한국어/영어 병렬 코퍼스에서 구 단위 정렬을 위한 단위 구 자동 추출

김기태[0] 김동주 김한우
한양대학교 컴퓨터공학과
{ktkim, djkim, kimhw}@cse.hanyang.ac.kr

Automated Unitary Phrases Extraction for Aligning Phrases in Korean-English Bilingual Corpus

Ki-Tae Kim[0] Dong-Joo Kim Han-Woo Kim
Dept. of Computer Science & Engineering, Hanyang Unviersity

요 약

정렬(alignment)은 병렬 코퍼스에서 원문서의 문단, 문장, 혹은 단어와 같은 단위 요소에 대해, 대역문서에서의 상응하는 단위 요소를 찾는 일로, 코퍼스 기반 기계번역 방식에서 매우 중요한 과정이다. 동일 어족간의 원문과 대역문에서는 어순이나 단위 요소들이 거의 일치하여 정렬에 큰 어려움이 없으나, 한국어와 영어와 같이 어족이 다른 언어간의 정렬은 언어의 단위 요소의 상이성과 어순의 차이 등으로 인해 많은 어려움이 존재한다. 본 논문은 어족이 다른 언어 사이의 정렬을 위해 상대 구문 고립성(Relative Syntactic Isolativity)이라는 개념을 적용하여 언어 단위의 상이성을 극복할 수 있는 단위 구를 제안하고 이들을 추출하는 방법에 대해 보인다.

1. 서 론

코퍼스 기반 기계번역 방식에서 어떤 언어가 다른 언어로 번역되는 경향을 살피고, 그 경향을 학습하기 위해 선행되어야 할 작업이 정렬이다. 코퍼스로부터 대역 사전의 자동 구축과 동일한 과정으로 생각할 수 있는 정렬은 병렬 코퍼스에서 한 언어로 작성된 문서의 문단, 문장 혹은 단어와 같은 단위 요소에 대해, 다른 언어로 작성된 문서 내에서 상응하는 요소를 찾는 것이다. 동일 어족간의 정렬에서는 많은 연구가 있었으며 많은 성공적인 방법들이 제안되었다. 그러나 동일 어족간의 정렬과는 달리 한국어와 영어와 같은 서로 다른 어족간의 정렬은 단위 요소들의 상이성과 어순의 차이 등과 같은 이유로 많은 어려움이 존재한다[1, 2]. 이와 같은 상이성을 극복하기 위해 신중호[1]에서는 한국어와 영어에 대하여 셉트(cept) 개념을 도입하여 정렬 단위를 구로 확장하였으나 단순히 공기 빈도만을 이용하여 인접한 형태소의 나열을 초기 구로 설정하고 인접한 형태소들로 구의 범위를 확장하여 정렬을 행하였기에 구 단위 정렬의 정확도에 한계가 있었으며, 구를 언어학적인 구가 아닌 단순히 인접한 형태소들의 나열로 정의함으로써 결과물로 제시된 대역사전에는 많은 문제점을 내포하고 있다. 따라서 본 논문에서는 어족이 다른 언어 사이의 정렬을 위해 언어 단위의 상이성을 극복할 수 있는 단위 구를 제안하고 이들을 추출하는 방법에 대해서 보인다.

