한국어의 이형태 표준화를 통한 구 기반 통계적 기계 번역 성능 향상

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

Improve Performance of Phrase-based Statistical Machine Translation through Standardizing Korean Allomorph

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

요 약

한국어는 형태론적으로 굴절어에 속하는 언어로서, 어휘의 형태가 문장 속에서 문법적인 기능을 하게 되고, 형태론적으로 풍부한 언어라는 특징 때문에 조사나 어미와 같은 기능어들이 다양하게 내용어들과 결합한다. 이와 같은 특징들은 한국어를 대상으로 하는 구 기반 통계적 기계번역 시스템에서 데이터 부족 문제(Data Sparseness problem)를 더욱 크게 부각시킨다. 하지만, 한국어의 몇몇 조사와 어미는 함께 결합되는 내용어에 따라 의미는 같지만 두 가지의 형태를 가지는 이형태로 존재한다. 따라서 본 논문에서 이러한 이형태들을 하나로 표준화하여 데이터부족 문제를 완화하고, 베트남-한국어 통계적 기계 번역에서 성능이 개선됨을 보였다.

주제어: 한국어 통계적 기계번역, 기계학습, 데이터부족문제, 한국어 이형태

1. 서론

번역이란 자연언어를 대상으로 한, 특정 언어(원시언어)에서 같은 의미를 가진 다른 언어(대상언어)를 생성하는 작업이며, 기계번역이란 컴퓨터를 통해 자동화된 번역 과정을 의미한다. 오늘날 국제화시대가 도래하면서 국가 간 교류가 활발해지고, 이로 인해 대량의 문서를 빠르고 안정적으로 번역해 줄 수 있는 기계번역의 중요성은 더욱 부각되고 있다.

초기 기계번역 방법론은 규칙에 기반을 두었으나, 언어쌍마다 규칙을 정의하는 데 많은 비용이 드는 단점이 있었다. 따라서 예제 기반, 통계 기반 방식을 거쳐 최근에는 인공신경망 기반 방식이 연구되고 있다. 하지만 인공신경망 기반 방식은 속도 문제 등 아직 상용화 단계에 이르지 못해, 아직까지 통계 기반 방식이 주로 사용되고 있다.

통계적 기계번역은 대량의 병렬 말뭉치로부터 학습시킨 확률모델을 통해 가장 높은 확률을 가지는 번역문을 생성시키는 방법이다. 하지만 대량의 병렬 말뭉치는 구축하는데 많은 비용이 들고, 충분하지 않은 병렬 말뭉치는 데이터 부족 문제(data sparseness problem)를 일으킨다.

일반적으로 데이터 부족 문제는 가짓수가 많고 등장 횟수의 편차가 심한 내용어에서 더욱 부각되고 있는 것으로 알려져 있으나, 구 기반 통계적 기계번역(Phrase-Based Statistical Machine Translation)의 번역 단위인 구(phrase)는 내용어와 기능어를 함께 포함하며, 특히 한국어와 같이 형태론적으로 풍부한(Morphologically-rich) 언어는 다양한 조사나 어미 등의 기능어가 내용어와 결합하기 때문에, 기능어에서도 데이터 부족 문제가 발생할 수 있다.

본 논문에서는 이러한 문제점들을 보완하기 위해 한국어에 존재하는 결합되는 어휘에 따라 뜻은 같지만 모양이 다른 이형태(異形態, Allomorph)들을 추출하여 하나의 형태로 표준화시키고 형태소 단위의 번역을 적용하여 최종적으로 번역 성능이 개선됨을 보이고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 통계적 기계번역에서의 데이터 부족 문제를 해결하기 위한 관련 연구를 살펴본다. 3절에서는 구 기반 통계적 기계번역 시스템에 대해 간단히 설명한다. 4절에서는 이형태의 표준화 및 형태소 단위의 기계번역의 방법 및 필요성에 대해 설명한다. 5절에서는 실험 및 평가에 대해 분석하고, 마지막으로 6절에서 결론 및 추후연구에 대해 다룬다.

