등급 재현율: 이중언어 사전 구축에 대한 평가 방법

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

Rated Recall: Evaluation Method for Constructing Bilingual Lexicons

Hyeong-Won Seo○, Hong-Seok Kwon, Jae-Hoon Kim
Korea Maritime University, Computer Engineering Institute

요 약

이중언어 사전 구축 방법을 평가하는 방법에는 정확률, 재현율, MRR(Mean Reciprocal Rank) 등이 있다. 이들 방법들은 평가 집합에 있는 대역어를 정확하게 찾는 것에 초점을 맞추고 있다. 그러나 어떤 대역어가 얼마나 많이 사용되는지는 전혀 고려하지 않는다. 즉 자주 사용되는 대역어를 빨리 찾을 수 있는 방법이 좋은 방법이라고 말할 수 있다. 이와 같은 문제를 해결하기 위해서 본 논문에서는 이중언어 사전 구축의 새로운 평가 방법인 등급 재현율을 제안한다. 등급 재현율(rated recall)은 대역어가 학습 말뭉치에 나타난 정도를 반영하는 재현율이며, 자주 사용되는 대역어를 얼마나 정확하게 찾는지를 파악할 수 있는 좋은 측도이다. 본 논문에서는 문맥벡터와 중간언어를 이용한 이중언어 사전 구축 시스템의 성능을 평가하고 기존의 방법과 비교 분석하였다.

주제어: 이중언어 사전, 문맥벡터, 중간언어, 등급재현율

1. 서론

이중언어 사전(bilingual lexicon)은 자연언어처리(Natural Language Processing), 기계 번역[1], 다중언어(Multilingual) 정보검색[2] 등의 분야에서 주요한 자원으로 활용되고 있다[3]. 영어와 불어과 같이 널리 사용되는 언어에 대해서는 이중언어 사전의 구축이 그다지 어렵지 않다. 그러나 모든 언어 쌍마다 이중언어 사전을 구축하는 것은 많은 시간과 노력이 소요된다[4-7]. 이런 연구들은 주로 병렬 말뭉치(parallel corpora)나 비교 말뭉치(comparable corpora)를 이용하고 있다. 병렬 말뭉치는 통계기반 기계번역에서 널리 사용되는 단어 정렬(word alignment)을 이용하고 병렬 말뭉치는 문맥벡터 기반 방법을 이용한다. 이 방법들은 초기 사전(seed dictionary)이나 말뭉치의 존재 유무가 매우 중요하다. 그러나 어떤 언어 쌍에 대해서는 초기 사전뿐 아니라 병렬 및 비교 말뭉치조차도 쉽게 구할 수 없다. 예를 들면 한국어(KR)와 스페인어(ES)/불어(FR)로 공개된 이중언어 사전도 없고, 더 나아가 공개된 병렬 혹은 비교 말뭉치도 없다. 이런 환경에 맞서 적은 노력과 시간을 투자하면서도 이중언어 사전을 구축할 수 있는 간단하면서도 효과적인 방법이 제안되었다[8]. 이 방법은 사전을 구축할 때 생길 수 있는 도메인 문제를 완벽하게 해결할 수는 없지만 병렬 말뭉치와 중간언어(pivot language)를 사용함으로써 초기 사전이 필요하지 않다는 점과 구축이 어려운 언어 쌍에서도 충분히 이중언어 사전을 구축하기에 용이하다는 장점이 있다.

