말뭉치 자동 확장을 통한 SMT 성능 향상에 대한 연구

최규현○, 신종훈, 김영길
과학기술연합대학원대학교○, 한국전자통신연구원, 한국전자통신연구원
choko93@ust.ac.kr○, jhshin82@etri.re.kr, kimyk@etri.re.kr

Research about SMT Performance Improvement Through Automatic Corpus Expansion

Gyu-Hyun Choi○, Jong-Hun Shin, Young-Kil Kim
University of Science and Technology○
Electronics and Telecommunication Research Institute

요 약

현재 자동번역에는 통계적 방법에 속하는 통계기반 자동번역 시스템(SMT)이 많이 사용되고 있지만, 학습 데이터로 사용되는 대용량의 병렬 말뭉치를 수동으로 구축하는데 어려움이 있다. 본 연구의 목적은 통계기반 자동번역의 성능을 향상시키기 위해 기존 다른 언어쌍의 말뭉치와 SMT 자동번역 기술을 이용하여 대상이 되는 언어쌍의 SMT 병렬 말뭉치를 자동으로 확장하는 방법을 제안한다. 제안 방법은 서로 다른 언어 B와 C의 병렬 말뭉치를 얻기 위해, A와 B의 SMT 자동번역 시스템을 구축하고 기존의 A-C 말뭉치의 A를 SMT를 통해 B로 번역하여 B와 C의 말뭉치를 자동으로 확장한다. 실험을 통해 확장한 병렬 말뭉치가 통계기반 자동번역 시스템의 성능을 향상시킬 수 있음을 확인한다.

주제어: Parallel 말뭉치, Pivot, Pseudo 말뭉치, 병렬 말뭉치, 통계 기계 번역

1. 서론

자동번역(Machine Translation) 기술은, 의존하는 언어 지식의 유형과 그 활용 방법에 따라 크게 규칙 기반 기법과 말뭉치(Corpus) 기반 기법으로 나누어진다. 말뭉치 기반 기법인 경우 지식기반과 통계기반 기법으로 나누어 볼 수 있는데, 현재 통계기반 기법으론 통계기반 자동번역 시스템 SMT(Statistical Machine Translation)가 많이 사용되고 있다. SMT를 기반으로 한 자동번역 시스템은 두 언어사이의 단어, 구(Phrase), 문법 등에 대한 통계 모델을 기반으로 구축될 수 있다.

통계기반 자동번역은 번역 대상의 언어에 종속된 문법 및 규칙에 의존하지 않고 각 언어의 번역 결과에 대한 통계 정보를 사용하기 때문에 규칙기반 자동번역보다 개발 시간을 단축할 수 있고 특정 언어에 대해 의존할 필요가 없어서 어떤 언어 쌍에도 사용할 수 있다는 장점이 있다. 하지만, 의미있는 통계정보를 추출할 수 있을 만큼의 대용량 병렬 말뭉치가 필요하다는 단점이 있다. 대용량 병렬말뭉치를 수동으로 구축하는 것은 시간적으로 비용적으로 어려운 일이다. 적은 양의 데이터를 사용하여 통계기반 자동번역 시스템을 구축하게 되면 번역 성능이 떨어지게 된다. 따라서 본 연구에서는 통계기반 자동번역에 사용될 말뭉치를 자동으로 확장할 수 있는 방법에 대해 소개한다.

제안 방법은 서로 다른 언어 B와 C의 병렬 말뭉치를 얻기 위해 (1)A와 B의 SMT를 구축하는 단계 (2)A와 C의 말뭉치를 구축한 SMT를 통해 A를 B로 번역하는 단계, (3)최종적으로 B와 C의 말뭉치를 구축하는 단계로 구성된다. 그리고 구축한 말뭉치를 활용하여 통계 기반 자동번역 시스템의 성능에 어떤 영향을 주는지 확인한다.

본 논문은 MOSES 시스템을 통계 기반 번역 시스템으로 사용하였으며[1], BLEU (Bilingual Evaluation Understudy)를 이용해 번역 성능을 평가하였다[2]. BLEU는 IBM 연구소에서 제안한 번역 성능 평가 측도이며, n-gram의 공기빈도를 이용하여 번역의 질을 측정하는 방법으로 현재 자동번역의 성능 측정에서 널리 사용되고 있다.

