제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

북한 영한 기계번역 시스템의 문제점 및 개선 방향

이병희*, 전성진*, 서정현*, 류범종*
*한국과학기술정보연구원 정보시스템연구실
e-mail : {bhlee, sjjhun, jerry, ybj}@kisti.re.kr

The Problems and Suggestions in a North Korean E-K Machine Translation System

Byeong-Hee Lee*, Sung-Jin Jhun*, Jeong-Hyeon Seo*, Beom-Jong You*
*Dept. of Information Systems, KISTI

요 약

KISTI 는 최신 해외 과학기술 정보를 우리 글로 쉽게 접근할 수 있도록 온라인 상에서 영문 제목을 한글로 번역하여 서비스하는 시스템을 중국 단동에 소재한 하나프로그램 센터와 남북한 과학기술 정보교류 차원에서 공동으로 개발해 왔다. 본 연구에서는 북한의 하나프로그램 센터에서 개발한 영한 번역 시스템을 가지고 각종 문서의 제목을 번역하였을 때 발생하는 번역 오류를 분석하고, 알타비스타 번역 서비스를 이용하였을 때의 번역 결과와 비교하고 분석하였다. 본 영한번역 시스템이 실용적으로 쓰이기 위해서는 문법 구조에 관한 연구에 많은 노력이 필요하다.

1. 서론

정보화 시대를 사는 많은 사람들이 일상 생활에서 자연스럽게 사용하고 있는 자연 언어를 컴퓨터를 이용해서 번역하려는 노력이 1950 년대부터 여러 국가에서 시도되어 현재도 진행 중에 있다[1].

한국과학기술정보연구원(KISTI)은 해외 각국의 최신 과학기술 정보를 수집, 분석, 종합한 자료를 DB 화하여 국내외의 많은 정보 이용자들에게 서비스하고 있다. 영문 자료가 대부분을 차지하는 해외 과학기술 자료를 정보 이용자들이 최신 해외 과학기술 정보를 우리 글로 쉽게 접근할 수 있도록 KISTI 는 온라인 상에서 영문 제목을 한글로 번역하여 서비스하는 시스템을 중국 단동에 소재한 하나프로그램 센터와 남북한 과학기술 정보교류 차원에서 개발해 왔다.

현재 국내에서 영어를 한국어로 번역하는 영한 번역 시스템에는 앙코르, 트래니, 번역마당, 클럭큐, 이지맨, 인가이드, 스마트랜, 알타비스타 번역 서비스 등이 있다. 또한 기존의 제품보다 나은 고품질의 번역 시스템을 개발하기 위해 번역 분야를 제한하든가, 입력 방법을 쉽게 하기 위해 음성 인식 및 문자 인식 기술을 접목하는 제품들이 개발되어 시판되는 추세이다.

본 연구와 유사한 문서들이 여러 번역 소프트웨어 회사나 연구 기관 내에서 내부 자료로 사용되고는 있으나 공식적으로 발표하기를 꺼리고 있는 실정이며, 각각의 번역 시스템들이 시스템마다 고유한 특성(system dependent)들을 가지고 있어 여러 번역 시스템들을 객관적으로 분석하기가 쉽지 않다.

본 연구에서는 북한의 하나프로그램 센터에서 개발한 영한 번역 시스템을 가지고 각종 문서의 제목을 번역하였을 때 발생하는 번역 오류를 분석하고, 여러 번역 시스템 중에서 무료로 서비스되고 있는 알타비스타 번역 서비스(http://babel.altavista.com/tr?)를 이용하였을 때의 번역 결과와 비교하고 분석한다.

2. 남북한 언어 사용의 차이

과학기술 문헌의 영한 번역에 있어서 남북한의 사용자에게 가장 큰 부담을 주는 부분이 남북한 언어 사용의 차이이다.

최현규[2]는 남북한의 언어 차이를 맞춤법의 차이

(자모의 명칭과 순서, 두음법칙, '폐'는 '페'로 등), 어미 류의 차이('어'와 '여', 사이시옷에 관련된 표기 등) 띄어 쓰기의 차이(하나로 묶어지는 덩이, 부속적 단어는 붙여쓰기로 등), 어휘의 차이(방언에서 유래되는 차이, 사회 제도에서 유래되는 차이, 외래어와 관련된 차이 등), 의미가 상이한 단어, 이음 동의어와 이철 동의어, 북한에만 존재하는 단어로 나누어 설명하였다.

