문장 길이 축소를 이용한 구 번역 테이블에서의 병렬어휘 추출 성능 향상

정선이, 이공주
충남대학교 정보통신공학과
syjeong@cnu.ac.kr, kjoolee@cnu.ac.kr

Performance Improvement of Extracting Bilingual Term from Phrase Table using Sentence Length Reduction

Seon-Yi Jeong, Kong-Joo Lee
Dept. of Information and Communication Engineering, Chung-Nam University

요 약

본 연구는 대량의 특정 도메인 한영 병렬 말뭉치에서 통계 기반 기계 번역 시스템을 이용하여 병렬어휘를 효과적으로 추출해 낼 수 있는 방법에 관한 것이다. 통계 번역 시스템에서 어족이 다른 한국어와 영어간의 문장은 길이 및 어순의 차이로 인해 용어 번역 시 구절 번역 정확도가 떨어지는 문제점이 발생할 수 있다. 또한 문장 길이가 길어짐에 따라 이러한 문제는 더욱 커질 수 있다. 본 연구는 이러한 조건에서 문장의 길이가 축소된 코퍼스를 통해 한정된 코퍼스 자원 내 구 번역 테이블의 병렬어휘 추출 성능이 향상될 수 있도록 하였다.

주제어: 기계 번역, 구 번역 테이블, 용어 추출

1. 서 론

최근 통계기반의 기계 번역에 대한 연구가 활발히 진행되고 있다. 통계 기반의 기계 번역 프로그램인 MOSES[1]는 두 언어의 병렬 코퍼스를 이용하여 GIZA++[2]로부터 두 언어의 어휘의 번역 정보가 담긴 구 번역 테이블을 생성한다. 구절 테이블은 통계 기반 기계 번역에서 자동 번역을 구축할 시 어휘의 번역 정보로 이용된다. 그러므로 구 번역 테이블의 정확도 및 추출 성능은 통계 기반 기계 번역 시스템의 성능과도 직결된다. 본 논문에서는 통계 기반의 기계번역 프로그램인 MOSES를 이용하여 추출된 구 번역 테이블의 용어 추출 성능을 향상시키기 위한 연구를 시도하였다.

한영 기계번역의 경우 단어 번역(word translation) 과정뿐만 아니라 재배열(reordering)과정에서 긴 문장으로 인한 오류가 발생한다. 통계 정보만을 이용하는 단어 번역 단계의 경우 문장의 길이 및 어순 정보는 구 번역 테이블 내 적합한 번역어의 선택에 영향을 미칠 수 있다. 'SOV' 어순의 영어의 경우 문장의 길이가 주어와 서술어를 제외한 내용어들을 중심으로 길어진다고 할 때 서술어에 해당하는 영어의 동사와 한국어의 동사가 극단적으로 멀어질 수 있는 경우가 발생하게 된다. 이 경우 단어가 아닌 구절의 번역 정확도는 떨어질 수 있다. 이러한 문제점을 보완하기 위해 코퍼스와 동일한 도메인의 단어 및 구절 정보를 이용하여 문장 내 병렬 어휘 정보를 삭제하고 문장 길이를 축소하여 유효 번역 단어 간 거리를 줄여 용어 추출 성능을 높이고자 하였다. 본 연구는 기계 번역에서 사용되는 구절 번역 테이블을 활용한 용어 추출 시 용어 및 대역어 추출 성능을 향상시키기 위한 연구이다. 최종 평가 데이터로 구절 번역 테이블로부터 추출된 용어 셋을 사용함을 밝힌다.

논문 구성은 2 장에서 관련 연구를 설명하고 3장에서 본 논문에서 제안한 문장 길이 축소방법에 관해 설명한다. 4 장에서는 평가를 위한 실험환경 및 평가 환경을 설명하고 평가 셋에 대한 기존 결과와 제안 방법의 평가 결과를 보여준다. 5 장에서는 논문 내용을 결론짓고 향후 연구 방향에 대해 논의한다.