2. 정렬의 대상 단위와 정렬

정렬은 대상에 따라 크게 문단 혹은 문장간의 정렬과 문장을 이루고 있는 구성 요소들에 대한 정렬로 나눌 수 있다. 문단 혹은 문장간의 정렬은 서로 다른 언어로 구성된 두 개의 코퍼스에서 주어진 문장의 대역 문장을 찾는 것이고, 문장을 이루고 있는 구성 요소들에 대한 정렬은 주어진 문장을 이루고 있는 구성 요소에 대응하는 대역 문장의 구성 요소를 찾는 것이다. 문장을 이루고 있는 요소들에 대한 정렬은 언어간의 언어 단위의 상이성에 따라 정렬의 성능과 난이도가 달라지며 정렬의 단위 설정에 따라서도 마찬가지다. 정렬의 대상 단위로는 형태소, 단어, 구, 구문 트리 등이 있는데, 정렬의 대상 단위가 작아질수록 언어 사이의 단위의 상이성에 직면하게 되어 정렬이 어려우며 단위가 커질수록 단위의 상이성이 작아져 정렬은 쉬워지나 정렬의 대상이 되는 단위의 추출이 힘들어 진다. 특히 구문 트리의 정렬과 같은 경우 사용된 문법의 특성에 따라 구문 트리의 모양이 달라지므로 두 언어 모두 동일한 특성을 지닌 문법을 사용하여 구문 트리를 만들어야 하며 문법이 지닌 모호성에 의해 구문 트리의 모양이 크게 달라지므로 정렬의 대상이 되는 단위의 추출에 어려움이 많다.

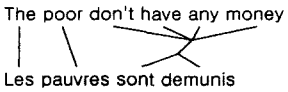
2.1 동일 어족에 속하는 언어간의 정렬

영어와 불어, 영어와 독일어의 경우 거의 비슷한 언어 단위를 지니며 이들 언어가 동일 어족에 속하는 까닭에 어순 또한 비슷하다. 따라서 많은 경우에 있어 원문의 한 단어가 번역 대상 언어의 한 단어와 정렬이 가능하다. 또한 원문의 단어와 번역 대상 언어의 단어 사이의 정렬에 필요한 단어 비가 1:n, n:1, n:n인 나머지 경우에 있어서도 [그림 1]과 같이 어순이 비슷한 까닭에 Brown[3]에서처럼 공기 빈도가 높은 인접한 단어들을 셉트로 묶은 후 정렬을 수행하고 다시 인접한 셉트들로 구를 확장하는 것만으로도 효율적인 정렬이 가능하며 효용성 있는 대역 쌍의 획득이 가능하다.

2.2 서로 다른 어족에 속하는 언어간의 정렬

중국어와 영어, 일본어와 영어와 같이 어족이 서로 다른 언어들 사이에는 언어 단위의 상이성 정도가 클 뿐만 아니라 어순 또한 차이가 많다. 따라서 각 언어의 언어 단위를 정렬의 단위로 할 경우 동일 어족에 속하는 언어간의 정렬에 비해 정렬의 성능이

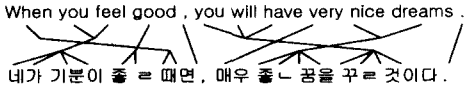
크게 떨어진다. 따라서 이들 언어들 사이의 정렬 성능을 높이기 위해서는 언어 단위의 상이성을 극복할 수 있는 정렬의 대상 단위 설정이 필요하며 이와 같은 노력으로 구문 트리의 정렬을 통해 언어 단위의 상이성을 극복하려는 시도가 있었다[2, 4].



[그림 1] 영어/불어 사이의 정렬

2.2.1 한국어와 영어 사이의 정렬

한국어와 영어 사이의 정렬은 신중호[1]에서 단어 단위 및 구 단위 정렬을 Brown[3]의 모델을 한국어와 영어 사이의 정렬에 맞게 확장 이식함으로써 이루어졌으며 68.7%라는 비교적 양호한 구 단위 정렬의 정확도를 보였다. 그러나 이것은 어순이 서로 다른 한국어와 영어에 대해 단순히 공기 빈도에만 의존하여 빈도가 높은 인접한 형태소들로 셉트(cept)를 형성하고 이웃한 셉트로 정렬 단위 구를 확장하였기 때문에 정렬의 정확도에 한계가 있었다.



