어휘별 중의성 제거 규칙과 통계 정보를 이용한 한국어 품사 태깅

Korean Part-of-Speech Tagging using Disambiguation Rules for Ambiguous Word and Statistical Information

안광모, 한규열, 서영훈
충북대학교 컴퓨터 공학과

Kwang-Mo Ahn(ahnmo@nlp.cbnu.ac.kr), Kyou-Youl Han(sept102@nlp.cbnu.ac.kr),
Young-Hoon Seo(yhseo@chungbuk.ac.kr)

요약

규칙 정보와 통계 정보를 이용하는 복합적 품사 태깅은 통계를 기반으로 하는 방법의 견고함과 확장성을 가지고, 통계 정보에 벗어나는 언어현상들을 규칙 정보를 이용하여 해결함으로서 높은 정확도를 가질 수 있다. 하지만 기존의 연구는 규칙 정보의 제한적인 적용범위 때문에 통계 정보에 벗어나는 언어 현상을 처리할 수 없는 경우가 발생하게 된다. 본 논문에서는 이를 해결하기 위하여 어휘의 사전적 의미와 문맥적 관계를 반영할 수 있는 "어휘별 중의성 제거 규칙"을 제안한다. 어휘별 중의성 제거 규칙은 세종 말뭉치로부터 말뭉치 데이터를 형태소 분석하여 상위 50%의 중의성 어휘에 대한 사전적 의미와 문맥적 관계를 고려한 품사 태깅 정보를 추출하고 이것을 규칙으로 만든 것이며, 현재까지 총 1,815개로 구성되어 있다. 어휘별 중의성 제거 규칙을 기존의 복합적 품사 태깅 시스템에 적용하여 품사 태깅의 정확도를 높일 수 있었다.

■ 중심어 : | 품사 태깅 | 중의성 제거 |

Abstract

A hybrid part-of-speech tagging approaches may be robust, easily extendable, and accurate because they can have the advantages of both statistical approach and rule-based approach. But conventional hybrid part-of-speech tagging systems hardly resolve some morphological ambiguities which can't be resolved by statistical information. It is because the coverage of rules is narrow. So, we define disambiguation rules for individual ambiguous word based on syntax and semantics of surround words. We select words from which the top 50% of ambiguities are occurred in Sejong corpus and build 1,814 rules for them. The accuracy of our hybrid part-of-speech tagging system using those rules is 98.28%.

■ keyword : | Part of Speech | Disambiguation |

* 본 논문은2007년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었습니다.
(This work was supported by the research grant of the Chungbuk National University in 2007)

접수번호 : #081008-004
접수일자 : 2008년 10월 08일
심사완료일 : 2008년 10월 28일
교신저자 : 서영훈, e-mail : yhseo@chungbuk.ac.kr

I. 서 론

자연언어처리를 어렵게 하는 가장 큰 원인은 자연언어가 중의성(ambiguity)을 가지고 있기 때문이다. 여기서 중의성이란 하나의 어절이나 문장 등이 여러 가지 해석을 갖는 경우를 말하며, 중의성 제거는 자연언어처리에 있어서 가장 중요한 논점 중 하나이다. 자연언어의 중의성은 어휘 중의성(lexical ambiguity), 구문 중의성(syntactic ambiguity), 의미 중의성(semantic ambiguity)으로 나눌 수 있으며, 이 중 어휘 중의성을 제거하는 자연언어처리의 과정이 품사 태깅(part-of-speech)이다. 예를 들면, “watch”라는 단어는 품사가 명사 또는 동사가 될 수 있는 중의성 어휘이다. 하지만, “I have a watch.”라는 문장에서 “watch”는 문맥상 명사로 쓰였다는 것을 알 수 있으며, 이와 같이 어떤 단어가 가질 수 있는 품사 중 문맥에 가장 적절한 품사를 결정하여 어휘 중의성을 제거하는 과정이 품사 태깅이다.

품사 태깅을 수행하는 방법에는 크게 규칙 기반의 방법[1]과 통계 기반의 방법[2-4], 그리고 이 둘의 장·단점을 상호보완적으로 사용하는 복합적 방법[5-9]가 있다. 이 중 복합적 방법에 의한 품사 태깅은 규칙 정보로 처리할 수 있는 언어 현상을 높은 정확도로 처리하고, 규칙 정보를 적용할 수 없는 언어 현상은 통계 정보를 이용한 확률 값을 이용하기 때문에 높은 확장성을 갖게 되어 최근에 주로 사용되는 품사 태깅 방법이다. 하지만 이 방법은 규칙으로 처리할 수 없는 언어 현상을 통계 정보를 이용하여 해결하기 때문에, 통계 정보에 벗어나는 언어 현상에 대하여 잘못된 결과를 낼 가능성이 여전히 존재한다.

