中韓 기계 번역을 위한 中國語 單語 自動 분리

宋 寅 聖*

〈 目 次 〉

1. 기계 번역과 방법론[1)]

이 글에서는 기계 번역을 간략히 소개하고, 中韓 기계 번역을 위한 첫 단계로서의 단어 자동 분리 방법과 관련 문제점을 토론하기로 한다.

機械 飜譯(中-機器飜譯/英-Machine Translation)은 컴퓨터 시스템을 이용하여 한 언어의 텍스트[2)]를 다른 언어의 텍스트로 치환하는 것이라고 정의할 수 있다. 이것은 자연언어를 대상으로 한다는 점에서 인간

* 고려대 민족문화연구원 연구조교수(e-mail: sis99kr@yahoo.co.kr)

1) 이 부분은 주로 한정한(2001)을 참조하였다.

2) 최근 Phone-to-Phone MT 에서처럼 음성 입력을 음성 출력으로 번역하는 경우, 음성의 치환도 포함될 수 있다.

번역(Human Translation)과 매우 유사하나, 後者가 認知的 번역 중에서도 情感的 번역까지도 수행하는 等價 飜譯을 지향하는데 비해,前者는 認知的 번역 중에서도 對應 가능한 水準에서만 번역이 이루어지는 對應 번역이라는 점에서 차이가 있다.

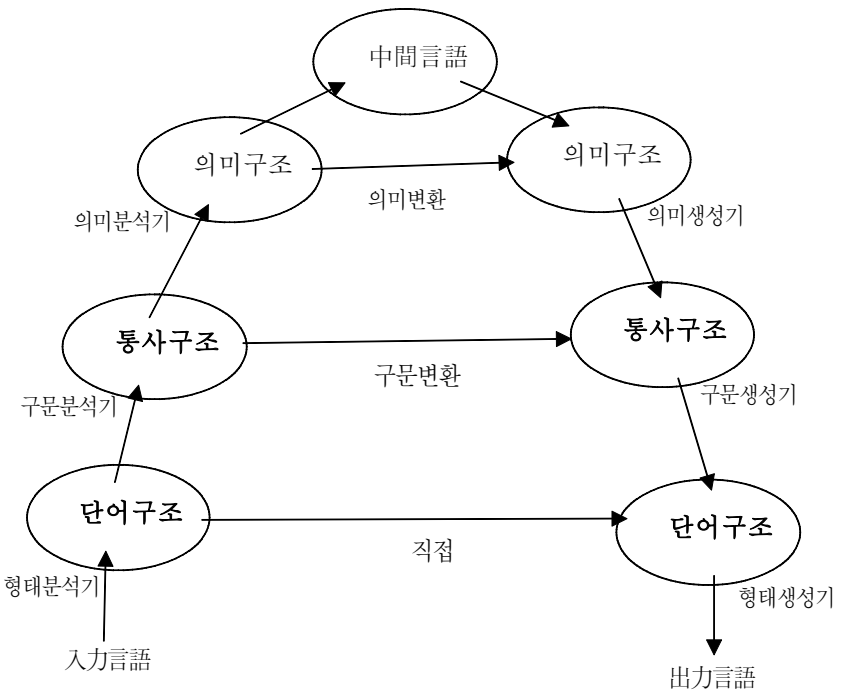
또한 機械 飜譯은 사람이 하는 것이 아니라, 컴퓨터 시스템을 이용한다는 점에서 아래와 같이 人間 飜譯과 다른 몇 가지 중요한 차이가 있다.

1) 入力 言語와 出力 言語를 컴퓨터 可讀型 부호로 바꾸어야 한다.
2) 入力 言語와 出力 言語의 문법(음성, 형태, 구문, 의미) 정보들이 언어별 電子 辭典에 저장되어 있어야 한다.
3) 入力 言語와 出力 言語의 생성을 위해 문법 구문 해석 규칙(Parsing)이 필요하다.
4) 入力 言語와 出力 言語의 對應 관계를 처리하기 위하여, 兩者 사이의 변환 처리규칙이 필요하다.
5) 多國語 기계 번역 시스템을 만들 때 번역의 경제성을 증대하기 위한 방법, 즉 人工 言語가 필요하다.

여기서 1)은 문자 코드의 統一 問題를 가리킨다. 中.日.韓 三國에 걸쳐 있는 漢字 코드 統一 問題가 여기에 해당되고, 한글 옛글자의 入出力 처리도 아직 Unicode 상에서 구현되고 있지 않다. 2)는 종이 사전이 人間 飜譯에서 보조적인 역할에 그치는 것과 달리, 機械 飜譯의 電子 辭典은 오직 辭典에 등재된 것만 번역할 수 있다는 점에서 필수적이라는 것을 뜻한다. 또 言語 분석에 사용되는 단어간 결합 정보(連語 3) 辭典)나, 세계 지식에 사용되는 워드넷(Wordnet), 코퍼스(Corpus)와 같은 것도 電子 辭典에 포함될 수 있다. 3)은 음성인식기(생성기),형태소 분석기(생성기),구문 분석기(생성기)와 같이 구체적으로 言語 분석과 생성에 사용할 수 있는 단위화된 문법 구문해석 규칙 기술들을 말한다. 이와 반대로, 人間 飜譯은 직관적 문법 지식을 사용한다. 4)는 入力 言語의 분석 결과를 出力 言語에 對應시키는 對譯語 선택 과정으로, 한

3) 2개 이상의 단어가 결합하여 더 복잡한 관념을 나타내는 언어.collocation이라고 함.