2. 관련 연구

통계 기반 기계 번역의 성능 향상을 위해 긴 문장에 대한 문제 인식을 통해 문장을 줄이거나 분할하려고 시도

표 1 영한 예시 문장

no.	길이	문 장
1	33	Recently, with the rapid spread of high-speed communication network the *population* using online photo print service is *rocking up* photograph taken by digital camera is transmitted via the internet and printed and delivered.
2	24	최근 **초고속 통신망이** 빠르게 보급됨에 따라 **디지털 카메라로** 찍은 사진을 인터넷에 전송한 후 사진을 인화하여 배송 받는 **온라인 사진 인화 서비스를** 이용하는 *인구*가 *급증하고 있다*.

했던 연구는 [3, 4, 5]가 있다. [3]의 연구에서는 통계 기반의 기계 번역 시스템에서 20 어절 이상의 긴 문장들을 보다 정확히 분석하기 위해 복수개의 의미 있는 절로 문장 분할을 시도하였다. 문장 내 분할 지점을 인식하기 위해 SVM을 사용하여 분할 지점의 특성을 학습한 후 분할 지점을 탐색한다. [4]의 연구에서는 입력 문장이 길어지면 통계 기반의 기계 번역 성능이 떨어짐을 지적하고 이를 완화하기 위해 긴 문장을 같은 의미의 짧은 문장들로 분할하여 기계 번역의 성능을 향상 시킬 수 있도록 하였다. 분할 방법으로는 변환 규칙을 학습하는 변환 기반 문장 분할 방법을 사용하였다. [5]에서는 긴 문장이 기계 번역에 미치는 영향에 대한 연구를 수행하였다. [5]의 연구에서는 문장의 길이가 길수록 기계 번역에서 발생하는 변형이 많아짐을 지적하고 문장의 길이가 길수록 번역의 성능이 떨어짐을 보였다. 이와 같이 문장 길이에 대한 문제 인식과 기계 번역 성능간의 연구는 존재하지만 긴 문장이 단어 번역 과정 및 구절 테이블에 미치는 영향에 관한 연구는 이루어지지 않고 있다.

3. 병렬 어휘 정보 삭제에 의한 문장 길이 축소

본 장에서는 구절 번역 테이블의 용어 추출 성능 향상을 위한 코퍼스의 문장 길이 축소 방법 및 효과를 설명한다.

3.1. 특정 도메인

기계 번역에서 이용할 문서 자원에서 특정 도메인 특성을 가지는 병렬 말뭉치의 경우 동일한 도메인 특성의 어휘정보가 말뭉치에 상당량 분포하게 된다. 이때 번역 어휘 쌍이 병렬 문장 쌍에서 동시에 존재하게 된다.

3.2 문장 길이 축소

같은 도메인을 공유하는 병렬 코퍼스의 경우 번역 어휘 쌍이 병렬 문장 쌍에서 병렬적으로 발견된다. 이러한 병렬 어휘 쌍을 이용하여 병렬 문장 쌍의 길이를 축소할 수 있다. 표 1은 병렬 코퍼스의 일부 정렬된 문장쌍이다.

문장 쌍 내부에 단일 어절 및 다중 어절의 어휘가 병렬적으로 존재한다. 이 중 단일 어절을 제외한 다중 어절 어휘의 길이를 1로 대치한다면 문장의 길이 축소 효과를 볼 수 있다. 문장 1의 경우 'high-speed communication network'와 'online photo print service', 'digital camera'가 길이 축소를 위한 병렬 어휘로 선택될 수 있다. 이에 해당하는 문장 2의 번역 어절은 '초고속 통신망', '디지털 카메라', '온라인 사진 인화 서비스'이다. 예시와 같이 문장 1과 문장 2에서 병렬 어휘로 대응된 2어절 이상의 어절을 하나의 어절로 축소시킴으로써 전체적인 문장 길이 축소 효과를 기대할 수 있다.

