어절별 중의성 해소 정보를 이용한 품사 태깅의 성능 향상

박희근, 서영훈
충북대학교 컴퓨터공학과
pinetree@nlp.chungbuk.ac.kr, yhseo@chungbuk.ac.kr

Improving Part-of-speech Tagging by using Resolution Information for Individual Ambiguous Word

Hee-Geun Park, Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University

요 약

품사 태깅 시스템에서 규칙 정보와 통계 정보는 상호보완적으로 사용되어 품사 태깅의 성능을 향상시킨다. 하지만, 두 가지 정보로는 품사 태깅의 성능을 향상시키기에는 한계가 있다. 이에 본 논문에서는 어절별 중의성 해소 정보를 이용하여 품사 태깅 시스템의 정확률을 향상시키는 방법에 대해서 기술한다. 통계 정보는 21세기 세종계획의 천만 어절 균형 말뭉치와 태그 부착 말뭉치에서 추출한 trigram 형태의 중의성 어절 및 품사 태그열 출현 빈도 정보를 이용하여 구축하였고, 규칙 정보는 보조용언, 숙어, 관용적 표현 등을 이용하여 구축하였다. 어절별 중의성 해소 정보는 세종 천만 어절 균형 말뭉치의 중의성 어절에서 고빈도 상위 50%에 해당하는 어절을 대상으로 해당 어절의 의미정보와 문맥정보를 고려하여 구축되었고, 이것은 통계 정보를 이용한 품사 태깅 전에 적용되어 분석 후보를 줄여준다. 또한, 학습을 통하여 어절별 중의성 해소 정보를 수정 및 보강하여 잘못된 품사 태깅 결과를 보정해준다. 이와 같이 통계 정보와 규칙 정보를 이용한 품사 태깅 시스템에 고빈도 중의성 어절에 대한 어절별 중의성 해소 정보를 이용함으로써 품사 태깅의 성능을 향상시킬 수 있었다.

1. 서론

한국어의 정보처리과정에서 형태소 분석과 품사 태깅은 가장 기본이 되는 처리과정이다. 형태소 분석은 어절 단위로 분리된 입력 문장에 대하여 각 어절이 분석할 수 있는 모든 품사 정보를 나타내어 주는 과정이다. 각 어절의 형태소 분석 결과는 여러 개의 분석 결과를 가질 수 있으며, 이를 중의성(重義性, ambiguity)을 갖는 어절이라 한다. 이런 중의성을 갖는 형태소 분석 어절을 구문 분석, 정보 검색, 기계 번역 등의 여러 자연언어처리 응용 분야에 사용한다면, 정확한 결과를 얻기 힘들다. 예를 들면, 어절 '나는' 은 '나[대명사]+는[조사]' 와 '나[동사]+는[어미]' 그리고 '날[동사]+는[어미]' 의 형태로 형태소 분석된다. 문장 "나는 하늘을 나는 새를 보았다." 에서 첫 번째 어절은 '나[대명사]+는[조사]' 의 형태소 분석 결과를 선택해야 하고, 세 번째 어절은 '날[동사]+는[어미]' 의 형태소 분석 결과를 선택하여야 한다. 이처럼 형태소 분석 결과의 중의성을 해소하고, 문맥에 맞는 적합한 품사를 결정하는 작업을 품사 태깅(part-of-speech tagging)이라고 한다.

이렇게 형태소 분석 결과에 대해 품사 태깅을 함으로써 형태소 분석 결과를 이용하는 자연언어처리응용 분야의 시스템에서 보다 정확한 결과를 얻을 수 있다. 또한, 품사 태깅의 성능이 다른 응용 시스템의 성능에 영향을 미치기 때문에 품사 태깅의 성능을 향상시킬 필요가 있다.

기존의 품사 태깅 방법에는 통계적 접근법[1-6]과 규칙 기반 접근법[7-9] 및 복합적 접근법[10-18]이 있다.

통계적 접근법은 대량의 태그 부착 말뭉치로부터 추출한 통계 정보를 이용하여 품사 태깅을 수행하기 때문에, 확장성이 좋고 적용범위가 넓으며 전체적인 정확성이 비교적 높다는 장점이 있다. 그러나 말뭉치에 의존적이고, 의미 있는 통계 정보를 추출하기 위해서는 일정 크기 이상의 태그 부착 말뭉치가 구축되어 있어야 하기 때문에, 말뭉치 구축에 시간과 노력이 많이 요구되고, 말뭉치가 편중되어 있거나 불충분한 경우에는 통계 자료 부족(data sparseness)으로 인하여 신뢰도가 떨어진다는 단점이 있다.