본 논문의 2장에서 말뭉치 구축에 대한 관련 연구에 대해 소개하고, 3장에서 제안하는 말뭉치 구축 방법을 설명한다. 4장에서는 제안한 방법으로 구축한 말뭉치를 사용하여 실험한 결과를 살펴보고 5장에서 결론을 설명한다.

2. 관련 연구

병렬 말뭉치를 확장하는 연구는 앞서 다양하고 많은 방법으로 연구가 진행되었다. 특히, 논문과 관련되어 있는 기술인 Pivot Language 기술을 사용하여 병렬 말뭉치

를 구축하는 기술이 많이 연구되었으며, 기본적으로 잘 알려진 3가지 종류의 구축방법이 있다.

- Transfer Method (Sentence Pivoting, Sentence Translation Strategy)
- Triangulation Method (Phrase Pivoting, Phrase table Multiplication)
- Synthetic Method (Pseudo-corpus approach)

Transfer 방법은 두 개의 분리된 자동번역 시스템이 존재하고, 이를 사용하여 단순히 A를 B로 번역 한 뒤, B를 C에 대한 번역을 차례로 수행 하여 A-C를 수집하는 방법이다. 그러나, Transfer 방법에서는 서로 다른 도메인으로 구축된 번역 시스템을 사용될 경우, 번역 오류가 전파되어 완성된 병렬 말뭉치의 성능에 문제가 발생할 수 있다.

Triangulation 방법은 A-B와 B-C의 Phrase Table을 병합 후 정제(Filtering)하면서 질 좋은 A-C의 Phrase를 수집하는 방법이다. [3]는 페르시아어-아랍어의 Phrase Table을 획득하기 위해 이 방법을 사용하였다. 그리고 Triangulation 방법은 정제를 통해 생성된 Phrase Table의 크기가 기하급수적으로 커진다는 문제점이 있으며, 이를 해결하기 위해 별도의 작업이 필요하다.

Synthetic 방법은 A-B와 B-C 말뭉치가 있을 때, B-C로 MT를 구축하고 A-B의 B부분을 번역하여 A-C 말뭉치를 수집하는 방법이다. [4]는 카탈로니아어-스페인어, 스페인어-영어를 사용하여 이 방법의 성능을 증명하였다. Synthetic 방법으로 구축한 말뭉치를 의사(Pseudo) 말뭉치라고 부른다.

문서 단위를 기반으로 같은 주제로 정렬된 비교 가능 말뭉치(comparable corpus)를 얻는 방법들도 많이 연구되었다. [5]는 한•중 병렬 말뭉치를 구축하기 위해서 중국어 번역문을 서비스하는 국내의 일간지와 인터넷 기사를 활용하였다. [5]가 제시하는 방법은 자동 수집이 가능한 언론사를 대상으로 중국어 번역문이 있는 사설과 칼럼을 언론사 사이트에서 제공하는 자동 수집 버튼을 사용해 단순하게 수집하였다. HTML(HyperText Markup Language)문서를 활용한 말뭉치 구축에 대한 연구에서 [7][8]은 웹상에서 번역문서 후보를 추출한 다음 HTML문서 구조를 비교하여 번역문서인지를 판별하고 문장 단위 정렬을 이용하여 병렬 말뭉치를 구축하는 방법을 제시하였다.

외부 인터넷에서 말뭉치를 수집하는 또 다른 연구로는 최근 위키피디아를 사용하는 연구들이 많다. 광범위한 분야에서 위키피디아를 원천 자원으로 하여 병렬 문장을 자동으로 추출하는 연구들이 진행되고 있다. 위키피디아는 많은 사람들이 공유하고 자원 역시 공개되어 있다. 위키피디아 데이터는 하나의 주제에 대해 여러 언어로 설명되어 있기 때문에 비교 말뭉치로서 중요하게 활용될 수 있지만, 대용량의 병렬 말뭉치를 수집하기는 어렵다.