言語의 辭典 정보 뿐 아니라, 모든 文法 정보까지 변환된다. 5)는 두 가지 言語 機械 飜譯에서 多種의 言語 機械 飜譯 에로의 전환 과정에서 제기되는 변환 과정의 복잡성 문제로, 어떻게 변환 과정의 수를 현저히 줄일 수 있는가 하는 문제이다. 이 경우의 人工 言語는 自然 言語일 수도 있고, 인공적인 中間 言語 일 수도 있다.



기계 번역을 하기 위해서는 原文에 대한 분석이 先行되어야 하는데, 分析을 어느 정도 하는가에 따라서 直接(direct), 變換(transfer), 中間言語(interlingua) 방식 등으로 분류되는데, 위 그림은 分析 水準과 번역 방식의 관계를 보여 준다.

直接 方式은 단어(또는 형태소) 사이의 直接 번역에 의존한다. 이것은 英-佛, 英-獨, 日-韓 등과 같은 구조가 類似한 言語 간에 효과적으

로 사용될 수 있다. 그러나 入力語의 구문, 의미 分析을 하지 않으므로, 특히 異質的인 言語 유형간의 고품질 번역을 기대할 수 없다.

이에 반해 變換 方式은 入力 言語와 出力 言語의 형태·구문 분석을 수행하므로, 언어간 번역 과정에서 나타나는 형태·구문 重意性을 처리해 줄 수가 있다. 그러나 相異한 語族간의 심층적 의미 분석이 필요할 경우와 다른 새로운 言語가 계속해서 번역 시스템에 추가될 경우에, 그 때마다 추가 번역 규칙을 만들어야 한다는 短點이 있다.

中間 言語 방식은 1)형태,구문 분석은 물론, 심층적인 의미 분석과 어휘 해체까지를 거치므로 入出力 言語가 바뀔 때마다 새로운 變換 규칙을 만들 필요가 없다. 2)入力 言語와 出力 言語가 固定되지 않기 때문에 기존의 번역기와 달리 多國語 번역기 개발에 容易하다. 3)相異한 言語 구조를 가지는 言語들간의 자연스러운 번역을 보장해 줄 수 있다. 등의 長點이 있다. 그러나 동시에 眞正한 의미의 言語 보편적인 中間 言語를 위한 통일된 의미 표시 구축 方法이 없고, 광범위한 의미 영역으로 확대되지 못한다는 短點도 존재한다.

이상의 직접. 변환. 중간 언어 방식 외에, 통계 기반 방식. 예제 기반 방식. 신경망 기반 방식[4]등이 있다.

2. 中國語 문법의 言語 단위

中國語 문법의 체계는 學者에 따라 여러 가지 학설이 있으나, 本文에서 사용한 中國語 文法의 기본 체계는 中國 大陸에서 1984년에 공포된 中.高校 學校 文法 體系인 "中學教學語法系統提要(試用)"[5]로, 이에 따른다면 문법상의 言語 단위는 1)형태소(語素) 2)단어(詞) 3)구(短語)

4) 자세한 것은 김영택외(2001:283-293)을 참조할 것.
5) 中國 大陸의 中.高校 학교 문법 체계로는 1956년에 공포된 "暫擬漢語教學語法系統"이 있고, 이를 修正.補完하여 1984년에 공포한 "中學教學語法系統提要(試用)"가 있다.後者는 教學 方案이 아니고,文法 체계와 명칭을 설명한 개요에 지나지 않으나,현재 중국 대륙의 中.高校 語文 교재는 이를 적극 反映하고 있음.

4)문장(句子) 5)문장 결합(句群)이다. 아래에서는 이를 간략히 소개하여, 뒤에서 中國語 단어 자동 분리를 토론할 때 참고가 되게 한다. 또한 中國語 文法 用語의 우리말 對譯語는 1991.6 "韓國中國言語學會"에서 통과된 '中國語文法用語 統一試案"[6]을 참조하였다.

2.1 형태소(語素)

형태소는 소리(音)와 뜻(義)이 결합되어 있는 최소의 문법상의 언어 단위이다. 형태소는 여러 가지 기준에서 분류할 수 있는데, 보통 아래와 같이 크게 3가지로 분류한다.

1) 음절 형식에 의한 분류: 아래와 같이 單音節. 2音節. 多音節 형태소로 분류된다.

가) 單音節 형태소는 漢語형태소의 기본 형식이다.(예:天.民.語.化.員)
나) 2音節 형태소는 주로 古代漢語[7] 에서 전해 내려온 "聯綿詞"[8](예: 惆悵 / 葫蘆)와 音譯된 외래어(예:琵琶 /琵琶 /尼龍 /咖啡)이다.
다) 多音節 형태소는 주로 音譯된 외래어이다
예:凡士林-바셀린(vaseline), 托拉斯-trust. 奧林匹克-올림픽 , 布爾什維克-볼세비키.

