CRFs를 이용한 구문분석기의 오류 분석 및 자질 추천

신창욱○, 차정원
창원대학교
papower1@changwon.ac.kr, chajeongwon@gmail.com

Error analysis and feature recomandation of dependency parser using CRFs

Chang-Uk Shin○, Jeong-Won Cha
Changwon National University

요 약

우리는 CRFs를 이용한 구문분석기에서 자질부족으로 인해 흔히 발생할 수 있는 오류에 대해서 분석하였다. 피지배소의 기능태그와 지배소의 구문태그 쌍이 코퍼스에서 어떤 분포로 나타나는지 분석하고, 그 분석 결과를 바탕으로 자질을 추천하였다. 또한, 조용사를 활용한 자질에 대해서 추천하였다. 앞으로 이런 자질을 이용한 구문분석기를 작성하여 성능을 향상시킬 수 있으리라 생각된다.

1. 서 론

CRFs(Conditional Random Fields)[1]를 이용한 자연어처리는 높은 성능으로 꾸준한 관심을 받고 있다. 하지만, CRFs를 이용한 구문분석에는 뚜렷한 한계가 존재한다. 자질의 수를 무한히 늘릴 수 없음이나, 경우에 따른 자질 확대 축소가 제한되는 CRFs의 특성상 일부 특정 조건에 따른 수식구조를 분석해내기 힘들기 때문이다. 다양한 수식구조가 나타날 수 있는 한국어의 특성 때문에도 그렇다. 본 논문은 이러한 구문분석 오류에 대해서 조사하고, 그 것들을 해결하기 위해 사용할 수 있는 자질을 제안한다.

2. 이전연구

CRFs는 2001년도에 [1]로부터 제안되었다. HMM의 경우 서로 연관된 여러 개의 자질에 대해서 관측 확률을 계산하는 것이 많은 양의 코퍼스를 필요로 하기 때문에 사용이 까다로운데, 그러한 이유로 CRFs가 대안으로써 인기를 끌게 되었다. 다단계 구단위화(Cascaded Chunking)[2] 방법은 1991년에 처음 제안되어 [3]에서 한국어에도 적용된 바 있다. 다단계 구단위화 모델을 base모델로 삼고, base모델과 CRFs에서 발생하기 쉬운 오류를 분석하고 해결하기 위한 자질을 제안한다.

3. 본 론

우리는 먼저, CRFs의 한계에 초점을 맞추고 연구를 진행하였다. 세종 구문분석 코퍼스에서 피지배소-지배소 태그별 출현 빈도를 분석하였고, 한국어 문법에 위배된다고 판단되는 4쌍에 대해서 상세히 분석하였다. 이 때, 이미 알고 있는 CRFs의 특징과 다단계 구단위화의 특성을 고려하였다. 분석을 통해 상위 4쌍 총 4,433 어절의 후보 중 많은 수의 어절이 같은 유형을 보임을 확인하였고, 그 것을 처리하기 위한 자질을 추천하였다. 또한 첫 형태소가 NNG 또는 XR일 때, XSA, XSV 태그로 VP가 되는 어절들을 따로 처리할 수 있도록 구문태그를 일부 수정하여 성능을 향상시키고자 한다.

4. 실 험

사용된 코퍼스는 세종 구문코퍼스를 직접 정제한 후 남은 50,450문장이다. 학습은 5-fold cross validation을 수행하였다. base모델을 먼저 학습/평가하여 성능을 확인하고, 발생한 오류를 분석하여 자질을 선정하였다. 평가 성능은 어절의 LAS(Labeled Attachment Score)를 사용하였다.

[표 1] base모델의 성능

	문장 수	어절성능	문장성능
1	10090	84.36%	26.64%
2	10090	84.30%	26.21%
3	10090	84.50%	26.92%
4	10090	84.45%	26.27%
5	10090	84.33%	26.92%
평균		84.39%	26.59%

base모델은 자질로써 각 어절의 첫 번째 형태소와 마지막 형태소, 구문표지와 용언 자질을 사용하였다. 용언 자질은 다음 어절의 구문표지가 VP인 경우 '1', 그렇지 않은 경우 '-'을 사용하였다. 조용사 자질은 해당 어절에 조용사가 존재하는 경우 해당 태그를 자질로 주었다. 이렇게 5개의 자질을 적절히 조합하여 총 32개의 자질을 사용하였다.

[표 2] base모델의 자질

형태소분석	1	2	구문표지	용언	조용사
그/NP+는/JX	NNG	JKB	NP	-	-
물/NNG+이/JKS	NNG	JKS	NP	-	-
필수/NNG+처럼/JKB	NNG	JKB	NP	1	-
말/NNG+하/XSV+어/EC	VV	EC	VP	-	XSV

다음은 base모델에서 오류로 많이 나타날 법한 태그쌍을 선정하기 위해서, 먼저 정답 코퍼스의 태그별 출현빈도를 조사하였다. 아래의 표는 피지배소의 기능태그별 해당 어절의 지배소의 구문태그를 센 것이다.

[표 3] 정답 코퍼스 내 구문/기능태그 출현빈도

	AP	DP	IP	NP	VNP	VP	X
-	1003	136	127	82674	10752	130477	499
AJT	1064	8	4	**2201**	1844	76750	2
CMP	2	0	3	62	194	10089	0
CNJ	27	0	0	15216	457	376	18
INT	0	0	3	38	26	336	0
MOD	62	5	0	109450	16526	1973	4
OBJ	**22**	0	2	**911**	491	60588	0
PRN	0	0	0	31	7	3	3
SBJ	**679**	9	6	**1255**	5487	72303	2

여기서 우리는 오류를 만들어 낼 것이라고 예상되는 태그쌍을 다시 추려서 정리하였다.