3. 제안 방법

이 장에서는 Synthetic Method를 활용한 말뭉치 확장 방법에 대해 설명한다. 제안 방법은 서로 다른 언어쌍 B-C 병렬 말뭉치를 얻기 위해 A-B 병렬 말뭉치와 A-C 병렬 말뭉치를 활용하여 원하는 병렬 말뭉치를 얻는 방법에 대해 설명한다. A-C의 SMT를 구축한 후 A-B를 사용해 B-C를 얻는 방법과 A-B의 SMT를 구축한 후 A-C를 사용해서 B-C를 얻는 방법이 있는데, 어떤 방법을 사용하던지 상관은 없지만, 문법적 구조가 유사한 언어쌍의 SMT 구축을 먼저 수행하는 것이 양질의 유사(Pseudo) 말뭉치를 얻을 수 있다. 본 논문에서는 A를 영어, B를 스페인어 그리고 C를 한국어 대상으로 하며, 스한 말뭉치를 자동으로 확장한다.

3.1. Pseudo 말뭉치 구축

본 논문에서는 스한 SMT 자동번역 시스템 개발에 필요한 스한 말뭉치의 자동 확장을 대상으로 하며, 기 구축되어 있는 영한 병렬말뭉치를 활용한다.

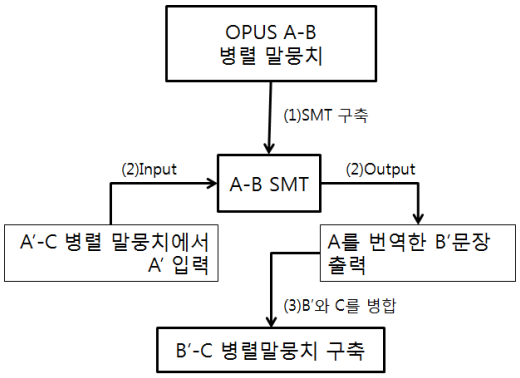

그림 1 Pseudo 말뭉치 구축 과정

첫 번째, 영어-스페인 병렬 말뭉치를 이용하여 영스 SMT를 구축한다. SMT 구축에 사용되는 병렬 말뭉치로는 공개 데이터인 Open Parallel Corpus(OPUS)를 이용하였다[10]. 데이터를 학습하기 전에 MOSES 시스템에 알맞은 형태로 만드는 전처리 단계가 필요하다. 예를 들면 형태소 분석을 통해 구를 형태소 단위로 분리하는 것과 같은 단계가 필요하다. 본 논문에선, MOSES 시스템이 인식하지 못하는 영어-스페인에 사용된 특수문자를 제거하는 과정만 수행하고 어휘 단위로 SMT를 학습하였다.

두 번째, 영어-한글 병렬 말뭉치의 영어를 영어-스페인 SMT를 통해 스페인어 문장으로 자동번역한다. 그리고

스페인어로 번역된 문장을 한글과 문장 단위로 1대1로 정렬한 후, 기존 스페인어-한국어에 병합하였다. 이 방법을 사용하여 수집하기 힘든 스페인어-한국어 병렬 말뭉치를 확장할 수 있다.

마지막으로 Pseudo 말뭉치가 SMT 성능 향상에 기여하는 것을 확인하기 위해, 기존 말뭉치로 학습한 SMT와 기존 말뭉치와 자동으로 구축한 말뭉치를 병합한 확장 말뭉치로 학습한 SMT의 성능을 비교 평가하였다.

4. 실험

본 논문에서는 OPUS 영어-스페인 병렬 말뭉치 약4,000만 문장을 영스 SMT 자동번역 시스템의 학습 데이터로 사용하였고, 기존 말뭉치로는 한국어-영어 말뭉치 약 240만 문장을 사용하였다. 말뭉치 확장을 통한 SMT 성능 향상을 확인하기 위해 수동으로 구축한 말뭉치를 Pseudo 말뭉치와 병합하여 스한 SMT 자동번역의 학습 데이터로 사용하였다. SMT 성능을 평가하기 위해 두 가지 도메인의 평가문을 사용하였다. 사용한 평가문은 일반 대화문 3,000문장과 여행 대화문 300문장이다.