이와 달리 규칙 기반 접근법은 언어 정보를 생성 규칙의 형태로 표현하여 품사 태깅을 수행하기 때문에,

규칙이 적용되는 언어 현상에 대해서는 높은 정확도를 보이지만, 규칙으로 해결하지 못하는 예외적인 언어 현상이 존재하고 규칙의 구축과 관리에 많은 시간과 노력이 요구되는 단점으로 인해 처리 범위가 넓지 못하다. 한국어의 경우 규칙으로만 해결하기 어려운 언어 현상이 많이 존재하기 때문에 규칙적 접근법만을 이용하여 한국어 품사 태깅에 적용한 사례는 드물다.

최근의 품사 태깅에 대한 연구는 복합적 접근법을 이용한 품사 태깅이 주로 연구되고 있으며, 이는 높은 정확도와 넓은 적용 범위라는 규칙 기반 접근법과 통계적 접근법의 장점을 동시에 만족시키기 위해서이다.

지금까지 연구된 복합적 접근법을 이용한 품사 태깅에 사용된 규칙 정보는 보조용언, 숙어, 관용구 등의 정보를 이용한 것이다. 예를 들면, [17]에서 보조용언은 앞 어절의 마지막 품사가 '어', 게', '지', '고' 등의 연결어미일 경우에만 사용되어야 한다는 것과 [15]에서 'ㄹ/을 수~ 있/없' 등의 관용구 정보가 사용된 것이다. 또한, 대부분의 규칙 정보는 통계 정보를 이용하여 품사 태깅된 결과에 대한 수정 정보(correctional information)를 규칙으로 구축하여 품사 태깅에 이용한 것이다[12, 14-18]. 통계 정보의 경우에는 21세기 세종계획의 천만 어절 균형 말뭉치와 태그 부착 말뭉치가 발표되기 전에는 한국전자통신연구원의 29만 어절 태그 부착 말뭉치를 사용하였다. 이 말뭉치의 경우 크기가 충분히 크지 않기 때문에 신뢰도가 높다고 할 수 없었다.

이에 본 논문에서는 복합적 접근법을 이용한 품사 태깅의 정확률 향상을 위하여 세종 천만 어절 균형 말뭉치로부터 추출한 trigram 통계 정보와 일반적인 보조용언, 숙어, 연어, 관용구 정보를 이용한 규칙 정보 외에 어절별 중의성 해소 정보를 이용하였다. 어절별 중의성 해소 정보는 세종 천만 어절 균형 말뭉치의 중의성 어절에서 고빈도 상위 50%의 어절을 대상으로 해당 어휘의 사전적 의미와 문맥적 관계정보를 이용하여 구축하였다.

본 논문에서 제안하는 어절별 중의성 해소 정보는 규칙 정보가 적용된 후에 적용되며, 통계 정보가 적용된 품사 태깅 결과를 이용하여 학습한 후 새로운 정보를 추가적으로 생성한다.

2. 규칙 정보 및 통계 정보의 구축

본 논문에서 제안하는 어절별 중의성 해소 정보를 이용하여 품사 태깅을 향상시키기 위해서 복합적 접근법을 이용한 품사 태깅 시스템을 구현하였다. 이 장에서는 이 시스템에 사용된 규칙 정보와 통계 정보에 대하여 기술한다.

2.1 규칙 정보

본 논문의 실험을 위해 구현된 복합적 접근법을 이용한 품사 태깅 시스템에 입력된 문장의 형태소 분석 결과는 가장 먼저 규칙 정보에 의해 품사 태깅이 실시된다. 규칙 정보는 한국어에서 사용되는 보조용언의 구성 원리나 숙어, 관용구 정보 또는 양태, 연어 정보 등을 이용하여 규칙으로 작성한 것을 말한다. 규칙 정보의 예로는 '~ㄹ/을 수 있/없다', '~하지 않을 수 없다' 와 같은 관용구 정보를 이용한 것이 있다. 그리고 '~게 되다' 와 같이 '되다' 라는 보조용언의 앞에 '게' 라는 어미정보를 이용하여 중의성을 가지는 '되다' 라는 앞 어절은 [본용언]+[어미]로 이루어져 있음을 추측하여, 어휘 '되다' 가 [보조용언]으로 사용되었음을 알 수 있는 보조용언의 구성 원리를 이용한 것이 있다.

