뉴스 타이틀 번역을 위한 중한 기계번역 시스템[1]

황금하*,** 송희정* 김지현* 송영미* 강원석* 서충원* 채영숙* 최기선*
*한국과학기술원 전산학과 전문용어언어공학연구센터, 첨단정보연구센터
**중국 길림성 연변과학기술대학
korterm@korterm.kaist.ac.kr

Chinese-Korean Machine Translation System for News Title Translation

Jin-Xia Huang*,**, Hee-Jeong Song*, Ji-Hyoun Kim*, Yong-Mi Song*,
Won-Sek Kang*, Chong-Won Seo*, Young-Souk Chae*, Key-Sun Choi*
* AITrc, KORTERM, Dept. of Computer Science, KAIST
**Yanbian University of Science and Technology

요 약

본 논문은 근 몇 년간 꾸준히 진행되어진 중한 기계번역시스템에 대한 연구의 기초 위에서, 뉴스 타이틀 번역이라는 특정 도메인에 초점을 맞추어 이의 언어적 특성을 살펴보고, 중한 언어적 유사성에 기반 한 뉴스 타이틀 번역을 위한 중한 기계번역시스템에 대하여 설명한다.

1. 서론

중,한 양국간의 정치, 경제, 문화 등 분야에서의 교류가 날로 활발하여 짐에 따라 중한 기계번역에 대한 수요도 날로 많아지고 있다. 이런 상황에 비추어, 우리는 1998년 말 중한 기계번역 원형 (prototype) 시스템을 구축하여 1999년 9월 MT SUMMIT'99에서 정식으로 데모하였고 (장민, et.al, 1999), 그 후에도 중한 기계번역에 대한 연구를 계속하여 왔다. 기계번역의 실용화와 보다 양질의 번역 결과를 위하여 금년 봄부터는 뉴스타이틀 번역이라는 특정 도메인에 초점을 맞추어 우리의 연구를 진행하기로 하였다.

본 논문은 이미 구축된 MATES/CK (MAchine Translation nvironment System/Chinese-Korean) 원형 시스템의 기초에서, 뉴스 타이틀 번역이라는 특정 분야를 위한 기계번역 방식에 대한 토론을 진행하였다.

2장에서는 중국어와 한국어간의 언어적 차이성을 비교하면서 MATES/CK 원형 시스템에서 문장패턴을 사용 하게된 경위를 설명하였고, 동시에 이런 기존의 문장패턴 기반의 기계번역 시스템의 약점에 대하여 언급하였다. 3장에서는 뉴스타이틀 번역을 위하여 중국어-한국어 사이의 언어적 유사성 및 뉴스타이틀 번역에서의 특유한 언어 현상을 살펴보도록 하고, 이런 언어적 유사성과 뉴스타이틀 번역에서의 특징에 비추어, 4장에서는 뉴스 타이틀 번역을 위한 시스템에 대하여 설명하고, 5장에서는 시스템 구축 현황 및 향후 연구에 대하여 소개하도록 한다.

[1] 본 연구는 첨단정보기술연구센터 과제 "다국어 정보검색 연구"와 전문용어언어공학연구센터 과제 "중국어-한국어 정렬을 통한 번역지식의 획득"을 통하여 과학재단의 지원을 받았음.

2. 중-한 언어적 차이성에 기반 한 기존의 중한 기계번역 시스템

중한 기계번역은 중국어와 한국어 사이의 언어적 차이성 때문에 기계번역에서 많은 어려움을 봉착하게 된다. 이런 언어적 차이성과 기계번역에서의 어려움은 다음과 같다.

(1) 중국어 분석에서의 어려움. 중국어는 단어 분할(segmentation), 품사 부착, 구문구조 분석, 의존관계 분석 및 의미 분석 등 언어분석의 모든 단계에서 아직 극복하지 못하고 있는 어려움들이 존재한다.

(2) 중국어와 한국어는 서로 다른 어족에 속하는 두 개의 언어로, 중국어는 SVO어순인 반면 한국어는 SOV어순이다. 이는 구조변환에서의 어려움으로 나타난다.

(3) 중국어에서는 문법기능을 명시해 주는 조사 사용이 드문 반면 한국어에서는 조사의 형태가 다양하며 일반적으로 조사로 문법적 기능을 나타낸다. 중국어의 구구조내 및 구구조간 의존관계를 밝혀내는데 어려움이 존재하는 상황에서, 이런 차이는 한국어 기능어 생성에 어려움을 주고 있다.