2) 형태소의 활동성에 의한 분류: 아래와 같이 3가지로 분류한다.

가) 自由 형태소: 단독으로 단어를 이룰 수 있는 것.(예:我, 書 등)
나) 半自由 형태소: 단독으로는 단어를 이룰 수가 없고, 다른 형태소와 결합해야만 단어를 이룰 수 있는 것(예:技--技術/科技 등).

---

6) 全文은 『중국언어연구』1(1997 서울:학고방)의 310-315쪽에 게재되어 있으며, 또한 강식진외 3인이 共編한 『進明中韓辭典』(1997 서울:進明出版社)의 부록 3-5쪽에도 게재되어 있음.

7) 古代 漢語는 일반적으로 先秦.兩漢.六朝 시기의 漢語를 가리킴.

8) 2음절로 聯綴되어 이루어지고, 분리되어서는 의미를 갖지 못하는 단어.

다) 不自由 형태소: 단독으로 단어를 이룰 수가 없을 뿐더러, 다른 형태소와 결합할 때, 위치가 늘 고정된 것.(예: 第 --第一, 第十 등 /們 --你們,人們 등)

3) 의미에 의한 분류: 아래와 같이 3가지로 분류된다.

가) 實素: 의미가 비교적 實在적인 것.(예:天,大,走등)
나) 虛素: 實在적인 의미를 나타내지는 못하고, 문법 관계를 나타내는 것. (예:和,而 등)
다) 半實素(또는 半虛素):일정한 의미를 가지고 있으나, 實素처럼 선명하지 못한 것. (예:第,化 등)

2.2 단어(詞)

단어는 형태소가 결합된 것으로, 독립적으로 운용할 수 있는 最小의 단위로서, 크게 아래와 같이 實詞와 虛詞 2가지로 나뉜다.

1) 實詞: 名詞, 動詞, 形容詞, 代詞, 數詞, 量詞
2) 虛詞: 副詞, 전치사(介詞), 접속사(連詞), 助詞, 감탄사, 의성사(擬聲詞)

實詞는 實在적인 의미를 표시하며, 句나 문장의 성분이 될 수 있고, 단독으로도 문장을 이룰 수가 있는 반면, 虛詞는 이와 반면에 實在적인 의미를 표시하지 않으며, 句나 문장의 성분이 될 수 없고(단지 副詞는 부사어가 될 수 있다), 기본 용도는 어법관계를 표시한다.

2.3 구(短語)

구(短語)는 詞組라고도 하며, 단어로 구성되었다. 구는 문장 성분이 될 수 있고, 대다수의 句에다 일정한 語調가 첨가되면 문장이 될 수 있다. 句는 기능에 따라서 名詞句. 動詞句. 形容詞句. 主述句. 전치사구

등으로 나뉜다.

2.4 문장(句子)

문장은 언어의 사용 단위로서, 단어 또는 句로 구성된다. 모든 문장은 일정한 語氣. 語調가 있다. 정상적으로 계속 얘기할 때, 문장과 문장 사이에는 비교적 커다란 休止가 있어, 글에서는 일정한 문장 부호(마침표. 물음표. 느낌표)로 표시한다. 1개의 문장은 상대적으로 1개의 완전한 의미를 나타내고,1차례의 간단한 交際의 임무를 수행할 수 있다.

2개 혹은 2개 이상의 단문은 복문을 이룬다. (예:老師愛護學生,學生愛護老師. /我們需要把學到的某些知識記住,但是死記硬一些並不眞正理解的東西沒有用處.)

2.5 문장 결합(句群)

句群은 句組.語段이라고도 하며, 앞뒤로 여러 개의 문장이 이어진 것을 말한다. 1개의 문장 결합에는 1개의 명백한 중심 의미가 있다.

3. 中國語 단어 自動 분리의 기초

中國語는 英語나 한글과 달리 띄어 씌기를 하지 않기 때문에 단어를 자동으로 분리하는 것이 기계 번역을 위해 가장 시급히 해결해야 할 문제인데, 먼저 中國語에 있어서 단어 자동 분리가 중요한 이유를 3가지로 나누어 논술한다.

(1) 단어 자동 분리는 現代 中國語 구문 분석기의 기초 작업이기 때문이다.

中國語 자연 언어 처리에는 광범위한 응용 분야가 있는데, 예를 들

면 問答 시스템.中國語-外國語 기계 번역 등에서 대상 텍스트를 입력할 때 구문분석(Parsing)은 필수 불가결한 단계이다. 그런데 컴퓨터가 문장 분석을 처리할 때 의존하는 어법 지식은 電子 辭典과 문법 규칙 사전에 의존한다. 電子 辭典은 中國語 단어의 詞法(형태론).句法(통사론).語義 知識 등을 수록하였는데, 일반적으로는 통사 규칙은 품사 등의 기초 위에서 만들어 진 것이다. 따라서 중국어 문장은 우선 반드시 단어를 분리한 후에야 비로서 문장 분석을 할 수 있기 때문이다.

