한국마린엔지니어링학회지 제36권 제5호, pp. 683~693, 2012. 7(ISSN 2234-8352 Online) / (ISSN 2234-7925 Print)
Journal of the Korean Society of Marine Engineering http://dx.doi.org/10.5916/jkosme.2012.36.5.683

# MOSES를 이용한 한/일 양방향 통계기반 자동 번역 시스템

이공주[1] · 이성욱† · 김지은[2]

(원고접수일 : 2012년 4월 24일, 원고수정일 : 2012년 5월 8일, 심사완료일 : 2012년 5월 25일)

# A Bidirectional Korean-Japanese Statistical Machine Translation System by Using MOSES

Kong Joo Lee[1] · Songwook Lee† · Jee Eun Kim[2]

**요 약 :** 통계기반 자동 번역 시스템은 구현과 유지보수의 용이함으로 최근 많은 관심을 받고 있다. 본 연구의 목적은 MOSES[1] 시스템을 이용하여 통계기반의 한/일 양방향 기계번역시스템을 구축하는 것이다. 한/일 문장단위 병렬 코퍼스를 구축하여 번역모델 학습에 이용하였고, 한/일 각각 대량의 원시 코퍼스를 이용하여 언어모델 학습에 이용하였다. 시스템 구축 결과 기존의 규칙기반 번역 시스템의 성능에 근접하는 결과를 얻었으며, 발생하는 오류의 대부분은 각 처리 단계에서 발생하는 노이즈에 기인하였다.
**주제어 :** 기계번역, 통계적 번역, MOSES

**Abstract:** Recently, statistical machine translation (SMT) has received many attention with ease of its implementation and maintenance. The goal of our works is to build bidirectional Korean-Japanese SMT system by using MOSES [1] system. We use Korean-Japanese bilingual corpus which is aligned per sentence to train the translation model and use a large raw corpus in each language to train each language model. The proposed system shows results comparable to those of a rule-based machine translation system. Most of errors are caused by noises occurred in each processing stage.
**Key words:** machine translation, statistical machine translation, MOSES

## 1. 서 론

SMT(Statistical Machine Translation)은 통계 기반의 번역 모델이다. SMT를 기반으로 한 기계번역은 양 언어 사이에 문장 단위로 정렬화된 코퍼스만 있으면 기본적으로 번역 시스템을 구축할 수 있다. SMT 기술은 1991년 IBM의 Thomas J. Watson 연구소 연구원들에 의해 소개되면서부터 연구가 부활하여[2], 현재 가장 활발하게 연구되는 기계번역 기술이다[3]. 특정 언어에 대한 의존성 없이 어떠한 언어 쌍에도 적용할 수 있다는 장점 때문에 Google, IBM, ISI 등에서 N대 N의 언어쌍에 대해 실용적 활용을 목표로 개발을 해오고 있다.

SMT는 초기 개발 비용이 저렴하며 빠른 시간에 개발할 수 있다는 장점을 갖고 있다. 또한 해당 언어에 대한 전문가가 없어도 번역 시스템을 개발할 수 있다는 특성을 지니고 있다. 이런 반면에 언어적으로 전혀 의미가 없는 단위로도 번역 패턴이 학습되는 단점을 가진다. 또한 번역 오류가 발견되어도 시스템 튜닝이 어렵다는 단점을 갖고 있다.

본 연구에서는 Open Source인 MOSES 시스템[1]을 이용하여 통계기반의 한/일 양방향 기계번역 시스템을 구축하고자 한다. 기존의 규칙 기반의 번역 방식을 최대한 배제하고 대부분의 모듈을 통계 기반의 번역 방식을 도입하여 구축해 본다. 그래서

† 교신저자(한국교통대학교 컴퓨터정보공학과, E-mail: leesw@ut.ac.kr, Tel: 043-841-5464)
1 충남대학교 정보통신공학과, E-mail: kjoolee@cnu.ac.kr, Tel: 042-821-5662
2 한국외국어대학교 영어학과, E-mail: jeeeunk@hufs.ac.kr, Tel: 02- 2173- 3110

순수한 통계 방식의 기계 번역이 얼마나 실용 가능한지를 평가해 본다.

이와 같은 통계기반의 접근 방법은 영역에 상관없이 모든 분야의 문서에 동일하게 적용 가능하다. 그렇기 때문에 평소 번역 작업이 많이 필요한 해양/조선 분야의 문서 번역에도 그대로 적용 가능하다.

본 논문의 2장에서 SMT 시스템에 대해 소개하고, 3장에서 한/일 양방향 SMT시스템의 구현방법을 설명한다. 4장에서는 제안 시스템의 번역 결과를 살펴보고 5장에서 결론을 맺는다.

## 2. SMT 시스템

### 2.1 SMT 시스템의 구조

다음 Figure 1은 제안하는 시스템의 구조도이다. SMT는 Figure 1에서 보는 바와 같이 크게 세 가지 부분으로 나뉜다.

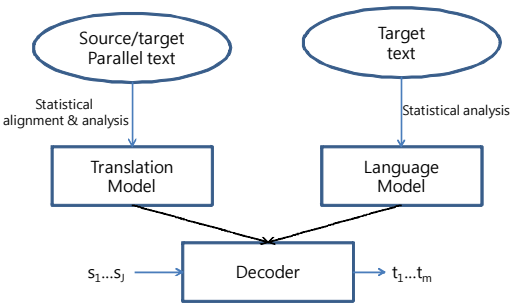


**Figure 1:** The architecture of SMT system

① Translation Model: 두 언어에 대해 문장 단위로 정렬된 코퍼스로부터 두 언어 사이의 대역 정보를 학습하게 된다. 보통 단어 단위나 구절 단위의 대역 정보를 학습한다.

② Language Model: 목적 언어를 생성할 수 있는 생성 정보를 학습한다. 대부분 N-gram 언어 모델을 이용한다.

