형태소 분석 결과를 사용한 통사 지표 자동 부여

박정열†, 함영균 ○‡, 임경태‡, 김영식§, 최기선‡§

†IRISA, 프랑스 렌느 1 대학 ‡KAIST, 웹사이언스공학 §KAIST, 전산학과

†jungyeul.park@univ-rennes1.fr ‡§{kyungtaelim, hahmyg, twilight, kschoi}@kaist.ac.kr

Assigning automatically syntactic tags using morphological analysis

Jungyeul Park†, YoungGyun Hahm○‡, Kyungtae Lim‡, Youngsik Kim§, Key-Sun Choi‡§

†IRISA, Université de Rennes 1 ‡KAIST, WebST §KAIST, Dept. of CS

요 약

이 논문에서는 형태소 분석 결과로 부터 통사 지표를 자동으로 부여하는 방법에 대해 설명한다. 세종 구문 분석 말뭉치에서 제공하는 통사 지표를 학습하고 테스트하여 94.78%의 정확도를 얻었다. 또한 세종 말뭉치에서 학습한 내용을 구조적으로 다른 카이스트 구문 분석 말뭉치에 적용하는 방법을 소개한다. 통사 지표 정보는 구문 분석 전처리로 사용되어 구문 분석 성능을 향상 시킬 수 있다. 실제로 새로이 통사 지표를 부여받은 카이스트 구문 분석 말뭉치는 구구조 구문 분석에서 성능 향상을 보였다.

1. 서 론

문장 성분이란 문장 내에서 어휘가 하는 통사적 역할로 한국어에서는 주어, 목적어, 보어, 관형어, 부사어, 서술어 및 독립어 등이 있다. 통사 지표는 이러한 문장 성분을 표시하는 지표로 구문 분석 말뭉치에서 통사적 정보를 표시하는 장치이다. 세종 구문 분석 말뭉치[1]의 경우에는 NP-SBJ, NP-OBJ와 같이 해당 어휘가 주어인지 목적어인지를 명시적으로 어휘의 통사적 역할을 표시한다. 따라서 “엠마누엘 웅가로가”와 같이 주격조사 “가”로 끝나는 경우에는 두 어절의 지배하는 노드에 NP-SBJ를 표시하며 문장에서 주어라는 것을 알 수 있다. 유펜 한국어 구문 분석 말뭉치[2]의 경우에도 비슷한 방법으로 문장 성분에 대한 표시를 하여 어떤 부분이 주어인지 목적어인지 명확히 알 수 있게 한다. 한국어와 같은 교착어는 격조사를 사용하여 문장 성분을 표시할 수 있는 외형적 표기 방법을 가지고 있기 때문에 쉽게 어휘의 문장 성분을 추측할 수 있다고 하지만 격조사의 자리에 보조사가 사용되는 경우에는 문맥을 이해하지 않고는 쉽게 문장 성분을 찾아내기가 어렵다.

이 논문에서는 구문 분석 말뭉치에서 제공하는 통사 지표를 학습하고 이를 자동으로 부여하는 방법을 연구한다. 형태소 분석 결과로 부터 통사 지표를 자동으로 부여하기 때문에 카이스트 구문 분석 말뭉치[3]와 같이 문장 성분을 표시하는 장치가 없는 구문 분석 말뭉치에도 응용할 수 있다. 또한 통사 지표 정보가 구문 분석 전처리로 사용될 수 있기 때문에 통사 지표를 부여하여 구문 분석 성능을 향상하여 말뭉치의 효능을 높일 수 있으리라 본다.

기존 연구[1]에서는 세종 구문 분석 말뭉치를 대상으로 학습하여 같은 말뭉치에 적용하여 실험하였다. 세종 구문 분석 말뭉치는 이미 통사 지표가 명시되어 있기 때문에 이러한 학습이 가능하다. 하지만, 카이스트 말뭉치에 통사 지표를 자동으로 부여하려면 세종 말뭉치와 같이 통사 지표가 있는 다른 말뭉치에서 학습하고 구조적 차이를 해결한 다음에 적용할 수 있다.