본 논문의 실험을 위해 구현된 시스템에 사용된 규칙 정보는 총 44개로 이루어져 있으며, 세부적으로 어절들의 출현 형태에 따라 어미-보조용언의 쌍으로 분석되는 22개의 규칙, 어미-명사(조사)-용언의 쌍으로 분석되는 20개의 규칙, 그리고 기타 2개의 규칙으로 이루어져 있다.

	어미 보조용언(E_ VX)
1	게_되
2	게_하
3	고_나가
…	
	어미 명사(조사) 용언(E_ N_+(J_) V_)
1	을_수가_있
2	을_수가_없
3	을_필요가_없
4	을_필요가_있
…	
	어미 명사(조사) 용언(E_ N_+(J_) V_)
1	는_것_같
2	는_일_있
3	는_일_없
4	는_적_있
…	

그림1. 규칙 정보 구축에 사용된 연어, 숙어 정보의 예

2.2 통계 정보

본 논문의 실험을 위해 구현된 복합적 접근법을 이용한 품사 태깅 시스템에서 통계 정보는 품사 태깅에 가장 많은 영향을 주는 정보로 규칙 정보에 의해 품사 태깅이 되지 않은 모든 중의성 어절의 품사를 결정한다. 이 통계 정보는 21세기 세종계획의 천만 어절 균형 말뭉치에서 추출한 모든 중의성 어절에서 출현 빈도수가 2 이상이고 5음절 이하인 중의성 어절 19,506개와 태그 부착 말뭉치에서 추출한 어절 단위 품사 태그열 7,418개를 이용하여 구축되었다. 각 통계 정보는 trigram 형식으로, 하나의 어절을 중심으로 앞, 뒤 어절에 대한 품사 태그열 정보를 추출하였으며, 중의성 어절에 대한

통계 정보와 품사 태그열에 대한 통계 정보로 구성되었다. 이때, 중의성 어절에 대한 통계 정보 구축 시 해당 중의성 어절의 총 빈도수가 1인 경우는 확률 100%의 위험으로 인하여 통계 정보 구축 대상에서 제외하였다.

다음은 추출된 통계 정보의 일부분이며, 각각의 통계 정보는 중의성 어절이나 어절 단위 품사 태그열을 중심으로 앞 어절의 품사 태그열 정보, 뒤 어절의 품사 태그열 정보, 출현 빈도수의 순서로 구성되었다. 이렇게 추출된 통계 정보는 어절 단위 HMM(Hidden Markov Model)을 이용하여 품사 태깅에 사용된다. 추출된 통계 정보 중 EOS는 문장의 마지막을 의미하며, 나머지 품사 태그에 대한 정보는 부록으로 첨부한다.

중의성 어절 "하는"의 추출 정보		어절태그 "NNG+VCP+EC"의 추출 정보		
...		...		
NNG+JK NNB+VCP+EF	10	MM	,	6
NNG+JK NNB+JK	19	MM	?	1
NNG+JK NNB+JX	11	MM	EOS	1
NNG+JK NNG	16	MM	MAG	8
NNG+JK NNG+JK	30	MM	NNG	1
NNG+JK NNG+JX	10	...		
...				

그림2. 추출된 통계 정보의 일부분

3. 어절별 중의성 해소 정보

본 논문에서 구현된 복합적 접근법을 이용한 품사 태깅 시스템에서는 지금까지 연구되었던 기존의 품사 태깅 시스템과 유사한 방법으로 규칙 정보와 통계 정보가 구축되었고, 이 시스템의 정확률은 95~97%로 기존에 연구되었던 시스템의 정확률과 비슷하였다. 이에 본 논문에서는 품사 태깅의 정확률을 향상시키기 위하여 일반적으로 사용되는 규칙 정보와 통계 정보 이외의 정보를 사용하고자 어절별 중의성 해소 정보를 구축하였다.

