한국어 중의성 사전 구성을 위한 기초 자료 연구

김 수 정 (서울대 어학연구소)

<차례>

Ⅰ. 들어가며
Ⅱ. 정보처리에 있어서 한국어 중의성의 특징
1. 중의성의 개념
2. 중의성의 유형
Ⅲ. NLP에서의 중의성 처리
1. NLP에서의 어절 중의성의 유형
2. 어절 중의성 처리의 문맥 정보
Ⅳ. 맺으며

Ⅰ. 들어가며

한국어 정보처리란 국어 문법이나 언어학 이론을 기반으로 하여 컴퓨터에서 한국어를 분석하기 위한 방법론 및 문법 체계의 정립 등을 연구하는 것이다. 인간이나 동물들이 의사소통을 위하여 사용하는 언어와 같이 자연발생적으로 생성된 언어를 '자연언어'(natural language)라 할 때, 인간이 인위적으로 만들어 낸 언어를 '인공언어'(artificial language)라고 한다. 인공언어의 대표적인 것으로 에스페란토어와 프로그래밍 언어가 있는데, 에스페란토어는 자연언어의 다양성으로 인하여 발생하는 정보 전달 문제를 해결하고자 국제 공용언어를 목적으로 고안되었다. 프로그래밍 언어에는 기계어(machine language)를 비롯하여 어셈블리어와 고급언어가 있으며 컴퓨터 소프트웨어를 제작하는데 사용한다.

자연언어처리(natural language processing)는 컴퓨터를 이용하여 자연언어를 분석하거나 생성하는 것으로 1950년대 기계번역에 관한 연구로부터 시작되었다. 형태소 분석과 구문분석, 의미 분석 등 원시언어의 분석은 필수적인 요소이며, 자연언어 처리의 핵심적인 기반 기술로 요구되고 있다. 자연언어처리 분야는 크게 기반 기술과 응용 기술 분야로 구성되는데 자연언어 처리의 핵심 부분인 기반 기술은 다시 전자사전(electronic dictionary) 구축, 자연언어 분석(natural language analysis), 자연언어 생성(natural language generation) 분야로 나뉜다.

이중에서 자연언어 분석 기술은 자연언어 처리의 핵심적인 부분으로 자연언어의 문법적 특성에 따라 단계별로 형태소 분석(lexical analysis), 구문 분석(syntactic analysis), 의미 분석(semantic analysis), 담화분석(discourse analysis), 중의성 해결(ambiguity resolution) 등으로 구분된다.

자연언어처리 기술은 기반 기술에 해당하므로 응용분야를 전제로 한다. 자연언어처리의 응용기술이란 기계번역, 정보검색 시스템, 맞춤법 검사, 자연언어 인터페이스 등이 있다.

자연언어의 분석 기법은 문법 기술 방법에 따라 다양한 문법 체계가 제안되었다. 확장 구구조 문법(APSG: Augmented Phase Structure Grammar), 일반 구구조 문법(GPSG: Generalized Phrase Structure Grammar), 어휘 기능 문법(LFG: Lexical Functional Grammar), 중심어 주도 구구조 문법(HPSG: Head-driven Phrase Structure Grammar), 의존 문법(Dependency Grammar), TAG(Tree Adjoining Grammar) 등이다.

자연언어 분석 기술의 가장 기초가 되는 분야는 바로 형태소 분석(lexical analysis)인데, 형태소 분석의 앞에 선행되어야 할 처리가 바로 중의성 해결(ambiguity resolution)이다.[1)]

1) 형태소 분석을 한다해도 중의성이 발생되는 어휘의 경우에 전자 사전(lexicon)에 등재되어 있는 우선 순위(priority)에 의해 결정되므로 형태소 분석 오류의 대부분을 이룬다.

한국어 형태소 분석기에 대한 노력은 1990년 이후 부단히 있어왔다. 그럼에도 불구하고 형태소 분석기의 개선점은 다음과 같다. 1) 분석실패나 과분석 같은 분석오류를 줄이고 정확성을 높여야 한다는 점 2) 분석의 효율성을 높여야 한다는 점 3) 다양한 응용분야에 적용될 수 있도록 설계되어야 한다는 점 4) 분석결과의 모호성이 해결되어야 한다는 점이다.[2] 이에 휴리스틱 정보 추출을 위한 연구의 대표적인 것이 바로 중의성 처리의 문제이다.

