조사와 어미의 문법 기능을 활용한 품사 태깅 시스템

안영민[0] 서영훈[00]
충북대학교 컴퓨터공학과
컴퓨터 정보통신 연구소
[0]manic@dcenlp.chungbuk.ac.kr [00]yhseo@cbucc.chungbuk.ac.kr

Part-Of-Speech Tagging System Using Grammatical Function of Josa & Eomi

Young-Min An Young-Hoon Seo
Dept. of Computer Engineering, Chungbuk National University
Research Institute for Computer and Information Communication

요 약

본 논문은 규칙과 통계 정보를 모두 적용하는 혼합형 품사 태깅 시스템에서 통계 정보를 이용하여 품사 태깅을 수행할 때 조사와 어미를 문법 기능에 따라 구분하여 사용하는 품사 태깅 시스템을 기술한다. 품사 태깅은 주로 주변의 품사열을 이용하게 되는데 품사 정보를 추출할 때 조사와 어미의 문법 기능인 조사의 격 정보와 어미의 활용형 정보에 따라 몇 가지로 분류하고 정보를 추출하여 품사 태깅에 적용하면 조사와 어미를 분류하지 않은 품사열 만을 사용한 태깅 방법 보다 더 나은 성능을 얻을 수 있다.

1. 서론

자연언어처리 응용 분야의 기본 단계인 형태소 분석을 마치면 언어의 특성 때문에 중의성이 발생하게 된다. 이런 중의성의 해소를 위해 품사 태깅이 필요하게 된다. 기존의 알려진 품사 태깅으로는 규칙을 이용하는 방법[1], 통계 정보를 이용하는 방법[2,3], 둘 모두를 이용하는 방법[4,5,6,7]이 있다. 규칙만을 이용하는 경우는 높은 태깅 정확도를 가질 수 있으나 규칙이 적용될 수 있는 제한된 문장에만 사용될 수 있다. 따라서 처리할 수 있는 범위가 매우 제한적이다. 그에 반해 통계 정보를 이용하는 경우는 대량의 말뭉치에서 통계 정보를 추출한 확률값으로 태깅에 적용하므로 데이터의 처리 범위가 매우 넓다. 그러나 실세계의 언어 현상을 반영할 수 있는 양질의 말뭉치를 구축하기가 힘들어 통계 정보의 자료가 부족하게 된다. 그리고 통계, 확률이라는 방법을 쓰기 때문에 정확도가 떨어지는 단점이 있다. 최근에는 규칙과 통계를 모두 이용하여 상호보완적으로 정확도와 처리범위를 향상시키는 방법이 주로 쓰이고 있으며 그 중에서도 통계를 이용한 방법에서 더 낳은 통계 정보를 추출하여 적용함으로써 좀 더 정확한 태깅 결과를 얻기 위해 연구하고 있다.

본 논문에서는 조사와 어미를 문법기능에 따라 분류하여 추출한 통계정보를 태깅에 적용하고 품사 태그 패턴의 보정값을 적용하여 태깅의 정확도를 높이는 방법을 기술한다.

한국어는 조사와 어미가 그 단어의 특성과 주변 단어들과의 관계를 나타내는 경우가 많으므로 통계 정보 추출 시 조사와 어미를 분류하여 추출하면 단어와 단어간의 관계를 나타내는 좀더 정확한 정보를 얻을 수 있다.

그리고 통계 정보를 이용한 태깅은 전체 말뭉치에서 뽑아낸 어절 단위 태그패턴의 빈도수를 기반으로 수행되기 때문에 전체 빈도수가 높은 태그패턴은 태깅의 옳고 그름에 관계없이 지역적으로 높은 빈도수를 가질 수 있다. 이것은 잘 균형 잡힌 양질의 말뭉치 구축이 어렵기 때문이기도 하다. 이를 보정하기 위해 지역적인 빈도수와 전체 빈도수를 고려한 보정값을 사용하여 태깅한다.

2. 조사와 어미의 문법 기능 활용

접속조사는 체언과 체언과의 관계를 규명하는 범주로서, 성분과 성분을 서로 연결하여 문장에서 동등한 자격

을 가지게 한다. 그러므로 접속조사나 관형격조사 다음에는 용언 보다는 체언이 나올 확률이 높다. 그리고 일반적으로 목적격 조사 다음에는 체언 보다는 용언이 나올 확률이 그만큼 높다. 따라서 이 조사들을 구분하여 통계 정보를 추출하는 것이 통계를 이용한 방법에서는 좀 더 정확한 태깅을 위한 통계 정보가 될 수 있다.

어미의 경우, 종결어미는 문장의 끝에 나올 가능성이 가장 크고 연결어미는 뒤따르는 문장이나 용언에 연결시키는 어미이고 관형사형 어미는 관형어의 역할로 다음에 체언이 나올 가능성이 크며, 명사형 전성어미가 붙으면 체언으로서 조사를 취할 수도 있다. 이렇듯 역할이 조금씩 다르므로 모두 통합하여 명사+조사 또는 동사+어미 하나로 통계 정보를 추출하는 것 보다는 역할에 따라 조사나 어미를 분류하여 통계 정보를 추출하고 태깅에 적용하는 것이 더욱 신빙성 있는 통계 정보를 추출할 수 있고, 태깅의 정확도를 높일 수 있게 된다.

