한국어-몽골어 구 기반 번역시스템에 대한 연구

김영미
몽골 후레 정보통신 대학교
컴퓨터과학과
e-mail : ymkim@hureeict.edu.mn, gloriaym@yahoo.co.kr

A Study of Korean-Mongolian Phrase-based Machine Translation System

Young-Mi Kim
Dept. of Computer Science, Huree University of ICT

요 약

한국어-몽골어 구 기반 기계번역시스템은 몽골어와 한국어간의 양방향 기계 번역시스템으로서 개발중인 시스템이다. 두 언어의 구조적 특성이 유사한 점에 기안하여 직접기계번역방식에 구단위 번역과 예제에 기반한 번역방식을 병행하여 문장단위의 번역이 가능하다.

1. 서론

최근 몽골과 한국의 교류가 활발해짐에 따라 두 나라간의 문서 번역의 요구가 많아지고 있다. 이에 한국어와 몽골어간의 번역 시스템에 관해 연구하고 후레정보통신 대학내의 공문서 번역을 예로 들어 구현하고 있는 바를 이 논문에 담고자 한다.

2. 한국어와 몽골어의 유사점 및 차이점

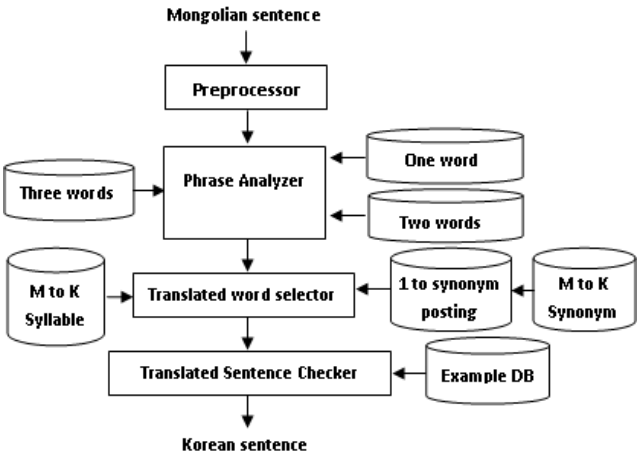
몽골어 문장은 한국어처럼 주어+목적어+서술어의 순서로 되어 있고 서술어 수식이 서술어 앞에 온다. 또한, 어절은 어간과 어미로 구성되어 있고 격조사를 사용하는데, 어간의 모음의 종류에 따라 어미의 형태가 달라지는 모음조화 현상이 있다.

반면, 몽골어는 7 개의 모음과 28 개의 자음으로 구성된 키릴문자를 사용하여 표기하며 모음과 자음의 갯수가 한글보다 많아서 한글로 표기할 때 중복표기 현상이 나타나게 된다. 또한 장모음 표기가 사용되고 한국어에서와 같은 존칭어 표현은 거의 없다.

이 논문에서는 몽골어와 한국어의 구조가 유사한 점에 기안하여 기계번역방식중 직접번역방식을 채택하였다. 또한, 모음조화 현상에 따라 많은 종류의 어미들이 사용되고 있고 한자어가 몽골어로 번역될 때, 두 단어 이상의 합성어로 표현되기 때문에 구단위 번역방식을 사용하였다.

3. 시스템 구성

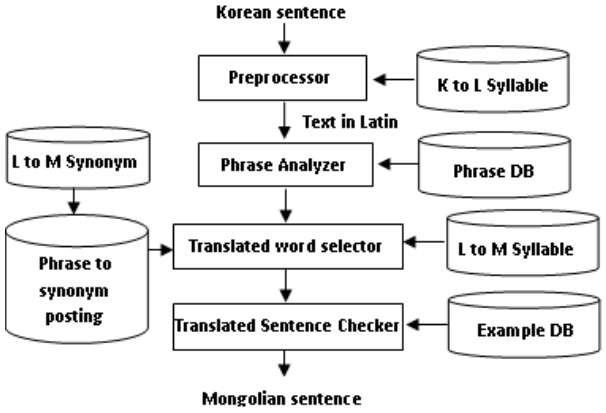
그림 1 은 몽한 번역시스템의 구성을 보여준다. 몽골어에서 한국어로의 번역과정은 네 단계를 거치는데 첫번째는 Preprocessor 단계로서 구단위 번역을 위해 문장을 어절로 절단하는 단계이다. 두번째는 Phrase Analyzer 단계로서 구단위 번역을 위해 Three words DB, Two words DB, One word DB 를 참조하여 절단된 어절들을 번역한다. 세번째는 Translated word selector 단계로 다의어의 경우 문장에 맞는 단어를 선택하고 만약 지명이나 인명인 경우는 그대로 한글로 옮기는 단계이다. 마지막은 Translated Sentence Checker 단계로서 예제문장을 참조하여 번역되어 나온 문장을 체크하고 문법에 맞게 수정하여 출력한다.



