Deep Bi-affine Network와 스택 포인터 네트워크를 이용한 한국어 의존 구문 분석 시스템

안휘진[O], 박찬민, 서민영, 이재하, 손정연, 김주애, 서정연

서강대학교, 자연어처리 연구실

aurorall@naver.com, cmPark@sogang.ac.kr, tkio1313@gmail.com, years6666@gmail.com,

sjy20131565@gmail.com, juaeKim@sogang.ac.kr, seojy@sogang.ac.kr

Korean Dependency Parsing Using Deep Bi-affine Network and Stack Pointer Network

Hwijeen Ahn, Chanmin Park, Minyoung Seo, Jaeha Lee, Jeongyeon Son, Juae Kim, Jeongyeon Seo

Sogang University, Natural Language Processing Lab

요 약

의존 구문 분석은 자연어 이해 영역의 대표적인 과제 중 하나이다. 본 논문에서는 한국어 의존 구분 분석의 성능 향상을 위해 Deep Bi-affine Network 와 스택 포인터 네트워크의 앙상블 모델을 제안한다. Bi-affine 모델은 그래프 기반 방식, 스택 포인터 네트워크의 경우 그래프 기반과 전이 기반의 장점을 모두 사용하는 모델로 서로 다른 모델의 앙상블을 통해 성능 향상을 기대할 수 있다. 두 모델 모두 한국어 어절의 특성을 고려한 자질을 사용하였으며 세종 의존 구문 분석 데이터에 대해 UAS 90.60 / LAS 88.26(Deep Bi-affine Network), UAS 92.17 / LAS 90.08(스택 포인터 네트워크) 성능을 얻었다. 두 모델에 대한 앙상블 기법 적용시 추가적인 성능 향상을 얻을 수 있었다.

주제어: 의존 구문 분석, Bi-affine Attention, Pointer Network, Ensemble

1. 서론

의존 구문 분석은 문장 성분간의 관계를 파악함으로써 문장의 구조를 이해하는 작업으로, 자연어 이해 영역의 대표적인 과제이다. 기존 의존 구문 분석 연구는 전이 기반 방식과 그래프 기반 방식으로 분류된다.

전이 기반 방식[1]의 경우, 파싱 과정 에서 생성되는 의존 구문 서브트리 정보를 자질로 사용할 수 있다는 장점이 있다. 하지만 문장의 전역적인 정보를 사용할 수 없다는 점과 분석 초기에 생성되는 에러가 문장 전체에 전파되는 문제점이 있다.

그래프 기반 방식은 가능한 모든 의존 구문 트리를 고려하는 전역 탐색 방식으로 의존 구문 분석에 접근한다. 최근 연구[2]에서는 두 방식의 특성을 딥 러닝과 결합하여 성능 향상을 이끌어냈다.

본 논문에는 Deep Bi-affine Network[3]와 스택 포인터 네트워크[2]를 사용한다. Deep Bi-affine Network는 그래프 기반 방식에 속하는 모델이며, 스택 포인터 네트워크는 두 방식의 특성을 모두 사용하는 모델이다. 스택을 이용해 순차적으로 구문분석을 하되 양방향 LSTM 인코더를 통해 문장의 전역적 정보를 활용하는 방식이다.

세종 구문 분석 데이터에 대한 실험 결과, Deep Bi-affine Network UAS 90.60/ 88.26, 스택 포인터 네트워크 LAS 92.17 / 90.08의 성능을 기록했다. 특성이 다른 두 모델을 앙상블에 사용하여 보다 향상된 결과인 UAS 92.23 / LAS 90.23을 얻었다.

2. 관련 연구

전통적인 의존 구문 분석은 그래프 기반 모델과 전이 기반 모델로 분류된다. 최근에는 두 가지 방법의 특성을 모두 활용한 연구가 이루어졌다. [2,5]는 전이 기반 방식에 인코더-디코더 모델을 활용하는 모델을 제안하였다.

한국어의 경우, 포인터 네트워크를 사용한 그래프 기반 의존 구문 분석 연구[6]가 가장 높은 성능을 보인다. 마찬가지로 포인터 네트워크를 사용한 [7]연구에서도 음절 단위 태그 분포를 자질로 사용하여 성능을 향상시켰다. 또한 Bi-affine Network을 사용한 그래프 기반 연구[8]도 높은 성능을 보였다.

