어절별 중의성 해소 규칙을 이용한 혼합형 한국어 품사 태깅 시스템

(Korean Part-of-Speech Tagging System Using Resolution Rules for Individual Ambiguous Word)

박 희 근 † 안 영 민 ††
(Hee-Geun Park) (Young-Min Ahn)

서 영 훈 †††
(Young-Hoon Seo)

요 약 본 논문에서는 어절별 중의성 해소 규칙과 trigram 통계 정보를 이용하는 혼합형 한국어 품사 태깅 시스템에 대하여 기술한다. 어절별 중의성 해소 규칙은 중의성을 가지는 어절들 각각에 대해 정의된 중의성 해소 규칙으로, 현재 중의성을 가지는 어절의 50%에 대해 작성되어 있다. 본 논문의 태깅 시스템은 먼저 보조용언, 숙어, 관용적 표현 등에 해당하는 공통규칙을 적용하고, 그 후에 어절별 중의성 해소 규칙을 적용한다. 마지막으로 중의성이 해소되지 않은 어절은 각 어절을 중심으로 하는 trigram 통계 정보를 이용하여 중의성을 해소한다. 실험 결과는 본 논문에서 제안하는 어절별 중의성 해소 규칙과 trigram 통계 정보를 혼합하여 중의성을 해소 시키는 방법이 높은 정확률과 넓은 처리 범위를 가지고 있다는 것을 보여준다.

키워드 : 한국어 품사 태깅, 품사 태깅, 중의성 해소

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원 사업의 연구결과로 수행되었음. IITA-2006-(C1090-0603-0046)
· 이 논문은 2007 한국컴퓨터종합학술대회에서 '어절별 중의성 해소 규칙을 이용한 혼합형 한국어 품사 태깅 시스템'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 충북대학교 컴퓨터공학과
pinetree@nlp.chungbuk.ac.kr
†† 비 회 원 : 충북대학교 컴퓨터공학과
maniacan@nate.com
††† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수
yhseo@chungbuk.ac.kr

논문접수 : 2007년 9월 27일
심사완료 : 2007년 10월 29일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제13권 제6호(2007.11)

Abstract In this paper we describe a Korean part-of-speech tagging approach using resolution rules for individual ambiguous word and statistical information. Our tagging approach resolves lexical ambiguities by common rules, rules for individual ambiguous word, and statistical approach. Common rules are ones for idioms and phrases of common use including phrases composed of main and auxiliary verbs. We built resolution rules for each word which has several distinct morphological analysis results to enhance tagging accuracy. Each rule may have morphemes, morphological tags, and/or word senses of not only an ambiguous word itself but also words around it. Statistical approach based on HMM is then applied for ambiguous words which are not resolved by rules. Experiment shows that the part-of-speech tagging approach has high accuracy and broad coverage.

Key words : Part-of-speech tagging for Korean, Resolution of ambiguous word, Resolution rules

1. 서 론

한국어의 일반적인 자연언어처리과정은 형태소 분석, 구문 분석, 의미 분석, 화용 분석의 단계를 가진다. 이중에서 가장 기본이 되는 형태소 분석을 세분화하면 형태소 분석과 품사 태깅으로 나눌 수 있다. 형태소 분석 시스템은 입력 문장에 대하여 어절 단위로 분리하고, 각 어절이 가질 수 있는 모든 품사에 대한 정보를 분석하여 출력해주는 시스템이며, 이때 각각의 어절은 여러 개의 분석 결과를 가질 수 있다. 이 분석 결과를 구문 분석, 정보 검색, 기계 번역 등의 자연언어처리 응용분야에 적용할 경우, 여러 개의 분석 결과를 가지는 형태소 분석 결과의 중의성(重義性, ambiguity)으로 인해 정확한 결과를 얻기가 어려워진다는 문제점을 가지고 있다. 예를 들면, 어절 "나는"의 경우 문장의 쓰임에 따라 "나[대명사]+는[조사]"와 "나[동사]+는[어미]" 그리고 "날[동사]+는[어미]"의 형태로 분석될 수 있다. 문장 "나는 하늘을 나는 비행기를 보았다."에서 첫 어절의 "나는"은 "나[대명사]+는[조사]"의 형태로 분석되어야 하고, 3번째 어절 "나는"은 "날[동사]+는[어미]"의 형태로 분석되어야 한다. 이처럼 형태소 분석 결과의 중의성을 해소하고 문맥에 맞는 적합한 품사를 결정하는 작업을 품사 태깅(part-of-speech tagging)이라고 하며, 품사 태깅을 수행하는 시스템을 품사 태깅 시스템(part-of-speech tagging system)이라 한다.

