세종 계획 말뭉치를 이용한 품사 태거의 성능 개선

김형준°, 임동희, 강승식, 은지현, 장두성
국민대학교 컴퓨터공학부, KT 미래기술연구소
dictions@condition-red.net°, dhlim@kookmin.ac.kr, sskang@kookmin.ac.kr,
jh06@kt.co.kr, dschang@kt.co.kr

Improving Part-of-speech Tagger by using Sejong Corpus

Hyung-Joon Kim°, Dong-Hee Lim, Seung-Shik Kang, Jihyun Eun, Du-Seong Chang
School of Computer Science, Kookmin University; Advanced Technology Laboratory, KT

요 약

품사 태거를 구축할 때 어휘사전 증축이나 변환을 통해 성능 개선을 시도하지만 적당한 품사 태깅 코퍼스의 부재와 태그셋 불일치로 인한 변환 과정에 어려움을 겪고 있다. 본 논문에서는 세종 말뭉치 품사 태깅 코퍼스를 이용하여 품사 태깅용 어휘사전을 증축하고 품사 태거에 적용하여 성능을 개선하는 과정을 기술하였다. 품사 태거의 성능을 개선하기 위하여 세종 코퍼스를 태거의 태그셋에 적합하게 변환하고, 변환된 코퍼스에서 추출된 통계 정보를 품사 태거에서 활용하였다. 세종 코퍼스를 이용하여 품사 태거를 위한 어휘사전을 보강함으로써 품사 태거의 성능을 향상시킬 수 있었다.

1. 서 론

단어의 품사 정보는 구문 분석, 기계 번역 시스템, 정보 검색 시스템 등 자연어 처리 시스템에서 매우 중요한 역할을 한다[1,2,3,10]. 품사 태깅은 문장 내 어휘의 품사를 결정하는 과정에서 각 어휘의 품사별 통계적 출현 비중을 반영하기 때문에 품사 출현빈도 사전이 필요하다. 품사 태거는 문장 내 각 어휘에 대한 형태소 분석 결과의 좌우 문맥 정보를 바탕으로 각 어휘 및 어휘의 품사 정보를 결정한다[4,5,6,12].

이 과정에서 각 어휘의 통계적 출현 비중이 반영되기 때문에 품사 태깅에는 어휘별 품사 출현빈도 사전이 필요하다[7,8,11,13]. 본 논문은 현재까지 구축되어 있는 세종 말뭉치 품사 태깅 코퍼스를 이용하여 품사 태깅 사전을 증축하고, 이 사전을 품사 태거에 적용하여 품사 태거의 성능을 개선하는 방법을 기술한다.

사전 구축 과정은 세종 말뭉치 품사 태깅 코퍼스를 사전 구축용으로 가공하는 작업에서 시작한다[9]. 코퍼스 가공 과정은 코퍼스 구축시 발생한 에러를 교정하고 사전 구축용으로 부적절한 품사나 어휘를 제거하는 작업을 포함한다. 세종 말뭉치의 품사 태그셋과 기존 품사 태거의 태그셋이 다르기 때문에 세종 코퍼스의 품사 태그를 기존 태거의 품사 태그로 변환하는 작업이 필요하다. 이렇게 변환이 끝난 코퍼스를 바탕으로 사전을 구축하고 품사 태거에 적용하면 테스트가 가능하다.

논문의 내용은 다음과 같이 구성하였다. 2장에서 세종 말뭉치를 기존 품사 태거에 적합하게 태그셋을 변환하는 과정을 설명한다. 3장에서는 종전의 사전과 세종 코퍼스로 구축한 사전을 적용했을 때 품사 태거의 성능을 비교하고 성능을 평가한다. 그리고 서로 다른 태그셋의 코퍼스를 변환하여 사전을 구축하는 과정에서 발생하는 문제점을 기술한다.

2. 말뭉치 변환 및 어휘사전 구축

품사 태깅용 어휘 사전을 구축하기 위해 사용된 말뭉치는 2002년도에 세종계획 프로젝트의 결과물로 배포된 1천만 어절 규모의 품사 태깅 말뭉치이다. 세종 말뭉치의 규모는 다음과 같다.