(그림 1) 몽한 번역시스템

그림 2 는 한몽 번역시스템의 구성을 보여준다. 한국어에서 몽골어로의 번역과정도 네 단계이다. 첫번째 단계는 Preprocessor 단계로서 입력된 한국어 문장을 라틴어로 된 문장으로 바꾸어준다. 두번째 단계는 Phrase analyzer 단계로서 라틴어로 되어진 문장을 구단위로 절단하고 Phrase DB 를 참조하여 구단위로 번역한다. 세번째는 Translated word Selector 단계로서 다의어와 고유명사를 처리하는 단계이다. 네번째는 Translated Sentence Checker 단계로서 구단위로 번역된

문장을 예제문장을 참조하여 체크하고 올바른 몽골어 문장을 출력하게 된다.



(그림 2) 한몽 번역시스템

3.1 Preprocessor

몽한번역에서는 입력된 몽골문장을 어절로 절단하는 단계이다. 어절로 절단된 단어들은 따로 따로 저장되어 다음단계인 Phrase Analyzer 단계에서 구단위 번역을 위한 입력자료가 된다.

한몽번역의 경우는 검색의 용이를 위해 K to L syllable DB 를 사용하여 한글을 라틴어로 바꾸어 준다. 표 1 은 "고프다"는 어절을 라틴어로 바꾸기 위해 참조하는 K to L Syllable DB 의 인용 예로서 "고프다"는 "kopheuta"로 변환된다.

<표 1> K to L Syllable DB 예

Korean	Latin
고	ko
프	pheu
다	ta

3.2 Phrase Analyzer

몽골어와 한국어가 구조적으로 유사점이 많으나 모음조화현상에 의해 어미의 형태가 달라지기 때문에 형태소로 나누어 번역할 경우 정확한 철자를 기대하기 어렵다. 또한, 두 단어나 세 단어로 형성된 합성어가 많아서 어절 단위로 번역하면 어색하고 애매한 번역이 되기 쉽기 때문에 이 시스템에서는 구단위 번역기법을 사용하였다.

몽한번역의 경우는 번역의 용이를 위해 한단어로 된 DB, 두단어로 된 DB, 세단어로 된 DB 를 따로 구축하여 입력문장을 세 단어씩 잘라 검색하고 없으면 두 단어, 한 단어 순으로 검색한다. 보통 몽골어 합성어의 경우 3 단어까지로 이루어져 있기 때문이다. 표 2 표 3, 표 4 는 각각 세 단어 DB, 두 단어 DB, 한 단어 DB 의 예를 보여준다.

<표 2> Three words DB 예

Id	Mon_exp	Kor_exp	Tag
1	Уралдааны үйл ажиллага	경연대회	N
2	Мэдээлэл Холбооны Технологи	정보통신기술	N

<표 3> Two words DB 예

Id	Mon_exp	Kor_exp	Tag
1	Номын сан	도서관	N
2	шилдэг оюутан	우수학생	N

<표 4> One word DB 예

Id	Mon_exp	Kor_exp	Tag
1	цаг	시계	N
2	даваа	월요일	olon
3	үнэтэй	비싸다	Adj0
4	үнэтэй	비싼	Adj1

표 5 는 한몽 Phrase DB 의 예를 보여준다.

<표 5> Phrase DB 예

ID	Latin	Mongolian	Tag
1	kopheuta	өлсөх	V
2	nun	цас	olon
3	be	гэдэс	olon

3.3 Translated word selector

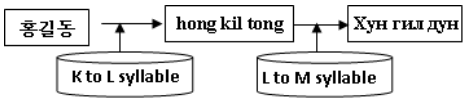
다의어의 경우 Synonym DB 에 저장되어 있는 연관어를 참조하여 문장에 적합한 단어를 선택한다. 표 6 은 몽한 Synonym DB 의 예를 보여준다. 표 4 에서, Id 가 2 인 경우 다의어의 의미로 olon 이란 Tag 를 사용하였다. 이경우 One-to-Synonym Posting table 을 통해 M to K Synonym DB 에서 Id 가 1,2 로 매칭됨을 알 수 있게 된다. 그 다음, Id 가 1 인 연관어를 검색해서 문장에 해당되는 단어가 있으면 "다와"라는 한국어 번역을 선택하게 되고 Id 가 2 인 연관어에 해당되면 "언덕"이라는 번역을 찾아내게 된다. 둘 다 연관어가 없으면 표 4 에서 Id 가 2 인 "월요일"이라는 번역으로 대체한다.