한편, 이중언어 사전 구축 방법을 평가하는 방법에는 정확률(accuracy), 재현율(recall), MRR(Mean Reciprocal Rank) 등이 있다. 이들 방법들은 평가 집합에 있는 대역어를 정확하게 찾는 것에 초점을 맞추고 있다. 그러나 어떤 대역어는 얼마나 자주 사용되는지는 고려하지 않는다. 자주 사용되는 대역어를 빨리 찾을 수 있는 방법이 좋은 방법이다. 예를 들면 한국어 '학교'는 영어 'school', 'college', 'institution'로 번역될 수 있다. 그러나 많은 경우에는 'school'로 번역된다. 이처럼 이중언어 사전 구축 시스템에서도 번역되는 대역어를 빨리 찾을 수 있는 시스템이 좋은 시스템이다. 이와 같은 문제를 해결하기 위해서 본 논문에서는 이중언어 사전 구축의 새로운 평가 방법인 등급 재현율을 제안한다. 등급 재현율(rated recall)은 대역어가 학습 말뭉치에 나타난 정도를 반영하는 재현율이며, 자주 사용되는 대역어를 얼마나 정확하게 찾는지를 파악할 수 있는 좋은 측도이다.

본 논문의 구성은 다음과 같다. 2장에서 문맥벡터에 기반을 둔 이전 연구에 대하여 간략히 기술하고, 3장에서는 새롭게 제안하는 평가 방법인 등급재현율에 대하여 기술한다. 4장에서는 실험에 대한 내용을 기술하고 마지막으로 5장에서 결론을 짓는다.

2. 관련 연구

2.1 이중언어 사전 자동구축

이중언어 사전을 효과적으로 구축하기 위해 다양한 기존 연구들이 진행되어 왔다[9]. 이런 연구들에는 중간언어를 이용하는 방법[4][5][9], 병렬 말뭉치를 이용한 방법[10][11], 그리고 비교 말뭉치를 이용한 방법[6][7][12] 등이 있다. 병렬 말뭉치를 이용하여 이중언어 사전을 만들었을 때 충분히 좋은 성능을 보였다는 연구 결과가 있다[13]. 하지만 영어를 제외한 언어에 대해서는 병렬 말뭉치가 공개된 것이 드물고, 만약 구축해서 효과적으로 쓰기 위해서는 상당히 많은 양의 말뭉치가

필요하다는 한계점이 있다.

이런 문제들을 해결하기 위해서 중간언어를 활용하는 연구들이 있었다[3]. 중간언어를 이용하면 언어 자원이 많지 않은 언어 사이에서도 보다 쉽게 병렬 말뭉치를 얻을 수 있다는 장점이 있다. 이에 반해, 비교 말뭉치는 특정 도메인에 대하여 언어가 서로 다르지만 문맥이 비슷한 문서들(일반적으로 집필 날짜가 겹치는 뉴스 문서)을 모아서 구축하였기 때문에 영어처럼 언어 자원이 풍부한 언어가 아닌 쌍에 대해서도 병렬 말뭉치보다 비교적 쉽게 구축할 수 있다는 장점이 있다. 또한 병렬 말뭉치보다 도메인 문제를 어느 정도 해결할 수 있다는 장점도 있다. 하지만 기존에 비교 말뭉치를 사용하여 이중언어 사전을 구축한 연구[14]를 보면 이 방법은 초기 사전이 필요하다는 것을 알 수 있다. 초기 사전은 원시 언어(SL: Source Language)나 대상 언어(TL: Target Language) 문맥벡터를 다른 언어로 번역할 때 사용하며, 그 단어의 양이 많을수록 좋지만 처음에 적은 양이여도 무방하다는 특징을 가지고 있다.

2.2 기본 문맥벡터 방법

본 절에서는 대표적인 이중언어 사전 구축 방법인 '문맥벡터 기반 방법'[14]을 살펴본다. Fung은 이중언어 사전을 구축하기 위해서 문맥벡터를 만들었고 이것의 대략적인 과정을 기술하면 다음과 같다. 먼저 원시 언어와 대상 언어의 비교 말뭉치에서 모든 단어를 대상으로 문맥벡터를 만든다. 이 때, 두 단어의 연관도를 측정하기 위해서 Chi-square Test[18]와 같은 연관성 측도(association measure)를 이용한다. 그 다음, 초기 사전을 이용하여 한 쪽 언어의 벡터를 다른 쪽 언어에 맞게 번역한다. 그러면 벡터 공간의 차원이 같아지기 때문에 원시 언어와 대상 언어의 벡터들을 서로 비교할 수 있게 된다. 원시 언어의 한 단어에 대한 벡터와 대상 언어의 모든 단어에 대한 벡터들을 그들 간의 유사도를 통해 서로 비교한 후, 그 유사도에 따라 정렬하고 상위 몇 개의 후보를 추출한다.

