이국어 병렬말뭉치와 중간언어를 활용한 이국어 사전 자동구축

서형원*, 권홍석**, 김재훈***
*한국해양대학교 컴퓨터공학과
wonn24@gmail.com, hong8e@naver.com, jhoon@hhu.ac.kr

Automatic bilingual lexicon construction via bilingual parallel corpus and pivot language

Hyeong-Won Seo*, Hong-Seok Kwon**, Jae-Hoon Kim***
*Dept of Computer Science, Korea Maritime University

요 약

본 논문은 한국어-스페인어와 한국어-불어 간의 양방향 이국어 사전(Bi-directional bilingual lexicon)을 자동으로 구축하기 위한 새로운 방법을 제안한다. 일반적으로 한국어와 스페인어/불어 간의 병렬 말뭉치를 직접적으로 구축하기에는 어려움에 따르기 때문에, 영어를 중심언어로 하는 영어(EN)-한국어(KR)/스페인어(ES)/불어(FR) 병렬 말뭉치를 이용하여 문맥 벡터를 만들고 그들 간의 유사도를 계산하는 변형된 문맥 벡터 방법을 제안한다. 영어는 다른 언어와의 이국어 병렬 말뭉치가 비교적 많이 공개되어 있기 때문에 이 방법을 이용하면 비교적 쉽게 KR-ES와 KR-FR 양방향 이국어 사전을 구축할 수 있다. 본 논문에서 제안한 방법으로 실험해본 결과 최고 85%(ES→KR)의 정확도를 얻을 수 있었다.

1. 서론

이국어 사전(Bilingual lexicon)은 일반적으로 두 개 이상의 언어로 표현된 단어 쌍을 정리해놓은 사전을 말하며[1], 많은 자연언어처리(NLP) 분야 특히 기계번역[2]이나 다중언어(multilingual) 정보검색[3] 등에 주요한 자원으로 사용되고 있다[4]. 하지만 이국어 사전을 구축하는 데에는 많은 시간과 노력, 돈 등이 들기 때문에 전 세계의 모든 언어 쌍에 대해서 만드는 것이 거의 불가능하다.

이국어 사전 구축의 어려움을 해결하기 위해 이미 많은 연구들이 진행되어 왔다[5-8]. 이런 연구들은 주로 이국어 말뭉치의 한 부류인 병렬말뭉치(parallel corpora)나 비교말뭉치(comparable corpora)에 대한 연구 혹은 중간언어를 활용하는 방법에 대한 연구다. 이국어 사전을 구축하기 위해서는 초기사전 혹은 이국어 말뭉치의 존재유무가 매우 중요하다. 하지만 본 논문에서 초점을 맞추고 있는 한국어와 스페인어/불어 사이에는 공개된 이국어 사전이 없을 뿐더러 공개된 병렬말뭉치나 비교말뭉치도 없는 실정이다.

따라서 본 논문에서는 위에 나열한 문제들을 해결하기 위해 새로운 방법을 제안한다. 제안하는 방법은 사전 구축 시 파생될 수 있는 동의어와 다의어 문제를 병렬말뭉치의 사용으로 다소 해결할 수 있다고 가정한다. 대신에 병렬말뭉치의 한계점인 구축의 어려움을 해결하기 위하여 중간언어를 이용한 한국어와 스페인어/불어 간의 양방향 이국어 사전 구축 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 이국어 사전과 구축 방법에 대하여 기술하고 3장에서는 본 논문이 성능 비교를 위하여 기본 방법으로 설정한 연구와 더불어 새롭게 제안하는 방법에 대해서 기술한다. 4장에서는 실험에 대한 내용을 기술하고 5장에서 결론을 짓는다.

