

품사별 자질을 이용한 한국어 품사부착의 성능 향상

최원종⁰ 이도길 임해창
 고려대학교 자연어처리연구소
 {wjchoi, dglee, rim}@nlp.korea.ac.kr

Improving Korean Part-of-speech tagging by Part-of-Speech specific features

Won-Jong Choi⁰, Do-Gil Lee, Hae-Chang Rim
 Natural Language Processing Lab. Korea University

요 약

한국어 형태소분석 및 품사부착에서 일부 품사는 높은 중의성으로 인하여 오류가 많으며, 일부 품사가 전체 오류의 대부분을 차지한다. 본 연구에서는 높은 중의성으로 인하여 오류가 많은 품사를 대상으로, 각 품사에 적합한 자질을 이용하여 학습한, 정확률이 높은 분류기를 통계적 방식의 태거와 순차 결합하여 형태소분석/품사부착 성능을 향상하였다. 2003년 세종계획 품사 부착 말뭉치 200만 어절에서 학습하여 평가를 한 결과 기존 통계적 품사 부착기에 비해 정확도는 0.62% 향상되었으며, 오류는 13.12% 감소하였다.

1. 서 론

한국어 형태소 분석 및 품사 부착은 자연어를 이해하여 처리하는데 있어서 가장 기본적인 기술이며 다른 상위 시스템의 성능에 큰 영향을 미치는 가장 중요한 부분이다.

이러한 한국어 형태소 품사 부착의 성능을 향상하기 위해서는 모든 품사의 중의성을 해소할 수 있는 자질을 모형에 반영하여야 하나, 널리 쓰이는 태거가 여러 자질을 이용하여 이미 95% 이상의 성능을 보이고 있기 때문에 다양한 오류를 반영하는 자질을 찾기란 쉽지 않다.

따라서 모든 품사의 성능을 향상할 수 있는 자질을 추가적으로 사용하여 성능을 향상 하는 것은 쉽지 않은 문제이다. 하지만 모든 품사에는 유용하지 않지만 일부 품사에는 유용한 자질을 이용하여 특정 품사의 부착에만 적용할 수 있다면 쉽게 성능향상을 할 수 있다.

표 1은 품사 부착 오류 중에서 비중이 높은 상위 9종류의 품사쌍이다. 표 1에서 알 수 있듯이 중의성이 높은 상위 9종류의 품사쌍이 전체 품사 오류 중에서 80% 이상을 차지한다. 따라서 모든 품사에 대한 성능을 향상하지 않고 중의성이 높은 일부 품사들만 성능을 향상해도 오류 중 대부분을 개선할 수 있다.

본 연구에서는 이러한 특성을 이용하여 특정 품사에 적합한 이진 분류기에 기반하는 품사부착 후처리를 제안한다. 이때 대상이 되는 품사는 표 1에 나온 품사처럼 중의성이 높고 전체 오류에서 많은 비중을 차지하는 품사를 대상으로 한다.

그리고 제안된 후처리 방식은 품사 부착 대상의 주변 문맥의 어휘 자질과 품사 자질을 이용하기 때문에 통계적 방식의 태거[1]로 품사 부착된 말뭉치를 입력으로 받아 대상이 되는 품사들만 다시 품사 부착하며 다른 태거와도 쉽게 결합할 수 있다.

표 1: 한국어 품사 부착에서 품사 오류

오류 품사 유형	오류 형태소 개수	품사 오류 중 비율
VV ↔ VX	945	21.203%
NNG ↔ NNP	814	18.263%
JKC ↔ JKS	530	11.891%
NNB ↔ NNG	287	6.439%
XSA ↔ XSV	315	7.068%
JC ↔ JKB	367	8.234%
MAG ↔ NNG	131	2.939%
MM ↔ NNG	91	2.042%
NNG ↔ XR	92	2.064%
계		80.144%

2. 기존 연구

품사 부착의 오류가 있는 부분을 효과적으로 찾고, 교정하여 성능을 향상하는 기존연구는 다음과 같은 것들이 있다.