이 방법을 활용한 몇 가지 다른 변형된 방법들을 기술하면 다음과 같다.

<문맥 범위 조절>
- 3문장[15]
- 3단어 25개[16]

<초기 사전의 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 혹은 Jaccard indexes[12][15]

2.3 중간언어 기반 문맥벡터 방법

2.2 절에서 기술한 방법에 대하여 정리하면 다음과 같다. 이전의 문맥벡터 기반 연구는 비교 말뭉치와 초기 사전을 사용한다. 초기 사전은 전체 성능에 영향을 주기 때문에 매우 중요한 요소이며, 이 사전이 문서에 포함된 단어들을 얼마나 포함하고 있는지도 중요한 문제가 될 수 있다. 또한 언어가 바뀔 때마다 사전도 구축해야 한다는 문제점이 있다.

이에 반해, 중간언어 기반의 문맥벡터 방법은 우리가 대상으로 하는 언어에 대해 비교 말뭉치가 아닌 병렬 말뭉치를 사용한다. 이 방법을 이용하면 원시 언어 문맥벡터를 대상 언어에 맞게 번역할 필요가 없기 때문에 초기 사전을 일일이 구축하지 않아도 된다는 장점이 있다.

이런 중간언어를 이용한 문맥벡터 방법을 좀 더 자세하게 묘사하면 그림 1과 같이 나타낼 수 있다. (1) 먼저

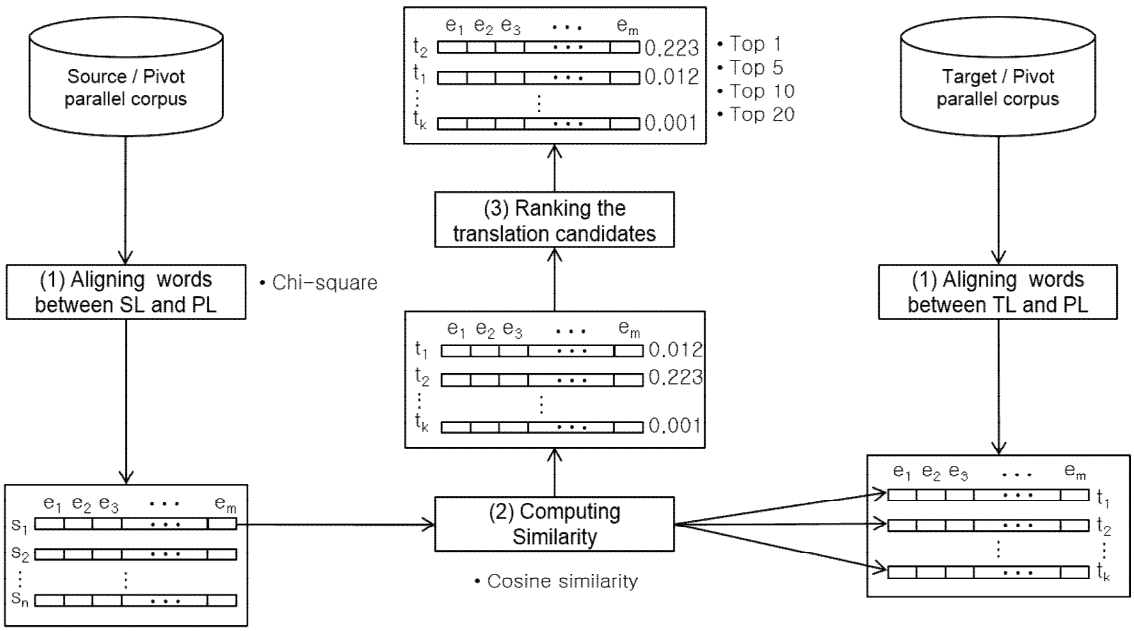


그림 1. 중간언어를 활용한 문맥벡터 방법 흐름도

각각의 병렬 말뭉치들에 포함된 모든 단어들에 대하여 연관성을 측정한다. 여기서 병렬 말뭉치는 SL와 중간언어(PL: pivot language) 그리고 TL와 PL로 구성된 2개의 말뭉치를 의미한다. 본 논문에서는 기호나 불용어를 제외한 나머지 단어를 대상으로 하며 명사, 동사, 형용사, 부사를 제외한 품사의 단어는 제외시킨다. 그 후, 남아있는 단어들 사이에 Chi-square test를 이용하여 두 단어(SL/TL과 PL의 단어)가 서로 얼마나 연관성이 있는지를 측정한다. 여기서 사용된 단어 빈도수는 DF(Document frequency)를 사용하고, 문장을 문서로 간주하여 단어를 포함하고 있는 문장의 수를 센다. (2) 이렇게 만들어진 문맥벡터들(SL-PL, PL-TL) 사이에 Cosine measure를 이용하여 벡터들 간의 거리 유사도를 계산한다. (3) 그 후 유사도가 높은 순으로 정렬하여 각 SL 단어마다 상위 k개의 TL 번역 후보들을 추출한다.