자연어의 의미는 주로 문맥에 의존한다. 자연언어 표현 자체는 많은 중의성을 포함하는데 일반적으로 자주 나타나는 중의성의 문제들로 단어 의미의 모호성과 수식 범위의 모호성을 크게 나눌 수 있다. 단어 의미의 모호성은 많은 단어들이 비록 하나의 문법 범주를 가졌다고 하더라도 의미적으로 모호한 경우가 많이 있다. 예를 들어 'go'라는 동사는 대부분의 사전에서 약 40여 개의 서로 다른 의미를 가지고 있다. 또 수식 범위의 모호성에서도 중의성이 발생하는데 구문적으로 모호하지 않은 문장에 대해서도 어떤 한정사의 수식 범위가 모호한 경우가 많다. 예를 들면, "Every boy loves a dog"와 같은 문장의 의미는 한정사인 'every'의 수식 범위에 따라서 두 가지로 해석될 수 있다.[3]

한국어는 교착어라는 특성으로 단어 의미의 모호성의 유형이 너무나 다양하다. 본고에서는 어절 중의성을 논의의 대상으로 삼는다.

Ⅱ. 정보처리에 있어서 한국어 중의성의 특징

1. 중의성의 개념

---

2) 임해창 외(1996) , "한국어처리의 고찰", 『계산의미론과 그 응용』, p. 192.

3) ① 모든 소년이 사랑하는 어느 특정한 한 마리의 개가 있다(every가 boy만 수식하는 경우)
② 각 소년들이 사랑하는 (여러 마리의) 개가 있다.(Every가 문장 전체를 수식하는 경우)

중의성(ambiguity)이란 자연언어의 커다란 특징 중의 하나로, 하나의 단어 즉 어절이 여러의미로 해석될 수 있음을 말한다. 자연어 처리과정에서 발생하는 중의성은 크게 구조적 중의성(structural ambiguity)과 어휘적 중의성(lexical ambiguity)으로 나뉜다. 구조적 중의성이란 규칙에 따라 문장의 구조를 해석할 때 두 가지 이상의 구조로 해석되는 문장을 말하고, 어휘적 중의성이란 한 단어가 두 가지 이상의 형태소로 분석되는 현상을 뜻한다. 어휘적 중의성은 다시 구문적 어휘 중의성(syntactic ambiguity)과 의미적 어휘 중의성(semantic lexical ambiguity)으로 나뉜다. 구문적 어휘 중의성은 주로 어절에 나타나는 단어가 여러 품사를 가질 수 있어서 발생하는 중의성이다.

fish n. 물고기
v. 낚시하다

의미적 어휘 중의성은 어절에 나타나는 어휘가 몇 가지 의미를 가질 수 있음으로 해서 발생하는 중의성이다.

배 n. pear
n. ship
n. stomach

구조적 중의성은 한 문장 또는 구에 대한 구문적 해석이 여러 가지 존재함으로써 발생하는 중의성을 뜻한다.[4)]

한국어는 조사와 어미의 기능에 의해 문장성분이 결정되고 불규칙 활용 및 축약 현상이 발생하는 첨가어에 속한다. 이러한 다양한 중의성의 형태는 한국어 특성 중의 하나인 교착어적인 특성, 즉 조사, 어미의 발달에 기인한다.

자연언어처리는 컴퓨터를 기반으로 하고 있기 때문에, 한국어 처리를

4) "Every boy loves a dog"에 대한 두 가지 해석이 그것이다.

위해서는 형태소를 원형으로 복원하고 어절의 형태소 결합을 생성해 내는 형태소 분석 단계가 반드시 필요하다. 한국어 화자의 직관적으로 파악하고 있는 품사 구별도 컴퓨터는 단어의 형태를 일정한 문법(syntax)으로 만들어 할당해 주지 않으면 전혀 알 수 없다. 형태소 결합의 차이, 다품사의 존재, 불규칙 활용 및 축약현상은 중의성 발생의 원인이 된다. 궁극적인 형태소 분석의 효율성과 정확성을 위해서는 형태소 분석에서 중의성 처리 단계가 설정되어야 하며, 이에 관해서는 3장에서 논의하고자 한다.

2. 중의성의 유형

자연어처리(NLP)에서 다룰 수 있는 중의성의 종류는 기준에 따라 다음과 같이 나뉜다. 1980년대에 한국어 형태소 분석의 논의는 주로 형태소 분리 문제와 형태소끼리의 결합제약, 단어의 검색방향이 관심의 대상이 되었다. 그후 1990년대에 이르러 자동색인이나 철자 검사, 기계번역, 사전 편찬 등 한국어처리와 관련된 응용분야에서 실용적인 형태소 분석 시스템이 요구되었다. 이 시기에 제안된 형태소 분석을 효율화하는 연구로는 휴리스틱을 이용한 방법과 기분석 사전을 이용한 방법론으로 분류할 수 있다.