2. 관련 연구

Zipf's law[1] 에 따르면, 문서에 등장하는 단어들의 분포는 일반적으로 고르지 않고 자주 등장하는 단어에 편향되는 특성을 가진다. 이러한 특성으로 인해, 통계적

기계번역의 확률모델이 등장빈도가 높은 단어들로 편향되어 학습되고, 이러한 단어들이 우선적으로 원시언어에 대한 번역으로 채택될 수 있다. 따라서 자주 등장하지 않는 단어들이 충분한 분포를 가지기 위해선 대량의 학습말뭉치가 필요로 하지만, 제한된 학습말뭉치는 데이터 부족 문제를 야기할 수 있다.

데이터 부족문제를 완화하기 위한 연구로 [2]에서는 영-한 통계적 기계번역에서 서로 다른 영어 구 $(e_1, e_2, ..., e_n)$ 가 하나의 한국어 구(c)로 정렬(번역)되는 즉, 같은 의미를 가지지만 여러 가지로 표현될 수 있는 패러프레이즈(paraphrase)들을 추출하고 여기에 가중치를 부여하여 특정 구로 번역되도록 함으로서 데이터부족 문제를 완화하였다.

[3]에서는 알파벳을 사용하는 언어에서 뜻은 같지만 대, 소문자의 차이를 가지는 단어들을 소문자 혹은 대문자로 통일시켜주는 Truecasing 방법을 이용해 데이터부족 문제를 완화하여 통계적 기계번역 시스템의 성능을 개선시켰다.

3. 구-기반 통계적 기계번역

3.1 통계적 기계번역

통계적 기계번역은 두 개의 확률모델인 번역모델(translation model)과 언어모델(language model)을 이용하는 방법으로 입력문장에 대한 번역문장(e_{best})은 베이즈 정리(Bayes' theorem)[4]를 통해 (식1)와 같은 조건부확률로 부터 결정된다.

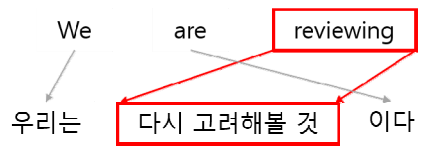
$$e_{best} = \arg\max_e p(e|f) = \arg\max_e p(f|e)p(e) \quad \text{(식 1)}$$

(식 1)에서 f는 원시문장(source sentence), e는 f에 대한 대상문장(target sentence)이며 $p(f|e)$는 번역모델, $p(e)$는 언어모델을 뜻한다.

3.2 구-기반 통계적 기계번역

통계적 기계번역에서의 단어(Word)의 정의는 번역의 최소단위를 뜻한다. 가령 영어의 경우 띄어쓰기를 기준으로 단어를 정의할 수 있고 한글의 경우 어절 혹은 형태소가 단어로 정의될 수 있다.

기계번역에서의 구(Phrase)는 언어학적 구를 의미하는 것이 아니며 (그림 1)와 같이 한 번에 번역될 수 있는 한개 이상의 단어열(Word sequence)을 말한다. 이와 같은 구를 기본 단위로 번역을 시도하는 것을 구-기반 통계적 기계번역이라 하며 구는 학습 말뭉치에서 단어정렬 모듈(Word-alignment)[5]로부터 추출된다.



(그림 1) 추출된 영어, 한국어 구 예시

3.3 번역모델

번역모델은 병렬 학습말뭉치로 부터 구축되며 단어정렬모듈을 통해 추출된 구는 번역확률과 함께 구 번역 테이블(Phrase-table)에 저장된다.

번역모델의 확률 $p(f|e)$는 e로 번역되는 f의 확률을 말하며 (식 2)와 같이 최대우도측정(MLE) 방법으로 구할 수 있다.

$$p(f|e) = \frac{\text{count}(e,f)}{\sum_f \text{count}(e,f)} \quad \text{(식 2)}$$

3.4 언어모델

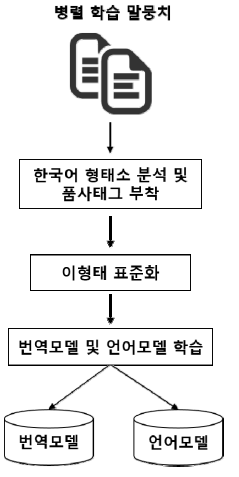
언어모델은 단일 언어 말뭉치(monolingual corpora)로부터 구축되며 어떤 번역문장(w_i)가 이전에 번역된 문장 $(w_0, w_1, ... w_{i-1})$과 결합했을 때 얼마나 자연스러운 지에 대한 확률을 나타낸다. 하지만 이전에 변역된 모든 문장을 볼 수 없기 때문에 보통 n-gram 방식이 사용되며 언어모델의 확률 $p(e)$는 (식 3)와 같이 최대우도측정 방법으로 구할 수 있다.