(4) 중국어와 한국어의 단어 단위가 달라 평균 1개의 중국어 단어에 1.9개의 한국어 형태소가 대응되며 일부 단어, 예 하면 중국어에서의 량사 (단위성의존명사)는 한국어에서 생략되는 경우가 많다.

이러한 어려움, 특히 중국어와 한국어간의 구조변환에서의 어려움을 극복하기 위하여, 기존의 중한 기계번역 시스템에서는 문장패턴 기반의 변환방식을 사용하였다. 그리고 이런 문장패턴의 적용률이 낮은 약점을 보완하기 위하여 규칙기반의 변환방식을 보조용으로 사용하였다. 다음은 문장패턴의 예이다:

[중] 增强+Cate1=[n|r|np]+Cate2=[w]
[한] Cate1=[n|r]+을 증강하다+Cate2=[w]

입력된 중국어 문장은 단어분할 및 품사부착 후 구문분석을 거치게 되며, 이런 중국어 입력 문장에 맞는 가장 합당한 문장패턴을 선택하여, 구문구조 변환을 진행한다. 이런 문장패턴 일치 (matching)에서는 정확한 일치을 요구한다. (Zhang & Choi, 1999)

문장패턴 일치과정에서 실패하는 경우, 구문규칙기반의 변환 방식으로 구구조 변환을 진행하는데, 구문분석에서 실패는 규칙기반의 변환을 불가능하게 하기 때문에 곧 번역실패로 이어지게 된다.

이런 방법은 문장패턴의 적용률이 낮은 데에 따른 문장패턴 부족현상이 심각하다는 약점이 있다. 또한 문장패턴에 의한 변환이 패턴 부족으로 실패할 경우 규칙기반에 의한 변환방식에 의존하게 되는데, 이런 규칙기반의 변환방식은 이미 언급된 중국어와 한국어간 구문구조의 차이성 때문에 정확도가 떨어지게 된다.

이런 방법의 또 하나의 약점으로, 문장패턴과 변환규칙의 양분으로 인한 구문구조 변환에서의 정보 부족이다. 예하면, "NP<--r+u"[2] 구조가 사용되는 언어적 표현에는 다음과 같은 예들이 있다:

[VP去(가다)/v[NP 你(너)/r的(의)/u]][3]: 저쪽으로 비켜

[VP是(..이다)/v[NP 你(너)/r 的(의)/u]]: 너의 것이다

위의 예에서, "[NP 你(너)/r 的(의)/u]"는 NP구조 밖에 오는 동사에 따라 완전히 다른 뜻으로 번역된다. 그러나 기존의 변환 규칙으로 이상의 차이를 기술할 수 없기에, 이 구가 하나의 긴 문장에 속하고, 해당 문장에 마침 문장패턴이 없을 경우 (문장패턴의 부족 때문에 이런 가능성은 아주 높다), 정확한 번역은 불가능한 것으로 된다. 물론 "去+你+的→저쪽 비켜"라는 문장패턴을 구축해 놓을 수는 있지만, 문장패턴 일치에서 완전 일치를 요구하기 때문에, 입력 문장이 "还是去你的吧"로 확장될 경우, 위의 문장패턴은 패턴일치에서 실패하여, 여전히 규칙기반 방식의 구조변환을 시도하게 된다. 앞에서 이미 언급된 이유로, 이는 결국 번역 오류로 이어지고, 이런 문제를 해결하는 방법은 " 去+你+的"라는 구가 들어가는 모든 문장패턴을 구축할 수밖에 없다.

즉, 문장패턴은 문장패턴이 가지고 있는 적용률 저하의 결함을 가지고 있는데, 이의 보완책으로서의 규칙기반의 변환은 정확도가 낮은 결함이 있어 번역 오류의 원인으로 되고 있다. 때문에 구구조 변환에서의 새로운 방법이 요구되고 있다.

---

[2] NP: 명사구, r: 대명사, u: 조사

[3] VP: 동사구, v: 동사

또한 구문 실패문장의 경우, 규칙기반의 변환이 불가능하기에 직접 번역 실패로 이어지는 점을 감안하면, 구문 실패하는 경우에 대한 보완책이 필요하다.