③ Decoder: 원시 언어 s의 입력에 대해 Translation Model과 Language Model의 파라미터(parameter)로부터 다음의 식 (1)을 만족하는 목적 언어 문장 t를 찾아내는 작업이다. 식 (1)에서 파이(Φ)는 통계 모델(translation model)을 의미하며, LM은 언어 모델(Language Model)을 의미한다.

$$\hat{t} = \arg_t \max P(t|s) = \arg_t \max \Phi(s|t) \times LM(t) \quad (1)$$

번역기(decoder)는 학습된 언어모델과 문장 정렬 코퍼스로부터 추정된 translation model을 이용해서 주어진 원시문장을 목적문장으로 번역한다. 통계기반 번역기는 주어진 원시문장이 입력되면 단어나 구절 정렬 모델을 이용해서 번역 가능한 후보단어나 구절을 선택하여 이들을 재정렬하는 방법을 이용해서 목적문장을 생성한다. 번역 가능한 후보 단어나 구절의 갯수는 일반적으로 아주 많은 편이며, 이들을 다시 재정렬하기 위해서 상당히 많은 양의 계산이 필요하므로, 효율적인 계산을 위해 다양한 알고리즘들이 사용된다. 일반적으로 사용되는 알고리즘으로는 A* 알고리즘이나 탐욕 알고리즘(greedy algorithm)이 있다[4]. 또 다른 방법으로는 빔 탐색 알고리즘(beam search decoding)이 있는데, 계산된 확률값이 충분이 클 때만 재정렬 가능성을 검증하고 그렇지 못할 경우에는 더 이상 가능성을 타진하지 않는다[5]. 그 외 방법으로는 FST와 동적프로그래밍(dynamic programming) 기법을 이용하는 방법들이 있다[6].

### 2.2 GIZA++

GIZA는 병렬 말뭉치로부터 단어정렬 모델을 학습하는 프로그램이다. GIZA++[5,7]은 IBM 모델 4, 5와 HMM 모델이 추가되었으며, 단어 군집화 기법을 이용해서 학습의 속도가 크게 개선된 GIZA의 확장판이다.

### 2.3 MOSES

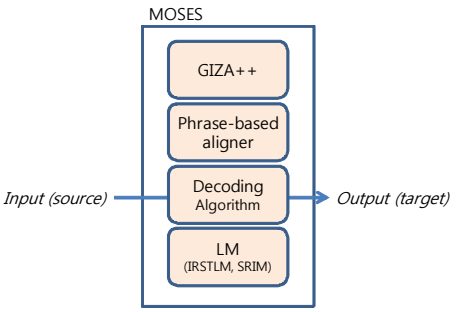


**Figure 2:** Components of MOSES

MOSES[1]는 phrase 단위의 번역 패턴을 학습할 수 있으며 동시에 decoder를 포함하고 있는 SMT 시스템이다. GIZA++가 학습한 단어 단위의 정렬을 기반으로 하여 phrase 단위의 정렬을 학습할 수 있으며, 빔 탐색(beam-search)을 통해 번역 속도를 개선하였다. 또한 학습하는 구절 정렬에서 단어뿐만 아니라 품사나 기타 번역에 유용한 정보를 함께 포함시켜 학습을 수행할 수 있다는 특성을 갖고 있다.

MOSES의 주요 구성 요소는 Figure 2와 같다.

### 2.4 언어모델(Language Model)

통계기반 기계번역에서는 자연스러운 목적 문장(target sentence)을 생성하기 위해 n-그램 언어모델을 이용하며 주로 3-그램을 사용하고 있다[8]. 즉 연속적인 3개 단어가 사용되는 확률이 가장 높은 문장을 생성하도록 결정한다.

본 연구에서는 언어모델 도구로 IRST를 사용하였고[9], 3-그램(trigram)을 기본적으로 사용하였다. 어휘정보에 대한 3-그램과 품사정보에 대한 3-그램을 동시에 구축하여, 만약 어휘정보에 대한 3-그램 정보가 없을 때, 품사정보의 3-그램으로 대치할 수 있도록 하였다.

## 3. 한/일 양방향 번역 모델

### 3.1 코퍼스 구축

SMT 시스템을 구축하기 위해서는 두 종류의 코퍼스가 필요하다. 첫 번째, 두 언어 간에 문장 단위로 정렬이 되어 있는 병렬 코퍼스(bilingual corpus)가 필요하다. 이 병렬 코퍼스로부터 한/일, 일/한에 대한 구 단위(phrase base) 정렬을 수행할 수 있으며, 이렇게 정렬된 정보가 대역어 정보로써 사용된다. 그렇기 때문에 양질의 문장 단위 병렬 코퍼스로부터 양질의 번역 시스템을 구축할 수 있다. 두 번째는 각각의 언어에 대한 언어 모델을 구축하기 위해 사용하는 단일 언어 코퍼스(monolingual corpus)이다. 각 언어에 대해 대량의 단일 언어 코퍼스를 수집함으로써 더욱 자연스러운 목적언어 문장을 생성할 수 있다.

한/일 문서 단위 병렬 코퍼스는 다른 병렬 코퍼스에 비해 비교적 수집이 용이하다. 가장 많은 출처로는 신문기사 사이트들이 제공하는 다국어 서비스 페이지들이다. 신문기사 사이트는 정치, 경제, 사회, 스포츠 등 다양한 도메인을 다루고 있기 때문에 한 도메인에 치우치지 않고 여러 도메인에 균형 잡힌 코퍼스를 구축할 수 있는 주요 소스이다. 신문기사의 경우에는 주로 기사 단위로 서비스되기 때문에 이를 문장 단위로 병렬화 시키는 부가 작업이 필요하다. 한/일 및 일/한 사전에 있는 예제 문장 또한 좋은 문장 단위의 병렬 코퍼스이다. 사전의 예제 문장은 기본적으로 문장 단위로 정렬이 되어 있으며 각 단어에 대한 가장 대표적인 쓰임을 제공하기 때문에 통계기반의 번역 시스템에서 가장 필요로 하는 학습 데이터가 될 수 있다. 그 이외에 병렬 코퍼스의 소스가 될 수 있는 것들로는 드라마나 영화의 자막, 한/일 일본어 학습서에서 제공하는 문장쌍들이다.

마이크로소프트의 MSDN 페이지는 동일한 페이지의 내용을 여러 개의 언어로 동시에 제공하고 있다. 이 동일한 내용을 담고 있는 페이지는 국가와 언어 정보 부분만 다르고 그 외 나머지는 동일한 URL을 갖고 있다. 예를 들어, http://msdn.microsoft.com/ko-kr/subscriptions/dd365189.aspx (한국어 페이지)와 http://msdn.microsoft.com/ja-jp/subscriptions/dd365189.aspx (일본어 페이지)처럼 동일한 URL을 갖고 있기 때문에 문서 단위의 정렬을 용이하게 할 수 있다. 그러나 기호나 고유명사, MS 고유 용어들이 많고, 부자연스러운 번역도 많다. 본 연구에서는 MSDN 문서로부터 기호나 숫자, 알파벳 등이 많이 사용된 문장들은 모두 제거한 후, 문장 단위의 병렬 코퍼스를 수집하였다.