휴리스틱을 이용한 형태소 분석 방법의 대표적인 것이 음절단위 분석법이다. 단어의 처리단위를 음절단위(syllable unit)로 삼음으로써 한글의 음절 특성을 이용하여 분석결과를 추론하는 알고리즘을 이용하여 형태소 분석을 한다. 이렇게 되면, ① 조사, 어미 분리 ② 선어말 분리 및 불규칙 원형 복원 ③ 접미사 분리 및 보조용언 처리 ④ 기분석 어절, 준말, 미등록어 처리 등의 문제들이 발생한다.

이러한 형태소 분석에서 나타날 수 있는 한국어 중의성의 유형은 다음과 같다.

<자연어처리 한국어 중의성 처리의 유형>

㈀ 어절 중의성 - ㉠ 의미론적 어휘 중의성
ⓐ 명사/명사
ⓑ 동사/동사
㉡ 구문적 어휘 중의성
ⓐ 명사/ 관형사
ⓑ 부사/명사
ⓒ 조사/명사
ⓓ 동사/형용사
㉢ Affix 결합 중의성
ⓐ 불규칙 중의성
ⓑ 축약형 중의성[5]
ⓒ 신조어 중의성
ⓓ 접사 결합형 중의성

㈁ 구문적 중의성

1) 어절 중의성 유형

(1) **의미론적 어휘 중의성**(semantic lexical ambiguity)[6]

의미론적 어휘 중의성은 어절에 나타나는 어휘가 몇 가지 의미를 가질 수 있음으로 해서 발생하는 중의성이다. 문맥 내에서 단어의 올바른 의미를 선택하는 작업에 단어와 문맥에 대한 다양한 지식을 필요로 한다. 다시 말해 동음이의어(Homonymy)라 불리운다. 즉, 관련이 있는 의미들(meaning)로 구성된 의미(sense) 집합으로 형태는 같으나 뜻이 다른 단어를 의미한다. 해당하는 의미 분류에 따라 동음이의어(Homonymy) 수준의 분류와 다의어(Polysemy) 수준의 분류로 나뉜다.[7]

---

5) 요즘 통신언어의 발달로 언어의 축약형과 신조어의 발생은 나날이 급증하고 있다. 자연어 처리가 컴퓨터를 기반으로 한 자연언어의 인터페이스를 목표로 할 때, 이와 같은 언어 현상은 무시할 수 없는 것이다.

6) 다음은 연세 말뭉치를 가공하여 만든 3만 단어 사전 중에서 상위 빈도 1000개 중에서 뽑은 것이다.

7) (1) 동음이의어(Homonymy) 수준의 분류
① MH(Mono-Homonymy) : 한 단어 유형에 대해 하나의 동음이의어만을

㈎ 체언 중의성

① 명사 중의성[8)]

중의어	고려	고리	고무	고수	고름	성적	굴	벌	눈
의미	考慮	高利	鼓舞	高手	옷고름	成績	窟	罰	eye
의미	高麗	고리	고무	固守	fus	性的	oyster	bee	snow

② 명사/의존명사 중의성

중의어	개	일	해	병	줄	수
의미	個	日	sun	病	rope	數
의미	dog	一	year	bottle	probability	首
의미		work				probability

㈏ 용언 중의성

① 타동사/타동사 중의성

중의어	꾸다	대다	물다	빌다	빨다
의미	꾸다(dream)	대다(touch)	물다(bite)	빌다	빨다(suck)
의미	꾸다(borrow)	대다(stop)	물다(repay)	빌다(borrow)	빨다(wash)

② 타동사/ 자동사 중의성

갖는 경우로, 단어에 대한 표제어 수와 단어 유형의 수가 1개로 일치한다.
② PH(Poly-Homonymy) : 한 단어 유형에 대해 둘 이상의 동음이의어를 갖는 경우로, 단어에 대한 단어 유형은 1개, 표제어의 수는 2개 이상이다.
⑵ 다의어(Polysemy) 수준에 따른 분류
① MS(Mono-Sense) : 하나의 표제어가 하나의 의미만 갖는다.
② PS(Poly-Semous) : 하나의 표제어가 둘 이상의 의미를 갖는다.

