의미 분석에 기반을 둔 한글-한자 변환 시스템

o

정 일형, 이 종혁

포항 공과 대학 전자계산학과

Hangul-Hanja Translator Based on Semantic Analysis

o

Il-Hyung Jung, Jong-Hyeok Lee

Dept. of Computer Science, POSTECH

Tel : (0562) 79-2906, Fax : (0562) 79-2299

E-mail : jung@lion.posteck.ac.kr

요 약

본 논문은 한글-한자 변환에 있어서 여러 대응 한자를 갖는 동형이의어의 모호성 해소 방법을 제안한다. 기존의 변환 방법은 사용자의 개입으로 이루어지므로, 사용자에게 많은 부담을 주고 변환 효율을 떨어뜨린다. 한자 선택에 있어서 동형이의어 문제의 근본적 해결을 위해, 본 시스템에서는 의미 분석을 이용한 한글-한자 변환기를 제안한다. 이를 위해 격문법과 관련어 지식 베이스(thesaurus)를 사용한다. 격문법을 사용하여 서술어를 중심으로 관련된 격틀의 의미를 분석한다. 그리고 합성어의 경우에 합성어의 구성 형태에 따라 격문법을 사용하거나 관련어 지식 베이스에서의 의미 근접성을 사용한다. 본 논문은 이와 같이 의미 분석 및 개념 정보를 기반으로 하는 동형이의어의 모호성 해결 방안을 제시하고 이를 반영한 한글-한자 변환 시스템의 설계 및 구현에 관하여 기술한다.

I. 서론

적극적인 한글화 운동에도 불구하고 아직도 한자가 많이 사용되고 있다. 그리고 한글-한자 변환은 현재의 대부분의 워드-프로세서에서 제공되는 주요한

본 연구는 문교부 지원 한국학술진흥재단의 자유 공모 과제 학술 연구 조성비에 의하여 연구 중에 있음('92 7월 - '93 7월).

기능이다. 그러므로 사용자에게 편의를 제공하는 보다 정확하고 높은 효율의 변환기가 요구된다. 이러한 한글-한자 변환에서 가장 문제가 되는 것은 동형이의어의 모호성을 해결하는 알맞은 한자 선택이다. 이러한 동형이의어 모호성을 해결하기 위한 기존의 방법들은 모호성이 있는 단어들에 여러가지 제한을 두거나 [5,6,7,8,9] 휴리스틱 [4]을 사용하고 있다. 그러나 이러한 방법들은 의미 분석을 하지 않기 때문에 동형이의어의 문제를 근본적으로 해결하지 못한다 [1]. 본 논문에서는 의미 분석에 기반을 둔 한글-한자 변환기를 제안한다.

본 연구의 의미 분석은 두 부분에서 이루어진다. 그 중 하나는 서술어를 중심으로 이와 관련된 단어들에 대한 의미 분석이고 다른 하나는 명사 합성어에 있어서 이를 구성하는 단어들의 의미 분석이다. 한 문장안에 나타나는 각각의 서술어와 이에 관련된 단어들 간의 의미적 관계를 격문법을 사용하여 분석한다. 합성어에 있어서는 그 합성어를 이루는 단어들의 구성 형태에 따라서 격문법을 사용하거나 관련어 지식 베이스를 사용한 의미 분석이 이루어진다.

II. 기존의 한글-한자 변환 방법

사용자의 부담을 덜어 주기 위해서 국내에서 현재 사용되고 있는 대부분의 한글-한자 시스템에서는 한글의 입력후 대응하는 한자를 대치시키는 방법을 채택하고 있다 [2,3]. 한글의 음을 이용한 한자의 대치 입력은 변환되는 단위에 따라서 음절 단위, 단어 단위, 어절/문장 단위의 3가지로 나뉘어진다. 이 중에서 어절/문장 단위의 변환에서는 사용자의 개입없이 시스템이 한 어절 안에서 한자로 변환될 수 있는 부분을 자동적으로 추출하여 대응되는 한자로 변환한다. 본 시스템은 이러한 어절/문장 단위의 변환 방법을 사용하여 동형이의어의 모호성을 해결하는 방안을 제시한다.

