한베 통계기계번역의 성능 향상을 위한 내포문 추출 및 복원 기법

조승우[O+], 김영길[#], 권홍석[+], 이의현[+], 이원기[+], 조형미[+], 이종혁[++]
포항공과대학교 컴퓨터공학과[+], 한국전자통신연구원[#]
{itswc[O+], hkwon[+], jekyllbox[+], wklee[+], hyungmi[+], jhlee[++]}@postech.ac.kr, kimyk[#]@etri.re.kr

Embedded clause extraction and restoration for the performance enhancement in Korean-Vietnamese statistical machine translation

Seung-Woo Cho[O+], Young-Gil Kim[#], Hong-Seok Kwon[+], Eui-Hyun Lee[+],
Won-Ki Lee[+], Hyung-Mi Cho[+], Jong-Hyeok Lee[++]
Pohang University of Science and Technology, Department of Computer Science & Engineering[+]
Electronics and Telecommunications Research Institute[#]

요 약

본 논문에서는 기호로 둘러싸인 내포문이 포함된 문장의 번역 성능을 높이는 방법을 제안한다. 입력 문장에서 내포문을 추출하여 여러 문장으로 나타내고, 각각의 문장들을 번역한다. 그리고 번역된 문장들을 복원정보를 활용하여 최종 번역 문장을 생성한다. 이러한 방법론은 입력 문장의 길이를 줄여주며, 그로 인하여 문장 구조가 단순해져 번역 품질이 향상된다. 본 논문에서는 한국어-베트남어 통계 기반 번역기에 대하여 제안한 방법론을 적용하고 실험하였다. 그 결과 BLEU 점수가 약 1.5 향상된 것을 확인할 수 있었다.

주제어: Statistical Machine Translation(SMT), MOSES, Sentence simplification

1. 서론

최근 들어, 국가와 기업뿐만이 아니라 개인 차원에서도 외국과 소통할 수 있는 기회가 많아지면서 서로 다른 언어 간 번역의 필요성이 크게 늘어났다. 특히 베트남은 최근 한류열풍과 사회 기반 정책의 변경으로 경제가 크게 발전해 대한민국와의 교류가 날로 늘어나고 있다. 하지만 그동안 한국어-베트남어 간 기계번역 기술은 다른 언어쌍에 비해 주목받지 못했고 진행된 연구가 미비한 실정이다.

더욱이 베트남어는 한국어와 문법적으로 크게 다른 언어라 번역을 더욱 어렵게 만든다. 특히 어순이 한국어와 매우 다른데, 한국어는 Subject-Object-Verb(SOV) 언어이며 앞에서 수식하고 상대적으로 어순이 자유로운 언어다. 반면, 베트남어는 Subject-Verb-Object(SVO) 언어이며 뒤에서 수식하고 어순이 고정된 언어이다. 일반적으로 통계기계번역 성능은 원시언어와 대상언어가 문법적으로 다를수록 떨어지고 문장의 길이가 길어질수록 어순 조정이 힘들게 되어 성능이 떨어진다. 한국어와 베트남어는 어순 자체의 차이가 크기 때문에 이를 보완할 수 있는 방법론이 필요하다.

문법적인 구조 및 어순 차이에서 발생할 수 있는 어려움을 줄여주기 위해 문장의 길이를 짧게 만드는 문장 단순화 기법이 기계번역에 도입되어 왔다. 본 연구는 이러한 문장 단순화 기법의 일환으로 기호로 둘러싸인 내포문(직접인용문, 부연설명문, 강조문 등)에 대해 문장 분리를 함으로써 문장길이를 줄여 번역 성능을 높이려고 한다. 또한 동음이의어 문제를 해소하기 위해 품사를 활용한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 문장 분할 기법과 품사 활용법에 대한 관련 연구를 서술한다. 3장에서는 제안하는 문장 분할 알고리즘과 품사 정보 활용법을 설명한다. 4장에서는 제안한 방법론을 한국어-베트남어 번역 시스템에 적용하여 평가한다. 마지막으로 5장에서는 결론을 맺는다.