3. 뉴스타이틀 번역의 특성

중-한 기계번역에서의 어려움을 해소하고, 불필요한 작업량을 감소하기 위하여 중국어와 한국어 사이의 유사성도 고려하여야 한다. 다음은 중한 양국어에서 구조적 유사성을 가지는 구구조의 예이다. 이런 구구조에서 중국어와 한국어는 같은 어순을 나타낸다[4]:

[중]NP←a+n: 伟大(위대)+祖国(조국)→위대한 조국

[중]NP←MP+NP: [2+个(개)]+国家(나라)→두 (개) 나라

[중]MP←m+q: 2000+个(개)→2000개

[중]AP←d+a: 更(더)+好(좋다)→더 좋다

[중]VP←d+v: 终于(끝내)+实现(실현)→끝내 실현하다

이상의 구조적 유사성에는 일부 예외가 존재한다, 예하면 "AP←adv+adj", "VP←adv+verb"의 경우 "□", "□□"와 같은 일부 특정된 부정형 부사는 어순이 바뀌게 된다.

다음 예문들은 중국어와 한국어 사이의 구구조 유사성을 반영하고 있다:

(예1) 俄(러) : 不(아니) 会(..수 있다) 重演(재연) 切尔诺贝利(체르노베리) 事故(사고). → 러: 체르노베리 사고 재발하지 않을 것.

(예 2) 1600(1600) 万(만) 非洲(아프리카) 人(인) 6(6) 月份(월) 将(장래, 곧) 断粮(식량 끊김). → 1600 만 아프리카인 6 월 식량 기근 예상.

(예3) 保密(보안) 工作(공작,업무) 漏洞百出(빈틈 많음) → 보안 업무 빈틈 속출

위의 예문에서 밑 줄 친 부분은 문장의 술어 역할을 하는 동사구인데, 이런 동사구 내에서의 어순 변화를 생각하지않고, 단지 전체 문장에서의 어순 변화만 고려할 때, 예1에서 동사구와 목적어의 순서가 바뀌고, 예2에서는 술어구가 목적어를 가지고 있지 않기에 전체 문장 어순이 바뀌지 않았으며, 예문3에는 동사구가 없고 성구 " 漏洞百出" 가 술어역할 하는데 문장 어순은 역시 바뀌지 않고 있다.

이상의 예들은, 중국어와 한국어의 번역에서 어순의 변화는 주로 술어와 목적어 사이에서 일어나는 것을 볼 수 있다.

중-한 기계번역에서의 또 다른 하나의 어려움은 중국어 구 내부 의존 관계가 밝혀지지 않는 경우에서의 한국어 기능어 생성 어려움인데, 뉴스타이틀 번역에서는 기능어를 생략 가능한 경우가 많다. 다음은 그 예이다:

(1) 广州(광주)/n 军区(군구)/n 师(사)/n 级(급)/n 指挥员(지휘관)/n 平均(평균)/a 年龄(연령)/n 43/m 岁(세)/q

→광주 군구 사급 지휘관 평균 연령 43세

(2) 46(46)/m 人(인)/n 要求(요구하다)/v 2600(2600)/m 万(만)/m 元(원)/q 赔偿(배상)/v →46인 2600만원 배상 요구

위 예문에서 예문(1)은 어순 변화 없고, (2)에는 어순 변화 존재하는데 두 문장의 한국어 번역문 모두 기능어 생략이 가능한 것을 볼 수 있다.

중한 뉴스타이틀 번역에는 또 다음과 같은 특징이 있다:

(1) 조사 사용이 일반 문장보다 적다. 이는 구문분석 재현률 (recall) 및 정확도 저하의 원인으로 되는데, 일반 문장 구문 분석 재현률[5]이 69.6%인 반면 뉴스 타이틀의 구문분석 재현률은 29.6%밖에 안되고, 구문분석된 문장에서의 정확도[6]는 일반 문장이 61.8%인 반면 뉴스타이틀은 36.2%이다. 전체 문장에서의 구문분석 정확도[7]는 더욱 낮아 10.7% 밖에 안되며 이는 일반 문장의 43.3%보다 훨씬 낮은 것이다.

(2) 문장구조가 상대적으로 단순하며 기능어 생략 가능. 규칙에 의한 변환에서, 뉴스 타이틀의 번역 정확도[8]는 29.8%로서 일반문장의 0.54%보다 훨씬 더 높으며, 더욱이 구문분석 재현률 및 정확도가 일반문장보다 낮은 것

---

[4] 아래에서 사용된 기호들의 의미는 다음과 같다:
a:형용사, d:부사, m:수사, n:명사, q:단위성 의존명사,
AP: 형용사구, MP: 수량사구

[5] 재현률 = 구문분석된문장/전체문장

[6] 구문분석된 문장에서의 정확도 = 구문분석정확문장/구문분석된 문장

[7] 전체 문장에서의 구문분석정확도=구문분석정확문장/전체문장

[8] 번역 정확도 = 번역정확문장/번역된 문장
번역정확문장: 추측으로 의미 파악 가능한 문장

을 감안하면 이런 번역 정확도는 주목할 만하다. 이는 우선 뉴스타이틀은 문장 길이가 짧고, 문장구조가 상대적으로 단순하기 때문이며, 또한 이런 문장구조의 단순성 때문에 생성단계에서 기능어가 생략된다고 하더라도 이해 가능한 문장이 많기 때문이다.