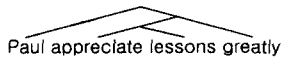
[그림 2] 한국어/영어 구단위 정렬(신중호[3])

또한, [그림 2]의 경우 "you"와 "네가"가 "feel good"과 "기분이 좋ㄹ"이 각각 구 단위가 되어 정렬되는 것이 직관적이겠으나 구를 언어학적인 구가 아닌 단순히 인접한 형태소들의 나열로 정의한 까닭에 언어학적인 요소가 반영되지 않아 "you feel"과 "네가 기분이"가 구 단위가 되는 등 정렬 결과의 효용성이 제한되었다.

3. 정렬을 위한 단위 구와 단위 구의 추출

3.1 정렬을 위한 단위 구

언어 단위의 상이성을 극복하고 정렬의 정확성과 정렬 결과의 효용성을 높이기 위해서는 문장을 이루는 언어 단위들 사이의 관계와 그런 관계가 의미적으로는 어떤 역할을 하는지 역시 고려되어야 한다. 본 논문에서는 단어들 사이의 관계에 대한 척도로서 상대 구문 고립성(Relative Syntactic Isolativity)이라는 개념을 도입한다. 문장내의 인접한 각각의 단어와 단어, 단어와 구, 구와 구 사이의 관계는 서로 상대적으로 다른 연관성을 지니며 연관성이 높은 요소들이 먼저 결합된다. 이와 같은 과정이 반복되면서 주변의 다른 요소들과 기능이나 의미의 구분이 명확한 상대적인 단위를 형성하게 되는데 이런 특성을 상대 구문 고립성이라 하며 문장내의 요소들간의 관계를 규명할 수 있는 구문 규칙들에 의해 구문 고립성이 표현될 수 있다. [그림 3]과 같은 문장에서 주술 관계에 있는 Paul과 appreciate, 술부를 수식하고 있는 greatly와 lessons보다는 함께 술부를 이루고 있는 appreciate와 lessons가 보다 강한 상대 구문 고립성을 나타내는 까닭에 구문 구조에 있어서도 이 두 단어가 먼저 결합하게 되어 의미적으로 고립된 단위를 형성해 나가게 되며 마지막에는 문장이라는 구문 고립성을 지닌 최대 단위를 형성하게 된다.



[그림 3] 구문 고립성과 구문 구조

상대 구문 고립성에 의해 구가 형성되는 과정에서 문장 내에 독자적으로는 의미를 형성할 수 없는 어휘가 남지 않는 최초의 구문 고립적인 단위들을 추출하게 되면 이들은 문장내의 다른 요소의 영향을 받아 그 의미나 기능이 변질되지 않는 지역적으로 폐쇄적인 성질을 갖게 될 것이다. 예를 들어, "with a telescope"라는 단어의 나열이 있을 때 "with a telescope"는 의미적으로 폐쇄적인 구이지만 "a telescope"는 "with"에 의해 그 의미와 기능이 변질되므로 의미적으로 폐쇄적인 구가 아니다. 이와 같은 성질을 고려하여 다음과 같이 직관적으로 생각을 해볼 수 있다.

"*대역 관계에 있는 상이한 언어로 된 두 문장이 있을 때, 두 문장의 의미는 동일한 것이므로 의미가 부분적으로 폐쇄적인 구를 추출하게 되면 두 문장에서 추출된 구들은 두 언어에 대해 일치하는 단위를 형성하게 되며 그럼으로써 이들 사이에는 대역 쌍이 존재하게 된다.*"

지금까지 설명한 구문적인 특성들에 따라 다음과 같이 정렬을 위한 단위 구를 정의할 수 있다.

정의 : 문장의 한 부분으로서 폐쇄적인 의미와 기능의 단위를 지니는 문법적인 구로서 하나의 단위 구는 또 다른 단위 구를 내포할 수 없다.