2. 관련 연구

통계기계번역에서는 성능 향상을 위해 다양한 문장 분할 기법들이 연구되어 왔다. [1]은 영어-힌두어 병렬 말뭉치 단어 정렬 성능 향상을 위해서 각 언어 말뭉치에서 나타나는 Cue word들을 활용했다. Cue word는 영어의 Because 같은 연결 표현으로 문장에서 발생한 담화 정보를 확장시켜 준다. 이를 이용하여 긴 문장에서 존재하는 내포된 구와 절(Clause)들을 추출하여 단어 정렬의 성능을 향상시켰다. [2]에서는 일본어 문장 내 형태소 품사 정보와 마침표 위치 정보를 활용하여 문장을 여러 개 절로 분할한다.

기존 문장 분할 기법을 통계기계번역에 적용하려면 아직 많은 문제점들이 존재한다. 문장 분할 기법은 대부분 원시언어에만 집중적으로 적용되어 왔다. 그러나 통계기계번역 시스템의 번역 모델은 대용량 병렬 말뭉치를 통해 구 단위 대역 정보를 학습한다. 원시언어 문장과 목표언어 문장의 대역 관계가 유지된 상태에서 번역 모델을 학습시켜야만 올바른 번역 모델을 생성해 낼 수 있다. 그렇기 때문에 원시언어에서 나타나는 특성만으로는 문장 분할의 경계를 설정하기 어려우며 양 쪽 언어에서 발견할 수 있는 공통적인 특성에 주목해야 한다.

더욱이, 통계기계번역에서는 분할된 원시언어문장의 원래 위치를 별도로 표기하는 것만으로는 완벽한 원문 복원을 기대할 수 없다. 원시언어 말뭉치에서 얻어낸 위치 정보를 번역된 목표언어 말뭉치에서 사용할 수 있다고 보장할 수 없기 때문이다. 따라서 복원 과정도 양방향으로 진행되어야만 한다.

3. 방법론

본 논문에서 제안하는 내포문의 추출은 몇 가지 사전 정의된 특수 기호들을 기준으로 수행된다. 그 종류는 아래 표 1과 같으며, 한국어-베트남어 번역기 구축을 위한 병렬 말뭉치로부터 자주 등장하는 기호들로 선정되었다.

표 1. 내포문 추출을 위한 특수 기호

여는 기호	“, ‘, <, {, [, (, 「, 『
닫는 기호	”, ’, >, },],), 」, 』

3.1 내포문 추출

내포문을 포함하고 있는 문장들은 다양한 형태로 표현이 가능하다. 단순히 내포문이 하나만 있지 않고 그 안에 또 다른 내포문이 존재할 수도 있기 때문에 재귀적으로 이를 처리할 수 있어야 한다. 또한, 내포문 추출 이후의 병렬 말뭉치 대역 상태를 보존하기 위해 원시언어 문장에서 추출된 내포문 개수와 목표언어 문장에서 추출된 내포문 개수를 대조한다. 대조 결과가 같지 않다면 내포문 추출 작업을 취소하고 원문을 사용하도록 한다.

추출 알고리즘은 스택 자료구조를 사용해서 간단하게 구현한다. 앞에서부터 문장을 읽어나가다 여는 기호를 보면 여는 기호 위치를 스택에 저장하고 닫는 기호를 만나면 스택에서 가장 최근에 저장한 여는 기호 위치를 가져와 내포문 내용을 추출한다.

내포문이 추출된 이후에는 병렬 말뭉치에 등장하지 않은 특수기호로 추출 위치를 표기한다. 또한, 기계 번역 이후에 진행될 복원 과정을 위해서 원문과 추출된 문장의 줄 번호를 기록한다. 아래 그림 1은 추출 결과에 대한 간단한 예제를 보여준다.

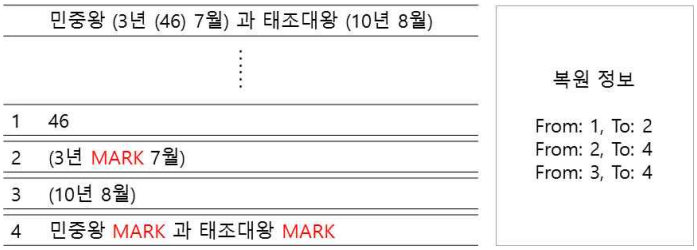

그림 1 내포문 추출 결과 예제

3.2 내포문 복원

실험 말뭉치에 대한 번역을 완료한 후, 원래 분할되었던 문장들을 다시 하나로 복원 하는 과정이 필요하다. 번역 성능을 평가할 수 있어야 하므로 목표언어 말뭉치 뿐만이 아니라 원시언어 말뭉치도 복원할 수 있는 알고리즘이 제시되어야 한다.