따라서 품사 태깅의 정확도를 높이기 위해서는 통계 정보를 적용하기에 앞서 어절이 갖는 사전적 의미와 문맥 관계를 충분히 고려할 수 있는 규칙 정보를 적용할 필요성이 있다. 하지만 지금까지 연구된 복합적 품사 태깅 시스템[5-9]에 적용된 대부분의 규칙 정보는 사전적 의미나 문맥 관계를 충분히 반영한 수준의 것이 아니다. 이에 본 논문에서는 어절의 사전적 의미와 문맥 관계를 모두 고려한 “어휘별 중의성 제거 규칙”을 제안한다. 어휘별 중의성 제거 규칙은 중의성 어휘의 사전적 의미와 문맥 관계를 고려한 규칙 정보로서 21세기 세종 천만 어절 균형 말뭉치에서 추출한 중의성 어휘 중 빈도수가 상위 50%인 어휘들을 선별하여 구축되었다.

본 논문에서는 통계 정보에 벗어나는 언어 현상들의 예를 들고, 이를 어휘별 중의성 제거 규칙으로 처리할 수 있음을 보인다. 그리고 어휘별 중의성 제거 규칙을 기존의 복합적 품사 태깅 시스템에 적용하여 정확도의 향상이 있었음을 실험을 통해 보인다.

II. 기존의 품사 태깅

지금까지 연구된 품사 태깅의 방법들에서 사용된 통계적 모델에는 크게 HMM(Hidden Markov Model), 최대 엔트로피 모델(Maximum Entropy Model)이 있다. 이 모델들은 그 이론적인 바탕이 견고하고, 확률 값을 이용함으로서 비문과 같이 언어 규칙으로 적용할 수 없는 현상에도 적용할 수 있는 견고함과 확장성이 있으며, 말뭉치로부터 양질이 통계 정보를 수집하면 어느 정도 높은 정확도를 갖는 품사 태깅을 수행할 수 있다. 게다가 2007년에는 21세기 세종 최종 성과 발표가 있었고, 대규모의 말뭉치를 활용할 수 있는 환경 또한 갖추어져 품사 태깅에 있어 통계를 기반으로 한 방법은 매우 좋은 방법임에는 틀림이 없다.

하지만 자연언어에서는 통계정보에 벗어나는 언어현상이 많이 발생한다. 따라서 통계정보에 벗어나는 언어현상으로 인한 품사 태깅 오류의 가능성을 줄이기 위해 문맥 정보를 통계정보에 반영하려는 방법들도 있었다[2][4]. 예를 들어, [2]의 경우는 관형형어미나 관형형조사 뒤에는 명사가 나올 확률이 높다거나 목적격 조사 뒤에는 명사가 나올 확률이 높다는 등의 조사와 어미의 특성을 고려하였으나, “아름다운 너와 멋진 그는 …”, “사과를 책상 위에 놓아라.”의 경우처럼 이 방법을 적용하기에는 예외현상이 많이 존재한다. [4]는 연어를 통하여 확장된 문맥 정보를 이용한 통계 정보를 이용하거나 동사의 중의성을 해소하기 위해 원시 말뭉치 등을 이용하는 방법을 제안하기도 했지만, 그 처리 범위가 협소

하여 품사 태깅의 전처리 정도로만 이용될 수 있는 방법이다. 그리고 무엇보다 이러한 방법들의 근본적인 문제점은 통계 정보를 바탕으로 하여 통계 정보에 벗어나는 언어현상에 대한 오류를 해결할 수 없다는 것이다.

따라서 이러한 문제를 해결하기 위해서는 규칙 정보가 통계 정보에 벗어나는 언어 현상을 최대한 해결할 수 있어야 한다. 즉, 품사 태깅의 성능 향상을 위해서는 기존의 규칙 정보보다 어휘의 사전적 의미와 문맥 관계를 충분히 반영할 수 있는 규칙 정보가 필요하다. 하지만 지금까지의 복합적 품사 태깅 방법은 이런 요소를 충분히 반영할 수 있는 규칙 정보를 적용하지 못하고 있다.

예를 들어, [5]의 경우, 1어절 이내에 결합 가능한 품사열 규칙이나, 각 품사가 올 수 있는 위치들의 제약, 보조 용언 앞에 올 수 있는 형태소의 제약을 이용하는 등 품사 정보나 연결어미의 형태소 정보 정도만을 규칙 정보에 활용하는 수준이며, [6-8]의 경우는 말뭉치로부터 확률적인 요소를 기반으로 하여 규칙 정보를 추출하여 많은 양의 규칙을 빠른 시간에 수집할 수 있지만, 이러한 방법들은 어휘의 사전적 의미 및 문맥 관계를 충분히 반영할 수 있는 규칙의 수집이 어렵다. 그리고 [9]은 어휘 정보를 이용하여 중의성을 해소하고는 있지만, 어휘의 표층 구조만을 이용하여 어휘의 사전적 의미가 충분히 반영되지 못하고, 하나의 어휘에 대하여 규칙의 수가 많아지는 문제가 있으며, 말뭉치로부터 규칙을 추출하기 위해 앞 뒤 어휘만을 보아 문맥적 관계를 충분히 고려하지 못한다.