표 1. 세종계획 품사 태깅 말뭉치 규모

세종 말뭉치		
크 기	구 어	23MB
	문 어	221MB
		약 244MB
문장 개수		81만 문장 (41개 폴더, 511 파일)
어절 개수		1천만 어절 (중복 제거 150만 어절)

품사 태깅 사전은 세종 말뭉치를 품사 태거의 태그셋 형태로 변형하고 사전으로 구축하여 이진탐색을 통해 사전을 사용하도록 구현하였다.

2.1. 세종 말뭉치 가공 및 어휘사전 구축

사전 구축에 앞서 세종 말뭉치를 정규화하고 사전 구축의 자동화가 가능하도록 가공할 필요가 있다. 1천만 어절이라는 거대한 규모로 구축된 말뭉치에 다수 포함된 에러 보정도 이 과정에서 함께 이루어진다.

먼저 500여 개에 이르는 세종 말뭉치 파일들을 하나로 병합하고, 어절 별로 분리되어 있는 말뭉치를 라인당 한 문장으로 묶으면서 유실된 숫자, 문장 기호 등의 에러를 복원한다. 그 다음 어절간의 연결 구분자를 '+'에서 공백 문자로 대치하고 일부 잘못된 태그 이름을 보정한다. 이 단계에서 세종 말뭉치의 보정 및 가공은 완료된다.

그러나 형태소 분석기의 데이터로 사용하기 위해서는 형태소 분석 결과 형식에 맞춰 분석 결과를 일부 변형해야 한다. 변형된 어휘 및 변형 빈도는 아래와 같다.

표 2. 세종계획 말뭉치의 분석 결과 변형

세종	KT	빈도
위하/XSA 아	위하/XSA 여	10,685
위하/XSV 아	위하/XSV 여	57,730
위하/VA 아	위하/VA 여	2,075
위하/VX 아	위하/VX 여	9,876
위하/VV 아	위하/VV 여	8,586
에/JKB 대한/NNP	에/JKB 대하/VV ㄴ/ETM	18,936
에/NF 대한/NNP	/NF 에/JKB 대하/VV ㄴ/ETM	154

형태소 분석 결과 변형 과정까지는 단순 작업의 반복이지만 형태소 분석 결과는 연구자의 모듈에 따라 다를 수 있으므로 모든 연구자에게 공통된 작업은 아닐 수 있다. 보정 후 형태소 분석 결과까지 변형된 말뭉치는 사전 구축을 위해 최적화하게 된다. 이 과정은 한자, 숫자, 문장부호를 제거하고 태그가 불명확한 어휘를 삭제한다. 품사가 불명확하여 삭제된 어휘 및 유형은 아래와 같다.

표 3. 불명확한 품사 어휘

NF(명사 추정)	106,108개
NV(용언 추정)	5,719개
UNIK(미등록어 추정)	3,043개

불필요한 어휘를 제거한 후 태그셋 변환에 사용된 세종 태그셋과 품사 태거의 태그셋의 상호 변환표는 아래와 같다. 이 태그셋 변환표는 세분류를 단일화한 축약형이다.

표 4. 세종 태그셋 변환표

세종 태그셋			KT 태그셋
대분류	소분류	세분류	
체언	명사 NN	일반명사 NNG	Mv*
		고유명사 NNP	Mg
		의존명사 NNB	Me*
	대명사 NP		T
	수사 NR		S
용언	동사 VV		D*
	형용사 VA		H
	보조용언 VX		d/h
수식언	관형사 MM		G
	부사 MA	MA*	B*
독립언	감탄사 IC		K
관계언	격조사 JK	JK*	i*
	보조사 JX		jv
	접속조사 JC		jj
의존형태		E*/X*	s/z/p/M*/e*
기호		S*/N*	m*/무시

불명확한 어휘를 삭제한 후 전체 세종 어휘 213,857개(숫자/영문자 제외) 중 98,715개의 항목을 변환된 태그에 맞춰 어휘사전으로 구축하였다. 품사태그 변환 과정에서 원래 다른 태그였던 어휘가 동일한 태그로 변환되는 경우도 고려하였으므로 중복된 어휘는 없다.