(3) 뉴스 타이틀에서 동사 없거나 동사 1개인 문장은 38.5%이고 동사 2개인 문장은 36.1%로 전체 뉴스 타이틀의 74.6%를 차지한다. 또한 2개 동사를 가진 문장의 50.0%는 휴리스틱을 이용하여 2개 동사를 하나의 동사구로 묶을 수 있다. 즉 휴리스틱 사용시 1개 이하의 동사만 포함한 문장은 전체 뉴스타이틀의 50.6%를 차지하게 된다.

(4) 뉴스타이틀 문장의 구조적 단순성의 또 하나의 특징으로, 1개 동사만 포함한 문장의 85.2%는 번역시의 단순한 동사구 후치로 구조 변환을 완성할 수 있다. 즉 동사구 이외의 부분은 원 어순을 유지하면 된다. 이는 양국어 구조적 유사성에 의한 것이기도 하다. 다음은 이런 문장의 예이다:

[湖北(호북)/n 少年(소년)/n] [滑(빠짐)/v 入(들어감)/v] [桃色/n 电话/n(색정 전화) 陷阱(함정)/n] →호북 소년 색정 전화 함정 빠짐

위에서 "入(들어감)/v"은 조동사로 휴리스틱으로 앞의 동사 "滑(빠짐)/v"와 묶어 하나의 동사구로 처리될 수 있다. 중국어에서의 조동사는 한국어로의 변환에서 일반적으로 생략된다.

(5) 뉴스 타이틀에는 중국인/외국인 인명, 지명 등 고유명사가 존재하기에 단어 분할 및 품사 부착 오류가 심각하며, 단어 사용에서 약어를 많이 사용하기에 어소(g)로 품사 태깅되는 단어가 많다. 단어 분할 및 품사 부착 오류는 구문분석 재현률 및 정확도가 낮은 원인으로 되고 있다. 예 하면, 고유명사를 포함한 뉴스 타이틀은 전체 뉴스 타이틀의 79.3%를 차지하고, 이런 고유명사를 포함한 뉴스 타이틀의 37.5%는 품사 태깅 오류를 포함하고 있다.

4. 뉴스타이틀을 위한 중한 기계번역

3장에서 서술한 뉴스타이틀 번역에서의 특징에 비추어, 시스템을 다음과 같이 개선하기로 한다.

(1) 고유명사 인식

구문구조 부분분석 전에 중국어 및 외국어 인명/지명 인식을 위한 고유명사 인식기를 추가한다.

(2) 구문 분석 실패를 대비한 구문구조 부분분석

구분분석 실패 문장에 대하여 문장의 어순에 영향을 주는 접속사 (连词) 패턴 인식, 전치사 (介词) 패턴 인식 및 중심술어 인식을 진행하여 구문구조에 대한 부분분석(chunking) 을 진행한다.

즉, 이를 위해서는 동사구 인식기와 중심술어 인식기가 필요하다.

(3) 구패턴을 이용한 기능어 생성

중국어 단어 어법정보를 이용하여 동사 위주의 패턴을 구축함으로 패턴의 대량 증가를 방지한다.

번역의 전반 과정을 살펴보면 다음과 같다.

(1) 단어 분할 (segmentation) 및 품사 부착

(2) 고유명사 인식

(3) 구문분석 진행, 성공시 (5)으로, 실패시 (4)로

(4) 기본 문형 및 중심술어 인식을 통한 구문구조 부분인식, 인식된 블록을 (3)으로

(5) 구문분석된 구: 문장패턴 및 변환규칙 이용한 구조변환

부분분석된 구: 구패턴을 이용한 구조 변환.

(6) 대역어 선택 및 한국어 생성

다음 각 절에서는 시스템 전반에 대한 설명(장민 등, 1999)보다는 고유명사 인식기와 동사구/중심술어 인식기을 위주로 설명을 진행하고자 한다.

4.1 고유명사 인식

중국어 및 외국어 인명/지명은 무한 집합이고 절대부분 사전 미등록어이기 때문에 형태소 분석에 어려움을 더해주며 구문분석 실패의 원인으로 되기도 한다. 이런 어려움은 앞의 3장에서도 언급하였듯이, 뉴스타이틀 번역에서 더욱 뚜렷이 나타나기도 한다. 중국어 인명 및 지명 인식에 대한 연구는 근년에 진행되고 있으나 (Wang, et.al, 1999; Tan, et.al, 1999) 중국어로 음차 표기 (音

译)된 외국어 인명/지명에 대한 인식 연구는 거의 이루어 지지 않은 상태이다.