이후 이 논문에서는 2장에서 세종 구문 분석 말뭉치에서 통사 지표를 자동으로 부여하는 방법에 대해 논의한다. 3장에서는 세종 구문 분석 말뭉치에서 학습한 내용을 카이스트 말뭉치에 적용하는 방법을 제안하고 해당 결과를 응용하는 방법을 모색한다. 마지막으로 4장에서는 결론 및 향후 과제를 제안한다.

2. 세종 구문 분석 말뭉치에서 통사 지표 자동 부여

이 절에서는 세종 구문 분석 말뭉치를 사용하여 자질 선택을 통해 통사 지표를 학습하고 자동으로 부여하는 방법을 테스트한다.

[1] http://www.sejong.or.kr

[2] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T09

[3] http://www-nlp.kaist.ac.kr

기본 시스템	복합 자질 (거리 1)	복합 자질(거리 2)
0.9303558	0.94786656	0.94686913
-SBJ: 0.9330519	-SBJ:0.9193297	-SBJ:0.91600233
-OBJ: 0.8607627	-OBJ:0.96486485	-OBJ:0.9586133
-AJT: 0.9073965	-AJT:0.96212894	-AJT:0.96179914
NP: 0.9033422	-NP::0.925195	-NP::0.9233353

표 1. 세종 구문 분석 말뭉치에 통사 지표 부착 실험: 전체 정확도, -SBJ, -OBJ, -AJT 개별 통사 지표 부착 정확도, NP 관련 지표 부착 정확도(통사 지표 부착된 지표 포함).

2.1 자질 선택

[1]의 경우에는 기본 모델의 자질로 구문 분석 말뭉치에서 제공하는 어절의 형태소 분석을 사용한다. 어절 끝에 위치한 문장 부호를 제외한 n개의 형태소 m_i ($1 \le i \le n$)가 하나의 어절을 구성한다고 가정할 때, 세가지 자질 [$Feature_1$, $Feature_2$, $Feature_3$]은 m_1, m_{n-1}, m_n의 위치에 있는 품사(POS)를 기본 자질로 사용하였다.[4] 예를 들어, "웅가로/NNP+가/JKS"는 [NNP, -, JKS]를 기본 자질로 가진다. $Feature_2$의 위치에 표시된 [-]는 해당 자질이 현재 어절이 없다는 의미이다. 이 논문에서도 위와 동일한 방법으로 학습에 필요한 기본 자질을 구축한다. 또한, 복합 자질이라는 개념으로 통사 지표 부착 성능을 향상할 수 있는데 복합 자질이란 하나의 어절에서만 자질을 추출하는 것이 아니라 해당 어절에서 주변 어절의 품사까지 고려할 수 있다. 이 논문에서는 해당 어절에서 거리 k까지$(m+[1..k])_n$의 최후 형태소를 복합 자질로 사용한다. 예를 들어, "엠마누엘/NNP 웅가로/NNP+가/JKS"의 경우에 "엠마누엘"을 기준으로 거리 1 뒤에 오는 어절의 최후 형태소 $(m+1)_n$는 "가"의 JKS이다.

2.2 세종 구문 분석 말뭉치에서 실험

이 논문에서는 세종 구문 분석 말뭉치의 표준 학습 데이터셋을 사용하여 학습하고, 개발 데이터셋을 통해 최적의 자질을 찾은 다음, 평가 데이터셋을 사용하여 테스트한다. 세종 말뭉치의 표준 데이터셋은 말뭉치를 80:10:10의 비율로 분할하여 사용한다.[5] 개발 데이터셋에서 거리 2까지를 고려할 때 가장 좋은 결과를 얻었기 때문에 평가 데이터셋은 거리 2 뒤에 오는 어절까지 테스트한다. 표 1은 CRF[2]를 사용하여 학습한 결과로 어절 수준에서 94.78%의 정확도를 얻었다. 평가 데이터셋에서는 개발 데이터셋과 달리 거리 1 뒤에 오는 어절까지 학습할 때 가장 높은 결과를 얻었다. 또한, 이는 기존 연구[1]이 얻는 정확도 93.90% 보다 높은 결과이다. 전반적으로 평가할 때 주어에 대한 통사 지표가 다른 문장 성분에 비해 가장 정확도가 낮다. 이는 주어의 경우에는 모호성이 많은 보조사가 주격조사 역할을 자주하기 때문으로 분석된다.