2. 이전 연구

2.1 이국어 사전 자동구축

이국어 사전을 구축에 대한 기존 여러 연구들이 꾸준히 진행되어왔다[9]. 이런 연구에는 중간언어를 이용하는 방법[5,6][9], 병렬말뭉치를 이용한 방법[10,11], 그리고 비교말뭉치를 이용한 방법[7,8][12]등이 있다. 병렬말뭉치를 이용하여 만든 이국어 사전은 Lardilleux et al.[13]에 의해 충분히 좋은 성능을 보인다는 연구 결과가 있었다. 하지만 병렬말뭉치가 영어를 제외한 언어에 대하여 극히 제한적인데다가 상당히 많은 양을 필요로 한다는 한계점을 가지고 있다.

이런 문제점을 해결하기 위해 중간언어를 활용하려는 기존 연구들이 있었다[4]. 중간언어를 이용하면 제한적인 언어 사이에서도 보다 효과적으로 병렬말뭉치를 구축할 수 있다는 장점이 있다. 이에 반해, 비교말뭉치는 특정 도메인에 대하여 언어는 서로 다르지만 비슷한 문맥을 가지는 문서들(일반적으로 집필된 날짜로 구분해볼 수 있는 뉴스)을 모아놓은 것이기 때문에 병렬말뭉치보다 구축이

용이하고 굳이 영어가 아닌 다른 특정 언어에 대해서도 번역 쌍을 구축하기 쉽다는 장점이 있다. 하지만 기존에 비교말뭉치를 사용하여 이국어 사전을 구축한 Fung[14]의 연구는 초기사전을 사용한다는 특징이 있다. 여기서 말한 초기사전이란, 단어의 양이 많을수록 좋지만 처음에는 적은 양이어도 무방하고 주로 병렬말뭉치가 아닌 비교말뭉치를 이용할 때 사용된다.

본 논문에서는 비교말뭉치 대신에 병렬말뭉치를 활용함으로써 이국어 사전 구축 시 필요한 외부 자원의 사용을 최소화하였다.

3. 제안된 방법

3.1 기본 문맥 벡터 방법

본 절에서는 이국어를 구축하기 위한 가장 일반적인 Fung[14]의 연구를 '기본 문맥 벡터 방법'이라 칭하고 이에 대하여 기술한다. Fung은 이국어 사전을 만들기 위해 문맥 벡터를 이용하였고 그것의 전체적인 과정은 다음과 같다. 먼저 각 원시/대상 비교말뭉치에 대하여 모든 단어를 문맥 벡터로 변환한다. 이때 벡터의 값을 계산하기 위해서 문서 빈도(document frequency)와 연관성 측도(association measure)를 이용한다. 그 다음, 초기사전을 이용하여 원시언어 벡터를 대상언어 벡터로 변환한다. 그리고 변환된 원시언어 벡터와 대상언어 벡터들 간의 유사도(similarity)를 cosine measure와 같은 방법으로 계산한다. 마지막으로 단어마다 계산된 유사도 별로 순위를 매겨서 가장 유사도가 큰 벡터들을 일정 수만큼 추출한다.

위에서 설명한 기본적인 방법을 조금씩 변형한 다른 연구들을 나열해보면 다음과 같다.

<문맥의 범위 조절>

- 3문장[15]
- 3단어 25개[16]

<초기사전의 양 조절>

- 16,000개의 단어[17]
- 대략 2만개[12][14,15]

<벡터들 간의 유사도 계산 방법 조절>

- city-block measure[17]
- cosine [12][14-16]
- Dice or Jaccard indexes[12][15]

3.2 동기

3.1절에서 기술한 방법에 대한 문제를 정리해보면 다음과 같다. 먼저, 기본 문맥 벡터 방법은 비교말뭉치와 초기사전을 사용하는 방법이다. 즉, 초기사전이 필요하다는 것은 그에 따른 정확도가 전체 완성도에 영향을 준다는 것을 의미한다. 또한 사전이 없는 언어마다 구축을 해야 한다는 문제점도 있다.

