2018 국어 정보 처리 시스템 경진 대회: 한국어 의존 구문 분석

이창기*, 배장성*, 박천음*, 홍혜진**, 이승재**

강원대학교*, 국립국어원**

{leeck, jseffort, parkce}@kangwon.ac.kr, {honghyejin soundsj}@korea.kr

2018 Korean Information Processing System Competition: Korean Dependency Parsing

Changki Lee*, Jangseong Bae*, Cheoneum Park*, Hyejin Hong**, Seung-Jae Lee**

Kangwon National University, National Institute of Korean Language

{leeck, jseffort, parkce}@kangwon.ac.kr, {honghyejin soundsj}@korea.kr

요 약

의존 구문 분석은 문장 구조를 중심어와 수식어로 이루어진 의존 관계로 표현하는 방법이다. 표현 방식이 간단하고 자유 어순에 적합하기 때문에 한국어와 같이 어순이 자유롭고 문장 성분의 생략이 빈번한 언어에 적합한 것으로 알려져 있다. 본 논문에서는 한국어 의존 구문 분석 연구를 활성화하기 위해 개최된 2018 국어 정보 처리 시스템 경진대회에서 사용된 학습 및 평가 데이터를 간략히 소개하고, 각 참가자들이 제출한 시스템의 결과를 요약한다.

1. 서론

구문 분석은 문장 성분 사이의 관계를 분석하여 문장의 구조적 중의성을 해결하는 자연어처리 문제이다. 구문 분석은 구구조 구문 분석(Phrase structure parsing)과 의존 구문 분석(Dependency parsing)으로 나눌 수 있다. 최근 구문 분석 연구는 표현 방식이 간단하고 자유어순에 적합한 의존 구문 분석이 연구의 주요 흐름으로 자리 매김 하고 있다.

본 논문에서는 의존 구문 분석 연구의 활성화를 위해 개최된 2018 국어 정보 처리 시스템 경진대회에 사용된 의존 구문 분석 데이터를 소개하고, 각 참가자들이 제출한 시스템의 결과를 요약한다. 본 경진대회에서 사용한 학습 데이터는 세종 구구조 구문 분석 말뭉치를 [1]의 연구에서 제시한 방법을 이용하여 자동으로 변환된 의존 구문 분석 말뭉치이다. 경진대회 참가 시스템을 평가하기 위해 2017 국어 정보 처리 시스템 경진 대회 운영팀(한림대 김유섭 교수님)에서 만든 뉴스 기사 1,500 문장의 의존 구문 분석 데이터를 사용한다. 최근 의존 구문 분석 연구는 주로 딥러닝을 이용하여 이루어지고 있으며 [2-4], [5-9]는 이번 대회의 참가한 시스템들이다.

2. 한국어 의존 구문 분석

2.1 한국어 의존 구문 데이터

세종 구구조 구문 분석 데이터를 [1]의 연구에서 제시한 방법에 따라 변환한 데이터는 [그림 1]과 같다. 컬럼 1은 어절의 번호이고, 컬럼 2는 head 어절의 번호, 컬럼 3은 의존 구문 분석 레이블, 컬럼 4는 정답 형태소 분석 결과(세종 코퍼스), 컬럼 5는 강원대 형태소 분석기[10]를 이용한 형태소 분석 결과이다. [그림 2]는 2017년 국어 정보 처리 시스템 경진 대회 운영팀(한림대 김유섭 교수님)에서 만든 한국어 의존 구문 분석 데이터이다. 컬럼 5의 형태소 분석 결과가 없지만 그 외, [그림 1]의 데이터와 동일한 형태를 지니고 있다.

```
; 구두는 굽이3~4㎝ 정도가 속도감 있고 안전한 높이.
1   7   NP_SBJ  구두[NNG]|는[JX] 구두/NNG|는/JX
2   3   NP  굽이[NNG]|3[SN]|~[SO]|4[SN]|㎝[SW]   굽이/NNG|3/SN|~/SO|4/SN|㎝/SW
3   7   NP_SBJ  정도[NNG]|가[JKS]   정도/NNG|가/JKS
4   5   NP_SBJ  속도감[NNG]  속도감/NNG
5   7   VP  있[VV]|고[EC]   있/VV|고/EC
6   7   VP_MOD  안전[NNG]|하[XSA]|ㄴ[ETM]   안전/NNG|하/XSA|ㄴ/ETM
7   0   NP  높이[NNG]|.[SF] 높이/NNG|./SF
```