(I) (saw) (a boy) (in the car) (with a telescope)
(나는) (망원경을 가지고) (차 안에 있는) (소년을) (보았다)

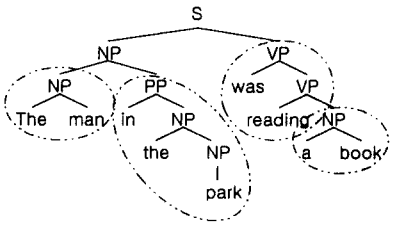
[그림 4] 정렬을 위한 단위 구의 추출 예

[그림 4]는 정의에 부합하는 정렬을 위한 단위 구의 한 예이며, 정렬을 위한 단위 구를 추출하기에 앞서 반드시 상기해야 할 한 가지 사실은 문장을 이루는 요소들의 관계들에 따라 상대적으로 결정되는 단위라는 사실이다. 예를 들어, "The man with long hair"와 같은 구문에서 "The man"은 정렬 단위 구이지만 "with the man"과 같은 구문에서는 "the man"이 아닌 "with the man"이 정렬 단위 구가 된다.

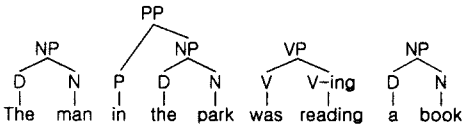
3.2 단위 구의 추출

정렬을 위한 단위 구의 추출은 구문 트리 상에서 단위 구의 정의에 부합하는 노드를 찾아 그 구성 요소들을 취하면 된다. 관찰한 바에 따르면 구문 트리 상에서 단위 구의 정의에 부합하는 노드들은 대략 뿌리 노드로부터 각각의 말단 노드들까지의 깊이의 이분의 일인 지점을 전후로 분포한다. 그러나 구문 구조의 모호성과 구 안에 구를 허용하는 문맥 자유 다시 쓰기 규칙(Context-Free rewriting rule)의 특성으로 단위 구의 추출이 쉽지 않은 문제점을 갖는다. [그림 5]에서는 이런 문제점으로 점선으로 묶여 있는 정렬을 위한 단위 구를 추출해내는 것이 매우 어렵다. 그러나 상대 구문 고립성을 고려하여 구문 규칙을 기술하고 이를 이용하여 정렬 단위 구가 형성되는 시점까지 부분 구문 분석을 수행하게 되면 [그림 6]과 같이 구문 구조의 모호성과 문맥 자유

다시 쓰기 규칙의 특성이 배제하고 정렬을 위한 단위 구를 추출할 수 있다.



[그림 5] CF다시쓰기 규칙의 구문 트리와 단위 구



T1: NP→D? N* N; VP→V-tns | Aux V-ing
T2: PP→P NP

[그림 6] 부분 구문 분석을 통한 단위 구 추출

4. 실 험

4.1 데이터

코퍼스는 대용량 음성/언어/영상 DB 구축 및 표준화 사업 발표회(KSURIMAL)에서 배포한 영한 기계번역시스템 평가세트를 사용하였다. 본 코퍼스는 중·고등학교 교과서와 여러 문법책들에서 수집된 대역 관계에 있는 한국어와 영어 1238문장 쌍(영어:10704개 어절, 한국어:8295개 어절)을 포함한다. 코퍼스에 포함된 영어 문장들은 평서문, 의문문, 가정법, 완료, 진행, 감탄문, 명령문 등 다양한 언어 현상들이 반영되어 있으며, 문장들의 구문 구조 또한 다양하여 상대 구문 고립성의 관찰을 통해 정렬을 위한 단위 구 추출에 사용할 구문 규칙들을 설계하고 시험해 보기에 적합하다. 코퍼스의 품사 부착은 영어와 한국어 각각에 대해 펜 트리 뱅크 태그 세트(Penn Tree Bank Tag Set)와 ETRI 태그 세트를 이용하여 품사를 부착하였다.