	전체 빈도수	다음어절 및 빈도수	다음어절 발생률
접속조사	2353	체언구 1801	76.5%
관형격조사	12237	체언구 9958	81.3%
목적격조사	21345	용언 16027	75.0%
연결어미	31007	용언 18153	58.5%
종결어미	22059	문장끝 19000	86.1%
관형형어미	32035	체언구 25693	80.2%

[표 1] 조사 및 어미에 따라 나타나는 다음 어절과 빈도수

위의 [표 1]은 태깅 시스템 구현에 사용된 대상 말뭉치에서 몇 개의 조사와 어미에 따라서 뒤따르게 되는 어절의 특징과 그 빈도수를 나타낸 것이다. [표 1]에서 나타나듯 조사나 어미의 활용에 따라 뒤따르는 어절이 특정한 품사패턴을 가지고 자주 발생한다는 것을 알 수 있다.

3. 통계 정보 추출

시스템에 사용되는 통계 정보는 신뢰성을 위해 이미 태깅된 말뭉치로부터 추출된다. 통계 정보 추출에 사용된 대상 말뭉치는 연구용으로 배포된 29만 어절의 ETRI 품사 태그 부착 말뭉치이다. 품사 태깅을 수행하기 위해 사용되는 형태소 분석기와 통계 정보 추출을 위한 대상 말뭉치 간에 품사 태그 집합이 다르기 때문에 통계 정보를 추출하기 위해 품사 태그 집합을 매핑 시키는 과정이 필요하게 된다.

통계 정보는 3어절의 어절 단위 품사 태그들로 구성된다. 통계 정보를 추출하기 위해 우선 대상 말뭉치 전체를 대상으로 한 어절 단위의 품사 태그 패턴과 그 패턴의 전체 빈도수를 추출한다.

```
NN          31029
NN+SF+ETM   7932
NN+PP       53576
AD          19991
VV+EC       20437
VX+EP+EM    2622
VV+ETM      15995
VV+EM       2946
NN+PM       10313
NN+PO       20378
VX+EC       2166
VV+EP+EM    6610
VV+EP+EC    798
VX+ETM      3419
DT          7199
NN+SF+EM    3933
VX+EM       1582
NN+SF+EC    6511
            .
            .
            .
```

[그림 1] 태그패턴과 출현 빈도

추출된 태그 패턴들의 모든 조합을 이용하여 통계 정보를 추출하게 된다. 그 각각의 조합이 통계 정보 추출에 사용되는 3 윈도우[8]의 좌우 양쪽을 구성하게 된다.

NN	추출될 태그 패턴	VV+EM

[그림 2] 통계 정보 추출에 사용되는 3개의 창을 가지는 윈도우

대상 말뭉치에 이 3 윈도우를 적용하여 한 어절을 중심으로 양쪽 어절의 태그 패턴이 동일하면 그 가운데 있는 어절의 태그 패턴이 빈도수와 함께 추출된다. 통계 정보를 추출할 때 조사와 어미를 각각 4가지로 구분함

으로써 조사나 어미에 따른 체언과 용언의 통계 정보를 구축한다.

조사는 PJ(접속조사), PM(관형격조사), PO(목적격 조사), PP(나머지) 이 네 가지로 구분하였고 어미는 EC(연결 어미), EM(종결 어미), ETN(명사형 전성어미), ETM(관형사형 전성어미) 이 네 가지로 구분함으로써 조사와 어미의 역할을 구분한 통계 정보를 구축한다.

```
NN+PP  NN
       NN        759
       AD        356
       VV+ETM    1135
       VV+EC     360
       VV+EP+EC  26
       NN+SF+ETM 447
       NN+SF+EM  9
       NN+SF+EC  86
       VV+ETN+PP 6
       NN+SF+EP+EM 6
       NN+PP     239
       DT        185
                 .
                 .
                 .
NN+PP  NN+SF+ETM
       AD        307
       VV+EC     149
       NN+PO     125
       VV+EC+VX+EC 7
       DT        43
       NN+PP     230
       NN+SF+ETM 25
       NN        43
       VV+ETM    74
       NN+PM     64
       NN+VV+ETM 5
                 .
                 .
                 .
```

[그림 3] 추출된 통계 정보의 형태

4. 태깅 시스템의 구조 및 실험

두개 이상의 형태소 분석 결과를 가진 중의성이 있는 어절을 만나게 되면 그 어절을 중심으로 좌우 양쪽 어절들의 품사 태그 패턴을 가지고 통계 정보를 얻어와 중의성이 있는 어절의 형태소 분석 결과들에 적용하여 scoring을 하게 된다. 이 형태소 분석 결과들 중 가장 큰 scoring 값을 가지게 되는 형태소 분석 결과가 중의성이 존재하는 어절의 형태소 분석 결과로 선택된다. 그리고 또한 현 태깅시스템은 태그패턴의 전체 빈도수와 좌우 양쪽 어절 사이의 태그패턴 빈도수를 고려한 보정값도 사용되어 태깅 결과를 얻게 된다.

