통계적 수정규칙을 이용한 한국어-중국어 단어정렬 개선방법

김장호[O1] 이금희[2], 나휘동[2], 김동일[3], 이종혁[2]
포항공과대학교 정보통신대학원 정보통신학과[1]
포항공과대학교 컴퓨터공학과[2], 중국 연변과학기술대학교 언어공학연구소[3]
{hchchh1, ljj2, leona2, jhlee2}@postech.ac.kr, dongil3@ybust.edu.cn

Using Statistical Correction Rule to Improve Word Alignment

Chang-hu Jin[O1], Jin-Ji Li[2], Hwidong Na[2], Dong-il Kim[3], Jong-Hyeok Lee[2]
Dept. of Graduate School for Information Technology, POSTECH
Dept. of Computer Science & Engineering, POSTECH
Language Engineering Institute, YUST, China

요 약

본 논문에서는 통계적으로 추출한 수정규칙을 이용하여 구 기반 한-중 통계기계번역 시스템(PBSMT)의 단어정렬 결과를 개선하는 방법을 제안한다. 논문에서 제안하는 수정규칙은 단어정렬의 결과를 사람이 만든 정답과 비교하여 통계적으로 추출하였다. 본 논문에서는 위에서 추출한 수정규칙을 이용하여 한-중 통계기계번역 시스템의 단어정렬의 결과에서 한국어 기능어(functional word)에 나타나는 오류를 수정함으로써 단어정렬의 결과를 개선하였고 최종적으로 기계번역의 성능을 제고하였다.

주제어: 단어정렬, 통계기계번역, 수정규칙

1. 서론

단어정렬의 좋고 나쁨은 구 기반 통계 기계번역 시스템에서 구 추출에 직접적인 영향을 미치는 중요한 요소이다. 때문에 단어정렬의 결과를 개선하는 작업은 최근 통계 기계번역 시스템의 첫 단계로서 큰 비중을 차지하며 연구되어 왔다. 하지만 현재 사용되는 많은 단어정렬 알고리즘은 언어학적 분석이 없이 오직 수학적인 모델과 확률만으로 두 단어사이의 매치확률을 계산한다. 그러므로 이런 알고리즘을 사용하여 단어정렬을 하면 두 언어의 언어학적 차이를 반영하지 못하게 된다.

한국어와 중국어는 언어학적 차이가 큰 대표적인 언어쌍이다. 한국어는 조사, 어미가 많이 발달되어 있지만 중국어는 이런 특징이 없다. 때문에 많은 한국어 조사, 어미는 중국어문장에서 대응되는 단어를 가지지 않거나 또는 대응되는 중국어 단어가 명확하게 표현되어 있지 않다. 하지만 일부 한국어 조사나 어미는 대응되는 중국어 단어가 있으며 중국어 문장에 따라 여러 가지 중국어 단어에 대응될 수 있다.[1] 기존의 단어정렬 알고리즘[2]은 이런 언어학적 차이를 고려하지 않고 오로지 해당 단어가 나타나는 빈도수를 기반으로 만들어졌기에 이런 방법으로 단어정렬을 하면 많은 오류가 생긴다.

Jin09에서는 기존 알고리즘으로 언어학적 차이가 큰 한국어-중국어 단어정렬을 하였을 때 나타나는 문제점을 한국어 형태소 유형에 따라 분석하였다.[1] 그 결과 한국어의 조사, 어미, 파생접사로 표현되는 기능어(functional word)에 의하여 생기는 오류가 가장 심각하게 나타나는 것을 알 수 있었다.

이런 문제점을 해결하기 위하여 본 논문에서는 통계적으로 추출된 수정규칙을 이용하여 단어정렬의 결과를 수정하는 방법을 제안한다. 논문에서 사용하는 수정규칙은 기존의 알고리즘으로 생성한 단어정렬의 결과를 사람이 만든 정답과 비교하여 자동적으로 추출하였다. 수정규칙은 정답에 기준하여 생성하였고 정답은 언어학적 지식을 바탕으로 사람이 만들었기에 수정규칙에는 한국어와 중국어의 언어학적 차이가 반영되었다.