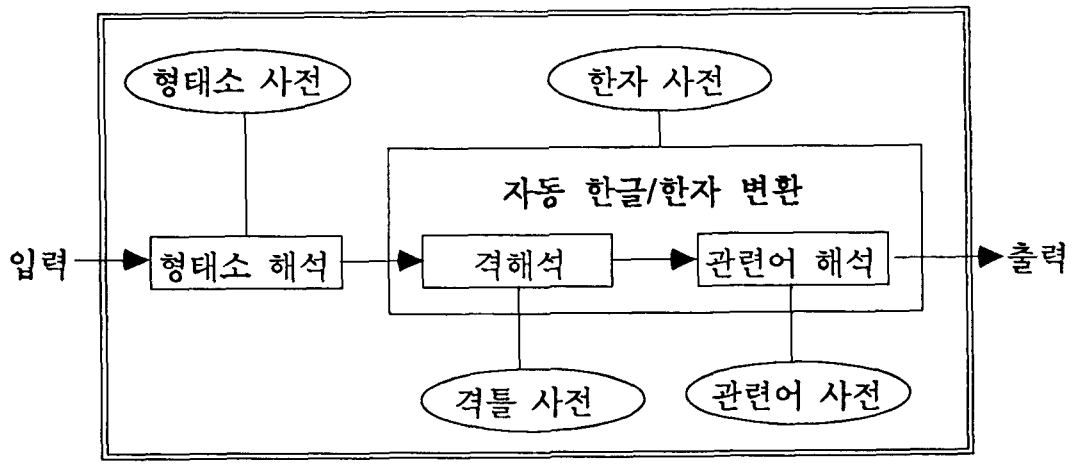
기존의 국내 한글-한자 변환에서의 수동 변환은 시간이 많이 들고 사용자의 부담이 크다. 이러한 단점을 줄이고자 일본에서는 문절(어절)/문장 단위의 변환이 가능한 자동변환 방식이 제안되었고 동형이의어 문제 해결을 위해 상용 시스템에서는 '최종 사용어 우선'과 '최다 사용어 우선'이 주로 사용되고 있다 [4]. 그러나 휴리스틱에 의존한 이러한 시스템은 동형이의어의 모호성 해결에 근본적으로 접근하지 못하므로 본 논문에서는 의미 분석에 기반을 둔 동형이의어 모호성 해결 방안을 제시한다. 일본에서도 가나-한자 변환에 격문법을 사용한 의미 분석을 제안하고 있지만 [10,11], 본 논문은 이보다 격문법을 폭 넓게 적용하고 격문법 외의 관련어 지식 베이스도 이용할 것을 제안한다.

그리고 주어진 문장의 불충분한 의미 제약으로 인하여, 본 논문에서 제안하는 의미 분석에 의하여도 모호성 해결이 이루어지지 않은 경우 이들 휴리스틱 방법들도 사용된다.

III. 본 시스템의 한글-한자 변환 방법

본 시스템은 사용자에게 효율성과 편리성을 제공하기 위해 기존의 메뉴를 이용한 수동 방식이 아닌 자동선택 방식을 사용한다. 자동선택 방식에 있어서 기존의 시스템은 자연언어의 모호성을 근본적으로 처리하지 못하는 단점이 있다. 본 시스템은 이러한 한글-한자 변환에서 발생하는 모호성, 즉 동형이의어 처리 문제와 형태소 분석에서 발생하는 모호성을 근본적으로 처리하려고 한다. 그리고 이러한 모호성의 근본적인 해결을 위해 체계적인 의미적 접근을 하려고 한다. 의미 분석 방법으로는 격문법의 적용과 관련어 사전의 이용이다.

본 시스템은 잘 정의된 격문법을 이용한 격해석과 잘 구축된 관련어 지식베이스를 사용한 관련어 해석을 통하여 한자에 대한 지식이 없는 사용자의 경우에도 동형이의어의 의미에 알맞은 한자를 결정할 때 '지능적'인 한글-한자 변환이 가능하다. 본 시스템의 논리적 모델은 [그림 1]과 같다.



[그림 1] 자동 한글-한자 변환 시스템

1. 격해석

격문법은 한 문장에서 의미 표시를 서술어와 문장내의 그와 연관된 명사

구들의 의미 역할을 설정하는 방법이다. 격문법을 구문 분석의 규칙으로 사용하는 격해석은 우리말과 같은 문맥 자유 어순인 언어에서 잘 적용될 수 있다. 또 격해석은 문법적 관계를 표시하는 조사가 발달한 첨가어에 잘 적용된다. 격해석 과정에 이용되는 본 연구의 격틀사전은 한국어에 발달한 조사를 사용하여 구성한다. 격해석을 <예문 1>에 적용하여 동형이의어의 모호성 해결의 한가지 방안으로 제시한다.

나는 꽃이 핀 공원에 갔다.