태그 변환 전(세종 태그셋)	태그 변환 후(품사 태거)
NR 하나 13463	S 하나 13463
NR 하나씩 418	S 하나씩 418
NR 하나쯤 50	S 하나쯤 50
NR 하나하나 132	S 하나하나 132
NR 한 543	S 한 543
NR 한둘 71	S 한둘 71
SH 래 1	_ 래 1
SL 가 2	Mg 가 2
SL 을 1	Mg 을 1
SL 져 2	Mg 져 2
VA ᄎᆞ 1	H ᄎᆞ 1
VA 가깝 2118	H 가깝 2118
VA 가냘프 64	H 가냘프 64
VA 가느다랗 109	H 가느다랗 109
VA 가느스름하 1	H 가느스름하 1
VA 가늘 361	H 가늘 361
VA 가당찮 13	H 가당찮 13
VA 가득차 168	H 가득차 168

그림 1. 품사 태그 변환 예

태그셋이 변환된 세종 말뭉치는 품사 전이 빈도를 계산하는데도 사용된다. 이렇게 가공된 세종 말뭉치를 사전으로 구축하고 이진탐색을 통해 탐색할 수 있도록 구현하였다. '태깅 어휘사전'은 기존 태거에서 해쉬 사전으로 구축되었다. 그러나 기존의 해쉬 사전에서 허용되는 어휘수는

최대 약 3만개로 제한되어 세종 말뭉치로부터 구축된 98,715개의 어휘사전을 동일한 플랫폼으로 구현할 수가 없다. 따라서 어휘사전 탐색 함수를 이진탐색 기법으로 구현하여 기존의 해쉬 사전 탐색 함수를 재정의하는 형태를 통해 사용할 수 있게 하였다.

2.2. 구축된 어휘사전의 빈도 통계

세종 말뭉치로부터 어휘사전을 구축할 때 빈도가 낮은 것을 제외하여 어휘사전의 크기를 줄이고 사전 탐색 효율을 높일 수 있다. 이러한 목적으로 98,715개 어휘의 빈도별 어휘수 통계를 조사하였으며, 그 결과는 표 5와 같다.

표 5. 세종 어휘사전의 고빈도 어휘수

빈도수	어휘수	누적 어휘수
5,331,521	1	1
100	73	10,610
50	185	16,301
40	248	18,498
30	401	21,775
20	664	26,838
10	1,671	37,715
9	1,843	39,558
8	2,257	41,815
7	2,685	44,500
6	3,199	47,699
5	4,006	51,705
4	5,298	57,003
3	7,196	64,199
2	11,631	75,830
1	22,880	98,710
0	5	98,715

고빈도 어휘 1만여 개 규모의 어휘사전을 구축하고자 할 때 빈도 100 이상인 10,610개의 어휘들만 어휘사전으로 구축할 수 있다. 위 표에서 빈도수 0인 것은 INI, FIN 등 실제 어휘가 아닌 것이다.

3. 실험 및 성능 평가

품사 태깅 데이터를 세종 말뭉치에서 추출한 어휘 사전을 적용하여 품사 태거의 성능을 실험한 결과는 표 6과 같다.[1] 표 6에서 Case1은 초등학교 교과서와 신문기사 등의 문서로 이루어져 있으며 Case2는 전산학 분야의 논문 모음으로 2,441 어절 규모이다. 첫번째 실험 문서 Case1에 대한 태깅 정확률은 "세종 말뭉치에서 추출한 어휘 사전"을 적용했을 때 더 높은 성능을 보인다. 그러나 두번째 실험 문서 Case2에 대해서는 원래 어휘 사전의 정확률이 높았다. 그 이유를 분석해 본 결과 세종 데이터의 경우 "본 논문은 …"이라는 문장에서 '본'에 대한 통계 데이터가 매우 적어서 '본'을 동사로 태깅하는 오류가 발생한다. 그런데 실험 문서가 논문이어서 '본'이 자주 출현했기 때문이었다. 이와 같이 세종 말뭉치의 태그셋 변환표에 따라 변환하여 어휘 사전을 구축할 때 발생하는 문제점은 아래와 같다.

표 6. 태깅 어휘사전 교체 실험 결과 – 정확률

	태거의 원래 어휘사전		세종 추출 어휘사전	
	공통 에러	비공통 에러	공통 에러	비공통 에러
Case1	92.89%		93.84%	
	55	72	55	55
Case2	95.82%		95.70%	
	66	36	66	39