본 논문의 목적은 앞에서 설명한 기본 문맥 벡터 방법의 문제점을 보완하기 위해 중간언어(본 논문에서는 영어)를 사용하여 한국어(KR)-스페인어(ES)와 KR-불어(FR) 이국어 사전을 자동으로 구축하는 것이다. 하지만 KR-ES/FR 병렬말뭉치가 공개된 것이 없고 비교말뭉치와 초기사전 역시 구축하는 것이 쉽지만은 않다. 이에 반해, 스페인어와 불어는 영어와 쌍을 이루는 병렬말뭉치가 이미 연구 목적으로 공개되어 있다. 따라서 본 논문에서는 중간언어에 기반을 둔 병렬말뭉치를 사용하는 변형된 문맥 벡터 방법을 제안한다. 제안된 방법을 이용하면 원시언어 문맥 벡터를 대상언어에 맞게 변환할 필요가 없기 때문에 초기사전과 같은 외부 요소를 사용하지 않아도 된다는 장점이 있다.

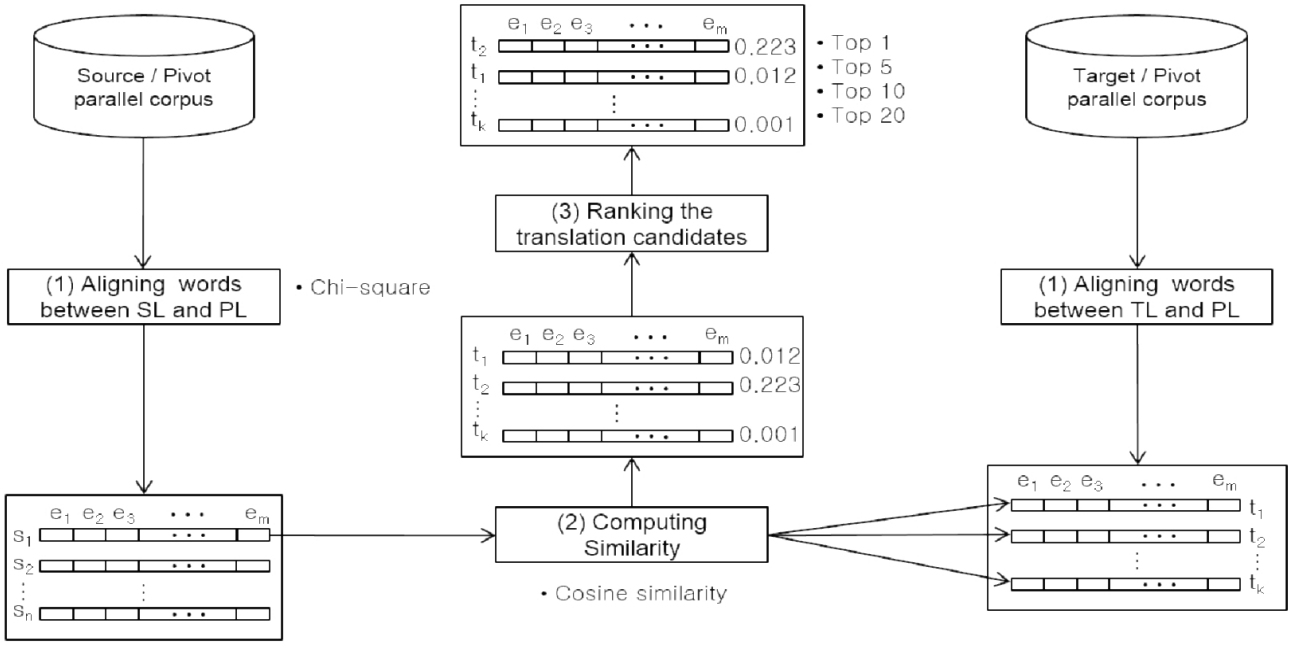

Figure 1: Proposed pivot language based approach

3.3 제안된 방법

본 논문에서 제안하는 방법은 그림 1과 같다. 먼저 (1) 각각의 병렬말뭉치들 즉, SL-PL(Source Language-Pivot Language)와 PL-TL(Pivot Language-Target Language)의 말뭉치에 포함된 의미 있는 단어들에 대하여 연관성을 측정한다. 여기서 의미 있다는 말은 불용어나 기호 등을 제외하고 남은 단어를 말한다. 예를 들어 SL을 KR(한국어), TL을 ES(스페인어)라고 가정하면 KR-EN(SL-PL), EN-ES(PL-TL) 병렬말뭉치에서 각각 기호나 불용어 등을 제거한 나머지 단어들을 모두 추출한다. 이 때, 명사, 동사, 형용사, 부사를 제외한 품사의 단어는 제외시킨다. 그리고 최종 추출된 단어들 간(SL의 단어와 TL의 단어)의 연관성을 측정한다. 본 논문에서는 Chi-square test[18]를 이용하여 두 단어의 연관성을 측정하였다. 여기서 사용되는 단어의 빈도수는 DF(Document Frequency)를 사용하며, 문장을 문서로 간주하여 단어를 포함하고 있는 문장의 수를 세었다. (2) 이렇게 만들어진 문맥 벡터들(SL-PL 문맥 벡터와 TL-PL 문맥 벡터) 사이에 Cosine measure와 같은 유사도 계산 방법을 이용하여 각 벡터들의 유사도를 계산한다. (3) 그 후 유사도가 가장 높은 순으로 벡터를 정렬하면 어떤 SL의 단어가 어떤 TL의 단어와 가장 유사한지를 살펴볼 수 있다.

본 절에서 기술한 방법을 이용하면 3.2절에서 언급한 것처럼 초기사전 혹은 단어정렬도구와 같은 외부 자원 없이도, 단어들 간의 연관성 측정값을 이용하여 문맥 벡터를 만들고 그것들 간의 유사도를 측정함으로써 SL와 TL의 단어들 중 병렬 후보들을 쉽게 추출할 수 있다.