보다 높은 품사 태깅의 성능을 위해 통계 정보의 연구만큼 규칙 정보의 연구도 필요하지만, 아직까지 규칙 정보에 대한 연구는 부족한 점이 많다. 따라서 본 논문에서는 이러한 문제점들을 해결하기 위한 넓은 처리범위의 규칙 정보인 "어휘별 중의성 제거 규칙"을 제안하며, 이에 대한 것은 다음 장에서 기술하도록 한다.

III. 어휘별 중의성 제거 규칙

2장에서 기술하였듯이, 규칙 정보를 기반으로 한 품사 태깅의 연구는 통계 기반 품사 태깅의 연구보다 아직 부족하다. 이번 장에서는 통계 정보에 벗어나는 대표적인 언어 현상들의 예를 들고, 이것을 본 논문에서 제안하는 어휘별 중의성 제거 규칙으로 해결할 수 있음을 보인다.

1. 통계 정보에 벗어나는 언어 현상

본 논문에서는 "통계 정보에 벗어나는 언어 현상"을 "어절열을 확률 모델에 적용하여 품사 태깅을 하였을 경우 잘못된 품사 태깅 결과를 갖게 되는 경우"라고 정의한다.

한국어는 교착어라는 특성 때문에 통계 정보를 추출할 때 자료 부족 문제가 발생하게 된다. 따라서 같은 양의 말뭉치에서 추출한 통계 정보가 하나의 어절이 하나의 품사를 갖는 영어보다 신뢰도가 떨어지게 된다. 즉, 통계 정보에 벗어나는 언어 현상이 영어보다 발생할 확률이 높다.

본 논문에서는 통계 정보에 벗어나는 언어 현상을 살펴보기 위하여 세종 태그 부착 말뭉치에서 trigram 통계 정보를 추출하여, 어절 단위 HMM(Hidden Markov Model)을 이용하여 품사 태깅을 수행하였으며, 확률 모델은 식 (1)과 같다.

$$P(e_{1,N}) \cong \underset{c_{1,N}}{\arg max} \prod_{i=0}^{N} \Pr(c_i|c_{i-1},\ c_{i+1})\Pr(e_i|c_i) \quad - (1)$$

이것은 "$N(N \geq 0)$개의 어절로 구성된 문장($e_{1,N} = e_1 e_2 ... e_N$)에서 가장 확률이 높은 품사열($c_{1,N} = c_1 c_2 ... c_X$)을 구하는 것"을 의미한다.

다음의 문장들은 통계 정보에 벗어나는 언어 현상의 예이고, [표 1]는 중의성 어휘에 대한 통계 정보이다.

S1-1. 그것을 **알(알/NNG)** 방법이 없다.

S2-1. **두(두/NNG)** 사람은 사이가 좋지 않다.

S3-1. 눈 **먼(멀/VA+ㄴ/ETM)** 장님 먼 산 보기.

표 1. 중의성 어휘에 대한 통계 정보

중의성 어휘에 대한 통계 정보				
중의성 어휘	통계 정보			
	왼쪽 어휘	해당 어휘	오른쪽 어휘	확률 값
알	NP+JK	VV+ETM	NNG+JK	0.0
	NP+JK	NNG	NNG+JK	1.3×10^{-3}
두	FOS	NR	NNG+JX	0.0
	FOS	NNG	NNG+JX	6.0×10^{-3}
먼	NNG	VV+ETM	NNG	2.6×10^{-5}
	NNG	VA+ETM	NNG	1.2×10^{-2}

[표 1]의 통계 정보에서 각 품사열은 첫 번째가 중의성 어휘의 왼쪽에 나타나는 어휘에 대한 품사열이고 가운데는 해당 중의성 어휘의 품사열이며, 마지막이 중의성 어휘의 오른쪽에 나오는 어휘에 대한 품사열이 된다. FOS는 문장의 앞을 뜻하며, 각 품사열에서 사용되는 태그의 정보는 부록에 첨부하였다.

