중간언어와 단어정렬을 통한 이중언어 사전의 자동 추출에 대한 성능 개선

권홍석○, 서형원, 김재훈
한국해양대학교, 컴퓨터공학과
hong8c@naver.com, wonn24@gmail.com, jhoon@kmou.ac.kr

Performance Improvement of Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool

Hong-Seok Kwon○, Hyeung-Won Seo, Jae-Hoon Kim
Korea Maritime and Ocean University, Department of Computer Engineering

요 약

본 논문은 잘 알려지지 않은 언어 쌍에 대해서 병렬말뭉치(parallel corpus)로부터 자동으로 이중언어 사전을 추출하는 방법을 제안하였다. 이 방법은 중간언어(pivot language)를 매개로 하고 문맥 벡터를 생성하기 위해 공개된 단어 정렬 도구인 Anymalign을 사용하였다. 그 결과로 초기사전(seed dictionary)을 사용한 문맥벡터의 번역 과정이 필요 없으며 통계적 방법의 약점인 낮은 빈도수를 가지는 어휘에 대한 번역 정확도를 높였다. 또한 문맥벡터의 요소 값으로 특정 임계값 이상을 가지는 양방향 번역 확률 정보를 사용하여 상위 5위 이내의 번역 정확도를 크게 높였다. 본 논문은 두 개의 서로 다른 언어 쌍 한국어-스페인어 그리고 한국어-프랑스어 양방향에 대해서 각각 이중언어 사전을 추출하는 실험을 하였다. 높은 빈도수를 가지는 어휘에 대한 번역 정확도는 이전 연구에서 보인 실험 결과에 비해 최소 3.41% 최대 67.91%의 성능 향상을 보였고 낮은 빈도수를 가지는 어휘에 대한 번역 정확도는 최소 5.06%, 최대 990%의 성능 향상을 보였다.

주제어 : Word alignment, Pivot language, Bilingual Lexicon Extraction, Parallel corpus

1. 서론

이중언어 사전은 기계 번역, 교차언어 정보검색 등 분야에서 중요한 자원으로 사용되고 있다. 이중언어 사전의 추출에 있어 가장 직접적인 방법으로는 병렬 말뭉치로부터 대역쌍(translation equivalence)을 추출하는 것이다[1]. 그러나 잘 알려지지 않은 언어 쌍에 대해서 병렬말뭉치를 모으는 일은 쉽지 않으며 특정 도메인에 제한되어 있다. 이러한 이유들 때문에 비교말뭉치(comparable corpus)로부터 이중언어 사전을 추출하는 연구[2][3][4]가 많이 이루어지고 있다. 그러나 잘 알려지지 않은 언어 쌍에 대해서는 비교말뭉치 또한 구축이 쉽지 않다. 이와 같은 문제를 해결하기 위해 중간언어를 매개로 이중언어 사전을 추출하는 연구들[5][6][7]이 있다.

다른 한편으로는 이중언어 사전 추출에 정보 검색 기법을 도입한 연구[8][9][10]도 있다. 정보 검색 기법을 사용한 방법은 간략하게 다음과 같다. 서로 다른 두 언어 L_1과 L_2 각각에 대해서 모든 어휘 단위들을 수집한다. 그리고 수집된 모든 어휘 단위에 대해서 문맥벡터 S와 T를 각각 생성하고 L_1과 L_2의 초기사전을 이용하여 각 문맥벡터 S와 T를 번역한다. 이때 초기사전은 사람에 의해서 수동으로 구축되어진 사전이며 그 용량이 클수록 정확한 번역이 가능해진다. 마지막으로 문맥벡터 S와 T의 유사도를 서로 비교하여 최종 번역 쌍을 추출한다.

본 논문에서는 잘 알려지지 않은 언어 쌍에 대해서 중간언어를 사용하고 정보검색 기법을 기반으로 이중언어 사전을 자동으로 추출하는 간단하고 효과적인 방법을 제안한다. 중간언어는 원시언어(source language)와 대상언어(target language)의 문맥벡터를 표현하는데 사용되고 정보검색 기법은 중간언어로 표현되어진 원시언어 문맥벡터와 대상언어 문맥벡터 사이의 유사도를 비교하는데 사용된다. 기존 연구와는 다르게 우리는 두 개의 병렬 말뭉치(예: 한국어-영어, 영어-스페인어)를 사용한다. 여기서 영어가 중간언어로서 사용된다. 그리고 우리는 문맥벡터를 쉽게 생성하기 위해 공개되어진 단어 정렬 도구인 Anymalign[11]을 사용한다.