(2) 단어의 計量 분석은 단어 빈도 통계, 새로운 단어의 식별, 컴퓨터 보조 사전 편찬, 단어 배합 연구, 문장 혹은 작자의 스타일 연구 등 여러 분야에 광범위하게 사용되고 있기 때문이다.

예를 들면, 사전 편찬시 등재한 해당 단어 아래의 용법 및 용례는 편집자가 임의로 삽입하는 것이 아니고, 대규모의 코퍼스에서 추출한 것이다. 그런데 이러한 코퍼스의 텍스트는 모두가 語料를 분리하여, 품사 태깅 등을 한 후에 사용하게 되는 것이다. 따라서 단어 자동 분리는 가장 기본적인 작업인 셈이다.

(3) 中國語 문헌의 자동 처리는 우선 단어 분리를 해야만, 단어 의미, 통사 구조 등 좀더 깊이 있는 언어 지식이 활용될 수 있게 되기 때문이다.

예를 들면, 자동 색인. 자동 요약. 자동 분류. 정보 검색 등에서 單語야말로, 가장 핵심적인 요소이다.

아래는 中國語 단어 분리에 대한 기본적인 처리 방법이다.

① 문장 중간에 공란이 있거나, 공간 및 문장 부호(마침표, 쉼표, 작은 쉼표, 쌍반점, 쌍점, 물음표, 느낌표 등)가 있을 경우, 분리한다.

② 2글자 또는 3글자

2글자 또는 3글자로 이루어진 단어 외에 결합이 긴밀하고 안정적으로 사용되는 2글자 또는 3글자로 구성된 구는 함께 분리한다.

예) 1)發展　　可愛　　紅旗　　自行車
　　2)對不起　　青霉素(페니실린)

③ 4글자

가) 成語는 함께 분리한다.

예)胸有成竹　　欣欣向榮

나) 4글자로 이루어진 단어 또는 긴밀하게 결합하고 안정적으로 사용되는 4글자로 이루어진 句도 역시 함께 분리한다..

예)社會主義　　春夏秋冬　　由此可見

④ 5글자 및 그 이상

5글자 및 그 이상으로 이루어진 속담.격언 등은 분리 후에 원래의 의미가 있을 경우는 각각 분리한다.

예)時間/就/是/生命　　失敗/是/成功/之/母

※그러나 긴밀하게 결합하고 안정적으로 사용되는 구이긴 하나, 분리 후에 원래 조합된 의미를 상실할 경우는 분리하지 않는다.

예)不管三七二十一(成語로 "앞뒤를 가리지 않다", "무턱대고 처리하다"는 의미)

⑤ 慣用語와 의미가 바뀐 단어나 句는 의미가 바뀐 言語 환경하에서는 함께 분리한다.

예) 半邊天 --- 婦女能頂半邊天(여자들이 이 세상의 반을 지탱한다.)
　　鐵公鷄 -他眞小氣, 象個鐵公鷄(그는 정말 인색해서, 구두쇠 같다.)
　　吃香　--他在公司裏很吃香(그는 회사에서 아주 인기가 있다.)

⑥ 약어는 함께 분리한다.

예)科技(科學技術)　　工農業(工業農業)

⑦ 단어에 兒가 첨부된 것은 그 단어와 합하여 함께 분리한다.
예)花兒 玩兒

⑧ 현대 중국어에서 출현하는 漢字가 아닌 부호, 예를 들면 아라비아 숫자, 수학 기호, 화학 부호 등은 원래 형태를 유지한다.
예)3.14 cm Co2

⑨ 현대 중국어 중 다른 언어의 音譯語는 분리하지 않는다.
예)巧克力(chocholate) ／吉普(jeep)

⑩ 다른 언어 환경에서의 同一한 單語는 언어환경에 의하여 相異하게 분리한다.
예) 가)把/手/抬起來.(손을 들어 올리다)
나)這个/把手/是/木製的.(이 손잡이는 나무로 만든 것이다.)

4. 中國語 품사별 단어 분리

아래에서는 중국어 품사를 아래 12가지로 나누어 자세히 살펴본다.

1) 實詞: 名詞.動詞.形容詞.代詞.數詞.量詞.
2) 虛詞: 副詞.전치사(介詞).접속사(連詞). 助詞. 감탄사. 의성사(擬聲詞)

4.1 名詞

4.1.1 보통 名詞

1) 2글자로 이루어진 名詞나 결합이 긴밀한 구는 함께 분리한다.
예)火車 牛肉 鋼鐵

2) 결합이 긴밀하나, 분리된 후에 원래의 의미를 상실하는 명사구

는 함께 분리한다.

예)有效功率　　被子植物

3) 形容詞에다 名詞가 더해진 구는 각각 분리한다.

예)綠/葉　　小/床

※ 단, 形容詞에다 名詞가 더해져서 뜻이 변경된 구는 함께 분리한다.