SMT 번역 시스템에서는 대역어 사전을 따로 사용하지 않는다. 그러나, 한/일 또는 일/한 대역어 사전은 번역에 있어서 가장 필수적인 요소이다. 따라서 본 연구에서는 기존의 대역어 사전을 문장 단위의 병렬 코퍼스로 구축하여 함께 학습에 사용하고자 한다. Tabel 1에서 본 연구에서 구축한 문장 단위의 병렬 코퍼스 규모를 제시한다. 문장 단위의 병렬 코퍼스는 모두 524,573 문장쌍이며, 한/일 대역어 사전에서 추가한 엔트리들이 347,820

쌍으로 모두 872,393의 병렬 문장이 학습 데이터로 사용된다.

**Table 1:** Size of bilingual corpus aligned by sentences

| 출처 | 파일 개수 | 문장 수 | 비율 |
|---|---|---|---|
| 중앙일보 | 1,110 | 17,366 | 18.8% |
| 조선일보 | 2,828 | 22,455 | |
| 동아일보 2005 | 390 | 6,887 | |
| 동아일보 2009 | 2,984 | 51,284 | |
| 네이버사전 예제 문장 | 1 | 119,854 | 22.8% |
| 세종 코퍼스 | 2 | 25,182 | 4.8% |
| MSDN | 13,189 | 237,651 | 45.3% |
| 기타 | 70 | 43,894 | 8.4% |
| **총합** | **20,574** | **524,573** | **100%** |
| 사전 엔트리 | 15 | 347,820 | |
| **총 합** | | 872,393 | |

다음은 구축된 문장 단위 병렬 코퍼스의 일부이다. 각 파일은 동일한 파일 이름 앞에 “kr_”과 “jp_”의 접두사로 동일한 내용을 담고 있는 파일임을 알려준다. 각 문장 앞에는 문장 번호가 부여되어 있어서 문장 단위의 병렬을 명확히 표시한다.

**파일 이름: kr_sports_20090623280088_news.txt**

00000001:FIFA랭킹 1위 스페인, A매치 15연승 질주
00000002:2008년 6월 30일 오스트리아 빈.
00000003:2008 유럽선수권대회 결승에 오른 스페인은 전차군단 독일을 1-0으로 꺾고 우승컵을 들어올렸다.
...
00000010:안드레스 이니에스타, 사비 에르난데스, 세스크 파브레가스 등이 버틴 라인은 미드필더의 교과서로 불린다.
00000011:90%에 육박하는 패스 성공률과 한 박자 빠른 패스, 뛰어난 개인기를 무기로 장착한 이들은 상대의 어떤 압박도 무력하게 만든다.
...

**파일 이름 : jp_sports_20090623280088_news.txt**

00000001:無敵艦隊スペインがＡマッチ１５連勝の大記録
00000002:０８年６月３０日、オーストリアのウィーン。
00000003:０８欧州選手権大会の決勝に進んだスペインは、「戦車軍団」ドイツを１対０で下して優勝カップを手に入れた。
...
00000010:アンドレス・イニエスタ、シャビ・エルナンデス、フランセスク・ファブレスガスらが構えているＭＦ陣営は、「ＭＦの教科書」と呼ばれる。
00000011:９０％に迫るパス成功率やワンテンポ速いパス、精彩に富む個人技を武器としている彼らは、相手のいかなるプレスも無力化してしまう。
...

단일 언어 코퍼스(monolingual corpus)는 한국어 13,483,715 문장과 일본어 9,239,001 문장으로 구성되어 있다.

### 3.2 정렬 기본 단위

본 연구에서는 한/일 양방향 SMT의 정렬 단위로 형태소를 사용한다. 즉, 입력된 문장을 형태소 단위로 모두 분석한 후, 형태소 단위로 정렬을 시킨 후, 목적 언어에 대한 문장을 생성해 내는 것이다. 단어나 구절 단위의 정렬을 수행하기 위해서, 우선 두 언어의 문장들에 대해 형태소 분석을 수행한다. 형태소 분석의 결과는 한 형태소에 대해 (surface, lemma, POS)의 세 가지 정보를 항상 유지한다.

구절 단위 정렬 시, 원시 언어에 대해서는 형태소 분석의 결과로부터 원형(lemma)과 품사정보를 사용하고, 목적 언어에 대해서는 표층형(surface)과 품사정보를 출력한다. 목적 언어에 대해 원형이 아닌 표층형을 사용하는 이유는 부가적인 생성 규칙 없이 바로 목적어 문장을 만들 수 있도록 하기 위함이다.


**Figure 3:** Factored model for phrase based alignment

구절 단위 정렬 과정은 원시언어의 “lemma|POS”에 대해 목적언어의 "surface|POS"를 출력하도록 설계하였다. 목적언어에 대해 표층형 이외에도 품사를 출력하도록 설계한 이유는 표층형에 대한 언어 모델이 미흡한 경우 품사에 대한 언어 모델이 동작하도록 하기 위함이다.

예를 들어 한국어 입력 문장 “한국에서 개최된 경기였다”에 대해 Figure 4와 같은 정렬이 가능하다.


**Figure 4:** An example of phrase based alignment

한국어 입력에 대해 형태소 분석을 통해서 각 형태소에 대해 표층형, 원형, 품사정보의 세 가지 정보를 얻는다. 그 중, 원시 언어에 대해서는 원형 정보와 품사 정보만이 사용되며, 정렬로 만들어지는 목적 언어에 대해서는 표층형과 품사 정보만이 사용된다. 다음 그림에서 보면 "경기|NNG 에서|J 이기|VV"라는 구절이 "試合|NCPV で|PJKG 勝っ|VIN"와 정렬되어 있음을 알 수 있다.

## 4. 한/일 양방향 번역 시스템 구축

### 4.1 문서 전처리기

일본어 문서에는 일반적으로 전각 문자와 반각 문자가 함께 사용된다. 전체 시스템에서 일괄적인 처리를 위해 우선 전각 문자들을 모두 반각 문자로 치환한다. 전각 문자로 사용될 수 있는 대상은 숫자, 알파벳, 각종 심볼이다. 한국어 문서에서도 1바이트짜리로 치환할 수 있는 2바이트짜리 심볼들을 모두 1바이트 심볼로 치환한다.

### 4.2 문장 분리기

문장 분리기는 입력 문서를 각각의 문장 단위로 분리하여 출력하는 기능을 담당한다.

① 한국어 문장 분리기

한국어 문장 분리기는 perl로 작성되었다. 문장 종결 부호와 종결어미를 기준으로 한국어 문장을 분리하였으며, 문장 분리에 사용된 종결어미는 세종코퍼스[10]의 형태소 부착 코퍼스의 종결어미 중 마지막 3음절만을 추출하여 사용하였다. 문장 분리 알고리즘은 발생빈도가 높은 종결어미, 종결어미 바로 다음에 위치한 어휘 정보와 문장부호 등의 조합을 고려한 휴리스틱으로 작성하였다.