본 논문은 아래와 같이 구성되었다. 2장에서 단어정렬의 결과를 개선하기 위한 선행연구에 대해 기술하고3장에서 본 논문에서 제안하는 단어정렬을 개선하는 방법에 대하여 설명하였으며 4장에서 구체적인 실험방법과 결과에 대하여 기술하고 5장에서 결과에 대해 분석하고 후속과제에 대하여 언급한다.

2. 관련연구

최근 단어정렬의 결과를 개선하는 연구는 몇 가지 특징을 가지고 있다. 첫째로 단어정렬의 결과를 개선하여 통계기계번역의 성능을 제고하는데 초점을 맞추고 있다. 둘째로 단어정렬의 결과를 개선하는 방법을 두 단계로 나누어서 진행한다. 우선 기존에 제안된 방법이나 공개된 단어정렬 도구를 이용하여 단어정렬을 진행한다. 다음 확률정보를 이용하여 생성된 단어정렬의 결과를 신뢰할 수 있는 부분과 나머지 부분으로 나누고 그 나머지 분에 대하여 수정작업을 진행한다. 셋째로 단어정렬을 개선하기 위하여 형태소 분석정보, 구문분석정보를 이용하여 두 언어의 언어학적 특징을 반영하였다.

Ma08, Ma09에서는 중국어-영어 통계기계번역 시스템에서 단어정렬의 결과를 생성하고 수정하는 하나의 통

합모델을 제안하였다.[3][4] 이 통합모델은 단어정렬의 결과를 생성하는 생성모델과 결과를 수정하는 수정모델로 이루어졌다. 본 논문은 생성모델에서 나온 단어정렬의 결과에서 매치확률을 이용하여 신뢰할 수 있는 부분을 추출하였다. 다음 위에서 추출한 신뢰할 수 있는 단어정렬의 결과와 두 언어의 구문분석정보 및 기타 통계정보를 수정모델의 자질로 사용하여 나머지 신뢰할 수 없는 부분의 결과를 수정하였다.

Victoria08에서는 소스언어의 구문분석정보와 기타 통계정보를 이용하여 단어정렬의 결과에서 단어 사이의 링크에 점수를 부여하는 모델을 제안하였다.[5] 이 모델은 점수가 부여된 단어정렬의 결과에서 일정한 점수이하의 링크를 삭제하는 방법을 사용하였다. 본 논문은 링크를 삭제하는 방법만 사용하였기에 정확률(precision)은 제고하였지만 재현율(recall)은 떨어뜨렸다.

이 외에도 Josep08에서는 일정한 어절범위를 초과하는 두 단어사이의 링크를 제거하는 방법으로 단어정렬의 결과를 개선하였고[6] Huang09에서는 서로 다른 모델을 이용하여 단어정렬의 결과를 여러 개 생성하여 통합하고 이 가운데서 일정한 확률이상의 결과만 선택하는 방법을 사용하였다.[7]

하지만 이런 방법은 형태소-구문 차이가 큰 언어 쌍에서 언어학적 차이를 고려하지 않고 오로지 두 단어사이의 통계적 매치확률에 기초하여 단어정렬을 진행하므로 많은 오류가 생긴다. 특히 기능어가 발달한 한국어나 아랍어의 경우 상대 언어에 명확한 대응어가 없는 경우가 많기에 단어정렬이 더욱 어려워진다. 이런 문제를 해결하기 위하여 ThuyLinh08, Lee04에서는 상대 언어에 존재하지 않는 형태소를 삭제하거나 병합하는 방법을 제안하였다.[8][9] 하지만 이런 방법은 기계번역의 다음 단계에 필요한 일부 정보를 단어정렬 단계에서 상실하게 된다.

