이질적인 언어 자원의 순차적 매칭을 이용한 문장 유사도 계산 기반의 위키피디아 한국어-영어 병렬 문장 추출 방법

천주룡○, 고영중
동아대학교
balendia@gmail.com, youngjoong.ko@gmail.com

Extracting Korean-English Parallel Sentences based on Measure of Sentences Similarity Using Sequential Matching of Heterogeneous Language Resources

Juryong Cheon○, Youngjoong Ko
DongA University, Computer Engineering

요 약

본 논문은 위키피디아로부터 한국어-영어 간 병렬 문장을 추출하기 위해 이질적 언어 자원의 순차적 매칭을 적용한 유사도 계산 방법을 제안한다. 선행 연구에서는 병렬 문장 추출을 위해 언어 자원별로 유사도를 계산하여 선형 결합하였고, 토픽모델을 이용해 추정한 단어의 토픽 분포를 유사도 계산에 추가로 이용함으로써 병렬 문장 추출 성능을 향상시켰다. 하지만, 이는 언어 자원들이 독립적으로 사용되어 각 언어 자원이 가지는 오류가 문장 간 유사도 계산에 반영되는 문제와 관련이 적은 단어 간의 분포가 유사도 계산에 반영되는 문제가 있다. 본 논문에서는 이질적인 언어 자원들을 이용해 순차적으로 단어를 매칭함으로써 언어 자원들의 독립적인 사용으로 각 자원의 오류가 유사도에 반영되는 문제를 해결하였고, 관련이 높은 단어의 분포만을 유사도 계산에 이용함으로써 관련이 적은 단어의 분포가 반영되는 문제를 해결하였다. 실험을 통해, 언어 자원들을 이용해 순차적으로 매칭한 유사도 계산 방법은 선행 연구에 비해 F1-score 48.4%에서 51.3%로 향상된 성능을 보였고, 관련이 높은 단어의 분포만을 유사도 계산에 이용한 방법은 약 10%에서 34.1%로 향상된 성능을 얻었다. 마지막으로, 제안한 유사도 방법들을 결합함으로써 선행연구의 51.6%에서 2.7%가 향상된 54.3%의 성능을 얻었다.

주제어: 병렬 문장, 위키피디아, 비교 말뭉치, 토픽 모델

1. 서론

웹에는 영어, 한국어 등 다국어로 작성된 방대한 양의 정보가 존재한다. 최근 웹 환경에서 정보의 양은 끊임없이 늘어나고 있고 국가 간 경계가 허물어짐에 따라 교차 언어 정보 검색의 연구가 활발하게 진행되고 있다. 이중 언어 혹은 다중 언어를 다루는 교차 언어 정보검색과 같은 분야는 질이 좋고 양이 풍부한 병렬 말뭉치가 필요하다.

병렬 문장으로 구성된 병렬 말뭉치는 언어 번역 및 분석에서 필수적인 원천 자원으로 사용된다. 그러나 병렬 말뭉치를 구축하는 작업은 시간과 비용이 많이 소요되는 작업이다. 효과적으로 병렬 말뭉치를 구축하기 위해서 위키피디아와 같은 언어 자원의 비교 말뭉치에서 병렬 문장만을 자동으로 식별하고 추출하기 위한 연구가 많이 이루어지고 있다[1]. 이러한 비교 말뭉치를 기반으로 자동 구축된 양질의 병렬 말뭉치는 기계 번역, 교차 언어 개체명 인식, 교차 언어 정보 검색 등과 같은 자연어 처리 연구에 많은 도움이 되고 있다.

본 논문에서는 언어 자원들의 순차적 매칭을 적용하여 위키피디아로부터 한국어-영어 간 양질의 병렬 문장들을 추출하는 유사도 계산 방법을 제안한다. 이를 위해, 한국어-영어 간 위키피디아 문서들 중 인터링크(Interlink)로 이루어진 문서 쌍을 비교 말뭉치로 이용하며, 문서 내에 존재하는 문장들을 유사도 계산을 적용할 병렬 후보 문장으로 활용한다.

본 논문은 양질의 병렬 문장 추출을 위해 선행 연구[2]에서 개선된 유사도 계산 방법을 제안한다. 제안하는 유사도 계산 방법은 크게 두 가지이며, 첫 번째는 사전과 같은 언어 자원을 이용하여 유사도를 계산하는 방식이다. 언어 자원은 위키피디아 제목으로 구성된 위키 사전과 다음 온라인 사전, 그리고 숫자 매칭을 이용한다. 본 논문에서는 이러한 이질적인 언어 자원을 이용하여 순차적으로 단어를 매칭하여 유사도를 계산하는 방법을 제안한다. 선행연구에서 언어 자원별로 문장 간의 유사

* 이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2009937)

도를 구하고 선형 결합하여 최종 유사도를 계산하는 방법을 제안하였으나, 이는 언어 자원들이 독립적으로 사용되어 각 언어 자원이 가지는 오류가 유사도 계산에 반영되는 문제가 있다. 이를 해결하기 위해 본 논문은 언어 자원들의 중요도를 고려, 언어 자원 간의 우선순위를 반영하여 순차적으로 단어를 매칭하고 문장 간의 유사도를 한 번에 계산함으로써 문제를 해결하였다.