8) 자연어처리에서는 장단음의 구별로 의미를 변별하기 어렵다.

중의어	개다	일다	바르다	고르다	지다
Vt	clear	clean out	paste	choose	carry on back
Vt	fold				
Vi	soften	happen	be straight	be even	fall
Vi					get defeated

③ 동사/ 형용사 중의성

중의어	쓰다	달다	세다	싸다	적다	짜다
V	write	hang	count	wrap	write	weave
V	use		become white	pus		squeeze
V	wear					
A	bitter	sweet	strong	cheap	small	salty

(2) 구문적 어휘 중의성

품사 중의성(parts of speech ambiguity)이라고도 불리우는 형태로 형태소 분석 결과에서 어떤 형태소의 품사가 두 가지 이상이 가능한 경우이다. 이 중의성은 그 적용 범위가 너무 넓기 때문에 두 분석 결과의 형태소 개수 및 각각의 형태소가 동일하고 단지 어휘형태소의 품사만 다른 것을 '좁은 의미의 품사 중의성'이라고 정의한다.

(가) 어근(stem)의 구문적 어휘 중의성

Stem 중의성	객관적	가다	곳	차다	적다	들
품사	명사	자동사	명사	동사	동사	명사
품사	관형사	보조용언	의존명사	형용사	형용사	의존명사

(나) 조사 · 어미(Affix)의 구문적 어휘 중의성

Affix 중의성	은	을	는	어	니
조사	절대격 조사	목적격 조사	절대격 조사		
어미	관형형 어미	관형형 어미	관형형 어미	종결어미	종결어미
				연결어미	시제 선어말어미
					연결어미

(다) 어근 · 조사 · 어미의 구문적 어휘 중의성

중의성	이	저	가	세
체언	지시 대명사	지시 대명사	명사	수관형사
	의존 명사	대명사	주격조사	단위성 의존명사
	주격 조사			
관계언			종결어미	종결어미

(3) Affix 결합형 중의성

Affix 결합형 중의성은 일종의 형태소 분리 중의성(morpheme segmentation ambiguity)이다. Affix는 조사 · 어미로 간주하는 바, 한국어의 교착어적 특성으로 인해 무수히 많은 Affix가 존재한다. 용언이나 체언에 결합하여 중의성을 발생시키는 일음절 Affix들은 다음과 같다.

Affix		Affix		Affix	
가	조사	다	어미	러	어미
게	어미	담	어미	려	어미
고	어미	데	어미	렴	어미
과	조사	도	조사	로	조사
군	어미	라	어미	를	조사
나	어미	람	어미	리	어미
네	어미	랑	조사	마	어미
는	조사/어미	래	어미	만	조사
니	어미	랴	어미	매	어미

Affix		Affix	
면	어미	서	어미
야	조사/어미	은	조사/어미
어	어미	을	조사/어미
에	조사	의	어미
어	어미	이	조사
에	조사	라	어미
오	어미	자	어미
와	조사	지	어미
요	어미	아	조사/어미

(가) 불규칙 용언 중의성

불규칙 용언 중의성은 용언의 어미 활용에서 기인된다. 불규칙 용언 중의성은 규칙용언과 'ㄷ', 'ㄹ' 불규칙 용언이 일정한 규칙을 이루며 중의성을 발생시킨다. 이때 affix인 'ㄹ', 'ㄴ', 'ㅁ', '고', '로', '도' 과 결합할 때에는 중의성을 발생시키는 명사가 합세하여 혼재되는 양상을 보이기도 한다.

① 규칙/ 'ㄹ'탈락/ 명사

중의어	사는	살	산	사고	사면	사서
분석 (Vt)	사(buy) +는	사(buy) + ㄹ	사(buy) +ㄴ	사(buy) +고	사(buy)+면	사다(buy) +어서
분석 (Vi)	살(live) +는	살(live) + ㄹ	살(live) + 은			
분석 (N)	사(死)+는					
	사(4)+는			사(4)+고	사(4)+면	
분석 (N)		flesh	mountain	accident	amnesty	a librarian
		age		thinking	the four sides	

② 규칙/ 'ㄹ'탈락/ 'ㄹ'탈락/ 명사

중의어	가는	간	갈	가니	감
분석 (Vi)	가(go) +는	가(go) +ㄴ	가(go) +ㄹ	가(go) +니	가(go) +ㅁ
분석 (Vt)	갈(grind) + 는	갈(grind) +ㄴ	갈(grind) +ㄹ	갈(grind) +니	
분석 (A)	가늘(thin) +은				
분석 (N)		肝			a persimmon

