연세대 형태소 분석기 morany:
말뭉치로부터 추출한 대량의 어휘 데이터베이스에 기반한 형태소 분석

윤준태
jtyoon@linc.cis.upenn.edu
IRCS
Univ. of Pennsylvania
Philladelphia, PA 19104, USA

이충희 김선호 송만석
{forever,pobi,mssong}@december.yonsei.ac.kr
서울시 서대문구 신촌동
연세대학교 공과대학 컴퓨터과학과
우편번호 120-749,

Morphological Analyzer of Yonsei Univ., morany:
Morphological Analysis based on Large Lexical Database Extracted from Corpus

Juntae Yoon
jtyoon@linc.cis.upenn.edu
IRCS
Univ. of Pennsylvania
Philladelphia, PA 19104, USA

Chunghee Lee Seonho Kim Mansuk Song
{forever,pobi,mssong}@december.yonsei.ac.kr
Dept. of Computer Science, College of Engineering
Yonsei Univ.
Seoul 120-749, Korea

ABSTRACT

본 논문에서는 연세대학교 컴퓨터과학과에서 연구되어 온 형태소 분석 시스템에 대해 설명한다. 연세대학교 자연 언어 처리 시스템의 기본적인 바탕은 무엇보다도 대량의 말뭉치를 기반으로 하고 있다는 점이다. 예컨대, 형태소 분석 사전은 말뭉치 처리에 의해 재구성되었으며, 3000만 어절로부터 추출되어 수작업에 의해 다듬어진 어휘 데이터베이스는 형태소 분석 결과의 상당 부분을 제한하여 일차적인 중의성 해결의 역할을 담당한다. 또한 복합어 분석 역시 말뭉치에서 얻어진 사전을 바탕으로 이루어진다. 품사 태깅은 bigram hmm에 기반하고 있으며 어휘 규칙 등에 의한 후처리가 보강되어 있다. 이렇게 구성된 형태소 분석기 및 품사 태거는 구문 분석기와 함께 연결되어 이용되고 있다.

1 서론

현재의 연세대학교 시스템의 특징은 대량의 말뭉치로부터 수집한 어휘 데이터에 있다고 할 수 있다. 어휘 데이터는 자동 획득 방법에 의해 그리고 수작업에 의해 모아졌다. 또한 거의 모든 시스템 모듈에서는 학습에 의한 방법과 규칙에 의한 방법이 함께 이용되고 있다. 이는 형태소 분석 시스템에 대해서도 동일하게 적용된다.

연세대학교 형태소 분석 시스템에 대해 설명하기 위해서는 먼저 전체 자연 언어 처리 시스템에 대해 알아볼 필요가 있으므로 현재 구축되어 있는 전체 시스템을 보이면 그림 1과 같다. 이들의 각각에 대해서는 다음 장에서 자세히 설명하도록 한다.

2 품사 분류

먼저, 형태소 분석 시스템과 태거가 이용하고 있는 품사 분류는 표 1과 같으며 이는 일반적으로 한국어 처리 시스템에서 이용되는 분류와 몇 가지를 제외하면 대체적으로 유사하다. 이 장에서는 본 분류 체계의 특징과 현재 표준화에서 제시하고 있는 분류 체계의 차이점 및 태깅 결과에 대해 알아보기로 한다.

전 절에서 언급한 바와 같이 이 분류 체계는 구문분석과 밀접하게 연관되어 있으며 또한 위의 분류는 실제 구문 분석에서 어휘적 중의성이 없는 한도에서 좀더 세분화되어 이용된다. 몇몇 태그에 대해서는 언급의 여지가 있는데, 먼저 ENCM은 어미와 조사의 복합 형태로서 구문분석에 들어갈 때, 전처리 단계에서 어미와 조사로 분리된다. 즉, 이들은 실제 품사 태깅에서 분석의 편이를 위해 복합어미로 다루어지며, 구문분석을 위해서는 둘로 나누어진다. 이를테면, '-다고'의 경우 실제 구문 분석에서는 '다'와 '라고'로 분리되어 입력되는데, 품사 태깅에서는 종속적 연결어미로 다루어지며, 구문 분석에서는 종결형 어미와 보문자[1]로 나누어진다.

