organizing Map for Feature Extraction", Journal of Geophysical Research, 111, C05018.

[6] N. Jiang, K. Cheung, K. Luo, P. J. Beggs, and W. Zhou. 2011. "On Two Different Objective Procedures for Classifying Synoptic Weather Types over East Australia", International Journal of Climatology, 32(10): 1475-1494.

[7] J. Shin and C. Ock. 2012. "A Korean Morphological Analyzer Using a Pre-analyzed Partial Word-phrase Dictionary", Journal of KIISE: Software and Applications, 39(5): 415-424.

[8] H. Schmid. 1994. "Probabilistic Part-of-speech Tagging Using Decision Trees", International Conference on New Methods in Language Processing, pp. 44-49.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality", Journal of Computing Research Repository, abs/1310.4546 [cs.CL].