③ 규칙/ 'ㄹ'탈락/ 'ㄷ' 불규칙/ 명사

중의어	기는	길	길어서	기고
분석 (Vt)	기(crawl) +는	기(crawl) + ㄹ	긷(pump) +어서	기(crawl) +고
분석 (A)		길(long) + ㄹ	길(long) +어서	
분석 (N)	기(flag) +는	길(road)		기고(a draft)

중의어	들어서	들을	들
분석 (Vt)	듣(hear) +어서	듣(hear) +어서	
분석 (Vt)	들(lift) +어서		들(lift) + ㄹ
분석 (N)		들(field) +을	들(field)

(나) 축약형 중의성

자연어 처리는 기본적으로 컴퓨터를 기반으로 하고 있으며, 궁극적으로는 on-line상에서 생성되는 언어들을 처리하는 것을 목적으로 하고 있다. 따라서 on-line상에서 구현되는 언어-특히, 구어체의 어휘들을 처리하는 것은 기본적인 자연어 처리의 수준을 뛰어넘는 것이 된다. 이렇게 구어에서 쓰이는 어휘에서 발생하는 축약형 중의성은 다음과 같다.

중의어	날	난	절	전	건
분석 (Vt)	날(fly) +ㄹ		절(limp) +ㄹ	절(limp) +은	걸(hang) +는
분석 (Vi)	나(happen) +ㄹ	나(happen) +ㄹ			
분석 (N)	날(a day)	난(orchid)	절(a temple)	전(food)	건(a case)
분석 (N)	날(a blade)		절(a clause)	전(before)	
분석 (축약형 N)	나(me) +를	나(me) +는	저(me) +를	저(me) +는	것(thing) +은

(다) 신조어 중의성

중의어	걸	걸을	깔	보이는
분석 (Vt)	걸(hang) +ㄹ	걸(hang) +을	깔(spread) +ㄹ	보이(let see) +는
분석 (Vt)		걷(walk) +을	까(peel) +ㄹ	
분석 (N)	것(thing) +을			
신조어 (N)	걸(girl)	걸(girl) +을	깔(lover)	보이(boy) +는

(라) 접사 결합형 중의성

피동 및 사동 접사 '이'는 서술격 조사 결합형[9]과 함께 중의성을 발생시킨다. 예를 들어, '죽이다'의 경우, '죽+이다'와 '죽이+다'로 2가지의 형태소 분석 가능성을 나타낸다. 이러한 유형들은 다음과 같다.

중의어	죽이다	죽이고	죽이니	죽이면
분석 (N)	죽(porridge) +이다	죽(porridge) +이고	죽(porridge) +이니	죽(porridge) +이면
분석 (V)	죽이(kill) +다	죽이(kill) +고	죽이(kill) +니	죽이(kill) +면

중의어	베이다	베이고	베이니	베이면
분석 (N)	베(cloth) +이다	베(cloth) +이고	베(cloth) +이니	베(cloth) +이면
분석 (V)	베이(get cut) +다	베이(get cut) + 고	베이(get cut) + 니	베이(get cut) +면

9) 서술격 조사는 조사 중에서 활용의 특징을 가지고 있어 사동·피동사의 활용형과 매우 유사성을 띤다.

중의어	물리다	물리면	물리니	물리면
분석 (N)	물리(physics) +이다	물리(physics) +이고	물리(physics) +이니	물리(physics) +이면
분석 (V)	물리(get bitten) + 다	물리(get bitten) +고	물리(get bitten) +니	물리(get bitten) +면

Ⅲ. NLP에서의 중의성 처리

1. 어절 중의성과 처리 방법

중의성 해결에 관한 연구는 다음과 같이 구분할 수 있다. 확률정보를 이용하는 통계적인 접근방법과 규칙기반의 접근방법으로 나눌 수 있다.[10)]

통계적인 접근 방법은 사람들이 실제로 사용하는 많은 데이터에서 확률정보 및 통계정보를 추출하고, 이를 이용하여 중의성을 해결하는 방법이다. 태깅된 코퍼스로부터 품사 태그들 간의 확률정보(probability information)을 추출하고, 이를 이용하여 중의성을 해결하는 방법이 그것이다.

