영한 및 한영 통계기반 기계번역에서의 이중언어 간 어순처리 및 단어정렬 최적화 방안 연구

정상원*
*고려대학교 컴퓨터정보통신대학원 소프트웨어공학과
e-mail : cswfox@korea.ac.kr

The study of Method for Optimization of Phrase Ordering Process and Word Alignment between Parallel Languages in Korean-English Statistic Based Machine Translation

Sang-won Chong*
*Dept. of Software Engineering, Korea University

요 약

통계기반 기계번역 시스템 (SBMT system)은 기계번역시스템 중에서 최근 활발히 연구되고 있는 분야이다. 통계기반 기계번역은 대용량의 말뭉치를 사용할 수 있어 특정 언어 쌍에 제한을 덜 받아 모델을 자동으로 학습할 수 있으며 다른 언어에 일반화하여 적용이 가능하다는 장점이 있다. 그러나 영어와 한국어 간 통계기반 기계번역에 있어서는 어순의 차이로 인한 문제를 해결할 필요성이 여전히 남아 있다. 이에 본 연구에서는 영어와 한국어 간 이중언어 말뭉치를 구축하고 통계기반 기계번역 훈련 시스템인 Moses 에 기반하여 구현한 베이스 시스템을 이용하여 이중언어 간 어순처리 및 단어정렬의 최적화 방안을 연구하였다.

1. 서론

글로벌라이제이션은 세계의 다양한 시각, 상품, 아이디어 및 서로 다른 문화를 국가 간에 교환함으로써 발생하는 국제적 통합 프로세스이다. 이는 EU 및 WTO 발족 등 지역 경제의 통합, SNS 및 미디어 기술의 발달 등 글로벌 정보에 대한 보다 쉽고 폭넓은 접근, 에너지 자원 및 첨단기술 개발을 위한 국가간 협력체계 강화 및 협정 체결을 가속화함으로써 이제 기업과 개인은 그들이 속한 국가를 넘어서는 문화, 경제, 사회적인 활동을 할 수 있게 되었다. 그러나 이러한 국제적 통합 프로세스 실현에 있어 국가나 민족마다 서로 다른 언어의 장벽을 넘기는 쉽지 않다.

이러한 언어의 장벽을 넘을 수 있는 것이 번역기술이며 기계번역 중 최근에 활발하게 연구되고 있는 것이 통계기반 기계번역이다[1]. 통계기반 기계번역은 어순이 같은 언어 간에는 호환성이 좋으나 반대의 경우에는 그렇지 못하다. 따라서, 본 논문에서는 어순이 다른 영어와 한국어 간 통계기반 기계번역을 구현하는 데 있어 두 언어 간 어순처리와 단어정렬의 최적화 방안을 연구하고자 한다.

본 논문의 2 장에서 통계기반 기계번역에 대하여 소개하고, 3 장에서 Moses 를 활용한 통계기반 기계번역 시스템 구현과 어순처리 및 단어정렬 최적화 방안을 설명한다. 4 장에서 구현 시스템의 번역품질을 분석하고 5 장에서 결론 및 향후 과제를 기술한다.

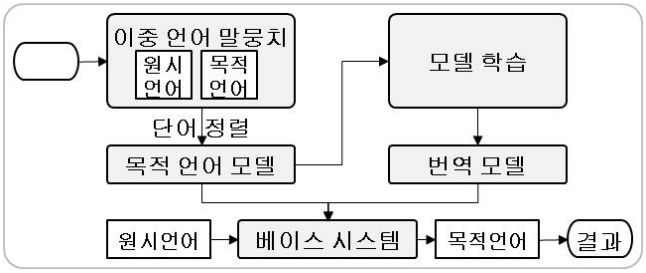
2. 통계기반 기계번역과 Moses

2.1 통계기반 기계번역의 개요

통계기반 기계번역은 기계학습을 사용하는 것으로 특징 지워지는 기계번역을 위한 접근방법이다[2]. 예제기반 또는 규칙기반 기계번역과 대응되는 것으로써 이중언어 말뭉치 분석결과로부터 추출되는 파라미터의 통계적 모델에 기반하여 번역을 하는 기계번역의 한 패러다임이다.

통계기반 기계번역은 1949 년에 Warren Weaver 가 처음 소개하였으며 한 동안 다른 기반 기계번역에 밀려났다가 1991 년 IBM 의 Thomas J. Watson Research Center 가 다시 연구하면서 지금도 연구가 활발히 진행되고 있다[3].