두 번째는 토픽 모델을 통한 단어의 토픽 분포를 이용하여 유사도를 계산하는 방식이다. 위키피디아 문서 안의 단어들은 토픽 모델 학습에 이용하여 단어들의 토픽 분포를 얻을 수 있으며 이를 벡터로 이용, 서로 유사한 단어를 찾을 수 있다. 본 논문에서는 이러한 단어들의 토픽 분포 중 서로 유사한 단어들의 토픽 분포만을 문장 간 유사도 계산에 적용하는 방법을 제안한다. 선행 연구에서는 문장 안의 모든 단어들의 토픽 분포를 고려함으로써 서로 관련이 적은 단어 간의 분포가 문장 간 유사도 계산에 반영되는 문제가 있었다. 본 논문은 단어 간 토픽 분포가 가장 유사한 단어의 순서대로 고려하여 비교적 서로 관련이 적은 단어 간의 분포를 제외하고 유사도를 계산함으로써 이런 문제를 해결하였다.

이 밖에, 선행 연구에서 제안한 방식에는 기존에 일부 구축된 병렬 말뭉치를 언어 자원으로 이용하여 유사도 계산에 적용한 방식이 있다. 이는 세종 병렬 말뭉치와 같은 병렬 자원으로부터 번역 확률을 추출하고, 문장 간의 유사도를 계산하는 방식이다. 실제로, 세종 병렬 말뭉치와 같이 기존에 구축된 병렬 말뭉치를 자원으로 활용하여 병렬 문장을 추출하는 방법[3]은 높은 정확도의 병렬 문장을 얻을 수 있지만, 현재 한국어를 포함한 병렬 말뭉치는 양이 적고 특정 영역을 중심으로 구성되었거나, 저작권 및 지적 소유권 등의 문제가 많이 존재하고 있다. 따라서, 본 논문은 병렬 말뭉치를 이용한 유사도 계산 방법은 고려하지 않는다.

본 논문은 제안한 유사도 계산 방법을 F1-score를 측정하여 평가하였으며, 언어 자원들을 이용하여 순차적으로 단어를 매칭한 유사도 계산 방법은 선행 연구의 48.4%의 성능보다 2.9%가 향상된 51.3%의 성능을 얻었고, 유사도가 높은 단어 간의 토픽 분포만을 이용한 유사도 계산 방법은 선행 연구보다 24.1%가 향상된 34.1%의 성능을 얻었다. 마지막으로, 제안한 유사도 방법들을 결합함으로써 선행연구보다 2.7%가 향상된 54.3%의 성능을 얻었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 소개하며, 3장에서는 논문에서 제안하는 방법을 상세히 기술한다. 4장에서는 실험 결과를 분석, 비교하며, 5장에서 결론 및 향후 연구 계획을 기술한다.

2. 관련 연구

타 언어 간의 유사한 문장을 추출하는 연구에 앞서 유사한 문서를 찾는 연구가 진행되었다. [4]은 원시 언어 문서에서 주요 키워드를 추출하고 이 키워드를 대상 언어의 단어로 번역한 뒤 대상 언어 문서들 중 번역한 단어들을 많이 포함하고 있는 문서를 비슷한 문서로 판단했다.

유사한 문장을 추출하는 연구에서 [5]은 원시 언어의 문서와 대상 언어의 문서를 각 언어에 맞는 형태소 분석을 한 뒤 비교 문장이 가지는 형태소 규칙을 세우고 후보 문장이 규칙에 매칭이 된다면 유사한 문장이라고 판단하였다. [6]는 정보 검색 방법으로 접근을 했으며, 원시 언어 문서와 대상 언어 문서의 한 문장 당 단어의 빈도수와 전체 문서에서 단어가 출현한 문장의 빈도수를 구한 후, 원시 언어 문장과 대상 언어 문장이 유사할 경우 문장들이 가지는 단어 벡터들이 유사할 것으로 가정하였고, 이는 코사인 유사도와 같은 유사도 측정 방식으로 유사도를 계산할 수 있다.

최근 들어, 광범위한 분야에서 위키피디아를 외부 자원으로 활용하는 연구가 활발히 진행되고 있다. 위키피디아는 누구에게나 공개된 자원이고 시간에 흘러감에 따라 정보를 가진 문서의 수가 기하급수적으로 늘어나며 현실을 즉각적으로 반영한 문서를 쉽게 얻을 수 있는 등 여러 가지 유용한 특징을 가지고 있다. 위키피디아 데이터에서 병렬 문장을 추출하는 연구는 위키피디아 문서 수가 많고 언어적으로 비슷한 특징을 가지고 있는 영어권에서 주로 진행되고 있다.