한국어 중의성 해결의 가장 기본적인 수준은 어절 중의성의 해결인데, 이는 어절을 단위로 삼아 발생하는 중의성을 해결하는 시스템으로 교착어로서의 특성 때문에 발생하는 형태소 분석의 복잡성을 피할 수 있는 형태의 해결방식이다. 어절 단위의 중의성 사전을 구축할 경우, 해당 어절에 대한 분석 정확도는 상당히 높은 편이다.

규칙기반의 방법은 중의성 해결에 사용될 수 있는 규칙을 추출하여 그 규칙에 따라 중의성을 해결하는 방법을 말한다. 품사가 발생하는 조건을 규칙(context frame rule)으로 기술하고, 이를 응용하여 중의성을 해결하는 방법이다. 주로 영어에 적용된 방법들로, 이 방법은 규칙추출

10) 최기선 외(1993), 『계산의미론과 그 응용』

이 어렵다는 문제점을 가지고 있다.

이렇게 규칙기반 방법과 통계적 기법에 기반한 방법의 통합이 중의성 해결에 도움을 줄 수 있는 가능성이 높은 편이다. 한국어에서 태깅(tagging)은 형태소 분석결과의 중의성을 제거하는 방법으로서 확률을 이용하는 방법, 퍼지망을 이용하는 방법, 신경망을 이용하는 방법 등 다양한 방법들이 동원되고 있다.

2. 어절 중의성 처리의 문맥 정보

이렇게 다양한 유형의 어절 중의성의 처리에 있어서 언어학적으로 유용한 도구가 전후 어휘 문맥을 통하여 중의성을 해소하는 것이다. '산'라는 어휘를 보면, '산[mountain]'이라는 의미와 '산[live]', '산[buy]' 등의 의미를 가진다. 만일 '어제 산 물건'이라는 구문을 해석할 때 비논리적인 논리 형태로 해석할 가능성은 분명히 존재한다. 이와 같은 비논리적 해석을 제거하기 위해서는 기본적인 배경 지식(Schema)이 필요하다. 즉, 단어 의미들간에 가능한 조합에 대한 제약 규칙을 가지고 있어야 한다. 예를 들어 '물건'이라는 단어가 취할 수 있는 논항(argument)이 있을 것이며, 전후 문맥 정보를 통해 발생하는 중의성을 해소할 수 있다.

1) 품사 정보를 이용한 중의성 해소

품사 정보를 이용한 중의성 해소에 따른 분류로 GD(Guaranteed Disambiguation),PD(Possible Disambiguation), ND(No Disambiguation)로 앞에서 논한 바가 있다. GD(Guaranteed Disambiguation) 의 대표적인 예로는 '익다'가 있는데, 이것의 표는 다음과 같다.

중의어	익다	익다
의미	get used to	get cooked
품사	A	V

중의어 '익다'는 해당 중의어 앞의 어휘 문맥 정보를 통해 중의성을 해결할 수 있다.

중의어	의미	품사	앞 정보		앞 정보 자질 부여 (품사 및 의미)		
익다	get used to	A	낯이	손에	[주격]	[처소격]	[body]
익다	get cooked	V	음식이	열매가	[주격]		[food]

PD(Possible Disambiguation)는 어느 품사로 태깅되는가에 따라 의미 중의성 해소가 되기도 하고 중의성 해소를 할 수 없기도 하다. 대표적인 예로 '쓰다'가 있는데, 타동사 '쓰다'에 대한 표는 다음과 같다.

기본형	품사	동작성 여부	의미	보어(complement)	의미자질
쓰다	Vt	+	write	+	[letter]
쓰다	Vt	+	wear	+	[clothes]
쓰다	Vt	+	use	+	[tool][human]
쓰다	A		bitter		[taste]

PD유형에 해당하는 어절들은 해당 중의어 자체의 의미자질 부여가 우선 되어야 한다. '맛이 쓰다'와 다른 타동사인 '쓰다'는 1차적으로 해결될 수 있지만, 동일한 타동사인 '쓰다'를 해결하기 위해서는 의미자질을 부여해야 한다. 이것은 전자사전(lexicon)에 품사자질 외에도 의미자질을 부여해야 한다는 뜻이다.

ND(No Disambiguation) 유형에 속하는 단어는 품사만으로 의미 중의성 해소를 할 수 없으며 전후 문맥을 이용하여 중의성 해소의 방법을 찾아야 한다.