기존의 품사 태깅 방법에는 통계 정보를 이용한 방법[1-4], 규칙 정보를 이용한 방법[5-6] 및 통계 정보와 규칙 정보를 혼합한 방법[7-14]이 있다.

통계 정보를 이용한 품사 태깅의 경우 수작업으로 품사 태깅된 대량의 태그 부착 말뭉치(Tagged Corpus)로

부터 추출한 통계 정보를 이용하여 품사 태깅을 수행하기 때문에, 확장성이 좋고 적용 범위가 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 떨어지는 단점이 있다.

이와 달리 규칙 정보를 이용한 품사 태깅 방법은 규칙이 적용되는 언어 현상에 대해 높은 정확률을 보이지만, 규칙으로 해결하지 못하는 예외적인 언어 현상이 존재하고 규칙의 구축과 관리에 많은 시간과 노력이 요구되는 단점으로 인해 처리 범위가 넓지 못하다. 한국어의 경우 규칙만으로 해결하기 어려운 언어 현상이 많이 존재하기 때문에 규칙만을 이용하여 한국어 품사 태깅에 적용한 사례는 드물다.

최근에는 규칙 정보와 통계 정보를 통합하여 높은 정확도를 갖고, 적용 범위가 넓은 혼합형 한국어 품사 태깅 시스템이 주로 연구되고 있다.

지금까지 연구된 혼합형 품사 태깅 시스템에 사용된 대부분의 규칙은 본 논문에서 제안하는 한국어 품사 태깅 시스템의 공통규칙 수준의 것이다. 예를 들면, [13]에서 규칙의 일부분인 보조용언은 앞 어절의 마지막 품사가 '어', '게', '지', '고' 등의 연결어미일 경우에만 사용되어야 한다는 것이고, [11]에서는 "~ㄹ/을 수~ 있/없~" 등의 관용구가 사용되었다. 이러한 규칙들은 본 논문에서 제안하는 어절별 중의성 해소 규칙과 같이 해당 중의성 어휘의 사전적 의미나 문맥 정보를 모두 고려한 수준의 것이 아니다. 또한 대부분의 규칙은 통계 정보를 통해 출력된 품사 태깅의 결과를 규칙을 이용하여 수정함으로써 오류를 줄여 정확률을 높이기 위해 사용된 것이었다[9,11-14].

이에 본 논문에서는 중의성을 가지는 어절에 대해 해당 어휘의 사전적 의미와 문맥적 관계정보를 이용하여 구축된 어절별 중의성 해소 규칙과 해당 어절의 출현 빈도와 앞, 뒤어절의 태그열 정보를 이용하여 구축된 trigram 통계 정보를 이용한 혼합형 품사 태깅 시스템을 제안한다.

본 논문에서 제안하는 시스템은 먼저 보조용언, 숙어, 관용어 등의 정보를 이용하여 작성된 공통 규칙을 적용하여 중의성을 해소한 다음, 어절별 중의성 해소 규칙을 적용하여 중의성을 해소한다. 그리고 마지막으로 중의성이 해소되지 않은 어절은 trigram 통계 정보를 이용하여 품사 태깅을 실시하게 된다.

2. 품사 태깅 정보의 구축

본 논문에서 제안하는 품사 태깅 시스템은 3가지 정보를 이용하여 중의성을 갖는 어절에 대해서 품사 태깅을 실시한다. 3가지 정보는 공통규칙, 어절별 중의성 해소 규칙, trigram 통계 정보이며, 이 장에서는 이들 3가지 정보에 대하여 기술한다.