실험에 사용한 데이터에서는 두 문장이 한 문장으로, 혹은 한 문장이 두 문장으로 번역되어 있는 경우와 관용문과 같이 의미의 중심이 문장 내에 존재하지 않아 정렬이 불가능한 경우, 그리고 본 논문에서 구현하여 사용하고 잇는 현재 구문 분석기의 성능 상 처리할 수 없는 인용문과 같은 경우 등의 245 문장 쌍을 제외한 1083 문장 쌍만을 사용하였다.

4.2 실험 결과

코퍼스로부터 무작위로 100문장 쌍을 선출하여 이들로부터 상대 구문 고립성을 관찰하고 상대 구문 고립성이 반영된 한국어와 영어 부분 구문 분석 규칙을 작성한 후 이를 다시 100문장 쌍에 적용하여 부분 구문 분석을 행하는 과정에서 정렬에 적합한 단위 구가 추출되도록 부분 구문 분석의 깊이를 조정하였다. 실험은 구문 규칙 작성에 사용한 100문장 쌍을 제외한 983문장 쌍에 적용하여 부분 구문 분석을 통해 정렬을 위한 단위 구를 추출하였고 실험 결과의 산출은 각 문장 쌍에서 자동 추출된 단위 구를 수작업으로 정렬하여 추출된 단위 구들이 실제 정렬에 어느 정도 기여할 수 있는지를 측정하였다.

[표 1] 실험 결과

정렬 가능한 단위 구	1:1	74.3%	90.1%
	1:n	15.8%	
정렬 불가능한 단위 구		9.9%	

[표 1]의 실험 결과는 추출된 단위 구가 정렬이 되는 것과 되지 않는 것으로 나누어 통계를 내었으며 정렬이 되는 경우에는 하나의 한국어 단위 구가 하나의 영문 단위 구와 정렬이 되는 이상적인 경우와 둘 이상의 영문 단위 구와 정렬이 되는 경우로 세분류하였다. 후자의 경우 하나의 한국어 단위 구가 두 개의 영문 단위 구와 정렬이 되는 경우가 전체의 12.3%였으며 그 상당 부분이 영어의 be 동사와 보어가 되는 명사구가 하나의 한국어 단위 구에 정렬되는 것이었다. 정렬이 되지 않는 단위 구들로는 관용구, 원문과 등가인 번역문이 존재하지 않는 경우, 부정의 의미를 지니는 단어들 (any, none, little, few 등)이 있다.

5. 결 론

정렬이라는 작업에서 대역 쌍을 찾는 행위는 번역이라는 작업에서의 행위와 동일하다[3]. 따라서 기계 번역에서의 어려운 점들은 정렬에서도 마찬가지로 처리하기 어려운 점들이며 이는 정렬을 위한 단위 구 추출에서도 그대로 반영되어 정렬이 되지 않는 단위 구들이 추출되는 근본 원인이 되었다.

본 논문에서는 추출된 단위 구에 대해 수작업으로 정렬을 수행하였으나 실제 정렬 작업에서도 정렬 대상 후보가 최소화되었으므로 정렬의 성능을 향상시킬 것으로 기대된다. [표 1]의 실험 결과에서 정렬이 되는 단위 구중 한국어와 영어 단위 구에 대해 1:n으로 대응되는 경우는 정렬 단위에 대한 품질을 향상시키기 위해 확률을 이용하여 추출된 대역 쌍들에 적합한 정렬 모델의 개발이 필요하다.

6. 참고문헌

[1] 신중호, "한국어/영어 병렬 코퍼스에 대한 단어단위 및 구단위 정렬 모델", 한국과학기술원 전산학과 석사학위논문, 1996.

[2] 최기선, 신중호, "한/영 정렬 시스템", 국어정보베이스 1차년도 최종 보고서, pp.137-156, 1995.

[3] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, "The Mathematics of Machine Translation: Parameter Estimation," Computational Linguistics, vol. 19(2):263-311, 1993.

[4] Dekai Wu, "Grammarless Extraction of Phrasal Translation Examples from Parallel Texts," Tech. Report, Hong Kong Univ., 1995.