<예문 1>

1.1. 서술어를 중심으로 하는 의미 체크

<예문 1>의 동형이의어 '공원'은 公園과 工員의 두 가지 의미를 가진다. <예문 1>를 [그림 2]와 같은 격틀을 사용하여 격해석하는 과정을 살펴보자. 주어진 문장의 어절들을 왼쪽에서 오른쪽으로 검색해 나간다. 이 과정에서 검색되는 모든 명사는 스택에 저장되고 서술어가 검색되면 스택을 팝(pop)하여 그 내용과 격틀의 내용을 비교해 나간다. <예문 1>에서 처음 만나게 되는 서술어는 '핀'이다. 그래서 '피다'의 격틀를 참조하여 주격과 장소격이 필요함을 알 수 있다. 이때 격의 판정에는 조사가 이용된다. 스택에 현재 들어 있는 명사 중에 '꽃'만이 격틀의 주격으로서 적당한 의미인 '식물' 또는 '꽃'과 일치되어 선택된다. 그리고 '갔다'라는 동사를 만나서는 '가다'의 격틀을 참조하여 스택의 내용 중에서 장소격인 '공원'을 얻는다. 그리고 이때 그 의미가 장소이기 때문에 공원의 두가지 의미 즉, 公園과 工員 중에서 公園만이 적합한 것으로 선택된다. 그리고 주격은 인간의 의미를 가진 '나'가 선택된다. 그 결과로 "나는 공원에 갔다"와 "꽃이 피다"의 의미절을 얻는다.

피다		
심층격	조사	의미
AGT	이, 가, 은, 는	식물, 꽃

가다		
심층격	조사	의미
AGT	이, 가, 은, 는	인간
LOC	에	장소

[그림 2] 격틀사전의 내용

스택에 필요한 격정보가 없는 경우에는 주어진 문장을 더 검색해 나간다. 한편, 문장의 끝까지 검색해도 적합한 격정보가 없는 경우에는 이미 팝된 단어들의 절을 찾고자 하는 격정보로 한다.

병원은 환자를 치료하는 의사들로 가득했다.

<예문 2>

<예문 2>를 격해석 할 때, 서술어 '치료하는'을 만난다. 그리고 '치료하다'의 격틀을 참조하여 주어가 필요함을 알지만 스택을 참조해도 적합한 것이 없다. 그 이유는 <예문 2>가 관형절을 안은 문장이기 때문이다. 이 경우에는 예문을 계속 검색하여 주어로서 적당한 '의사'를 찾게 된다. 서술어 '가득했다'의 경우 필요한 주격의 정보가 스택에 없다. 이 경우의 주격 정보는 이미 스택에서 팝된 단어들의 절 즉, '환자를 치료하는 의사들'이다.

두 가지 이상의 스택 내용이 격틀의 내용으로 적합할 경우에는, 주어진 문장의 어떤 서술어와 관련된 단어는 그 서술어와 더욱 근접하다는 근접의 원칙을 적용하여 스택에서 찾아지는 첫번째 것으로 선택한다.

우리는 미아가 결백하다고 믿는다.

<예문 3>

<예문 3>에서 결백한 것이 '우리'인지 '미아'인지를, '결백하다'의 주격에는 사람이 온다는 格정보만으로는 알 수 없다. 그러므로 이 경우 서술어 '결백하다'에 근접한 '미아'가 결백한 것으로 판단된다. '미아'는 '우리'보다 스택의 상위에 위치한다. 서술어 '믿는다'의 목적격은 '미아가 결백하다'이다.

1.2. 명사 합성어 처리

한자 복합어를 품사의 구성 형태에 따라 따라 [표 1]과 같이 나눌 수 있다. [표 1]에서 품사가 동사로 구분된 것은 '하다' 동사이다. 즉, '하다'가 붙어 동사가 될 수 있는 명사이다. [표 1] 중에서 B의 경우가 격해석을 적

용할 만하다. "논리정연'은 주어-서술어 관계, '공기오염'은 목적어-서술어의 관계가 있다. '공기오염'을 격해석하여 동형이의어 공기의 여러 의미 즉, 공기(놀이), 空氣 (자연), 空器 (빈 그릇) 중에서 오염시킬 수 있는 공기 (空氣)만이 선택된다. 한편 C의 경우에 '번역기계'처럼 격해석을 적용할 수 있는 경우는 상당히 드물기 때문에 격해석에서 제외된다. A의 경우는 III 장 2.1 절에서 설명되는 관련어 지식 베이스를 사용하여 해결한다.