이 방법을 이용하면 초기 사전과 같은 외부 자원 없이 병렬 말뭉치만으로도 SL과 TL 사이에 번역 후보들을 쉽게 추출할 수 있다.

3. 등급 재현율

일반적으로 이중언어 사전 구축 방법을 평가하는 방법에는 정확률, 재현율, MRR 등이 사용된다. 이들 방법들은 평가 집합에 있는 대역어를 정확하게 찾는 것에 초점을 맞추고 있다. 그러나 어떤 대역어는 얼마나 자주 사용되는지는 고려하지 않는다. 하지만 자주 사용되는 대역어를 빨리 찾는 방법이 좋은 방법이다. 예를 들면 한국어 '학교'는 영어 'school', 'college', 'institution'로 번역될 수 있다. 그러나 많은 경우에는 'school'로 번역된다. 이처럼 이중언어 사전 구축 시스템에서도 번역되는 대역어를 빨리 찾는 시스템이 좋은 시스템이다. 이와 같은 문제를 해결하기 위해서, 본 논문에서는 등급 재현율을 제안한다. 등급 재현율은 대역어가 학습 말뭉치에 나타난 정도를 반영하는 재현율이며, 자주 사용되는 대역어를 얼마나 정확하게 찾는지를 파악할 수 있는 좋은 측도이다.

먼저 재현율에 대해서 살펴보자. 재현율은 정답 단어를 시스템이 얼마나 찾았는가를 나타내며, 상위 n번째 후보 대역어의 재현율 R_n은 식 (1)과 같이 정의된다.

$$R_n = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C_i|}\sum_{j=1}^{n}c_{ij}, \quad c_{ij} = \begin{cases}1, \text{ if } t_{ij} \in C_i \\ 0, \text{ otherwise}\end{cases} \qquad (1)$$

여기서, N은 원시 단어의 총 수이며, C_i는 i번째 원시 단어의 정답 단어(대역어 집합)이고, $|C_i|$는 C_i의 개수이다. c_{ij}는 크로네커 델타 함수(Kronecker delta function)로서 i번째 원시 단어 s_i에 대한 j번째 시스템 결과 t_{ij}가 정답에 포함되면 1이 되고, 그렇지 않으면 0이 된다. 즉, 정답 단어를 시스템이 얼마나 정확하게 찾는가를 나타낸다.

