한국어 - 일본어 기계번역 시스템의 수식어 처리와 중문처리

°주인숙, 백모현, 진중화, 임선태, 임인칠
한 양 대 학 교

Modifiers and Compound Sentences Processing of a Korean-Japanese Machine Translation System

°I.S.Joo, M.H.Paik, J.H.Jin, S.T.Lim, I.C.Lim
Hanyang University

Abstrct

This paper proposes a Korean-Japanese Machine Translation System that processes unregistered words, modifiers and compound sentences.

In mophological analysis, the unregistered words are processed by using unregistered word processing algorithm.

The modifiers are processed by consulting noun-attributes and grammar rules.

The compound sentence processing algorithm recognizes whether the sentence that includes commas is compound sentence or not.

This system performs on IBM-PC/AT DOS using Prolog-1.

Ⅰ. 서 론

지식의 습득이나 추론과 같은 인간의 지식행위를 컴퓨터상에서 실현하는 지식 기반 시스템(Knowledge Based System)에 관하여 많은 연구가 행해지고 있으며, 기계번역은 자연어 처리의 응용분야로서 통신망의 발달로 국제교류가 빈번해지고, 각 언어간에 각종 정보의 유통량이 크게 증대됨에 따라 그 필요성이 더욱 부각되고 있다.[2]

자연언어의 기계번역은 형태소 분석, 구문분석, 의미분석, 사전구성등으로 크게 구분할 수 있는데, 기존의 형태소 분석단계에서는 미등록어 처리가 미흡하고 의미해석이 미비하여 사전구성이 단어의 일대일 대응인 경우가 많아 사전이 방대해지므로 사전검색시간이 오래 걸릴 뿐만 아니라 동음어나 어미활용의 처리에 문제점이 있다.

본 논문에서는 이를 해결하기 위해 한국어-일본어 기계번역 시스템의 실현에 있어서, 미등록어는 미등록어 처리 알고리즘을 이용하여 모든 단어를 검색하고 사전에 없는 단어일 경우 사용자가 직접 등록할 수 있도록 한다. 또한 수식어 처리에서는 앞 단어의 품사속성으로 문법규칙을 참조하여 동음어미를 판별할 수 있도록 하는 방법과 콤마를 이용해 중문을 번역하는 방법을 제안한다. 즉, 문장내에서 명사를 수식하는 수식어를 처리하고자 할 때 한국어인 경우는 형용사로 품사가 결정되는데 비해 일본어에서는 품사가 형용동사활용과ㅠ행변격동사활용으로 구분됨으로 동음어라 하더라도 대응어를 정확히 번역하도록 설계한다. 그리고, 중문을 처리할 시에는 콤마가 문장내에서 단어의 나열을 나타내는지 아니면 중문을 나타내는지를 인식하게 한다.

본 논문에서 제안한 알고리즘을 MS-DOS상에서 Prolog-1으로 프로그램하여 알고리즘의 유효성을 확인한다.

Ⅱ. 한국어-일본어 기계번역 시스템의 구성

기계번역 시스템은 그림 1과 같으며 크게 형태소 분석, 구문분석, 사전, 문법규칙, 번역발생등으로 구성되어 있다.

형태소 분석은 한국어의 특성중 하나가 띄어쓰기이므로 입력문에 대해 각 어절 단위로 행하여 단어 리스트를 생성하며, 사전에 등록되어 있지 않은 단어가 입력되면 미등록어 처리를 행한다.

구문분석은 문법규칙을 참조하여 각 단어에 대한 품사, 품사속성 리스트를 생성하여 다의성 조사에 대한 의미결정 및 중문처리를 위한 문형의 생성과 수식어처리를 실행하여 번역발생을 한다.

최종 번역 발생시에는 이들 과정에서 생성된 정보와 문법규칙과 어간사전을 참조하여 중문을 번역하고 이에 대응하는 일본어 출력을 한다.

기계번역 시스템의 구성을 그림 1과 같이 표시한다.