우리가 제안한 방법은 다음과 같은 장점들이 있다. 첫째로 영어와 같은 중간언어를 사용하여 잘 알려지지 않은 언어 쌍에 대해서도 쉽게 적용이 가능하다. 둘째로 복합단어로 표현된 어휘에도 쉽게 확장이 용이하며 마지막으로 큰 규모의 초기사전을 수동으로 구축해야하는 수고스러움이 없다.

본 논문의 구성은 다음과 같다. 2장에서는 우리가 제안한 방법의 전체구성과 각 단계에 대해서 설명하고, 3장에서는 실험과 그 결과에 대해서 설명한다. 마지막으로 4장에서는 결론 및 향후연구에 대해서 논의한다.

2. 제안 방법

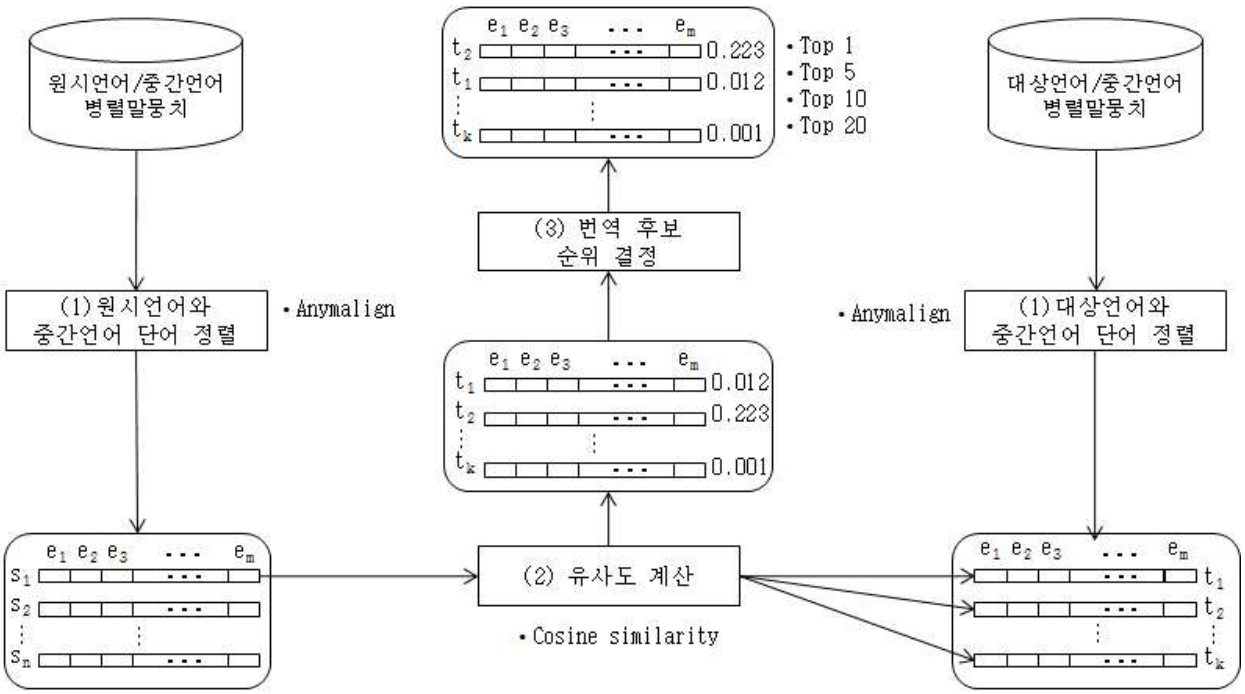

그림 1. 이중언어 사전 구축 전체 시스템 다이어그램

본 논문에서는 잘 알려지지 않은 언어 쌍에 대해서 중간언어와 정보검색 기법을 사용하여 간단하고 효과적으로 이중언어 사전을 추출하는 방법을 제안한다. 우리는 더 정확한 어휘 정렬 정보를 얻기 위해서 비교말뭉치가 아닌 병렬말뭉치를 사용한다. 그러나 잘 알려지지 않은 언어 쌍들에 대해서 병렬말뭉치를 모으는 일은 쉽지 않다. 이러한 이유들 때문에 우리는 잘 알려진 영어를 중간언어로써 사용한다.