등급재현율은 RR_n은 식(1)를 약간 수정하여 식 (2)와 같이 정의한다.

$$RR_n = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}c_{ij}\,r(t_{ij}), \quad c_{ij} = \begin{cases}1, \text{ if } t_{ij} \in C_i \\ 0, \text{ otherwise}\end{cases} \qquad (2)$$

여기서, $r(t_{ij})$는 i번째 원시단어 s_i에 대한 후보 대역어 t_{ij}가 학습 말뭉치에 출현한 비율을 의미하며, $\sum_{t \in C_i} r(t)$은 1이다. 예를 들어, 스페인어 단어 decisión에 대한 정답 사전이 표 1과 같고, 시스템이 낸 결과가 표 2와 같다면 각 순위에 따른 등급 재현율의 값은 표 3과 같다.

정답 단어	빈도수	$r(t)$
결정	6,007	0.752
결심	173	0.022
결의	369	0.046
결단	130	0.016
결단력	10	0.001
재정	880	0.110
판정	414	0.052
합계	**7,983**	**1.000**

표 1. 스페인어 decisión의 정답 단어

순위	번역 후보	$r(t)$
1	결정	0.752
2	통합력	
3	결단	0.016
4	여부	
5	최종판단	
6	의사결정	
7	판정	0.052
8	판단	
9	결정권한	
10	관망자세	
11	확정	
12	독자위성	
13	유격대	
14	앙케트	
15	결심	0.022
16	사업확장계획	
17	개인신용대출	
18	판결	
19	재정능력	
20	사항	

표 2. decisión에 대한 시스템 결과

순위	등급 재현율(RR_n)	재현율(R_n)
1	0.752	1/7 = 0.143
3	0.752 + 0.016 = 0.768	2/7 = 0.286
10	0.768 + 0.052 = 0.820	3/7 = 0.429
20	0.820 + 0.022 = 0.842	4/7 = 0.571

표 3. 각 랭크에서의 등급재현율과 재현율

표 1에서 알 수 있듯이 실제 문서에서 얼마나 자주 출현되느냐에 따라 각 단어의 비율이 결정되고, 이 비율이 실제 등급 재현율을 계산할 때 더해지게 된다. 표 2는 원시 언어인 스페인어 decisión에 대한 시스템 도출 결과를 나열한 것이다. 그리고 정답 사전으로부터 각 단어의 비율을 나타낸다. 비율이 표시되어 있지 않은 단어는 정답 사전에 없는 것으로써 대부분 합성명사인 것을 알 수 있다. 표 3은 실제 등급재현율을 계산하는 과정을 표로 나타낸 것이고 세 번째 줄에서 빈 칸은 등급 재현율이 0.0이다. 각 순위에서의 등급재현율은 상위부터 해당 순위까지 포함된 단어 중 정답 사전에 포함됨과 동시에 실제로 문서 안에 얼마나 많이 출현하였는지를 고려한 재현율이 계산되게 된다. 여기서 의미하는 비율은 하나의 원시 언어를 기준으로 하는 군집으로써의 비율이므로 일반적인 재현율 계산 때의 N_i로 나누는 작업은 생략하게 된다.

4. 실험 및 결과

4.1 실험 환경

4.1.1 데이터

본 논문에서는 실험을 위해 한국어(KR)-영어(EN), 스페인어(ES)-EN, 불어(FR)-EN의 병렬 말뭉치를 사용하였다. KR-EN은 433,151개의 문장으로 구성되었고 Seo et al.[19]의 연구에서 사용된 말뭉치를 보완하여 만든 뉴스 기사다. 이 말뭉치의 문장 당 평균 단어(한글의 경우에는 형태소)의 수는 각각 42.46(KR), 36.02(EN)이다. 반면에 ES-EN과 FR-EN 병렬 말뭉치는 Eurorarl(European parliament proceedings) 병렬 말뭉치이며 각각 약 160만, 200만 문장을 포함하고 있다. 이들의 문장 당 평균 단어 수는 각각 29.40(ES-EN에서 추출된 ES), 28.65(ES-EN에서 추출된 EN), 31.17(FR-EN에서 추출된 FR), 28.68(FR-EN에서 추출된 EN)이다.