2.2 통계기반 기계번역의 프로세스



(그림 1) 통계기반 기계번역 프로세스

통계기반 기계번역은 이중언어 간 단어를 정렬하여 목적 언어 모델을 생성하고 목적 언어 모델을 입력 값으로 하여 모델 학습 과정을 거치면 번역 모델 및 재배치 테이블이 생성되는데 튜닝 및 탐색 알고리즘을 거쳐 번역을 자동으로 수행한다.

2.3 Moses 의 개요

Moses 는 언어 쌍을 이용하여 자동으로 번역 모델을 학습할 수 있도록 해주는 통계기반 기계번역 시스템이다[4]. 다양한 이중언어 쌍을 제공하고 외부 통계기반 기계번역 소프트웨어들과의 연계가 쉬우며 원시 언어 및 목적 언어에 대한 효율적인 탐색 알고리즘으로 번역 확률이 높은 문장을 빠르게 검색하도록 도와준다[4].

Moses 에서의 학습 프로세스는 두 언어 간의 Correspondence 를 일으키기 위하여 병렬 데이터 및 단어와 구문의 Co-occurrence 를 사용한다[4]. 또한 대량의 번역 작업에 적용 가능한 학습, 정련 도구들을 제공한다[5]. Moses 를 적용한 연구도 활발히 진행 중이다[6,7].

3. 시스템 구현

3.1 베이스 시스템 구현

3.1.1 이중언어 말뭉치 구축

베이스 시스템을 구축하기 위하여 먼저 한국어와 영어 간 문장단위로 정렬된 이중언어 말뭉치를 <표 1>과 같이 구축하였다.

<표 1> 이중언어 말뭉치 구축 내역

출처	문장 수	단어 수	글자 수
SWRC[1]	72,484	601,448	3,347,758
영자신문	444	10,795	68,497
계	72,928	612,243	3,416,255

규칙기반 기계번역 시스템과 달리 통계기반 기계번역에서 이중언어 말뭉치의 양과 번역 품질은 번역 성능을 좌우한다고 해도 과언이 아닐 정도로 그 역할이 중요하다. 이중언어 말뭉치는 수집하거나 직접 작성하였는데 수집한 것은 번역 정확도가 많이 떨어져 정화나 수정이 필요했다.

3.1.2 통계기반 기계번역 시스템 구현

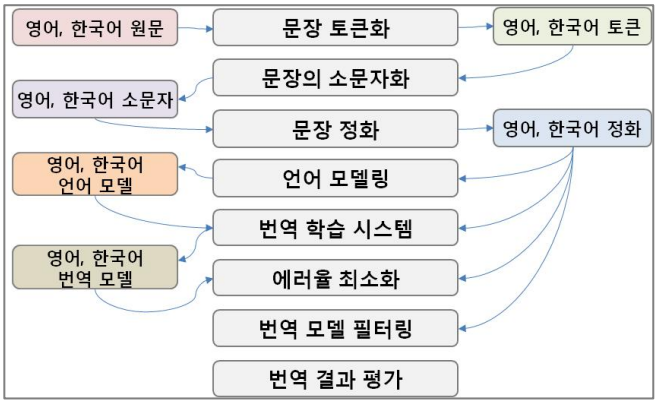
VM 기반 ubuntu v12.10 플랫폼에 오픈 소스 기반의 Moses 를 적용하여 시스템을 구현하였다. 언어 모델링은 SRILM 을 적용하였고[8] 번역 속도의 향상을 위하여 이진처리를 하였다. 번역 모델링 시 품사 tagging 을 위하여 <표 2>와 같이 영어는 Mxpost 및 Tree tagger 를, 한국어는 초고속 한국어 형태소 분석기 Mach 2.0 을 사용하였다[9].

[1] SWRC (Semantic Web Research Center): 영어↔한국어 이중언어 말뭉치 제공, KAIST

<표 2> 품사가 tagging 된 한국어와 영어

구분	내용
영어	After_IN further_JJ research,_NN the_DT study_NN added_VBD that_IN not_RB just_RB the_DT eye_NN color_NN but_CC the_DT face_NN shape_NN seem_VBP to_TO take_VB on_RP an_DT important_JJ role._NN
한국어	진전(NN) 되(SV) ㄴ(EM) 연구(NN) 후(NN) 에(JO) ,(SY) 그(NP DT) 연구팀(NN) 은(JO) 단지(NN AD) 눈(NN) 색깔(NN) 이(JO) 아니(AJ) 라(EM) 그러나(AD) 그(NP DT) 얼굴(NN) 형태(NN) 가(JO) 하나(NU) 의(JO) 중요(NN) 하(SV SJ) ㄴ(EM) 역할(NN) 을(JO) 뜨(VV AJ) 아(EM) 맡(VV) 을(EM) 것(NX) 같(AJ) 다고(EM) 덧붙이(VV) 었(EP) 다(EM) .(SY)

또한 가장 시간이 많이 소요되는 에러율 최소화 작업을 위해 CPU 를 병렬로 처리하였다.