[1]이 분류는 구문 분석을 위해 세분화된 체계이므로 표 1에는 나타나 있지 않다

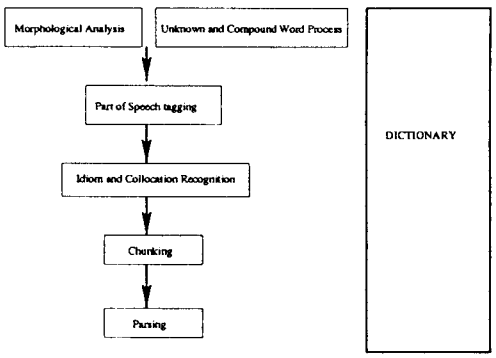

Figure 1: 시스템 개략도

또 다른 특징은 외국어에 대한 분류가 없다. 외국어는 모두 명사로 다루어지며, 표준화 분류와는 달리 'M-16'이나 '4.19'와 같은 단어는 모두 한 단어로 취급되고 있다. 또한 본 시스템에서는 동사화 접미사와 형용사화 접미사가 인정되지 않아 분류 체계에서는 제외되어 있다. 하지만 이들의 생산성은 매우 높기 때문에 형태소 분석 단계에서 고려는 되었으나 자동 태깅에 앞서 모두 하나의 용언으로 간주 변환된다. 따라서 대부분의 '명사+하' 형태는 '동사'로 등록되어 있으며 추정될 때는 어미의 활용을 고려하여 형용사나 동사 혹은 형용사, 동사로 태거에 입력된다. 이번 시험 과정에서는 이들을 재분리해야 했으나 분리 과정에서 오류가 있었다.

그리고 실제 품사 태깅된 결과에는 많은 차이가 있는 편인데, 일례가 시간 관련 품사 통용어에 관한 것이다. 이들에 대한 관찰 결과 품사 통용어의 경우 부사인지 명사인지에 대한 결정은 거의 어휘적으로 결정되며 개별 어휘마다 다른 특징을 가진 것으로 나타났다. 또한 이들은 수십만 수준의 품사 부착 말뭉치로는 올바르게 처리될 수 없음을 알 수 있다. 즉, 형태소 분석과 구문 분석의 중간 단계에서 이들을 인식하고 처리해 주는 부분이 필요하며 본 시스템에서는 chunker가 이 역할을 하고 있다. 때문에 실제 시험 과정에서는 이들이 모두 오류로 나타나게 되었다.

이외에도 의미가 많이 바뀌었다고 생각되는 부분은 형태소별로 세세히 나누지 않고 하나의 형태소로 분리한다. 예컨대 '때늦은'의 경우 형태소별로 나누게 되면 '때/nc+늦/pa+은/etm'이겠으나 하나의 관형사로 보았으며, 더욱이 '예뻐하다/귀여워하다'와 같이 '형용사+하다'로 이루어진 어휘는 모두 하나의 동사로 처리되어 오류가 발생했는데, 이와 같이 하나의 동사로 다룬 이유는 구문분석의 용이함을 위해서이다. 예컨대 '예쁘다'와 '예뻐하다' 주위에서 발생하는 문맥은 상당한 차이를 보이게 되며 이들은 각각 다른 용언으로 파악하는 것이 낫다고 판단되었기 때문이다. 따라서 실제로 표준화에서 제시된 태깅 분류와 많은 차이를 보인 것은 아니나 태깅 결과와의 차이는 매우 많아 이들을 충분히 고려하기는 어려웠다.

3 시스템에 대한 개요

이 장에서는 시스템의 각 부분에 대해 간략히 설명하기로 한다.

3.1 형태소 분석 및 태깅

형태소 분석 시스템은 규칙 기반 형태소 분석으로 기본 형태소 분석, 복합명사 분석, 미등록어 추정이 한데 묶여 있다. 먼저 형태소 분석에서 이용되는 사전에 대해 알아보면, 약 89,000 어휘의 형태소 단어 사전이 있다. 일반적으로 알려진 기본 어휘 수보다 좀 많은 수인데, 이는 전 절에서 말한 바와 같이 가능한 한 파생된 형태의 어휘를 많이 담고자 하였기 때문이다. 이러한 파생된 형태는 기본 어휘의 형태소 사전을 이용하여 말뭉치에 적용하고 이로부터 발견된 새로운 어휘들이 새로이 등록됨으로써 추가되었다.