본 논문에서는 [2,3]에서 제안한 모델을 한국어 어절 특성을 고려하여 한국어에 적용하였다. 또한 전체 의존구문 시스템의 성능을 향상시키기 위해 앙상블 기법을 적용하였다.

3. 제안 방법

본 논문에서는 Deep Bi-affine Network와 스택 포인터 네트워크를 앙상블한 모델을 제안한다. 세종 의존 구문 분석 데이터의 기본 입출력 단위는 어절이다. 한국어의 어절은 여러 개의 형태소로 구성되기 때문에 입력 형태소를 구성하는 다양한 방식이 존재한다. 본 논문에서는 어절당 4개의 형태소 및 품사정보를 입력 자질로 사용하였다.

3-1. Deep Bi-Affine Network

Deep Bi-affine Network[3]는 그래프 기반 방법으로 Bi-affine attention을 사용하여 의존소와 의존 관계명을 예측한다. LSTM의 은닉 표상 $\boldsymbol{r_t}$ 에 서로 다른 가중치를 가진 MLP 계층을 적용하여 $\boldsymbol{h_t^{arc-dep}}$, $\boldsymbol{h_t^{arc-head}}$ 를 생성한다. 이와 같은 연산을 입력 문장 전체에 적용하여 $\boldsymbol{H^{arc-dep}}$, $\boldsymbol{H^{arc-head}}$를 생성하고 다음과 같은 Bi-affine attention 수식 (1,2)를 통해 입력 단어 $\boldsymbol{x_t}$에 대한 지배소 확률 $\boldsymbol{s_t^{arc}}$을 계산한다.

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

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

위 수식 (1,2)에서 u는 bias를 의미한다. 의존 관계명도 위와 같은 방법으로 별도의 서로 다른 가중치를 갖는 $\boldsymbol{h_t^{dep-dep}}$, $\boldsymbol{h_t^{dep-head}}$를 생성하여 입력 단어 $\boldsymbol{x_t}$ 에 대한 의존 관계명 확률 $\mathrm{s}_\mathrm{i}^{rel}$을 생성한다. 단 의존 관계명의 경우, 가능한 arc(L)에 대해서 모든 의존 관계명(N)에 대한 확률 분포가 존재하기 때문에 (L×N×L) 사이즈를 갖는 학습 가중치 $\boldsymbol{U^{rel}}$를 사용한다.

3-2. 스택 포인터 네트워크

스택 포인터 네트워크[2]는 인코더와 디코더 두 부분으로 구성된다. Bi-LSTM으로 구성되는 인코더는 각 단어의 은닉 표상을 생성한다. 이때, 음절 단위의 정보를 추가하기 위해 CNN이 사용되며, 문법정보를 활용하기 위해 형태소 자질이 추가된다.

포인트 네트워크[9]를 사용하는 디코더는 지배소를 예측하는 역할을 한다. 디코더는 매 time step마다 스택의 최상위에 있는 단어에 대한 의존소를 예측한다. 의존소 예측에는 attention 매커니즘이 활용된다. 선택된 의존소는 스택에 PUSH되고, 이후 동일한 방식으로 의존소 예측이 진행된다. 만약 디코더가 스택의 최상위에 있는 단어를 의존소로 예측할 경우, POP 연산이 실행된다. 이와 같은 전이 기반 방식에 기반하여 디코더는 매 time step 마다 의존 구문 서브 트리를 생성한다. 생성된 서브 트리는 다음 의존소 및 의존 관계명을 예측하는데 사용된다. 추가적인 입력 자질로 지배소의 sibling, grand parent의 은닉 표상이 사용되었다.

3-3. 앙상블

본 논문에서는 Deep Bi-affine Network, 스택 포인터 네트워크 모델의 결과를 앙상블한 의존구문 분석 시스템을 제안한다. 각각의 모델은 그래프 기반, 전이 기반의 모델로, 서로 다른 특성을 갖고 있다. 그렇기 때문에 동일한 특성을 갖고 있는 모델들을 앙상블하는 것보다 향상된 성능을 가질 것이라고 기대할 수 있다.

표 2와 같이 총 10개의 서로 다른 가중치를 갖는 3종류의 모델을 학습하여 앙상블을 진행하였다. 앙상블 학습 데이터는 세종 테스트셋을 8:2 비율로 분할한 데이터를 사용하였다.