어절별 중의성 해소 정보를 구축하기 위해서 먼저 품사 태깅의 정확률이 낮은 어절을 분석하였다. 그 결과, 자주 사용되는 중의성 어절에서 사용 빈도가 낮은 품사 태그열이나 통계적으로 출현 빈도가 낮은 품사 태그열에 대해서 비교적 낮은 품사 태깅 정확률을 보였다. 예를 들면, 중의성을 가지는 어휘 '지난'은 '지난[일반명사]'와 '지나[동사]+ㄴ[관형형어미]'의 두 가지 형태소 분석 결과를 갖는다. "… 특히 지난 4월 …"이라는 문장에서 기존 시스템의 규칙 정보와 통계 정보를 이용하면 중의성 어절 '지난'은 '지난[일반명사]'로 품사 태깅이 된다. 왜냐하면, 중의성 어절 '지난'의 앞 어절 '특히[접속부사]'와 뒤 어절 '4[수사]+월[의존명사]'를 이용하여 적용할 규칙 정보가 존재하지 않고, 이를 통계 정보를 적용하여 품사 태깅을 실시하면 품사 태그열 '[접속부사] [일반명사] [수사]'는 품사 태그열 '[접속부사] [동사]+[관형형어미] [수사]'보다 더 높은 확률을 가지기 때문이다. 하지만 중의성 어절 '지난'은 "… 특히 지난 4월 …"이라는 문장에서 의미상 또는 문맥상 '지나[동사]+ㄴ[관형형어미]'로 품사 태깅이 이루어져야 올바른 결과가 된다. 이와 같이, 규칙 정보와 통계 정보를 이용해서는 사전적 의미 또는 문맥적 관계를 고려해서 품사 태깅을 할 수 없는 상황이 발생한다.

따라서, 본 논문에서는 각 중의성 어절별로 사전적 의미와 문맥적 관계정보를 분석하여 일반적인 규칙 정보와 통계 정보 이외의 품사 태깅 정보를 구축하였고, 이를 어절별 중의성 해소 정보라고 명명하였다.

어절별 중의성 해소 정보의 구축 과정은 다음과 같다. 세종 천만 어절 균형 말뭉치를 본 연구실의 형태소 분석 시스템인 CBKMA V3.1로 형태소 분석을 실시한 결과에서 모든 중의성 어절 396,454개를 추출하였다. 이 중 전체 중의성 어절 발생 빈도의 약 50.7%에 해당되는 상위 500개의 어절에 대하여 1차적으로 품사 태깅에 적용되는 1,348개의 어절별 중의성 해소 정보를 구축하였다. 이 정보는 해당 중의성 어절의 어휘에 대하여 국립국어원의 표준국어대사전과 민중국어사전에 공통적으로 제시된 사전적 의미 및 발생 형태와 문맥 정보를 이용하였으며, 각 중의성 어절에 대한 중의성 해소 정보들은 세종 천만 어절 균형 말뭉치에서 발생한 빈도수 중심으로 우선순위를 부여받아 품사 태깅에 적용된다.

다음으로 현대소설, 수필, 뉴스, 희곡, 인터넷 신문기사 등의 내용을 포함한 2만 어절 학습 말뭉치 5SET를 이용하여 구현된 복합적 접근법을 이용한 품사 태깅 시스템에 어절별 중의성 해소 정보를 적용하여 품사 태깅을 실시하였다. 이러한 학습을 통하여 2차적으로 어절별 중의성 해소 정보 467개를 구축하였고, 기존에 작성된 정보들을 수정 및 보완하기도 하였다.

다음은 구축된 어절별 중의성 해소 정보의 일부분이며, 각 정보는 중의성 어절, 해당 중의성 해소 정보, 품사 태깅 결과의 순서로 작성되었다.

수	[ㄹ:을:는]/ETM @ [있:없]/VA	수/NNB
	[ㄴ:은:는]/ETM @	수/NNG
이	@ /NNB	이/NR
	@ [학년:학기]/NNG	이/NR
	@ /NNG	이/MM
한	[ㄴ:은:는]/ETM @	한/NNG
	@ [일]/NNG [두]/NR [나라]/NNG	한/NNG
	[이:그:저:어떤:이런:저런:다른]/MM @	한/NNG
	[중]/NNB @	한/NNG
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

그림3. 어절별 중의성 해소 정보의 일부분

각 어절별 중의성 해소 정보에서 @기호는 해당 중의성 어절의 위치를 나타내며, 대괄호([어휘1:…:어휘n]) 안의 어휘는 비교 시 차례대로 비교되며, default는 우선순위가 높은 규칙들이 전부 적용되지 않았을 경우 적용되는 품사 태깅 결과를 나타낸다. 또한, 각각의 품사 태그는 본 연구실의 형태소 분석 시스템인 CBKMA V3.1의 품사 태그 집합을 이용하였으며, 각 품사 태그에 대한 정보는 부록으로 첨부한다.