둘째, 형태소 기분석 사전이 있다. 이를 개발한 것은 근본적으로 한국어 형태소 분석의 결과가 결코 중의성을 많이 포함하지 않는다는 점에 근거한다. 하지만 한국어 기능어의 수를 고려할 때 모든 어절에 대해 기분석 사전을 만들 수는 없다. 그러나 300만, 1000만, 3000만 어절 말뭉치를 조사해 보았을 때, 일정량의 어절 수가 전체의 65-70% 가량을 이룸을 알았다. 이 어절의 수는 약 6만 어절이며 경계가 되는 빈도수는 3000만 어절에서 빈도 49인 어절이었다. 이들에 대해서는 수작업 교정이 가능하며 형태소 분석의 효율을 높이는 효과가 있다. 하지만 무엇보다도 중요한 사실은 이들을 면밀하게 수작업할 경우, 중분류 수준[2]에서 85% 가량이 중의성을 가지지 않는

[2]명사, 대명사, 수사, 부사, 관형사 등으로 나눌 때

명사 (noun)	일반명사(independent)	고유명사(proper)	NNIN1
		보통명사(common)	NNIN2
	의존명사(dependent)	비단위성(common)	NNDE1
		단위성(unit)	NNDE2
대명사(pronoun)			PN
수사(number)			NU
동사(verb)			VBMA
형용사(adjective)			AJMA
지정사(copula)			CO
보조용언(auxiliary verb)			AX
부사(adverb)	성분부사(constituent)		ADCO
	문장부사(sentential)		ADSE
	접속부사(conjunctive)		CJ
관형사 (adnominal)	성상관형사(configurative)		ANCO
	지시관형사(demonstrative)		ANDE
	수관형사(numeral)		ANNU
감탄사(exclamination)			EX
조사 (postposition)	격조사(case)	주격조사(nominative)	PPCA1
		목적격조사(accusative)	PPCA2
		관형격조사(possessive)	PPCA3
		호격조사(vocative)	PPCA4
		부사격조사(adverbial)	PPAD
	접속조사(conjunctive)		PPCJ
	보조사(auxiliary)		PPAU
어미 (ending)	종결어미	평서형(assertive)	ENTE
	연결어미 (conjunctive)	대등적(coordinate)	ENCO1
		종속적(subordinate)	ENCO2
		보조적(auxiliary)	ENCO3
	전성어미 (transformative)	관형형어미(adnominal)	ENTR1
		명사형어미(nominal)	ENTR2
		부사형어미(adverbial)	ENTR3
	복합어미(ending+pp)		ENCM
선어말어미(pre-ending)			PE
접미사(suffix)			SF
접두사(prefix)			PF
쉼표(comma)			CM
종결부호(termination)			SC
왼쪽괄호(left quotation mark)			LQ
오른쪽괄호(right quotation mark)			RQ
심볼(symbols)			SY

Table 1: 형태소분석 및 구문분석을 위한 품사 분류 (연세대학교)

다는 사실이다. 중의성이 많이 발생하면 발생할 수록 오류의 가능성이 높아짐을 고려할 때, 이러한 필터 작용은 1차적인 품사 태깅의 역할을 하며 품사를 제한하게 된다.

형태소 기분석 사전의 또 하나의 특징은 매우 일반적인 오류가 포함되어 있다는 점인데, 흔히 발생하는 띄어쓰기 오류 같은 것이 그것들이다. 이들은 그 사용 빈도를 인정하여 올바른 형태소 분석 결과를 넣어두었다. 또한 흔히 발생하는 구어체나 방언 등에 대한 분석도 포함하고 있다.

한편 개발된 형태소 사전을 이용하는 형태소 분석기는 기본적인 형태소 전이망을 따르는 형태소 분석을 시행한다. '명사+조사'나 '동사+선어말어미+어미', '명사+접미사+지정사+어미' 등이 그 예이다. 그러나 한국어의 경우 띄어쓰기가 매우 불규칙적으로 일어나는 특징이 있기 때문에 이