문장 축소 과정은 다음과 같다.

- 정렬된 병렬 어휘가 존재하는 사전을 준비한다.
- 코퍼스 내 모든 병렬 문장 쌍 내 존재하는 번역 어휘 쌍 을 탐색한다.
- 탐색된 병렬어휘는 다음과 같다.

$$Phs = \{Phs_i : i = 1 \cdots n\}$$

$$Pht = \{Pht_i : i = 1 \cdots n\}$$

- 탐색된 번역 어휘 쌍인 모든 Phs, Pht 에 대해 각 언어의 문장에서 '[MARK]'로 대치한다. (token length = 1)
- 원시 언어 문장 S의 길이가 $Len(s)$, 번역 언어 문장 T의 길이가 $Len(t)$ 이고 병렬 어휘 Phs_i, Pht_i의

각각의 길이가 $Len(Phs_i)$, $Len(Pht_i)$ 일 경우
S의 길이는

$$Len(\bar{s}) = Len(s) - \sum_{i=1}^{n}(Len(Phs_i) - 1)$$

T의 길이는

$$Len(\bar{t}) = Len(t) - \sum_{i=1}^{n}(Len(Pht_i) - 1)$$

로 축소된다.

문장 길이 축소를 위한 병렬 어휘 쌍은 기존의 대역어 사전을 사용하거나, 동일한 도메인의 MOSES의 초벌결과도 사용이 가능하다.

문장 길이 축소의 목적은 유효 단어 간 상이 어순에 의한 거리 차를 줄여 구절 테이블의 용어 추출 성능을 높이는 것이다. 또한 문장 구조 보존을 위해 병렬 어휘는 문장 내에서 삭제하지 않고 길이 1의 태그를 이용해 대치하였다. 표 1의 예시에서 다중 어절 병렬 어휘를 이용할 경우 문장길이는 각각 28, 20으로 축소된다.

문장 길이 축소에 의한 기대 효과는 다음과 같다. MOSES/GIZA++가 허용하는 다중 어절(phrase)의 최대 길이는 7이다[1]. GIZA++는 문장 내 7의 단어 길이 제한 내에서 두 언어의 구절의 적합한 번역어를 탐색한 뒤 통계정보를 이용하여 테이블을 만든다.

표 1의 문장 2의 '인구가 급증하고 있다'를 구 번역 테이블의 유효 다중 어절이라고 가정할 때 다음과 같은 번역어 후보들을 얻을 수 있다.

a	인구가 급증하고 있다 ||| the|DT **population|NN** using|VBG *online|NN photo|NN print|NN service|NN*
b	인구가 급증하고 있다 |||**population|NN** using|VBG *online|NN photo|NN print|NN service|NN* is|BE
c	인구가 급증하고 있다 ||| using|VBG *online|NN photo|NN print|NN service|NN* is|BE **rocking|VBG**

표 2 길이 축소 이전의 구 번역 테이블 예

e	인구가 급증하고 있다 ||| the|DT **population|NN** using|VBG *[MARK]|dummy* is|BE **rocking|VBG up|RB**
f	[MARK] 이용하는 인구가 급증하고 있다 ||| the|DT **population|NN** using|VBG *[MARK]|dummy* is|BE **rocking|VBG up|RB**

표 3 길이 축소된 문장의 구 번역 테이블 예

표 2의 결과를 보면 구 번역 테이블이 추출하는 엔트리 a, b, c 모두 어순 차와 내용어의 길이가 길어짐에 따라 '인구'의 유효 번역어인 'population'과 '급증하고 있다'의 유효 번역어인 'rocking up'을 동시에 포함하지 않고 있다. 구절 번역 테이블에서 이러한 엔트리들은 용어 추출 과정에서 정확한 번역어를 표시한 엔트리가 아니기 때문에 구 번역 테이블에서 용어를 추출할 때 정확도를 떨어뜨릴 수 있다. 반면 표 3의 e, f는 유효 거리 7이내에서 적합한 번역 어휘들을 모두 포함할 수 있게 된다. 이러한 과정을 통해 문장 길이가 축소된 코퍼스를 사용한 구절 테이블에서 노이즈 감소 효과와 새로운 용어 추출 효과를 얻을 수 있을 것이다.