2.1 공통규칙

공통규칙은 본 논문에서 제안하는 품사 태깅 시스템에서 가장 먼저 적용되는 품사 태깅 정보로 보조용언이나 숙어, 관용구 또는 양태, 연어 정보로 이루어진 규칙을 말한다. 이것은 영어의 조동사와 비슷한 의미를 지니고 있는 어휘들을 한국어에 적용하여 규칙으로 정리한 것이다. 예를 들면, "~할 수 있다"(조동사 can), "~할 수 없다"(조동사 can not), "~하지 않을 수 없다"(조동사 must)와 같이 관용적으로 쓰이는 어휘들을 말한다. 이렇게 연속적으로 사용되는 관용적 표현들을 이용하여 품사 중의성을 제거 할 수 있다. 또한, "~게 되다"와 같이 "되다"라는 보조용언의 앞에 "게"라는 어미정보를 이용하여 중의성을 가지는 "되다"라는 어절의 앞 어절은 [본용언]+[어미]로 이루어져 있음을 추측할 수 있고, 이에 따라 "되다"라는 어휘는 [보조용언]으로 사용되었음을 알 수 있게 된다.

따라서 본 논문에서는 이러한 연어, 숙어, 관용구, 보조용언 정보를 품사 태깅에 이용하여 형태소 분석 결과의 중의성을 해소하고자 한다. 어절들의 출현 형태에 따라 어미-보조용언의 쌍으로 나타나는 정보와 어미-명사(조사)-용언의 순서쌍으로 나타나는 정보를 구축하였다. 그 결과로 어미-보조용언의 쌍으로 분석되는 공통규칙이 22개, 어미-명사(조사)-용언의 쌍으로 분석되는 공통규칙이 20개, 그리고 2개의 기타 정보가 구축되었다.

2.2 어절별 중의성 해소 규칙

형태소 분석 시 여러 개의 분석 결과가 나오는 어절, 즉 여러 개의 품사로 분석될 수 있는 어절을 중의성 어절이라 한다. 기존의 혼합형 한국어 품사 태깅 시스템들은 공통규칙 수준의 규칙과 bigram이나 trigram 통계 정보를 이용하여 품사 태깅을 실시하였다. 그 결과, 자주 사용되는 중의성 어절에서 사용 빈도가 낮은 품사나 통계적으로 출현빈도가 낮은 품사 태그열에 대해서 상당히 낮은 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 공통규칙이나 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

```
        어미 보조용언(e_ px)
1       게_되
2       게_하
3       고_나가
...
        어미 명사(조사) 용언(e_ n_+(j_) p_)
1       을_수가_있
2       을_수가_없
3       을_필요가_없
4       을_필요가_있
...
        어미 명사(조사) 용언(e_ n_+(j_) p_)
1       는_것_같
2       는_일_있
3       는_일_없
4       는_적_있
...
```

그림 1 공통규칙의 작성에 사용된 연어, 숙어 정보의 예

[수사]+월[의존명사]"의 형태소 분석 결과 중의성이 존재하지 않고, 이를 한국전자통신연구원의 29만 어절 태그 부착 말뭉치에서 trigram 통계 정보를 적용하여 품사 태깅을 실시하면 품사 태그열 "[접속부사] [일반명사] [수사]"은 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, 공통규칙으로 중의성이 해소되지 않고 통계 정보를 이용해서는 사전적 의미 또는 문맥적 관계를 고려해서 품사 태깅을 할 수 없는 상황이 발생한다.

본 논문에서는 이러한 중의성 어절의 해소를 위해 각 중의성 어절별로 사전적 의미와 문맥적 관계정보를 분석하여 개별 규칙을 구축하였으며, 이를 어절별 중의성 해소 규칙이라 명하였다.

한국전자통신연구원에서 연구용으로 배포한 29만 어절태그 부착 말뭉치를 본 연구실의 형태소 분석 시스템인 CBKMA V3.0으로 형태소 분석을 실시한 결과에서 중의성 어절을 추출하였다. 16,684어절이 중의성 어절로 추출되었고, 그중 상위 300개의 어절이 전체 중의성 발생 빈도의 약 50%에 해당되었다. 이에 따라, 상위 300개의 어절에 대하여 1,347개의 어절별 중의성 해소 규칙을 작성하였다. 어절별 중의성 해소 규칙은 해당 어절의 사전적 의미 및 발생 형태와 문맥 정보를 이용하여 작성하였으며, 가장 많이 출현한 형태에 대한 규칙에 우선순위를 부여하였다.

