멀티헤드 어텐션과 포인터 네트워크 기반의 음절 단위 의존 구문 분석

김홍진○, 오신혁, 김담린, 김보은, 김학수
강원대학교 컴퓨터정보통신공학과
jin3430@gmail.com, achromatically@naver.com, ekafls33@naver.com, boeun613@naver.com,
nlpdrkim@kangwon.ac.kr

Multi-head Attention and Pointer Network Based Syllables Dependency Parser

Hong-jin Kim○, Shin-hyeok Oh, Dam-rin Kim, Bo-eun Kim, Hark-soo Kim
Kangwon National University Department of Computer and Communications Engineering

요 약

구문 분석은 문장을 구성하는 어절들 사이의 관계를 파악하여 문장의 구조를 이해하는 기술이다. 구문 분석은 구구조 분석과 의존 구문 분석으로 나누어진다. 한국어처럼 어순이 자유로운 언어에는 의존 구문 분석이 더 적합하다. 의존 구문 분석은 문장을 구성하고 있는 어절 간의 의존 관계를 분석하는 작업으로, 각 어절의 지배소를 찾아내어 의존 관계를 분석한다. 본 논문에서는 멀티헤드 어텐션과 포인터 네트워크를 이용한 음절 단위 의존 구문 분석기를 제안하며 UAS 92.16%, LAS 89.71%의 성능을 보였다.

주제어: 의존 구문 분석, 멀티헤드 어텐션, 포인터 네트워크

1. 서론

구문 분석은 문장을 구성하는 어절들 사이의 관계를 파악하여 문장의 구조를 이해하는 기술이다. 구문 분석은 구구조 분석과 의존 구문 분석으로 나누어지는데, 한국어처럼 어순이 자유로운 언어에는 구구조 분석 보다 의존 구문 분석이 더 적합하다[1-2]. 의존 구문 분석은 문장을 구성하고 있는 어절 간의 의존 관계를 분석하는 작업으로, 각 어절의 지배소와 의존소를 찾아내어 의존 관계를 분석한다. 기존 의존 구문 분석 연구는 사람이 직접 자질(Feature)을 수동으로 설계해야 하는 그래프 기반(Graph Based) 방법과 전이 기반(Transition Based) 방법을 사용하였다. 최근에는 자질을 자동으로 추출하고 학습하는 딥러닝(Deep Learning)을 이용하여 높은 성능을 보이고 있다[2-4]. 의존 구문 분석을 수행할 때에는 주로 형태소 단위로 품사가 부착된 단어들이 입력되기 때문에, 정확한 의존 구문 분석을 위해서 형태소를 분석하는 과정을 거쳐야한다[2-4]. 음절 단위 품사를 입력으로 받으면 형태소 분석을 과정을 거치지 않고 각 음절에 품사를 태깅하므로 전처리 과정이 단순화되고 품사 태깅 및 자동 띄어쓰기와 관련된 기능을 결합하여 통합모델을 설계할 수 있는 장점이 있다[5-6]. 본 논문에서는 딥러닝 기법인 멀티헤드 어텐션(Multi-head Attention)[7]과 포인터 네트워크(Pointer Network)[8]를 이용하고 [9]에서와 같이 음절의 의존 관계명 태그 확률 분포를 이용한 음절 단위 의존 구문 분석을 제안한다.

2. 관련 연구

최근 딥러닝을 이용한 의존 구문 분석 연구에서는 포인터 네트워크를 사용하여 높은 성능을 보이고 있다. 포인터 네트워크는 주의 집중 방법(Attention mechanism)[10]을 통해 입력 열에 대응되는 위치를 학습한다. [2]에서는 포인터 네트워크를 이용하여 의존 관계와 의존 관계명을 학습하는 멀티 태스크 학습 모델을 제안하였다. [3]에서는 포인터 네트워크와 멀티헤드 어텐션(Multi-head attention)을 이용해 의존 관계를 분석하는 모델을 제안하였다. 음절을 이용하여 형태소 벡터나 어절 벡터를 생성하는 의존 구문 분석 연구도 있었다[9]. [9]에서는 각 음절의 의존 관계명 태그 확률 분포를 이용하여 어절 벡터를 유도하고 단어 표상을 확장하였다. 본 논문에서는 [3]에서와 같이 멀티헤드 어텐션과 포인터 네트워크를 이용하지만 음절 단위를 사용하여 형태소 분석의 오류를 보완한 의존 구문 분석을 수행한다.