S1-1에서 중의성 어휘 '알'은 형태소 분석 결과가 '알/VV+ㄹ/ETM'와 '알/NNG'이고, S2-1의 중의성 어휘 '두'는 형태소 분석 결과가 '두/NR'와 '두/NNG'인 중의성 어휘들이다. 이런 어휘는 어휘의 다음에 일반명사가 나올 경우 문제가 된다. 왜냐하면, 한국어의 경우 'NNG NNG(+J_)'의 형태로 나열된 품사열이 'VV+ETM NNG(+J_)'나 'NR NNG(+J_)'의 형태로 나열된 품사열보다 확률 값이 높게 나오고([표 1] 참고), 그 결과 이런 형태의 품사열을 갖는 어휘열은 통계 정보에 의해 'NNG NNG(+J_)'로 태깅이 되게 된다. 하지만 S1-1의 '알'은 '알/VV+ㄹ/ETM'로 품사 태깅이 되어야 하고, S2-1의 '두'는 '두/NR'로 품사 태깅되어야 올바른 경우이다.

어휘 '먼'은 '멀/VV+ㄹ/ETM'와 '멀/VA+ㄹ/ETM'를 형태소 분석 결과로 갖는다. 이런 경우 통계 정보를 이용하여 얻은 확률 값에 따라서 품사가 결정되게 되는데 이 경우도 태깅 결과가 잘못될 가능성이 있다. S3-1은 두 개의 어휘 '먼' 중 첫 번째는 '멀/VV+ㄹ/ETM'로 두 번째는 '멀/VA+ㄹ/ETM'로 분석되어야 되는 경우인데, 두 가지 모두로 '멀/VA+ㄹ/ETM'로 품사 태깅되었다. 이것은 중의성 어휘 '먼'이 'VV'보다는 'VA'로 쓰이는 경우가 많기 때문이다.

위의 예와 같이 통계 정보에 벗어나는 언어 현상이 존재하고, 이것은 통계 정보를 기반으로 하는 품사 태깅 시스템에서는 해결할 수 없는 문제이다. 이것을 해결하기 위해 규칙 정보를 통계 정보가 적용되기 이전에 적용하여 중의성을 제거해야 하는 데, 기존의 연구로 위의 예를 모두 적용할 수 있는 규칙 정보가 없으며, 이를 해결하기 위해서는 어휘의 사전적 의미와 문맥 관계가 충분히 반영되어야 한다.

2. 어휘별 중의성 제거 규칙을 이용한 중의성 제거

어휘별 중의성 제거 규칙은 어휘의 사전적 의미와 문맥적 관계를 모두 고려하여 작성된 중의성 제거 규칙으로서 다음과 같은 정보를 포함한다.

1. 중의성 어휘의 앞 어휘에 대한 형태소와 품사 정보: 중의성 어휘가 문장의 가장 앞이라면 <fos> 값을 갖게 된다.
2. 중의성 어휘의 뒤 어휘에 대한 형태소와 품사 정보: 중의성 어휘가 문장의 가장 뒤라면 <eos> 값을 갖게 된다.
3. 정보 1, 2 대신 default 값이 오게 되면, 상위 규칙이 적용되지 않을 때의 기본값으로 적용되는 규칙이다.
4. 중의성 어휘의 품사 태깅 결과

여기서 정보 1, 2는 각각 0개 이상이 올 수 있고, 정보 3은 0개 또는 하나가 올 수 있다. 그리고 각 규칙은 중의성 어휘별로 하나 이상을 가질 수 있으며, 상위에 오는 규칙이 높은 우선순위를 갖는다. 이 규칙은 중의성 어휘의 앞 뒤 어휘에 대한 형태소와 품사 정보를 모두 고려하며 형태소와 품사 정보가 각 어절에 포함되어있을 경우 적용되어 어휘의 사전적 의미를 충분히 고려할 수 있고, 중의성 어휘에 대하여 앞 뒤 어휘의 개수를 충분히 살펴보아 문맥 관계를 충분히 고려할 수 있게 된다. 또한, 어휘별 중의성 제거 규칙은 긍정 정보이기 때문에 품사 태깅의 결과가 결정적이고 신뢰성이 높다.

다음 표 2은 앞 절의 문장들에 대하여 어휘별 중의성

제거 규칙을 적용하여 올바르게 품사 태깅된 결과를 보여준다.

표 2. 어휘별 중의성 제거 규칙을 이용한 중의성 제거

어휘별 중의성 제거 규칙			
중의성 어절	왼쪽 어휘의 정보	오른쪽 어휘의 정보	태깅 결과
알	–	방법/NNG	알/VV+ㄹ/ETM
두	–	/NNG	두/NR
먼	[눈|귀]/NNG	–	멀/VV+ㄴ/ETM
규칙 적용 후 품사 태깅 결과			
S1-2. 그것을 알(알/VV+ㄹ/ETM) 방법이 없다.			
S2-2. 두(두/NR) 사람은 사이가 좋지 않다.			
S3-2. 눈 먼(멀/VV+ㄴ/ETM) 장님 먼 산 보기.			