2) 어휘 문맥 정보를 이용한 중의성 제거

NLP에서 중의성 어절이 들어오면 일차적으로 앞 어절과 뒤 어절의

문맥 정보를 참고하게 된다. 한국어의 조사·어미(Affix) 결합형 중의성의 경우에는 GD유형 보다는 GD유형과 PD유형이 혼재된 경우가 많이 나타나게 된다. 따라서 1차적으로는 품사 정보를 통해 중의성을 해소하고, 그 다음 PD유형은 어휘 문맥 정보를 통해 해결해야 한다. 어휘 문맥 정보를 이용할 때의 고려해야 할 사항은 다음과 같다.

1) 앞 어절이 NULL일 경우
2) 연속된 중의성 어절이 나타날 경우
3) 뒤 어절이 NULL일 경우

해당 중의성 어절의 전후 문맥 정보를 얻을 경우 우선 선행되어야 할 작업은 의미자질의 선정이다. 의미자질(semantic features)은 의미망(word-net)을 참조로 하여 거의 모든 품사에 두루 적용될 수 있는 것으로 선정해야 한다. 중의성 어절 처리는 기존의 사전(lexicon)과 의미자질 선정을 전제로 한다. 따라서, 문맥 정보와 품사 정보를 이용하여 형태소 분석 과정에서 이루어지는 단계인 것이다.

앞 어절이나 뒤 어절이 NULL인 경우에는 코퍼스의 용례 중 빈도수가 높은 것으로 처리할 수밖에 없다. 또한 연속된 중의성 어절이 나타날 경우에도 마찬가지이다.

어휘 문맥 정보를 이용한 중의성 제거 유형은 다음과 같다.[11]

(1) 앞의 어절을 보아야 하는 유형

중의성	품사정보	문맥 정보	
		전	후
쓰니 [WRITE]	[Vt]	[N][LETTER]	
쓰니 [WEAR]	[Vt]	[N][CLOTHES]	
쓰니 [USE]	[Vt]	[N][CONCRETE]	
쓰니 [TASTE]	[A]	[N][SOMETHING-TO-EAT]	

11) KAIST 용례 추출기 (www.csfive.kaist.ac.kr/kcp/)를 이용하였다.

(2) 앞 어절과 뒤 어절을 모두 보아야 하는 유형

중의성	품사정보	문맥 정보	
		전	후
세	단위성 명사	[N][NUMBER]	
세	관형사		단위성 명사
세	형용사	주어	

Ⅳ. 맺으며

이상에서 한국어 중의성 사전 구성을 위한 언어적 기초자료들을 살펴보았다. 한국어 어절 중의성은 어휘적 중의성(lexical ambiguity)을 비롯하여, 조사·어미(이상 Affix)가 결합하여 미세하게 변이된 어절 중의성 등을 포함한다. 특히 Afffix 결합 중의성은 ① 불규칙 용언 중의성 ② 축약형 중의성 ③ 신조어 중의성 ④ 접사 결합형 중의성 등으로 세분화된다. 이외에도 상태성 명사와 동작성 명사[12]의 중의성 명사 처리 문제는 차후의 해결 과제로 남겨둔다.

어절 중의성 해결의 첫 번째 절차는 품사 정보를 이용한 것이다. 동사/형용사 중의성의 경우에는 서술어가 요구하는 논항(Argument), 즉 격 정보를 알 수 있으므로 쉽게 해결될 수 있다. 그 다음의 절차는 해당 중의어의 의미자질 부여와 전후 어휘 문맥 정보를 이용하는 방법이다. 특히 관형사 '한, 두, 세, 네, 열' 등은 모두 어절 중의어를 가진 특징이 있다. 이러한 중의어의 처리는 전후 어휘 문맥 정보의 비중이 크게 차지하게 된다.

NLP(Natural Language Processing)는 컴퓨터를 기반으로 한(Computer-based) 정보 처리 기술이다. 따라서 컴퓨터를 대상으로 하여 대명사, 수사, 어미, 접사, 조사, 관형사 등을 비롯한 한국어 자체의 음소들까지도 품사 및 의미자질을 부여해야 한다는 점이 특징적이다. 따라

12) '하다'나 '되다', '스럽다', '롭다' 등이 결합하여 동사나 형용사를 생성하는 명사들이다. 예를 들어 고소(告訴)하다/ 고소하다 (tasty) 등이 그 예이다

서 한국어에 대한 미시적이고 섬세한 정보를 부여하지 않으면 안 된다. 한국어 중의성 처리는 이러한 미시적이고 섬세한 한국어의 정보를 컴퓨터에 교육적으로 접근해야만이 인공지능(Artificial intelligence)이 가능하다. 따라서 중의성 처리의 선결 과제는 해당 어절 중의어의 의미자질 부여와 전후 어휘 문맥 정보를 위한 의미자질의 선정이 필요하다.