예)小媳婦(새댁)　　戴／高帽兒(남을 추켜 세우다)

4) 접두사에 명사가 첨가된 것은 함께 분리한다. 예)阿哥 / 非金屬

5) 명사에 家, 手, 性, 員, 子, 化, 長, 頭, 者 등의 접미사가 첨가된 것은 함께 분리한다. 예)科學家 革命性 木頭 學者

※ 名詞에 여러 개의 접미사가 첨부되었을 경우, 모두를 1개의 분리 단위로 간주한다. 예)物理學家

6) 명사의 앞뒤에 접두사 또는 접미사가 첨가될 때는 모두를 1개의 분리 단위로 간주한다. 예)非黨員 初導性(초전도성)

7) 각 전문 분야의 기본 용어는 1개의 분리 단위로 간주한다.

예)加速度 中央處理機

8) 方位詞는 단독으로 분리한다. 예)卓子/上 長江/以北

9) "人們"이외에 단지 복수를 의미하는 "們"은 단독으로 분리한다.

예)朋友/們 學生/們

10) 時間詞 또는 구는 아래와 같이 처리한다.

가) 1년의 12개월 및 매주의 7일은 함께 1개의 단어로 분리한다.

예)五月 元月 3月 星期日 禮拜三

나) 年. 月. 日. 分. 秒 등은 각각 분리단어로 간주한다.

예)2002/年 4/月 12/日 2/時 30/分 10/秒

다) "前.后.上.下.大前.大后"등 時間詞 또는 量詞와 직접 결합할 경우, 함께 분리한다.
예)前天 后年 上星期 大后年

라) "初"에 10이내의 숫자가 더해지면 일괄적으로 함께 분리한다.
예)初一 初八

4.1.2 固有名詞

(1) 人名.칭호 등은 아래와 같이 처리한다.
가) 漢族 人名의 姓과 이름은 각각 따로 분리한다.
예)張/大年 諸葛/亮

나) 기타 국가와 민족의 人名은 습관에 따라 분리한다.
예)卡爾／馬克思(karl Marx) 牛頓(Newton)

다) 직책의 칭호는 일괄적으로 나누어 분리한다.
예)張/教授 王/部長

라) 약칭.존칭 등은 함께 분리한다.
예)小李 郭老 老張

마) 長幼有序의 순서를 나타내는 가족 간의 순서는 단독분리한다.
예)三/叔 大/女兒

(2) 민족명, 지명 중의 "族,省,市,州,縣,鄉,區,江,河,山" 등은 단독으로 분리한다.
※ 단 '族,省,市,州,縣,鄉,區,江,河,山" 등을 포함하여 2글자가 되

는 것은 함께 분리한다.

예)漢族　長江　北京/市　浙江/省

※ 街.路.村鎭명칭과 각 大洋.大海는 모두 함께 분리한다.

예)長安街　學院路　大西洋　地中海

(3) 국가명은 전체를 분리한다.

예)中華人民共和國　大韓民國

(4) 조직. 기구 등의 명칭은 전체를 함께 분리한다.

예)聯合國　中國　共産黨

(5) 상품의 상표. 품종.상 품의 시리즈명 등 고유명사와 보통 명사는 따로 분리한다.　예) 牧丹/III/型

4.2 動詞

(1) 動詞의 중첩 형식은 아래와 같이 분리한다.

가) 1글자의 동사 중첩은 함께 분리한다.

예)看看　動動

나) 2글자로 이루어진 動詞의 중첩 방식이 AABB일 경우, 함께 분리한다.

예)來來往往　拉拉扯扯

다) AAB.ABAB 중첩 형식의 동사구는 각각 분리한다.

예)說說/看　研究/研究

라) A一A.A了A.A了一A 등 중첩 형식의 동사구는 각각 별도 분리한다.

예)談/一/談　　想/一/想　　讀/一/讀
想/了/想　　想/了/一/想

(2) 동사 앞의 否定 副詞는 별도 분리한다.
예)不/能　　未/完成　　沒/吃

(3) 肯定+否定 형식의 의문 표시 동사구는 별도 분리하되, 완전한 의미를 나타내지 못할 경우는 함께 분리한다.
예)看/不/看
相信/不/相信 -- 相不相信

(4) 動詞+目的語 형식의 단어 또는 결합이 긴밀하고 안정적으로 사용되는 2글자로 이루어진 動詞+目的語 구는 함께 분리한다.
예) 跳舞　　開會
孩子/該/念書/了　　解決/吃飯/問題
※ 단, 결합이 긴밀하지 않은 것은 별도로 분리한다.
예)學/英文(中文.滑氷.....)　　寫/信(文章.書.論文.....)
※ 動詞+目的語 형식의 단어 또는 구 중간에 기타 성분이 삽일될 경우, 단독 분리한다.
예)吃/一/頓/飯　　跳/新疆/舞

(5) 動詞+補語 구조의 2글자로 구성된 단어 또는 결합이 긴밀하고 안정적으로 사용되는 動詞+補語 구는 함께 분리한다.
예)提高　　打倒
※ 단 2+1 또는 1+2 글자로 이루어진 動補구는 별도 분리한다.
예)整理/好　　說/淸楚　　解釋/淸楚
※ 動補 구조의 단어 또는 구 사이에 "得""不"가 첨부되면 각각 별도 분리한다.
예)打/得/倒　　提/不/高

(6) 수식 구조의 단어 또는 결합이 긴밀하고 안정적으로 사용되는 수식 구조의 구는 함께 분리한다.