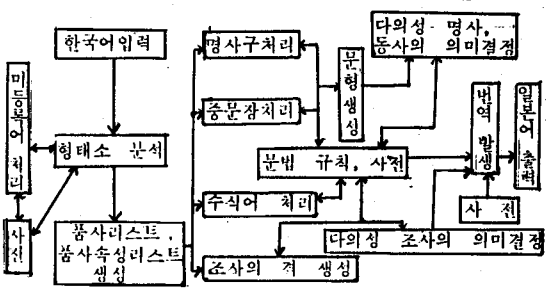
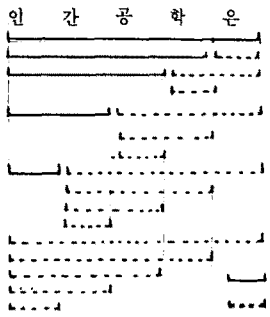


그림 1. 기계번역 시스템 흐름도

Ⅱ-1. 형태소 분석과 미등록어 처리

한국어 문장은 띄어쓰기를 하므로 각 어절 단위로 형태소 분석을 하는 방법에 있어서는 최장일치법을 사용하여 아래 예)와 같은 방식으로 사전에 검색하면서 단어를 찾아 나간다.[1] 한 어절이 분석되면 다음 어절을 계속해서 형태소 분석을 하여 각각의 단어를 argument로 하는 단어리스트를 생성한다.

예) 인간공학은 인간을 활용한 학문입니다.



우선 한 어절이 사전에 존재하는 지를 검색하여 존재하면 다음 어절을 계속하지만 존재하지 않는다면 한 단어씩 줄여가면서 존재여부를 검색하고, 사전에 존재하면 단어는 단어리스트의 argument로 저장되고 그 나머지 단어들은 다시 검색을 반복한다. 그러나 만약에 입력문에서 미등록어가 입력되어 형태소 분석이 어려울 경우는 형태소 분석과 미등록어 처리를 병렬처리해준다. 단어를 검색할 때 사전에 등록하지 않은 경우라면 미등록어 처리 알고리즘을 적용시켜서 한 어절에 대한 매칭될 수 있는 모든 단어들은 화면상에 나타낼 수 있게 하여 직접 사용자가 사전에 등록할 수 있게 한다.

미등록어 처리 알고리즘

단계 1 : 한 어절의 어소가 n개라면 n번째 어소가 사전에 존재하는 검색한다. 있으면 리스트에 저장한 후 단계 2로 가고 없으면 그대로 단계 2로 간다.

단계 2 : (n-1)(n)번째 어소가 사전에 검색되면 리스트에 저장하고, 아니면 (n-1)번째 어소만 다시 검색하여 존재하면 리스트에 저장한 후 단계 3으로 간다. (n-1)번째 어소가 존재하지 않으면 단계 3으로 간다.

단계 3 : (n-2)(n-1)(n)번째 어소가 사전에 검색되면 리스트에 저장하고 없으면 n을 뺀 (n-2)(n-1)두 어소를 검색하여 있으면 리스트에 저장한다. 두 어소가 없다면 (n-2)만 검색하여 있다면 리스트에 저장한 후 단계 4로 가며 (n-2)번째 어소가 없다면 단계 4로 간다.

단계 4 : 단계 1,2,3과 같이 반복한 후 한 어절에 대한 매칭될 수 있는 모든 단어들은 리스트에 저장되며 그 리스트를 화면상에 나타냄으로서 찾을 수 없는 어소에 대해서 사용자가 직접 사전에 등록할 수 있도록 한다.

Ⅱ-2. 품사, 품사속성 리스트

형태소 분석의 출력인 단어리스트를 생성해서 사전을 검색하고 각각에 대응되는 품사리스트와 품사속성리스트를 생성한다.

다품사를 취하는 단어는 문법규칙을 참조하여 하나의 품사에 매칭되게 한다. 즉, 품사속성만을 가지고 문법규칙을 형성하여 하나씩 매칭시켜 나간다.