평가를 위한 정답 사전은 다음과 같이 구축하였다. 사전을 구성하는 단어를 정하기 위해 각 병렬 말뭉치로부터 빈도수가 높은 순서대로 정렬한 후 가장 빈도수가 높은 100개의 명사(High)와 빈도수가 낮은 100개의 명사(Low)를 웹 사전을 참조하여 사람이 직접 구축하였다. 최종적으로 구성된 정답 단어 당 평균 번역 단어의 수는 각각 11.41(FR-KR), 10.3(ES-KR), 5.79(KR-FR), 7.36(KR-ES)개이다. 불어와 스페인어의 한글 번역 단어의 개수가 그 반대인 경우보다 상대적으로 많은 것을 확인할 수 있다.

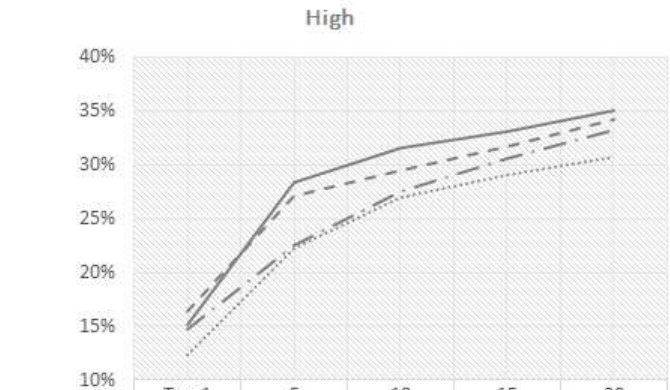
4.1.2 전처리

KR의 경우 한나눔 태거(KAIST Tagger)[1)][20]를 이용하여 품사를 부착하는 전처리만 수행하였다. EN, ES, FR의 경우에는 Tree Tagger[2)][21]를 이용하여 토큰 분리, 원형 분리(lemmatization)를 한 후 품사를 부착하였다. 품사가 모두 부착된 상태에서 각각의 언어에 맞는 불용어(KR은 제외)와 특정 품사를 제외하였다. KR 말뭉치에서는 총 69개의 품사 중 보통명사, 고유명사, 용언 그리고 수식언을 제외한 나머지 51개의 품사에 해당하는 단어들은 모두 배제하였다. 나머지 언어에 대해서도 같은 작업을 하여 EN은 61개 품사 중 19개, ES는 72개 중 33개, 마지막으로 FR은 36개 중 18개를 배제하였다.

전처리 후에 남은 단어의 타입 수는 각각 KR-EN 67,210(KR의 경우에는 형태소)/41,719(EN), ES-EN 12,926(ES)/28,764(EN), FR-EN 47,220(FR)/51,245(EN)이다. 이 중에서 같은 성격의 말뭉치임에도 불구하고 두 영어 문서(ES-EN과 FR-EN)의 타입 수(각각 28,764와 51,245)가 차이가 나는 이유는 실제 FR-EN의 영어 문서에 다수의 불어 문장이 포함되어 있기 때문이다.

4.2 실험 결과

본 논문에서 제안한 방법으로 실험한 결과는 다음과 같다. 그림 2와 3은 연관성 측도와 유사도 측정 방법을 각각 Chi-square test와 Cosine measure로 정해놓고, 각 병렬 말뭉치로부터 만든 문맥벡터 간에 유사도 계산 결과를 등급 재현율로 나타낸 것이다. 그림 2는 빈도수가 높은 단어(High), 그림 3은 빈도수가 낮은 단어(Low)에 대한 결과이다.