(그림 2) 베이스 시스템 구축 프로세스

3.2 영어와 한국어 간 어순처리

일반적으로, 다른 언어로 번역 시 Distortion[2] 현상이 발생하게 된다. 특히, 영어와 한국어는 어순이 다르기 때문에 Distortion 이 더 크게 발생한다.

<표 3> Distortion 산출 공식[4]

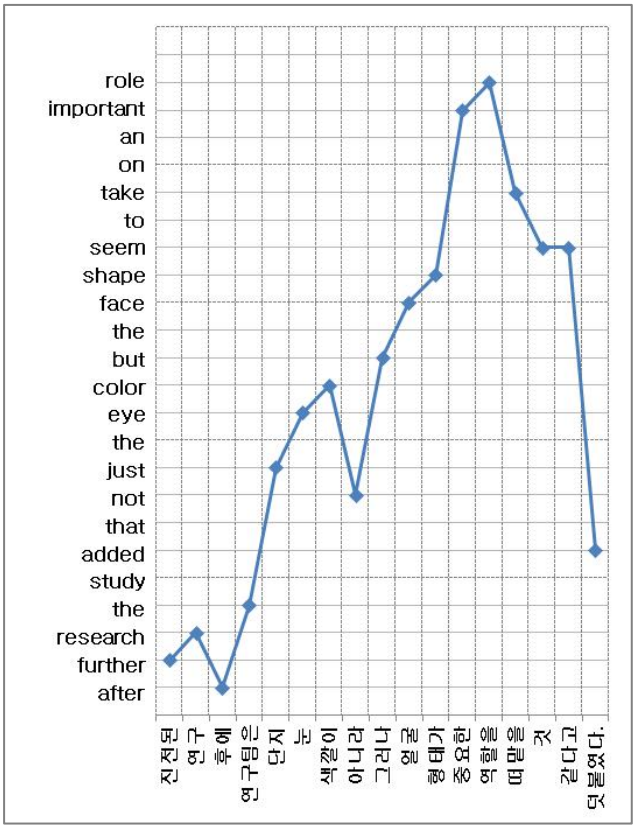
$$D(e,f) = -\sum(d_i)$$

d_i = abs(이전 번역된 문구의 마지막 위치 - 새로 번역된 문구의 첫번째 위치)

D: Distortion, e: 원시언어, f: 목적언어

<표 3>의 Distortion 산출공식을 적용하여 구한 (그림 3)의 Distortion 값은 -53 으로써 그만큼 번역 시 어순 재배치를 필요로 한다.

[2] Distortion: 번역 시 언어가 재배치되는 현상



(그림 3) 영어와 한국어 간 Distortion의 예시

3.3 영어와 한국어 간 단어정렬

3.3.1 영어와 한국어 간 단어 수 차이

영어와 한국어의 이중언어 간에는 단어 수에 차이가 발생하는데 그 이유는 영어의 정관사, 부정관사 및 관계대명사는 의미상 큰 차이가 없다면 번역되지 않거나 영어의 전치사, be 동사, to 부정사 및 조동사는 함께 쓰이는 단어와 합쳐져서 번역되기 때문이다. 그리고 의역된 번역문은 상기의 경우일수록 이 차이가 더 커지는데 그 이유는 다른 단어들이 누락되고 추가되거나 의미를 다르게 번역하기 때문이다.

<표 4>와 같이 영어를 한국어로 번역 시 의역이 될수록 ①과 같이 매칭되는 영어단어의 수가 줄어든다. 특히, 의역된 한국어에는 ② 및 ④와 같이 영어에 없었던 새로운 의미가 추가 되거나 ③과 같이 의미가 변형되었다.