[2]은 오류 가능성이 높은 부분을 자동으로 찾기 위하여, 널리 쓰이는 통계적 태거가 내어주는 각 품사에 대한 확률의 차이가 근소하거나, 미리 구축된 품사열에 대한 정규표현식을 이용하여 통계적 태거의 결과가 일치하지 않는 경우를 오류가 높은 부분으로 판단하는 두 가지 방법을 사용하였다.

[3]은 통계적으로 오류가 많다고 판단된 어절을 대상으로 해당 어절에 대하여 좌우어절내의 대표 형태소 어휘 문맥 정보에 기반하여 오류를 정정하는 방식이다. 하지만, 어절 어휘를 대상으로 하기 때문에 coverage에 한계가 있다.

[4]는 통계적 방식으로 품사 부착을 한 결과를 Eric

Brill스타일로 학습된 규칙을 이용하여 오류를 교정하는 방식이다.

3. 분류기



그림 1: 품사별 분류기를 이용한 후처리

기존의 태거가 형태소 분석 결과를 입력으로 받는 반면에 본 연구에서는 다른 태거의 품사부착 결과를 입력으로 받는다. 이때 사용할 수 있는 태거로는 규칙기반, 통계 기반의 모든 태거를 사용할 수 있다.

결국 입력으로 받은 결과에서 후처리 대상과 주위 문맥을 추출하여 가장 적절한 품사를 결정하는 분류이기 때문에 자연어처리에서 높은 성능을 보이며 여러 자질을 결합하기 편리한 최대 엔트로피 모델 (Maximum Entropy Model)을 이용하여 학습된 분류기를 학습한 후, 품사 부착을 한다.

4. 분류기에 따른 자질

품사들의 오류를 개선하기 위하여 사용한 자질을 설명한다.

4.1 본용언(VV) / 보조용언(VX)

① 어미 : 본용언과 보조용언 사이에는 어미가 들어가게 되며, 본용언 앞에 올 수 있는 어미는 일반서술(-아/어), 상대서술(-게), 부정서술(-지), 진행서술(-고) 등 4종류가 있으며, 이러한 어미 다음에 용언이 나타난다면, 그것은 보조용언일 가능성이 높다.

② 용언 : 보조용언은 보통 본용언의 뒤에 나타난다. 따라서 후처리 대상의 이전에 용언이 있다면 그것은 보조용언일 가능성이 높다.

③ 동사파생접미사, 형용사파생접미사 : 이들 접미사 뒤에는 보조용언이 나타날 가능성이 높다.

④ 종결어미 : 용언이 다른 용언과 종결어미의 사이에 있다면 해당용언은 보조용언일 가능성이 높다.

4.2 주격조사(JKS) / 보격조사(JKC)

① 어휘 : 어휘는 조사의 격에 대한 중의성 감소에 큰 도움이 된다. 주격조사는 "이, 께서, 가, 에서" 등이며, 목적격조사는 "을, 를", 보격조사는 "이, 가" 등으로 격에 따라 어휘가 달라지기 때문이다. 일부 어휘가 여러 가지 격으로 사용가능하기는 하지만, 어휘는 조사의 격을 결정하는데 중요한 자질이다.

② 다음 형태소의 어휘, 품사 : "되-, 아니-"의 앞서나오는 체언에 붙는 조사 "이, 가"는 보격조사이다. 따라서 다음 형태소가 동사이면서 "되, 아니"의 어휘를 가진다면 보격조사로 분류할 수 있기 때문에 주격조사, 보격조사 구분에 아주 중요한 자질이다.

4.3 동사파생접미사(XSV) / 형용사파생접미사(XSA)

① 이전 형태소의 어휘와 품사 : 그림 2에서와 같이, 동사로 파생되는지, 형용사로 파생되는지는 이전 형태소의 성질에 의해서 결정이 된다. 따라서 이전 형태소의 어휘와 품사는 중의성 해결에 중요한 자질이다.

