제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

문법 규칙과 어절 상관도를 이용한 품사 태깅 시스템

도미숙, 최호섭, 옥철영
울산대학교 컴퓨터정보통신공학과
{msdos98, hoseop, okcy}@mail.ulsan.ac.kr

Parts-Of-Speech Tagging System Using Grammar Rule and Eojeol Relativity

Mi-sook Do, Ho-seop Choe, Cheol-young Ock
Dept. of Computer Engineering and Information Technology, University of Ulsan

요 약

본 논문에서는 문법 규칙과 어절 상관도를 이용한 품사 태깅 시스템을 제안한다. 원시 말뭉치와 품사태그 부착 말뭉치에서 중의 어절(ambiguity eojeol)의 앞뒤 어휘와 품사 정보를 파악하여 문법 규칙을 마련하였으며, 한국어의 품사와 문장성분적 요소를 고려한 7 개의 어절 태그를 설정하여 이 어절 태그간의 확률값을 이용해 어절 간의 상관도를 구하였다. 이러한 방법들을 이용하여 품사 태깅을 실험한 결과, 150 만 어절의 학습 말뭉치와 3 만 어절의 실험 말뭉치에서 각각 평균 92%와 91%의 정확률을 보였다.

1. 서론

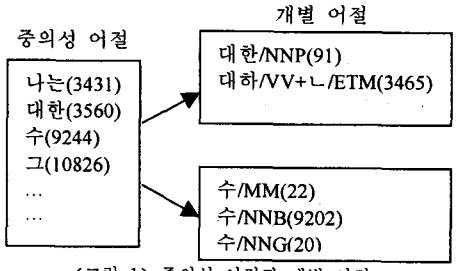
컴퓨터의 사용이 늘어남에 따라 사용자의 편의를 위한 인터페이스의 개발이 중요한 사항이 되고 있다. 그 중 자연어 처리를 요구하는 정보검색이나 기계번역 등의 응용 프로그램들이 좀더 높은 정확률과 속도 향상을 위해 여러 가지 방법을 동원한 연구가 진행중이다. 한국어 처리에서는 형태소 분석이 가장 기초 단계라 할 수 있는데, 이 과정에서부터 많은 중의성이 발생해 이런 중의성을 줄이고자 여러 가지 방법들이 제안되었다. 이런 형태소 분석 후 발생하는 중의성을 해결하는 즉, 품사를 하나로 결정시켜 주는 시스템을 품사 태깅 시스템이라 한다. 기존의 품사 태깅 시스템의 방법으로는 규칙을 이용한 방법([4],[6]), 확률 또는 통계를 이용한 방법([3],[8],[10],[16]), 그리고 이 둘을 혼합한 방법([1],[5],[9],[12],[15]) 등이 있다.

본 논문에서는 UCMA 형태소 분석기의 형태소 분석 결과를 바탕으로, 문법 규칙과 어절 상관도를 이용한 품사 태깅 시스템을 제안한다. 먼저 어절간의 상관도만 이용해 품사를 부착해 본 결과 약 88~90%의 정확률을 얻을 수 있었다. 다음으로 품사 부착한 결과에 대한 오류 분석을 통해 일반적인 규칙을 설정함으로써 정확률을 3% 정도 올릴 수 있었다.

2. 규칙 추출

어순이 자유로운 한국어의 모든 어휘에 대한 규칙을 추출할 수 없기 때문에 규칙 추출 대상을 선정하기 위하여 중의성을 일으키는 어절을 파악하였다. 이는 세종 150 만 원시 말뭉치(학습 말뭉치)에서 중의성 어절의 개수는 6,200 여 개이고, 총 32 만여 개의 빈도를 가진다. 그리고 이 중의성 어절에 대한 품사 태깅 결과를 살펴볼 수 있는 150 만 어절 품사태그 부착 말뭉치에서는, 품사태그 부착 형태가 다른 것인 개별 어절 개수가 13,000 여 개로 나타났다. 이러한 중의성 어절을 중심으로 문법 규칙을 통해 형태소 분석 결과를 줄여줌과 동시에, 형태소 분석 결과 중 문법 규칙에 확실히 부합하는 결과에 대해서는 품사 태그를 부착한다. 또한 문법 규칙에는 중의성 어절 중 고빈도 중