중간언어는 원시언어와 대상언어의 문맥벡터를 표현하기 위해 사용된다. 비교말뭉치를 사용하는 기존 연구와는 다르게 우리는 두 종류의 한국어-영어 그리고 영어-스페인어 병렬말뭉치를 사용하며 중간언어로 표현되어진 원시언어 문맥벡터와 대상언어 문맥벡터의 유사도 비교에 정보검색 기법을 사용한다.

기존 연구에서는 큰 용량의 초기사전을 사용해 문맥벡터들을 번역하는 과정이 필요했다. 그러나 본 논문에서는 문맥벡터들을 번역하는 과정은 더 이상 필요하지 않다. 게다가 공개된 단어 정렬 도구를 사용하여 문맥벡터들을 생성했다. 제안한 방법의 전체적인 구조는 그림 1과 같다. 그리고 제안한 방법은 크게 세 단계로 설명 할 수 있다.

(1) 원시언어와 중간언어 단어 정렬

두 종류의 독립적인 병렬말뭉치(한국어-영어, 영어-스페인어)에서 원시언어(예. KR)와 대상언어(예. ES) 각각의 단어들에 대해 원시언어 문맥벡터와 대상언어 문맥벡터를 각각 생성한다. 여기서 문맥벡터의 요소가 되는 모든 단어(영어)들은 단어 정렬 도구인 Anymalign에 의해 가중치가 결정된다. Anymalign이 제공하는 정보의 예는 표1과 같다. $T(t|s)$는 원시단어가 주어졌을 때 대상단어로 번역될 확률이며 $T(s|t)$는 대상단어가 주어졌을 때 원시단어로 번역될 확률을 의미한다. 본 논문에서는 문맥벡터의 요소로써 Anymalign이 제공하는 모든 단어 정렬 정보를 사용하지 않고, 양방향 번역확률 $T(t|s)$, $T(s|t) \geq x$ 특정 임계값($0 \leq x \leq 1$) 이상이고, 양방향 번역확률의 차 $|T(t|s) - T(s|t)| \geq 0.5$일 경우 올바른 단어 정렬 정보라고 판단하고 $\frac{T(t|s)+T(s|t)}{2}$를 가중치로 사용하였다. 이러한 이유는 문맥벡터의 잡음을 줄이기 위해서이다. 아래 표 1의 예를 보면 올바른 단어 정렬 정보는 양방향 번역확률의 차이가 크지 않고 번역확률 또한 높은 것을 확인할 수 있다. 그러나 잘못된 정렬 정보의 경우 양방향 번역확률의 차이가 크고 어느 한쪽의 번역확률이 낮거나 양방향 번역확률이 모두 낮은 것을 볼 수 있다.

표 1. Anymalign을 통한 단어 정렬 결과

원시단어	대상단어	$T(t\|s)$	$T(s\|t)$	빈도수
경찰	police	0.890421	0.877882	389447
경찰	scenario	1.000000	0.000167	8
정부	balance	0.000004	0.000506	11

(2) 유사도 계산

모든 원시언어 문맥벡터와 모든 대상언어 문맥벡터들 사이의 유사도를 계산한다. 여기서 유사도 비교에 코사인 유사도(cosine similarity)를 사용한다.

(3) 번역 후보 순위 결정

계산된 코사인 유사도의 값에 따라 번역 후보 중 상위 k

개의 단어 쌍을 추출한다.

본 논문에서는 문맥벡터를 생성하기 위해 중간언어(영어)를 공유하는 두 개의 병렬말뭉치를 사용했다. 그 이유는 공개된 한국어-스페인어의 병렬말뭉치가 없었기 때문이다. 그리고 공개 단어 정렬도구인 Anymalign은 문맥벡터의 모든 단어들에 가중치를 결정하는데 사용되었다. Anymalign은 무작위 샘플링과 문자열 차이를 기반으로 단어들을 정렬하며, 낮은 빈도수를 가지는 단어들에 대해서 정확한 번역 후보들을 추출하는데 높은 정확도를 보였다.