[7]에서는 위키피디아 문서들이 가지는 링크 자질을 이용했다. 문서 내의 문장들에 링크 표시가 된 단어나 구들은 그 단어와 구들이 문서 제목이 된다. 이와 같이 링크들이 인터-위키를 가지고 있고 원시 언어 문장 내에 링크와 대상 언어 링크들이 연결되어 있다면 두 문장은 같은 내용을 말하는 유사한 문장이라고 판단한다. 구하고자 하는 두 문장 안에 인터-위키 정보를 가지는 링크가 있다면 이 방법은 번역기와 비슷한 성능을 낸다.

최근에는 토픽 모델을 이용하는 연구들이 많이 진행되고 있으며 주로 LDA(Latent Dirichlet Allocation)[8]으로 학습한다. [9]에서는 원시언어 문서와 대상언어 문서를 학습하는 BiLDA(Bilingual LDA)를 사용하여 유사한 문서를 판단했으며 [10]은 위키피디아 문서와 병렬 말뭉치를 이용하여 BiLDA를 학습 시킨 후 교차언어 정보검색에 적용하였다.

3. 제안 방법

이 장에서는 사전과 같은 언어 자원을 이용한 순차적 단어 매칭의 유사도 계산 방법과 토픽 모델을 통해 추정된 단어의 토픽 확률 분포를 이용하여 유사도를 계산하는 방법에 대해 설명한다. 마지막으로 제안한 두 유사도의 결합 방법에 대해 설명한다.

3.1 언어 자원 기반의 순차적 단어 매칭을 이용한 유사도 계산 방법

선행 연구에서는 세 가지의 언어 자원을 이용하여 각각 독립적으로 유사도를 계산하고 선형 결합하였으나, 이는 각 언어 자원이 가지는 오류가 유사도 계산에 모두 반영되는 문제가 일어날 수 있다. 즉, 각 언어 자원에 따라 유사도를 계산하고자 하는 두 문장에 포함되어 있는 중요한 단어들의 중복 계산이 있을 수 있다. 본 논문

에서는 이러한 문제를 해결하기 위해, 유사도 계산에 이용할 언어 자원으로 위키피디아에서 추출한 위키 사전을 이용한 단어 매칭, 숫자 매칭, 다음 온라인 사전을 이용한 단어 매칭을 순차적으로 진행한다. 즉, 언어 자원 간의 우선순위를 주어 순차적으로 단어들을 매칭하고 문장 간의 유사도를 한 번에 계산하는 방법을 제안한다.

첫 번째 언어 자원으로 단어 매칭을 위해 사용하는 위키 사전은 위키피디아 한국어 문서의 제목들과 영어 문서의 제목을 쌍으로 하는 사전을 의미한다. 위키 사전의 내용은 인명, 지역명, 영화 제목 등 많은 개체명과 같은 용어들을 포함하고 있으며, 한국어 단어와 영어 단어를 직접적으로 매칭하는 언어 자원이다. 본 논문에서는 구축된 위키 사전의 단어들이 가장 중요한 단어라고 판단하여, 첫 번째로 위키 사전을 이용해 단어들을 매칭한다. 두 번째는 위키 사전을 이용하여 진행된 단어 매칭에서 매칭되지 않은 단어들을 기준으로 숫자 매칭을 진행한다. 이는 서수나 날짜 등으로 표현된 단어들을 일반 숫자로 바꾸어 숫자를 매칭하는 방식이다. 그 다음으로 진행하는 단어 매칭은 다음 온라인 사전을 세 번째 언어 자원으로 이용한다. 다음 온라인 사전은 다음에서 제공하는 한영 사전과 영한 사전에 동시에 출현하는 단어의 쌍으로 구성되어 있으며, 이는 일반적인 단어의 매칭에 적합하다.

언어 자원을 기반으로 순차적인 단어 매칭 유사도 계산에 사용하는 한국어-영어 문장 간의 유사도 계산 방법은 자카르트(Jaccard) 유사도를 사용한다. 자카르트 유사도는 기본적으로 두 집합(A, B)의 교집합의 크기를 합집합의 크기로 나눈 것이다. 이를 본 논문의 두 문장 간 유사도에 적용하면 식(1)과 같이 정의된다. A는 원시 문장에 포함된 단어 집합, B는 대상 문장에 포함된 단어 집합, 그리고 $J_D(A,B)$는 사전을 이용한 두 문장 간의 유사도를 가리킨다.