와 함께 복합명사 분석과 띄어쓰기 오류 검사가 함께 시스템내에 고려되어야 한다. 복합명사 분석은 형태소 분석과 함께 수행되나 모든 복합명사가 형태소 분석에서 고려될 수는 없으므로 복합명사 분석에서는 가장 그럴듯한 분석 하나만이 형태소 분석기에 제안된다. 복합명사 분석의 알고리즘은 위에서 설명한 6만 어절 기분석 형태소 분석 사전으로부터 추출된 단일명사와 빈도수를 기본으로 이루어지며 또한 사전에는 이를 보완하기 위한 추가적으로 형태소 분석 사전으로부터 얻은 명사가 추가되어 있다. 그리고 기분석 사전이 약 5만 어휘로 이루어져 있다. 기분석 사전에는 현재의 형태소 분석 시스템에서 처리 곤란한 기분석도 포함되어 있다. 한편, 띄어쓰기 오류는 기본적인 형태소 분석이 실패했을 경우 고려되며, 두 어절의 띄어쓰기 오류만이 고려된다.

태깅은 HMM 모델을 이용하며, 어절간의 전이확률 어절 내부의 전이확률에 의해 결과가 결정되며 bigram 모델을 채택하고 있다. 또한 이에 의해 나온 결과는 어휘 규칙에 따라 보정된다. 현재 이러한 어휘 규칙은 주로 수작업에 의해 만들어져 있다.

3.2 파서

처음에 설명된 바와 같이 현재 형태소 분석기는 파서와 매우 밀접한 관계가 있다. 그러한 이유로 결과를 자유로이 바꾸기가 다소 어려운 딱딱한 시스템이 되어 있다. 이 장에서는 구현되어 있는 파서에 대해 간략하게 알아보면서 형태소 분석 결과와의 연관성에 대해 기술한다.

먼저 형태소 분석기로부터 나온 결과는 숙어 및 연어 인식기를 거치게 된다. 여기서는 '올-수-있'과 같은 연어적인 성격을 가지는 어휘들이 다중어로 간주되어 하나의 어휘로 묶이게 된다. 둘째, 여기서 나온 결과는 chunking 단계로 들어가게 되는데, 이 단계에서 시간 부사 인식이 이루어진다. 예를 들어 '오늘 저녁'과 같은 형태소 열이 하나의 구로 인식되게 된다. 이 단계에서 '오늘'과 같은 시간 명사가 시간 부사인지 시간 명사인지 판별되기에 형태소 분석 결과로부터 나오는 모든 '오늘'은 명사로 출력된다. 이들은 결과에서 모두 오류의 원인이 되었다.

이러한 품사 통용어에 대한 연구 결과 이들은 어떤 어휘가 주어지는지에 따라 구문적으로 상당히 다르며 따라서 이들에 대한 별도의 처리가 필요하다는 사실을 주지할 필요가 있다.

현재 구문 분석기는 역시 3000만 어절 말뭉치로부터 추출된 공기 데이터와 규칙으로 이루어져 있다. 추출된 공기 데이터의 수는 약 250만 쌍이며 2이상 발생한 쌍의 수가 약 70만 쌍 정도가 되었으며 이들이 구문 분석과 보조사나 관형절에 의한 격 추정에 이용되고 있다. 여기서 주의할 점은 위에서 밝힌 바와 같이 '예쁘하다'와 '예쁘다'에 주어지는 명사구는 다르므로 각각을 다른 어휘로 간주했다는 점이다. 이들을 분리하는 것이 불가능한 것은 아니나 표준화 이전에 이 경우의 '하다'는 보조용언으로 간주하지 않고 구성한 탓에 현재 이들을 분리하여 형태소 분석하기에는 처리해야 할 문제가 너무 많아져 그대로 태깅하였다. 이들 역시 모두 태깅 오류로 나타났다. 그에 비해 '느껴지다' 역시 '느끼다'와는 다른 분포를 보이는데 '지다'는 과거 표준 문법에 따라 보조용언으로 간주하여 '느끼+어+지+다'로 분리하고 '느끼어지다'와 명사의 분포를 찾는 형식을 취했다.