앞서 언급했듯이 기존 연구에서는 원시언어와 대상언어의 문맥벡터들을 초기사전을 이용하여 번역하는 과정이 필요했으나 본 논문이 제안하는 방법에서는 더 이상 필요하지 않다. 또한 양방향 번역정보를 활용하고 이전 연구[12]에서 보인 불용어들이 번역후보로 추출되는 문제점을 품사태깅 후 제거함으로서 해결하였다. 그리고 문맥벡터들이 생성되고 나면 모든 원시언어 문맥벡터들과 모든 대상언어 문맥벡터들 사이 코사인 유사도를 통해 두 문맥벡터 간의 유사도가 측정된다. 마지막으로 유사도가 큰 순으로 정렬하고 상위 k개의 단어 쌍을 번역후보로 추출한다.

3. 실험 및 결과

본 논문에서는 두 개의 서로 다른 언어 쌍인 한국어(KR)-스페인어(ES) 그리고 한국어-프랑스어(FR)에서 명사들을 대상으로 임계값 x에 따른 이중언어 사전 추출 실험을 하였고 이전 연구(양방향 번역확률, 양방향 번역확률의 차를 고려하지 않은 방법)와 그 성능을 비교 실험하였다.

3.1. 실험 설정

3.1.1. 병렬 말뭉치

본 논문에서 사용된 병렬말뭉치로는 뉴스기사로부터 수집된 433,151 문장으로 구성된 한국어-영어[13] 병렬말뭉치와 Europarl[14] 병렬말뭉치로부터 무작위로 500,000 문장을 추출한 스페인어-영어 그리고 프랑스어-영어 병렬말뭉치이다. 각 언어에서 하나의 문장이 구성하고 있는 단어의 수는 표 2와 같다. 표 2를 보면 ES-EN과 FR-EN에서 각 언어의 문장 당 단어의 개수는 비슷하나 KR-EN의 경우는 조금 차이가 있다는 것을 볼 수 있다. 그 이유는 영어와는 다르게 한국어의 경우 하나의 단어가 한 개 또는 그 이상의 형태소로 구성되기 때문이다. 만약 형태소의 개수가 아니라 단어의 개수를 세면 한 문장 당 단어의 수는 비슷할 것이다.

표 2. 문장 당 평균 단어의 수

KR-EN		ES-EN		FR-EN	
KR	EN	ES	EN	FR	EN
31	19.2	26.4	25.4	29.7	27.1

3.1.2. 데이터 전처리

각 언어들은 다음과 같은 도구에 의해서 토큰이 분리되었다. 한국어의 경우 창원대학교 형태소분석기 Espresso[15]를 사용하여 토큰을 분리하고 품사 태깅을 하였다. 영어, 스페인어 그리고 프랑스어는 Tree-Tagger[16]를 사용하여 토큰이 분리되고 원형 복원되었다. 또한 모든 언어에서 명사를 제외한 품사들은 문장에서 제거되었다. 그 이유는 앞서 언급했듯이 Anymalign은 단어 정렬 시에 문맥을 고려하는 것이 아니라 무작위 샘플링과 문자열의 차이를 이용하기 때문에 본 논문에서의 이중언어 사전 구축대상인 명사를 제외한 품사들은 단어 정렬 시에 잡음으로 작용할 수 있기 때문이다.

3.1.3. 평가사전의 구축

본 논문이 제안하는 방법의 성능을 평가하기 위해 우리는 네 집합(KR-ES, KR-FR, ES-KR, FR-KR)의 평가사전을 인터넷 사전[1]으로부터 수동으로 구축하였다. 각 평가사전은 단방향으로 구축되어있으며, 즉 한 언어에서 다른 언어로의 번역을 말하며, 말뭉치 내에서 가장 빈도수가 높은 100개의 명사(HIGH)와 가장 빈도수가 낮은 100개의 명사(LOW)들로 구성되어 있다. 표 3는 각 평가사전에서 하나의 명사에 대한 평균 번역 수를 나타낸 것이다. 여기서 평균 번역 수는 그 명사의 모호성 정도를 의미한다고 볼 수 있다.