참고 문헌

강승식(1999), 『한국어 정보처리와 정보 검색』.

강승식(1999), 『한국어 형태소 분석』.

고창수(1999), 『한국어와 인공지능』, 태학사.

권종성(1996), 『조선어 정보처리』, 한국문화사.

김재훈 · 김길창(1996), "언어지식을 이용한 형태소 해설의 모호성 축소", 『제8회 한글 및 한국어 정보처리』, 한글 및 한국어 정보처리 학술대회 논문집.

김지형 외(1998), 『인공지능 이론 및 실제』, 희중당.

소길자 외(1998), "어휘적 중의성 제거 규칙과 부분 문장 분석을 이용한 한국어 문법 검사기 성능 향상", 『제10회 한글 및 한국어 정보처리』, 한글 및 한국어 정보처리 학술대회 논문집.

송도규 저(1996), 『인지언어학과 자연언어 자동처리』, 홍릉과학출판사.

임해창 외(1993), "한국어처리의 고찰", 『계산의미론과 그 응용』, 민음사.

장석진(1995), 『정보기반 한국어 문법』, 한신문화사.

한국정보과학회 한국어 정보처리 연구회(1999), 『자연언어처리 튜토리얼』, 한국정보과학회 한국어 정보처리 연구회.

Makoto Nagao 저, 황도삼외(1999), 『자연언어이해』, 홍릉과학출판사.

www.csfive.kaist.ac.kr/kcp/

426 국어교육학연구 10 (2000. 7)

<초록>

한국어 중의성 사전 구성을 위한 기초 자료 연구

김 수 정

중의성(ambiguity)이란 자연언어(NLP)의 커다란 특징 중의 하나로, 하나의 단어(word)가 여러 의미(meaning)로 해석될 수 있음을 말한다. 자연어 처리(Natural Language Processing)에서 발생하는 중의성(Ambiguity)은 크게 구조적 중의성(structural ambiguity)과 어휘적 중의성(lexical ambiguity)으로 나뉜다.

한국어의 중의성은 조사와 어미(Affix)의 기능에 의해 문장 성분(sentece component)이 결정되고 불규칙 활용(irregular conjugation) 및 축약 현상이 발생하는 첨가어적인 특성, 즉 조사, 어미의 발달에 기인한다. 본고에서 해결하고자 하는 한국어 중의성은 어휘적 중의성(lexical ambiguity)을 비롯하여, 조사·어미(이상 Affix)가 결합하여 미세하게 변이된 어절 중의성 등을 포함한다. 특히 Afffix 결합 중의성은 ① 불규칙 용언 중의성 ② 축약형 중의성 ③ 신조어 중의성 ④ 접사 결합형 중의성 등으로 세분화된다. 어절 중의성 해결을 첫 번째 절차는 품사 정보를 이용한 것이다. 동사/형용사 중의성의 경우에는 서술어가 요구하는 논항(Argument), 즉 격 정보를 알 수 있으므로 쉽게 해결될 수 있다. 그 다음의 절차는 해당 중의어의 의미자질 부여와 전후 어휘 문맥 정보를 이용하는 방법이다. 특히 관형사 '한, 두, 세, 네, 열' 등은 모두 어절 중의어를 가진 특징이 있다. 이러한 중의어의 처리는 전후 어휘 문맥 정보의 비중이 크게 차지하게 된다. 중의성 처리의 선결 과제는 해당 어절 중의어의 의미자질 부여와 전후 어휘 문맥 정보를 위한 의미자질의 선정이라고 할 수 있다.

<Abstract>

Basic Contents for Korean Ambiguity Lexicon

Kim, Soo Jung

This study focuses on constructing basic contents for developing Korean ambiguity lexicon. Korean ambiguity consists of structural ambiguity and lexical ambiguity. Especially it is one of characteristic solving problems of NLP(Natural Language Processing). This paper focuses on lexical ambiguity including affix-bound ambiguity words. Korean ambiguity words are appeared by combined affixes(Josa and Omi). Especially affix-bound ambiguity words include ① irregular-verb ambiguity ② contraction-type ambiguity ③ coinage ambiguity ④ suffix-bound ambiguity.

Korean disambiguation of NLP have three steps ; First, Using parts of speech information is available for unit of sentence construction disambiguation. Second, it needs assigning semantic features to its ambiguity words. Finally, context informations are available for solution of ambiguity words concerning NLP.