먼저 원시언어 말뭉치에서 내포문을 복원할 때는 몇 번째 줄의 추출 기호가 몇 번째 줄의 내용으로 치환되어야 하는지에 대한 정보가 필요하다. 그림 2은 원시 언어에서의 복원 과정을 나타내고 있다. 파란색 글씨가 복원 순서를 의미하며 빨간색 글씨가 추출 기호이다.

그림 2 원시언어 문장 복원 예제

목표언어 말뭉치에서의 내포문 복원은 다소 복잡하다. 원시언어 문장에서 남긴 추출 기호가 번역 과정에서 사라질 수도 있다. 또한, 원시언어 문장에서 다수의 추출 기호가 존재한다면 추출된 내포문과 추출 기호를 대역하는 것이 쉽지 않다. 따라서 원시언어 문장에 존재하던 추출 기호가 목표언어의 어떤 표현으로 번역되었는지 혹은 번역되지 않았는지 확인할 수 있어야 한다. 추출 기호의 번역 경로를 따라가기 위해 MOSES[3]가 생성해주는 단어 정렬 정보를 활용하기로 했다. MOSES의 단어 정렬 표시 형태는 그림 3과 같다. MOSES에서 사용하는 구분자는 붉은 원으로 표시하였으며 구분자 오른쪽에 있는 표기 사항이 단어 정렬 정보이다. 만약, 단어 정렬 정보가 0-0이면 원시언어 문장의 1번째 표현이 목표언어 문장의 1번째 표현으로 번역되었음을 의미한다.

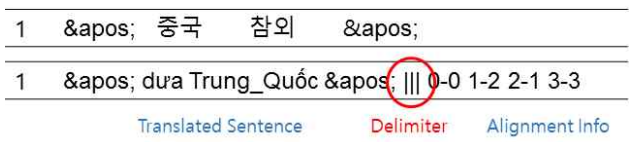

그림 3 MOSES 단어 정렬 표시 예제

따라서 목표언어 말뭉치에서 내포문 복원은 크게 2가지로 나눌 수 있다. 원시 언어의 추출 기호가 목표 언어로 번역된 경우와 그렇지 않은 경우이다. 추출 기호가 목표 언어로 번역된 경우라면, 먼저 원시언어 문장의 추출 기호 위치를 확인한다. 그리고 원시언어 문장과 대역되는 목표언어 문장의 단어 정렬 정보를 확인한다. 마지막으로, 추출 기호를 복원 정보에 표기된 문장으로 치환하면 복원이 완료된다. 그림 4는 원시 언어의 추출 기호가 목표 언어의 한 표현으로 번역되었을 때의 예제이다.

그림 4 추출 기호가 목표언어로 번역되는 예제

만약, 추출 기호가 목표 언어로 번역되지 않았다면 추출 기호의 단어 정렬 정보를 사용할 수 없으므로 복원 과정을 진행할 수 없다. 따라서 일정 문맥 크기를 설정하고 주위 단어의 번역 결과를 이용하여 복원 과정을 진행한다. 한국어의 수식 방향을 고려하여 문맥 범위 안에 있는 추출 기호의 왼쪽 단어들이 먼저 번역되었는지 확인한다. 만약, 왼쪽 단어들이 모두 복원되지 않았을 경우에는 문맥 범위 내 추출 기호의 오른쪽 단어들을 확인한다. 주위 단어가 번역되었음을 확인한다면 복원 정보를 이용해서 복원해준다. 그림 5은 추출 기호가 목표언어로 전혀 번역되지 않을 때의 예제이다.

그림 5 추출 기호가 목표언어로 번역되지 않는 예제

4. 실험 및 결과

본 논문에서는 한국어-베트남어 통계기계번역 시스템을 MOSES로 구축하고 제안한 방법론을 적용하여 실제 번역 성능을 향상시키는지에 대한 실험을 수행하였다.