어절별 중의성 해소 규칙은 1차적으로 적용된 공통규칙에 의해서 해소되지 않은 중의성 어절에 적용되며, 특별한 제약이 없는 경우에는 해당 중의성 어절을 기준으로 하여 앞, 뒤 한 어절을 주로 비교하며, 한 어절 이상

```
수      [ㄹ:을:는]/etm @ [있:없]/pa     수/nb
        [ㄴ:은:는]/etm @                 수/nc
이      @ /nb                           이/nn
        @ [학년:학기]/nc                 이/nn
        @ /nc                           이/mm
한      [ㄴ:은:는]/etm @                 한/nc
        @ [일]/nc [두]/nn [나라]/nc      한/nc
        [아:그:저:어떤:이런:저런:다른]/mm @ 한/nn
        [중]/nb @                       한/nn
우리    @ [안:속:밖]/nc                  우리/nc
        @ /nn [개]/nb                   우리/nc
        @ /nn [명]/nb                   우리/np
        [개:소:닭:토끼:돼지]/nc @우리/nc
        default                         우리/np
```

그림 2 구축된 어절별 중의성 해소 규칙의 예

을 비교하는 경우도 있다.

다음의 그림 2에 예로 나타낸 어절별 중의성 해소 규칙은 중의성 어절, 규칙, 품사 태깅 결과의 순서로 작성되어 있다. 규칙에서 @기호는 해당 중의성 어절을 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 CBKMA V3.0의 품사 태그셋을 이용하였으며, /etm은 관형형어미, /mm은 관형사, /nb는 의존명사, /nc는 일반명사, /np는 대명사, /pa는 형용사를 나타낸다.

2.3 trigram 통계 정보

trigram 통계 정보는 어절별 중의성 해소 규칙을 작성할 때 사용된 한국전자통신연구원의 29만 어절 태그 부착 말뭉치에서 중의성 어절을 대상으로 하였다. 각 통계 정보는 trigram 형식으로, 하나의 어절을 중심으로 앞, 뒤 어절에 대한 품사 태그열 정보를 추출하였다. 이렇게 추출된 통계 정보는 중의성 어절에 대한 통계 정보, 품사 태그열에 대한 통계 정보로 구성되어 있다.

다음의 표 1은 추출된 trigram 통계 정보의 예를 나타내며, 각각의 통계 정보는 해당 중의성 어절을 중심으로 하여 앞 어절의 품사 태그열 정보와 뒤 어절의 품사 태그열 정보 및 출현 빈도에 대한 정보로 구성되어 있다. 이렇게 추출된 trigram 통계 정보는 어절 단위 HMM을 이용하여 품사 태깅에 사용된다. 추출된 trigram 통계 정보 중 EOS는 문장의 마지막을 의미하고, 각 품사 태그는 /co는 지정사, /ec는 연결어미, /ef는 어말어미, /jc는 격조사, /jx는 보조사, /mag는 일반부사를 나타낸다.

3. 품사 태깅 시스템의 설계 및 구현

3.1 시스템 설계

표 1 추출된 trigram 통계 정보의 예

중의성 어절 "하는"의 추출 정보			어절태그 "nc+co+ec"의 추출 정보		
...			...		
nc+jc	nb+co+ef	10	mm	,	6
nc+jc	nb+jc	19	mm	?	1
nc+jc	nb+jx	11	mm	EOS	1
nc+jc	nc	16	mm	mag	8
nc+jc	nc+jc	30	mm	nc	1
nc+jc	nc+jx	10		...	
	...				

본 논문에서 제안하는 한국어 품사 태깅 시스템은 본 연구실이 보유하고 있는 형태소 분석 시스템 CBKMA V3.0을 이용하였다. CBKMA V3.0은 시스템 사전, 어미 사전, 조사 사전, 기분석 사전 등으로 구성되어 있으며, 품사 태그 집합은 26개로 이루어져 있다.

제안한 시스템은 어절 단위 품사 태깅 시스템으로써 태그 부착 말뭉치를 이용하여 구축된 어절별 중의성 해소 규칙과 trigram 통계 정보를 이용한 복합적 접근법(Hybrid Approach)에 기반한 한국어 품사 태깅 시스템이며, 2장에서 제안한 3가지 정보에 의해 품사 태깅을 실시하게 된다.

