어절간 주품사 정보와 제약 규칙을 이용한 한국어 품사 태깅 시스템

강 유 환† 서 영 훈††
충북대학교 컴퓨터공학과
† pingkey@dcenlp.chungbuk.ac.kr
†† yhseo@cbucc.chungbuk.ac.kr

Korean Part-of-Speech Tagging using Constrained-Rule and Main POS Information among Words

Yuhwan Kang Younghoon Seo
Dept. of Computer Engineering, Chungbuk National University

요약 본 논문에서는 품사 태깅을 위한 방법으로 어절간 품사 패턴 정보를 이용하는 방법을 제안한다. 품사 태깅을 위하여 여러 어절들 간의 품사 패턴 정보를 통계 정보로 구축하고 품사 태깅시에 품사 패턴 정보를 이용하여 품사 태깅을 수행한다. 이때 품사 패턴 적용시 몇가지 제약 규칙을 둠으로써 품사 태깅의 정확률을 높이는 방법을 연구하였다.

1. 서론

형태소 분석기는 입력된 단어를 모든 분석 가능한 형태로 출력해 줌으로써 형태소 분석기가 내준 결과에는 많은 중의성을 포함하는 어절을 갖게 된다. 따라서 형태소 분석기의 결과를 그대로 구문 분석이나 정보 검색 등에 이용할 경우 중의성 해결을 위한 오버로드가 들게 되므로 시스템이 복잡해지고 올바른 결과를 얻기가 어렵게 된다. 품사 태깅(Part-of-Speech Tagging) 시스템은 형태소 분석기가 내주는 여러 결과들 중 각 어절에 알맞은 결과를 선택하여 중의성을 해결해 주는 시스템으로써 품사 태깅 방법은 크게 규칙 기반 접근법(rule-based approach)과 통계 기반 접근법(stochastic approach)으로 구분된다. 규칙 기반 접근법은 각 어절에 적용될 수 있는 공통된 원리나 결정적 규칙을 찾아내고 이를 품사 태깅에 적용하는 방법이고, 통계 기반 접근법은 대량의 원시(raw) 또는 태그된(tagged) 코퍼스(corpus)로부터 추출한 확률 및 통계 정보를 이용하여 품사 태깅을 수행하는 방법이다. 규칙 기반 접근법은 규칙이 적용되는 어절에 대하여 거의 100%에 해당하는 정확률을 보이지만 규칙을 찾기가 어렵고 적용 범위가 넓지 못한 단점이 있는 반면에 통계 기반 접근법은 처리 범위가 넓지만 규칙 정보를 이용한 방법보다 정확률이 낮은 단점이 있다. 그렇기 때문에 규칙 기반 접근법과 통계 기반 접근법의 장점을 상호 보완함으로써 품사 태깅의 정확률을 높이려는 혼합 기반 접근법이 널리 쓰이고 있다. 통계 정보를 이용한 품사 태깅 방법에는 어휘 확률만을 이용하는 방법, HMM(Hidden Markov Model)의 자율 학습을 이용하는 방법, N-gram의 문맥 확률과 어휘 확률을 이용하는 방법이 있고 이외에도 신경망을 이용하는 방법과 퍼지망을 이용하는 방법 등이 있다. 이중 은닉 마르코프 모델은 통계 정보를 이용하는 방법 중 널리 쓰이는 품사 태깅 모델로 은닉 마르코프 모델을 이용한 품사 태깅법은 형태소와 각 형태소에 대응하는 품사 태그 및 품사 태그들 간의 통계 정보를 구축하고 이들에 대한 확률값을 가지고 품사 태깅을 수행하는 방법이다. HMM을 이용한 한국어 품사 태깅 시스템은 어절 단위 품사 태깅 시스템과 형태소 단위 품사 태깅 시스템으로 나눌 수 있으며, 어절 단위 품사 태깅 시스템은 문장 단위의 통계 정보 추출은 거의 불가능 하므로, 어절 단위로 통계 정보를 추출한다. 그러나 어절의 형태가 매우 다양하게 발생하는 한국어의 특성상 어절 단위의 통계 정보 획득에는 자료 부족 문제가 매우 심각하게 발생하게 된다. 형태소 단위 품사 태깅 모델은 형태소 단위의 통계 정보만을 필요로 하므로 어절 단위 품사 태깅 모델보다 자료 부족 문제가 심각하게 나타나지 않는다.