3. 음절 단위 의존 구문 분석 모델

3.1 모델 구조도

그림 1은 제안 모델의 전체 구조도이다. 어절 벡터(e_t)가 양방향 GRU(Bidirectional Gated Recurrent Unit)[11-12]를 통해 인코딩(Encoding) 된다. 인코딩된 출력(S_t)이 멀티헤드 어텐션에 입력되고 주의 집중 정보가 반영된 문맥 벡터가 생성된다. 디코더에서는 문맥 벡

터와 인코더의 출력 값을 이용하여 입력 어절 벡터들의 위치 분포를 생성하고 의존 관계명을 출력한다.

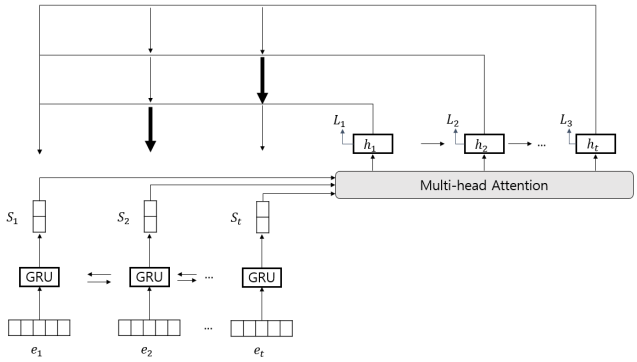

그림 1 음절 단위 의존 구문 분석 모델 구조도

3.2 임베딩

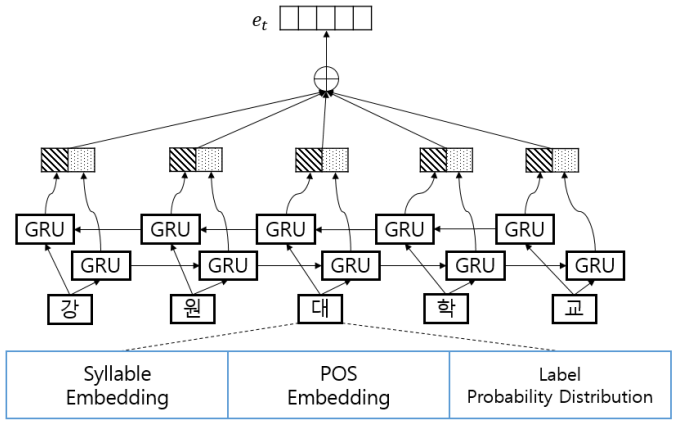

그림 2 어절 벡터 생성 방법

본 논문에서는 음절 임베딩과 음절 단위 품사 임베딩을 사용하며 모두 무작위로 초기화(Random Initialize)한다. 포인터 네트워크의 입력은 어절을 표현한 벡터이기 때문에 단어가 가진 고유한 의미를 잘 표현한 벡터를 생성하는 것이 중요하다. 따라서 본 논문에서는 음절 임베딩과 품사 임베딩 그리고 의존 관계명 태그 확률 분포를 이용하여 어절 벡터를 생성하였다. 그림 2는 본 논문에서 어절 벡터를 생성하는 과정이다. 각 음절 벡터는 음절 임베딩, 품사 임베딩, 의존 관계명 확률 분포를 연결(Concatenation)하여 사용하고 음절 벡터가 양방향 GRU를 통해 인코딩된다. 본 논문에서 사용한 의존 관계명 확률 분포는 학습 데이터에서 각 음절이 어떤 의존 관계명으로 많이 분류 되는지에 대한 분포를 구한 값이다. 양방향 GRU의 각 단계(Step)에서의 출력 값을 모두 연결하여 어절 벡터를 생성한다.

3.3 의존 구문 분석

의존 구문 분석 단계에서는 포인터 네트워크를 이용하여 의존 관계와 의존 관계명을 분석한다. 포인터 네트워크는 주의 집중 방법(Attention Mechanism)을 응용하여 디코더의 각 스텝에서 중요 위치가 강조된 인코더 위치 분포를 계산한다. 계산된 위치 분포는 입력 어절의 지배소 위치를 예측하는데 사용된다. 또한 주의 집중 가중치를 이용해 문맥 벡터(Context vector)를 생성하고 이를 기반으로 의존 관계명을 분석한다. 입력 문장의 가장 앞 부분에 <ROOT> 태그를 추가하여 지배소 후위 원칙에 의해 남는 마지막 어절은 <ROOT>를 지배소로 갖도록 하였다.