<표 4> 영어와 한국어 간 매칭되는 단어 수 비교

구분	내용	단어 수
단어 기반 번역	After further research, the study added ①~~that~~ not just the eye color but the face shape seem ~~to~~ take on an important role.	23
	진전된 연구 후에, 그 연구팀은 단지 그 눈 색깔이 아니라 그러나 그 얼굴 형태가 하나의 중요한 역할을 떠맡을 것 같다고 덧붙였다.	21 (-2)
일반적인 번역	After further research, ~~the~~ study added ~~that~~ not just ~~the~~ eye color ~~but~~ ~~the~~ face shape seem ~~to~~ take on ~~an~~ important role.	
	진전된 연구 후에, 연구팀은 단지 눈 색깔이 아니라 얼굴 형태가 중요한 역할을 떠맡을 것 같다고 덧붙였다.	17 (-6)
의역	~~After~~ ~~further~~ ~~research~~, ~~the~~ study added ~~that~~ not just ~~the~~ eye color ~~but~~ ~~the~~ face shape ~~seem~~ ~~to~~ ②take on ~~an~~ important ~~role~~.	23
	연구팀은 단순히 눈동자 색이 아닌, 눈동자 색에 ③따라 ④달라지는 얼굴형이 중요하다고 덧붙였다.	12 (-11)

3.3.2 단어정렬과 번역모델 영향도

<표 5>의 {문장 1}과 {문장 2}는 베이스 시스템에 이중언어 말뭉치가 존재함에 따라 번역모델을 그대로 이용하여 정확한 번역결과가 도출되었다.

그러나 {문장 3}은 {문장 1}과 “songs”라는 단어만 다를 뿐인데도 베이스 시스템에 이중언어 말뭉치가 존재하지 않아 번역모델을 그대로 이용하지 못하고 다른 이중언어 말뭉치에서 추출된 번역모델 문구들을 이용함으로써 정확하지 않은 번역결과가 도출되었다. 단어정렬의 평균점수는 2.52561e-05 이었다.

<표 5> 예시문장 별 단어정렬과 번역모델

{문장 1}	she burst into song .
단어정렬	NULL ({ }) 그녀는 ({ 1 }) 갑자기 ({ }) 노래를 ({ }) 부르기 ({ 2 3 45}) 시작했다 ({ }) . ({ 5 })
번역모델	she burst into song. ||| **그녀는 갑자기 노래를 부르기 시작했다.** ||| 0.5 0.0016975 1 1.41538e-05 2.718 ||| 0-0 1-1 3-2 2-3 3-3 3-4 4-5 ||| 2 1 1
번역결과	그녀는 노래를 부르기 시작했다 .
{문장 2}	she suddenly burst into song .
단어정렬	NULL ({ }) 그녀는 ({ 1 }) 갑자기 ({ 2 }) 노래를 ({ }) 부르기 ({ 3 4 5 }) 시작했다 ({ }) . ({ 6 })
번역모델	she suddenly burst into song. ||| **그녀는 갑자기 노래를 부르기 시작했다.** ||| 0.5 0.000664754 1 4.83121e-05 2.718 ||| 0-0 1-1 4-2 2-3 3-3 4-3 4-4 5-5 ||| 2 1 1
번역결과	그녀는 갑자기 노래를 부르기 시작했다 .
{문장 3}	she burst into songs .
단어정렬	없음
번역모델	she burst ||| **그녀는 갑자기** ||| 0.5 0.0406262 1 0.119342 2.718 ||| 0-0 1-1 ||| 4 2 2
	into ||| **속으로** ||| 0.111111 0.119403 0.015873 0.0059835 2.718 ||| 0-0 ||| 9 63 1
	song . ||| . ||| 4.64878e-05 6.59812e-05 1 0.979389 2.718 ||| 1-0 ||| 86044 4 4
	songs . ||| **곡의 노래를 녹음했다 .** ||| 1 0.532873 0.25 0.000222892 2.718 ||| 0-0 0-1 0-2 1-3 ||| 1 4 1
번역결과	그녀는 갑자기 속으로 곡의 노래를 녹음했다 .

3.4 어순처리와 단어정렬 최적화

3.4.1 어순처리 최적화

(그림 3)의 문장에서 Distortion 이 비교적 큰 것은 Added (14), not (12), seem (5), just(5)이며 Distortion 값을 줄이기 위하여 어순을 변경하였다.

진전된 연구 후에 그 연구팀은 **{덧붙였다.} {아니라} {단지}** 그 눈 색깔이 그러나 그 얼굴 형태가 **{떠맡을 것 같다고}** 하나의 중요한 역할을

3.4.2 단어정렬 최적화

3.3.1 에서 제시한 영어와 한국어 간 단어 수 차이를 줄이기 위하여 원시언어에서 부정관사, 정관사를 삭제하였다.