4 결과의 분석

전체적으로 볼 때 분석의 오류는 주어진 학습 말뭉치와 본 시스템이 가지고 있는 형태소 사전이 많은 부분에서 다른 데다 시스템의 교정시 거리상의 제약으로 인해 의사 소통이 원활하지 못해 제대로 학습이 되지 못했거나 학습말뭉치가 시스템에 충분히 반영되지 못했다는 점을 들 수 있다. 이 점은 향후 표준화에서 제시된 어휘들과의 비교를 통해서 재검토되어야 할 부분으로 나타났다. 또한 학습 말뭉치의 교정을 좀더 충실히 할 필요가 있었다. 즉, 본 시스템과 표준화에서 제안된 품사나 태깅 방식이 일치하지 않는 부분에 대해 좀더 정교하게 다듬어야 할 필요가 많음이 밝혀졌다.

4.1 명사 추출

명사 추출 오류는 크게 3가지 원인에 의해 나타났다. 첫째는 미등록어 추정과정에서의 오류이다. 이의 원인은 학습말뭉치에서 주어진 단어가 현재의 사전에 제대로 반영되지 않은 이유가 하나이다. 주어진 학습 말뭉치를 이용해 본 시스템의 태거를 학습시키기 위해서는 위에서 밝힌 여러 가지 차이점을 수정해야 하나, 시험날까지 모든 결과가 분석되고 수정이 되지 못한 이유로 인해 재학습을 시키지 못했다. 따라서 어휘 확률이나 사전에 등록되어야 할 단어들이 등록되지 못한 경우가 상당히 많았다. 이들은 명사 추출에 있어서 오류의 주요 원인이 되었는데, 예를 들어 '신한국당'의 경우 '신한국당' 하나의 명사로 분석되어야 하나 미등록어 추정시 '신한국/nc+당/sf'로 분석되었다. 이러한 오류는 한번 발생하면 자연 언어 문서의 특성상 지역성 및 반복성을 가지기 때문에 일단 잘못 분석되면 많은 오류를 발생시켰다. 그러나 이러한 오류는 오류로 인한 시스템의 비정확성에도 불구하고 현재 시스템의 보완 및 어떻게 분석하는 것이 형태소 분석기에 도움을 줄 수 있는가에 대한 방향을 제시해 주었다고 할 수 있다. 다음은 그러한 이

유로 오류로 분석된 예이다.

단어	morany 분석 결과
신한국당	신한국/nc+당/xsn
첫소식	첫소식/nc
종금당	종금/nc+당/xsn

오류에 대한 또 다른 이유는 형태소 분석기의 오류로 인해 잘못된 결과를 출력한 경우이다. 이 유형의 오류는 다시 2가지의 경우로 나뉘는데, 첫째 형태소 분석 사전에 어휘가 잘못 등재된 경우와 둘째 미등록어 추정이 오류를 발생한 경우로 나누어 볼 수 있다. 예컨대, '억원'의 경우 명백히 '억/nn+원/nd'으로 분석되어야 하나 사전에 '억원'이 등록되어 있고 '억원'을 복합어로 추정하지 못했기 때문에 발생한 경우이다. 또, '류근찬'과 같은 인명의 경우 '류근차+ㄴ'과 같은 후보어들을 생성하게 되는데, 이러한 과생성이 오류를 유발하기도 하였다. 지나친 과생성이 문제가 되는 부분이라고 할 수 있다.

단어	분석 결과
억원	억원/nc
류근찬	류근차/nc+는/jx
네자리수	네자리수/nc

세째는 표준화에서 제안된 부분과 본 시스템이 일치하지 않는 부분을 극복하지 못한 것이다. 이는 충분한 시간을 두고 튜닝하지 못했기 때문이기도 하다. 예를 들어보면, '-씨'의 경우 본 시스템에서는 접미사로만 간주하였으나 표준 말뭉치에서는 많은 경우에 이들의 띄어쓰기가 교정되어 있었다. 따라서 이들이 띄어져 표기된 경우 일반 명사로 출력되었다. 표 2에서 이러한 오류의 몇몇 예들과 잘못 분석된 이유에 대해 간략하게 기술했다. 이들중 '지금'이나 '처음'이 조사의 도움없이 나타난 경우 전 절에서 언급한 시간명사와 달리 처리를 하였다. 즉, 하나의 부사로 간주했으나 이들도 시간명사와 동일하게 취급되어야 할 필요가 있을 것으로 생각된다.