※ 보정내용

```
NN       100
NN+SF+ETM  24
NN+PP    53
AD       20
         .
         .
         .
```

태그패턴과 출현 빈도의 예

```
NN       VV+EM
         NN       12
         AD       10
         NN+PP    9
                  .
                  .
                  .
```

통계 정보의 예

[그림 4] 보정내용을 위한 예

[그림 4]에서 보면 전체 말뭉치에서 태그 패턴 NN이 100번의 빈도수를 가지고 있고 태그 패턴 NN+PP와 VV+EM사이에서 NN이 12번이 사용되었다. 그리고 태그 패턴 AD는 전체 말뭉치에서 20번의 빈도수를 가지고 있고 태그 패턴 NN+PP와 VV+EM사이에서 10번이 사용되었다. 이 경우 빈도수로는 NN이 12번으로 AD의 10번 보다 많지만 AD의 경우는 말뭉치 전체 중의 절반이 NN+PP와 VV+EM의 사이에 쓰인 것 이다. 이에 AD에는 그 통계값 만큼의 가중치를 주어 계산함으로써 통계 정보를 보정하여 사용한다..

실험

현 시스템은 규칙 및 통계 정보를 모두 이용하는 품사 태깅 시스템이다. 본 논문에서 기술한 조사와 어미의 문법기능을 이용한 품사 태깅 방법을 테스트 하는데 있어서 굳이 규칙 정보를 이용한 품사 태깅 루틴 부분을 삭제하지 않고 통합하여 테스트 하였다.

실험결과

	중의성을 가진 어절 수	옳은 태깅 결과	정확률
조사와 어미를 분류하지 않은 태깅	3849	2578	66.9%
조사와 어미를 분류한 태깅	3849	3651	94.8%

[표 2] 실험결과

5. 결론 및 향후 연구

본 논문에서는 조사와 어미를 분류하여 통계 정보를 추출하고 태깅에 적용하는 방법을 제안하였다. 조사와 어미의 쓰임에 따라 몇 가지로 분류하여 통계적 방법의 태깅에 적용함으로써 그렇지 않은 방법보다 더 향상된 태깅 결과를 얻을 수 있다는 것을 실험결과에서 볼 수 있었다. 또한 전체 출현 빈도수와 좌우 양쪽 어절 사이에 나타난 빈도수를 감안하여 통계 값을 보정함으로서 향상된 품사 태깅 결과를 얻을 수 있었다.

통계 정보를 얻고 유지하는 전 과정이 통계 정보 도구를 사용하여 자동으로 수행되기 때문에 대상 말뭉치에 대해 유동적이며 품사 태그 매핑 테이블만 변경해주면 품사 태거의 소스를 약간만 수정하는 것으로써 범용적으로 사용할 수 있게 된다.

오분석된 결과의 주된 원인으로는 중의성을 가진 어절이 연속으로 나타날 경우이다. 중의성을 가진 어절이 연속으로 나타날 경우 현재 시스템에서는 앞뒤 어절의 모든 경우를 계산하여 가장 높은 값을 선택하도록 되어 있는데 뒤쪽 어절 태그 패턴이 옳지 않은 경우가 태깅 스코어링에 영향을 주는 경우가 있어 정확한 결과를 얻기가 어렵다. 따라서 향후 연구에는 연속된 어절이 중의성을 가질 경우 좀 더 좋은 성능을 갖도록 하는 연구와 조사와 어미의 특성에 대한 연구 그리고, 현 시스템에서 고려하고 있지 않은 동품사 중의성에 대한 연구도 병행하게 될 것이다.

6. 참고 문헌

[1] 이중영, 이기영, 김한우, "어절간 규칙을 이용한 형태소 중의성 해결", 한양 대학교 전자계산학과 인공지능연구실

[2] 이상주, 임희석, 임해창, "은닉 마르코프 모델을 이용한 두단계 한국어 품사 태깅", 제 6회 한글 침 한국어 정보처리 학술대회 발표 논문집, pp.305-312, 1994.

[3] 임철수, "HMM을 이용한 한국어 춤사 태깅 시스템 구현",한국과학기술원 전산학과 석사학위 논문, 1994.

[4] 신상현, 이근배, 홍남희, 이종혁, "확률과 규칙을 사용한 품사 태깅", 제 6회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.318-321, 1994.

[5] 신상현, 이근배, 홍남희, 이종혁, "TAKTAG: 통계와 규칙에 기반한 2단계 학습을 통한 품사 중의성 해결", 제 7회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.169-174, 1995

[6] 임희석 김진동, 임해창, "언어 지식과 통계 정보의 보완적 특성을 이용한 품사 태깅", 제 9회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.102-108, 1997.

[7] 임희동, 서영훈, "어절간 문맥 정보를 이용한 혼합형 품사 태깅", 제 12회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp.376-380, 2000.

[8] Christopher D,Manning, Hinrich Sch\tze, "Foundations of Statistical Natural Language Processing", fourth printing, The Mit Press, 2001, p353-356.