② 일본어 문장 분리기

일본어 문장 분리기는 flex를 이용하여 구현되었다. 먼저 정규 문법과 문장 부호(예: 。, ．, ？, ！, ゜, °, ., ?, !)를 이용하여 일본어 문장을 분리하되 문장 기호 다음에 」, 』, ） , ), ], }, ", ', ` 등의 부호가 이어서 나타나는 경우는 인용 문장을 포함하는 경우이므로 분리하지 않는다.

### 4.3 문장 자동 정렬기

문장 자동 정렬기는 champollion-1.1[11] 툴킷을 사용하였다. 이 문장 자동 정렬기는 각 문장의 길이, 문서 내의 문장 위치와 더불어 어휘 정보를 사용한다. 또한, 대역어 사전을 사용하기 때문에 자동 정렬의 정확도가 높다는 특성을 갖고 있다. 한/일 문장 자동 정렬을 위해 한/일 대역어 사전을 추가하였다. 자동 정렬의 결과 1:1문장으로 정렬된 경우만을 학습 데이터로 사용하였다. 그 외 1:2나 2:1 등으로 대응된 병렬 문장의 경우, 번역 패턴 학습에 부정적인 영향을 끼치기 때문에 모두 제거하였다.

### 4.4 한국어 형태소 분석기

본 연구에서 한국어 형태소 분석기는 자체 개발한 차트파싱에 기반한 형태소 분석기를 사용하였다. 품사체계는 기본적으로 <21세기 세종계획>[10]의 품사 체계를 약간 수정하여 사용하였다.

다음은 심볼과 한글이 섞인 입력에 대한 한국어 형태소 분석기의 분석 결과의 예를 나타낸다.

입력: '현존 최강' 美 F-22, 동해 연합훈련 때 뜬다

출력: '|SW 현존|현존|NNG 최강|최강|NNG '|SW 美|美|SW F|F|SL -|-|SW 22|22|SN ,|,|SP 동해|동해|NNG 연합|연합|NNG 훈련|훈련|NNG 때|때|NNG 뜨|뜨|VV ㄴ다|ㄴ다|E

### 4.5 일본어 형태소 분석기

일본어 형태소 분석기는 공개용 SW의 하나인 MeCab(http://mecab.sourceforge.net/)을 사용하였다. MeCab은 Kyoto 대학과 일본 전신전화 주식회사(NTT)가 공동으로 개발한 오픈 소스 형태소 분석 엔진으로 대상 언어, 사전, 코퍼스에 의존하지 않는 범용적인 설계를 기본방침으로 하고 있다. SMT를 위해 Mecab의 분석결과에서 4단계로 세분화된 일본어 품사정보는 perl을 사용하여 69개의 대표 품사로 매핑(mapping) 하여 출력한다.

### 4.6 실패 완화기

① 일본어 실패 완화기

일한 번역과정에서 학습되지 않은 일본어 단어나 구절은 Phrase Table 탐색에 실패하여 일본어 원문이 그대로 출력된다. 이들은 주로 한자 또는 카타가나 형태의 인명, 지명, 조직명 등의 고유명사,

즉 개체명(Named Entity) 표현들이다. 한자나 카타가나 고유명사의 상당 부분은 한글 음독으로 변환하면 한국어로 의미가 전달된다. 일본어 실패완화기는 일본어 미등록어(한자, 카타가나, 히라가나)를 한국어 음독으로 철자변환(Transliteration)하여 출력한다.

일본어 실패 완화기는 일본어 JIS에서 정의한 일본어한자 (KANJI) 6,715개의 한글 음독과 카타가나 발음규칙 600여 개로 이루어져 있으며 flex로 구현되었다.

카타가나/히라가나 규칙은 카타가나/히라가나의 기본 발음 이외에도 복모음 처리 규칙(예: “デュ”=> "듀"), 받침 처리 규칙 (예: “テキ”=> “텍”) 그리고 예외 처리 규칙들로 구성된다. 예를 들어 “アルバム”을 음절단위로 기본적인 음독처리를 하면 “아루바무”로 음독 처리되지만 받침 처리 규칙과 예외 처리 규칙을 적용하여 “앨범”의 형태로 출력한다.

② 한국어 실패 완화기

일한 번역과정과 마찬가지로 동일한 경우가 한일 번역과정에서도 발생한다. 즉, 한글 인명, 지명, 조직명 등의 고유명사는 Phrase Table에 등록되어 있지 않을 수 있다. 이 경우 최종 번역 결과에는 미등록어가 한글로 그대로 출력된다. 한국어 실패 완화기는 한글에 익숙하지 않은 일본인을 위해 한국어 미등록어를 카타가나로 음독 처리하여 출력한다.

한국어 실패 완화기는 KSC-5601(1987)에서 정의한 2,350개의 한글 음절을 일본어 카타가나로 음독 처리한다. 예를 들어 "가"는 "ガ"로 처리한다. 종성 복자음의 경우에는 자모 한 글자 한 글자를 그대로 변환하지 않고 최종 발음 형태로 출력한다. 예를 들어, “갉”의 경우 "ガルク"(가루쿠)”가 아닌 "ガル(갈)"로 변환한다.

### 4.7 숫자 변환기

일본어 숫자 변환기는 히라가나와 한자로 표현된 숫자 표현(기수와 서수 표현)을 아라비아 숫자로 변환한다. 일본어 숫자 변환기는 flex 와 GNU Bison (GNU yacc)를 이용하여 구현되었다. 아라비아 숫자 출력시 천 단위로 콤마(,)를 삽입하여 출력한다.

한국어 숫자 변환기는 일본어와 달리 성분 어휘가 대부분 1음절로 구성되므로 (예: 일|이|삼|…|천|백|만|…|정|재|극) 일본어에 비해 부작용을 방지하기 위해 성분 어휘의 수가 4개를 초과하는 경우에만 변환 결과를 출력한다.

### 4.8 형태소 생성기

본 연구에서는 두 언어에 대한 정렬 과정에서 목적언어의 경우 표층형의 형태로 정렬이 이루어진다. 그렇기 때문에 원시 언어에서 목적 언어로의 디코딩 과정이 완료되었을 때, 목적 언어의 표층형태가 나타나게 되므로 따로 형태소 생성 과정이 필요하지 않다. 다만, "움직이+ㄴ"과 같이 종성 자모가 따로 출력되었을 때 이를 조합시키는 한글 자모 조합 과정만이 부가적으로 요구된다.