예)胡鬧 死背

(7) 복합 방향 動詞는 함께 분리한다.

예)出去 進來

※ 단 중간에 "得""不"이 삽일 될 경우는 각각 별도 분리한다.

예)出/得/去 進/不/來

(8) 動詞와 방향 동사가 결합한 구는 각각 별도 분리한다.

예)寄/來 跑/出去

(9) 1글자 또는 여러 글자의 動詞로 각각 독립적인 의미를 지닐 때는 각각 별도 분리한다.

예)聽/說 讀/寫

調查/研究 宣傳/鼓動

4.3 形容詞

(1) 형용사 중첩 형식인 AA.AABB.ABB.AAB. A里AB는 모두 함께 분리한다.

예)大大 高高興興

綠油油 馬里馬虎

※ 단 ABAB 형식의 形容詞는 별도 분리한다. 예)雪白/雪白

(2) 一A一B.一A二B.半A半B.半A不B.有A有B 등의 형용사성 구는 함께 분리한다.

예)一心一意 一淸二楚 半生不熟 有條有理

(3) 形容詞 병렬 구조는 아래와 같이 처리한다.

가) 2글자로 이루어진 形容詞로 품사가 변경된 것은 함께 분리한다.

예)長短　　深淺　　大小

나) 각자 원래 의미를 지니고 있는 구는 별도 분리한다.
예)光榮/偉大　　大/小/尺寸

다) 색깔 관련 형용사 또는 구는 함께 분리한다.
예)淺黃　　橄欖綠

라) 肯定+否定 형식의 의문 표시 形容詞구는 별도 분리한다. 단 불완전한 것은 함께 분리한다.
예)容易/不/容易　　--　容不容易

4.4 代詞

(1) 1글자로 이루어진 代詞에다가 們 이 첨부되면 함께 분리한다.
예)我們　　他們　　你們

(2) 지시 및 의문 대사 "這。那。哪" 에 "些。樣。么。里。邊"등이 첨부될 때는 함께 분리한다.
예)這个　這么　這邊　　那个　那么　那邊　　哪个　哪么　哪邊

(3) 지시대사 "這。那。哪" 에 數詞.量詞.名詞性 단어가 첨부될 때는 각각 별도 분리한다.
예) 這／十／天　　那／種

(4) 의문 대사 또는 구는 함께 분리한다.
예)多少／怎樣／爲什么／什么

(5) "各.每.某.本.該.此.全"등의 代詞와 그 뒤의 量詞는 별도 분리한다.
예)每／種　　該／單位　　全／校

4.5 數詞

(1) 數詞와 量詞는 별도 분리한다.
예)三/個 一/種

(2) 數詞는 하나 하나 별도로 분리한다.
예)三/百/五/十/四

(3) 序數를 표시하는 "第"와 그 뒤의 數詞는 별도 분리한다.
예)第/三 第/十五

(4) 分數 가운데 "分之"는 별도로 분리한다.
예)百/分之/三

(5) 숫자가 병렬하여 대략적인 숫자를 표시할 경우, 개략적인 것을 의미하는 숫자는 함께 분리한다.
예)十/七八/歲

(6) 대략적인 숫자를 표시하는 "多.來.几"등이 數詞 혹은 量詞 뒤에 첨가될 때 함께 분리한다.
예) 兩点多 十來人 十几个

(7) "些,一些,點兒"등 개략을 표시하는 단어가 형용사 또는 동사 뒤에 첨부될 때는 별도 분리한다.
예) 大/些 懂/一些 快/点兒

(8) "近,約,數"등이 數詞앞에 첨부되어 개략적인 숫자를 표시할 때는 별도 분리한다.
예) 近/百/人 約/四/千 數/萬
※ 단, "成.上"등이 數詞 앞에 첨부되어 함께 개략적인 숫자를 표시

할 때는 함께 분리한다.

예)成百　　上千

4.6 量詞

(1) 중첩하여 사용하는 量詞는 함께 분리한다.

예)年年　　家家戶戶　　天天

(2) 복합 量詞 혹은 구는 함께 분리한다.

예)人次　人年

4.7 副詞

(1) 副詞는 모두 별도 분리한다.

예)很/好　　剛/走　　互相/協助

(2) 자주 사용되고, 副詞 작용을 하는 구는 함께 분리한다.

예)越來越　　不能不　　不得不

※ 단 연결 작용을 하는 "越...越...". "又....又...."는 별도 분리한다.

예)越/來/越/熱　　又/便宜/又/好吃

4.8 전치사(介詞)

전치사는 모두 별도 분리한다.

예)按照/規定　　走/向/勝利

4.9 접속사(連詞)

접속사는 모두 별도 분리한다.

예)光榮/而/偉大　　工人/和/農民

4.10 助詞

(1) 構造 助詞 : "的.地.得.之.所"는 모두 별도 분리한다.
예)說/得/快　　成功/之/路
慢慢/地/走　　美麗/的/城市
所/認識

(2) 動態助詞 : "了.着.過"는 별도 분리한다.
예)看/了　　看/着　　看/過

(3) 語氣助詞 :모두 별도 분리한다.
예)你/呢?　　快/去/吧!