본 논문에서는 통계 정보 추출에 있어 나타나는 자료 부족 문제를 줄이고 어절 단위의 한국어 품사 태깅을 위한 새로운 시스템을 제안한다. 통계 정보를 어절 전체의 분석 결과를 가지고 추출하는 것이 아니라 각 어절에서 일부 품사만을 추출하는 방법을 사용하고, 여러 어절을 하나의 단위로 하여 품사 정보를 추출한다. 어절간에 추출한 품사 정보는 하나의 품사 패턴으로 만들어 지며, 이렇게 구축된 품사 패턴 정보를 품사 태깅에 적용할 수 있는 방법에 대하여 연구하였다. 2장에서는 품사 패턴 정보의 추출 방법에 대하여 설명하고 3장에서는 품사 패턴 정보를 이용한 품사 태깅 방법에 대해 설명한다. 4장과 5장에서는 각각 실험 결과와 결론 및 향후 연구 과제에 대하여 설명하겠다.

2. 어절간 품사 패턴 정보의 추출

어절간 품사 패턴 정보를 추출하기 위하여 품사 태그들 중 각 어절을 대표할 수 있는 품사 태그를 주품사 태그로 정의하였다. 본 시스템에서 사용하는 품사 태그의 수는 총 21개이며 이들 중 주품사 태그로 정의된 품사 태그는

<빈도수>:<품사 패턴 정보>위치정보-조사 태그 수 n-<조사 리스트 1><조사 리스트 2>..<조사 리스트 n>

(그림 1) 품사 패턴 형태

명사, 대명사 등 체언류에 대한 품사 태그와 동사, 형용사 등 용언류에 대한 품사 태그, 부사, 관형사, 감탄사 등 독립언에 대한 품사 태그 및 문장 부호에 대한 품사 태그로 총 13 개의 품사 태그를 주품사 태그로 정의하였으며 조사, 어미와 같은 문법 형태소에 대한 품사 태그는 주품사 태그에서 제외하였다. 통계 정보 추출에는 이들 주품사 정보와 조사, 접미사 정보만을 이용하게 된다. 통계 정보 추출시 문법 형태소를 추출하지 않고 주품사 태그만을 추출하게 되면 어절의 형태가 다양하게 나타나더라도 품사 패턴 정보가 크게 증가되지 않게 됨으로써 어절 단위 품사 태깅 시스템에서와 같은 심각한 자료 부족 문제의 발생을 줄일 수 있게 된다.

2.1 주품사 태그의 선정

본 시스템에서는 21 개의 품사 태그 셋을 갖고 있으며 이미 언급 했듯이 13 개의 품사 태그를 주품사 태그 셋으로 선정하였다. 한 어절에서 어절을 대표할 수 있는 품사 태그를 주품사 태그로 선정하였으며, 그외의 태그는 주품사 태그에서 제외하였다. 어절간 품사 패턴 정보를 추출함에 있어 주품사 태그만을 선정하는 것은 어절의 다양한 형태에 따른 자료 부족 문제를 줄이기 위해서이다. 또한 어미와 같은 정보들은 한 어절에 대해 형태소 분석기가 동일한 정보를 출력해 준다고 가정하였기 때문에 패턴 정보 추출에서 제외하였다.