4. 실험 및 결과

본 논문에서는 의존 구문 분석 학습 및 평가를 위해 세종 데이터[13]를 사용했다. 총 데이터는 59,659 문장이며 90%인 53,842 문장을 학습하고, 10%인 5,817 문장을 평가에 사용했다. 의존 구문 분석의 평가 척도로는 UAS(Unlabeled Attachment Score)와 LAS(Labeled Attachment Score)를 사용했다.

표 1 의존 구문 분석 성능 비교

	UAS	LAS
박천음 외[2]	91.79	89.48
박성식 외[3]	92.85	**90.65**
임준호 외[4]	**93.38**	90.42
안재현 외[9]	90.69	87.5
제안 모델	92.16	89.71

표 1은 기존에 연구된 의존 구문 분석 모델과 제안 모델의 성능을 비교한 표이다. 표 1의 모델들이 사용한 평가 데이터 말뭉치는 동일하다. [2]는 포인터 네트워크를 사용하여 의존 관계와 의존 관계명을 멀티태스크 방법으로 학습한 모델이다. [3]은 멀티헤드 어텐션과 포인터 네트워크를 사용하였고, [4]는 자가 주의 집중 방법을 사용하여 의존 구문 분석을 수행한다. 본 논문에서 제안하는 모델은 UAS 92.16%, LAS 89.71%의 성능을 보였다. 이는 가장 높은 성능은 아니지만 비교모델들은 형태소의 정답을 사용하였고, 제안모델에서는 [6]의 음절 품사 태깅 방법을 사용하여 형태소 분석 오류가 포함되어 있을 수 있기 때문에 조금 낮은 성능을 보인 것으로 판단된다.

5. 결론

본 논문에서는 멀티헤드 어텐션과 포인터 네트워크를 이용한 음절 단위 의존 구문 분석 모델을 제안하여 UAS 92.16%, LAS 89.71%를 보였다. 향후 연구로 CNN(Convolution Neural Network)를 이용하는 방법 등을 사용하여 어절 벡터를 확장하여 단어 의미를 더 잘 반영하는 연구를 진행할 예정이다.

감사의 글

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2016R1A2B4007732)

참고문헌

[1] 최맹식, 정석원, 김학수, “CRFs를 이용한 의존구조 분석 및 의존 관계명 부착”, *정보과학회논문지 : 소프트웨어 및 응용*, 제41권 제4호, pp. 302-308, 2014

[2] 박천음, 이창기. “포인터 네트워크를 이용한 한국어 의존 구문 분석”, *정보과학회논문지*, 제44권 제8호, pp. 822-831, 2017

[3] 박성식, 오신혁, 김홍진, 김시형, 김학수. “ELMo와 멀티헤드 어텐션을 이용한 한국어 의존 구문 분석”, *제 30회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 40-44, 2018

[4] 임준호, 김현기. “Self-Attention 지배소 인식 모델을 이용한 어절 단위 한국어 의존 구문분석”, *정보과학회논문지*, 제46권 제1호, pp. 22-30, 2019

[5] 이충희, 임준호, 임수종, 김현기, “기분석사전과 기계학습 방법을 결합한 음절 단위 한국어 품사 태깅”, *정보과학회 논문지*, 제43권 제3호, pp. 362-369, 2016

[6] 김홍진, 김담린, 김보은, 김학수, “딥러닝을 이용한 한국어 띄어쓰기 및 품사 태깅, 의존 구문 분석 통합 모델”, *한국정보과학회 2019 한국컴퓨터종합학술대회 논문집*, pp. 1755-1757, 2019

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin. “Attention Is All You Need”, *Neural Information Processing Systems (NIPS)*, pp. 5998-6008, 2017

[8] O. Vinyals, M. Fortunato and N. Jaitly. “Pointer Networks”, *Advances in Neural Information Processing Systems*, pp. 2674-2682, 2015

[9] 안재현, 고영중, “음절 단위 태그 분포와 멀티 태스크 학습 기반 포인터 네트워크를 이용한 한국어 의존 구문 분석”, *한국정보과학회 2017 한국소프트웨어종합학술대회 논문집*, pp. 613-615, 2017

[10] D. Bahdanau, K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, *arXiv preprint arXiv:1409.0473*, 2014

[11] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks”, *IEEE Transactions on Signal Processing*, 45.11, pp. 2673-2681, 1997

[12] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling”, *arXiv:1412.3555*, 2014

[13] The National Institute of the Korean Language, The 21 century Sejong plan, 2012. (in Korean)I. Mani and T. Maybury, Advances in Automatic Text, The MIT Press, 1999.