4.11 감탄사

감탄사는 모두 단독 분리한다.
예)唉呀／他／走／了!　　啊／眞／美!

4.12 의성사(擬聲詞)

의성사는 모두 단독 분리한다.
예)嘟(기적. 나팔. 피리 따위의 소리)　汪汪(개 짖는 소리)

5. 中國語 단어 분리의 문제점

위에서 中國語 단어 분리 문제를 기초적인 분리와 품사별 분리를 중

심으로 살펴보았지만 아직도 적지 않은 문제가 존재하는데, 여기서는 未登錄語와 重義 현상, 2가지를 중심으로 살펴보기로 한다.

5.1 未登錄語

未登錄語란 電子 辭典은 물론, 종이 辭典에도 등재되지 않은 中國 및 外國의 人名.地名.기구 조직명.사건명.화폐명.약어.파생어.각종 전문용어 외에 현재에도 계속 출현하고 있는 新造語들을 포함한다. 아래에 몇 가지로 나누어 고찰하기로 한다.

(1) 中國 人名의 自動識別

中國語 人名은 아래와 같은 특징이 있어, 컴퓨터가 식별하기에는 복잡하다.

가) 서양인의 人名과 같은 형태 특징(大小 文字)이 없다.

나) 姓名의 출현 형식이 다양하여 구조가 복잡하다.

(예: 姓氏+이름/姓氏 또는 이름만 출현/ 別名/筆名/어렸을 때의 이름 등)

다) 어떤 사람들은 다른 사람과 이름이 중복되는 것을 피하기 위해서 잘 사용되지 않고 심지어는 辭典에도 등재가 되지 않은 글자를 사용하여 이름을 짓는다.

예) 喆(哲) /犇(奔)

※ 어떤 사람들은 西洋式의 이름을 지으며, 어떤 사람들은 姓氏는 사용하지 않고, 이름만 사용한다. 예)王瑪麗 /李約翰

※ 몇 가지 오류의 실례

가) 인명의 약칭에 대한 처리 능력 부족--예)阿春/阿蓮/小曾/老葉 등

나) 이름이 2글자인데, 하나로 잘못 식별함--예)民警劉海東送老人回家。(劉海東)

다) 이름이 1글자인데, 2글자로 잘못 식별함---예)部長强衛和烈士崔大慶(强衛)

라) 이름을 국가이름으로 잘못 식별함--예)盛中國小提琴獨奏音樂會(盛中國)

(2) 中國 地名의 자동식별

中國地名委員會가 編(1994)한『中華人民共和國地名錄』(北京:中國社會出版社)에 10만 개 정도의 地名이 수록되어 있긴 하나, 아직도 수많은 도시의 도로명, 골목명. 시골마을 명등이 수록되어 있지 않아, 향후 수많은 地名을 수집하여 수록해야 한다.

또한 동일한 地名이 同義語.略語 외에 古今이 상이한 형식으로 출현하기 때문에 더욱 인식이 어렵다.(예 北京--京/天津--津)

(3) 외국의 번역명

중국어는 외국의 人名.地名등의 고유명사를 그대로 쓰지 않고, 中國語로 번역하여(주로 音譯 /義譯 /音譯+義譯의 형식임)사용하기 때문에 中國人 사이에도 漢字 표기가 相異할 정도이다.

예)可口可樂(coca cola) 卡車(트럭--car+車)

可利福尼亞/州(캘리포니아/주)--可利福尼/亞州(/아시아주)

5.2 重義 현상

1개 문장에서 동일한 단어를 어떻게 분리.분석하는 가에 따라서 한국어 번역이 相異해 진다. 아래에 몇가지 예를 들어 본다.

예)

1) 把:

你/把/兄弟/看成/什麽/人了---他們/是/把兄弟(의형제)

2) 在:

가)동사--他/在/家

나)전치사--他/在/家/吃/飯.

다)부사---他/在/吃/飯

3) 研究生/會/採取/行動-- 研究生會/採取/行動

6. 맺는 말

배재석(2001)은 "현재 국내에서는 중한 기계 번역이 전무한 상태로, 앞으로 이 분야에 관심을 가지고 연구를 진행해야 한다고 본다"라고 하였지만, 中韓 기계 번역은 이미 수 년 전부터 韓國科學技術院(KAIST). 포항공대 등과 코난테크놀로지.두레소프트.한국전자통신연구원(ETRI)등 주로 컴퓨터 학계를 위주로 연구가 진행되어 왔다.

韓國科學技術院(KAIST)는 이미 1995년에 『한국어.중국어간 기계 번역 시스템』이라는 보고서를 과학기술정책관리연구소에 제출한 바 있으며, 현재 中國人 學者와 함께 中韓 기계 번역 시스템을 개발 중이고, 포항공대에서는 유니소프트와 함께 규칙및 예제 기반의 서버용 中韓 기계 번역기를 제작하여 年內에 시판할 예정이다. 두레소프트에서는 인터넷 채팅용 英.日.中.韓 4개 국어 동시 번역 엔진(DMTS)을 개발하였으며, 코난테크놀로지 및 한국전자통신연구원(ETRI) [9]은 각각 中韓 및 韓中 기계 번역기를 개발 중이다. 인문계에서는 (주) 언어과학 및 고려대 민족문화연구원에서 중국어 기계 번역을 위한 연구를 진행중이다.