[표 4] 오류를 생성할 것이라 예상되는 태그 쌍의 빈도

피지배소	지배소	개수	피지배소 수	비율
AJT	NP	2,201	81,873	2.69%
OBJ	AP	22	62,014	0.04%
	NP	911		1.47%
SBJ	NP	1255	79,739	1.57%
계		4,433	223,626	1.27%

※ 비율은 (개수/피지배소 수)

분석결과 총 5개의 태그쌍, 4,433개의 어절이 오류를 발생시킬 수 있으리라 생각되었다. 그것들은 다시 피지배소-지배소쌍에 따라서 오류 유형이 구분될 수 있었다.

OBJ-AP쌍은 "남부지방은 물론 북부지방까지"와 같은 예제나, "지속적인 성장을 거듭, 1위로 발돋움..."과 같은 예제가 있다. 첫 예제에서 '물론'은 일반부사이지만 앞의 '남부지방은'을 목적격 피지배소로 갖는다. 두 번째 예제에서는 '성장을'이 '거듭,'의 목적어 절이 된다. SBJ-AP쌍도 마찬가지인데, "수 없이 많은 사람들"에서 '수'가 '없이'를 수식하는 것과 같다.

이런 문제들은 테스트환경에서 AP(부사구)가 체언구로부터 주격, 목적격 수식을 받을 확률을 증가시킨다. 따라서 어떤 경우에 AP를 수식하는지, 어떤 자질이 그것을 구분하는데 도움이 될지를 고민할 필요가 있다. 우리는 'NNG 없이/MAG'를 '없/VA+이/EC'로 수정하여 이 문제를 해결하고자 한다. 또한, 이 OBJ(SBJ)-AP쌍에서 AP에 해당하는 태그나 어휘를 정리하거나, 지배소가 쉼표를 가지는지 여부를 자질로 추가하면 성능을 향상시킬 수 있을 것이라 생각한다.

AJT-NP, OBJ-NP, SBJ-NP는 다소 유사한 형태를 보인다. AJT-NP는 "서울에서 부산까지" 등의 예를 들 수 있다. '서울에서'는 '부산까지'라는 부사를 수식하는데, 부사가 대부분의 경우에 용언구를 수식하기 때문에 자질이 부족하면 오류를 발생시킬 수 있다.

OBJ-NP쌍은 지배소가 조사 '으로'를 대동하는 경우 등에서 발생하는데, 예를 들면 다음과 같은 문장이다. "사회자를 중심으로 에워쌌다". 이 문장에서 '사회자를'은 '중심으로'를, '중심으로'는 '에워쌌다'를 수식한다. '사회자를'이 NP_OBJ이고 '중심으로'는 NP태그를 갖는다. 목적어절은 대부분의 경우 용언구를 지배소로 갖는 경향이 있다. 그러므로, 학습 코퍼스 내 목적어 절 중 체언구를 지배소로 갖는 데이터가 오류를 발생시킬 가능성이 있다.

SBJ-NP의 경우도 마찬가지이다. "류큐는 일본의 소수민족으로 ..."와 같은 문장이 있을 때, '류큐는'은 '소수민족으로'를 지배소로 갖는다. 하지만, '소수민족으로'가 NP이기 때문에, 주격절이 NP를 수식할 확률이 발생하고, 이는 오류가 될 가능성이 있다.

AJT-NP, OBJ-NP, SBJ-NP쌍에서 공통적으로 나타나는 현상은 '으로' 등의 조사가 명사절인 지배소절을 용언절처럼 활용되게 만든다는 점이다. 이런 조사를 자질로 사용할 수 있을 것이라 생각된다.

VP만을 따로 떼어놓고 분석해보면, NNG 또는 XR이 XSA, XSV와 합하여 VP가 되는 경우가 있다. 이 경우는 가장 앞 형태소 NNG, XR이 다른 피지배소로부터 수식받을 가능성과, XSA, XSV와 합쳐져 용언으로 동작하는 특징 둘 모두를 갖고 있다. CRFs와 다단계 구 단위화를 사용하는 상황에서 우리는 XSA, XSV가 어절을 단순히 VA, VV로 만들어 주는 것이 아니라, 용언구와 체언구 둘 모두의 특징을 가지는 게 아닌가 생각하였다. 따라서 우리는 VCP와 마찬가지로 XSA, XSV가 포함된 어절을 VNP로 수정하여 성능을 향상시키고자 한다.

5. 결 론

우리는 정답 코퍼스 내에 존재하는 여러 형태의 데이터 중 일부가 실제 환경에서 의도치 않는 결과를 만들어 낼 수 있다고 생각하였다.

그런 결론을 내리기 위해서 우리는 정답 코퍼스 내에 문제를 발생시킬 법한 피지배소 기능태그 - 지배소 구문태그 쌍을 수집 및 분석하였다.

그것들 중에는 일반부사가 목적격 피지배소를 가지는 경우나, 목적격, 주격, 부사격 구가 명사구를 수식하는 경우가 다수 포함되어 있었다. 각각의 경우에 대해 도움을 줄 수 있는 자질에 대해서 기술하였다.

또한 조용사가 CRFs 구문분석에 미칠 수 있는 영향에 대해서 이야기하였다.

향후에는 이러한 자질을 학습해 결과를 도출해보고, 효과를 검증해 볼 예정이다.

참 고 문 헌

[1] John Lafferty, Andrew McCallum and Femando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", ICML, pp. 282-289, 2001

[2] Steven Abney, “Parsing By Chunking,” In PrincipleBased Parsing. Kluwer Academic Publishers, 1991.

[3] 오진영, 차정원, “다단계 구단위화를 이용한 고속 한국어 의존구조 분석”, 한국어시뮬레이션학회논문지, Vol. 19, No. 1, pp. 103-111, 2010