의성 어절(빈도 1,000 이상)에 대하여 실질 어휘에 대한 규칙을 설정하여 품사 태그를 부착한다. 그 중 3장에서 살펴볼 어절 상관도에 의해 해결되지 않는 어절에 대한 규칙을 추가적으로 설정함으로써 규칙을 수집하는 데 따른 부담을 덜고자 하였다. <그림 1>은 중의성 어절과 개별 어절의 예를 든 것이다.



<그림 1> 중의성 어절과 개별 어절

다음으로 한국어에서 발견되는 일반적인 품사 부착 및 연결 규칙들과, 어절 상관도를 고려한 문법 규칙을 설정하였다. 아래의 규칙은 본 논문에서 설정한 규칙의 일부를 기술한 것이다.

<표 1> 문법 규칙

① 1어절 내의 결합 가능한 품사열 규칙

- NNG
- NNB
- …
- VV … EC
- VA ETN … JKS
- … ETM JKO
- MM EF …
- MAG XSV … EC
- … … ETN
- IC ETM
- …

② 의존명사, 보조용언은 문장의 첫머리에 올 수 없다.
③ 관형어, 부사어는 문장의 끝에 올 수 없다.
④ 보조 용언은 앞 어절이 '어', '게', '지', '고'로 끝난 연결어미 뒤에서 사용된다.
⑤ 종결어미는 기호가 포함되지 않은 한 문장의 가장 마지막 어절에 있어야 한다.
⑥ 부사격 조사 '에게' 앞에 오는 품사는 대명사의 결과가 있으면 대명사로 선택한다.
⑦ '연결어미+보조사'에 나타날 수 있는 연결어미는 제한적이다.→ '는지' , '는가' ,…
……

3. 어절 태그의 비율 추출

2 장에서 살펴 본 문법 규칙들을 통해, UCMA 형태소 분석기를 통한 형태소 분석 결과를 줄여주거나, 품사를 결정하여 하나의 품사태그를 부착할 수 있는 경우가 있음을 알 수 있다. 그러나, 문법 규칙 적용 단계에서 해결할 수 없는 형태소 분석 결과들이 다수 존재한다.

이러한 문제를 해결하기 위하여 본 논문에서는 어절 상관도(eojeol relativity)를 이용하고자 하였다. 어절 상관도에 이용되는 것이 바로 어절 태그이다. 본 논문에서의 어절 태그는 총 7 개로 이루어진다. 어절 태그는 한국어의 품사와 문장성분적 성격을 고려하여, 비슷한 성질을 갖는 어절 내의 품사열들을 묶은 집합 태그(set tag)이다. 예를 들어 '관형사'와 '용언 + 관형형 전성어미'를 하나의 'ME'라는 어절 태그로 설정하는 것으로, 이는 체언을 수식하는 역할을 하고, 바로 뒷 어절이 체언으로 시작하는 품사열이 일반적으로 나타난다는 특징을 가진다. 이러한 것을 고려하여 본 논문에서는 <표 2>와 같이 7 개의 어절 태그를 정의하였다. 또한 어절 태그는 한 어절 내의 품사열 구성과는 상관없이 어절 내의 마지막 품사에 의해 그 성질이 결정되도록 하였다. 즉 '명사+조사, 명사+명사, 동사+명사형 전성어미+조사' 등과 같이 격조사의 경우는 체언류가 격조사의 앞에 위치함으로 이를 체언형 어절(NE)로 판단하는 것이다. 마지막으로 기호(sign)로 끝난 어절도 어절 태그에 포함시켰으며, 공백 어절은 문장의 첫 번째 어절 앞에, 비어 있는 어절을 상정하여 이를 공백 어절로 임시적으로 설정하였다. 문장의 마지막 어절 역시 마찬가지이다. 이러한 공백 어절을 이용함으로써 문장의 앞에 자주 나오는 품사의 비율과 문장의 끝에 쓰이는 품사의 비율을 구할 수 있다.