지금까지의 연구는 대개 컴퓨터와 中國語學 전공자가 따로 따로 연구를 해 오고 있는 상황으로, 앞으로는 서로 힘을 합쳐 연구를 한다면 짧은 기간 내에 보다 좋은 결과가 있으리라 생각된다.

中韓 기계 번역은 英韓 및 日韓 기계 번역과는 달리, 앞으로 해야 할 일이 무궁무진하여 적지 않은 시간을 투자해야만 어느 정도 번역의 정확율을 확보하게 될 것으로 여겨져, 특히 젊은 中國語學 전공 대학원생들이 이 방면에 많은 관심을 갖게 되길 희망한다.

9) ETRI는 2001.3--2003.3 까지 "東洋 4國 言語(우리말외에 英語.中國語.日本語 등을 가리킴)에 대한 TM 기반 통합 시스템"을 수행 중으로, 자세한 것은 황은하 외 2명(2002)를 참조할 것.

■ 參考文獻 ■

강승식(1993) 『음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석』
서울: 서울대학교 컴퓨터 공학과 박사 논문

김영택(1994) 『자연 언어 처리』 서울:敎學社

김영택 외(2001) 『자연 언어 처리』 서울:생능출판사

송도규(1996) 『인지언어학과 자연 언어 자동 처리』 서울:홍릉과학출판사

한국과학기술원전산학과 인공지능센터(1995) 『한국어. 중국어간기계번역시스템』 대전:한국과학기술원

조관희 편역(2000) 『영중한(英中韓) 대조 컴퓨터 용어집』 서울: 도서출판 시놀로지

배재석(2001) 「중국에서의 중국언어정보처리 연구 현황」, 『中語中文學』 29:195-215

강원석 외 6명(2000) 「중한기계 번역을 위한 형태소 분석기」,제 12회 한글 및 한국어정보 처리 학술대회(2000.10.13-10.14) 발표 논문

송영미 외 6명(2000) 「METES/CK 중한기계 번역 시스템의 구문분석 규칙」, 제12회 한글 및 한국어 정보 처리 학술대회 (2000.10.13-10.14) 발표 논문

황금하 외 7명(2000) 「뉴스 타이틀 번역을 위한 중한 기계 번역 시스템」, 제 12회 한글 및 한국어 정보 처리 학술대회(2000.10.13-10.14) 발표 논문

장민 외 3명(1999) 「중한 기계 번역기 METES/CK:파이프라인 번역」, 제11회 한글 및 한국어 정보 처리 학술대회 (2000.10.8-10.9) 발표 논문

김나리(2001) 「中韓 기계 번역--시스템 구조와 연구 과제」, 고려대 민족문화연구원 전문가 초청 강연(2001.8.30) 발표 논문

한정한(2001) 「다국어 기계번역 시스템을 위한 의미표시 방법」,『민족문화연구』35

황은하 외 2명(2002) 「동사 패턴에 기반한 韓中 기계 번역」, 고려대 민족문화연구원 전문가 초청(2002.4.12) 발표 논문

沈小喜(1998) 『漢語的節奏單位與語法結構』 北京:北京大學中文系博士論文-----(1999) 「한국인의 중국어 문장 끊어 읽기에 대한 고찰」, 『중국언어연구』 8:213-239

傅永和(1999) 『中文信息處理』 廣州:廣東敎育出版社
詹衛東(2000) 『面向中文信息處理的現代漢語短語結構規則硏究』 北京:淸華大學出版社
俞士汶 외 3명(1998) 『現代漢語語法信息詞典』 北京:淸華大學出版社
馮志偉(1992) 『中文信息處理與漢語硏究』北京:商務印書館
----(1996) 『自然語言機器飜譯新論』 北京:語文出版社
----(1999) 『應用語言學綜論』 廣州:廣東敎育出版社
----(2002) 「中國的機器飜譯硏究, 英漢機器飜譯示例, 日漢機器飜譯示例」, 고려대 민족문화연구원 전문가 초청(2002.4.12) 발표 논문

■ 中文摘要 ■

題目:漢語自動分詞--中韓機器飜譯之初步階段

隨着最近中韓兩國的頻繁交流,非常需要中-韓或韓-中機器飜譯之開發.因此本文初步探討漢語自動分詞.因爲漢語自動分詞是中文信息處理的基礎工程,也是中韓機器飜譯之初步階段.

第1章, 簡單介紹機器飜譯及其方法論. 第2章, 簡介漢語語法的語言單位. 第3.4章, 討論漢語分詞的基本規律及具體方法.該章分12個詞類, 比較詳細探討漢語分詞的方法. 第五章, 討論漢語分詞的一些問題, 如未登錄詞及重義現象. 第6章, 略述國內的中-韓或韓-中機器飜譯硏究情況. 到目前爲止,這門學科仍有待積極開發,希望以後年靑的硏究生多加入機器飜譯硏究之行列.

중심어: 기계 번역. 中-韓 . 中國語. 單語. 自動 분리