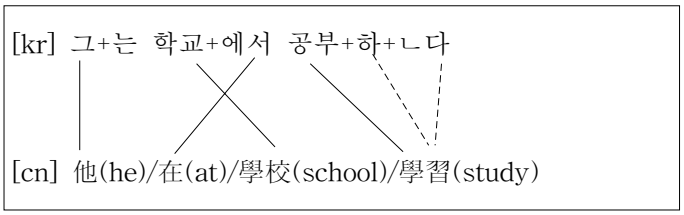
Hermjakob09에서는 형태소 차이가 큰 아랍어-영어 단어정렬에서 아랍어 기능어가 상대 언어에서 명확한 대응어가 존재하지 않는 문제를 해결하기 위하여 단어정렬의 결과가 잘못된 형태소의 패턴을 분석하여 세 가지 수정규칙으로 통합한 다음 이를 이용하여 단어정렬의 결과를 개선하였다.[10] 이 논문에서는 아랍어 기능어의 품사종류를 세 그룹으로 나누고 각 그룹에 대응되는 수정규칙을 만들어 단어정렬을 개선하였다. 하지만 이런 수정규칙은 같은 품사를 가진 형태소는 같은 형태의 오류가 생긴다는 전제를 기반으로 제안되었기에 다양한 형태의 오류가 나타나는 언어 쌍에서는 큰 효과를 얻기 어렵다.

3. 통계적 수정규칙을 이용한 단어정렬 개선방법

본 논문은 통계적으로 수정규칙을 추출하여 기존 알고리즘[2]으로 생성된 단어정렬의 결과를 수정하는 방법을 제안한다.

본 논문에서 제안하는 수정규칙은 정답을 기준으로 만들어 진다. 논문에서 사용한 정답은 사람이 한국어와 중국어의 문법적 지식을 바탕으로 단어정렬을 한 결과이다. 정답 작업에 관한 지침은 Li08을 참조하였다.[11]

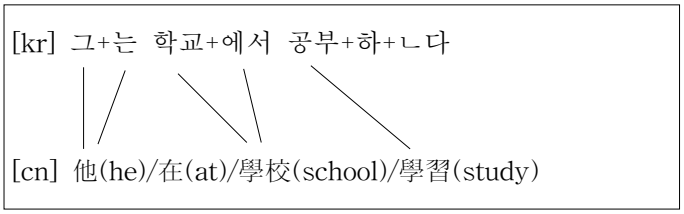
예제1: (영어단어는 각 중국어 단어의 뜻을 나타낸다)



예제1은 정답에서 표시한 링크정보이다. 실선으로 표시된 링크는 명확한 대응을 나타내고 점선으로 표시된 링크는 의미적인 대응을 나타낸다.[11] 명확한 대응과 의미적인 대응 여부는 사람이 언어학적 지식을 바탕으로 결정하였다.

본 논문에서는 GIZA++[1]를 단어정렬의 도구로 사용하여 위에서 표시된 정답을 가지고 있는 한국어-중국어 문장에 대하여 단어정렬을 하였다.

예제2: (영어단어는 각 중국어 단어의 뜻을 나타낸다)



예제2는 GIZA++를 이용하여 단어정렬을 한 결과이다.

예제2에서 볼 수 있듯이 한국어 기능어 형태소 “는”, “에서”, “하”, “ㄴ다”는 정확하게 연결되지 않았다. 이런 한국어 기능어 오류는 자주 틀리는 기능어와 그렇지 않는 기능어가 존재한다. 또 한국어 기능어는 그 종류가 제한되어 있다. 때문에 샘플 말뭉치에서 많이 틀리는 기능어는 대량의 말뭉치에서도 많이 틀리게 된다. 그러므로 정답이 표시된 샘플 말뭉치를 기준으로 추출한 수정규칙을 대량의 말뭉치에 적용하여도 효과적으로 작용할 수 있다.