4. 품사 태깅의 성능 향상

본 논문에서 제안하는 품사 태깅의 성능 향상을 위한 방법은 기존에 연구되었던 복합적 접근법에 기반한 한국어 품사 태깅 시스템에서 사용된 규칙 정보와 통계 정보 이외의 정보를 이용하는 것이다. 그 정보가 3장에서 제안한 어절별 중의성 해소 정보이며, 이 정보는 다음과 같이 품사 태깅에 적용된다.

다음의 그림4는 품사 태깅의 과정을 나타낸다.

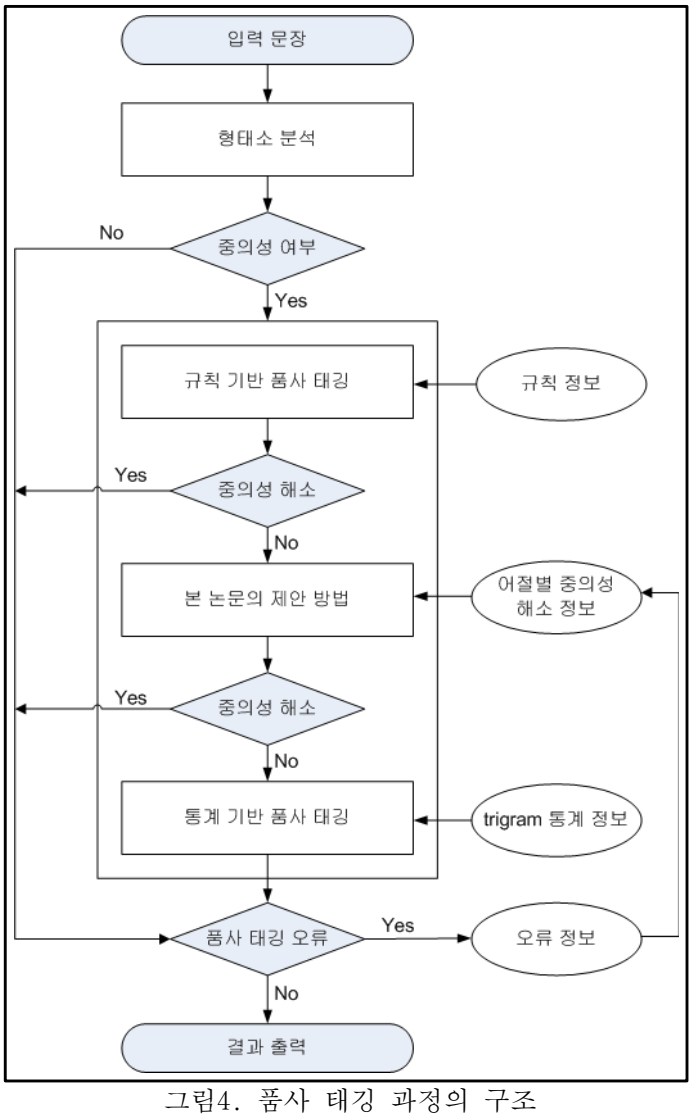

그림4. 품사 태깅 과정의 구조

입력 문장이 형태소 분석 시스템인 CBKMA V3.1에 의해 형태소 분석 단계를 거치고, 형태소 분석이 완료된 문장에서 중의성을 가진 어절은 품사 태깅의 과정을 거치게 된다. 품사 태깅은 1차적으로 규칙 정보에 의해 이루어지며, 규칙 정보에 의해 중의성이 해소 되지 않은 어절은 2차적으로 어절별 중의성 해소 정보에 의해 품사 태깅이 이루어진다. 규칙 정보와 어절별 중의성 해소 정보를 이용하여 중의성이 해소 되지 않은 어절은 마지막으로 통계 정보에 의해 품사 태깅이 이루어진다.

이런 과정을 학습 말뭉치를 통해 반복하면서 잘못된 품사 태깅 결과에 대한 오류 수정 정보를 구축한다. 오류 수정 정보는 어절별 중의성 해소 정보에 적용되어 품사 태깅의 성능을 향상시키게 된다.