기존 방법에서 중국어 인명/지명 인식은 주로 대용량 지명 사전과 통계정보를 이용한 규칙기반 방법을 사용하였으나 (Wang, et.al, 1999; Tan, et.al, 1999), 통계정보가 결핍한 상황에서 우리는 주로 규칙 기반 방법을 사용하기로 하였다. 다음은 인명 인식에서 후보를 선정하는 주요 휴리스틱이다:

(1) 특정 동사 앞에 나오는 단어들로, 이런 단어들은 모두 1개의 문자로 구성되면, 후보로 선정한다. 이러한 동사에는 "访问(방문하다), 会见(회견하다), 会晤(회견하다), 宣称(선포하다), 称(말하기를), 宣布(선포하다), 当选(당선되다), 表态(표시하다), 揭露(적발하다), 口求(요구하다)" 등이 있으며 중국어 유의어사전을 이용하여 이런 동사 셋을 쉽게 구축할 수 있었다.

(2) 특정 동사 뒤에 나오는 단어들로, 이런 단어들은 모두 1개의 문자로 구성되면, 후보로 선정한다. 이러한 동사에는 "会见(회견하다), 会晤(회견하다), 要求(요구하다)" 등이 있으며 (1)에서 사용되는 동사 집합의 부분집합으로, 사람에 의하여 구축한다.

(3) 문장 부호 ":" 앞에 나오는 단어들로, 이런 단어들은 모두 1개의 문자로 구성되면, 후보로 선정한다.

(4) 특정 명사의 앞, 혹은 뒤에 나오는 단어들로, 이런 단어들은 모두 1개의 문자로 구성되면, 후보로 선정한다. 이런 명사에는 "主席(주석), 女王(여왕), 国王(국왕), 大王(대왕), 总统(대통령), 团长(단장), 大使(대사)" 등이 있으며, 중국어 유의어 사전을 이용하여 자동구축 후 사람에 의하여 선별하도록 한다.

(5) 연이어 2개 이상의 단어 길이가 모두 1이며, 이런 단어들 중 적어도 하나 이상의 단어 품사는 어소(g)로 태깅되어 있으면, 후보로 선정한다.

이런 후보 선정 과정에서 1개 이상의 조건에 부합되면 가산점을 부여한다. 선정된 후보를 최종 인명으로 판단하는 휴리스틱은 다음과 같다:

(1) 2개 이상의 후보 선정조건에 부합되는 후보.

(2) 후보의 첫 번째 문자가 중국인 성씨 사전에 포함되어 있으며, 성씨 외의 문자는 1, 혹은 2개이다. (예: "诸葛 亮", "李 鹏", "江 泽民")

(3) 후보를 구성하는 문자 중에는 외국인 인명에서 사용되는 문자가 2개 이상 포함되어 있다 (예: "比埃尔霍夫"). 이런 외국인 인명 상용 문자 사전 영중 사전을 이용하여 구축한 다음 말뭉치를 이용하여 자료 보완하였다 (예: "夫", "森", "希").

우리는 이상의 비교적 단순한 휴리스틱을 이용한 고유명사 인식기에 의한 중국인 인명/지명 인식 정확도는 49%이고, 고유명사기를 추가한 전체 단어분할 품사태거에서의 고유명사 인식 정확도는 72%로 나타났다. 특히 이중 중국인 인명 지명 인식에 대한 정확도는 69%로 나와 10만 등록어 중국 지명사전을 사용하고 180만자 중국 인명/지명이 태깅되어 있는 말뭉치를 이용한 통계적 방법에서 86.7%의 정확도를 얻은 결과 (Tan, et.al, 1999)와 비교할 때 비교적 고무적인 결과라고 할 수 있다. 비교적 간단한 방법으로 이런 결과를 얻을 수 있는 것은 뉴스타이틀 번역이라는 특정 도메인에서의 비교적 단순한 언어현상 때문이기도 하다.

4.2 접속사 패턴, 전치사 패턴 및 중심술어 인식을 통한 구문 부분분석

뉴스 타이틀의 구문분석이 성공률이 낮은 점에 비추어, 구문 분석 실패 시를 대비한 접속사패턴, 전치사 패턴 및 동사구/중심술어 인식을 진행한다.