2.2 품사 패턴 형태

(그림 1)은 품사 패턴 정보의 추출 형태에 대해 보여주고 있다. 품사 패턴 정보에는 패턴의 출현 빈도수, 품사 패턴 정보, 위치 정보, 조사 품사 태그의 수, 조사 리스트 등을 포함한다. 이중 위치 정보는 품사 패턴 정보가 추출된 위치를 나타낸다. 예를들어, 0 은 문장의 가장 처음 위치, 1 은 문장 중간, 2 는 문장의 가장 끝 위치에서 품사 패턴 정보가 추출되었음을 나타낸다. 품사 패턴 정보에는 주품사 정보뿐 아니라 조사 정보도 포함하는데 조사 태그 수는 품사 패턴 정보에 들어 있는 조사의 태그 수를 나타낸다. 즉, 조사 태그가 품사 패턴 정보에서 두 번 나타났다면 조사 태그의 수는 2 가 된다. 이때, 각 조사 태그와 대응하는 조사 리스트 항목이 존재하게 되며, 조사 태그의 수가 2 라면 조사 리스트 항목도 2 개가 된다. 또한 각각의 조사 리스트에는 조사로 쓰이는 형태소 항목들의 리스트가 포함되게 된다.

2.3 품사 패턴의 추출

(그림 2)는 7 개의 어절로 된 문장에 대해서 품사 패턴 정보를 추출하는 예를 보여주고 있다. 본 논문에서는 품사 패턴 정보의 추출 단위를 4 어절로 하고 있기 때문에 총 5 개의 품사 패턴 정보가 추출되게 된다. Step 1 에서 첫번째 어절에서는 NP 와 PP 태그가 추출되었고, 두번째 어절에서는 EF 를 제외한 NN 과 SV 태그만이 추출된 것을 알 수 있다. 또한 첫번째 어절에서 조사로 쓰인 형태소와 세번째 어절에서 조사로 쓰인 형태소가 조사 리스트에 각각 등록된 것을 볼 수 있다.

2.4 조사 리트스의 추가

품사 패턴 정보 추출 과정에서 동일한 품사 패턴이 발생하게 되면 패턴의 출현 빈도수를 증가시켜 주게 된다. 또한 이때 품사 패턴에 조사 태그가 포함되어 있을 경우 조사로 쓰인 형태소가 조사 리스트에 들어 있는지 검사하고 조사로 쓰인 형태소가 조사 리스트에 포함되어 있지 않을 경우 조사 리스트에 새로이 추가해 주게 된다. 예를들어 <그림 2>의 step 1 에서 추출한 품사 패턴 정보가 다시 나타났고, 첫번째와 세번째 어절에 나타난 조사 형태소가 각각 '가'와 '은'이라면 step 1 에서 추출된 품사 패턴 정보는 아래와 같이 바뀌게 된다.

2:NP+PP$NN+SV$NN+PP$AD-0-2-<가,이><은>

즉, 빈도수가 하나 증가되고 새로 나타난 조사 형태소 '가'가 첫번째 조사 리스트 항목에 추가되게 된다.

[입력 문장]
그것이 형성되는 길은 대개 두 가지에 의해서다.

그것이 : 그것(NP) + 이(PP)
형성되는 : 형성(NN) + 되(SV) + 는(EF)
길은 : 길(NN) + 은(PP)
대개 : 대개(AD)
두 : 두(NU)
가지에 : 가지(NX) + 에(PP)
의해서다 : 의하(VV) + 어서(EF) + 다(EF)
.:.(SY)

[통계 정보 추출]
step 1 : 1:NP+PP$NN+SV$NN+PP$AD-0-2-<이><은>
step 2 : 1:NN+SV$NN+PP$AD$NU-1-1-<은>
step 3 : 1:NN+PPADNU$NX+PP-2-<은><에>
step 4 : 1:ADNUNX+PP$VV-1-<에>
step 5 : 1:NU$NX+PP$VV$SY-2-1-<에>

(그림 2) 패턴 정보 추출 예

2.5 동일한 주품사(명사 품사)의 반복 출현

'기생(NN) + 관광(NN) + 사업(NN) + 의(PP)'와 같이 동일한 주품사가 연속해서 등장하는 경우 현재 어절에 대해서만 품사 패턴 정보를 추출하게 되면 'NN+NN+NN+PP' 형태를 갖게 된다. 그러나 이처럼 동일한 주품사 태그가