5. 실험 및 결과

5.1 실험 대상

본 논문에서 제안하는 어절별 중의성 해소 정보의 구축을 위하여 21세기 세종계획의 천만 어절 균형 말뭉치를 사용하였고, 중의성 어절 중심의 통계 정보 구축을 위하여 21세기 세종계획의 천만 어절 균형 말뭉치, 어절 단위 품사 태그열 중심의 통계 정보 구축을 위하여 21세기 세종계획의 품사 태그 부착 말뭉치를 사용하였다.

어절별 중의성 해소 정보를 2차적으로 구축하기 위한 학습에 사용된 학습 말뭉치는 21세기 세종계획 사이트에 공개된 원시 말뭉치를 이용하여 2만 어절 단위로 중복된 문장이 없도록 하여 5SET의 학습 말뭉치를 구축하였으며, 학습 말뭉치의 분량은 총 9,342문장, 100,217어절이다. 학습 과정은 구축된 학습 말뭉치를 이용하여 구현된 품사 태깅 시스템으로 품사 태깅하고, 그 결과를 분석하여 오류 수정 정보를 구축하고, 이를 어절별 중의성 해소 정보에 적용하는 방식으로 진행되었다.

최종 실험은 품사 태깅에 사용된 규칙 정보와 통계 정보, 어절별 중의성 해소 정보를 구축하는데 사용되지 않고, 학습 말뭉치에도 사용되지 않은 문장을 대상으로 실험 말뭉치를 구성하였다. 실험 말뭉치의 내용은 현대소설, 수필, 뉴스, 희곡, 성경, 백과사전, 사설, 인터넷 신문기사 등으로 다양하게 포함하도록 하였으며, 총 1,914문장 20,598어절로 구성하였다.

5.2 실험 결과

실험 말뭉치를 본 논문에서 제안한 어절별 중의성 해소 정보를 적용한 복합적 접근법 기반의 품사 태깅 시스템을 이용하여 품사 태깅한 결과는 다음과 같다.

형태소 분석의 정확률은 99% 이상이었으며, 각 품사 태깅 정보별 중의성 해소율은 규칙 정보 적용 단계에서 약 18%, 통계 정보 적용 단계에서 약 46%, 어절별 중의성 해소 정보 적용 단계에서 약 36%를 보였다.

실험 말뭉치 (1,914문장)		20,598어절
형태소 분석 오류		86어절
어절별 중의성 해소 정보에 의한 품사 태깅 오류		7어절
통계 정보에 의한 품사 태깅 오류		366어절
정확률	형태소 분석 오류 포함	97.77%
	형태소 분석 오류 포함하지 않음	98.19%

그림5. 품사 태깅 실험의 결과

실험 결과를 토대로 본 논문에서 제안하는 어절별 중의성 해소 정보를 이용한 품사 태깅 시스템의 정확률이 기존에 연구되었던 통계 기반 품사 태깅 시스템[1-6]의 평균 정확률(약 95%)이나 혼합형 품사 태깅 시스템[10-18]의 평균 정확률(약 97%)보다 향상되었다는 것을 알 수 있다.

품사 태깅 오류는 주로 통계 정보를 이용하여 품사 태깅이 이루어진 중의성 어절에서 주로 발생하였다. 그 이유는 품사 태깅에 적용되는 trigram 통계 정보가 문맥상 관계 정보나 사전적 의미 정보가 고려되지 않았고, 품사 태그열 정보만을 고려하여 구축되었기 때문이다. 품사 태깅 오류의 종류를 살펴보면, 용언이 세 개 이상 연속해서 출현하는 경우와 한 어절에 대한 모든 형태소 분석 후보가 전부 같은 품사 태그열로 이루어진 경우가 대부분이었다. 그러나 본 논문에서 제안하는 어절별 중의성 해소 정보에 의한 품사 태깅 오류는 해당 정보에 대한 중의성 해소 과정에서 올바른 품사 태깅의 경우가 훨씬 더 많았기 때문에 생긴 오류이며, 그 수가 상당히 적기 때문에 품사 태깅의 전체적인 성능 향상을 위해 충분히 감수할 수 있는 정도의 것이다.

6. 결론 및 향후 연구

본 논문에서는 어절별 중의성 해소 정보를 이용하여 한국어 품사 태깅의 성능을 향상하는 방법에 대하여 제안하였다. 어절별 중의성 해소 정보는 기존에 연구되었던 혼합형 한국어 품사 태깅 시스템에서 사용되었던 규칙 정보와 통계 정보에서는 고려되지 않았던 해당 중의성 어절의 사전적 의미와 문맥 정보를 모두 고려하여 구축되었다. 실험 결과 본 논문의 제안 방법을 이용하여 기존 품사 태깅의 성능을 향상 시킬 수 있었다.