4.1 말뭉치 구성

한국어-베트남어 번역기 구축 및 성능 평가를 위하여 표 2과 같이 말뭉치를 구성하였다. 먼저 학습 말뭉치는 통계기계번역의 파라미터들을 학습하는데 사용되어지고, 개발 말뭉치는 학습된 파라미터들을 세부 조정하는데 사용되어진다. 그리고 실험 말뭉치는 번역기의 성능을 평가하는데 사용되어지며 학습 말뭉치와는 중복 없이 구성되었다. 또한 제안 방법론의 실효성을 평가하기 위하여 내포문이 포함된 문장만으로 구성되었다.

표 2. 초기 말뭉치 통계 수치

종류	문장 수	평균 단어 수
한국어 학습 말뭉치	936,267	16.10
한국어 개발 말뭉치	1,000	31.61
한국어 실험 말뭉치	1,000	31.24
베트남어 학습 말뭉치	936,267	13.28
베트남어 개발 말뭉치	1,000	26.11
베트남어 실험 말뭉치	1,000	27.28

실험 말뭉치에서 추출된 내포문들의 통계는 표 3으로 정리하였다. 내포문들이 원문의 평균 단어 수와 비교하여 비교적 적은 단어들을 가지고 있음을 알 수 있다.

표 3. 실험 말뭉치 내포문 통계 수치

종류	문장 수	평균 단어 수
한국어 실험 말뭉치	1,251	7.08
베트남어 실험 말뭉치	1,251	6.47

4.2 말뭉치 전처리

본 논문에서는 형태소 번역 단위를 가지는 번역기를 구축하였다. 한국어는 포항공과대학교 지식 및 언어 공학 연구실의 한국어 형태소 분석기(KoMA)[4]를 사용하였고, 베트남어는 호치민 대학교의 단어 분리기[5]와 품사 태거[5]를 사용하였다. 표 4는 실험에서 사용한 전처리 분석기들의 성능이다.

표 4. 전처리 분석기 성능

분석기	Precision
KoMA	93.1%
CLC 단어 분리기	99.28%
CLC 품사 태거	96.55%

또한 동음이의어의 번역 애매성을 해소하기 위하여 [6]에서와 같은 방법으로 품사 정보를 활용하였다. 그림 6은 형태소 품사 정보를 활용하는 예제를 보여준다. 첫 번째 줄은 원문, 두 번째 줄은 형태소 분석만 처리된 문장, 마지막 세 번째 줄은 품사 정보까지 활용한 문장이다.

Original text	가격은 얼마입니까
Analyzed text	가격 은 얼마 이 ㅂ니까
PoS tagged text	가격/CMCN 은/fjb 얼마/CMCN 이/fpd ㅂ니까/fmofq

그림 6 품사 정보 활용 예제

4.3 실험 결과

기계 번역 이후, 원시언어 문장의 추출 기호가 어떻게 번역되었는지 확인하고 알고리즘을 달리 적용하여 목표 언어 문장을 복원했다. 번역 결과에서 추출 기호가 번역된 문장, 추출 기호가 번역되지 않고 사라진 문장들을 세어 표 5에 표시하였다. 추출된 내포문 수에 비하여 크지 않은 숫자지만 번역 과정에서 추출 기호가 없어지는 문장이 있음을 확인하였다.

표 5. 전처리 분석기 성능

종류	문장 수
추출 기호 번역	1,169
추출 기호 상실	82

실험을 통해 내포문 추출 및 복원과 품사 정보 활용 기법에 대한 성능을 평가하였다. 평가 방법으로는 BLEU[7] 점수를 사용하였다. 아래 표 6은 여러 실험 조합에 대한 실험결과를 나타낸 것이다.

표 6. 실험 케이스 및 실험 결과

	말뭉치 전처리 적용 사항	BLEU
Case 1	Baseline	32.96
Case 2	품사 정보 활용	32.48
Case 3	내포문 추출 및 복원	34.36
Case 4	내포문 추출 및 복원 + 품사 정보 활용	**34.55**

Case 1은 단순히 각 언어에 대하여 형태소 분석 및 단어 분리기만을 적용하여 평가한 결과이다. Case 2는 baseline에 품사 정보를 부착하여 실험한 결과를 나타낸다. Case 3은 baseline에 내포문 추출 및 복원을 적용한 실험이며, 마지막 Case 4는 baseline에 내포문 추출 및 복원과 품사 정보까지 적용된 실험 결과를 보여준다.