표 3은 태그를 바꾸는 과정에서 생긴 오류이다. 본 시스템에서 접미사 '-적-' 관련 명사의 경우 하나의 명사로 간주하고 있는데, 현재 시험을 위해 수정하는 과정에서는 3글자 이상의 것만을 대상으로 하였고 이에 따라 오류가 발생하기도 했다. 또, '-하-'나 '-되-'의 경우에도 형태소 분석에서는 하나의 동사나 형용사로 간주하였는데, 이를 분리할 때 모든 명사가 아닌 적은 수의 명사만을 이용하여 제대로 분리가 되지 않은 예이다.

전체적으로 볼 때, 첫번째나 두번째 예와 같은 피할 수 없는 오류들도 있고 여전히 시스템에 대한 수정의 여지를 가지고 있으나, 이들도 상당 부분 실제로는 수정가능하며 표준화 방법론과의 튜닝과정에서 불충분한 정보로 인한 오류 등을 고려할 때, 좋은 성능을 보였다고 할 수 있다.

4.2 품사 태깅

자동 태깅은 명사의 오류를 포함하고 있다. 명사 분석에서 판명된 오류는 역시 전체 태깅에서도 오류를 유발하였다. 품사 태깅은 전체적으로 비교적 낮은 정확률을 보였으나 이들의 상당부분은 ETRI의 표준화 태깅 방식에로의 튜닝 과정에서의 오류임을 감안할 때 나름대로의 방식 및 오류 유형을 파악해 볼 필요가 있다. 실제로 이러한 오류에도 불구하고 명사 추출이 매우 높은 정확률을 보이고 있음은 태깅이 상당히 정확함을 반증하는 것이다. 본 시스템의 관점에서 파악하기 위하여 학습 말뭉치로부터 임의의 문장들로부터 2055 어절을 추출하고 수작업으로 분석하였다. 학습말뭉치로부터 학습시키지 않은 채로 태깅 결과를 얻었으며 그 결과 어절 단위에서 약 94.7%의 정확성을 보였다.

우선 태깅 오류는 흔히 발생하는 태깅 오류를 포함한다. 명사 추출에서도 보인 바와 같이 이들은 시스템이 오류를 발생시킨 것으로 '신한국당'을 명사와 접미사로 분석하는 것과 같이 시스템에 악영향을 미치는 부분들이다. 이들이 사전으로 제공되었다면 오류를 막을 수는 있겠으나 이러한 예는 어디에서나 발생가능하다. 따라서 이를 제대로 해결할 수 있는 방법론이 필요할 것이다. 또한 이들중에는 '지금'과 '오늘'에서 보이는 태깅의 불일치와 같이 시스템 자체적으로 다듬어져야 하는 부분이 있다.

다음으로는 시험에서는 오류로 판정되었으나 본 시스템의 관점에서 오류가 아닌 부분들이다. 본 시스템에서 추구하는 바는 무엇보다도 가능한 한 하나의 의미를 가진 단어는 하나의 어휘로 간주하는 것이다. 따라서 이들이 재조절되기 위해서는 좀더 세분화된 형태소의 단위로 나누는 변환 과정을 거쳐야 하는데 이 과정에서 변환을 제한하거나 새로이 바꾸는 과정이 필요한 관계로 오류를 피할 수 없었다. 실제로는 이들이 상당히 많이 분포되어 있었으며, 다음 표 4는 그 예들을 보여준다. 표에서 일부 형태소들은 분리 과정에서 제대로 분리되지 않았으며, 표준화에 맞추기 위해 분리하는 과정에서 일관성을 유지하려 하는 과정에서 달리 분리가 된 경우가 있다. 또 어미류의 이형태에 대한 부분도 오류로 나타났는데, 이들은 실제 파싱 과정에 들어갈 때는 하나의 대표형으로 변환되는 과정을 거치게 된다. 이외에도 표준화 학습 말뭉치가 오류를 포함하는 경우 역시 많았다.