4. 실 험

본 연구에서는 실험에 사용할 코퍼스로 '전자 뉴스' 영한 병렬 코퍼스를 사용하였다. 이 코퍼스의 특성은 전문적인 기술 용어 및 개체명(Named Entity)을 나타내는 어휘가 상당량 분포한다. 또한 다양한 길이의 문장이 존재한다. 실험에 사용한 코퍼스의 특성을 표 4에 나타내었다.

표 4 실험에 사용한 코퍼스 특성

코퍼스	문장수	평균 길이
한국어 (Baseline)	10000	21.16
영어 (Baseline)	10000	36.16
한국어(SLR)	10000	20.59
영어 (SLR)	10000	34.01

본 연구에서는 가장 코퍼스 특성에 부합되는 병렬어휘 정보를 사용하기 위해 같은 도메인의 다른 코퍼스로부터 얻은 MOSES 구 번역 테이블로부터 추출된 병렬어휘 정보를 사용하였다.

표 5 병렬어휘 사전 특성

엔트리	22040
참조된 병렬 어휘	3951
참조된 병렬 어휘 길이 평균 (영어)	2.32
참조된 병렬 어휘 길이 평균 (한국어)	1.22

표 5의 병렬 어휘 사전으로부터 코퍼스 내 병렬 문장에서 일치하는 병렬 어휘가 나타날 경우 모두 길이 1의 토큰으로 대치하고 길이를 축소할 수 있도록 하였다. 실험 결과 실험 코퍼스에 대해 참조된 병렬 어휘는 3951개로 나타났다. 참조된 병렬 어휘 길이의 평균은 영어와 한국어 각각 2.32와 1.22로 나타나는데 다중 어절의 영어 용어의 대역어가 한국어에서 단일 어절로 번역되거나 한국어 띄어쓰기 문제로 인해 영어에 비해 한국어 문장의 길이 축소 효과가 미비하게 나타났다. 표 6에 참조된 번역 어휘의 일부를 제시하였다.

표 6 참조된 병렬 어휘의 일부

embedded linux	임베디드 리눅스
bio industry	바이오산업
role-playing game	롤플레잉게임
offline company	오프라인 기업
wireless internet user	무선 인터넷 사용자
application software	응용 소프트웨어
location information	위치정보
web agency	웹에이전시
wireless lan market	무선랜 시장
bluetooth module	블루투스 모듈
embedded linux operating system	임베디드 리눅스 운용체계
shooting game	슈팅게임
online content	온라인 콘텐츠

실험 과정을 그림 1에 나타내었다.

실험에 사용한 구 번역 테이블의 추출 도구는 MOSES의 GIZA++를 사용하였다. 영어의 POS정보를 사용하기 위해 MOSES의 factored 모델을 이용하였다. 베이스라인 실험은 영어 쪽에만 POS정보가 부착된 한영 병렬 코퍼스를 GIZA++를 이용하여 구 번역 테이블을 추출하고 명사구 추출 필터를 이용하여 병렬 어휘를 추출한다.

명사구 추출 필터는 품사 정보를 이용하여 명사구만을 용어로 추출할 수 있도록 하였으며 구 번역 테이블의 확률 정보를 이용하여 번역 후보를 필터링 하였다.