<표 2> 어절 태그

- NNG+JX, VV+ETN+JKS,…: 체언형 어절(NE)
- VV+EF, NNG+XSV+EC,… : 용언형 어절(PE)
- MM, NNG+JKG, VV+ETM,..: 관형어형 어절(ME)
- MAG, VV+ETN+JKB… : 부사어형 어절(AE)
- IC : 독립어형 어절(IE)
- 기타 기호 : 기타 어절(EE)
- 공백 어절 : null

3.1 어절 태그(ET1) : 어절 태그(ET2)

어절 태그(ET1)과 어절 태그(ET2)간의 공기 확률값은 조건부 확률(수식 1)로 나타낼 수 있다.

[수식 1] $P(pre) = P(ET1 \cap ET2)$
$= P(ET1 \mid ET2) * 1/P(ET2)$

[수식 1]에 가중치 w 를 곱한다. 이 때, 가중치는 값을 보정하는 역할을 한다. 예를 들어 체언형 어절은 모든 경우에 가장 많은 수치를 가지므로, 위 조건부 확률만 이용하게 되면 체언형 어절이 가장 높은 확률을 가지게 된다. 여기에 가중치 w = 1/P(NE)를 곱하게 되면 값이 보정된다. 그리고, 독립어형 어절은 자체 빈도가 가장 낮기 때문에 모든 경우에 가장 낮은 확률을 가진다. 그러나 가중치 w = 1/P(IE)를 곱하게 되면 독립어형 어절(IE) 다음 어절의 어절 태그 중 가장 높은 확률을 가지게 된다. 따라서 최종 수식은 [수식 2]와 같다.

[수식 2] **w*P(pre) = 1/P(ET) * P(pre)**

어절 태그간의 공기 확률을 아래에 정리하였다.

① 체언형 어절(NE)과 다른 어절 태그와의 확률

앞 어절 태그

	P(pre)	w *P(pre)
AE	0.103	0.131
EE	0.079	0.202
IE	0.000	0.129
ME	**0.327**	**0.288**
NE	0.326	0.158
PE	0.080	0.078
null	0.084	0.015

뒤 어절 태그

	P(next)	w*P(next)
AE	0.169	0.208
EE	0.065	0.161
IE	0.000	0.082
ME	0.210	0.179
NE	**0.326**	0.153
PE	0.229	**0.217**
null	0.001	0.000

② 용언형 어절(ME)과 다른 어절 태그와의 확률

앞 어절 태그

	P(pre)	w *P(pre)
AE	0.212	**0.274**
EE	0.017	0.044
IE	0.001	0.168
ME	0.091	0.081
NE	**0.462**	0.228
PE	0.203	0.202
null	0.014	0.003

뒤 어절 태그

	P(next)	w*P(next)
AE	0.066	0.107
EE	0.038	0.122
IE	0.000	0.134
ME	0.182	0.203
NE	0.161	0.099
PE	0.203	**0.253**
null	**0.349**	0.082

3.2 어절 태그(ET) : 어절의 첫 품사 태그(Hpos)

어절 상관도를 측정하기 위한 방법의 하나로 어절 태그간의 확률값을 측정하였다. 그러나 어절 태그간의 확률값만으로는 한국어에서의 어절 상관도를 측정하기 힘들다. 그리하여 본 논문에서는 어절 상관도 측정 시 어절 태그간의 확률 값뿐만 아니라 어절 태그와 다음 어절의 첫 품사와의 공기 확률을 함께 사용한다. 다음 어절의 첫 품사는 어절의 처음에 나타날 수 있는 품사들로 NNG, NNB, NNP, NP, NR, PV, PA, PX, MM, MAG, MAJ, IC 의 12 가지이다. 이 품사와 앞 어절에 나타나는 어절 태그와의 공기 확률을 구해 보면 어절 태그 다음에 올 수 있는 요소들이 확연히 몇 가지로 축약됨을 알 수 있다.