연속해서 나타날 경우 주품사를 모두 추출하여 품사 패턴 정보를 추출하게 되면 품사 패턴 정보의 유형이 많아지게 되며, 명사와 같은 경우는 다양한 복합명사의 결합 형태가 가능하므로 모든 복합명사 유형에 대한 품사 패턴을 통계 정보로 추출하는 것은 거의 불가능한 일이다. 따라서 동일한 주품사 태그가 연속해서 나오는 경우는 주품사 태그를 모두 패턴 정보에 포함하는 것이 아니라, 하나의 주품사 태그만을 패턴 정보에 추가함으로써 품사 패턴 정보의 수를 줄이도록 한다. 즉, 'NN+NN+NN+PP'를 'NN+PP'로 간략히 하여 품사 패턴 정보를 구축하게 된다.

2.6 주품사 태그의 분리

어절간 품사 패턴 정보를 추출할 때 한 어절 내에 주품사 태그가 여러 개 나올 때는 각각의 주품사 태그를 독립된 어절로 간주하고 품사 패턴을 추출한다. (그림 3)은 한 어절 내에 주품사 태그가 2개 이상 나올 경우 주품사 태그를 분리하여 추출하는 과정을 보여주고 있다.

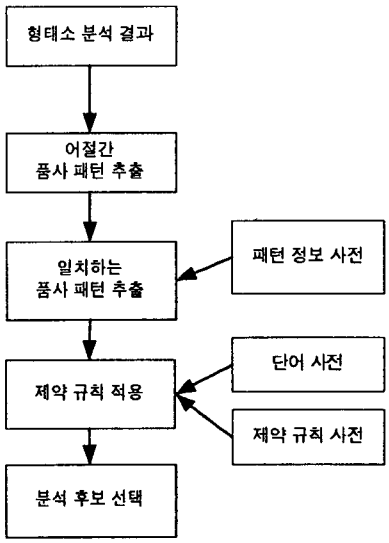
```
오후 : 오후(NN)
2 시쯤에 : 2(NU) + 시(NX) + 쯤(NX) + 에(PP)

[품사 패턴 정보]
      NN$NU$NX$NX+PP-1-1-<에>

감작스럽게 : 감작(NN) + 스럽(SJ) + 게(EF)
맥박이 : 맥박(NN) + 이(PP)
빨라진다면 : 빠르(VJ) + 아(EF) + 지(VX) + ㄴ다면(EF)

[품사 패턴 정보]
      NN+SJ$NN+PP$VJ$VX-1-1-<이>
```

(그림 3) 주품사 태그의 분리

(그림 3)에서 처럼 주품사 태그를 분리하여 추출하게 되면 띄어 쓰기 오류로 인해 한 어절로 잘 못 분석된 어절에 대해서도 올바른 패턴 정보를 추출할 수 있을 뿐만 아니라 어절의 형태가 다양하게 나타나더라도 품사 패턴 정보를 단순화하기 때문에 추출되는 품사 패턴의 양을 줄일 수 있게 된다. 또한 품사 패턴의 형태가 단순해 짐으로써 다양한 어절의 입력 형태에 따른 자료 부족 문제도 줄일 수 있게 된다.

3. 품사 태깅

(그림 4)는 본 시스템의 품사 태깅 구성도를 보여주고 있다. 품사 태깅은 입력된 분석 어절을 어절쌍(본 시스템에서는 기본적으로 4어절을 어절쌍으로 한다)으로 나누고 이들로부터 품사 패턴을 구한 후 통계 정보 사전에서 일치하는 품사 패턴을 찾게 된다. 일치하는 품사 패턴이 여러 개일 경우는 제약 규칙을 적용하여 가장 알맞은 후보를 선택하게 되고, 제약 규칙 적용 후에도 일치하는 품사 패턴이



(그림 4) 품사 태깅 구성도

여러 개일 경우는 빈도수가 가장 큰 것을 선택하게 된다. 품사 패턴이 통계 정보에 존재하지 않을 경우는 주품사가 패턴 정보와 가장 많이 일치하는 품사 패턴을 선택하게 된다.