향후 연구로 많은 수의 올바른 품사 태깅을 위해 감수해야만 했던 어절별 중의성 해소 정보에 의한 품사 태깅 오류가 더 이상 발생하지 않도록 어절별 중의성 해소 정보를 더욱 견고하게 구축해야 한다. 또한, 대부분의 품사 태깅 오류가 통계 정보에 의해서 발생하고 있기 때문에 좀 더 신뢰도가 높은 태그 부착 말뭉치를 이용하여 통계 정보를 구축하는 것이 필요하다. 어절별 중의성 해소 정보가 견고해질수록, 통계 정보의 신뢰성이 높아질수록 품사 태깅의 성능은 향상될 것으로 기대된다.

7. 참고문헌

[1] 이하규, 김영택, “통계 정보에 기반을 둔 한국어 어휘 중의성 해소”, 한국통신학회 논문지, 제19권, 제2호, pp.265-275, 1994.

[2] 신중호, 한영석, 박영찬, 최기선, “어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사태깅”, 제6회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.389-394, 1994.

[3] 김재훈, 임철수, 서정연, “은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅”, 정보과학회논문지(B), 제22권, 제1호, pp.136-146, 1995.

[4] 김진동, 임희석, 임해창, “Twoply HMM: 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델”, 정보과학회논문지(B), 제24권, 제12호, pp.1502-1512, 1997.

[5] 강인호, 김재훈, 김길창, “최대 엔트로피 모델을 이용한 한국어 품사 태깅”, 제10회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.09-14, 1998.

[6] 강인호, 김도완, 이신목, 김길창, “어휘 정보의 자동 추출과 이를 이용한 한국어 품사 태깅”, 제11회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.117-122, 1999.

[7] Eric Brill, “A simple rule-based part-of-speech tagger”, Proc. of the 3rd Conference on Applied NLP, Trento, Italy, pp.153-155, 1992.

[8] Eric Brill, “Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging”, Proc. of the 3rd Workshop on Very Large Copora, pp. 1-13, 1995

[9] Atro Voutilainen, “A syntax-based part-of-speech analyser”, Proc. of the 7th Conference of the European Chapter of the ACL, pp.157-164, 1995.

[10] M Zhang, S. Li and T. Zhao, “Tagging Chinese Corpus Based on Statistical and Rule Techniques”, Proceedings of the Int. Conference on Computer Processing of Oriental Language (ICCPOL-97), pp.503-506, 1997

[11] 신상현, 이근배, 이종혁, “통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템”, 정보과학회논문지(B), 제24권, 제2호, pp.160-169, 1997.

[12] 임희석, 김진동, 임해창, “통계 정보와 언어 지식의 보완적 특성을 고려한 혼합형 품사 태깅”, 정보과학회논문지(B), 제25권, 제11호, pp.1705-1715, 1998.

[13] 심준혁, 김준석, 차정원, 이근배, “통계와 규칙을 이용한 강인한 품사태거”, 제11회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.60-75, 1999.

[14] 강유환, “어절간 주품사 정보와 제약 규칙을 이용한 이용한 통합 기반 한국어 품사 태깅 시스템”, 충북대학교 컴퓨터공학과 석사학위 논문, 2000.

[15] 임희동, “어절간 문맥 정보를 이용한 통합 기반 한국어 품사 태깅 시스템”, 충북대학교 컴퓨터공학과

석사학위 논문, 2001.
[16] 안영민, “문법 형태소를 이용한 통계 정보와 규칙에 기반한 한국어 품사태깅 시스템”, 충북대학교 컴퓨터공학과 석사학위 논문, 2002.
[17] 도미숙, 최호섭, 옥철영, “문법 규칙과 어절 상관도를 이용한 품사 태깅 시스템”, 제20회 한국정보처리학회 추계학술발표대회 논문집, 제10권, 제2호, pp.481-484, 2003.
[18] 이동훈, 강미영, 황명진, 권혁철, “규칙과 비감독 학습 기반 통계정보를 이용한 품사 태깅 시스템”, 한국컴퓨터 종합학술대회 2005 논문집, pp.445-447, 2005.

부록. CBKMA V3.1의 품사 태그 집합

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