여기에 사용된 수식도 마찬가지로 가중치를 부여한 조건부 확률(수식 3)을 이용한다.

[수식 3] $P(pre) = P(ET \cap Hpos)$
$= P(ET \mid Hpos) * 1/P(Hpos)$
$w*P(pre) = 1/P(ET) * P(pre)$

첫 어절의 품사와 앞에 오는 어절 태그와의 공기 확률을 표로 나타내면 다음과 같다.

① 일반명사(NNG)의 앞 어절 태그 확률 - P(NNG)

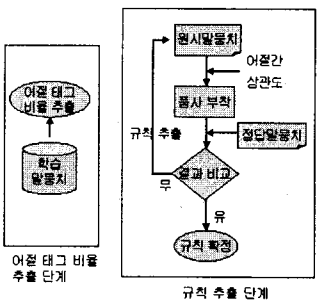
	P(pre)	w * P(pre)
ME	0.306	**0.226**
NE	**0.371**	0.155
PE	0.080	0.101
EE	0.071	0.185
AE	0.122	0.131
null	0.050	0.118
IE	0.000	0.084

② 동사(PV)의 앞 어절 태그 확률 - P(PV)

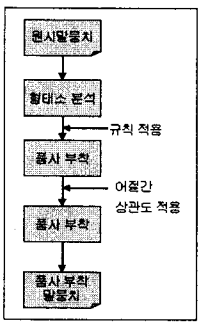
	P(pre)	w * P(pre)
ME	0.006	0.004
NE	**0.541**	0.244
PE	0.131	0.178
EE	0.013	0.037
AE	0.298	**0.346**
null	0.011	0.027
IE	0.001	0.163

4. 시스템 구성도

어절 태그간의 비율을 구하는 단계(어절 상관도)와 규칙을 추출하는 단계는 <그림 1>과 같다. 2~3 장에서 기술하였듯이, UCMA 형태소 분석기에서 나온 형태소 분석 결과를 바탕으로, 문법 규칙과 어절 상관도를 이용하여 품사 태깅 시스템을 구현하였다. 먼저 문법 규칙을 설정하고, 다음으로 어절 태그를 이용한 어절 상관도를 구하여, 시스템의 기본적인 정보로 활용하였다.



<그림 2> 어절 태그 비율과 규칙 추출 단계



<그림 3> 품사부착단계

다음으로 품사 부착 단계의 과정으로, 먼저 형태소 분석을 거친 후 중의성 어절에 대해서 다음의 방법으로 품사를 부착한다. 앞에서 언급하였듯이, 학습 말뭉치를 중심으로 문법 규칙을 설정하여, 문법 규칙을 이용하여 1 차 품사 부착을 한 후, 어절 상관도 비율을 적용해 2 차 품사 부착을 시행한다.

5. 실험 및 평가

본 실험에서 사용된 말뭉치는 150 만 어절의 세종 원시·품사 태그 말뭉치와 3 만 어절의 ETRI 원시·품사 태그 부착 말뭉치이다. 세종 말뭉치를 학습 말뭉치로 활용하였으며, ETRI 말뭉치를 실험 말뭉치로 이용하였다.

먼저 어절 상관도만을 이용한 시스템의 정확률은 다음과 같다.

	학습 말뭉치	실험 말뭉치
소설류	90%	89%
비소설류	91%	90.6%
뉴스기사류	90%	90%

다음은 규칙이 적용된 후 나머지 중의 어절에 대하여 어절간의 상관도를 적용한 시스템의 정확률이다.