전망이 우세하다.	전망/NNG+이/JKS 우세/NNG+하/XSV+다/EF+./SF
전망이 우세하다.	전망/NNG+이/JKS 우세/NNG+하/XSA+다/EF+./SF

그림 2: XSV와 XSA 품사 부착 오류 예[오류(위), 정답(아래)]

4.4 부사격조사(JKB) / 접속조사(JC)

① 형태소 다음에 따라오는 동사파생접미사 / 형용사파생접미사 : 부사격 조사와 접속조사 간에 오류가 발생하는 경우는 형태소의 어휘가 "와/과"인 경우에 발생한다. 또한 이전 형태소와 다음 형태소가 명사인 경우에 다음 형태소에 전성어미가 붙어서 동사나 형용사로 쓰이는 경우 오류가 발생한다.

4.5 관형사(MM) / 일반명사(NNG)

① 이전형태소의 어휘, 품사 : 그림 3 경우처럼 형태소 어휘 "전"은 관형사/일반명사 중의성이 심한 형태소이지만, 이전 형태소의 쓰임에 따라 그 쓰임이 결정되기 때문에 이전 형태소의 어휘와 품사를 자질로 이용할 수 있다.

얼마전 얼마/NNG+전/NNG

그림 3 "전/NNG"의 예

② 현재 형태소의 어휘 : 관형사/일반명사 중의성이 심한 형태소는 "전-, 온-"처럼 어휘 자체의 중의성이 심한 경우가 많다. 따라서 어휘는 이 형태소가 관형사/일반명사 중의성이 심한 형태소인지를 구별할 수 있는 자질이 된다.

③ 다음 형태소의 어휘, 품사 : 형태소가 관형사로 쓰인다면, 다음에는 명사가 올 가능성이 높다.

5. 실험 및 평가

세종계획 2003년도 품사부착 말뭉치 데이터를 사용하였으며, 학습에는 1813491개의 어절을, 실험에는 202369개의 어절을 사용하였다. 결국 기존 통계적 태거보다 풍부한 자질을 사용하여 특정 품사에 정확한 분류 모델이 학습이 되는 것이므로 약 180만 개의 학습 집합의 개수는 충분하다고 할 수 있다. 최대 엔트로피 확률을 추정하기 위해 Zhang Le의 Maximum Entropy Modeling Toolkit[5]을 이용하여 학습을 수행하였다. 파라미터 추정은 L-BFGS 알고리즘을 사용하고, 반복회수는 150으로 하였다.

5.1 학습된 분류기의 성능

4장에서 설명한 자질을 이용하여 학습된 분류기의 성능을 평가하기 위하여 다음과 같은 정확률 평가 척도를 이용하여 실험을 하였다.

$$\text{정확률} = \frac{\#(\text{형태소 분석과 품사가 맞는 형태소})}{\#(\text{형태소 분석이 맞는 형태소})}$$

표 2는 통계적 방식 품사태거의 품사별 품사부착 정확률과 품사별 자질을 이용하여 학습된 분류기의 품사별 품사부착 정확률을 비교한 것이다.

각 품사에 대하여 통계적 방식 품사태거의 정확도보다 품사별 자질을 이용하여 학습된 분류기가 대상 품사에 대해서 향상된 성능을 내는 것을 알 수 있다. 최대 94.55%(4.34%→0.24%, JKC/JKC)가 감소되는 등 품사 부착 문제에 있어서 추가된 자질이 매우 효과적이었다. 즉, 추가된 품사별 자질은 품사 부착의 성능에 도움이 됨을 알 수 있다. 따라서 이러한 향상된 분류기를 적용한다면, 전체 성능을 향상할 수 있다.

오류를 많이 발생하는 상위 9종류의 품사들 중, NNB, NNP, NNG등은 사용되는 어휘, 품사 문맥이 상당히 흡사하여 성능 향상에 도움이 되는 자질을 찾지 못하였다. 어휘, 품사이외에 다른 자질을 이용하여 성능을 향상할 필요가 있는 품사들이다.