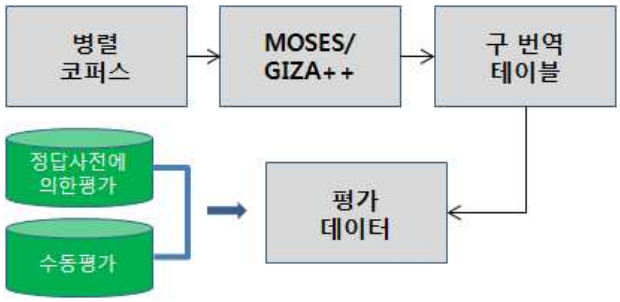

그림 3 데이터 추출 및 평가

제안한 방법의 실험은 기존의 코퍼스를 문장 길이 축소 방법을 이용하여 10000 문장쌍의 길이 축소가 적용된 새로운 병렬 코퍼스를 얻는다. 이 병렬 코퍼스 또한 영어 쪽에만 POS정보가 부착되어 있다. 베이스라인의 POS정보와 제안한 방법의 POS정보는 동일하다. 이 새로운 병렬 코퍼스와 기존 코퍼스를 합쳐 20000 문장쌍의 병렬 코퍼스를 생성한다. 생성된 20000 문장쌍의 병렬 코퍼스를 GIZA++를 통해 구절 번역 테이블을 얻고 명사구 추출 필터를 이용하여 병렬 어휘를 추출한다.

5. 평가 및 결과

추출된 병렬 어휘에 대한 평가는 기존 사전 용어들과 네이버 지식백과[6] 및 두피디아[7]의 웹 정보로 구축해 놓은 도메인 정답사전을 이용하였다. 그러나 정답 사전의 어휘는 제한되어 있으므로 추출된 모든 용어에 대한 정확한 정확률(precision) 값과 재현율(recall) 값을 얻을 수 없다. 그러한 문제의 보완을 위해 정답사전에 의한 평가와 정답 사전에 없는 엔트리는 수동 평가 결과로 평가를 진행하였다. 평가 대상이 될 엔트리는 코퍼스 내에서 빈도가 높아 통계 정보만으로도 정확한 번역 결과를 얻을 수 있는 어휘는 제외하고 빈도 20 이하의 엔트리들 중 명사구 필터를 통해 추출된 용어로 구성되었다. 추출 된 용어 가운데 길이 축소를 위해 참조된 병렬 어휘 사전에 포함되는 엔트리는 제외하였다.

표 7 Baseline과 제안 방법의 구 번역 테이블 크기 비교

구 번역 테이블	크 기
Baseline	43859
SLR	58757

표 8 추출된 병렬 어휘 결과

	Baseline	SLR
추출된 병렬어휘	2289	2443
다중 어절	1545	1767
단일 어절	744	676

표 9 정답사전에 의한 단일 어절 용어 추출 평가 결과

	Baseline	SLR	Baseline +SLR
Accuracy	62% (364/582)	58% (300/514)	52% (462/893)

표 10 정답사전의 의한 다중 어절 용어 추출 평가 결과

	Baseline	SLR	Baseline +SLR
Accuracy	66% (48/73)	73% (52/71)	62% (71/115)

표 11 수동평가에 의한 다중 어절 용어 추출 평가 결과

	Baseline	SLR	Baseline +SLR
Accuracy	68% (68/100)	76% (76/100)	72% (144/200)