그림 1. 한국어 의존 구문 분석 데이터(세종 구구조)

```
2015년 12월, 문재인 새정치민주연합 대표와 호남 지역을 중심으로 한 비주류 의원들과의 갈등 국면에서 그는 자신이 만든 당을
뛰쳐나왔다.

1   2   2015년  NP
2   18  12월,  NP_AJT
3   5   문재인  NP
4   5   새정치민주연합   NP
5   18  대표와  NP_CNJ
6   7   호남NP
7   8   지역을  NP_OBJ
8   9   중심으로 NP_AJT
9   11  한  VP_MOD
10  11  비주류  NP
11  13  의원들과의  NP_MOD
12  13  갈등NP
13  18  국면에서 NP_AJT
14  18  그는NP_SBJ
15  16  자신이  NP_SBJ
16  17  만든VP_MOD
17  18  당을NP_OBJ
18  0   뛰쳐나왔다.  VP
```

그림 2. 한국어 의존 구문 분석 데이터(2017 정보 처리)

2.2 Deep bi-affine network

Deep bi-affine network는 그래프 기반 딥 러닝 모델로 bi-affine attention을 사용하여 의존소와 의존 관계명을 학습 및 예측한다. 입력열에 대하여 LSTM으로 인코딩을

수행하고, hidden state $\mathbf{r}_t$에 서로 다른 가중치를 가진 feed-forward neural network (FFNN)를 적용하여 $\mathbf{H}^{arc-dep} = \mathrm{FFNN}(\mathbf{r}_t)$를 생성한다. 그 후에, bi-affine attention을 이용하여 alignment score를 계산하며, 수식은 다음과 같다.

$$\mathbf{s}_t^{arc} = \mathbf{H}^{arc-head}\mathbf{U}^{arc}\mathbf{h}_t^{arc-dep} + \mathbf{H}^{arc-head}\mathbf{u}^{arc} \quad (1)$$

$$\mathbf{s}_t^{rel} = \mathbf{H}^{arc-head}\mathbf{U}^{rel}\mathbf{h}_t^{arc-dep} + \mathbf{H}^{arc-head}\mathbf{u}^{rel} \quad (2)$$

여기서 U는 $\mathbf{H}^{arc-head}$와 $\mathbf{h}_t^{arc-dep}$에 대하여 bi-affine을 수행하기 위한 가중치 매트릭스이며, 식 (1)의 arc와 식 (2)의 rel에 따라 각각 계산되고 u는 bias를 의미한다.

2.3 Stack-pointer networks

Pointer networks는 RNN encoder-decoder에 어텐션 매커니즘을 적용한 모델이다. 인코더(encoder)는 RNN으로 입력열에 대한 hidden state를 인코딩하고, 디코더(decoder)는 현재까지 생성된 디코더의 hidden state와 인코딩 hidden state에 대한 어텐션 가중치를 계산하여 가장 높은 점수를 가지는 위치를 출력한다. Stack-pointer networks는 전이 기반 의존 구문 분석 모델로, pointer networks를 기반으로 양방향 LSTM을 이용하여 인코딩을 수행하고 단방향 LSTM을 이용하여 디코딩을 수행한다. 디코더의 매 time step은 스택의 최상위 단어와 지배소의 sibling, grand parent의 hidden state를 기반으로 의존 구문 부분 트리 정보를 생성하며, 생성된 부분 트리는 다음 의존소 및 의존 관계명을 예측하는데 사용된다. 예측된 의존소는 스택에 push되는데, 이때 예측된 결과가 스택의 최상위에 있는 의존소인 경우, pop을 수행한다.

2.4 한국어 의존 구문 분석 모델

본 장에서는 기존에 발표된 의존 구문 분석 시스템[4]과 참가 시스템의 모델을 요약하여 설명한다.

기존 최고 모델 [4]: pointer networks를 기반으로 의존 구문 분석을 수행한 모델로, 디코더에서 의존소의 hidden state가 인코더 입력열에 위치한 지배소를 포인팅 하는 방법으로 문제를 해결한다. 인코더는 multi-layer로 구축되어 문장 정보의 추상화를 시도한다.

강원대 [5]: Transformer의 multi-head attention과 pointer networks 기반으로 의존 구문 분석을 해결한다.

서강대 [6]: 전이 기반 모델인 stack pointer networks과 그래프 기반 방식에 사용된 bi-affine attention에 따른 단일 모델은 제안하고, 이 둘의 결과에 앙상블 모델을 적용하여 의존 구문 분석을 해결한다. Stack-pointer 모델 5개, bi-affine 모델 5개의 각각 다른 가중치에 대하여 학습하고 앙상블을 진행한다.

충북대 [7]: 그래프 기반 방식에 사용된 bi-affine attention을 기반으로 의존 구문 분석을 해결한다.