수정규칙을 생성하기 위한 정답 말뭉치는 한국어와 중국어의 언어학적 지식을 가진 사람이 만들었기에 정답에 기준한 수정규칙은 두 언어의 언어학적 차이를 반영하였다. 수정규칙은 자동으로 생성되기에 정답 말뭉치에서 추출된 수정규칙은 단어정렬 시스템에 곧바로 반영될 수 있다. 또 수정규칙을 추출하기 위한 정답말뭉치는 단어정렬 시스템에 관계없이 독립적으로 작업하여 얻을 수 있다. 때문에 추후 이 말뭉치를 많이 만들수록 수정규칙을 추출하기 위한 학습 자료가 늘어나므로 더욱 정교한 확률을 학습할 수 있고 따라서 수정규칙의 적용범위도 더 늘어날 수 있다.

3.1 수정규칙의 생성

본 논문에서는 GIZA++를 이용하여 얻은 단어정렬 결과와 정답을 비교하여 오류 링크를 가진 집합을 추출하였다. 오류 링크의 집합은 한국어 기능어에 대해서만 추

1) http://www.fjoch.com/GIZA++.html

출하였다. 그 이유는 기능어가 한-중 단어정렬에서 가장 많은 오류를 발생시키기 때문이다.

그림1: 시스템과 정답의 비교

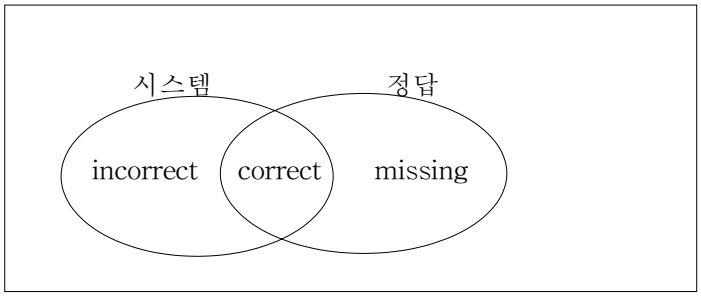


그림1에서 incorrect부분은 시스템이 잘못 생성한 링크이고 missing부분은 시스템에서 생성하지 못한 링크이다. incorrect 부분이 적을수록 단어정렬의 정확률이 높아지고 missing부분이 적을수록 단어정렬의 재현율이 높아진다. 본 논문에서는 정확률과 재현율을 모두 제고하기 위하여 incorrect 부분과 missing 부분의 링크를 모두 오류 링크로 보고 정확하게 수정하는 규칙을 생성하였다.

본 논문은 위에서 추출한 오류 링크의 집합에 나타나는 모든 한국어 기능어에 대하여 정답에서 표시한 대응되는 중국어 단어를 모두 찾아내어 한국어 형태소중국어 단어 쌍의 매치확률을 통계적으로 학습하였다.

대응되는 중국어 단어는 정답에 표시된 결과에 의하여 생성하기에 "NULL", "CONTENT", "실제 단어" 이 세 가지 타입을 가진다.

예제1에서 조사 "는"은 대응되는 중국어 단어가 없다. 때문에 이런 경우 한국어 형태소 "는"에 대응되는 중국어를 "NULL"로 표시하였다. 파생접사 "하"와 종결어미 "ㄴ다"는 중국어 단어 "學習"에 의미적으로 대응되었다. 이런 경우에는 대응되는 중국어 단어를 "CONTENT"로 표시하였다. 조사 "에서"는 대응되는 중국어 단어 "在"와 명확하게 대응된다. 이런 경우에 대응되는 중국어 단어는 정답에서 나타나는 단어 "在"를 그대로 사용하였다.

예제3: 수정규칙의 예

한국어 형태소(품사)	중국어 단어	매치확률
하(동사형 파생접사)	CONTENT	1.0
였(과거형 선어말어미)	了	1.0
다(평서형 종결어미)	了	1.0
에서(일반부사격 격조사)	在	0.7
에서(일반부사격 격조사)	中	0.3
는(보조사)	NULL	1.0

예제3는 정답을 기반으로 추출한 수정규칙의 예제이다. 수정규칙을 생성하기 위하여 본 논문에서는 품사를 고려한 한국어 형태소에 대하여 대응되는 중국어 단어와의 매치확률을 학습하였다. 수정규칙에서 표시한 매치확률은 한국어 형태소(품사)가 전체 정답 말뭉치에서 대응되는 중국어 단어를 여러 개 가질 때 그들 사이에 매치되는 상대적 확률이다. 이 확률은 위에서 추출한 한국어 형태소(품사)-중국어 단어 쌍의 빈도수로 계산된다. 예를 들면 한국어 형태소 "에서(일반부사격 격조사)"가 중국어 단어 "在"에 7번 매치되고 "中"에 3번 매치되었으면 그 확률은 각각 0.7과 0.3이다.