입 력 문 : 〔본 논문은 기계번역 시스템에 관한 연구입니다〕

단어리스트 : 〔본, 논문, 은, 기계, 번역, 시스템, 에, 관, 한, 연구, 입니다〕

품사리스트 : 〔pren, n, aux, n, n, n, aux,n, omi, n, suf〕

품사속성리스트 : 〔mod,con,ind,con,abs,con,cau,ver,omil,abs, agg〕

Ⅱ-3. 조사의 격 생성과 다의성 분석

격조사는 체언에 붙어 그의 문법적인 기능을 나타내는 조사로[6] 다른 어휘와의 격관계를 표시해주는 기능을 가지며 1 : N형의 해석 패턴이 발생한다. 격조사의 다의성을 분석하기 위해서는 각 조사의 패턴별 분석이 필요하다.

다의성 조사를 처리하기 위하여 전접체언에 품사속성을 규정해줌으로서 의미를 제한한다. 즉, 한 예를 든다면 "종로에 간다"와 "망치로 친다"를 번역할 때 전접체언의 의미속성에 따라 격조사는 방향격(に)와 도구격(で)로 구분된다.

명사의 품사속성은 표1과 같이 분류하였다.

분 류	분 류 기 호
추 상 명사	abs
구 체 명사	con
생 물 명사	ani
무 생 물 명사	noa
인 칭 명사	hum
시 간 명사	tim
장 소 명사	pla
방 향 명사	dir
수 량 명사	num
지 시 명사	ind
동 사 형 명사	ver
형용동사형 명사	ave

표1. 명사의 품사속성

Ⅱ-4. 수식어 처리

한국어 문장에서 명사를 수식하는 수식어 중에서 형용사 활용어미가 동음어인데 비해 대응되는 일본어는 サ행변격동사활용과 형용동사 활용으로 각각 다르게 표기된다.

본 논문에서는 이것을 해결하기 위하여 명사의 품사속성을 12개로 세분하였다. 즉, 명사에 "하다"라는 접미사가 후접하여 동사로 나타낼 수 있는 동사형 명사와 형용사의 의미를 가진 형용동사형 명사로 품사속성을 분류하고 품사속성으로 이루어진 문법규칙을 이용하여 올바른 번역을 할 수 있도록 한다.

두 가지 예문으로 수식어의 동음활용어미의 처리를 설명하고자 한다.

예 1) 이것은 편리한 방식입니다.

예 2) 기계번역은 자연어 처리를 응용한 분야입니다.

이때 명사를 수식하는 수식어로 예 1) 에서는 "편리한"이 "방식"을 수식하고 있고 예 2) 에서는 "응용한"이 "분야"를 수식하고 있다. "편리한"과 "응용한"을 형태소 분석하고자 할 때 "편리하다"는 대응하고자 하는 일본어의 품사로서 그 상태를 나타내는 형용동사로 "편리" 자체를 형용동사형 명사로 인식하고 "한"은 활용어미로 분석한다. "응용하다"에서 "응용"이 명사의 속성과 サ행변격동사를 후접시켜서 동사의 속성도 갖고 있기 때문에 동사형 명사로 품사속성을 인식하고 "한"은 활용어미로 분석하여, 활용어미의 정확한 번역을 위해 전접체언과 활용어미의 품사속성을 가지고 문법규칙을 참조한다.

동사형 명사(ver)+활용어미(omi)일 때는 활용어미 대응어를 "した"로 번역발생하고, 형용동사형 명사(ave)+활용어미(omi)는 활용어미 대응어를 "な"로 번역 발생한다. 이때 "한"이 수사인지 활용어미인지의 구분은 전접하고 있는 단어여부로 판별하여 없을 때는 수사있을 때는 활용어미로 인식하게 한다.

단어리스트 : 〔이것, 은, 편리, 한, 방식, 입니다〕

품사리스트 : 〔 pron, aux, n, omi, n, suf 〕

품사속성리스트 : 〔 ind, ind, ave, omil, con, agg 〕

단어리스트 : 〔기계, 번역, 은, 자연어, 처리, 를, 응용, 한, 분야, 입니다〕

품사리스트 : 〔 n, n, aux, n, n, aux, n, omi, n, suf 〕

품사속성리스트 : 〔 con, abs, ind, abs, con, obj, ver, omil, abs, agg 〕

수식어 처리 알고리즘

단계 1 : 품사속성이 omil이면 앞에 단어가 있는지 여부를 검색하여 없으면 대응어 리스트의 argument1에 대응하고, 있으면 단계 2로 간다.