M_{11} : A, B의 각 단어 중 교차하는 단어의 수

M_{10} : A에 교차하지 않고 남아있는 단어의 수

M_{01} : B에 교차하지 않고 남아있는 단어의 수

$$J_D(A,B) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}} \quad (1)$$

한글 문장과 영어 문장의 유사도 계산은 다음과 같이 진행된다.

단계1 : 두 문장 안의 링크된 단어들을 모두 추출하며, 이렇게 추출된 링크 단어에서 영어 링크 단어는 모두 위키 사전을 통해 한글로 번역한다. 영어 문장에서 번역된 영어 링크 단어와 한글 문장의 한글 링크 단어 간에 교차하는 개체 수를 식 (1) 의 M_{11}에 추가하고, 두 원본 문장에서 교차되었던 링크 단어는 제거한다.

단계2 : 한글 문장에서 숫자를 추출하고, 영어 문장에서는 서수나 날짜를 숫자 매칭을 이용하여 숫자로 변환, 추출한다. 위와 마찬가지로 추출된 한글 문장의 숫자 단어와 영어 문장의 숫자 단어 간의 교차하는 개체 수를 식 (1)의 M_{11}에 추가하며, 교차된 숫자 단어를 제거한다.

단계3 : 마지막으로 남아 있는 두 문장의 단어들에서 중요한 품사인 명사, 형용사, 동사 단어들을 각각 추출한다. 여기서 영어 단어들을 모두 다음 온라인 사전으로 한국어로 번역을 하고, 두 문장에서 교차하는 개체 수는 식 (1)의 M_{11}에 추가하고, 교차하지 않고 남아있는 단어들은 모두 M_{10}와 M_{01}에 포함하여 수식 (1)을 계산한다.

3.2 단어의 토픽 확률 분포 기반의 유사도 계산 방법

LDA 모델은 Blei[9]에 의해 제안된 대표적인 토픽 모델 중 하나로, 하나의 문서의 여러 토픽이 있으며 각 토픽과 연관된 단어들이 생성된다고 가정한다. 이러한 토픽들은 다항분포로서 정의되며 각 문서는 자신이 가진 토픽의 분포와 각 토픽들이 가진 단어들의 분포에 기반한다. 그리고 각 토픽별로 생성된 단어들이 나열되어 해당 문서가 작성되는 것으로 가정한다.

본 논문에서는 BiLDA 모델을 사용한다. BiLDA는 두 개의 언어로 확장한 LDA 모델이다. 토픽 모델에 사용한 데이터는 위키피디아에서 한글 문서와 인터위키로 대응되는 영어 문서를 사용했으며 이 두 문서는 비교 말뭉치로 활용할 수 있다. 그리고 토픽 모델을 통해 얻어진 단어들의 토픽 확률 분포를 단어 벡터로 이용한다. 즉, 단어 w에 대한 벡터는 $V(w) = (t_1, t_2, t_3, \cdots t_n)$로 표현할 수 있는데, n은 총 토픽 수를 의미하며, t_i는 단어 w에 대한 i번째 토픽 확률을 의미한다. 이리하여 모든 단어를 벡터로 표현할 수 있으며, 이를 이용하여 서로 유사한 단어를 찾을 수 있다.

선행 연구에서는 이러한 단어 간의 토픽 확률 분포를 이용하여 문장 간 유사도를 계산하는 방법으로 두 문장 간 모든 단어들의 토픽 확률 분포를 고려하여 유사도에 적용하였다. 하지만, 이 방법은 모든 원시 언어 단어와 대상 언어 단어 간의 유사도를 반영하기 때문에, 문장 간 서로 유사한 단어들의 중요도를 고려하지 않는 문제가 존재한다. 이런 이유로, 본 논문에서는 단어 간의 토픽 분포가 유사한 단어를 가장 높은 순서대로 고려하여 서로 관련이 적은 단어 간의 분포를 제외한 유사도를 계산함으로써 이런 문제를 해결하였다.