After further research, study added that not just eye color but face shape seem to take on an important role.

그리고 목적언어에서 복수명사, 주어 등 다른 경우의 수가 포함된 문장을 추가하였다.

1) She suddenly **bursts** into song. 2) She suddenly bursts into **songs**.

4. 번역품질 분석

번역품질을 분석하기 위하여 BLEU[3]를 이용하였다. BLEU 는 자연어에서 다른 언어로 기계 번역되는 문장의 품질을 평가하기 위한 알고리즘이다[10]. 어순처리 및 단어정렬 최적화 전· 후의 영어와 한국어 간 BLEU 점수를 비교하였다<표 6>. 유럽지역 언어 및 영어 간 BLEU 최고점수 평균이 22.3<표 7>인 것을 감안하면 본 논문에서의 최적화 후 영어와 한국어 간 BLEU 점수는 비교적 양호한 것으로 분석되었다.

<표 6> 영어와 한국어 간 BLEU 점수 비교

구분	최적화 전	최적화 후	향상율
영어→한국어	19.4	22.4	15.5%
한국어→영어	24.3	25.6	5.3%

<표 7> 유럽지역 언어와 영어 간 BLEU 점수[11]

	목적언어				
원시언어	체코어	15.3	23.6	22.4	19.7
	13.8	독일어	25.7		
	16.3	17.3	영어	33.5	29.9
	14.7		34.5	스페인어	
	14.2		31.5		불어

5. 결론 및 향후 과제

통계기반 기계번역은 이중언어 말뭉치를 모델링하고 이를 효율적으로 검색할 수 있는 알고리즘을 통하여 최적화된 번역결과를 찾아낸다. 이 과정에서 영어와 한국어의 어순이 다르고 관사, to 부정사 등 한국어에 없는 단어들이 영어에 존재하기 때문에 어순처리와 단어정렬을 최적화할 필요가 있었다.

본 논문에서는 한국어와 영어 간 이중언어 말뭉치에서 어순처리와 단어정렬 최적화를 위하여 단어정렬의 경우, 한국어에 없는 영어의 부정관사 등을 제거하였고 어순처리의 경우, Distortion 값이 큰 단어들의 어순을 변경하였다. 이를 통하여 영어와 한국어 간 통계기반 기계번역 품질 향상의 가능성을 보았다.

그러나 한국어와 영어 간 어순처리와 단어정렬 최적화에 있어 보다 근본적인 해결방안이 필요하다. 이를 위하여 향후 이중언어 말뭉치의 대용량 구축 자동화 및 단어정렬의 일관성 향상을 위한 연구를 지속적으로 수행할 것이다.

[3] BLEU: Bilingual Evaluation Understudy

참고문헌

[1] ETRI, “The Trends of Machine Translation Technology and Case Study” Vol. 23, No. 1, pp. 91, 2008

[2] Adam Lopez, “Statistical Machine Translation” ACM Computing Surveys, Vol. 40, No. 3, Article 8, pp. 8:2, 2008

[3] Peter F. Brown, “A Statistical Approach To Machine Translation”, pp. 1, 1990

[4] Philipp Koehn, “Statistical Machine Translation System User Manual and Code Guide”, pp. 11-12, 39, 2013

[5] Philipp Koehn, “Moses: Open Source Toolkit for Statistical Machine Translation” Proceedings of the ACL 2007 Demo and Poster Sessions in Prague, pp. 177, 2007

[6] Kong-Joo Lee, Song-wook Lee, and Jee-Eun Kim, “A Bidirectional Korean-Japanese Statistical Machine Translation System by Using MOSES”, 2012

[7] The Kyoto Free Translation Task (KFTT): http://www.phontron.com/kftt, 2012

[8] Stolcke, Andreas, "SRILM - An Extensible Language Modeling Toolkit" Intl. Conf. on Spoken Language Processing, pp. 2, 2002

[9] 심광섭, 양재형, “인접 조건 검사에 의한 초고속 한국어 형태소 분석” 한국정보과학회, 31 권 1 호, pp. 89-99, 2004

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation” Proceedings of the 40th Annual Meeting of the ACL in Philadelphia, pp. 311-318, 2002

[11] Viewing Matrix: http://matrix.statmt.org, 2012