단계 2 : omil 앞의 품사속성이 형용동사형 명사이면 대응어 리스트에서 argument2를 대응하고, 품사속성이 동사형 명사이면 대응어 리스트에서 argument3로 대응한다.

단계 3 : omil 앞에 전접체언이 있고 후접하는 어미가 있는 경우는 동사어간 품사로 변환한다.

Ⅲ. 중문처리와 문형생성

Ⅲ-1. 중문처리

문형이란 문장을 이루는 낱말이나 문법소의 배열유형이라 할 수 있다.[9] 의미기능어인 특수조사는 문형의 생성요소는 되지 못하지만 체언에 후접하는 격조사가 있을 경우는 격조사에 의해 문형생성이 이루어진다.

기존의 논문에서는 간단한 단문형식을 번역예로 하였으나 본 논문에서는 콤마로 인한 중문형식을 해결한다. 우선 중문은 단문과 단문사이에 콤마로 연결되어 있어 입력문장이 중문인지 아니면 명사의 나열을 나타내는 나열단문인지를 인지할 수 있도록 문법규칙에 적용시켜서 매칭되는지를 알아본다.

품사리스트의 입력배열 중에 콤마가 있으면 다음 콤마가 나올 때까지의 품사를 검색한다. 그래서 품사가 명사로 되어 있으면 나열문으로 처리해 주고 다음 콤마가 없고 콤마 앞이 명사에 접미사나 동사활용이 후접된 경우는 중문으로 처리하여 콤마의 앞문장과 뒷문장으로 나누어 단문형식을 처리한 후 번역발생시 다시 연결해 준다.

예) 이 논문은 VIDEOTEX분야이고, 그 논문은 PC 통신입니다.

단어리스트 : 〔이, 논문, 은, VIDEOTEX, 분야, 이고,', 그, 논문, 은, PC, 통신, 입니다〕

품사리스트 : 〔 pron, n, aux, n, n, suf,', pron, n, aux, n, n, suf 〕

품사속성리스트 : 〔 ind, ads, ind, con, abs, agg,', ind, abs, ind, con, abs, agg 〕

분리한 품사속성리스트 : 〔〔 ind, abs, ind, con, abs, agg 〕, 〔 ind, abs, ind, con, abs, agg 〕〕

문형생성 〔SV〕

Ⅲ-2. 동사어간과 어미의 활용처리

한국어와 일본어의 특징중의 하나로 어미에 붙은 어미의 활용이 다양하기 때문에 동사의 시제는 현재형으로 한정지었다.

일본어 동사는 오단, 상일단, 하일단, サ행변격, カ행변격등의 동사에 따라 각기 어미활용이 다르므로 활용에 적합하게 동사어간 사전내의 한 argument로 어미리스트를 형성하고, 대응어리스트와 매칭되어 처리하는 방식을 취한다. 또한 "근거하다", "처리하다"를 형태소 분석하고자 할 때 국문법으로만 볼 때에는 명사와 "하다"라는 접미사로 구분될 수 있으나 대응되는 일본어로 표현할 때 "근거하다"는 "よる"인 오단동사로, "처리하다"는 "しょうする"라는 명사에 サ행변격동사가 연접하여 다르게 표기되므로 사전을 이용하여 이를 해결하였다. 즉 사전을 검색할 때에 한국어를 key자체로 보기 때문에 "근거하다"는 word('근거하', v,0,v11, yo, cauact, 〔'다', '지', '면'더라도', '든지'〕, 〔ru,ra,ri,reba,deattemo,youga〕). 로 사전에 등록

시키고, "처리하다"는 사전에 명사가 등록되어 있으므로, "하다"에 대한 サ행변격동사를 검색하여 인식한다.