표 3. 평가사전에서의 평균 번역단어의 수

평가사전	HIGH	LOW
KR-FR	5.79	2.26
KR-ES	7.36	3.12
ES-KR	10.31	5.49
FR-KR	10.42	6.32

3.1.4. 평가 방법

제안한 방법의 이중언어 사전 구축 성능을 평가하기 위해서 정보검색의 평가방법과 유사하게 정확도(accuracy)와 MRR(Mean Reciprocal Rank)[17]을 이용하였다. 정확도는 모든 평가 단어에 대하여 평가기준 상위 k개 이내에 정답이 적어도 한 개 있는 평가 단어들의 수의 조화평균을 구한 것이며, MRR은 모든 평가 단어에 대해 평가기준 상위 k개 이내에서 처음으로 나온 정답순위의 역순을 구하고 그것들의 조화평균을 구한 것이며 번역후보결과의 순위화 성능을 평가하기 위해 이용되었다.

3.2. 결과

본 논문에서는 임계값 x를 변경해 가면서 이중언어 사전 추출 실험을 하였고, 표 5는 각 언어쌍에서 이중언어 사전을

1) dic.naver.com

표 4. 평가사전 HIGH와 LOW에 대한 정확도 결과

HIGH								
	accu.Top 1		*accu.Top 5*		*accu.Top 10*		*accu.Top 20*	
	이전연구	제안방법	이전연구	제안방법	이전연구	제안방법	이전연구	제안방법
KR-ES	0.577	0.683(↑18.37%)	0.718	0.704(↓1.94%)	0.758	0.714(↓5.80%)	0.778	0.724(↓6.94%)
KR-FR	0.487	0.707(↑45.17%)	0.703	0.727(↑3.41%)	0.738	0.727(↓1.49%)	0.763	0.727(↓4.71%)
ES-KR	0.293	0.492(↑67.91%)	0.751	0.743(↓1.06%)	0.804	0.779(↓3.10%)	0.872	0.825(↓5.38%)
FR-KR	0.282	0.459(↑62.76%)	0.736	0.709(↓3.66%)	0.789	0.775(↓1.77%)	0.842	0.790(↓6.17%)

LOW								
	accu.Top 1		*accu.Top 5*		*accu.Top 10*		*accu.Top 20*	
	이전연구	제안방법	이전연구	제안방법	이전연구	제안방법	이전연구	제안방법
KR-ES	0.244	0.324(↑032.78%)	0.449	0.492(↑09.57%)	0.469	0.555(↑18.33%)	0.489	0.575(↑17.58%)
KR-FR	0.164	0.269(↑064.02%)	0.299	0.543(↑81.60%)	0.412	0.624(↑51.45%)	0.463	0.680(↑46.86%)
ES-KR	0.030	0.297(↑990.00%)	0.299	0.459(↑53.51%)	0.391	0.459(↑17.39%)	0.433	0.459(↑06.00%)
FR-KR	0.062	0.251(↑404.83%)	0.260	0.415(↑59.61%)	0.395	0.415(↑05.06%)	0.489	0.415(↓15.13%)