3.1 제약 규칙의 적용

제약 규칙은 품사 태깅시에 올바른 품사 패턴을 선택하기 위하여 사용하게 된다. 빈도수가 높게 나타난 패턴이라고 하더라도 제약 규칙에 맞지 않는 패턴은 분석 후보에서 제거된다. (그림 5)에서 규칙 1은 조사 '을' 또는 '를'이 나오고 동사 태그가 나올 경우 해당 동사가 타동사이어야 한다는 제약 규칙을 나타내고 있으며, 규칙 2는 규칙 1과 유사하게 조사 '을' 또는 '를'이 나올경우 해당 명사가 '하다 타동사'이어야 함을 나타내고 있다.

```
Rule 1. PP[을,를] →  VV[VT]
Rule 2. PP[을,를] →  NN+SV[HT]
```

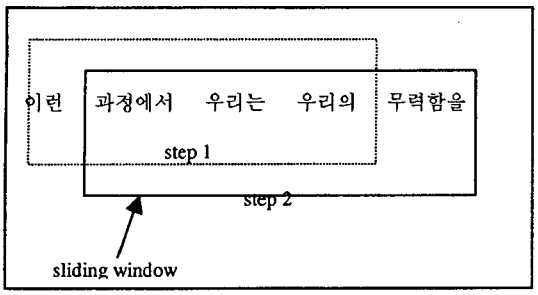
(그림 5) 제약 규칙 예

제약 규칙은 품사 패턴 형태에 대하여 공통적으로 적용할 수 있는 것을 규칙으로 선정하게 된다.

3.2 품사 태깅

본 시스템에서 품사 태깅은 입력 문장에 대하여 4어절을 기본 단위로 하여 수행하게 되며 입력 문장을 4어절씩 끊어 품사 태깅을 수행하고 다음 어절로 한 어절 건너뛴 후 다시 반복하여 품사 태깅을 수행한다. 이렇게 각 어절에 대해 품사 태깅을 반복 수행함으로써 보다 정확한 태깅을 가능하게 한다. 각각의 어절들은 품사 태깅 창에 여러 번 노출되므로 품사 태깅이 반복적으로 수행되고 여러 개의 분석 결과중 가장 빈번히 적용된 분석 결과가 해당

어절의 태깅 결과로 사용되게 된다.



(그림 6) sliding window 방식의 품사 태깅

4. 실험 결과

본 시스템에서는 21 개의 품사 태그를 가지고 품사 패턴 정보를 이용한 품사 태깅을 수행하였다. 품사 패턴 정보는 태그된 코퍼스 약 100 만 어절에 대하여 추출하였으며 사용된 제약 규칙의 개수는 9 개이다.

<표 1> 실험 코퍼스 통계

코퍼스 종류	어절수	중의적 어절수	중의도
내부 코퍼스	378	155	2.85 개
외부 코퍼스	365	134	2.72 개

실험 방법은 test suit 으로 각각 내부 평가 코퍼스 20 문장과 외부 평가 코퍼스 20 문장에 대해서 수행되었다. <표 1>은 실험에 사용된 문장의 전체 어절수와 중의성을 갖는 어절수 및 중의성을 갖는 어절의 평균 중의도를 나타내고 있다. 중의적 어절 수는 중의성을 갖는 어절의 수를 의미하며 중의도는 중의성을 갖는 어절의 평균 분석 개수를 나타낸다. <표 2>는 본 논문에서 제안한 시스템을 이용하여 품사 태깅한 결과를 나타내고 있다. 정확률은 중의성을 갖는 어절수에 대해 올바르게 태깅된 어절수의 비율로 나타내었다.

<표 2> 품사 태깅 결과

코퍼스 종류	중의적 어절수	중의성 해소 어절수	정확률
내부 코퍼스	155	135	87.1%
외부 코퍼스	134	116	86.6%

실험에서 내부 코퍼스와 외부 코퍼스에 대한 품사 태깅 결과 정확률이 각각 87.1%와 86.6%로 나타났다. 정확률이 이처럼 낮게 나타난 것은 아직 시스템이 완성되지 않았고, 제약 규칙도 몇 가지만 적용하였기 때문이다. 그러나 내부 코퍼스와 외부 코퍼스에 대한 정확률이 비슷하게 나옴으로써 품사 패턴 정보가 다양한 입력 문장에 대하여 이용 가능함을 보여주고 있다. 또한 품사 패턴 정보의 추출에 있어 현재는 4 어절 단위로 하고 있으나 품사 패턴 정보의 단위를 5 어절 혹은 6 어절로 늘리면 보다 높은 정확률을 얻을 수 있을 것이다.