S1-1, S2-1는 앞 절에서도 기술하였듯이, 통계 정보를 이용한 품사 태깅을 할 경우 조사가 붙지 않은 명사 다음에 명사가 오는 경우가 높은 확률 값이 적용되기 때문에 나타나는 결과였으며, S3-1의 경우는 형용사와 동사 둘 다 될 수 있는 어휘에서 어느 한쪽의 확률 값이 더 높을 경우 발생하는 품사 태깅 오류였다. 이러한 오류는 각 중의성 어휘의 사전적 의미를 문맥을 통하여 알 수 있어야 정확하게 태깅을 할 수 있으며, [표 3]에서와 같이 앞 어휘들과 뒤 어휘들에 대한 형태소나 품사 정보를 통해서 파악하거나, 어휘들 간의 문맥 관계를 통해서 파악할 수 있게 된다.

IV. 품사 태깅 시스템의 설계

본 논문의 품사 태깅 시스템은 기존의 복합적 품사 태깅 시스템에 어휘별 중의성 제거 규칙을 적용하여 구축되었으며, 어휘별 중의성 제거 규칙은 통계 정보 이전에 적용되어 품사 태깅을 실시한다. [그림 1]은 본 논문의 품사 태깅 시스템 구조를 나타낸다.

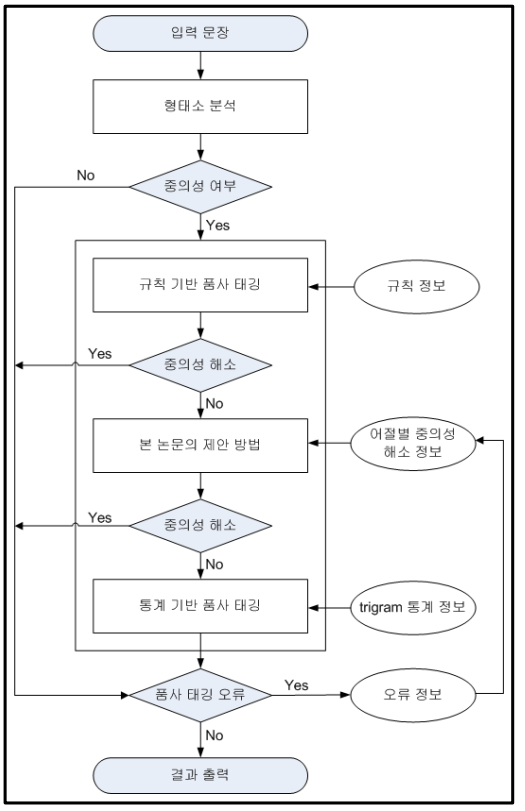

그림 1. 품사 태깅 시스템의 구조

1. 규칙 정보의 구축

본 논문의 품사 태깅 시스템에서 사용된 규칙 정보는 규칙 정보와 수정 정보로 구분된다.

이 중 규칙 정보는 보조용언의 구성 원리, 관용구 정보, 양태, 연어 정보 등을 이용하여 규칙으로 작성한 것을 말하며, 수정 정보는 중의성이 없는 형태소 분석 결과 중 오류가 있는 어절의 정보를 수집하여 구축한 것이다.

본 논문의 품사 태깅 시스템에서 사용된 규칙 정보는 총 44개로 이루어져 있다. 그리고 이것은 22개의 '어미-보조용언' 규칙, 20개의 '어미-명사(조사)-용언' 규칙, 그리고 기타 2개의 규칙으로 구성되어 있다. 다음 [표 3]은 그 규칙의 예이다.

표 3. 규칙 정보의 일부분

어미-보조용언 규칙	어미-명사(조사) 용언 규칙
게_되	을_수가_있
게_하	을_수가_없
고_나가	을_필요가_있
…	…

2. 통계 정보의 구축

통계 정보를 이용한 품사 태깅 시스템은 확장성이 높아 견고한 시스템을 구축할 수 있는 장점이 있으나, 신뢰성 있는 말뭉치 구축에 많은 시간과 노력이 든다는 단점이 있다. 하지만 21세기 세종 계획의 결과, 말뭉치로부터 어느 정도 신뢰성 있는 통계 정보를 추출할 수 있게 되었다.

본 논문의 품사 태깅 시스템에서는 21세기 세종 계획의 천만 어절 균형 말뭉치로부터 추출한 모든 중의성 어절에서 출현 빈도수가 2 이상이며, 음절수가 5음절 이하인 중의성 어절 19,506개와 태그 부착 말뭉치에서 추출한 어절 단위 품사 태그열 7,418개를 이용하여 통계 정보를 구축하였다. 이 통계 정보는 trigram 형식으로서, 한 어절의 앞, 뒤 어절에 대한 품사 태그열 정보를 추출하고, 중의성 어절에 대한 통계 정보와 품사 태그열에 대한 통계 정보로 구성되어 있다. [표 4]는 중의성 어절 '먼'에 대한 통계 정보의 일부이다.