명사	오류원인
한번	명사로 출력
달러값	달러를 의존 명사로만 간주
내일,오늘	오늘,내일 등의 시간명사는 명사로 간주
내, 측	내, 측, 상, 하 등은 조사성 위치명사로 모두 일반명사 처리
나흘째	'나흘/nc+째/xsn'으로 분석
올들어	'올들어'를 하나의 부사로 간주
이번, 이날	모두 하나의 명사로 간주
지금	부사로 간주
한꺼번에	하나의 부사로 간주

Table 2: 제안된 방법과 본 시스템의 차이로 인한 오류 1

관련단어	추출되어야 할 명사
법적	법
폭락하	폭락
급등하	급등
망연자실하	망연자실
공명정대하	공명정대
타결되	타결
불분명하	불분명

Table 3: 제안된 방법과 본 시스템의 차이로 인한 오류 2

5 토의 및 향후 연구

전체적으로 볼 때, 본 시스템은 비교적 안정적인 좋은 결과를 보이고 있었다. 그러나 5%의 오류는 여전히 문제점을 안고 있는데, 이는 100어절 중에 5개의 오류가 있음을 의미한다. 이들의 일부는 현재의 적은 양의 학습 말뭉치로는 극복하기 힘든 오류들도 있으나 시스템에 대한 손질로서 극복가능한 오류도 꽤 많아 여전히 전체 시스템은 희망적이라 할 수 있다. 이들 중에는 내용의 파악에 중요한 명사의 오류도 포함되어 있으며 이는 지속적인 연구가 필요한 부분의 하나이다. 또한 전절에서 언급한 바와 같이 태그간 불일치도 본 시스템의 문제점으로 지적될 수 있으며 사전 어휘의 고찰을 통해서 수정되어야 할 부분이다.

연세대학교 시스템은 기본적으로 어휘 정보를 기반으로 하고 있으며 이들이 전체 시스템의 성능을 향상시키도록 유도하고 있으며, 이러한 노력은 앞으로도 계속될 것이다. 따라서 적은 양의 품사 부착 말뭉치 뿐만 아니라 대량의 원시 말뭉치로부터의 비교사 학습에도 많은 노력이 기울여질 것이다.

다음으로, 이번 ETRI 표준화 과제로부터 제안된 태깅 방식에 약간의 수정이 필요하지 않을까 생각된다. 현재 배포된 말뭉치는 비교적 정확하게 품사가 부착이 되어 있으며, 특히 품사의 분류에는 대부분의 시스템들이 큰 차이가 없으리라 생각되나 태깅 방식에는 개별적으로도 차이가 많을 뿐 아니라 이들은 이미 모두 개발되어 있는 상태여서 섣불리 수정하기 어려운 점을 가지고 있다. 대부분의 시스템은 표준 문법 이론을 따르고 있으므로 이를 충분히 고려하는 것이 바람직하지 않을까 생각해본다.

수정이 가해질 여지가 있는 부분은
첫째, 어미에 관한 부분인데 지나치게 분리되는 경향과 함께 형태소 자체를 바꾸는 문제를 언급해야 할 것 같다. '다며'를 '다+면서'로 분리하거나 '지만'을 '지마는'으로 바꾸는 등은 검증되지 않은 것으로 위험한 방법이 될 수 있음이 우려된다. 이와 함께 고려되어야 할 점은, 현재 표준화 작업이 사실상(de facto) 표준을 유도하고 있는 만큼 지나치게 새로운 방식의 태깅은 문제가 될 수 있다는 점이다. 이번 형태소 분석기의 평가에 참여한 많은 기관들이 말뭉치의 태깅 방식을 그대로 따랐다기보다는 오히려 전체적인 구조는 그대로 유지한 채 평가에 맞추어 출력을 바꾸었다는 점은 고려를 해봐야 할 문제이다. 즉, 평가나 표준화 이후에 이를 그대로 따르지 않을 가능성이 상당히 많다는 점이다. 특히 문제가 될 수 있는 부분은 태깅 방식인데, 문제로 지적될 수 있는 태깅 방식에 대해 현재 시스템들의 전체적인 태깅 방식을 조사하고 무리가 없는 경우 전체적으로 통일될 수 있는 방향을 따르는 것이 바람직하지 않을까 생각된다.