4. 실험 및 결과

4.1 실험 환경

4.1.1 데이터

본 논문에서는 실험을 위해 3개의 병렬말뭉치(KR-EN, EN-ES, EN-FR)를 사용하였다. 여기서 사용된 말뭉치들은 모두 병렬말뭉치이고 KR-EN은 뉴스 도메인, EN-ES는 국회 도메인에 관련된 말뭉치이다. KR-EN 병렬말뭉치는 Seo et al.,[19]의 연구에서 사용한 말뭉치를 보완[1]하여 433,151개의 문장으로 구성된 것이다. 문장 당 평균 단어(한글의 경우에는 형태소)의 수는 각각 19.2(KR), 31(EN)이다. EN-ES 병렬말뭉치와 EN-FR 병렬말뭉치는 European parliament proceedings[2]로부터 50만 문장씩을 무작위로 추출한 것이다. 이들의 문장 당 평균 단어의 수는 각각 25.4(EN from EN-ES), 26.4(ES from EN-ES), 27.1(EN from EN-FR), 29.7(FR from EN-FR)이다.

실험 결과를 평가하기 위한 정답 사전(KR-ES, KR-FR)은 다음과 같은 방법으로 구축하였다. 사전을 구성하는 단어들을 선별하기 위해 각 병렬말뭉치로부터 빈도수가 높은(상위 50% - KR:319~68685번, ES:402~56465, FR:372~55187) 명사 단어를 임의로 200개씩 추출하고, 이것을 사람이 웹 사전[3]을 이용하여 직접 구축하였다. 최종적으로 추출된 정답 사전에 포함된 단어 당 평균 번역단어의 수는 각각 11.41(FR→KR), 10.3(ES→KR), 5.79(KR→FR), 7.36(KR→ES)개이다. 불어와 스페인어를 한글 사전에서 찾은 의미의 개수가 그 반대 방향인 경우보다 상대적으로 많은 것을 확인할 수 있다.

4.1.2 전처리

영어, 스페인어 그리고 불어 모두 TreeTagger[4]를 이용하여 토큰 분리, 원형·사전등재·표제어(lemmatization) 분리를 한 후 품사를 부착하였다. 한글은 한나눔 태거(Kaist Tagger)[5]를 이용하여 품사를 부착하는 전처리만 수행하였다. 그 다음, 품사가 모두 부착된 말뭉치로부터 각각의 언어에 맞는 불용어(한국어 제외)와 특정 품사를 제외하였다. 한국어에 대해서는 총 69개의 품사 중 보통명사, 고유명사, 용언 그리고 수식언을 제외한 나머지 51개의 품사에 해당하는 단어들을 말뭉치로부터 모두 제외시켰다. 나머지 언어에 대해서도 같은 작업을 하여 영어에 대해서 각각 61개 품사 중 19개, 스페인어에 대해서 74개 품사 중 33개, 마지막으로 불어에 대해서 36개 중 18개를 제거하였다.