$$p(e) = p(w3|w1,w2) = \frac{count(w1,w2,w3)}{\sum_w count(w1,w2,w)} \quad \text{(식 3)}$$

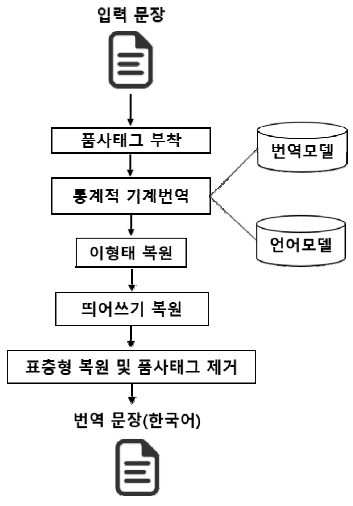
4. 제안하는 방법

본 절에서는 한국어를 대상으로 하는 구-기반 통계적 기계번역에서 발생하는 데이터부족 문제와 연산의 복잡도를 완화해줄 수 있는 방법으로 형태소 단위의 번역과 이형태의 표준화 방법을 제시한다.

통계적 기계번역 시스템 구축에 대한 전반적인 과정은 (그림 2)와 같고 번역문(한국어) 출력에 대한 과정은 (그림 3)와 같다.



(그림 2) 통계적 기계번역 시스템 구축과정



(그림 3) 번역문 출력 과정

4.1 형태소 단위의 번역

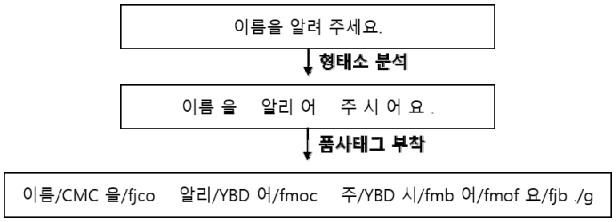
통계적 기계번역에서 적절한 단어를 정의하는 것도 데이터 부족문제를 완화시켜주는데 영향을 줄 수 있다. 한

국어는 하나 이상의 형태소가 결합된 어절을 기준으로 띄어쓰기가 이뤄지는데, 한국어의 특성상 기능어에 해당하는 조사나 어미가 내용어와 다양하게 결합하여 수많은 형태의 어절을 생성하여 문장에서 문법적인 역할을 한다.

따라서 어절단위의 번역은 데이터 부족문제를 야기할 수 있어 최소의미를 가지는 형태소를 단어로 정의하여 확률모델을 학습하는 것이 적절하다.

또한, 각 형태소에 품사태그(Part-Of-Speech tag)를 부착함으로서 동음이의어를 구분할 수 있고 번역모델 학습과정 중 구 정렬에서 더욱 적절한 구를 추출해 낼 수 있어 번역 품질을 향상시켜 줄 수 있으며[6] 이형태를 탐지하고 복원하는데 중요한 역할을 한다.

이와 같은 일련의 처리 과정의 예는 (그림 4와)와 같으며 학습말뭉치를 구축하는 전처리 단계에서 적용된다.



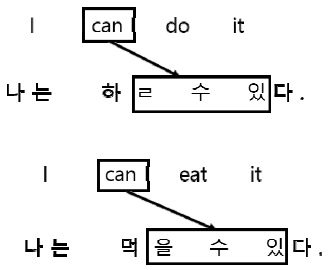
(그림 4) 학습말뭉치의 한국어 처리과정

(그림 4)의 형태소 분석 및 품사태그는 포항공과대학교 지식 및 언어 공학 연구실에서 제작한 KoMA 형태소 분석기 및 KLE 품사태그[7]를 사용하였다.

4.2 이형태 표준화의 필요성

이형태란 뜻은 같지만 환경에 따라 다른 형태를 가지는 형태소를 말하며 예를 들어 한국어의 경우 '-이/-가', '-을/-를' 영어의 경우 '-d/-ed' 와 같은 것들을 말한다.