단계1: 우선 기본 문형 인식을 진행한다. 이런 기본문형인식에서는 "虽然… 但是… (비록… 지만,…)", "不但… 而且… (… 뿐만 아니라, … 도)"등과 같은 접속사패턴이 이용되며, 접속사 패턴이 적용된 문장은 이에 의해 몇 개의 블록으로 나누어 진다. 예:

"虽然 希望 再度 入选, 但 被 认为 可能性 不 大" (비록 재차 당선되기를 희망하지만, 가능성이 희박한 것으로 여겨지고 있다) → "[虽然 [希望 再度 入选]], [但 [被 认为 可能性 不 大]]"

이런 접속사패턴은 중국어 허사사전에서 두 개 이상 단어로 구성된 허사 연어를 사람에 의하여 선별함으로 구축되었다.

단계2: 전치사 패턴 및 문장부호를 이용한 구문 부분분석을 진행한다. 이런 전치사 패턴에는 "把… V(… 을

V+하다)" , "从… V(… 로 부터 V+하다)", 등이 있으며, 문장부호에는 "" "", "[ ]", " ‘’ ", "( )"등이 있다. 주로 전치사 (介词)구 및 문장부호사이에 있는 명사구에 대한 인식 과정이다 (Zhou, 1999).

단계3: 중심 술어인식을 진행하는데 우선 동사구 인식을 진행하고 다음 이런 동사구 중에서 중심술어를 선택한다.

동사구 인식을 위한 규칙은 다음과 같다[9]:

(1) ‘ 부사+동사’ ,’ 전치사+동사’ , ‘ 형용사+동사’ , ‘ 동사+조동사’ , ‘ 조동사+동사’ 등 의 동사 구패턴에 부합되면 하나의 동사구로 묶는다.

(2) ‘ 동사+특정명사’ 시 해당 동사는 술어로 되지 않기에 동사구 인식에서 제외함.

예: 动物园(枪击/v)案/n→동물원 (총기 난발) 사건

(3) 특정 동사(报告-보고하다/보고서, 报-보고하다/신문)은 형태소 분석에서 애매성 때문에 동사로 잘못 태깅 되는데 다른 동사와의 공기 상황에 따라 명사로 인정함. 예:

美报告(称)“ 导弹(防御)系统” (将掀)军备(竞赛). → 미 보고서는 “ 유도탄 (방어) 시스템” 은 군비(경색) (일으킬 것)이라고 (언급)

동사구를 인식 한 다음 이런 동사구 중에서 하나의 중심 술어를 선택하게 된다. 이를 위해 우선 중국어 어법정보사전 (Yu, 1998)에서 제공하는 동사의 특징(feature) 중에서 어순에 영향을 주는, 즉 목적어와 관계되는 특징을 살펴보면 다음과 같다:

- 문장목적어 가질 수 있는 동사인가
- 겸어(兼语) 동사인가
- 두개 목적어 가질 수 있는 동사인가
- 목적어 가질 수 있는 동사인가

이외에도 4자 성구도 중심술어의 판단에 영향 주기에 이도 중심술어판단에서 하나의 참고 요소로 삼는다.

중심술어 인식은 다음의 순서로 동사구에 우선 순위를 주어 인식하기로 한다.

(1) 문장목적어(小句宾)/두개 목적어(双宾)/겸어 목적어(兼语)를 가질 수 있는 동사 중 인간이 그 주어로 될 수 있는 동사. 예:

"表示(표시하다) ","希望(희망하다) ", "要求(요구하다) ","警告(경고하다) "…

(2) 성구가 문장의 마지막 단어로 될 때

(3) 자동사가 문장의 마지막 단어로 될 때

(4)“ 被/p+v” ,“ 遭/v+v” 가 문장의 마지막 구로 될 때

(5)“ 부사+v” , “ 형용사+v” 의 출현

(6)“ 타동사1+x+타동사2” 패턴에서 타동사1이 주술어

(7)“ 타동사1+x+타동사2+y” 패턴이면 동등 위치.

이런 휴리스틱을 이용한 중심술어 인식의 예는 다음 같다:

美国警方(已逮捕)动物园枪击案(嫌犯). (main: 逮捕)
미 경찰 동물원 총기 난발사건 (혐의범) (이미 체포)

(给)华最惠国待遇(是)义务(不是)礼物. (main: 是, 不是)
중국에 최혜국대우 (주)는 것은 의무(이지) 선물(이 아니다)

주어진 구내의 중심 술어 이외 부분은 하나의 구로 인식하고 중심술어의 동사 유형에 따른 변환규칙에 의해 구조 변환이 이루어 지며, 동사구 내부의 구조변환은 동사구패턴을 이용하여 진행한다.