### 4.9 정렬 테이블 필터

정렬 작업이 완료되면 한->일 구절 단위 정렬 테이블(Phrase translation table)과 일->한 구절 단위 정렬 테이블이 구축된다. 이 정렬 테이블에는 불필요한 정렬들도 있고 수정을 가해주어야 하는 정렬 패턴들도 있다. 가령, 동일한 숫자, 동일한 심볼, 동일한 알파벳에 대한 정렬 정보는 정렬 테이블에 없어도 번역에 아무런 영향을 미치지 않는다. 따라서 이러한 정보들은 정렬 테이블에서 모두 제거한다. 또한, 숫자나 알파벳 정보 중에도 분명하게 잘못 정렬된 정보들이 존재한다. 다음 예를 보자.

(23)로부터 ||| (22)から

(26개)선두다 ||| (22個)で首位だ

0.5 ||| 0.6

이와 같이 확실히 잘못된 정렬은 이를 제거한다. 그러나, 다음과 같은 경우는 제거해서는 안 된다. 정렬 테이블에서 정렬 패턴을 제거하는 것은 신중하게 수행해야 한다.

8강 라운드 ||| 2次ラウンド

0.25초 ||| 1/4秒

그 외, 목적 언어의 인용부호 쌍이 서로 일치하지 않는 번역 결과가 있을 수 있는데, 이럴 때에는 인용부호를 원시 언어쪽의 인용부호로 통일시켜 해

결한다.

## 5. 번역 시스템 결과

### 5.1 구절 정렬 테이블(Phrase Translation Table)

본 연구에서 구축한 구절 단위의 정렬 테이블은 다음과 같은 규모를 갖는다.

한->일 방향: 11,142,163 phrases; 1.8 GB

일->한 방향: 11,088,784 phrases; 1.8 GB

다음은 구절 단위 정렬 테이블의 일부를 보여주고 있다. 하나의 정렬 패턴은 5개의 필드로 나뉘어져 있으며, 각 필드는 "|||"로 구분된다. 첫 번째와 두 번째 필드는 각각 원시 언어 구절과 목적 언어 구절이며, 세 번째와 네 번째 필드는 각각 원시언어 대 목적언어로의 단어 매핑 정보, 목적언어 대 원시언어로의 단어 매핑 정보이다. 마지막 다섯 번째 필드는 5개의 확률로 구성되는데 각각 구절 정렬 확률(phrase translation probability) Φ(s|t), 어휘 가중치(lexical weight) lex(s|t), 구절 정렬 확률 Φ(t|s), 어휘 가중치 lex(t|s), 마지막으로 구절 페널티(phrase penalty)이다.

```
사각|NNG 형|XSN 의|J ||| (|NCPV 四角形|NG の|PCS ||| (1) (1) (2) ||| () (0,1) (2) ||| 1 0.0610178 0.0150376 0.000360211 2.718
...
사과|NNG 우유|NNG 를|J 배|NNG ||| ,|SYP りんご|NG ,|SYP 牛乳|NG を|PJKG お腹|NG ||| (1) (3) (4) (5) ||| () (0) () (1) (2) (3) ||| 1 0.203582 0.5 0.000147429 2.718
...
사과|NNG 의|J ||| おわび|NCPV の|PCS ||| (0) (1) ||| (0) (1) ||| 1 0.10051 0.0714286 0.0313168 2.718
사과|NNG 의|J ||| りんご|NG の|PCS ||| (0) (1) ||| (0) (1) ||| 0.571429 0.279992 0.285714 0.0872395 2.718
사과|NNG 의|J ||| リンゴ|NG の|PCS ||| (0) (1) ||| (0) (1) ||| 0.666667 0.236494 0.285714 0.0939502 2.718
사과|NNG 의|J ||| 謝罪|NCPV の|PCS ||| (0) (1) ||| (0) (1) ||| 0.5 0.198926 0.357143 0.297509 2.718
...
```

### 5.2 번역 결과

번역 결과의 성능을 살펴보기 위해 다음의 3가지 시스템의 번역 결과를 함께 비교해 보았다. 첫째, 본 연구에서 구현한 SMT 시스템, 둘째 네이버(Naver)에서 서비스하고 있는 기계번역 시스템, 그리고 구글(Google)의 번역 결과를 함께 비교해 본다(S:제안 시스템, N:네이버, G: 구글로 각각 표기).

| | |
|---|---|
| 원문 | 菅首相は5日、松山市内で街頭演説し、消費税率の引き上げに關連し、「消費税の話は好きで言っているのではない。 |
| S | 스가 총리는 5 일, 송산 시내에서 가두 연설하면서 소비 세율 인상과 관련해 ' 소비세의 이야기는 좋아하여 말하고 있는 것은 아니다. |
| N | 칸 수상은 5일, 마츠야마시내에서 가두연설해, 소비세율의 끌어올려에 관련해, 「소비세의 이야기는 좋아하고 말하는 것은 아니다. |
| G | 칸 총리는 5 일, 마츠야마 시내에서 가두 연설하고 소비 세율 인상과 관련, "세금 이야기 마음에 말하고 있는 것은 아니다. |
| 원문 | 860兆円の大借金を作った責任は、自民党にも公明党にも（あるが）民主党にも、なにがしかの責任はある。 |
| S | 860 조 엔의 큰 빚을 만든 책임은 자민당에도 공명당에도 (하지만) 민주당에도 얼마간의 책임은 있다. |
| N | 860조엔의 대빚을 만든 책임은, 자민당에도 공명당에도 (있지만) 민주당에도, 아무개인가의 책임은 있다. |
| G | 860 조엔의 대규모 부채를 만든 책임은 자민당도 공명도 (그러나) 민주당도 기대가 단의 책임이다. |
| 원문 | 最近出版した『田舎の村の韓國戰爭』（石枕社）は、 |
| S | 최근 출간한 시골 마을의 한국 전쟁 ' (이시 마크라사)는 |
| N | 최근 출판한 「시골의 마을의 한국전쟁」(돌베개사)은, |
| G | 최근 출판한 『시골 마을의 한국 전쟁 '(石枕 사)는 |
| 원문 | 거의 완벽한 연기를 한 19세의 김연아 |
| S | ほぼ完璧な演技を終えた19歳の金妍兒 |
| N | ほとんど完璧な延期をおおよそ 19歳のキム・ヨナ |
| G | ほぼ完璧な演技をした19歳の金妍兒 |
| 원문 | 우리는 주 5일제 근무를 싫어한다. |
| S | 我々は週休2日制を嫌っている. |
| N | 私たちは株 5日制勤務を嫌やがる. |
| G | 私たちは、週休2日制勤務を嫌っている。 |

다음은 잘못된 어휘가 선택된 번역 결과의 예들이다.