	Top 1	5	10	15	20
KR→FR	0.1228	0.2222	0.2700	0.2908	0.3074
FR→KR	0.1638	0.2710	0.2948	0.3171	0.3415
KR→ES	0.1473	0.2255	0.2754	0.3053	0.3318
ES→KR	0.1518	0.2840	0.3157	0.3317	0.3512

그림 2. 빈도수 높은 단어(High)에 대한 등급재현율

그림 2와 3에서 볼 수 있듯이 KR 번역 후보를 찾는 경우, 즉 대상 언어가 KR인 경우는 High에서 높은 결과를 보였고, 그 반대인 경우(원시 언어가 KR)에는 Low에서 높은 결과를 보였다는 것을 알 수 있다. 4.1.1절에서 기

1) http://kldp.net/projects/hannanum
2) http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

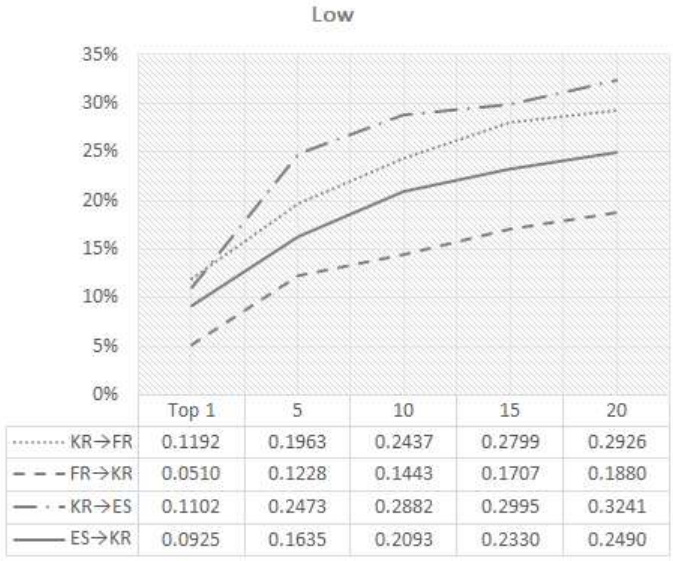


그림 3. 빈도수 낮은 단어(Low)에 대한 등급재현율

술된 정답 사전은 대상 언어가 KR일 경우가 아닌 경우에 비해 상대적으로 많은 정답 단어를 포함하고 있다. 이것으로 볼 때, 여러 뜻을 가지는 단어일수록 문서에 많이 포함된다는 사실을 알 수 있다. 또한 상위 1위부터 5위까지 그래프가 급격하게 기울었다가 이후로 갈수록 점점 기울기가 낮아지는 것을 볼 수 있다. 이런 특징은 Low보다 High에서 좀 더 두드러지지만 대부분의 중요한 단어들(빈도수가 높거나 정확한 번역 후보)은 중하위보다 주로 상위에 포진되어 있다는 것을 의미한다.

그림 4와 5는 모두 일반적인 재현율과 등급재현율로 시스템을 평가한 것으로써 각각 원시 언어 혹은 대상 언어가 KR↔FR, KR↔ES인 것에 대한 결과를 통합하여 평균낸 결과이다. 그림에 나타난 결과들을 보면 심한 경우(KR-ES, 상위 20위)에 15% 정도의 차이가 난다는 것을 알 수 있다. 이는 일반 재현율로 봤을 때 정답 단어의 20% 정도에 가까운 단어들을 시스템이 도출해냈지만, 사실 이 단어들은 사전 안에서 35% 정도를 차지하고 있다는 사실을 의미한다. 또한 상위 1위에서 재현율과 등급재현율의 성능 차이를 살펴보면, KR-ES와 KR-FR 모두의 경우에 High는 대략 10%, Low는 대략 5% 정도의 차이를 확인할 수 있다. 즉, 사전에 있는 정답 단어 중 5% 혹은 10% 정도만을 시스템이 도출해내었다고 생각할 수 있지만 이 단어들이 실제 문서에서 차지하는 중요도는 그 이상이라는 점을 의미한다. 오히려 문서에 별로 나오지 않은 단어를 시스템이 도출해낸 것보다는 문서에서 빈번하게 등장하는 단어를 시스템이 도출해내는 것이 훨씬 중요하기 때문이다. 따라서 이런 경향으로 봤을 때, 본 논문에서 제안하는 평가 방법이 큰 의미가 있다.