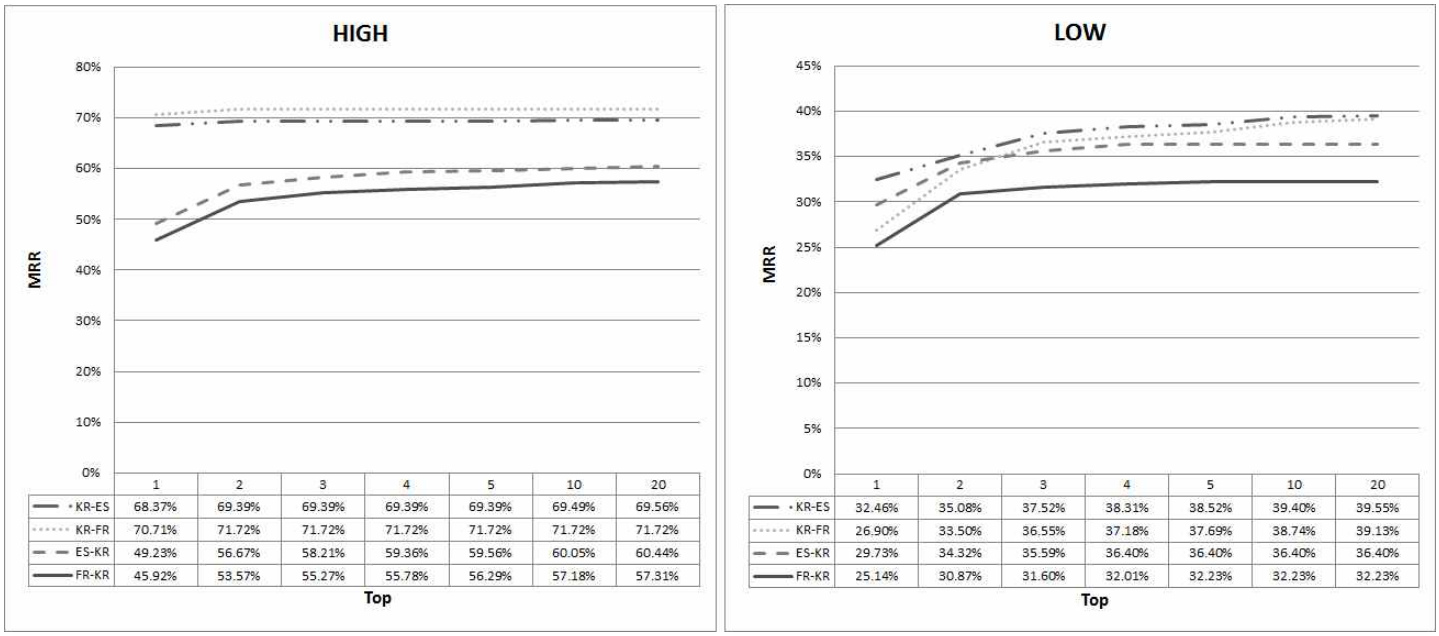

그림 2. 평가사전 HIGH와 LOW에 대한 MRR 결과

추출할 때 가장 좋은 성능을 보여준 임계값 x를 보여준다.

표 5. 각 평가사전에서 사용한 임계값 x

평가사전	HIGH	LOW
KR-FR	0.1	0.00005
KR-ES	0.1	0.00005
ES-KR	0.01	0.05
FR-KR	0.01	0.05

평가단어 HIGH와 LOW에 대한 정확도는 표 4와 같다. 양방향 번역확률을 고려한 결과 평가사전 HIGH에서는 최상위에서 최대 67.91%의 성능 향상을 얻을 수 있었다. 그러나 상위 5위를 넘어서면서 상위 20위에서 최대 6.94%의 성능 하락이 있었다. 그 이유는 상위 5위 이내의 성능을 향상시키기 위해 임계값 x를 높게 설정하면서 문맥벡터의 요소가 되는 단어 정렬정보들이 많이 소실되었기 때문이다. 같은 실험으로 평가사전 LOW에서는 최상위에서 최대 990%의 성능 향상이 있었고 상위 10위에서 최소 5.06%의 성능 향상을 얻을 수 있었다.

표에서도 볼 수 있듯이 최상위에서 KR에서 ES와 FR로 번역되는 정확도에 비해 그 역방향인 ES-KR과 FR-KR의 번역 정확도가 낮은 것을 볼 수 있다. 그 이유를 분석해 보면 한국어의 경우 다른 언어들과는 다르게 형태소 분석과 정이 필요하고 또한 앞서 표 2에서 보았듯이 한국어의 문장 당 단어의 수가 영어의 문장 당 단어 수보다 많다는 것에 있다. 즉 다른 언어에서 한국어로의 번역이 한국어에서 다른 언어로의 번역보다 더 모호하다는 것이다.

표 6는 'exercice(운동)'의 상위 5위의 번역후보를 보여주는 예이다. 이전 연구에서의 문제점으로 불용어들이 최종 번역후보로 추출되는 문제점들은 앞서 언급했듯이 데이터 전처리 과정을 통해 해결되었다. 그러나 최상위와 3위를 보면 형태소분석과 품사태깅의 오류로 인한 잘못된

번역 후보들이 추출되는 문제점들을 발견할 수 있었다.