[4] if (n==1) {$FEATURE_3 = POS_n$;} else if (n==2) {$FEATURE_1 = POS_1$; $FEATURE_3 = POS_n$;} else if (n>2) {$FEATURE_1 = POS_1$; $FEATURE_2 = POS_{n-1}$; $FEATURE_3 = POS_n$;}

[5] 세종 구문 분석 말뭉치를 표준 데이터셋으로 분할하는 스크립트를 제공할 예정이다.

3. 카이스트 구문 분석 말뭉치에 통사 지표 자동 부여 적용 및 응용

3.1 두 구문 분석 말뭉치의 차이점

세종 말뭉치에서 학습한 내용을 카이스트 구문 분석 말뭉치에 적용하기 위해서는 다음 두가지 사항을 먼저 고려해야 한다. ㄱ) 세종 말뭉치와 카이스트 말뭉치에서 사용하는 태그셋이 다르다. ㄴ) 세종 말뭉치와 카이스트 말뭉치는 분석 구조가 다르다. 세종 말뭉치와 비교하여 카이스트 말뭉치의 가장 큰 특징으로 문법 형태소가 분리되어 분석되어 있다. 또한, 세종 구문 분석 말뭉치는 하나의 어절이 하나의 구구문 지표의 지배를 받지만 카이스트 구문 분석 말뭉치의 경우에는 하나의 어절에서 어휘 형태소와 문법 형태소가 분리되어 해당 문법 행태소가 같은 어절의 어휘 형태소의 구구문 지표와 함께 지배를 받는게 아니라 구구문 지표의 부모 지표들에 의해 지배를 받는다.

3.2 태그셋 및 구조적 차이 해결 방법

이 논문에서는 두 구문 분석 말뭉치에서 사용하는 태그셋의 차이는 테그셋 매핑으로 해결하였다. 두 말뭉치의 구조적 차이는 다음과 같은 방법으로 통사 지표를 자동으로 부여한다. ㄱ) 먼저, 카이스트 구문 분석 말뭉치에서 종단 노드와 종단 노드의 부모 노드(문장 성분 표시가 생략된 통사 지표)를 추출한다. ㄴ) 이를 세종 태그셋에 맞게 변환하고, 어절 단위로 합친 다음, CRF로 학습한 내용을 적용한다. ㄷ) 새로이 부착된 통사 지표에서 체언구 관련 통사 지표(NP-*)를 카이스트 구문 분석 말뭉치에 부여한다. 세종 구문 분석 말뭉치에서 용언이 명사형 전성 어미가 붙어 명사형으로 전성되고 조사가 붙어 격을 받는 경우에는 용언구의 흔적이 표시되는데(VP-SBJ, VP-OBJ, VP-AJT), 카이스트 구문 분석 말뭉치에는 NP-SBJ, NP-OBJ, NP-AJT로 변환하여 부여한다. 그외, 예외로 "있을 수 있다"의 수/NNB와 같은 경우에는 세종 말뭉치에서는 주어(NP-SBJ)로 표시하지만, 카이스트 말뭉치에서는 AUX의 일환으로 표시한다. 따라서 수/NNB와 같은 의존 명사가 격조사 없는 상황에서 통사 지표를 부여받은 경우에는 이를 무시한다. 그림 1은 위의 과정을 예문으로 설명한다. 그림 2은 최종 변환된 내용을 말뭉치에 적용한 예이다.

통사 지표 정보를 전체 카이스트 구문 분석 말뭉치에 새로이 부여를 하였지만, 이미 정의된 평가 데이터셋이

없기 때문에 이를 자동으로 평가하기가 힘들다. 따라서 이 논문에서는 카이스트 말뭉치에서 첫 번째 파일을 대상으로 수작업으로 평가하였다. 새로이 부여된 NP 관련 비종단 노드 252개를 사람이 평가한 결과 80.56%의 정확도를 얻었다.