전 처리 후 남은 단어 종류의 수는 각각 KR-EN 말뭉치에서 67,210(KR의 경우는 형태소)/41719(EN), EN-ES 말뭉치에서 28,764(EN)/12,926(ES), EN-FR 말뭉치에서 51,245(EN)/ 47,220(FR)이다. 이 중에서 같은 Euro-parl 병렬말뭉치임에도 불구하고 두 영어 문서(EN-ES의 영어 문서와 EN-FR의 영어 문서) 안에 서로 중복되지 않는 단어의 총 개수(28,764와 51,245) 차이가 많이 나는 이유는 실제 EN-FR의 영어 문서에 불어 문장이 다수 포함되어 있기 때문이다.

4.2 실험 결과

본 논문에서 제안한 방법으로 실험한 결과는 그림2와 같다. 그림 2는 연관성 측도와 유사도 계산 방법을 각각 Chi-square test, Cosine measure로 정해놓고, 각 병렬말뭉치로부터 얻어진 두 언어 문맥 벡터 간에 유사도 계산 결과의 정확도를 나타낸 것이다. 그림 2에서 볼 수 있듯이, 상위 랭크일수록 성능 차이는 미비하지만 15위 아래로 갈수록 ES→KR과 FR→KR이 다른 결과와 큰 차이를 보인다는 점을 눈으로도 확인할 수 있다. 이것은 4.1.1 절에서 언급한 정답 사전의 분포에 따른 결과라고 판단된다. 그리고 상대적으로 불어에 대한 성능이 스페인어에 대한 성능보다 낮은 것은 EN-FR 병렬말뭉치 중 영어 문서에

1) 동아 뉴스: http://www.donga.com, 조인스 뉴스: http://www.joins.com
2) http://www.statmt.org/europarl/
3) http://dic.naver.com/
4) http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
5) http://kldp.net/projects/hannanum

무수히 많은 불어 문장들이 포함된 것 때문에 벡터 간의 유사도 계산 시 악영향을 받은 것이라고 해석할 수 있다.

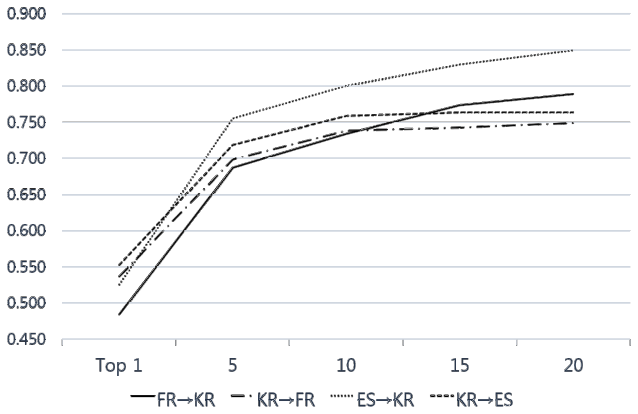

Figure 2: Comparison accuracy of N-best candidates

5. 결론 및 향후 연구

본 논문은 양방향 이국어 사전을 자동으로 구축하기 위한 새로운 방법을 제안하였다. 사전을 만들기 위한 병렬 말뭉치는 특정 언어 사이에 구축하는 것이 쉽지 않기 때문에, 중간언어를 기본으로 포함하는 두 개의 병렬말뭉치를 이용하여 문맥 벡터를 만들고 이들의 유사도를 계산하는 방법을 기술하였다. 제안된 방법을 이용하면 초기사전이나 단어정렬도구와 같이 특별한 다른 외부 자원 없이도 비교적 쉽게 특정 언어에 대하여 이국어 사전을 구축할 수 있다는 장점이 있다. 하지만 병렬말뭉치를 사용함으로써 특정 도메인에 특화된 단어들만 추출할 수 있다는 점과 두 병렬말뭉치의 도메인 역시 비슷해야 한다는 점이 단점으로 지목된다.