품사	예
A. 명사 + 명사	과학지식, 전화번호
B. 명사 + 동사	논리정연, 공기오염
C. 동사 + 명사	폭행사건, 번역방식
D. 동사 + 동사	투신자살, 시정명령
E. 형용사 + 명사	기기묘묘

[표 1] 명사 합성어의 구성

2. 관련어 해석

격해석 과정에서 불충분한 격정보로 인하여 동형이의어의 모호성이 완전히 제거되지 못하는 경우가 있다. 본 논문에서는 관련어 지식 베이스를 제안하고 그 구성 방법과 이를 동형이의어의 모호성 해결에 적용하는 방법에 관하여 논한다.

관련어 지식 베이스 (thesaurus)는 개념적으로 연관성이 있는 단어들을 모아 놓은 사전이다. 관련어 지식 베이스에는 단어들이 다양한 의미 관계를 가지고 무리지어 있다. 이들 단어들은 전체적으로는 상위-하위 관계를 이루면서 부분적으로는 전체-부분 관계뿐만 아니라 동의 관계, 이의 관계, 유의 관계, 반의 관계, 환유 관계 등을 가진다. 그리고 관계는 있으되 그 관계를 정의할 수 없는 단어들까지도 체계적으로 구성되어져야 한다.

[그림 3]은 관련어 지식 베이스의 전체적인 구성도이다. 이와 같은 관련어 지식 베이스는 격해석에서 해결하지 못한 명사 합성어의 동형이의어 모호성 해결을 위해 사용된다. 동형이의어에 해당하는 각각의 의미의 단어들은 다른 그룹에 속해 있으며, 이들과 합성어를 형성하는 다른 단어와의

의미적 근접성이 계산될 수 있다. 의미적 근접성 계산의 기본 전제는, 관련어 지식 베이스에서 단어들이 더 근접하게 무리지어 있을수록 그들 단어들은 의미적으로 더욱 가깝다는 사실이다.

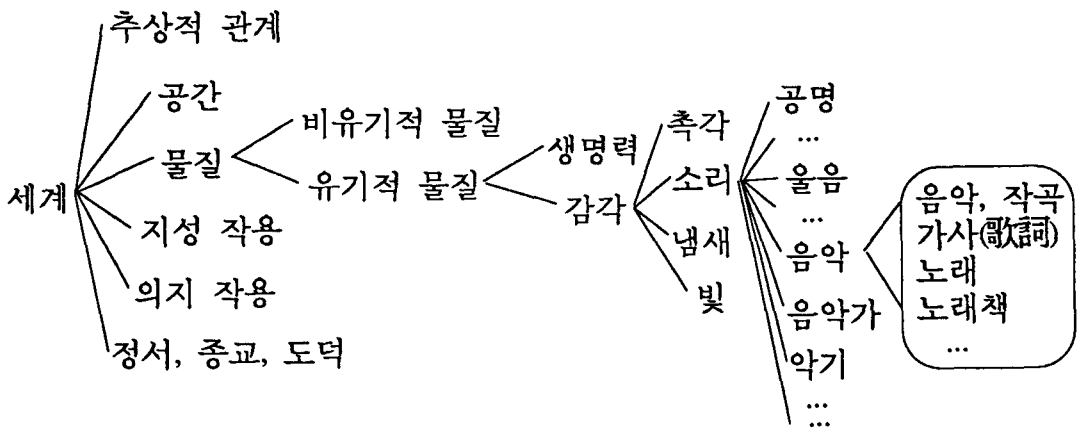
2.1 명사 합성어 처리

[표 1]에서 타입 A (명사 + 명사)의 명사 합성어의 동형이의어 모호성을 [그림 3]과 같은 관련어 지식 베이스를 사용하여 해결할 수 있다.

그 노래 가사는 아름답다.

<예문 4>

<예문 4>에는 '가사'라는 단어가 있고 이의 의미에는 家事, 袈裟, 歌詞, 歌辭의 4가지가 있다. <예문 4>와 같은 문장은 격해석을 통해서 동형이의어 '가사'의 의미적 중의성을 해결하지 못한다. 서술어 '아름답다'가 歌詞, 歌辭, 그리고 袈裟를 구분할 수 있을 정도의 강한 의미적 제약을 줄 수 없기 때문이다. 이러한 경우에 [그림 3]과 같은 관련어 지식 베이스를 참조하여 가사를 歌詞로 결정할 수 있다. 노래와 가사(歌詞)가 음악이라는 클래스에 함께 무리지어 있어서 의미적으로 歌辭나 袈裟 보다 더욱 근접하다고 판단하여 歌詞라는 의미의 '가사'를 선택하게 된다. 즉, 개념적 연관성이 가장 높은 단어를 선택한다.