5. 결론

본 논문은 기존에 문맥벡터를 이용하여 이중언어 사전을 자동으로 구축하는 방법이 얼마나 효과적인지 새롭게 평가하였다. 일반적인 재현율은 단순히 얼마나 많은 단

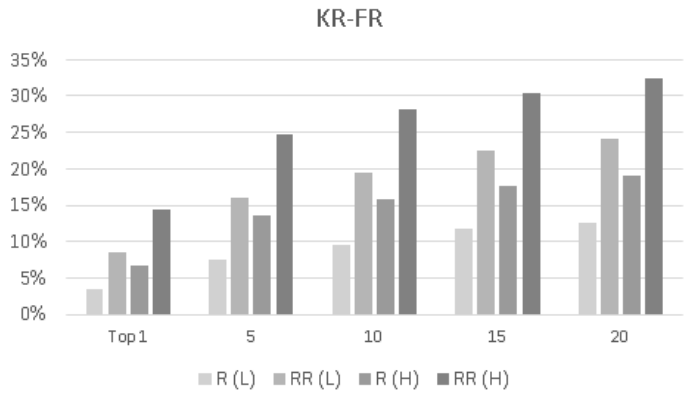


그림 4. 일반 재현율과 등급재현율의 비교1 (KR-FR). L은 Low, H는 High, R은 Recall, RR은 Rated Recall을 의미한다.

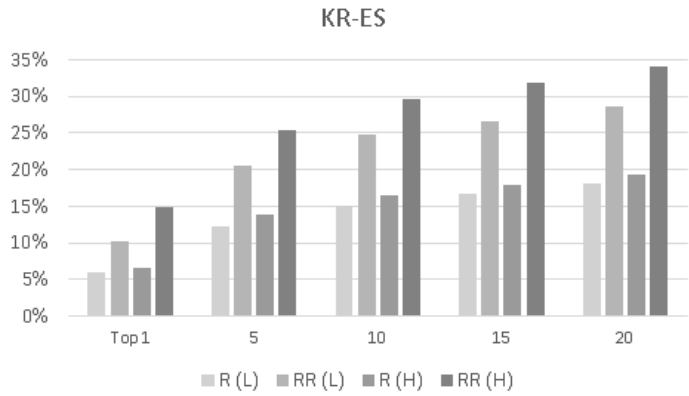


그림 5. 일반 재현율과 등급재현율의 비교2 (KR-ES).

어를 실제 시스템이 결과로 도출해내는지에 대한 평가 방법이지만 모든 단어들에 대하여 똑같이 비율을 지정하여 실제 얼마나 중요한 단어들을 찾아냈는지 판단하기 어렵다. 하지만 본 논문에서 제안한 등급 재현율을 이용하여 평가해보면 시스템이 도출해낸 번역 후보 단어들이 실제 문서에서 얼마나 큰 영향을 끼치는지를 알 수 있기 때문에 시스템이 얼마나 효과적인지 파악할 수 있다는 장점이 있다.

향후 연구로는 스페인어와 불어 이외의 언어에 대하여도 이중언어 사전을 구축해보는 것과 다중단어(multi-word expression)에 대한 연구도 해볼 수 있을 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

[18] 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.

[19] 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 theWeb" In Proc. of the ACM SIGIR pp. 74-81.

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

[21] 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.

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

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

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

[25] 서형원, 권홍석, 김재훈 2013 "이중언어 병렬말뭉치와 중간언어를 활용한 이중언어 사전 자동 구축" 한국정보처리학회 춘계학술발표대회 논문집. 제20권. 제1호. pp. 307-310.

[26] 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.

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

[28] 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.

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

[30] 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.

[31] 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.

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

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

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

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

[36] 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.

[37] 박상원, 최동현, 김은경, 최기선 2010 "플러그인 컴포넌트 기반의 한국어 형태소 분석기" 한글 및 한국어 정보처리 학술대회 (HCLT) Poster. pp. 197-201.

[38] H. Schmid 1995 "Improvements in Part-of-Speech Tagging with an Application to German" In Proc. of the ACL SIGDAT-Workshop. pp. 47-50.