표 4. trigram 통계 정보의 일부분

중의성 어절 '먼'의 통계 정보			
앞 어절의 품사열	해당 어절의 품사열	뒤 어절의 품사열	통계 정보
…			
VX+EC	VV+EC	VX+EP+EF	1
…			

어절 품사열 'VV+EC'에 대한 통계 정보		
앞 어절의 품사열	뒤 어절의 품사열	통계 정보
…		
NNG	MM	789
…		

[표 4]와 같이 추출된 통계 정보는 어절 단위 HMM(Hidden Markov Model)을 이용하여 품사 태깅에 사용된다.

3. 어휘별 중의성 제거 규칙의 구축

어휘별 중의성 제거 규칙은 다음과 같이 구축되었다. 세종 천만 어절 균형 말뭉치를 본 연구실의 형태소 분석 시스템인 CBKMA v3.1로 형태소 분석을 실시하고, 여기서 모든 중의성 어절 396,454개를 추출한 후, 전체 중의성 어절 발생 빈도의 약 50.7%에 해당되는 상위 500개 어절에 대하여 1,348개의 1차 어휘별 중의성 제거 규칙을 구축하였다. 이 규칙은 국립국어원의 표준국어대사전과 민중국어사전에 공통적으로 제시된 사전적 의미 및 발생 형태와 문맥 정보를 이용하여 구축했으며, 세종 천만 어절 균형 말뭉치에서 발생한 빈도수를 기준으로 우선순위를 부여하여 품사 태깅에 적용된다. 그리고 1차 어절별 중의성 제거 규칙을 적용한 복합적 품사 태깅 시스템을 이용하여 현대소설, 수필, 뉴스, 희곡, 인터넷 신문 기사 등의 내용을 포함한 2만 어절 학습 말뭉치 5셋을 학습한 후 467개의 2차 어휘별 중의성 제거 규칙을 구축하였으며, 기존의 규칙들을 수정 및 보완하였다. [표 5]는 어휘별 중의성 제거 규칙의 일부분이다. '@'은 중의성 어휘의 위치이며, ':'으로 구분된 '[]'안의 어휘들은 중의성 어휘의 앞 또는 뒤에 올 수 있는 어휘들을 의미한다. default는 앞의 우선순위가 높은 규칙이 적용되지 않았을 때 적용되는 품사 태깅 결과를 나타낸다.

표 5. 어휘별 중의성 제거 규칙의 일부분

중의성 어휘	중의성 제거 정보	품사 태깅 결과
진	@ [밥:안주:죽]/NNG	질/VA+ㄴ/ETM
우리	@ [안:속:밖]/NNG	우리/NNG
	@ /NR [개]/NNB	우리/NNG
	@ /NR [명]/NNB	우리/NP
	[개:소:닭:토끼:돼지]/NNG @	우리/NNG
	default	우리/NP

V. 실험결과

1. 분석 결과

본 논문에서는 어휘별 중의성 제거 규칙을 적용하지

않은 시스템과 적용한 시스템으로 정확도를 실험하였다. 실험 대상은 세종 원시 말뭉치 데이터에서 임의로 1,000개의 중복되지 않은 문장을 추출하였다. 추출된 문장은 충북대학교 자연언어처리 연구실의 형태소 분석기인 CBKMA v3.1로 형태소 분석을 하였으며, 띄어쓰기 오류, 복합명사 및 미등록어는 정확도에서 제외하였다. 이를 제외한 후 총 어절 수는 14,095개이다. [표 6]의 실험 결과는 형태소 분석 오류는 포함하지 않은 어절 단위의 결과를 나타낸다. 여기서 '기존 복합적 품사 태깅 방법에 대한 오류 감소'는 기존의 복합적 품사 태깅 방법에서 품사 태깅이 잘못된 어절과 그것에 대하여 어휘별 중의성 제거 규칙을 적용하였을 때 품사 태깅이 바르게 된 어절의 비율이고, '어휘별 중의성 제거 규칙으로 새롭게 발생한 오류'는 전체 실험 어절에 대하여 어휘별 중의성 제거 규칙으로 인하여 새롭게 발생한 오류의 비율이다.