남북한의 언어가 달라졌다 하더라도 아직은 공통적인 면이 많으며, 약간의 차이점을 제외하면 비슷한 모습을 보이고 있다. 이는 남북한 맞춤법이 모두 1933년 제정된 한글 맞춤법 통일안에 기반을 두고 있기 때문이다.

이러한 언어 사용의 차이를 위해서는 남북한의 대표자들이 만나 표준화된 어휘와 언어를 만들어야 할 것으로 보인다.

3. 북한 영한 번역 시스템

과학기술 문헌 온라인 영한 기계번역 시스템을 구축하기 위해서는 영한 어휘 사전과 과학기술 용어사전의 설계 및 구축 영한 자동번역 엔진의 연구개발이 필요하다[3,4].

이러한 목적을 위해 일반 단어 사전과 KISTI 에서 수집, 정리한 과학기술 용어 30 만개를 과학기술 용어 사전에 수록하였고, 어휘/결합 수준 사전, 관계 부사와 접속 부사의 처리, 문법 수준 사전을 구축하여 Trie 검색 방법을 이용하여 검색하는 방식을 취하였다.

또한 명사, 형용사, 동사, 부사 사전 관리 도구를 개발하였고, 과학기술 및 일반적인 영어에 적용할 수 있도록 구문 유형 및 특성을 반영하는 구문 해석 엔진을 개발하고, 클라이언트/서버 형으로 온라인 번역을 할 수 있도록 하였다.

그리하여 과학기술 논문 제목 번역에 대하여는 90%이상의 번역율을 보장하도록 하며, 일반 문장에 대해서는 50%이상의 번역율을 보장하도록 하였다.

4. 오류 유형 및 개선 방향

본 연구에서는 오류의 종류를 남북한의 단어 사용상의 차이에서 오는 오류, 어휘 번역 오류, 문법 구조 오류, 띄어 쓰기/기호/구두점 및 기타 오류로 나누어 비교 분석 한다.

(가). 남북한의 단어 사용상의 차이에서 오는 오류

	단동 하나	알타비스타
Italian	이딸리아	이탈리아
Energy	에네르기	정력
mechanics	력학	기계공

(나). 어휘 번역 오류

● 어휘 미등록 오류

	단동 하나	알타비스타
AI	AI	AI
Data Mining	자료는 mine 합니다	자료 광업

● 전문 용어 번역 오류

	단동 하나	비고
Technological Paper	기술적인 종이	기술적 논문

● 고유 명사 번역 오류

	단동 하나	비고
SENT-specimen	쎈뜨-표본	대문자로 구성된 약어
MAX.3 antigen	최대값.3 항원	대문자로 구성된 고유명사

● 어휘 중의성에서 오는 오류

	단동 하나	비고
Linear elastic	선형 고무줄	선형 고무줄
population	주민들	사람을 지칭하는 주민이 아닌 밀도, 분포와 같은 전문적인 표현으로 해석 필요

● 관용적 표현 오류

	단동 하나	비고
Getting personal	개인용을 구하기	숙어 및 관용어구에 대한 인식률이 떨어짐 : 단어 단위로 해석하는 문제 발생

(다). 문법 구조 오류

영한 기계 번역 시스템의 성능을 평가한 것에는 시스템공학연구소 보고서[5]와 김향숙[6]의 논문이 있다. 논 연구에서는 김향숙의 문법 구조 오류 중에서 제목에 자주 나오는 문법 구조만을 가지고 검사해 보았다.

● 동명사

	단동 하나	알타비스타
I object to **his receiving** an invitation.	초청 나는 그의 수신기에 반대합니다.	나는 안내장을 받으면 그에게 반대한다.
It was tough answering questions.	그것은 곤난한 대답처리 질문이었습니다.	험한 응답 질문 이었다.

● 부정사

	단동 하나	알타비스타
How do you expect me **to respond** to this?	어떻게 당신이 이것에 응하기 위하여 나를 대기하는가 하는 것?	어떻게 너는 나에게를 이것에 반응하는가 예기하는가?
Pete wanted **to be** the greatest hitter of all time.	Pete 은 모든 시간의 가장 훌륭한 타격수로 되는 것을 바랐습니다.	Pete 모든 시간의 가장 중대한 hitter 이고 싶 ss 다.

● 분사

	단동 하나	알타비스타
The only car **repaired** by that mechanic is mine.	그 기계공에 의하여 고쳐 진 유일한 승용차는 나의 것입니다.	저 기계공이 고치는 유일한 차는 나의 것이다.
The man **writing** the obituaries is my friend.	사망기사를 저술하는 남자는 나의 친우입니다.	obituaries 쓰면 남자는 나의 친구 이다.