<표 6> M to K Synonym DB 예

Id	Mon_exp	Kor_exp	Linkage words
1	даваа	다와	Нэр,нэртэй,овог,овогтой,гэдэг
2	даваа	언덕	Цаана,давах,наана,гүвээ

지명이나 인명과 같은 고유명사의 경우 그대로 한글로 옮기기 위해 M to K Syllable DB 를 참조한다. 한글로 옮기는 과정에서 몽골어 어절을 6 자씩 절단하여 M to K Syllable DB 에서 검색하는 방법을 사용하였는데 이는 몽골어 단어의 평균길이가 6 자인 것에서 착안한 것이다.

한몽번역의 경우 "홍길동"이라는 인명을 번역하는 과정은 그림 3 과 같다.



(그림 3) 한국어 인명의 번역과정

3.4 Translated sentence checker

세번째 단계에서 번역된 문장으로부터 모호성을 제거하고 문법구조에 맞게 번역되었는지 체크하고 수정

하는 단계이다. 이를 위해 많은 예제가 담겨진 몽한 Bilingual DB 를 구축하였다. 이 논문에서는 사용범위를 후레대학 공문서로 한정하였기 때문에 수년간 후레대학에서 사용한 공문서를 Example DB 에 담았다.

3.5 Tag

<표 7> 태그의 종류

종류	설명
N	명사
V	동사
olon	다의어
Adj1	형용사
Adj0	형용동사
syl	고유명사
numbi	말로 쓰여진 숫자
num	아라비아 숫자
S	주어

몽골어의 경우, 형용사와 형용동사의 형태가 같은 반면 한국어의 경우는 어미의 형태가 달라지므로 정확한 번역을 위해 형용사와 형용동사를 구분하였다. 표 4 에서 “үнэтэй”라는 형용사가 “цаг”라는 단어의 앞에서 수식하는지 서술어로 쓰이는지에 따라 한국어의 어미가 달라진다. 문장 1 과 2 에서 몽골어 표현은 같은 반면 한국어 표현은 “비싼”과 “비싸다”와 같이 어미의 형태가 달라짐을 알 수 있다.

Үнэтэй цаг　　비싼 시계　　(1)

Цаг үнэтэй　　시계가 비싸다.　　(2)

숫자의 경우, 말로 쓰여진 숫자를 그대로 한국어로 번역하면 번거로움이 있어서 일단 숫자로 바꾸기위해 태그가 필요하였다.

몽골어의 구조가 한국어와 유사하고 격조사를 사용하지만 동사마다 사용하는 격이 한국어로 번역할 경우 일치하지 않는 경우가 있어서 예외처리를 위해 주어를 따로 구분하였다.

4. 결론

한국어와 몽골어간의 번역시스템을 구축하기 위해 연구 조사할 때, 어간과 어미로 구성되어 있는 면이 두 언어간의 공통점이서 형태소단위 번역에 관심이 있었다. 한몽간에는 형태소 단위의 직접번역시스템도 충분히 훌륭한 번역시스템이 될 수 있지만 숙어나 합성어의 경우 자칫 자연스럽지 못한 번역이 되기 쉽고, 현대에 이르러 컴퓨터의 용량이 커지고 있는데다 속도도 빨라지고 있어서 보다 자연스러운 번역을 위해 구단위 번역기법을 사용하였다.

용이한 이식을 위해 프로그램은 자바언어로 작성하였고 데이터베이스는 MySql 을 사용하여 구축중에 있다. 번역시스템이 완성되면 웹사이트에서 무료로 제공할 예정이다.

현재 몽골에서는 한몽, 몽한 번역시스템이 없기 때문에 성능비교를 할 수 없지만 번역할 문장의 범위가 제한되어 있어서 계속하여 많은 예제와 데이터가 수집되고 추가되어야 할 것이다. 또한, 이 번역시스템을 시작으로 더 좋은 시스템이 한국와 몽골에서 개발될 것을 기대해본다.

참고문헌

[1] Tanveer Siddiqui and U.S. Tiwary, Natural Language Processing and Information Retrieval

[2] 황도삼, 최기선,김태석 공역, Natural Language Processing

[3] Purev Jaimai, USE OF COMPUTER-AIDED SYSTEM IN THE STUDIES OF THE MONGOLIAN LANGUAGE: TREE ADJOINING GRAMMARS FOR MONGOLIAN

[4] H.S.PARK, THE KOREAN CORE LANGUAGE ENGINE

[5] Gantur Togtokh, Rule-based Machine Translation from Korean Verb to Mongolian Verb

[6] Euiseok Kang, Design and Implementation of A Korean-English Translation System Using N-Gram

[7] Changhu Kim, Improving Korean-to-Chinese Phrase-based Statistical Machine Translation Using Enhanced Word Alignment