표 6. 품사 태깅 실험 결과

실험 말뭉치(어절수) ※ 띄어쓰기 오류, 복합명사, 미등록어 제외	14,095
어휘별 중의성 제거 규칙 적용 전 정확도(%) $= \frac{\text{올바른 품사태깅 결과 어절수}}{\text{전체 어절수}} \times 100$	94.79
어휘별 중의성 제거 규칙 적용 후 정확도(%) $= \frac{\text{올바른 품사태깅 결과 어절수}}{\text{전체 어절수}} \times 100$	98.28
기존 복합적 품사 태깅 방법에 대한 오류 감소(%) $= \frac{\text{본 논문의 방법으로 해결된 오류}}{\text{복합적 방법의 오류}} \times 100$	74.39
어휘별 중의성 제거 규칙으로 새롭게 발생한 오류(%) $= \frac{\text{본 논문의 방법으로 발생한 오류}}{\text{전체 어절수}} \times 100$	0.39

실험 결과, 어휘별 중의성 제거 규칙을 적용하기 전은 정확도가 94.79%가 나왔고, 어휘별 중의성 제거 규칙을 적용하였을 때의 정확도는 98.28%로 나와 약 3.49%의 정확도 향상이 있었으며, 어휘별 중의성 제거 규칙에 의하여 통계 정보에서 발생한 오류를 74.39% 줄일 수 있어, 기존의 복합적 방법에서 발생한 오류를 상당히 줄일 수 있음을 알 수 있다.

[표 7]은 기존의 연구와 정확도를 비교한 것이며, 성능이 각 시스템의 품사 집합 수나 실험 말뭉치에 따라 다르게 나타날 수 있으므로 객관적인 비교는 힘들지만, 본 논문에서 제안한 방법을 적용하였을 때의 정확도가 다른 시스템에 비하여 높음을 알 수 있다. 여기서 [4]의 방법은 특정 어절에 대하여 측정한 것으로 비교 대상이 되지 않아 제외하였다.

표 7. 타 시스템과의 성능 비교

비교 시스템	정확도
안영민의 방법[2]	약 94%
김영길의 방법[3]	약 97%
도미숙의 방법[5]	약 92%
신상현의 방법[6]	약 93%
심준혁의 방법[8]	약 97%
임희석의 방법[7, 9]	약 95%
본 논문의 방법	약 98%

2. 오류 분석

어휘별 중의성 제거 규칙으로 기존의 복합적 품사 태깅에서 발생한 오류를 감소시킬 수 있었다. 하지만 오히려 어휘별 중의성 제거 규칙으로 새롭게 발생한 오류가 전체 어절 수에 대해서 0.39%였으며, 그에 대한 원인은 다음과 같다.

1. 규칙에 영향을 미치는 어절에 오류가 있을 경우.
2. 규칙의 우선순위가 잘못된 경우.
3. 잘못된 규칙

그리고 어휘별 중의성 제거 규칙으로도 기존의 복합적 품사 태깅에 대한 오류를 해결하지 못한 경우도 있었으며 그 원인은 다음과 같다.

4. 통계 정보에 벗어나는 언어 현상을 일으키는 어휘에 대한 규칙이 없는 경우
5. 기존의 복합적 품사 태깅 시스템의 규칙 정보가 잘못된 경우.

원인 1이 발생하는 경우는 대부분 미등록어, 복합명사 및 오타 등에 의해서 잘못된 형태소 분석 결과 로 인하여 발생했으며, 이것은 미등록어나 복합명사를 처리하면 오류를 감소할 수 있다. 원인 2의 경우는 어휘별 중의성 제거 규칙의 우선순위를 바꾸어 주면 대부분 해소가 되나, 우선순위를 부여하기 힘든 규칙들 간에는 문제가 발생할 가능성이 있으며, 규칙의 우선순위에 대한 연구가 필요하다. 원인 3의 경우는 수작업을 통한 사람의 실수로 발생한 오류로서 어휘별 중의성 제거 규칙을 관리할 수 있는 워크벤치 등의 구현을 통해서 오류를 어느 정도 감소할 수 있을 것이라 본다. 원인 4의 경우는 해당 어휘에 대한 어휘별 중의성 규칙을 추가하는 방법이 있으며, 원인 5의 경우는 복합적 태깅 시스템의 잘못된 규칙을 수정함으로서 해결이 가능하다.

VI. 결론

품사 태깅은 다른 자연언어처리 및 응용분야의 전처리로 사용되기 때문에, 품사 태깅의 정확도는 매우 중요한 문제가 아닐 수 없다. 품사 태깅에서 발생한 작은 오류 하나가 다른 언어처리 시스템의 성능에 크게 영향을 미치게 되므로 품사 태깅의 정확도가 100%에 준하는 성능이 될 때까지 계속적인 노력이 필요하다. 최근의 연구들은 말뭉치를 이용하여 품사 태깅 시스템을 구축하는 것이 용이하기 때문에, 대부분 통계 정보 및 규칙 정보를 말뭉치를 이용하여 추출한다. 하지만 이러한 방법들은 통계 정보에 벗어나는 언어 현상에 대하여 중의성 어휘의 사전적 의미와 문맥 관계를 충분히 고려한 규칙 정보를 적용한 품사 태깅이 힘들며, 따라서 여러 언어 현상에 적용 가능한 규칙 정보에 대한 연구가 필요하다. 이를 위해 본 논문에서 제안한 “어휘별 중의성 제거 규칙”은 좋은 방법론이 될 수 있을 것이라 생각하며, 품사 태깅 시스템의 더 높은 정확도를 위하여 규칙의 추가와 오류 수정을 통해 더욱 충실한 어휘별 중의성 제거 규칙을 구축해야 할 것이다.