품사 태깅 과정을 살펴보면, 사용자에 의해 입력된 원시 문장이 CBKMA V3.0에 적용되어 형태소 분석 단계를 거치고, 형태소 분석이 완료된 문장 중 중의성을 가진 어절은 다음의 품사 태깅 과정을 거치게 된다. 중의성을 가진 어절은 먼저 공통 규칙에 적용되어 1차적으로 중의성을 해소하게 된다. 공통 규칙에 의해서 중의성이 해소 되지 않은 어절은 어절별 중의성 해소 규칙에 적용되어 2차적으로 중의성을 해소하게 되고, 마지막까지 중의성이 해소되지 않은 어절은 trigram 통계 정보를 적용하여 중의성을 해소함으로써 최종 품사 태깅 결과를 출력하게 된다. 또한, 결과를 통해 추출된 오류 정보를 이용하여 오류 수정 정보를 구축함으로써 시스템의 신뢰도를 높인다.

3.2 시스템 구성도

3.1절의 설계를 바탕으로 하여 시스템을 구성하면 그림 3과 같은 시스템 구성도를 갖게 된다.

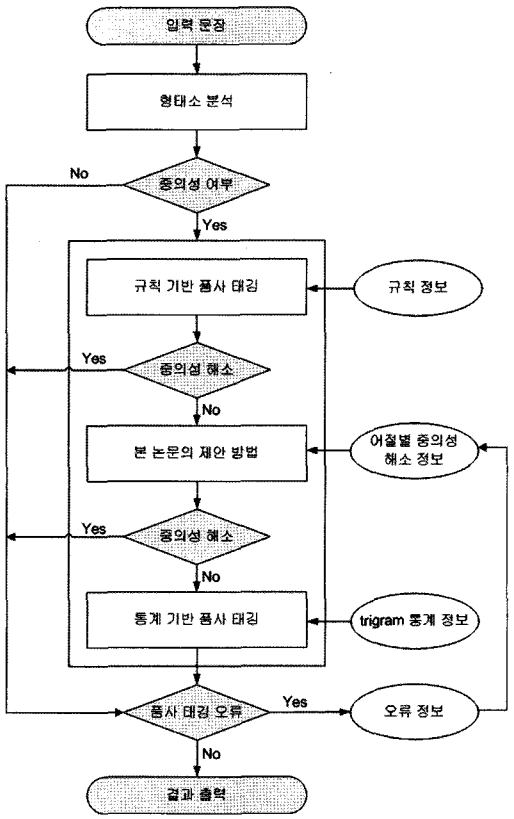

그림 3 제안한 한국어 품사 태깅 시스템 구성도

4. 실험 및 결과

4.1 실험 대상

제안한 한국어 품사 태깅 시스템에 필요한 어절별 중의성 해소 규칙과 trigram 통계 정보를 구축하기 위해서 한국전자통신연구원의 29만 어절 태그 부착 말뭉치를 학습 말뭉치로 사용하였다.

실험 말뭉치는 소설, 뉴스, 수필, 성경, 설명서 등의 내용을 포함하도록 하여, 한국전자통신연구원의 29만 어절 태그 부착 말뭉치의 원문에서 200문장과 세종 21계획의 결과로 제공받은 세종 천만어절 말뭉치의 원문에서 800문장을 추출하여 총 1,000문장, 17,384어절을 대상으로 하였다.

4.2 실험 결과

실험 말뭉치를 제안한 한국어 품사 태깅 시스템에 적용하여 품사 태깅을 실시한 결과는 표 2와 같다.

형태소 분석 결과와 품사 태깅 결과의 오류는 어절 단위로 판단하였으며, 정확률은 실험 말뭉치 전체 어절을 기준으로 계산하였다. CBKMA V3.0은 99% 이상의 정확률을 보이고 있으며, 실험 말뭉치에 대한 형태소 분석 오류는 127어절이었다. 또한, 품사 태깅 정확률에서 형태소 분석 오류를 고려하지 않는 경우는 형태소 분석 오류가 품사 태깅에 영향을 미치지 않는다는 가정을 한 결과로써 순수한 품사 태깅 결과의 정확률을 의미하고, 형태소 분석 오류를 고려한 경우는 형태소 분석 오류가 품사 태깅의 결과에 영향을 미친다고 가정하여 형태소 분석 오류도 품사 태깅의 일부라고 판단한 품사 태깅의 정확률을 의미한다.