● 동사

	단동 하나	알타비스타
It won't **matter**.	그것은 관계될것이지 않습니다.	그것은 중요하지 않을 것이다.
This orange **peels** well.	이것 귤은 잘 벗겨 집니다.	이 오렌지는 잘 거피한다.

● 조동사

	단동 하나	알타비스타
Can I say something?	내가 어떤 물건을 말한다는 것은 수 있습니까?	나는 무언가를 말하는가 Yorosi_un_ka?
Will you marry me?	당신이 나와 결혼한다는 것은 것입니까?	너는 나에게를 결혼할 것인가?

● 시제

	단동 하나	알타비스타
I'll be asking you some questions.	나는 일부 질문을 당신이 물어 보다게 될 것입니다.	I?ll 은 너에게 얼마간 질문을 묻고 있다.
I've asked you some questions.	나는 일부 질문을 당신이 물어 보았습니다.	나는 너에게 얼마간 질문을 물었다.

● 문장 종류

	단동 하나	알타비스타
I **can hardly** wait.	나는 거의 기다리지 않을 수 있습니다.	나는 단단하게 기다릴 없다.
Don't give up.	포기하지 말아 주십시오.	위로 주지 말라.

● 접속사

	단동 하나	알타비스타
Jack painted the kitchen white **and** the bathroom blue.	남자는 푸른 부엌의 흰색과 목욕실을.	잭은 부엌 백색이라고및 목욕탕 파랑을 그렸다.
I'd like it to be black **or** gray.	나는 흑인 혹은 회색정복을 입은 사람이기 위하여	I?d 은 까말 I 그것또는 회색이 좋아한다.
	그것을 좋아 할 것입니다.	

● 관계사

	단동 하나	알타비스타
I have two sons, **who** are teachers.	나는 선생들있는 2 아들들을 가지고 있습니다.	나에는 교사 인 두 아들이 있다.
I need something **with which** to write.	나는 어느 것을 저술하여야 하는 가와 함께 어느 정도 필요가 있습니다.	나는 쓴 위하여 무언가를 필요로 한다.

● 수동태

	단동 하나	알타비스타
It **is said** that he was reared abroad.	그가 국외로 뒤발로 섰다는 것이 말합니다.	그가 해외로 길렀다 고 말한다.
He **is said** to have been reared abroad.	그 남자는 국외로 길러 졌기 위하여 말합니다.	그는 해외로 기르라고 말한다.

● 특수 구문

	단동 하나	알타비스타
What **the hell** is going on?	도대체 계속하고 있습니까?	무엇 나락은에 가고 있는가?
It is me **that** the dog scared.	그것은 개가 놀래운 나입니다.	개가 위협했다고 그것은 나에게 이다.

● 전치사

	단동 하나	알타비스타
I'll be there **by** noon.	나는 정오까지 거기에 있을 것입니다.	나는 정오의옆에서 거기서 것이이다.
I go to church **on** Sundays.	일요일에 우에(서) 나는 교회에 갑니다.	나는 일요일에 교회에 간다.

● 대명사

	단동 하나	알타비스타
The pleasure is all **mine**.	기쁨은 나의 것 전체입니다.	쾌락은 모든 광산 이다.
Few focused on improving quality.	거의 품질을 개선하면 집중시키지 않았습니다.	거의 없은 질을 개량하기에 초점을 맞췄다.

● 형용사

	단동 하나	알타비스타
She pours **a ton of** cream and	그녀는 그의 커피로의 크림과	그는 크림의 톤을 퍼붓고 그의

sugar into his coffee.	설탕 따른다. 신청	커피로 설탕을 칩니다.
There are a **bunch of** guys in the classroom.	교실에(서) 당김 바줄의 송이가 있습니다.	교실안에 녀석의 낱단 있다.

● 부사

	단동 하나	알타비스타
Let' go. Come on.	갑시다. 다가 오다.	가자. 위에 오는.
Come on. Jim.	다가 오다. 짐.	위에 오는. Jim.

● 비교 구문

	단동 하나	알타비스타
I'm **as happy as** you are.	당신이 있는껏 나는 행복합니다.	너것과 같이 행복한 I?m 은 만큼 이다.
He is **bigger than** me.	그 남자는 나보다 더 성장합니다.	그는 나에게보다는 더 크다.