표 6. 'exercice(훈련)'의 상위 5위 이내의 번역 후보의 예

Rank	원시언어	대상언어	유사도
1	exercice	훈련에	0.350
2	exercice	운동	0.281
3	exercice	중요하다	0.265
4	exercice	합동	0.256
5	exercice	군사훈련	0.145

평가단어 HIGH와 LOW에 대한 MRR은 그림 2와 같다. 그림에서 볼 수 있듯이 HIGH에서의 MRR은 상위 2위까지 급격히 증가하다가 점차 증가하는 것을 볼 수 있으며 LOW에서의 MRR은 상위 3위까지 급격히 증가하다가 그 이후로는 점차 증가하는 것을 확인 할 수 있다. 이는 본 논문이 제안하는 방법을 사용했을 때 HIGH에서는 상위 2위 이내에서 대부분의 올바른 번역후보들이 추출된다는 것을 의미하며 LOW에서는 상위 3위 이내에서 대부분의 올바른 번역후보들이 추출된다는 것을 의미한다.

4. 결론

본 논문에서는 정보 검색 기법을 기반으로 중간언어를 사용하여 이중언어 사전을 병렬말뭉치로부터 자동 구축하는 방법을 제안하였다. 우리는 이전 연구에서 보인 문제점을 데이터 전처리를 통해 해결하였고 양방향 번역확률 정보를 고려하여 상위 5위 이내의 정확도를 크게 향상시켰다. 향후 연구로는 스페인어와 프랑스어 이외에 다른 언어에 대해서도 이중언어 사전을 자동으로 구축해 보는 것과 복합단어(Multi-word expression)으로 확장해 나가는 연구를 해볼 것이다.

감사의 글

본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041807, 지식학습 기반의 다국어 확장이 용이한 관광/국제행사 통역률 90%급 자동 통번역 소프트웨어 원천 기술 개발]

참고문헌

[1] D. Wu and X. Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA 1994, Columbia, Maryland, USA, October), pages 206-213.

[2] P. Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In Proceedings of the Third Workshop on Very Large Corpora (VLC' 95), pages 173-183.

[3] K. Yu and J. Tsujii. 2009. Bilingual dictionary extraction from Wikipedia. In Proceedings of the 12th Machine Translation Summit (MTS 2009), Ottawa, Ontario, Canada.

[4] A. Ismail and S. Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In Proceedings of the International Conference on Computational Linguistics, pages 481-489.

[5] K. Tanaka and K. Umemura. 1994. Construction of a Bilingual Dictionary Intermediated by a Third Language. In Proceedings of the 15th International Conference on Computational Linguistics (Coling' 94), Kyoto, Japan, August, pages 297-303.

[6] H. Wu and H. Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, pages 856-863.

[7] T. Tsunakawa, N. Okazaki, and J. Tsujii. 2008. Build-ing Bilingual Lexicons Using Lexical Translation Probabilities via Pivot Languages. In Proceedings of the International Conference on Computational Linguistics, Posters Proceedings, pages 18-22.

[8] P. Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In Proceedings of the Parallel Text Processing, pages 1-16.

[9] E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte and H. Dejean. 2004. A geometric view on bilingual lexcion extraction from comparable corpora. In Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, pages 527-534.

[10] A. Hazem and E. Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), pages 288-292.

[11] A. Lardilleux, Y. Lepage, and F. Yvon. 2011. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. International Journal of Advanced Intelligence, 3(2):189-217.

[12] H. Kwon, H. Seo and J. Kim, 2013. Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, August 8, 2013, pages 11-15.

[13] H. Seo, H. Kim, H. Cho, J. Kim and S. Yang, 2006. Automatically constructing English-Korean parallel corpus from web documents. Korea

Information Proceedings Society, 13(2):161-164.
[14] P. Koehn. 2005. EuroParl: A parallel corpus for statistical machine translation. In proceedings of the Conference on the Machine Translation Summit, page 79-86.
[15] J. Hong and J. Cha, 2008. A New Korean Morphological Analyzer using the Eojeol Pattern Dictionary. In Proceedings of the Korea Computer Congress, pages 279-283.
[16] H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, pages 44-49.
[17] E. Voorhees. 1999. The TREC-8 Question Answering Track Report. In 8th Text Retrieval Conference (TREC-8), pages 77-82.