고려대 [8]: stack-pointer network를 기반으로 의존 구문 분석을 해결하며, 단어 표현과 음절-태그 단위, 형태소 단위, 평태소 품사 정보, 문자 등의 추가 자질을 함께 사용한다.

NHN [9]: 그래프 기반 방식인 bi-affine attention을 이용하여 의존 구문 분석을 해결한다. 미등록어 문제 해결과 stacked Bi-LSTM에서 표현하는 각 어절의 은닉 표상의 품질을 높이기 위해 형태소+품사, 형태소, 품사, 음절 정보를 이용한다. 의존 관계를 결정하는 단계에서는 다층 신경망을 추가하여 각 어절의 head를 찾기 위해 의존 정보뿐만 아니라 의존 태그 정보도 사용한다.

A 시스템: 다양한 단어 표현과 한국어 의존 구문 분석 모델들을 결합한 앙상블 모델을 제안한다. ELMo와 다양한 자질들을 사용한다. 앙상블 모델은 stack-pointer networks와 RNN 기반 전이 모델, CNN 기반 전이 모델로 구성된다. 각 분석 모델들의 결과들을 종합하여 최종적으로 배깅 기법을 통하여 결과를 생성한다.

B 시스템: arc-eager 방식의 전이 기반 모델로 의존 구문 분석을 해결한다.

C 시스템: arc-eager 방식의 전이 기반 모델로 의존 구문 분석을 해결한다.

3.1 실험 결과

본 대회에서는 한국어 의존 구문 분석 실험을 위해 [1-4]와 동일한 데이터 셋을 사용하였다. 모델 개발을 위한 데이터셋은 총 59,737 문장으로, 53,920 문장을 학습에 사용하고, 5,817 문장을 개발 데이터로 사용하였으며, 각 모델의 평가를 위해 사용된 데이터는 신규로 구축된 신문 기사 1,500 문장이다. 한국어 의존 구문 분석 결과의 평가 척도는 Unlabeled Attachment Score(UAS)와 Labeled Attachment Score(LAS)를 사용하였다. [표 1]은 각 팀별 한국어 의존 구문 분석 실험 결과이다.

표 1. 한국어 의존 구문 분석 결과

Team	Model	UAS	LAS
기존 최고 모델 [4]	Multi-layer Pointer networks	85.55	83.07
강원대[5]	Multi-head attention and pointer networks	85.31	82.72
서강대[6]	Deep bi-affine attention and stack pointer networks	85.24	83.02
충북대[7]	Deep bi-affine attention	85.19	82.89
고려대[8]	Stack pointer networks	84.69	82.02
NHN[9]	Deep bi-affine attention	84.43	80.55
A 시스템	Stack pointer networks	83.93	80.41
B 시스템	Arc-eager	69.22	-
C 시스템	Arc-eager	47.80	21.81

참고문헌

[1] 이창기, 김준석, 김정희. 딥 러닝을 이용한 한국어 의존 구문 분석. 한글 및 한국어 정보처리 학술대회, 2014.

[2] 박천음, 이창기. 멀티 태스크 학습 기반 포인터 네트워크를 이용한 한국어 의존 구문 분석, 한국정보과학회 동계학술발표회, 2016.

[3] 박천음, 황현선, 이창기, 김현기. 멀티 레이어 포인터 네트워크를 이용한 한국어 의존 구문 분석, 제29회 한글 및 한국어 정보처리 학술대회, 2017.
[4] 박천음, 이창기. 포인터 네트워크를 이용한 한국어 의존 구문 분석, 정보과학회논문지, 제44권 제8호 pp 822-831, 2017.8.
[5] 박성식, 오신혁, 김홍진, 김시형, 김학수, ELMo와 멀티헤드 어텐션을 이용한 한국어 의존 구문 분석, HCLT 2018.
[6] 안휘진, 박찬민, 서민영, 이재하. 손정연, 김주애, 서정연, Deep Bi-affine Network와 스택 포인터 네트워크를 이용한 한국어 의존 구문 분석 시스템, HCLT 2018.
[7] Uygun Shadikhodjaev, 민태홍, 윤준영, 이재성, Biaffine 한국어 의존파서, HCLT 2018.
[8] 차다은, 이동엽, 임희석, Stack-Pointer Network를 이용한 한국어 의존 구문 분석, HCLT 2018.
[9] 오동석, 우종성, 이병우, 김경선, NHN_한국어 의존 구문 분석을 위한 개선된 Deep Biaffine Attention, HCLT 2018.
[10] 이창기. Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델. KCC, 2013.