이런 간단한 휴리스틱을 이용한 중심술어 인식 및 한정된 접속사패턴, 전치사패턴, 술어패턴을 이용한 구구조 변환이 가능한 것은 앞의 3장에서 언급되었던, 문장구조가 상대적으로 단순하고 기능어 생략이 가능하다는 뉴스타이틀 번역의 특성 때문이다.

실제 실험에서 총 118개의 뉴스타이틀 중 구문분석 실패 및 문장패턴 매칭 실패로 구구조 변환 실패한 문장 한 총 90문장에 대하여 동사구 인식 및 중심술어 인식을 진행한 결과 68개 문장이 정확히 인식되어 75.6%의 정확도를 얻을 수 있었다.

3장에서 기술한 뉴스타이틀 번역의 특성을 이용하여 비교적 높은 정확도의 중심술어인식을 달성했지만 실제 중국어 일반 문장에서의 중심술어 인식은 매우 복잡한 문제로 통계기반 방법 및 더 많은 지식의 도움이 있어야 한다(穗志方& 俞士汶, 1998).

[9] 괄호 안 부분은 동사구 인식기에 의해 인식되어진 동사구임.

5. 기타 연구

위에서 언급한 고유명사 인식기 및 중심술어 인식을 통한 문장 구조 부분분석 외에 우리는 통계기반의 중국어 단어 분할 및 품사부착 모듈에 규칙을 도입하였고 중국어 구문분석의 정확도를 향상하기 위한 노력으로 구문분석규칙 제약조건 강화하는 작업을 진행하였다.

(1) 단어분할 및 품사부착에서의 규칙 사용

중국어 단어분할 및 품사부착은 주로 통계기반 방법에 규칙을 사용하는 방법을 사용하는데 어떤 시스템은 선 규칙 후 통계이고 어떤 시스템은 선 통계 후 규칙 등 서로 다른 방법을 사용하고 있다.

MATES/CK에서는 선 통계 후 규칙의 방법을 사용하여 우선 FB (Forward-Backward) 알고리즘에 의한 단어분할 및 품사부착 (장민 등, 1999) 후 여전히 애매성을 가지고 있는 어휘에 대해서는 규칙을 도입하여 해결하기로 하였다 (강원석 등, 2000; 김지현, 2000). 중국어 품사부착에서 우리는 북경대 품사부착 지침에 따랐다 (송희정, 2000).

(2) MATES/CK의 구문분석기에서의 제약규칙 강화

구문분석기에서 사용하는 CFG (Context Free Grammar) 규칙의 제약 조건이 완전하지 못한 부분에 대하여 수정 보완하는 작업을 진행하였는데 실제로 닫힌 실험 (close test)에서 수정 보완된 규칙을 사용한 후 구문분석 정확도가 현저히 향상되는 것을 볼 수 있었다 (송영미 등, 2000).

(3) 예제기반 변환방식에 대한 실험

구문분석에서 실패한 문장패턴의 변환을 위하여 처음에 예제기반 방식을 도입하려고 계획하였고 이 때문에 예제기반 변환방식에서 최적 예제 매칭에 대한 모듈을 구축하고 실험하였는데 좋은 결과를 얻었다. 그러나 이런 방법을 사용함으로 기존의 문장패턴의 적용률을 높이는 데는 실패하였는데, 이는 예문 (패턴)의 적용률 향상을 위해서는 결국 예문이 구구조를 가져야 하고, 또한 구문단위로 정렬되어야 하는 제약이 있기 때문이다 (Sata & Nagao, 1990; Kaji et al.1992). 결국 이런 방법보다는 구문분석 실패에 대한 보완책으로 부분분석을 도입하고, 예제가 아닌 구패턴을 이용한 구구조 변환 방식을 채택하기로 하였다 (4장 참조).

6. 결론 및 향후 연구

본 논문은 이미 구축된 중한기계번역 원형 시스템의 기초 위에서 중국어와 한국어의 언어적 차이성뿐만 아니라 유사성을 살펴보고, 뉴스 타이틀이라는 특정 도메인에 한정하여 중국어 분석 및 변환 정확도 향상을 위하여 진행되어진 지난 한 동안의 작업에 대한 기록이다.

중국어 및 음차표기된 외국 인명/지명 인식과, 접속사/전치사 패턴 인식 및 중심술어 인식을 이용한 구문분석 실패문장에 대한 부분분석에 초점 맞추어 연구를 진행하였으며 단어분할 및 품사부착에서의 규칙 도입, 보다 나은 효과의 변환방식에 대한 탐색과 시도도 있었다.