실험 결과 Baseline과 제안방법(SLR)의 구 번역 테이블의 크기는 표 7과 같다. MOSES/GIZA++로부터 생성된 구 번역 테이블을 명사구 추출 필터를 이용하여 평가 데이터로 쓰일 용어를 추출하였다. 추출된 용어는 표 8과 같다. 추출된 용어 중 단일 어절과 다중 어절의 평가를 다르게 하기 위하여 데이터를 구분하여 평가하였다. 단일 어절의 경우 정답사전에 의한 평가만을 수행하였고 다중 어절의 경우 정답 사전에 의한 평가 엔트리가 적어 정답 사전에 의해 평가된 엔트리를 제외한 엔트리 중 임의로 뽑은 100개를 수동 평가 하였다. 정답 사전의 단일 어절 평가 결과 추출한 baseline과 SLR에서 각 0.62의 과 0.58의 정확률을 보이는 582개와 514개의 용어가 추출되었다. 다중 어절의 경우 Baseline보다 SLR에서 더 많은 정확한 엔트리를 추출하여 추출한 엔트리의 정확률 값이 0.73으로 0.66의 Baseline보다 높음을 보였다. 수동 평가의 경우 임의의 100개의 용어를 수동 평가 한 결과 각각 0.68 및 0.76의 정확률을 보였다. 자동 평가 방법에 의해 병렬 어휘를 평가한 경우 수동 평가에 의한 방법보다 정확도가 떨어짐을 알 수 있다. ‘전자 뉴스’ 코퍼스로부터 추출된 어휘는 다양한 상용적인 번역어를 포함하고 있다. 이 대역어가 정답 사전에 존재하는 대역어와 일치하지 않은 경우 정확도를 떨어뜨리게 된다. 예를 표 12에 나타내었다.

표 12 정답 사전과 매칭되지 않은 병렬 어휘

어휘	추출된 대역어	사전의 대역어
home automation system	홈오토메이션 시스템	홈자동화시스템
video memory	비디오메모리	비디오저장장치
video conference	화상회의	영상회의
security level	보안성	보안 수준

평가에 사용된 데이터 셋은 기존 방법과 SLR의 방법으로 뽑은 용어들 중 서로 중복되지 않는 용어들이 다수 포함되어 있다. 그러므로 Baseline의 방법 외에 SLR의 방법을 사용할 경우 Baseline에서 얻을 수 없는 단일 어절 및 다중 어절 용어들을 확보할 수 있음을 증명하였다. Baseline의 방법과 SLR의 방법을 병행하여 사용할 경우 Baseline만을 사용할 경우보다 한정된 코퍼스로부터 좀 더 풍부한 어휘 추출을 기대할 수 있다. 다만 단일어절과 다중어절 Baseline과 SLR의 데이터 특성이 다르고 실험 결과의 정확도 및 재현율 등의 값을 고려할 때 각 방법과 다중 어절 및 단일 어절 데이터 셋을 추출하는 과정에서 명사구 추출 필터에 최대한 정확도가 좋은 용어 선별을 위한 각 데이터에 맞는 휴리스틱 적용이 요구된다.

6 . 결 론

문장 길이가 축소된 코퍼스를 이용하여 통계 기반의 기계 번역 시스템의 구절 테이블의 용어 추출 성능을 높이는 실험과 그에 따른 결과를 제시하였다. 그 결과 베이스라인에서 얻을 수 없는 코퍼스의 용어가 다량 추출되었고 그의 정확도도 나쁘지 않음을 실험 결과에서 알 수 있다. 실험 결과를 통해 통계 기반이 기계 번역 시스템에서 한정된 코퍼스 자원을 가지고 보다 풍부한 어휘 정보의 번역을 요구할 때 제안한 방법이 도움을 줄 수

있을 것으로 기대한다. 또한 실험에 나타난 정확도와 재현률 값의 보완을 위해 용어 추출 시 사용되는 필터에 각 데이터 특성에 맞는 휴리스틱 적용이 필요하다.

참고문헌

[1] http://www.statmt.org/moses/
[2] http://www.statmt.org/moses/giza/GIZA++.html
[3] 김유섭, "지지 벡터 기계를 이용한 긴 문장의 효과적인 분할", 10-19, 한국정보기술학회논문지, 2007.
[4] 이종훈, 이동현, 이근배, "통계적 기계 번역을 위한 변환 기반 문장 분할 방법", 한국정보과학회 언어공학연구회 학술발표논문집, 276-281,2007.
[5] 조희영, 서형원, 김재훈, "문장 길이가 한영 통계기반 기계번역에 미치는 영향 분석", 한국정보과학회 학술발표논문집 34(1C), 199-203, 2007.
[6] http://terms.naver.com/
[7] http://www.doopedia.co.kr/