본 논문에서 단어 간 유사도의 계산은 코사인 유사도를 이용하였다. 코사인 유사도를 적용하여 두 문장 간의 유사도 계산은 식 (1)을 변형하였으며, 각 원시 언어 문장과 대상 언어 문장은 모두 중요한 품사인 명사, 형용사, 동사만을 추출한 단어를 사용한다. 그림 1은 단어의 토픽 분포 확률을 이용하여 단어 간 코사인 유사도를 적용하고 문장 간의 유사도를 계산하는 알고리즘의 의사코드(pseudo code)이다.

```
1.  cossum ← 0                          //단어 간 코사인 유사도의 최대값들의 합을 저장할 변수
2.  K ← ⟨k_1 ··· k_kl⟩                  //K는 한글 단어 집합, k는 한글 단어, kl은 한글 단어의 개수
3.  E ← ⟨e_1 ··· e_el⟩                  //E는 영어 단어 집합, e는 영어 단어, el은 영어 단어의 개수
4.  k_i ← ⟨kt_1 ··· kt_n⟩               //kt는 한글 단어 k의 토픽 확률 분포, n은 토픽 개수
5.  e_j ← ⟨et_1 ··· et_n⟩               //et는 영어 단어 e의 토픽 확률 분포
6.  for iter = 1 to Min(kl, el) do
7.    max ← 0                           //두 문장 간 단어들 중 코사인 유사도가 가장 큰 값을 저장할 변수
8.    for all pairwise (k_i, e_j), k_i ∈ K, e_j ∈ E do
9.      if max < Cos(k_i, e_j)          //두 단어의 코사인 유사도 계산
10.       max ← Cos(k_i, e_j)           //단어들 간 코사인 유사도 중 가장 큰 값을 저장
11.     end if
12.   end for
13.   cossum ← cossum + max             //단어들 간 코사인 유사도 중 가장 큰 값을 더함
14.   K ← K − ⟨k_max⟩                   //코사인 유사도의 max값을 가지는 단어를 단어 집합에서 제거
15.   E ← E − ⟨e_max⟩
16. end for
17. return cossum
```

그림.1 코사인 유사도를 적용한 문장 간 유사도 계산 알고리즘 의사코드

그림 1에서 한글 문장은 한글 단어의 집합, 영어 문장은 영어 단어의 집합으로 보고, 두 문장 간 단어들의 코사인 유사도를 구한다. 여기서, 단어들 간 코사인 유사도 값이 가장 높은 순서대로 그림 1의 *cossum*에 더해지며, 가장 높은 코사인 유사도 값으로 선택된 한글 단어와 영어 단어는 각 원본 문장에서 제외된다. 이는 유사도를 계산하고자 하는 문장 안의 단어가 없어질 때까지 반복된다. 단어 간의 토픽 분포가 가장 비슷한, 가장 높은 코사인 유사도 값이 순서대로 더해진 값의 합 *cossum*은 식 (1)의 M_{11} 대신에 추가되며, M_{10}과 M_{01}은 각각 한글 문장 단어의 수와 영어 문장의 원래 단어의 수를 따른다.

3.3 언어 자원 기반과 단어의 토픽 분포 기반의 결합

본 논문에서 제안하는 언어 자원 기반 유사도 계산 방식과 단어의 토픽 분포 기반의 유사도 계산 방식을 결합한 방법은 언어 자원 기반의 방식의 유사도 계산을 수행하고, 두 문장에서 교차하지 않고 남아 있는 단어들을 단어 간의 토픽 확률 분포를 이용, 단어 간 코사인 유사도 값을 구하여 가장 높은 순서대로 추가하는 방법이다.

식 (2)는 위의 두 유사도 방식의 결합에 적용되는 수식으로 자카르트 유사도의 변형으로 A문장과 B문장의 유사도 $J_{DT}(A,B)$를 계산한다.

M_1 : A, B의 각 단어 중 교차하는 단어의 수

M_2 : A, B의 단어로부터 순서대로 더한 코사인 값의 합

M_3 : A에 교차하지 않고 남아있는 단어의 수

M_4 : B에 교차하지 않고 남아있는 단어의 수

$$J_{DT}(A,B) = \frac{M_1 + M_2}{M_1 + M_3 + M_4} \qquad (2)$$

그림 2은 앞서 제안한 두 유사도 계산 방식을 결합한 방법의 이해를 돕기 위해 전체적인 과정을 설명한다.

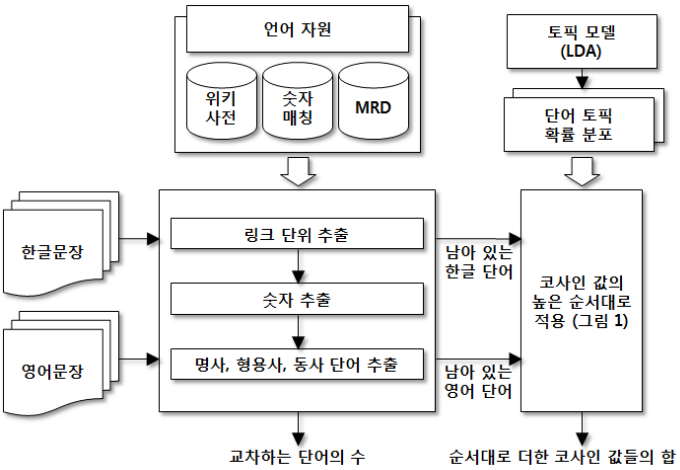

그림.3 언어 자원과 단어의 토픽 분포의 결합 과정

여기서, A와 B는 각각 한글과 영어라고 가정한다. 언어 자원 기반에 사용된 각 자원은 위키 사전, 숫자 매칭, 다음 온라인 사전의 순서대로 적용되어 교차하는 단어의 수가 M_1에 적용된다. 단어 중, 매칭되지 않고 남아있는 각 문장의 한글 단어의 수를 M_3, 영어 단어의 수를 M_4에 적용한다. 두 문장 안에 남아있는 단어는 토픽 모델을 통한 단어 간의 토픽 확률 분포를 이용하여 코사인 유사도를 계산하고, 가장 높은 순서대로 추가한 코사인 유사도 값의 합이 M_2에 적용된다.