이러한 이형태는 한국어의 조사나 어미에서 매우 다양하게 존재하여 구-기반 통계적 기계번역 시스템의 번역모델과 언어모델에서 데이터부족 문제를 발생시킬 수 있다. 가령 (그림 5)처럼 'can'이 'ㄹ 수 있' 혹은 '을 수 있'으로 구 정렬이 될 경우 번역모델은 두 개의 구에 대해 각각 확률 값을 가지게 된다. 결과적으로 같은 의미를 가지는 구에 대한 확률 값이 양분되어 데이터부족 문제가 발생하고, 구 번역 테이블의 엔트리가 증가해 연산복잡도가 증가할 수 있다.



(그림 5) 이형태로 번역되는 예

이러한 문제는 N-gram 방식의 언어모델에서도 동일하게 발생할 수 있다. 예를 들어, (그림 6)에서처럼 이전에 등장하는 단어는 같지만 '-ㅂ니다/-습니다' 의 이형태로 인해 확률이 분산될 수 있기 때문이다.

사 랑 하 ㅂ 니 다.

사 랑 하 겠 습 니 다.

(그림 6) 언어모델에 등장하는 이형태 예

따라서 이형태들을 하나의 형태로 표준화하는 방법을 적용하여 확률모델 학습과정에서 발생하는 데이터부족 문제를 완화시키고, 번역문 출력과정에서 연산복잡도를 감소시키는 효과를 기대할 수 있다.

4.3 이형태 표준화 방법

우선, 앞으로 언급될 이형태의 표준형이란 아래에서 설명하는 방법을 통해 하나의 형태로 변환된 이형태를 말하는 것으로 정의한다.

한국어의 이형태는 조사와 어미에서 존재하며 이들을 표준화 하는 방법은 간단하다. 첫 번째로, 조사는 결합하는 체언의 종성 유무에 영향을 받는다. 예를 들어, '-과/-와' 의 경우 전자는 종성이 존재하는 체언과 후자는 종성이 없는 체언과 결합한다(예: 형님**과**, 친구**와**). 본 논문에서는 총 11개의 이형태를 가지는 조사에 대해 종성이 없는 체언과 결합하는 형태로 표준화하는 방법을 적용하였으며 (표 1)은 그 일부분에 대한 예이다.

(표 1) 이형태를 가지는 조사 예

이형태를 가지는 조사	표준화된 조사
-과/-와	**-와**
-이랑/-랑	**-랑**
-은/-는	**-는**
-이여/-여	**-여**
-이/-가	**-가**
-을/-를	**-를**

두 번째로, 어미는 그 종류가 매우 많아 용언과 다양하게 활용한다. 어미의 이형태는 용언 어간이 자음으로 시작하는 경우, 종성 유무에 영향을 받는 부류와 모음의 종류(양성 혹은 음성)에 영향을 받는 부류가 있다.

전자는 용언 어간의 끝 음절의 종성유무('ㄹ' 제외)에 영향을 받는다. 예를 들어, '-ㅂ니다/-습니다' 의 경우 '가+ㅂ니다: 갑니다' '먹+습니다: 먹습니다' 와 같이 활용한다. 이와 같은 특징을 가지는 174개의 이형태들에 대해 종성이 없는 용언과 결합하는 형태로 표준화 하는 방법을 적용하였으며 (표 2)는 그 일부분에 대한 내용이다.

(표 2) 어간의 종성 유무에 영향을 받는 어미 예

이형태를 가지는 어미	표준화된 어미
-습니다/-ㅂ니다	**-ㅂ니다**
-으랍니까/-랍니까	**-랍니까**
-으니까/-니까	**-니까**
-으니/-니	**-니**
-는다/-ㄴ다	**-ㄴ다**
-으려든/-려든	**-려든**

후자는 용언 어간의 끝 음절 모음에 영향을 받는다. 예를 들어 '-아/-어'의 경우 용언이 양성모음('ㅏ, ㅑ,ㅗ')으로 끝나면 '-아'와 결합하고, 그 외의 경우 '어'와 결합한다. 단, 끝 음절이 'ㅡ'일 경우 단음절의 용언 어간에는 '-어'가 결합하고, 2 음절 이상의 용언 어간에는 그 앞 음절의 모음에 따라 활용한다.

이 부류에 대한 예는 (그림 7)와 같으며 총 11개의 어미에 대해 양성모음이 아닌 용언과 결합하는 형태로 표준화를 하였고 (표 3)은 그 일부분에 대한 내용이다.