VII. 부록

표 8. 본 논문에서 사용된 품사의 태그셋

대분류	중분류	의 미
N_ (명사류)	NNG	일반명사
	NP	대명사
	NNB	의존명사
	NR	수사
V_ (용언류)	VV	동사
	VA	형용사
	VX	보조용언
	VCP	지정사
MA_ (부사류)	MAG	일반부사
	MAJ	접속부사
J_ (조사류)	JK	격조사
	JC	접속조사
	JX	보조사
	JKG	속격조사
E_ (어미류)	EC	연결어미
	EF	종결어미
	ETM	관형형전성어미
	ETN	명사형전성어미
	EP	선어말어미
XS_ (접미사류)	XSV	동사파생접미사
	XS	동사파생접미사를 제외한 접미사
	XP	접두사
	MM	관형사
	IC	감탄사
	S	기호
	SL	외국어
	NF	미등록어, 명사추정범주

참 고 문 헌

[1] B. Eric, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," Computational Linguistics, Vol.21, No.4, pp.543-564, 1995.

[2] 안영민, 서영훈, “조사와 어미의 문법 기능을 활용한 품사 태깅 시스템”, 제13회 한글 및 한국어 정

보처리 학술대회 논문지, pp.97-100, 2001.
[3] 김영길, 양성일, 홍문표, 박상규, “형태소 어휘 문맥에 기반한 태깅 오류 정정”, 제15회 한글 및 한국어 정보처리 학술대회 논문지, pp.63-68, 2003.
[4] 이충희, 윤준태, 송만석, “국소 문맥을 이용한 형태적 중의성 해소”, 제12회 한글 및 한국어 정보처리 학술대회 논문지, pp.48-55, 2000.
[5] 도미숙, 최호섭, 옥철영, “문법 규칙과 어절 상관도를 이용한 품사 태깅 시스템”, 제20회 한국정보처리학회 추계학술발표대회 논문집, 제10권, 제2호, pp.481-484, 2003.
[6] 신상현, 이근배, 이종혁, “통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템”, 한국 정보과학회 논문지(B), 제24권, 제2호, pp.160-169, 1997.
[7] 임희석, 김진동, 임해창, “어절 태그 변형 규칙을 이용한 한국어 품사 태거”, 한국 정보과학회 논문지(B), 제24권, 제6호, pp.673-684, 1997.
[8] 심준혁, 김준석, 차정원, 이근배, “통계와 규칙을 이용한 강인한 품사 태거”, 제11회 한글 및 한국어 정보처리 학술대회 논문집, pp.60-75, 1999.
[9] 임희석, 김진동, 임해창, “통계 정보와 언어 지식의 보완적 특성을 고려한 혼합형 품사 태깅”, 한국 정보과학회 논문지(B), 제25권, 제11호, pp.1705-1714, 1998.

저 자 소 개

안 광 모(Kwang-Mo Ahn) 준회원

- 2007년 2월 : 충북대학교 컴퓨터공학과(공학사)
- 2007년 3월 ~ 현재 : 충북대학교 컴퓨터공학과 석사과정

<관심분야> : 한국어 형태소분석 및 품사 태깅, 한국어 구문분석, 질의응답시스템

한 규 열(Kyou-Youl Han) 준회원

- 2007년 2월 : 충북대학교 컴퓨터공학과(공학사)
- 2007년 3월 ~ 현재 : 충북대학교 컴퓨터공학과 석사과정

<관심분야> : 한국어 형태소분석 및 품사 태깅, 정보검색, 기계번역, 질의응답시스템

서 영 훈(Young-Hoon Seo) 종신회원

- 1979년 ~ 1983년 : 서울대학교 컴퓨터공학과 졸업(학사)
- 1983년 ~ 1985년 : 서울대학교 컴퓨터공학과 졸업(석사)
- 1985년 ~ 1991년 : 서울대학교 컴퓨터공학과 졸업(박사)
- 1994년 ~ 1995년 : 미국 Carnegie Mellon 대학 기계번역센터 객원교수
- 1988년 ~ 현재 : 충북대학교 전기전자컴퓨터 공학부 교수

<관심분야> : 자연언어처리, 한국어 구문분석, 한영기계번역, 정보검색, 질의응답시스템