| | |
|---|---|
| 원문 | 15분 동안의 급유가 이뤄지는 동안 공중급유기와 전용기는 똑같은 속도에다 일정한 고도를 그대로 유지하며 공중 도킹의 장관을 연출했다. |
| S | 15分の間の給油が行われている間,空中給油機と専用機は同じ速度に加え,一定の高度を維持しながら空中ドッキングの長官を演出した |
| N | 15分間の給油が成り立つ間公衆給油機と専用機はまったく同じな速度に決まった高度をそのまま維持して公衆ドッキングの壯觀を演出した. |
| G | 15分間の給油が行われる間、空中給油機や専用機は同じ速度の横に一定の高さをそのまま維持し、公共のドッキングの壯觀を演出した |
| 원문 | 학생 수가 정원의 곱절까지 늘었다. |

| | |
|---|---|
| S | 生徒數が定員の2倍まで增加した. |
| N | 學生數が庭園の一倍まで增えた. |
| G | 生徒數が庭の2倍まで增えた。 |
| 원문 | 임금은 고개를 끄덕이며 다시 물었다. |
| S | 賃金はうなずいて,再び質問した. |
| N | 賃金は首をうなずいてまた問った. |
| G | 王はうなずいであり、再度頼んだ。 |
| 원문 | トラックや乘用車など計八台が絡む玉突き事故が起きた。 |
| S | 트럭과 승용차 등 총 8 대가 얽힌 당구 사고가 일어났다. |
| N | 트럭이나 승용차 등 합계8대가 관련되는 연쇄 추돌 사고가 일어났다. |
| G | 트럭과 승용차 등 모두 여덟 대가 둘러싼 玉突き 사고가 일어났다. |

다음은 형태소 분석 오류 때문에 잘못 번역되어진 문장의 예이다.

| | |
|---|---|
| 원문 | 군의 전력이 사적인 일에 전용됐다는 사실에 충격이 가시질 않는다. |
| S | 軍の戰力が私的なことに轉用されたという事實に衝撃が消えない. |
| N | 軍医全力が私的な事に專用された　という事實に衝撃がいらっしゃらない. |
| G | 軍の戰力が個人的な仕事に轉用されたという事實に衝撃をガシジルない。 |
| 원문 | 미국이 북한에 이익이 되는 행동을 더이상 하지 않으리란 것을 추측할 수 있다. |
| S | 米國が北朝鮮への利益になる行動をもうしないだろうということを推測することができる. |
| N | アメリカが北朝鮮に利益になる行動をこれ以上夏至アンウリと言うのを推測することができる. |
| G | 米國が北朝鮮の利益になる行動をもはやしないアンウリランことを仮定することができる。 |

### 5.3 비교 평가

본 연구에서 개발한 SMT 기반의 한/일 양방향 번역 시스템의 성능을 평가해 보기 위해 간단한 비교 평가를 수행해 보았다. 평가는 일본어 전문가 한 사람이 1~5점까지의 점수를 매기는 방식이다. 한/일, 일/한 모두 150 문장에 대해 평가를 수행했다. Table 2에서와 같이 모든 평가에서 네이버 서비스의 번역기가 가장 좋은 성능을 발휘했으며 본 연구의 결과는 구글 번역기보다는 좋은 성능을 보였다.

**Table 2:** Results of Korean/Japanese SMT system

| 한/일 번역 (150 문장) | | |
|---|---|---|
| 신문기사 (100문장) | SMT | 3.7 |
| | NAVER | 3.9 |
| | GOOGLE | 3.4 |
| 블로그 (50문장) | SMT | 3.28 |
| | NAVER | 3.7 |
| | GOOGLE | 3.0 |
| 일/한 번역 (150 문장) | | |
| 신문기사 (100문장) | SMT | 4.0 |
| | NAVER | 4.4 |
| | GOOGLE | 3.8 |
| 블로그 (50문장) | SMT | 3.6 |
| | NAVER | 3.9 |
| | GOOGLE | 3.3 |

### 5.4 SMT 번역의 문제점

① 병렬 문서의 지역화(localization)

정렬의 부정확성이나 또는 문서의 내용이 각 언어별로 지역화되었을 때, 다음과 같은 문제가 발생할 수 있다. 다음의 문장에 대해 통계기반의 번역방식을 취하는 SMT와 구글의 번역결과에서는 “万円”을 “만원”으로 번역한다.

| | |
|---|---|
| 원문 | 2000万ウォン（約140**万円**）以下の罰金に處せられる。 |
| SMT | 2000 만 원 (약 140 만원) 이하의 벌금에 처해 지다. |
| N | 2000만원( 약 140만엔) 이하의 벌금에 처해진다. |
| G | 2000 만원 (약 140 만원) 이하의 벌금에 처해진다. |
| 원문 | 高橋尙成投手（35）は18日（日本時間）敵地ヤンキースタジアムでのヤンキース戰に先發登板し |
| SMT | 타카하시 상 나리 투수 (35)는 18 일 (한국 시간) 적지 양키 스타디움에서 양키 경기에 선발 등판해 |
| N | 타카하시 나오나리 투수(35)는 18일(일본 시간) 적지 양키 스타디움에서의 양키스전에 선발 등판해 |
| G | 다카하시 히사 노리 투수 (35)는 18 일 (한국 시간) 적지 양키 스타디움에서 양키스 전에 선발 등판해 |

이는 학습 데이터가 각 언어별로 다르게 작성되어 있었기 때문이다. 가령, 두 병렬 문서에서 일본어 문서에서는 “엔”화로 계산하여 “(約163億円)”로 작성되어 있고 동일한 내용을 한국어 문서에서는 “원”화로 계산하여 “(약 1976억원)”로 작성되어 있다. 결과적으로 이러한 병렬 데이터 때문에 ‘万円’

이 '만원'으로 정렬되는 구절 매핑이 존재하게 되어 이런 오류를 발생시키는 것이다.