- 어기 명사 "적절"의 "적절한"이 "적절+하+ㄴ"으로 분석되지 않고 "적절하+ㄴ"로 태깅됨
- 보조용언 "뻔하다"(예:"~이 뻔하다") 본용언으로 태깅됨
- 관형사 '본', '볼' 이 세종 말뭉치에 출현하지 않아 대부분 명사로 태깅됨
- "~에 대한"의 "대한"이 명사로 태깅됨
- '칠'이 거의 수사로 태깅됨
- 동사/보조용언의 "있다"가 형용사/보조용언으로 태깅
- 접속조사/보조사의 "~이나"가 잘못 태깅되는 경우 다수
- 관형사 혹은 형용사로 태깅 되어야 할 "다른"이 대부분 관형사로 태깅됨
- "~이 아니다"의 '이'가 주격조사로 태깅됨

이번 연구에서는 이러한 문제점들을 감수할 수밖에 없었으나 차후 이러한 문제점들이 해결되면 성능 개선의 효과가 좀더 높을 것으로 예상된다. 다만, 이러한 문제점을 해결하는데 있어 단순히 어휘 사전의 데이터만을 바꿈으로써 해결하기 보다는 태깅 알고리즘 차원에서의 수정이 필요할 수도 있다.

태거의 성능 측정 결과로 정확률이 실제보다 높게 계산되었는데, 그 이유는 테스트 데이터를 문법적으로 옳은 문장으로 선별하지 실험을 수행한 것이 아니라, 논문의 장 및 절의 제목과 개조식 문장, 오류어가 포함되어 있는 일반 문서를 그대로 사용했기 때문이다. 따라서 문법적으로

[1] 성능 실험에 사용된 품사 태거는 KT 태거이다.

옳은 한글 문장들에 대해서만 실험을 할 경우에 위 결과보다 정확률이 약간 낮아질 수 있다.

4. 결론

본 논문은 세종 말뭉치 품사 태깅 코퍼스를 바탕으로 품사 태깅 사전을 구축하여 이를 품사 태거에 적용함으로써 성능 개선을 시도하였다. 구축한 사전은 품사 변환 과정을 거치면서 많은 문제점을 내포하였음에도 품사 태거의 태깅 정확률을 향상시키면서 단순히 사전만 교체하는 작업만으로 성능 개선을 기대할 수 있음을 확인하였다. 품사 태거를 구축하는데 사용된 태깅 코퍼스에 따라 품사 태거의 성능 개선 폭이 향상될 수 있을 것이다.

코퍼스의 변환 과정은 코퍼스 구축 과정에서 발생한 에러를 교정하고 사전 구축에 부적합한 어휘 및 품사를 제거하는 과정을 거친 후 기존의 태그를 새로운 태그셋에 맞게 변환표에 의해 변환한다. 이 과정에서 두 시스템의 출력 양식도 감안하여 태깅 결과를 변환하였다. 이렇게 체계적인 과정을 거쳐 태그셋을 변환하였음에도 문제점들이 발생하였다.

참고문헌

[1] 이하규, 김영택, “통계 정보에 기반을 둔 한국어 어휘 중의성 해소”, 한국통신학회 논문지, 제19권 제2호, pp.265-275, 1994.

[2] 이하규, “어말-어두 공기 정보를 이용한 한국어 어휘 중의성 해소”, 정보과학회논문지(B), 제24권 제1호, pp.82-89, 1997.

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

[4] 임희석, 김진동, 임해창, “어절 태그 변형 규칙을 이용한 한국어 품사 태거”, 정보과학회논문지(B), 제24권, 제6호, pp.673-684, 1997.

[5] 김진동, 임희석, 임해창, “Twoply HMM: 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델”, 정보과학회논문지(B) 제24권 제12호, pp.1502-1512, 1997.

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

[7] 김재훈, "가중치 망을 이용한 한국어 품사 태깅",정보과학회논문지(B), 제25권, 제6호, pp.951-959, 1998.

[8] 임해창, 임희석, 이상주, 김진동, “자연어 처리를 위한 품사 태깅 시스템의 고찰”, 정보과학회지, 제14권, 제7호, pp.36-57, 1996.

[9] 문화관광부, 21세기 세종계획 국어 기초자료 구축, 문화관광부, 1998-1999.

[10] Eric Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging", Computational Linguistics, Vol.21, No.4, pp.543-564, 1995.

[11] 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.

[12] Kenneth Ward Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", Proc. of 2nd Conference on Applied NLP, pp.136-143, 1988.

[13] Adwait Ratnaparkhi, "A Maximum Entropy Model for Part-of-Speech Tagging", Proc. of the Empirical Methods in Natural Language Processing Conference, pp.133-142, 1996.