[그림 3] 관련어 지식 베이스

자연언어를 처리하려면 의미 분석 단계가 반드시 필요하게 되며 이는 자연언어에 원초적으로 내재한 모호성의 해결을 위함이다. 본 논문에서는 관련어 지식 베이스를 제안하여 여러 단계에 걸쳐 일어날 수 있는 모호성을 해결하는 지식 기반으로 삼고자 한다.

2.2 서술격 조사 '이다' 처리

정의를 내리는 문장에서 처럼 'A는 어떠한 B이다.'로 표현되는 문장의 B와 A가 대개 상위-하위 관계에 있다. 즉 B는 A의 일반화 된 개념이다.

회의는, 여러 사람이 한 장소에 모여서 일정한 규칙에 따라
의견을 말함으로써 문제를 해결해 나가는 모임이다.

<예문 5>

<예문 5>의 서술격 조사 '이다'와 관련되어 "회의는 어떠한 모임이다."를 얻을 수 있다. 즉 '회의'와 '모임'은 주어-보어 관계에 있고 '회의'는 '모임'의 구체화된 개념이다. 그러므로 이들 두 단어의 상위-하위 관계를 관련어 지식 베이스에서 참조하여 여러 의미의 회의 (會議, 懷疑, 會意) 중에서 會議가 선택될 수 있다.

IV. 결론

본 논문은 의미 분석에 기반을 둔 한글-한자 변환의 동형이의어 모호성 해결 방안을 제시한다. 본 한글-한자 시스템은 격해석과 관련어 해석에서 의미 해석이 이루어진다. 격해석은 격틀 사전을, 관련어 해석은 관련어 지식 베이스를 참조한다. 격해석은 서술어를 중심으로 이와 관련된 단어들과 명사 합성어의 일부에서 이루어진다. 그리고 관련어 해석은 격해석 후에도 해결되지 못한 명사 합성어의 동형이의어 모호성을 해결한다. 본 논문의 한글-한자 변환기의 높은 성능을 위해서는 격틀사전과 관련어 지식 베이스가 충분히 잘 구성되어야 한다.

참고 문헌

[1] 정일형, 이종혁, "지능형 한글-한자 변환 시스템," 정보과학회지, 제 19권 1호, pp.655-658, 1992.

[2] 진민, "기계 사전과 어절 분석을 이용한 한글 한자 변환 시스템," 석사학위논문, 한국과학기술원, 1984.

[3] 정병수, "한글 한자 변환 시스템에서의 기계사전의 구성과 처리 방법," 석사학위논문, 한국과학기술원, 1985.

[4] "かな漢字變換方式の實現手法を探る," NIKKEI ELECTRONICS, 1983, 8.pp.180-195. (일어판)

[5] 荒木健治, ? 内香次, 永田邦一, "多段階分割法によるべた書き日本語文のかな漢字變換", 情報處理學會論文誌, Apr. 1987, pp.412-421. (일어판)

[6] 本間茂, 山階正樹, 小橋史?, "連語解析を用いたべた書きかな漢字變換," 情報處理學會論文誌, Nov. 1986, pp.1062-1067. (일어판)

[7] ? 内香次, 伊藤太?, 鈴木康?, "前後連接文字を利用した同音語選擇機能を有するかな漢字變換システム," 情報處理學會論文誌, Mar. 1986, pp.313-320. (일어판)

[8] Masaki YAMASHINA and Fumihiko OBASHI, "Kana-to-Kanji Translation Based on Collocational Analysis for Non-Segmented Input", REVIEW of the Electrical Communications Laboratories, Jan. 1989, pp.65-70.

[9] 김경서, 김대철, 정강석, 송만석, "말뭉치를 이용한 형태소 분석 단계에서의 중의성 해결에 관한 연구," 인간과 기계와 언어, 제 3 회 한글 및 한국어 정보처리 학술발표논문집, pp.36-43.

[10] M. Abe and Y. Ooshima, "A Kana-Kanji Translation System for Non-Segmented Input Sentences based on Syntactic and Semantic Analysis," Proc. of COLING '86, pp.280-285 (1986).

[11] H. Makino and M.Kizawa, "An Automatic Translation System of Non-Segmented Kana Sentences into Kana-Kanji Sentences and its Homonym Analysis," Transactions of Information Processing Society of Japan, Vol.22, No.1, pp.59-67 (1981) (일어판)