3.3 통사 지표를 사용한 구문 분석 적용

새로이 부여된 통사 지표 정보의 유용성을 테스트하기 위해 이 논문에서는 카이스트 말뭉치를 구구조 구문 분석에 적용하였다.[6] 카이스트 말뭉치의 표준 데이터셋은 말뭉치를 80:10:10의 비율로 분할하여 사용한다. 다음 표는 통사 지표가 있고 없고에 따른 테스트 데이터셋에 대한 구문 분석 결과의 차이를 보여준다.[7]

	기본 시스템	통사 지표 부착
전체	80.51%	82.35%
길이<=40	80.84%	82.75%

표 2. 통사 지표를 부착한 구문 분석 결과

4. 결론 및 향후 과제

이 논문에서는 형태소 분석 정보를 사용하여 통사 지표를 자동으로 부여하는 방법에 대해 설명하였다. 세종 구문 분석 말뭉치를 학습하여 테스트하였고 이를 카이스트 구문 분석 말뭉치에 적용하였다. 통사 지표는 구문 분석 전처리 또는 부가 정보로 사용되어 구문 분석 성능을 향상 시킬 수 있다. 특히, 카이스트 말뭉치에서는 이 논문에서 제안하는 방법을 통해 구구조 구문 분석에서 성능 향상을 보였다. 또한, 의존구조로 변환할 때 해당 통사 지표를 사용하여 레이블 정보를 얻을 수 있어 UAS 및 LAS를 모두 계산할 수 있다[5,6]. 이 논문에서 제안된 방법을 통해 앞으로 카이스트 구문 분석 말뭉치도 의존구조 변환할 때 통사 지표가 레이블 정보로 사용될 수 있을 것으로 예상한다.

참 고 문 헌

[1] Jin Young Oh, Yo-Sub Han, Jungyeul Park, Jeong-Won Cha. Predicting Phrase-Level Tags Using Entropy Inspired Discriminative Models. *Proceedings of ICISA 2011*.

[2] John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of ICML 2001*.

[3] Slav Petrov, Leon Barrett, Romain Thibaux, Dan Klein. Learning Accurate, Compact, and Interpretable Tree Annotation, *Proceedings of COLING-ACL 2006*.

[4] Tagyoung Chung, Matt Post, Daniel Gildea. Factors Affecting the Accuracy of Korean Parsing. *Proceedings of Workshop on SPMRL 2010*.

[5] 오진영,차정원. 다단계 구단위화를 이용한 고속 한국어 의존구조 분석.한국시뮬레이션학회 논문지. 19권 1호.

[6] Jinho D. Choi, Martha Palmer. Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing. *Proceedings of Workshop on SPMRL 2011.*

하기야/maj	IP
짐승/ncn	NP
+도/jxc	VP
잘/mag	ADVP
가르치/pvg	VP
+기/etn	NP
+만/jxc	VP
하/pvg	VP
+면/ecs	VP
어느/mmd	MODP
정도/ncn	NP
+는/jxt	VP
순치/ncpa+되/xsv	VP
+ㄹ/etm	AUXP
수/nbn	AUXP
있/paa	AUXP
+다/ef+./sf	S

ㄱ. 카이스트 말뭉치 형태의 종단 노드 및 통사 지표

하기야/MAJ	AP
짐승/NNG+도/JX	NP-OBJ
잘/MAG	AP
가르치/VV+기/ETN+만/JX	VP
하/VV+면/EC	VP
어느/MM	DP
정도/NNG+는/jxt	NP-OBJ
순치/NNG+되/XSV+ㄹ/ETM	VP-MOD
수/NNB	NP-SBJ
있/VA+다/EF+./SF	VP

ㄴ. 실험을 위해 세종 말뭉치 형태로 변환한 형태

그림 1. 변환 과정 예

```
(S (IP 하기야/maj )
  (VP (NP-OBJ 짐승/ncn )+도/jxc
      (VP (VP (NP (VP (ADVP 잘/mag) 가르치/pvg )+기/etn )+만/jxc
                  하/pvg )+면/ecs
          (VP  (VP (NP-OBJ (MODP 어느/mmd ) 정도/ncn )+는/jxt
                   순치/ncpa+되/xsv )
  +(AUXP ㄹ/etm 수/nbn 있/paa ))))+다/ef+./sf )
```

그림 2. 변환된 내용을 말뭉치에 적용한 예

[6] 구구조 구문 분석 실험은 버컬리 파서를 사용하였다[3].

[7] [4]에서도 유펜 한국어 구문 분석 말뭉치에서 통사 지표의 유무에 따른 구구조 구문 분석 성능의 차이를 실험하여 통사 지표가 있는 경우에 3.4% 향상된 56.18%의 F1 성능을 보였다.