5. 결론 및 향구 연구 과제

본 논문에서는 한국어 품사 태깅에 어절간 품사 패턴 정보와 제약 규칙을 이용한 품사 태깅 방법을 제안하였다. 아직 시스템이 완성되지 않은 관계로 보다 많은 코퍼스에 대한 실험을 하지 못하였지만 test suit 으로 실험을 해 본 결과 어절간 품사 패턴을 이용한 품사 태깅이 어느 정도 가능성이 있음을 알 수 있었다. 향후 연구 과제로 품사 패턴 정보를 좀더 구축하고 품사 태깅의 정확률을 높이기 위한 제약 규칙의 생성을 수행하고자 한다. 또한 품사 패턴 정보의 추출 단위를 5 어절, 6 어절로 확장한 후 품사 태깅을 수행해 보고 품사 태깅의 정확률과 품사 패턴의 유형이 얼마나 증가하는지를 알아볼 것이다. 본 시스템의 처리 범위를 알기 위해 코퍼스로부터 추출된 품사 패턴 정보가 다양한 입력 어절에 대해서도 잘 적용될 수 있는지를 알아보아야 할 것이다. 이외에도 품사 패턴 정보와 제약 규칙이 적용되지 않는 어절의 처리 문제에 대한 연구가 수행되어져야 할 것이다.

참고 문헌

[1] 김진동, 임희석, 임해창, "어절 단위의 문맥을 고려한 형태소 단위의 한국어 품사 태깅 시스템", 인지과학회 춘계 학술발표 논문집, pp, 97-106, 1996.

[2] 임희석, 김진동, 임해창, "어절 태그 변형 규칙을 이용한 한국어 품사 태거", 한국정보과학회 논문지(B), 제 24 권, 제 6 호, pp. 673-684, 1997.

[3] 김재훈, "가중치 망을 이용한 한국어 품사 태깅", 한국정보과학회 논문지(B), 제 25 권, pp. 951-959, 1998

[4] 이운재, "한국어 문서 태깅 시스템의 설계 및 구현", 한국과학기술원 전산학과, 석사학위 논문, 1993

[5] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사태깅", 한국정보과학회 논문지, 제 22 권, 제 1 호, pp.136-146, 1995.

[6] 임철수, "HMM 을 이용한 한국어 품사 태깅 시스템 구현", 한국과학기술원 전산학과, 석사학위 논문, 1997.

[7] 임해창, 임희석, 윤보현, "자연언어처리를 위한 품사 태깅 시스템의 고찰", 한국정보과학회지, 제 14 권, 제 7 호, pp. 36-57. 1996.

[8] 신상현, 이근배, 이종혁, "통계와 규칙에 기반한 2 단계 한국어 품사 태깅 시스템", 정보과학회논문지(B), 제 24 권, 제 2 호, pp. 160-169, 1997.

[9] P.Tapanainen, A. Voutilainen, "Tagging accurateley Don't guess if you know", Proc. Of the 7th Conference of the European chapter of the Association for Computational Linguistics, pp. 149-156, 1994

[10] A. Volutiliainen, "A syntax-based part-of-speech analyzer", Proc. Of the 7th conference of the European chapter of the ACL, pp. 157-164, 1995.

[11] M.Zhang, S. Li, T. Zhao, "Tagging Chiness Corpus Based on Statistical and Rule Techiniques", Proc. Of the Int. Conference on Computer Processing of Oriental Language(ICCPOL-97), pp. 503-506, 1997.

[12] Doug Cutting, Julian Kupiec, Jan Pedersen, Penelope Sibun, "A Practical Part-of-Speech Tagger", Proceedings of 3rd Conference on Applied NLP, pp. 133-140, 1992.