	학습 말뭉치	실험 말뭉치
소설류	92%	91%
비소설류	93%	92.6%
뉴스기사류	91%	91%

6. 결론 및 향후 연구

본 논문에서는 UCMA 형태소 분석기의 형태소 분석 결과를 바탕으로, 문법 규칙과 어절 상관도를 이용한 품사 태깅 방법을 제안하였다. 한국어의 일반적인 품사 부착 및 연결 규칙을 파악하여 문법 규칙을 설정함으로써 한국어의 특성을 반영하고자 하였으며, 다음으로 어절 태그를 설정하여 어절 상관도를 계산함으로써 문법 규칙으로 해결하지 못하는 품사 태깅 문제를 해결하고자 하였다.

현재 시스템은 여러 가지 부분에서 수정·보완 중에 있는 단계이므로, 수사, 복합명사, 미등록어 등 형태소 분석 단계에서의 일반적인 문제를 가지고 있다. 그러나 이 부분에 대한 처리가 진행되고 있으므로 어느 정도의 정확률을 가지는 품사 태깅 시스템을 구현할 수 있으리라 보인다.

7. 참고문헌

[1] 류원호, 이상주, 임해창. "어휘 문맥 의존 규칙과 통계 모델을 이용한 한국어 품사 부착 말뭉치 구축 도구", 정보과학회, 1998.

[2] 이정규, 이상주, 임희석, 임해창. "규칙 기반 한국어 품사 태깅을 위한 어휘 규칙 획득의 수작업 최소화 방안", 정보과학회, 1997.

[3] 이운재, 최기선, 김길창. "한국어 문서 태깅 시스템", 정보과학회, 1993.

[4] 이중영, 이기영, 김한우. "어절 간 규칙을 이용한 형태소 중의성 해결", 정보과학회, 1997.

[5] 심준혁, 김준석, 차정원, 이근배. "통계와 규칙을 이용한 강인한 품사 태거", 제 11 회 한글 및 한국어 정보처리 학술대회, 1999.

[6] 안영민, 서영훈. "조사와 어미의 문법 기능을 활용한 품사 태깅 시스템", 제 13 회 한글 및 한국어 정보처리 학술대회, 2001.

[7] 임해창, 임희석, 이상주, 김진동. "자연어 처리를 위한 품사 태깅 시스템의 고찰", 정보과학회지 제 14 권 제 7 호, 1996.

[8] 김재훈, 조정미, 김창현, 서정연, 김길창. "퍼지망을 이용한 한국어 품사 태깅", 제 5 회 한글 및 한국어 정보처리 학술대회 발표논문집, 1993 년.

[9] 김재훈, 김길창. "확률적인 품사태깅 모델에서 언어적 제약조건의 부여", 한국정보과학회 봄 학술대회 발표논문집 제 23 권 제 1 호, 1996 년.

[10] 김재훈. "가중치 망을 이용한 한국어 품사태깅", 한국정보과학회논문지 제 25 권 제 6 호, 1998 년.

[12] 임희동, 서영훈. "어절간 문맥 정보를 이용한 혼합형 품사 태깅", 한글 및 한국어 정보처리 학회, 2000.

[13] 김재한. "한국어 어휘 중의성 해소를 위한 태깅 시스템", 석사학위논문, 울산대학교, 1994.

[14] 임희석. "어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기", 석사학위논문, 고려대학교, 1994.

[15 신상현. "TAKTAG: 통계와 규칙에 기반한 혼합형 한국어 품사 태깅", 석사학위논문, 포항공과대학교, 1996.

[16] 이상주. "자동품사 부착을 위한 새로운 통계적 모형", 박사학위논문, 고려대학교, 1999.

[17 김충원. "의미 정보를 이용한 형태소 중의성 해결", 석사학위논문, 연세대학교, 1994.

[18] Steven Abney. "Part-of-Speech Tagging and Partial Parsing", 1996.

[19] Atro Voutilainen. "A syntax-based part-of-speech analyzer", 1995.

[20] Eric Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", 1995.