이런 연구는 뉴스타이틀이라는 특정 도메인에서 비교적 효과적인 것으로 나타났으며 향후 보다 넓은 도메인에서의 적용을 위하여는 통계 기반 방법의 도입에 대한 연구가 진행되어져야 한다.

구문분석 및 변환의 성능 향상을 위해서는 구패턴을 이용한 구문분석, 변환 등에 대한 연구도 진행하여야 하며, 형태소분석 및 구문분석 제약규칙 구축도 계속 진행되어져야 한다. 한국어 생성을 위한 필요한 중국어 정보의 제공에 대한 연구도 진행되어야 한다.

중한기계번역에 대한 평가 기준 및 평가 문장 집합 구축에 대한 연구가 이루어져야 하며, 중한 기계번역을 위한 보다 전면적인 언어학적 비교 연구가 시급하다.

중한 기계번역에 대한 연구는 한 일보다도 해야 할 일이 많은 작업으로, 티끌로 태산 모으는 인내와 노력이 필요하다.

7. 참고문헌

[1]. 강원석, 김지현, 송영미, 송희정, 황금하, 채영숙, 최기선, 2000, " 중한 기계 번역 시스템을 위한 형태소 분석기" , 제12회 한글 및 한국어 정보처리 학술대회

[2]. 김지현, 2000, " 중한기계번역 시스템 향상을 위

한 형태소 및 품사 태깅 분석", KAIST KORTERM & NLP Lab. 2000년도 하계 워크샵 논문지, 2000.7
(http://kibs.kaist.ac.kr/~hgh/research.html)
[3]. 송영미, 2000, "중한기계번역시스템의 구문분석", KAIST KORTERM & NLP Lab. 2000년도 하계 워크샵 논문지, 2000.7
(http://kibs.kaist.ac.kr/~hgh/research.html)
[4]. 송희정, 김지현, 송영미, 2000, "중국어 코퍼스 가공 지침", KAIST KORTERM & NLP Lab. 2000년도 하계 워크샵 논문지, 2000.7
(http://kibs.kaist.ac.kr/~hgh/research.html)
[5]. 장민,황금하,서충원,최기선. 중한 기계번역기 MATES/CK: 파이프라인 번역. 제11회 한글 및 한국어 정보처리 학술대회. pp. 121-127, 1999.10
[6]. 황금하, 2000, "정렬을 통한 다단계 변환패턴의 자동구축",석사학위논문
(http://kibs.kaist.ac.kr/~hgh/research.html)
[7]. 梅家驹,竺一鸣,高蕴琦,殷鸿翔, 1985, "同义词词林(유의어사전)", 중국상해사서출판사발행소 (in Chinese)
[8].穗志方, 俞士汶, 1998, "汉语单句谓语中心词识别知识的获取及应用(중국어 문장 술어 중심사 식별 지식의 획득 및 응용)", 북경대학 전산학과 & 북경대학 계산언어학연구소 俞士汶등 편찬 계산언어학문집 pp.166-175 (in Chinese)
[9]. Kaji, Hiroyuki, Yuuko Kida & Yasutsugu Morimoto, 1992, "Learning translation templates from bilingual text", in proceeding of COLING92, pp.645-651
[10]. Sato, Satoshi & Makoto Nagao, 1990, "Toward memory-based translation", in Proceedings of COLING 90, pp.247-252
[11] Tan Honggye, Zheng Jiaheng, Liu Kaiying, 1999, "Research of the Method of Automatic Recognition of Chinese Place", in proceeding of JSCL-99 (중국 제5기 계산언어학 연합학술회의), pp.174-179 (in Chinese)
[12] Wang Xing, Huang Degen, Yang Yuansheng, 1999, "Identifying Chinese Names based on Combination of Statistics and Rules", in proceeding of JSCL-99 (중국 제5기 계산언어학 연합학술회의), pp.155-161 (in Chinese)
[13]. Yu, Shiwen, Xuefeng Zhu, Hui Wang & Yungyung Zhang: 1998, "the Grammatical Knowledge-base of Contemporary Chinese --- A Complete Specification", Tsinghua University Press (in Chinese)
[14]. Zhang Min, Key-Sun Choi, 1999, "Pipelined Multi-Engine Machine Translation: Accomplishment of MATES/CK System", in Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation.
[15]. Zhou Ming, 1999, "J-Beijing Chinese-Japanese Machine Translation System", Compute Linguistic Paper Collocation, TsingHua University Press (in Chinese)