4. 실험 및 성능 평가

실험에 사용된 데이터는 세종 구문 분석 데이터로, 53,842 문장의 훈련 데이터와 5,817 문장의 테스트 데이터로 구성되어 있다. 단어 표상의 경우 50차원의 사전 학습 NNLM 임베딩[1]과 [10]에서 제안하는 100차원 사전 학습 임베딩을 사용하였다. 음절 표상은 임의로 초기화된 값을 사용하였다. 품사 태그 표상은 대용량의 뉴스 데이터로부터 학습하여 사용하였다. 의존 구문 분석 결과에 대한 평가 척도로는 UAS(Unlabeled Attachment Score)과 LAS(Labeled Attachment Score) 두 가지 기법을 사용하였다.

표 1는 앙상블에 성능과 이에 사용된 모든 의존 구문 분석기의 성능을 나타낸다. (1) ~ (5)는 활성화 함수와 은닉층 개수, 워드 임베딩 종류 등 본 논문에서 사용한 스택 포인터 모델의 경우, 표 2과 같이, 기존 한국어 의존 구문 분석 연구와 비교했을 때 가장 높은 성능을 보였다.

모델	UAS	LAS
Bi-affine (1)	90.43	88.09
Bi-affine (2)	90.60	88.26
Bi-affine (3)	90.41	88.14
Bi-affine (4)	90.54	88.22
Bi-affine (5)	90.42	88.06
스택 포인터(1)	91.65	89.55
스택 포인터(2)	92.05	90.00
스택 포인터 (3)	91.92	89.88
스택 포인터 (4)	**92.17**	**90.08**
스택 포인터 (5)	92.12	90.07
Ensemble	**92.23**	**90.23**

표 1. 제안 시스템의 실험 성능

의존 구문 분석	UAS	LAS
이창기[1]	90.37	88.17
안재현[7]	90.69	87.5
나승훈[8]	91.78	89.76
박천음[6]	92.16	89.88
스택 포인터	**92.17**	**90.08**

표 2 한국어 의존 구문 분석 성능 비교

5. 결론

본 논문에서는 Deep Bi-affine Network과 스택 포인터 네트워크를 사용한 한국어 의존 구문 분석을 제안하였다. 세종 구문 분석 데이터에 대한 실험 결과 각각 UAS 90.60 / LAS 88.26 (Deep Bi-affine Network), UAS 92.17 / LAS 90.08 (스택 포인터 네트워크)의 성능을 얻을 수 있었다. 앙상블을 통해 전이 기반, 그래프 기반 방식 모델을 앙상블하여 추가적인 성능 향상이 있음을 확인할 수 있었다.

사사의 글

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음 "(2015-0-00910)

참고문헌

[1] 이창기, 김준석, 김정희. “딥 러닝을 이용한 한국어 의존 구문 분석.” 제 26 회 한글 및 한국어 정보처리 학술대회 (2014): 87-91

[2] Ma, Xuezhe, et al. "Stack-Pointer Networks for Dependency Parsing." *arXiv preprint arXiv:1805.01087* (2018).

[3] Dozat, Timothy, and Christopher D. Manning. "Deep biaffine attention for neural dependency parsing." *arXiv preprint arXiv:1611.01734* (2016).

[4] 나승훈, 김강일, 김영길, “Stack LSTM을 이용한 전이 기반 한국어 의존 파싱,”KCC 2016

[5] Zhang, Zhirui, et al. "Stack-based multi-layer attention for transition-based dependency parsing." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

[6] 박천음, 황현선, 이창기, 김현기. “멀티 레이어 포인터 네트워크를 이용한 한국어 의존 구문 분석”, HCLT 2017

[7] 안재현, 고영중. (2017). 음절 단위 태그 분포와 멀티 태스크 학습 기반 포인터 네트워크를 이용한 한국어 의존 구문 분석. 한국정보과학회 학술발표 논문집, , 613-615

[8] 나승훈, 이건일, 신종훈, 김강일. (2017). Deep Biaffine Attention을 이용한 한국어 의존 파싱. 한국정보과학회 학술발표논문집, , 584-586.

[9] Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." *Advances in Neural Information Processing Systems*. 2015.

[10] 황현선, 이창기, 장현기, 강동호. 단어의 위치정보를 이용한 Word Embedding, 제29회 한글 및 한국어 정보처리 학술대회, 2017 (SK C&C 과제) (우수논문상)

[11] Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365* (2018).