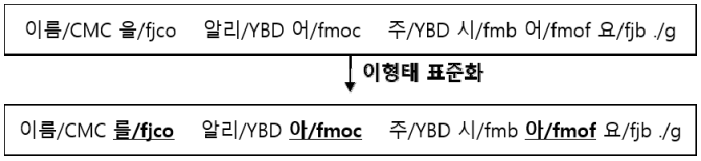
(표 3) 어간의 모음 종류에 영향을 받는 어미 예

이형태를 가지는 어미	표준화된 어미
-아/-어	**-어**
-아다/-어다	**-어다**
-아도/-어도	**-어도**
-아서/-어서	**-어서**
-아야/-어야	**-어야**
-아야만/-어야만	**-어야만**

막- + -아 → 막아, 얇- + -아 → 얇아
개- + -어 → 개어, 겪- + -어 → 겪어
끄- + -어 → 꺼, 바쁘- + -아 → 바빠

(그림) 7 어간의 모음에 영향을 받는 어미 예

이와 같은 일련의 처리과정을 거쳐 생성된 학습 말뭉치의 형태는 (그림 8)와 같다.



(그림 8) 이형태 표준화를 거친 말뭉치 예

4.4 이형태 복원

이형태가 표준화된 상태로 학습된 확률모델로부터 출력된 번역문은 이형태 복원이 필요하다.

이형태 복원은 변환에 비해 다음과 같은 점을 추가로 고려해야 한다.

먼저 조사는 활용이 일어나지 않기 때문에 4.3에서 언급된 규칙, 즉 결합된 내용어의 종성유무에 의해 간단히 복원가능하다.

반면 어미는 어간이 불규칙 활용을 가질 경우 예외처리가 필요하다. 규칙 어간의 경우 마찬가지로 4.3에서 언급된 자질(종성유무, 모음종류)로 복원한다. 하지만 불규칙 어간의 경우 그 활용형과 사전형 사이에 자질이 달라질 수 있다. 예를 들어 '돕다'의 경우 규칙 어간인 '잡다'와 달리 사전형은 '돕-'으로 유종성이기에 어미 '-음'와 결합해야하지만 활용형은 무종성인 '도우-'로 실질적으로 'ㅁ'이 결합되어 '도움'이 된다. 또한 어간이 합성어일 경우 규칙 어간이어도 예외처리가 필요한데 대표적으로 '본뜨다'의 경우가 있다. 어미 '-아/어'와 결합 시, 규칙에 따르면 양성모음인 'ㅗ'에 따라 '-아'가 결합해야하지만 이는 합성어로 '-아/어'는 '뜨-'와의 결합규칙을 따른다. 따라서 '본' + '뜨+어' = '본떠'가 활용형이 된다.

그 외에도 다음과 같은 추가 고려사항이 있다. 첫 번째로, 종성을 가지는 기호들도 고려를 해야 한다. 예를 들어 숫자 0은 '영'으로 발음될 수 있고, t는 '톤'으로, g는 '그램'으로 발음될 수 있다. 이와 같은 기호들은 종성이 있는 형태소로 취급하여 이형태 복원을 수행해야 한다.

두 번째로, 괄호로 둘러싸인 부가설명문이 등장할 경우 이형태와 연관된 내용어가 무엇인지 감지한 뒤 복원을 수행해야 한다. 예를 들어 다음과 같은 문장 "학교(공부하는 곳)**은** ..."에서 '은'은 '곳'이 아닌 '학교'와 연관되어 이형태 복원을 해야 한다.

5. 실험 및 결과

본 실험에서는 Moses 툴킷[8]과 구 정렬 도구인 GIZA++을 이용하여 베트남어-한국어를 대상으로 구-기반 통계적 기계번역시스템을 구축하였다.

한국어 말뭉치의 형태소 분석 및 품사태그부착에 대한 전처리를 위해 KoMA 형태소 분석기[7]를 사용하였고, 베트남어에 대해서는 호치민대학교의 CLC툴킷[9]을 사용하여 음절 단위 단어분할 및 품사태그 부착을 수행하였다.

학습 및 평가에 사용된 말뭉치는 ㈜Systran 으로부터 제공 받았으며, 문어체와 구어체가 혼용되어 있다. 이에 대한 통계는 (표 4)와 같다.