표 2 제안한 한국어 품사 태깅 시스템의 실험 결과

실험 말뭉치 (1,000문장)		17,384어절
형태소 분석 오류		127어절
올바른 품사 태깅		17,151어절
잘못된 품사 태깅		233어절
정확률	형태소 분석 오류 포함	97.93%
	형태소 분석 오류 포함하지 않음	98.66%

각 단계별 중의성 해소율은 공통 규칙 적용 단계에서 약 18%, 어절별 중의성 해소 규칙 단계에서 약 41%이고, 나머지 중의성 어절은 trigram 통계 정보 적용 단계에서 해소된다.

실험 결과를 토대로 본 논문에서 제안하는 시스템의 정확률이 기존에 연구되었던 통계 기반 품사 태깅 시스템[1-4]의 평균 정확률(약95%)이나 혼합형 품사 태깅 시스템[7-14]의 평균 정확률(약97%)보다 높다는 것을 알 수 있다. 그러나 기존의 규칙 기반 품사 태깅 시스템들[5-6]은 한국어를 대상으로 연구가 이루어지지 않았기 때문에 본 논문의 시스템과 직접적인 비교가 어렵다.

품사 태깅 오류는 공통규칙과 어절별 중의성 해소 규칙으로 처리가 되지 않은 중의성 어절에서 주로 발생하였다. 그 이유는 해당 중의성 어절에 대한 규칙이 존재하지 않았기 때문이며, 마지막으로 품사 태깅에 적용되는 trigram 통계 정보에 의해서 주로 발생한다. 그 이유는 trigram 통계 정보가 문맥상 관계 정보나 사전적 의미 정보가 고려되지 않았고, 품사 태그열 정보만을 고려하여 구축되었기 때문이다.

5. 결 론

본 논문에서는 어절별 중의성 해소 규칙을 이용한 혼합형 한국어 품사 태깅 시스템을 제안하였다. 기존에 연구되었던 혼합형 한국어 품사 태깅 시스템의 규칙 정보와 다르게 해당 중의성 어절의 사전적 의미와 문맥 정보를 모두 고려하여 한국전자통신연구원의 29만 어절 태그 부착 말뭉치에 대한 어절별 중의성 해소 규칙을 구축하여 품사 태깅에 사용하였다. 그리고 한국어에서 많이 쓰이는 관용 표현을 이용하여 공통규칙을 구축하여 어절별 중의성 해소 규칙을 적용하기 이전에 중의성 해소에 적용하였다. 또한, 공통규칙과 어절별 중의성 해소 규칙으로 해결되지 않은 어절에 대해서는 trigram 기반으로 추출한 통계 정보를 이용하여 어절 태그열과 해당 어절의 발생확률을 계산하여 품사 중의성을 해결하였다. 실험 결과에 따르면 본 논문에서 제안한 시스템이 기존에 연구되었던 통계 기반 품사 태깅 시스템이나 혼합형 품사 태깅 시스템보다 높은 정확률을 보이는 것을 알 수 있다.

향후 연구로 공통 규칙과 어절별 중의성 해소 규칙의 확장 및 더욱 신뢰할 만한 통계 정보의 구축을 들 수 있다. 정확성이 높은 공통 규칙과 어절별 중의성 해소 규칙의 수가 늘어날수록, 통계 정보의 신뢰성이 높아질수록, 제안된 시스템의 성능은 높아질 것으로 기대된다.

참 고 문 헌

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

[6] 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.

[7] 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.

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

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

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

[11] 임희동, "어절간 문맥 정보를 이용한 통합 기반 한국어 품사 태깅 시스템", 충북대학교 컴퓨터공학과 석사학위 논문, 2001.

[12] 안영민, "문법 형태소를 이용한 통계 정보와 규칙에 기반한 한국어 품사태깅 시스템", 충북대학교 컴퓨터공학과 석사학위 논문, 2002.

[13] 도미숙, 최호섭, 옥철영, "문법 규칙과 어절 상관도를 이용한 품사 태깅 시스템", 제20회 한국정보처리학회 추계학술발표대회 논문집, 제10권, 제2호, pp. 481-484, 2003.

[14] 이동훈, 강미영, 황명진, 권혁철, "규칙과 비감독 학습 기반 통계정보를 이용한 품사 태깅 시스템", 한국컴퓨터 종합학술대회 2005 논문집, pp. 445-447, 2005.