둘째는 말뭉치 품사 부착에 관한 것이다. 학습

	본 시스템	표준화 제안
제공하	제공하/pv	제공/nc+하/xsv
이질적	이질적/nc	이질/nc+적/xsn
으로부터	으로부터/jc	으로/jc+부터/jx
이라는	이/co+라는/etm	이라고/jc+는/etm
이다고	이/co+다/ef+라고/jc	이/co+라/ef+고/jc
었다고	었/ep+다/ef+라고/jc	었/ep+다/ef+고/jc
아쉬운	아쉽/pa+은/etm	아쉬/pa+ㄴ/etm
나타났	나타나/pv+ㅆ/ep	나타나/pv+았/ep
일부	일부/nc	일부/mm

Table 4: 본 시스템과 표준화 과정에서 제안된 내용의 차이점

말뭉치로 배포된 품사 부착 말뭉치에 대한 검토 결과 여전히 오류를 많이 포함하고 있다는 점이다. 이는 품사의 세분화로부터 온 결과라기보다는 오히려 품사 태깅의 복잡성에 있다고 생각된다. 이는 일관된 품사 부착에 장애가 되는데, 위에서 언급한 어미에 관한 문제도 이에 해당한다. 또 하나의 일례가 품사통용어에 관한 것이다. 품사 통용어에 대한 품사 선택은 실질적으로는 어휘적으로 결정되게 되는데, 이들 어휘에 대한 것이 명확히 밝혀져 있지 않아 실제로 비일관성이 지적될 수 있음이 학습 말뭉치에 보여졌다.

세째는 말뭉치의 양인데, 필요한 어휘 관계를 얻기 위해서 필요한 말뭉치의 양은 수십만으로는 너무 부족한 감이 있다. 처음 초기화 작업이 진행된 만큼 이를 바탕으로 300만 어절 수준까지 확장해야 하지 않을까 생각된다.

그럼에도 불구하고 이번 형태소 분석 평가 대회는 시스템을 재정비하는데 큰 도움이 될 수 있었다. 또한 자연 언어 처리에 있어서 어휘 자료 및 어휘 지식의 중요성이 매우 강조되고 있는 지금 표준화 등을 통한 자료의 공유는 이 분야의 발전에 매우 중요하다고 생각되며, 따라서 이러한 노력이 계속 되어야 함은 충분한 설득력을 가진다.

6 결론

본 논문에서는 현재까지 구축된 연세대학교의 형태소 분석 시스템에 대해 설명하였으며, 이를 표준 품사 부착 말뭉치에 적용한 결과에 대해 논의하였다. 결과적으로 볼 때, 현재 연세대학교의 시스템은 표준화 단계에서 제안된 태깅 결과와 상당한 차이를 보이고 있었으며 이의 수정이 그다지 용이하지 않았으며 이러한 결과는 많은 오류를 동반하는 결과를 낳았다. 그러나 자체 평가 결과 상당 부분은 관점의 차이로부터 온 것이 많다는 사실도 밝혀졌으며 결과는 훨씬 정확한 것으로 나타났다. 따라서 이들을 어떻게 보완하느냐의 문제가 남아있다.

또 하나, 결과적으로 몇 가지 문제점이 제기될 수 있는데, 표준화 품사 부착 말뭉치의 보완 및 수정은 가장 중요한 문제로 여겨진다. 특히 품사 분류보다는 태깅 방식에 대한 문제가 제기될 수 있는데 현재까지 구축되어 있는 대부분의 시스템은 상당수가 일정한 전통을 따르고 있으며 그 기준은 전통문법일 것이다. 따라서 이로부터 너무 벗어난 태깅은 구축된 많은 시스템에 큰 희생을 요구한다.

그럼에도 불구하고 이러한 표준화에 대한 노력 및 자료의 공유는 매우 중요하다고 생각되며 이를 더욱 확장하는 문제가 남아 있다.

References

권명국 1992. 말뭉치 품사태그 체계화 연구. 연세대학교 대학원 전산학과 석사학위 논문.

김병희 1995. 형태소 접속 특성과 인접 말마디 정보를 이용한 형태소 분석기. 연세대학교 대학원 전산학과 석사학위 논문.

박영환 1992. 말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현. 연세대학교 대학원 전산학과 석사학위 논문.

박혜준 1994. 말뭉치에서의 품사꼬리달기 시스템의 구현. 연세대학교 대학원 전산학과 석사학위 논문.

박혜준, 윤준태, 송만석. 1994. 말뭉치 품사꼬리달기 시스템 구현. 정보과학회 봄 학술발표 논문집, 1994

윤준태 1998. 공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석. 연세대학교 대학원 컴퓨터과학과 박사학위 논문.

임권묵 1995. 한국어 형태소 분석에서의 오분석 제거와 중의성 해결. 연세대학교 대학원 컴퓨터과학과 박사학위 논문.