그림 3은 실제로 위키피디아의 문서 쌍에 존재하는 한글 문장과 영어 문장의 유사도 계산에서 사전 기반 유사도 계산의 예를 보인 것이다.

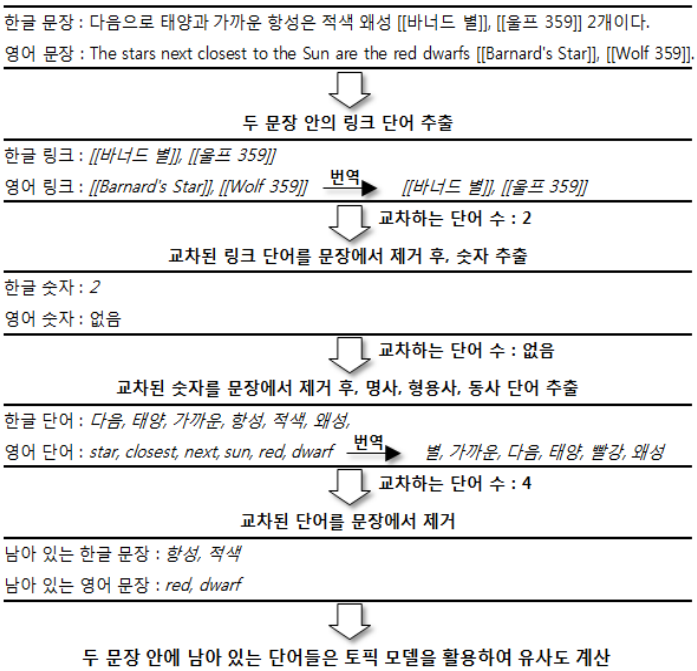

그림.2 언어 자원 기반의 유사도 계산의 실제 예

그림 3에서 언어 자원 기반으로 매칭이 되어 교차하는 단어의 수는 총 6개이며, 교차하지 않고 남아있는 단어의 수는 두 문장에서 각각 2개이다. 여기서, 남아있는 단어들은 단어 간의 토픽 분포를 이용한 코사인 유사도를 적용하여 유사도를 계산한다. 그림 4는 위의 과정에서 이어지는 단어 간의 토픽 분포를 이용한 코사인 유사도 계산의 예이다.

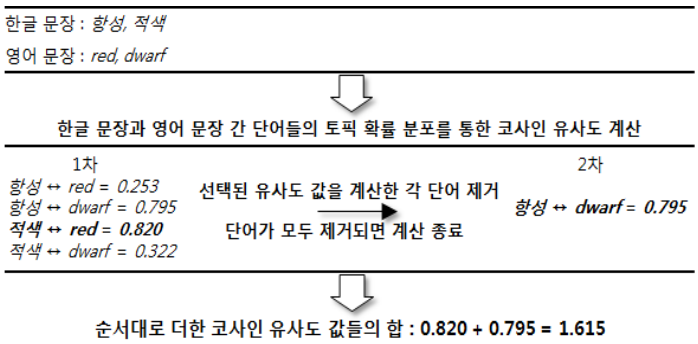

그림.4 토픽 모델 기반의 유사도 계산의 실제 예

그림 4에서 볼 수 있듯이, 두 문장의 남아있는 단어 중 가장 큰 코사인 유사도 값을 가지는 적색-red가 가장 유사한 두 단어로 선택되어 적용된다. 이런 과정은 계산할 단어가 없어질 때까지 반복되며, 높은 순서대로 모두 더해진 코사인 값의 합은 1.615가 된다. 결과적으로 식(2)에 대입하면 그림 3의 예를 통해 M_1과 M_3, M_4를 구할 수 있으며 그림 4의 예를 통해서는 M_2를 구할 수 있다.

4. 실험

4.1 실험 데이터

본 논문에서 비교 말뭉치로 사용한 위키피디아 데이터는 동일한 제목으로 연결되어 있는 한국어-영어 간 문서 쌍이 총 106,582개이며, 각 문서의 길이가 충분히 길며 양쪽 문서의 길이가 차이가 많이 나지 않는 임의의 100개의 문서 쌍을 선택하였다.