3.2 수정규칙의 적용

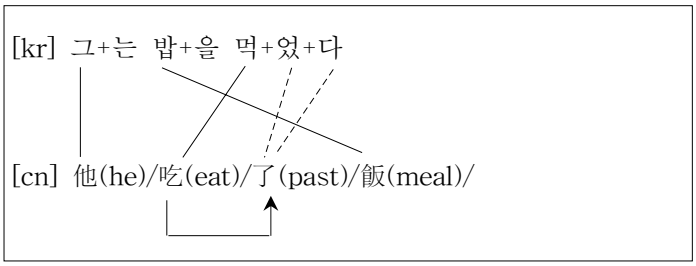
수정규칙은 기존 시스템에서 생성된 링크를 제거하거나 새로운 링크를 생성할 때 모두 사용된다. 본 논문에서는 수정규칙에 존재하는 한국어 형태소에 대하여 링크를 삭제하거나 생성하여 단여정렬의 결과를 수정하였다. 링크를 생성할 때 구문정보를 이용하는 방식은 Ma08에서 제안한 방법을 참조하여 사용하였다[3]

링크제거 단계에서는 수정규칙에 존재하는 모든 한국어 형태소에 대하여 기존 시스템에서 생성한 링크를 삭제하였다. 수정규칙이 틀린 링크를 기반으로 추출하였기에 오류가 생성될 가능성이 있는 링크를 미리 제거하면 단어정렬의 정확률을 높일 수 있기 때문이다.

링크를 생성하는 단계에서는 한국어 문장에서 수정규칙에 존재하는 한국어 형태소에 대하여 대응되는 중국어 단어를 연결해야 할 후보로 생성하였다. 다음으로 중국어 단어후보를 한국어 어절정보와 중국어 구문분석 정보를 이용하여 중국어 문장에서 찾아 아래의 규칙에 의하여 연결시켜 주었다.

- 만약 대응되는 중국어 단어후보에 "실제 단어"가 존재하면 매치확률이 가장 높은 "실제 단어"에 연결시켜 준다.
- 만약 대응되는 중국어 단어후보에 "실제 단어"가 없으면 대응되는 중국어 단어가 "CONTENT" 혹은 "NULL" 타입을 가지는 경우이다. 이 경우에는 "CONTENT"가 매치확률이 높으면 가장 앞쪽에 있는 "CONTENT"에 연결시켜 준다. "NULL"의 매치확률이 높은 경우에는 링크를 생성하지 않는다.

예제4: (영어단어는 각 중국어 단어의 뜻을 나타낸다)



예제4에서 실선은 신뢰할 수 있는 링크, 점선은 생성해야 할 링크, 화살표는 두 중국어 단어의 의존관계를 나타내는 구문분석 정보이다. 신뢰할 수 있는 링크는 GIZA++의 intersection결과를 사용하였고 그 상세한 설명은 4장 2절에서 기술한다.

수정규칙에 나타나는 대응되는 중국어 단어가 "실제 단어"이면 한국어 형태소가 포함된 어절정보, 중국어 단어의 구문분석정보를 이용하여 링크를 생성한다.

예제4에서 수정규칙에 나타나는 한국어 형태소 "었"에 대응되는 중국어 단어 "了"가 중국어 문장에 있으면 우

선 "었"을 포함하는 한국어 어절의 핵심 형태소 "먹"을 찾고 중국어 단어 "了"와 의존구문관계를 가지는 중국어 단어 "吃"도 찾는다. 다음 한국어 형태소 "먹"과 중국어 단어"吃"사이에 신뢰할 수 있는 링크가 있으면 수정규칙에 나타나는 형태소 "었"과 중국어 단어 "了"사이의 링크를 생성한다.