표 1 말뭉치별 문장 수 및 사용 목적

구분	말뭉치	문장 수	목적
학습용	OPUS 영스	40,030,835	영스 SMT
	한영 말뭉치	2,452,799	말뭉치구축
	스한 수동구축 말뭉치	1,242,433	스한 SMT
	Pseudo 말뭉치	2,452,799	스한 SMT
	Pseudo 말뭉치 +수동구축 말뭉치	3,695,232	스한 SMT
평가용	일반 대화 3,000	3,000	스한 SMT 성능 평가
	여행 대화 300	300	스한 SMT 성능 평가

실험을 위해 스페인어는 어휘 단위로 분리하고 한국어는 형태소 분할과 태깅을 동시에 수행하는 모델[12]을 적용하여 형태소 단위로 분리하였다. 번역 성능은 BLEU를 사용해 측정하였다.

표 2는 일반 대화문 3,000문장을 번역한 결과에 대한 성능표이다. 수동으로 구축한 말뭉치와 제안 방법으로 구축한 Pseudo 말뭉치, 수동 말뭉치와 Pseudo 말뭉치를 병합한 확장 말뭉치를 각각 SMT로 사용했을 때 성능을 비교하였다. 자동으로 구축한 말뭉치를 포함한 확장 말뭉치 기반의 SMT가 기존 수동 구축 말뭉치만으로 학습한 SMT에 비해 번역 성능이 향상되는 것을 볼 수 있다.

표 2 일반 대화문 SMT 번역 성능

SMT	BLEU
수동구축 말뭉치 SMT [스한]	17.97
Pseudo 말뭉치 SMT [스한]	19.12
확장 말뭉치 SMT [스한]	24.26

표 3은 여행 분야에서 사용되는 대화문 300문장을 번역한 결과에 대한 성능표이다. 표 2와 다르게 Pseudo 말뭉치만 사용하여 구축한 SMT의 성능은 하락하지만 확장 말뭉치를 사용하면 SMT 성능이 향상되는 것을 볼 수 있다.

표 3 여행 대화문 SMT 번역 성능

SMT	BLEU
수동구축 말뭉치 SMT [스한]	16.26
Pseudo 말뭉치 SMT [스한]	15.20
확장 말뭉치 SMT [스한]	20.48

그림 2는 병렬 말뭉치의 크기에 따른 SMT의 성능의 변화를 나타내며, 말뭉치의 양이 늘어날수록 SMT의 성능이 향상하는 것을 볼 수 있다. 일반 대화문 3,000문장에 대한 SMT 성능 변화는 말뭉치가 증가하면서 계속적으로 상승하는 것을 볼 수 있고, 여행 대화문 300문장에 대한 SMT 성능도 중간에 감소하기도 하지만 확장 말뭉치가 추가됨에 따라 성능이 향상됨을 보여준다.

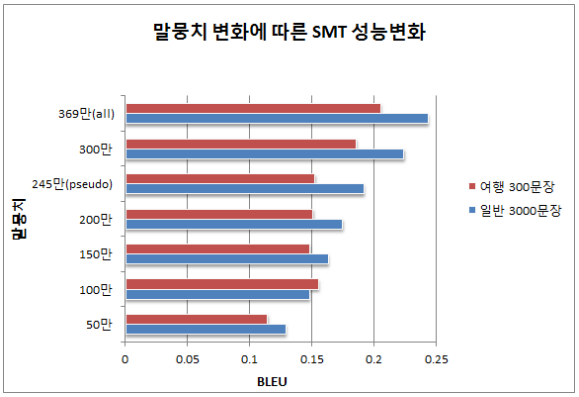

그림 2 말뭉치 변화에 따른 SMT 성능 변화

실험 결과에서 확인한 것처럼, 제안 방법으로 병렬 말뭉치를 확장하여 SMT의 학습 데이터로 사용한다면, 수동으로 구축하는데 어려움을 겪는 언어쌍에 대한 병렬 말뭉치를 추가로 확보할 수 있고 이를 통해 SMT의 성능을 향상시킬 수 있다.

표 4, 5는 자동번역 결과의 일례를 보여준다. 예를 들

어, 수동 구축 말뭉치 SMT의 결과로 보이는 '자칫하면 끊어지니까 부분입니다.' 문장은 확장 말뭉치 SMT를 통해 '그 부분을 자를 수 있어요.' 라는 자연스럽고 정확한 번역 결과가 나올 수 있었다.

표 4 번역할 스페인어 문장

스페인어
i Se puede cortar por secciones. (It can cut through sections.)
ii Estábamos caminando cuidadosmente fuera del camino. (We were walking carefully out of the way.)
iii ¿Está vendiendo todos estos productos? (He's selling these products?)