1개의 한국어 문장에 대해 2개 이상의 영어 문장, 혹은 2개 이상의 한국어 문장에 대해 1개의 영어 문장과 정답 문장 쌍을 이루는 경우도 가능하기 때문에, 5명의 어노테이터(annotator)는 복수 문장끼리의 일치 여부도 검토하면서 정답 문장 쌍을 수작업 태깅하였다. 그 결과 총 3,100개의 병렬 문장쌍이 정답으로 태깅되었으며, 실험 데이터의 문서와 평균 문장 수 및 단어 수는 표 1과 같다.

표.1 실험에 사용된 위키피디아 정답 문서의 수

개수	한국어	영어
문서 수	100	100
문서 당 평균 문장 수	65.1	77.8
문장 당 평균 단어 수	9.3	11.9

토픽 모델 학습은 mallet toolkit[12]에서 제공한 오픈소스를 사용하여 학습하였고, 토픽 수는 가장 좋은 성능을 보인 1000개를 사용, 파라미터 α와 β는 각각 0.01과 50/토픽 수로 지정해서 사용하였다. 토픽 모델 학습에 사용된 한국어-영어 간 문서 쌍의 수는 총 7400여개이다. 이는 실험 데이터로 사용한 100개의 문서 쌍에 존재하는 인터위키 정보인 링크를 제목으로 하는 문서들로 확장하는 과정을 두 차례 진행하여 얻어진 문서 쌍에서 각 문서의 길이가 충분이 길고, 양쪽 문서의 길이가 차이가 많이 나지 않는 문서 쌍 만을 추출한 수이다.

4.2 실험 결과

성능 평가 도구로는 정확률(precision), 재현율(recall), F1-score을 사용하였으며, 두 문장 간 유사도 값의 적절한 임계점을 찾아 정답으로 간주하였다.

모든 실험 결과는 선행 연구[2]를 Baseline으로 본 논문의 제안 방법과 비교하였다. 표 2는 언어 자원을 기반으로 문장 간 유사도를 계산한 실험 결과이다. 여기서 사전은 l(링크 사전), m(다음 온라인 사전), n(숫자 사전)이라 지칭한다.

표.2 언어 자원 기반 유사도 계산의 실험 결과

방법	정확률	재현율	F1-score
lmn(base)	49.1%	47.7%	48.4%
lmn(proposed)	46.2%	57.7%	**51.3%**

위 실험 결과에서 언어 자원마다 각각 문장 간의 유사도를 구하여 선형 결합한 기존 방법(base)보다 언어 자원 간의 우선순위를 반영하여 문장 간의 유사도를 한 번에 계산하는 제안 방법(proposed)이 2.9%의 성능 향상을 보였다. 표 3은 토픽 모델을 통해 단어의 토픽 분포를 기반으로 유사도를 계산한 실험 결과이다.

표.3 단어의 토픽 분포 기반 유사도 계산의 실험 결과

방법	정확률	재현율	F1-score
topic(base)	약 10%		
topic(proposed)	26.5%	47.8%	**34.1%**

선행 연구에서 문장 안의 모든 단어들의 토픽 분포를 고려하여 문장 간 유사도를 계산한 방법(base)은 성능이 약 10%로 낮았기 때문에 단독으로는 사용하지 않고 언어 자원과 결합한 방법만을 제안하였다.

표 3에서 알 수 있듯이 본 논문에서 제안한 단어의 토픽 분포를 기반으로 가장 유사한 단어의 순서대로 고려하여 문장 간 유사도를 계산하는 제안 방법의 성능이 기존의 성능에 비해 약 24%정도 향상된 것을 알 수 있다. 34.1%의 성능은 비록 다른 언어 자원을 사용하는 방법과는 차이가 크지만, 수작업이 필요 없고 별도 원천자원도 없이 오직 비지도 학습만을 이용하여 얻은 성능이라는

점에서 매우 높은 성과를 얻었다고 할 수 있다.

다음으로 표 4는 위의 두 유사도 방식인 언어 자원 기반의 유사도 계산 방법과 단어의 토픽 분포 기반의 유사도 계산 방법을 결합한 실험 결과이다.

표.4 언어 자원과 단어 토픽 분포를 결합한 실험 결과

방법	정확률	재현율	F1-score
lmn+topic(base)	45.8%	59.1%	51.6%
lmn+topic(proposed)	48.9%	61.1%	**54.3%**

실험 결과, 최종 성능 면에서도 언어 자원을 선형 결합한 유사도 계산과 단어의 토픽 분포를 이용한 모든 단어 간의 관계를 반영한 유사도 계산의 결합의 기존 방법(base)에 비해 약 3% 개선된 성능을 얻을 수 있었다. 이는 언어 자원 기반으로는 유사도를 한 번에 계산하고, 매칭이 되지 않고 남아있는 단어들에 대해서는 토픽 분포를 이용한 단어 간 코사인 유사도 방법을 적용한 유사도 계산 방법의 성능이 기존 방법에 비해 우수하다는 것을 알 수 있다.