수정규칙에 나타나는 대응되는 중국어 단어가 "CONTENT"인 경우에는 한국어 형태소가 포함된 어절의 핵심 형태소가 신뢰할 수 있는 링크를 가지고 있으면 그 링크에 연결된 중국어 단어에 연결시켜 준다.

예제3에서 한국어 형태소 "하"는 수정규칙에서 대응되는 중국어 단어 "CONTENT"를 가진다. 이런 경우 우선 한국어 형태소 "하"가 포함된 어절의 핵심 형태소 "공부"를 찾고 "공부"가 신뢰할 수 있는 링크를 가지고 있으면 그 링크에 연결된 중국어 단어 "學習"와 연결시켜 준다.

4. 실험결과와 분석

4.1 말뭉치와 실험환경

본 논문에서는 동아일보[1] 신문기사를 말뭉치로 사용하였다. 말뭉치는 약 10만 문장 정도이며 문장 당 평균 단어가 약 30개 전후로 긴 문장들이 많다.

표1. 말뭉치 문장 수와 각 부분 평균 문장길이

	문장 수	중국어 평균 문장길이	한국어 평균 문장길이
Training	99226	28.05	31.61
Dev	500	29.86	33.47
Test	367	23.29	26.36

표1은 단어정렬과 기계번역에서 사용한 말뭉치의 구체적인 수치이다. 표1에서 Training은 학습 말뭉치, Test는 평가 말뭉치, Dev는 기계번역 시스템의 자질 값을 최적화[15] 하는데 사용된 말뭉치이다.

본 논문에서는 수정규칙을 추출 적용하기 위하여 본 연구실에서 개발한 한국어 형태소 분석기를 이용하여 한국어 문장을 형태소 단위로 분리하고 품사를 부착하였다. 중국어 문장은 Stanford[2]에서 개발한 중국어 구문분석기를 이용하여 문장을 단어별로 분리하고 구문분석을 진행하여 의존관계 구문정보를 추출하였다. 본 논문은 기계번역의 결과를 생성하기 위하여 MOSES[3]를 디코더로 사용하였다.

4.2 단어정렬 실험

본 논문에서는 GIZA++를 이용하여 생성된 단어정렬의 결과에 대하여 수정규칙을 적용하여 개선된 단어정렬의 결과를 얻어내었다. GIZA++에서 제공하는 단어정렬 알고리즘에는 여러 가지 옵션이 있으며 기본옵션은 grow-diag-final-and이다. 본 논문에서는 이 기본옵션과 intersection옵션을 사용하여 단어정렬의 결과를 각각 생성하였다. intersection 결과는 양방향으로 정렬된 두 결과의 교집합이고 grow-diag-final-and는 intersection 결과에서 출발하여 이웃하고 있는 가장 가까운 단어부터 차례로 링크시켜 주어 만든 결과이다. intersection결과는 높은 정확률을 가지고 있기[1]에 신뢰할 수 있는 결과로 사용하였다.

수정규칙은 링크제거, 링크생성, 링크제거+생성 등 세 가지 방식으로 적용하였다. 링크제거와 링크생성은 독자적으로 이루어진다. 링크제거 방법은 모두 기본옵션으로 생성된 단어정렬 결과에서 intersection결과를 제외한 나머지 부분에 대하여 적용하였다. 링크생성 방법은 기본옵션으로 생성된 단어정렬의 결과에 대하여 적용하였다.

단어정렬을 평가하기 위하여 사용된 말뭉치는 사람이 정답을 작업하여 만든 367문장의 평가 말뭉치이다. 본 논문에서는 AER[12]과 F-Measure[13]로 단어정렬의 성능을 평가하였다. AER값은 낮을수록 단어정렬의 결과가 좋다.