(표 4) 말뭉치 통계

		한국어	베트남어
번역 모델 학습 말뭉치	문장 수	975,879	
	단어 수	15,166,987	13,237,687
병렬 테스트 말뭉치	문장 수	4,000	
	단어 수	51,625	45,966
언어모델 학습 말뭉치	문장 수	3,860,535	975,879
	단어 수	57,027,195	13,237,687

(표 5)와 (표 6)은 Baseline과 본 논문에서 제안한 모델의 번역문을 형태소 단위로 BLEU, RIBES, TER을 측정한 결과이다. baseline은 형태소 분석과 품사태그만을 이용한 모델이다.

(표 5) 베트남어-한국어 실험 결과

No.	실험환경	BLEU	RIBES	TER
1	Baseline	15.5	0.727	69.3
2	**제안하는 방법**	**16.5**	**0.741**	**67.6**

(표 6) 한국어-베트남어 실험 결과

No.	실험환경	BLEU	RIBES	TER
1	Baseline	16.7	0.759	59.2
2	**제안하는 방법**	**14.2**	**0.710**	**69.4**

베트남어-한국어 방향 번역에서는 모든 평가지표가 향상됨을 확인할 수 있었다. 반면 한국어-베트남어 방향 번역에서는 오히려 번역 성능이 하락했는데 다음과 같은 이유로 추정된다.

- 베트남어는 고립어로서 기능어의 역할을 어순이 대신하므로 한국어의 기능어는 NULL로 정렬된다. 따라서 기능어에 대한 기본형 변환이 최종 BLUE 점수 향상에 도움을 주지 않고 오히려 오류가 전파되어 성능이 떨어졌을 수 있다.
- 베트남어 언어모델 학습에 사용된 베트남어 단일언어 말뭉치 크기가 한국어 단일언어 말뭉치 크기에 비해 약 97만 문장으로 한국어 단일언어 말뭉치 크기(386만)에 비해 약 4배 적어 성능이 하락했을 수 있다.

6. 결론

데이터부족 문제는 기능어보다 내용어에서 빈번히 발생하지만, 구-기반 통계적 기계번역은 특성 구 단위로 번역을 하며, 한국어는 다양한 기능어가 내용어와 결합하므로, 기능어 개수를 줄이는 것이 데이터부족 문제에 도움이 될 것이라는 취지로 이형태 표준화 방법을 제안하였고, 실제로 베트남어-한국어 번역 성능이 향상되었음을 보였다.

하지만, 반대 방향(한국어-베트남어) 번역에서는 성능 변화가 거의 없었는데, 그 이유는 베트남어 단일언어 말뭉치 크기가 작으며, 베트남어는 고립어에 속하기 때문에 한국어의 조사나 어미가 NULL로 정렬이 되는 경우가 빈번하기 때문이라 예상한다.

본 논문은 기능어에 대해 일반화를 함으로써 데이터 부족 문제를 완화했다. 향후 개체명 인식 혹은 품사 정보를 활용하는 등 상대적으로 데이터 부족문제에 큰 영향을 미치는 것으로 알려진 내용어에 대해서도 이와같은 연구가 이루어진다면 보다 큰 성능 향상이 이루어질 것으로 기대한다.

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] Li, Wentian. "Random texts exhibit Zipf's-law-like word frequency distribution." IEEE Transactions on information theory 38.6 (1992): 1842-1845.

[2] 이형규, 김민정, and 임해창. “이중 언어 기반 패러프레이즈 추출을 위한 피봇 차별화 방법.” 인지과학 22.1 (2011): 57-78

[3] Lita, Lucian Vlad, et al. "Truecasing." Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.

[4] Mr. Bayes, and Mr Price. "An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs." Philosophical Transactions (1683-1775) (1763): 370-418.

[5] Zens, Richard, Franz Josef Och, and Hermann Ney. "Phrase-based statistical machine translation." Annual Conference on Artificial Intelligence. Springer Berlin Heidelberg, 2002.He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[6] Lee, Jonghoon, Donghyeon Lee, and Gary Geunbae Lee. "Improving phrase-based Korean-English statistical machine translation." INTERSPEECH. 2006.

[7] 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 (1999): 76-87.
[8] 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. Association for Computational Linguistics, 2007.
[9] Nghiem, Minh, Dien Dinh, and Mai Nguyen. "Improving Vietnamese POS tagging by integrating a rich feature set and Support Vector Machines." Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on. IEEE, 2008.