② 학습데이터 의존에 의한 오역

두 언어에 대한 문서 단위의 정렬 코퍼스가 한쪽 언어에 대해 우위를 갖고 있을 때 오역이 발생한다. 예를 들어 특정 사물을 3인칭의 표현을 사용하지 않고 1인칭이나 2인칭으로 표현하는 경우 다음과 같은 오역이 발생한다.

| 원문 | もちろん、**韓國**メーカーもシェアを守るために必死の努力を傾けています。 |
|---|---|
| S | 물론 **국내** 업체들도 시장 점유율을 지키기 위해 필사의 노력을 기울이고 있습니다. |
| N | 물론, 한국 메이커도 쉐어를 지키기 위해서 필사의 노력을 기울이고 있습니다. |
| G | 물론 한국 업체들도 시장 점유율을 지키기 위해 필사적인 노력을 기울이고 있습니다. |
| 원문 | 그러나 러시아측이 **우리**의 요청을 거부할 경우, |
| S | しかし,ロシア側が**韓國**の要求を拒否します. |
| N | しかしロシア側が私たちの要請を拒否する場合, |
| G | しかし、ロシア側が我々の要求を拒否する場合は、 |

③ 과적합(overfitting)

과적합으로 인하여 엉뚱한 정렬이 이루어질 수 있다. 이런 경우 이해하기 어려운 번역 결과가 발생될 수 있다. 다음의 예제를 보자. 첫 번째 문장의 경우 "フットワークが"에 대해 "흐쯔토와-크게"라는 대역어를 생성했다. 그러나 전혀 이해되지 않는 번역 결과이며, 이는 학습 데이터 중에 오류로 보이는 정렬이 남아 있었기 때문이다.

| 원문 | フットワークが輕いのとは使う神経もエネルギーもちょいと違うのですよ。 |
|---|---|
| S | 흐쯔토와 ー 크게 가벼운 것과는 쓰는 신경도 에너지도 조금 다를 것입니다. |
| N | 자세가 가벼운 것과는 사용하는 신경도 에너지도 조금 다른 거에요. |
| G | 풋워크가 가벼운 것과 사용 신경도 에너지도 잠깐 차이에요.. |
| 원문 | 본지가 LG 텔레콤 측에게 위의 상담사례를 추궁한 결과 |
| S | 本紙が同社にとって上の相談の事例を追及した結果 |
| N | 見てから LG テレコム側に上の相談事例を追窮した結果, |
| G | 本紙は、LGテレコムの側面に、上記の相談事例を追及した結果 、 |

두 번째 문장의 경우에도 학습데이터 중에 " LG 텔레콤 측"을 " 同社"로 번역했던 데이터가 존재했기 때문이다. 이와 같이 과적합으로 인해 오역이 발생할 수 있으며 이와 같은 오역은 원인을 파악하기도 매우 어렵다.

④ 원거리 관계(Long distance dependency)

통계 방식의 기계 번역의 경우 멀리 떨어져 있는 단어나 구절에 대한 처리를 조절하기가 어렵다. 그 중 가장 빈번하게 발생하는 경우가 문장 부호 중, 짝이 맞아야 하는 괄호나 따옴표의 처리이다. 다음은 그러한 오류의 예이다.

| 원문 | サングクの短編小説『偶像の涙』は、暴君のように振舞う生徒を、力ではなく知略によって追放する担任教師を通じ、 |
|---|---|
| S | 상국의 단편 소설 ' 우상의 눈물'은 폭군처럼 행동하는 학생을, 힘이 아니라 지략에 의해 추방하는 담임 교사를 통해 |
| N | 상크의 단편소설 「우상의 눈물」은, 폭군과 같이 행동하는 학생을, 힘은 아니고 지략에 의해서 추방하는 담임 교사를 통해서 |
| G | 산구쿠 단편 소설 **「우상의눈물」**은폭군처럼행동하는학생을힘이아닌지략에의해추방하는담임교사를통해 |
| 원문 | ネット系メディアの「購讀無料の廣告モデル」から距離を置き 「有料メディア」を重視する姿勢を鮮明にしている。 |
| S | 넷을 계 언론의 **"구독 무료의 광고 모델)**에서 거리를 두고 " 유료 미디어 "을 중시하는 태도를 천명하고 있다. |
| N | 넷계 미디어의「구독 무료의 광고 모델」로부터 거리를 두어「유료 미디어」를 중시하는 자세를 선명히 하고 있다 |
| G | 네트워크 기관 언론 "구독 무료 광고 모델"에서 거리를 두고 "유료 매체"를 중시하는 자세를 선명하게 하고있다. |

⑤ 단순 문법이나 규칙에 취약

통계 방식의 기계 번역은 단순 문법이나 규칙에 의해 쉽게 처리할 수 있는 부분에 대해 취약한 경우가 많다. 다음의 예는 단순한 규칙에 의해 쉽게 처리될 수 있으나 통계기반의 번역에서 종종 나올 수 있는 번역 오류이다. 첫 번째 문장의 경우에는 "보고 있고 있었다"로 "고 있다"가 두 번이나 사용되었다. 두 번째의 경우에는 "-되다"의 사역 표현이 제대로 생성되지 못한 경우이다.

| 원문 | 再現した展示に見入っていた。 |
|---|---|
| S | 재현한 전시를 열심히 보고 있고있었다 |
| N | 재현한 전시에 주시하고 있었다. |
| G | 재현한 전시 주시하고 있었다. |
| 원문 | 외신들은 김연아의 금메달이 확정된 직후인 이날 오후 1시 54분~56분 |
| S | 外國メディアはキム·ヨナの金メダルが確定した直後, 同日午後1時 54分~ 56分 |
| N | 外信たちはキム · ヨナの金メダルが確定された直後のこの日午後 1時 54分~56名 |
| G | 外信はキムヨナの金メダルが確定した直後の同日午後1時54分~56分 |

⑥ 문장 부호 쓰임

일본어의 경우 띄어쓰기가 없기 때문에 한국어에 비해 상대적으로 쉼표를 많이 사용한다. 이와 같은 정확한 문장 부호의 사용은 문장의 전체적인 구조를 파악해야지만 가능하다. 그러나 통계 기반의 번역에서는 전체적인 문장의 구조를 파악할 수 없기 때문에 문장 부호 사용의 정확도가 매우 가변적일 수 있다.

| 원문 | このような経験を基に、今回は、こうすれば、上手に報告書が書けるのではないか、というポイントをまとめてみました。 |
|---|---|
| S | 이러한 경험을 바탕으로 이번에는 이렇게 하면, 능숙하게 보고서가 쓸 수 있는 게 아닐까 하는 점을 정리해 보았습니다. |
| N | 이러한 경험을 기본으로, 이번은, 이렇게 하면, 능숙하게 보고서를 쓸수있는것은아닌지, 라고 하는 포인트를 정리해 보았습니다. |
| G | 이러한 경험을 바탕으로, 이번에는 이렇게하면, 잘보고가 쓸 것이 아닌가라는 점을 정리해 보았습니다. |

### 5.5 한/일 양방향 SMT 실용화 방안

① 대량의 병렬 코퍼스 수집: 양질이면서 대량의 문장 단위 병렬 코퍼스를 얼마나 구축할 수 있느냐가 SMT의 성패를 좌우한다. 지속적으로 문장 단위의 병렬 코퍼스를 수집할 수 있는 방안을 마련해야 한다.

② 문장 자동 정렬기의 정확도: 잘못된 정렬의 주된 원인은 문장 정렬이 잘못 이루어져 있었기 때문이다. 문장 자동 정렬기에서 정확하게 문장 정렬을 수행해야 한다.