표 5 말뭉치별 SMT의 번역 결과

사용 말뭉치	번역 결과
수동 구축 말뭉치	i 자칫하면 끊어지니까 부분입니다.
	ii 우리는 조심스럽게 걷는 길 밖으로 밀쳤다.
	iii 모든 제품들을 판매하고 있나요?
Pseudo 말뭉치	i 자를 수 있을 것이다.
	ii 우리는 조심스럽게 걷고 있는 길이다.
	iii 상품들을 판매하고 있나요?
확장 말뭉치	i 그 부분을 자를 수 있어요.
	ii 조심해서 길을 걷고 있었습니다.
	iii 이 제품을 팔고 있나요?

5. 결론

본 논문에서는 기존 말뭉치와 자동번역 방법을 활용하여 원하는 새로운 언어쌍의 병렬 말뭉치를 자동으로 확장하는 방법을 제안하고, 확장한 말뭉치의 활용 가능성을 확인하기 위해 통계기반 자동번역 시스템(SMT)을 사용하여 번역 성능을 측정하였다. 대량의 영어-스페인어 말뭉치를 이용하여 영스 SMT를 구축하였고 기 구축된 한영 병렬 말뭉치의 영어를 자동으로 번역하여 스한 병렬 말뭉치를 자동으로 구축할 수 있었다. 수동으로 구축하기 힘든 병렬 말뭉치를 기존의 자원을 활용하여 자동으로 확장할 수 있으며 번역 성능을 향상시킬 수 있음을 보였다. 본 논문에서는 한국어-스페인어 SMT 자동번역을 대상으로 했지만, 본 논문에서 제안하는 방법론을 통해 중국어, 불어, 독어, 러시아어 등 학습 데이터가 충분치 않은 다양한 언어쌍으로 확장 가능하다. 다음 연구로 SMT의 번역 품질에 대한 자동평가를 통해 품질이 낮은 번역문을 제외하는 방법을 추가로 연구할 것이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발]

참고문헌

[1] Philipp Koehn et al, "Moses:Open Source Toolkit for Statistical Machine Translation", ACL, June 2007.

[2] Papineni, K. et al, "BLEU: a method for automatic evaluation of machine translation", Proceedings of ACL, pp.311-318, 2002.

[3] El Kholy, "Language Independent Connectivity Strength Features for Phrase Piot Statistical Machine Translation", The 51st ACL, pp.412-418, 2013.

[4] Henriquez et al, "Pivot strategies as an alternative for statistical machine translation tasks involving iberian languages", ICL, 2011

[5] 황은하, "한•중 인터넷 신문 기사 표제 병렬망뭉치 연구 - 기계번역을 위한 시험적 연구", 번역학연구, 제10권, 제3호, pp.217~245, 2009.

[6] 이공주, "Moses를 이용한 한일 양방향 통계기반 자동 번역 시스템", 한국마린엔지니어링학회지, 제36권, 제5호, pp.683~693, 2012.

[7] 양주일, "웹 문서로부터 한-영 병렬 말뭉치 자동 구축과 문장 단위 정렬", 제11회 한글 및 한국어 정보처리 학술발표 논문집, pp.150-155, 1999.

[8] 김지형, "웹 번역문서 판별과 병렬 말뭉치 구축", 2004년 한국정보과학회 학술발표 논문집, 제31권, 제2호, pp.787-789, 2004.

[9] Adafre Sisay Fissaha et al, "Finding similar sentences across multiple languages in wikipedia," In Proceedings of EACL '06, pp.62, 2006.

[10] Jörg Tiedemann et al, "The OPUS corpus - parallel and free", In Proceedings of the Fourth International Conference on Language Resources and Evaluation(LREC), pp.26-28, 2004.

[11] Philipp Koehn et al, "Findings of the 2013 Workshop on Statistical Machine Translation", In Proceedings of the Eighth Workshop on Statistical Machine Translation, pp.1-44, 2013.

[12] 나승훈, 양성일, 김창현, 권오욱, 김영길, "CRF에 기반한 한국어 형태소 분할 및 품사 태깅", 한글 및 한국어 정보처리, 2012.