표 2: 품사별 품사 부착 성능 비교

	통계적 품사태거의 품사별 품사부착 정확률	학습된 분류기의 정확률	오류 감소율
VV ↔ VX	97.471%	98.608%	44.96%
JKC ↔ JKS	95.631%	99.762%	94.55%
JKB ↔ JC	98.103%	98.775%	35.42%
XSV ↔ XSA	97.201%	99.182%	70.78%
MM ↔ NNG	99.071%	99.112%	4.41%

5.2 품사 부착 후처리에 적용

높은 정확률을 가진 품사별 분류기를 적용한 결과가 가져오는 성능 향상을 실험하였다.

품사 부착의 성능은 어절단위 평가를 하였으며, 후처리 분류기를 적용하여 실험하였을 경우의 성능 향상은 다음과 같다.

$$\text{정확률} = \frac{\#(\text{정답과 일치하는 어절})}{\#(\text{전체 어절})}$$

표 3: 후처리 적용 후 성능 비교

품사	향상된 정확률
VV/VX	+0.207%
JKS/JKC	+0.245%
JKB/JC	+0.067%
XSV/XSA	+0.105%
MM/NNG	+0.001%
위의 품사들을 모두 적용하였을 경우	+0.623% (95.874%)

표 3은 품사별 자질을 이용하여 후처리를 적용하였을 경우 향상된 정확률이다. 본용언/보조용언은 0.2%, 주격조사/보격조사는 0.24%로 전체 성능 향상에 도움을 주었으며, 나머지 품사들도 성능향상에 도움이 되었다. 전체적으로 높은 증의성으로 인하여 오류가 많은 품사일수록 품사별 분류기를 이용한 후처리가 성능 향상에 도움이 많이 되고 있다.

본 연구에서 사용한 자질을 모두 적용하였을 경우 어절단위 평가로 전체적으로 0.623% (95.251% → 95.874%)로 13.11%의 오류가 감소하는 성능 향상을 하였다.

6. 결론 및 향후 연구

자질을 이용하여 품사 부착의 성능을 향상하고자 할 때 모든 품사에 유용한 자질을 찾는 것은 매우 어려운 일이다. 하지만 특정 품사에 유용한 자질을 특정 품사에만 적용한다면, 원하는 품사의 정확도를 향상시키고, 전체적인 성능 향상도 향상시킬 수 있다.

또한 품사 부착에서 증의성이 높은 일부 품사들의 오류가 전체 오류의 대부분을 차지하기 때문에 일부 품사만의 개선만으로도 전체의 성능을 효과적으로 향상할 수 있다.

본 연구에서는 이러한 특성을 이용하여, 다른 태거의 형태소 분석/품사부착 결과를 입력으로 받아 주변 어휘/품사 문맥 자질을 이용한 후처리를 통하여 성능을 향상하였다.

연구 결과, 일부 몇 가지 품사만으로도 의미 있는 성능향상을 할 수 있었으며, 앞으로 다른 품사에 대한 자질을 추가할수록 성능을 더욱 높일 수 있다.

References

- [1] 이도길, 한국어 형태소 분석과 품사 부착을 위한 확률 모형, 고려대학교 박사학위논문, 2005.
- [2] Elliott Macklovitch, "Where the Tagger Falts", In Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation, pp 113-126, 1992.
- [3] 김영길, 양성일, 홍문표, 박상규. "형태소 어휘 문맥에 기반한 태깅 오류 정정", 제15회 한글 및 한국어 정보처리 학술대회, 2003.
- [4] 신상현, 이근배, 이종혁, "통계와 규칙에 기반한 2 단계 한국어 품사 태깅 시스템", 정보과학회논문지 제 24권 제 2호, 1997
- [5] Zhang Le, "Maximum entropy modeling toolkit for Python and C++", 2004.