표2. 단어정렬의 성능

type	AER	precision	recall	F-measure
BASE	30.00	64.37	77.09	70.16
DE	28.12	79.62	65.28	71.96
IN	29.14	64.41	78.55	70.78
DE+IN	24.72	72.91	77.74	75.25

표2에서 BASE는 기존 시스템의 기본옵션으로 단어정렬을 한 결과이고 DE는 링크제거, IN은 링크생성, DE+IN은 링크제거를 실행하고 다시 링크생성을 한 결과이다.

수정규칙을 이용하여 단어정렬을 개선하면 AER로 평가되는 단어정렬의 결과는 모두 좋아진다. 하지만 단순히 틀린 링크를 제거하는 방법은 정확률은 제고하지만 재현율이 떨어진다. 반대로 링크를 생성하는 방법은 정확률이 떨어졌지만 재현율을 높이었다. 틀린 링크를 제거하고 다시 생성하는 방법은 정확률과 재현율을 모두 제고시켰고 가장 좋은 단어정렬의 결과를 보여주었다.

본 논문에서 사용된 수정규칙은 단어정렬을 평가하는 정답 말뭉치에서 추출하였다. 단어정렬의 결과를 객관적으로 평가하기 위하여 본 논문에서는 전체 정답 말뭉치를 10개 부분으로 나누어 교차평가(10-fold cross validation)하는 기법을 사용하였다. 기계번역에 사용된 수정규칙은 전체 정답 말뭉치에서 추출하였다.

4.3 기계번역 실험

기계번역 실험은 위에서 제안한 방법으로 단어정렬의 결과를 개선하여 구 추출에 사용하였다. MOSES시스템의 다른 옵션은 모두 기본으로 설정하여 사용하였고 학습문장과 테스트 문장은 단어정렬에서 사용하였던 것과 동일하다. 기계번역의 성능은 BLEU[14]로 측정하였다.

1) http://www.donga.com
2) http://nlp.stanford.edu
3) http://www.statmt.org/moses/

표3. 기계번역 성능

type	BLEU
BASE	21.38
DE	21.27
IN	21.52
DE+IN	**21.81**

기계번역 실험에서도 링크를 삭제하고 다시 생성하는 방법을 사용하였을 때 성능이 가장 좋았다. 링크를 삭제하는 방법만 적용했을 때는 단어정렬의 성능이 나아졌어도 기계번역의 성능은 오히려 떨어졌다. 링크를 생성하는 실험에서는 기계번역의 성능에서 뚜렷한 제고를 보이지 않았다.

4.4 단어정렬과 기계번역 결과의 비교분석

위에서 나타난 실험결과에서 알 수 있듯이 단어정렬의 결과를 하나의 값으로 표시한 AER과 F-Measure가 좋아진다고 꼭 기계번역의 성능이 좋아지는 것은 아니다. 링크제거 실험에서 단어정렬의 정확률을 높이고 재현율을 떨어뜨리면 기계번역의 성능은 떨어졌다. 또 링크생성 실험에서 단어정렬의 정확률을 떨어뜨리고 재현율을 높였을 때 기계번역의 성능은 뚜렷하게 제고되지 않았다. 하지만 링크제거+생성 실험에서 정확률과 재현율을 모두 높였을 때 기계번역의 성능은 좋아졌다. 이러한 결과에서 보았을 때 이후의 연구에서도 기계번역의 성능을 제고하기 위해서는 단어정렬의 정확률과 재현율을 모두 고려하여야 한다.

5. 결론

본 논문은 구 기반 통계기계번역 시스템에서 통계적으로 추출한 수정규칙을 이용하여 단어정렬의 결과를 개선하였고 기계번역의 성능을 제고하였다. 논문에서 제안한 수정규칙은 언어학적 지식을 바탕으로 통계적으로 추출하였기에 오류의 유형을 사람이 분석하여 대응되는 수정규칙을 제정하는 방법에 비하여 비교적 객관적으로 규칙을 제정할 수 있었다.