(라) 띄어 쓰기, 기호, 구두점 및 기타 오류

● 띄어 쓰기

	단동 하나	알타비스타
I'm as happy as you are.	당신이 있는껏 나는 행복합니다.	너것과 같이 행복한 I?m 은 만큼 이다.
I have two sons, who are teachers.	나는 선생들있는 2 아들들을 가지고 있습니다.	나에는 교사 인 두 아들이 있다.

● 기호/구두점/기타

	단동 하나	알타비스타
How do you expect me to respond to this?	어떻게 당신이 이것에 응하기 위하여 나를 대기하는가 하는 것?	어떻게 너는 나에게를 이것에 반응하는가 예기하는가?
	':' 구분자를 이용한 description 또는 형용구의 관계 및 의미에 대한 인식률이 떨어짐	
	'-' 기호 앞뒤에 오는 단어의 의미 연결 관계 설정 미숙	

5. 결론

단동 하나프로그램 영한 번역기에 있어서 전문용어, 고유 명사 번역, 관용적 표현 번역의 측면에서는 알타비스타에 비해 우수하지만, 남북한의 단어 사용상의 차이에서 오는 오류, 띄어 쓰기 오류 검사에서는 부족한 면이 있다. 또한 문법 구조 오류 측면에서는 아직 많은 정제 작업이 필요하다.

또한 단동 하나의 경우 전체적으로 제목을 해석하는 데 있어서 한글 어법과 영어 어법의 관계를 고려하여 문장 뒤쪽에서 해석을 하도록 하는 방식을 사용하는 문제가 있어 다음과 같은 문제가 발생하고 있다. "A high-rate(20~30%) of parental consanguinity in cytochrome-oxidase deficiency"의 경우 "시토크롬옥시다제 부족에서 어버이다운 동족의 (20~30%) 높은 비율"과 같이 '높은 비율'을 꾸며주는 '(20~30%)'의 위치가 적당하지 못한 위치에 있으며, "Fruit characteristics of 'Yourk' apples during development and after storage"의 경우 의미 및 문맥상 "개발 및 개발 후 저장 시에" 와 같은 방식으로 해석되어야 할 문장이 뒤쪽에서부터 하다 보니 "저장 후에 그리고 개발 동안"의 식으로 해석이 되는 문제가 있다.

이 외에도 자연어 처리에서 근본적인 문제점인 중의성(ambiguity) 제거를 위해서는 상황에 맞는 하나의 단어의 의미를 선택할 수 있는 기능을 갖춘 번역 시스템을 위해 앞으로도 많은 노력이 필요하다. 또한 문서의 제목을 번역하기 위해서는 문법 구조 오류의 종류에서 보인 바와 같이 아직도 해결하여야 할 문제점이 많으며 이러한 문제점 해결을 위해서 향후 문법 구조 연구에 많은 연구와 정제가 필요하다.

마지막으로 남북한의 언어가 많이 달라졌다고 하지만 아직은 이질적인 면보다 동질적인 면이 많다. 한 민족의 언어를 가지고 자유롭게 의사 소통을 하기 위해서는 남북한의 언어 표준화를 위한 구체적인 방안들이 나와야 한다. 이를 위해 남북한의 교류를 확대하고, 서로간의 언어 정책을 수용하여 장점들을 살려 우리의 언어를 발전시키며, 남북한의 통일을 대비하여 표준화된 언어를 만들고, 이를 널리 보급할 수 있는 기구를 설치하여 많은 노력을 하여야 할 것이다.

참고문헌

[1] 김태완, 박철제, "특집 한글공학: 기계번역 시스템," 한국정보처리학회지, 제 5 권 제 5 호, pp.29-36, 1998.

[2] 최현규, 한선화, "남북 과학기술용어의 차이 비교 연구," 한국과학기술정보인프라 워크샵 학술발표 논문집, pp.236-246, 2002.

[3] 단동 하나프로그램센터, "과학기술자료 온라인 영한 자동번역시스템," 한국과학기술정보연구원 최종 보고서, 2002.

[4] 단동 하나프로그램센터, "과학기술자료 온라인 영한 자동번역시스템," 한국과학기술정보연구원 사용 설명서, 2002.

[5] 이민행, 정소우, 지광신, "기계번역 시스템 평가 방안 연구," 시스템공학연구소 최종 보고서, 1998.

[6] 김향숙. "영한 기계 번역 시스템의 번역 오류 유형 분석," 성신여대 영어교육과 석사학위 논문, 1999.