먼저 Case 2의 실험 결과를 보면 Case 1에 비하여 BLEU 점수가 약 0.5 정도 하락한 것을 확인 할 수 있었다. 그 이유를 분석하여 보면 내포문이 포함된 문장들의 경우 한국어의 형태소 분석기 및 베트남어의 단어 분리기의 성능이 좋지 않아 오히려 성능을 하락시키는 요인이 되었기 때문이다.

Case 3의 결과를 보면 baseline 대비 약 1.5 정도 BLEU 점수가 향상된 것을 확인할 수 있었다. 따라서 제안한 방법론이 내포문을 포함한 문장의 구조를 단순화시키고 그에 따라 번역 시스템의 여러 자질(Feature)학습(어순 조정, 단어 정렬 등)에 긍정적인 효과를 보여준 것이라 볼 수 있다.

마지막 Case 4의 실험결과를 보면 Case 2와는 다르게 품사 정보를 활용하였을 때 더 좋은 성능을 낸 것을 확인하였다. 이는 Case 3에서의 경우와 마찬가지로 내포문 추출로 인한 문장 구조의 단순화가 말뭉치 전처리에 도움이 되었고 그 결과로 번역 품질 또한 향상되었다고 생각할 수 있다.

5. 결론

본 논문은 통계기계번역 시스템의 성능 향상을 위한 내포문 추출 및 복원 알고리즘을 제안하였다. 그리고 제안한 방법론이 실제로 번역 시스템의 성능을 향상시킨다는 것을 실험으로 확인하였다. 또한 특정 언어쌍에만 적용할 수 있는 것이 아니라 언어쌍에 독립적으로 적용할 수 있다는 장점이 있다.

그러나 내포문을 가지고 있는 문장들의 수가 말뭉치 전체 문장 수와 비교하여 상대적으로 적고 추출 기호를 원문에 삽입하는 것이 부자연스러운 원문을 만들어내므로 아직 완벽하지 않다. 또한, 특수 기호에 대한 단어 정렬 성능을 확인하지 않았기 때문에 이를 검증할 필요가 있다.

향후 과제로는 문장 단순화(text simplification) 방법을 활용하여 내포문뿐만 아니라 일반 문장도 처리할 수 있게 적용 범위를 넓히고 추출 기호를 생략하면서 내포문을 복원할 수 있는 알고리즘을 실험할 계획이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신 · 방송 연구개발사업(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발), ICT명품인재양성사업(R0346-16-1007) 및 (주)시스트란인터내셔널의 지원을 바탕으로 수행하였습니다.

참고문헌

[1] Srivastava, Jyoti, and Sudip Sanyal, Segmenting long sentence pairs to improve word alignment in english-hindi parallel corpora, Advances in Natural Language Processing, pp.97-107, 2012

[2] Goh, Chooi-Ling, and Eiichiro Sumita, Splitting Long Input Sentences for Phrase-based Statistical Machine Translation, Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing, pp.802-805, 2011

[3] Koehn, Philipp, et al, Moses: Open source toolkit for statistical machine translation, Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pp.177-180, 2007

[4] 권오욱, et al, 음절단위 CYK 알고리즘에 기반한 형태소 분석기 및 품사태거, 1999 년도 제 11 회 한글 및 한국어 정보처리 학술대회 및 제 1 회 형태소 분석기 및 품사태커 평가 워크숍, pp.76-87, 1999

[5] Dien, Dinh, and Vu Thuy, A maximum entropy approach for Vietnamese word segmentation, 2006 International Conference onResearch, Innovation and Vision for the Future. IEEE, 2006.

[6] Lee, Jonghoon, Donghyeon Lee, and Gary Geunbae

Lee, Improving phrase-based Korean-English statistical machine translation, INTERSPEECH, 2006

[7] Papineni, Kishore, et al, BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th annual meeting on association for computational linguistics, pp.311-318, 2002