③ 영역(domain)에 따른 병렬 코퍼스 수집, 학습, 적용: 통계기반의 번역기 구축 시, 대상 영역에 따라 나누어 코퍼스를 수집하고, 영역에 따라 나누어 학습시킨다. 번역하고자 하는 대상 문서가 입력으로 들어 왔을 때, 우선 대상 문서의 영역을 자동으로 파악하고 영역에 맞는 번역 패턴과 언어 모델을 이용하여 번역을 수행한다면, 번역 어휘 선정의 정확도를 높일 수 있다.

④ 형태소 분석기의 지속적인 튜닝: 형태소 분석기가 사용하는 사전 및 규칙을 지속적으로 보강하고 형태소 분석기를 꾸준히 튜닝해야 한다.

⑤ 원문에 대한 전처리 작업: 두 언어의 원문에 대한 전처리 작업을 수행해야 한다. 개체명을 우선 인식하여 하나의 덩어리로 처리할 수 있어야 한다. 또한 숫자, 날짜, 시간, 화폐 등의 표현에 대해 정규화 작업을 통해 하나의 단일화된 표현으로 변환시켜야 한다. 또한, 가장 어려운 문제 중의 하나인 문장 내에 내포된 문장이나 구절에 대한 처리가 필요하다. 특히 괄호 안에 쓰인 문장이나 긴 구절에 대해서 어떻게 처리할 것인지에 대해 고려해야 한다. 통계 기반의 번역기의 경우에는 문장의 구조를 파악할 수 없기 때문에 내포 문장과 바깥쪽 문장과의 관계를 고려할 수 없다. 그러다 보니 문장의 경계를 구분 짓는 심볼이 생략되기도 하여 이해하기 어려운 번역이 되는 경우가 있다.

⑥ 목적 언어에 대한 후처리(post-processing): 통계 기반의 번역기로 생성한 목적 문장에 대해 튜닝 작업을 수행함으로써 번역 결과를 향상시킬 수 있다. 예를 들어, 형태소 생성 규칙을 도입하여 형태소 생성을 점검하고 여는/닫는 괄호나 따옴표의 사용 등을 규칙을 적용하여 점검해 본다.

⑦ 구절 정렬 테이블에 대한 필터링: 불필요한 정렬이나 부정확한 정렬을 찾아내어 제거한다. 또한 잘못된 정렬 패턴을 수정하여 번역 결과를 향상시킬 수 있다.

“http://nlp.cnu.ac.kr/~smt/web/smt_test_POS.php”에서 본 연구에서 제안된 통계 기반의 한/일 양방향 기계 번역의 결과를 살펴볼 수 있다.

## 6. 결론 및 향후과제

본 연구에서는 통계적 자동 번역 기술을 이용한

한/일 양방향 기계 번역 시스템을 구축하였다. 비교적 짧은 개발 기간에 기존의 규칙 기반 번역기의 성능에 근접한 성과를 얻을 수 있었다. 한국어와 일본어 사이에 언어 유사성이 있었기 때문에 통계 기반의 번역 방식이 더 잘 적용될 수 있었다. 통계 기반의 한/일 양방향 번역이 더 높은 성능을 얻기 위해서는 각 처리 단계에서 더욱 엄격한 기준을 적용하여 노이즈를 걸러 낼 수 있어야 한다. 부정확한 학습 데이터를 배제하고 양질의 학습 데이터를 대량으로 꾸준히 공급받을 수만 있다면 규칙 기반 번역기의 성능을 쉽게 뛰어 넘을 수 있을 것이다.

한/영이나 한/중과 같이 언어 유사성이 적고 어순이 많이 달라지는 언어쌍에 대해서는 한/일 SMT 번역과 같은 성능을 빠른 시간에 얻기 힘들 것으로 예상된다. 이런 언어쌍의 경우에는 inter-phrase의 어순 변화를 처리하기 쉽게 고안된 다른 디코더 도입을 고려해 보아야 한다. 또한 구절 단위의 번역이 아닌 구문(syntactic) 단위의 번역이나 트리(tree) 기반의 번역을 고려해 보아야 한다.

## 후 기

본 연구는 2010년 충남대학교 공무국외여행 학술연구비에 의해 지원되었음.

## 참고문헌

[1] http://www.statmt.org/moses/

[2] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, “The mathematics of statistical machine translation: Parameter Estimation”, Computational Linguistics, vol. 19, no. 2, pp. 263-311, 1991.

[3] Yun Kim and et al, “The trends of machine translation techonology and case study”, Electronics and Telecommunications Trends, vol. 23, no. 1, 2008.

[4] U. Germann, “Greedy decoding for statistical machine translation in almost linear time”, Proceedings of HLT-NAACL pp. 1-8, 2003.

[5] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models”, Computational Linguistics, vol. 29, no. 1, pp. 19-51, 2003.

[6] K. Yamada, and K. Knight, “A syntax-based statistical translation model”, Proceedings of The Association for Computational Linguistics 2001, pp. 523-530, 2001.

[8] R. Rosenfeld, “Two decades of statistical language modeling: where do we go from here?”, Proceedings of IEEE, vol. 88, no. 8, pp. 1270-1278, 2000.

[10] The 21st Century Sejong Project, http://sejong.or.kr/sejong_kr/index.html, 2006.

[11] Xiaoyi Ma, “Champollion: A robust parallel text sentence aligner”, Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genova, Italy, 2006.

[12] C. Bannard and C.B. Callison, “Paraphrasing with Bilingual Parallel Corpora”, Proceedings of The Association for Computational Linguistics 2005, pp. 597-604, 2005.

[13] Y.S. Hwang, Y.K. Kim, and S.K. Park, “Paraphrasing depending on bilingual context toward generalization of translation knowledge”, Proceedings of the Third Int’l Joint Conf. on Natural Language Processing, pp. 327-334, 2008.

[14] Daniel Marcu and William Wong, “A phrase-Based, joint probability model for statistical machine translation”, Proceedings of Empirical Methods on Natural Language Processing 2002, pp. 133-139.

[15] Nicola Ueffing and Hermann Ney, “Using POS information for statistical machine translation into morphologically rich languages”, Proceedings of The European Chapter of the Association for Computational Linguistics 2003.

[16] Philipp Koehn, Franz Josef Och and Daniel Marcu, “Statistical phrase-based translation”, Proceedings of Human Language Technologies-the North American Chapter of the Association for Computational Linguistics 2003, pp. 347-354.

[17] Eleftherios Avramidis and Philipp Koehn, “Enriching morphologically poor languages for statistical machine translation”, proceedings of Association for Computational Linguistics 2008, pp. 763-770.