이와 같이 언어 자원을 이용하는 것은 두 문장의 유사도를 계산함에 있어 기본적으로 가장 큰 영향을 미치며, 단어의 토픽 확률 분포를 이용하는 것은 완벽하게 일치하는 단어를 제외하고 의미적으로 비슷한 단어들을 추가적으로 찾아낼 수 있다는 점에서 긍정적으로 작용한다고 볼 수 있다.

그림 5는 기존 방법과 제안 방법의 성능을 전체적으로 보여주고 있다.

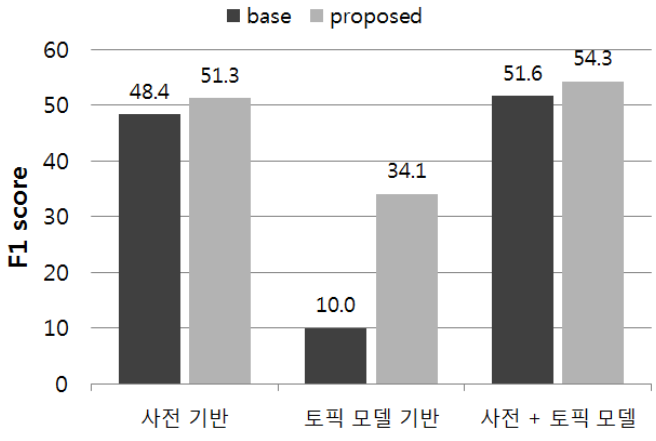

그림.5 기존 방법과 제안 방법의 성능 비교(%)

5. 결론

본 논문에서는 위키피디아라는 신뢰 있는 원천 자원을 비교 말뭉치로 이용하여 병렬 문장 자동 추출을 위해 제안하는 유사도 계산 방법에 대한 실험을 수행하였다. 이는 주위에서 쉽게 활용할 수 있는 자원만을 활용하여 병렬 문장 추출의 정확도를 높이고자 하는 목표였으며 비지도 학습인 토픽 모델을 적극 활용하여 성능 향상을 이루었다.

본 논문에서 제안한 방법의 기여는 크게, 1)위키피디아와 온라인 사전과 같은 주위에서 쉽게 얻을 수 있는 언어 자원만을 활용하였다는 점, 2)비지도 학습인 토픽 모델만을 단독으로 이용한 병렬 문장 추출의 성능이 기존 방법에 비해 약 24%정도 향상되었다는 점, 3)마지막으로 위의 두 방식을 결합한 방법의 성능 또한 약 3%정도 향상되었다는 것이다.

향후 과제로는 일부 위키피디아 데이터가 아닌 전체 한국어-영어 위키피디아 데이터로부터 대용량의 병렬 문장을 추출하는 작업을 병행할 것이며, 또한 한국어-일본어, 한국어-중국어 등 다양한 다른 언어들에 대해서도 연구를 지속할 계획이다.

참고문헌

[1] Teubert Wolfgang, “Comparable or parallel corpora?,” International journal of lexicography, 9(3), p.238, 1996.

[2] 김성현, 양선, 고영중, “위키피디아로부터 한국어-영어 병렬 문장 추출”, 정보과학회논문지 : 소프트웨어 및 응용, 제41권 제8호, p.580-585, 2014.

[3] Dragos Stefan munteanu and Daniel Marcu, “Improving machine translation performance by exploiting non-parallel corpora,” Computational Linguistics, 31(4), p.477, 1995.

[4] Tao Tao and ChengXiang Zhai, “Mining comparable bilingual text corpora for cross-language information integration,” In proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery in data mining (KDD-2005), p.691, 2005.

[5] Ramirez Jessica C and Yuji Matsumoto, “A Rule-Based Approach For Aligning Japanese-Spanish Sentences From A Comparable Corpora,” arXiv preprint arXiv:1211.4488, 2012.

[6] Utiyama Masao and Hitoshi Isahara, “Reliable measures for aligning Japanese-English news articles and sentences,” In proceedings of ACL ‘03, p.72, 2003.

[7] Adafre Sisay Fissaha and Maarten De Rijke. "Finding similar sentences across multiple languages in wikipedia," In Proceedings of ACL ‘06, p.62, 2006.

[8] David M. Blei, Andrew Y. Ng and Michael I.Jordan, "Latent dirichlet allocation," The Journal of Machine Learning research, 3, p.993, 2003.

[9] Zede Zhu, Miao Li, Lei Chen and Zhenxin Yang, "Building Comparable Corpora Based on Bilingual LDA Model," In Proceedings of ACL ‘13, p.278, 2013.

[10] Ivan Vulic, Wim De Smet, and Marie-Francine Moens, "Cross-language information retrieval with latent topic models trained on a comparable corpus," Information Retrieval Technology, Springer Berlin Heidelberg, p.37, 2011.

[11] Mallet toolkit, http://mallet.cs.umass.edu/download.php