예) "보면"

word ('보', v,1,v12, mi, percepst, ['다', '지', '고', '면', '더라도', '든지'], [ru, [ ], ri,reba,deattemo, youga]).

"보면"을 인식할 때 우선 "보"에 대한 어간사전을 검색하여 존재하는 지를 보고 다음에 "면"이라는 단어가 '보' 어간사전내에 있는 어미리스트의 argument와 매칭되는 지를 보고 매칭된다면 대응어리스트에서 같은 순서의 argument가 대응되도록 한다.

Ⅳ. 번역 발생

번역 발생시 입력정보는 한국어 문장의 형태소 리스트, 품사리스트, 품사속성리스트, 문형리스트등이며 이 정보를 기초로 하여 사전구성은 각 품사별로 단어를 분류하였고, 동사는 어간사전으로 구분하여 빠르게 검색하고 번역을 실행한다.

Ⅴ. 실행 결과

1) 인간공학은 인간을 활용한 학문입니다.
ningenkougakuwaningenogatsuyousitagakumondesu

2) 본 논문은 기계번역 시스템에 관한 연구입니다.
honronbunwakikaihonyakusisutemudesekisitakengkyuudesu

3) 이것은 편리한 방식입니다.
korewabenrisitahousikidesu

4) 기계번역은 자연어 처리를 응용한 분야입니다.
kikaihonyakuwasizengoshorioouyousitabunyadesu

5) 이 논문은 VIDEOTEX분야이고, 그 논문은 PC통신입니다.
konoronbunwaVIDEOTEXbunyade, sonoronbunwaPCtsuusindesu

Ⅵ. 결 론

본 논문에서는 형태소 분석단계에서 미등록어 처리를 병렬로 수행하였으며, 수식어 처리에 있어서 품사속성과 문법규칙으로 동음어미를 정확히 번역할 수 있도록 대응어를 구분하였다. 또한 콤마를 인식하여 중문을 번역하고, 동사의 다양한 어미활용은 규칙동사의 어간사전을 구성할 때 어미활용을 하나의 리스트로 작성하여 보다 빠른 시간내에 사전을 검색할 수 있도록 하였다.

본 시스템에서는 Prolog언어를 사용함으로써 상관적인 데이타 베이스(relational data base)를 논리적으로 다룰 수 있었으며, 여러 규칙들에 의해 사실들을 조합하여 새로운 사실을 유도(deduction)할 수 있었고 IBM-PC상에서 실행하였다. 또한 문장이 난해한 중문이나 복문 처리와 불규칙동사의 활용에 관해서는 연구가 진행중이다.

참 고 문 헌

1. 임선태, 백모현, 진중화, 주인숙, 임인철, "Videotex 상호접속을 위한 한국어-일본어 기계번역 시스템의 설계", 대한 전자공학회 추계 종합 학술대회 논문집, vol,9 no,2 86/12
2. 임경심, 박용진, 임인철, 박창호, 이기석, "한국어-일본어 기계번역 시스템에 있어서의 격패턴에 의한 동사의 의미해석"'대한 전자공학회 추계 종합 학술대회 논문집, vol,8,no,2,85/11
3. 안동언, 최기선, 김길창, "기계번역을 위학 한국어 해석에서 형태소로 부터 구문요소의 형성에 관한 연구", 인공지능 연구회 발표록, 1987/3
4. 한광록, 박영서, 이주근, "한-일 기계번역 system(Ⅲ)", 대한 전자공학회 하계 종합 학술 대회, 논문집, 1986/6
5. kosho SHUDO, Toshiko NARAHARA, and Sho YOSHIDA, "A Structural Model of Bunsetsu for Machine Processing of japanese", 일본 정보처리학회 논문집, 1979/12
6. 용사만, 국어특수조사론, 학문사, 1983
7. 최창열, 한국어의 의미구조, 한신문화사, 1983
8. 김형규, 국어학 개론, 일조각, 1982
9. 서정수, 문형의 습득과 문장작성, 한양대 국어국문학과