향후 연구로는 스페인어와 불어 이외의 언어에 대하여 한국어와 이국어 사전을 구축하는 것과 병렬말뭉치로 얻어진 사전을 이용하여 비교말뭉치를 별도로 수집하여 기존의 사전을 보강하는 연구가 있겠다.

참고문헌

[1] A. Haghighi, P. Liang, T. Berg-Kirkpatrick and D. Klein 2008 "Learning Bilingual Lexicons from Monolingual Corpora" In Proc. of the ACL-HLT pp. 771-779.

[2] P. Brown, J. Cocke, Stephen A. Della Pietra, V. Pietra, F. Jelinek, J. Lafferty, R. Mercer and P. Roossin 1990 "A statistical approach to machine translation" Coling'90 16(2) pp. 79-85.

[3] J. Nie, M. Simard, P. Isabelle, and R. Durand 1999 "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the-Web" In Proc. of the ACM SIGIR pp. 74-81.

[4] T. Tsunakawa, N. Okazaki, and J. Tsujii 2008 "Building Bilingual Lexicons Using Lexical Translation Probabilities via Pivot Languages" In proc. of LREC.

[5] K. Tanaka and K. Umemura 1994 "Construction of a Bilingual Dictionary Intermediated by a Third Language" In Proc. of the Coling'94 pp. 297-303.

[6] L. Nerima and E. Wehrli 2008 "Generating Bilingual Dictionaries by Transitivity" In Proc. of the LREC'08 pp. 2584-2587.

[7] P. Fung 1995 "Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus" In Proc. of the VLC'95 pp. 173-183.

[8] K. Yu and J. Tsujii 2009 "Bilingual dictionary extraction from Wikipedia" In Proc. of the MT Summit XII pp. 379-386.

[9] F. Bond, R. Binti Sulong, T. Yamazaki, and K. Ogura. 2001 "Design and Construction of a machine-tractable Japanese-Malay Dictionary" In Proc. of the MT Summit VIII pp. 53-58.

[10] D. Wu and X. Xia 1994 "Learning an English-Chinese lexicon from a parallel corpus" In Proc. of the AMTA'94 pp. 206-213.

[11] P. Fung and K. Church 1994 "K-vec: A New Approach for Aligning Parallel Texts" In Proc. of the Coling'94 2 pp. 1096-1102.

[12] Y. Chiao and P. Zweigenbaum 2002 "Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora" In Proc. of the Coling'02 pp. 1208-1212.

[13] A. Lardilleux, J. Gosme and Y. Lepage 2010 "Bilingual Lexicon Induction: Effortless Evaluation of Word Alignment Tools and Production of Resources for Improbable Language Pairs" In Proc. of the LREC.

[14] P. Fung 1998 "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora" In Proc. of the Parallel Text Processing, pp. 1-17.

[15] B. Daille and E. Morin 2005 "French-English Terminology Extraction from Comparable Corpora" Natural Language Processing - IJCNLP 3651.

[16] E. Prochasson and E. Morin 2009 "Anchor points for bilingual extraction from small specialized comparable corpora" TAL 50(1) pp. 283-304.

[17] R. Rapp 1999 "Automatic Identification of Word Translations from Unrelated English and German Corpora" In Proc, of the ACL'99 pp. 519-526.

[18] T. Dunning 1993 "Accurate methods for the statistics of surprise and coincidence" Coling'93 19(1) pp. 61-74.

[19] H.-W. Seo, H.-C. Kim, H.-Y. Cho, J.-H. Kim and S.-I. Yang 2006 "Automatically Constructing English-Korean Parallel Corpus from Web Documents" Korea Information Processing Society 13(02) pp. 0161-0164.