수정규칙을 생성하기 위해서는 정답이 표시되어 있는 말뭉치가 필요하다. 하지만 이런 말뭉치를 구축하는 작업은 기존의 시스템과 분리하여 진행할 수 있고 이런 말뭉치만 있으면 수정규칙을 자동으로 추출하여 시스템에 곧바로 반영할 수 있다. 때문에 이 방법은 수정규칙을 쉽게 보완하고 지속적으로 단어정렬의 결과를 개선할 수 있다.

이후의 연구에서는 기능어를 포함한 모든 오류에 대하여 수정규칙을 적용하는 작업을 하려고 한다. 또 단어정렬의 개선이 구 추출에 미치는 영향도 함께 연구할 것이며 이를 기반으로 최종적으로 구 기반 통계기계번역의 성능을 제고하려 한다.

6. 감사의 글

본 논문은 2009년도 두뇌한국21사업과 한국과학재단 기초연구사업(No. 2009-0075211)의 지원으로 수행되었습니다.

7. 참고문헌

[1] 김장호, 이금희, 나휘동, 이종혁: 한국어 형태소유형에 따른 한국어 중국어 단어정렬 결과분석. 2009.7.1-3 제주대학교. 2009한국컴퓨터종합학술대회. 논문지C pp. 325-330.

[2] Franz Josef Och & Hermann Ney: A systematic comparison of various statistical alignment models. Computational Linguistics 29 (1), pp.19-51.

[3] Yanjun Ma, Sylwia Ozdowska, Yanli Sun, & Andy Way: Improving word alignment using syntactic dependencies. Second ACL Workshop on Syntax and Structure in Statistical Translation (ACL-08 SSST-2), Proceedings, 20 June 2008, Columbus, Ohio, USA; pp.69-77.

[4] Yanjun Ma, Patrik Lambert & Andy Way: Tuning syntactically enhanced word alignment for statistical machine translation. EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation, ed. Lluís Màrquez and Harold Somers, 14-15 May 2009, Universitat Politècnica de Catalunya, Barcelona, Spain; pp.250-257.

[5] Victoria Fossum & Kevin Knight: Using bilingual Chinese-English word alignments to resolve PP-attachment ambiguity in English. AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai'i, 21-25 October 2008; pp.245-253.

[6] Josep M. Crego, Nizar Habash: Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT; Columbus, Ohio, USA, June 2008;Proceedings of the Third Workshop on Statistical Machine Translation, pp. 53-61.

[7] Fei Huang: Confidence Measure for Word Alignment: Proceeding of the 47th Annual Meeting fo the ACL and the 4the IJCNLP of the AFNLP, suntec, Singapre, 2-7 August 2009; pp. 932-940.

[8] ThuyLinh Nguyen & Stephan Vogel: Context-based Arabic morphological analysis for machine translation. CoNLL: proceedings of the Twelfth Conference on Computational Natural Language Learning, 16-17 August 2008, Manchester, UK; pp.135-142.

[9] Lee, Yong S Morphological Analysis for Statistical Machine Translation. In HLT-NAACL 2004; Boston, Massachusetts, USA; short papers pp. 57-60.

[10] Hermjakob Ulf: Improve Word Alignment with Statistics and Linguistic Heuristics; Proceeding of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6-7 August 2009. pp. 229-237.

[11] Jin-Ji Li, Dong-Il Kim, & Jong-Hyeok Lee: Annotation guidelines for Chinese-Korean word alignment. LREC 2008: 6th Language Resources and Evaluation Conference, Marrakech,Morocco,26-30 May 2008; pp. 7.

[12] Paul C. Davis, Zhuli Xie, & Kevin Small: All links are not the same: evaluating word alignments for statistical machine translation. MT Summit XI, 10-14 September 2007, Copenhagen,Denmark.Proceedings; pp.119-126.

[13] Alexander Fraser & Daniel Marcu: Measuring word

alignment quality for statistical machine translation. Computational Linguistics 33 (3), pp. 293-303.
[14] Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu: BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, July 2002; pp.311-318.
[15] Franz Josef Och: Minimum error rate training in statistical machine translation ACL-2003: 41st Annual meeting of the Association for Computational Linguistics, July 7-12, 2003, Sapporo, Japan.