XLNet을 이용한 한국어 구문분석

김민석, 신창욱, 오진영, 차정원
창원대학교
{20143109, papower1, psycheojy, jcha}@changown.ac.kr

Korean Syntactic Parsing with XLNet

Min-Seok Kim, Chang-Uk Shin, Jinyoung Oh, Jeong-Won Cha
Changwon National University

요 약

문맥기반 사전학습 단어 임베딩이 다양한 분야 적용되어 훌륭한 성능을 보여주고 있다. 본 논문에서는 사전학습한 XLNet 모델을 구문분석에 적용하였다. XLNet은 문장에서 생성 가능한 모든 후보에 대해 트랜스포머를 기반으로 하는 사전학습을 진행한다. 따라서 문장 전체 정보를 필요로 하는 구문분석에 특히 유용하다. 본 논문에서는 한국어 특성을 반영하기 위하여 형태소 분석을 시행한 107.2GB 크기의 대용량 데이터를 사용해 학습을 진행하였다. 본 논문에서 제안한 모델을 세종 구문 코퍼스에 적용한 결과, UAS 91.93% LAS 89.30%의 성능을 보였다.

주제어: 의존구문분석, XLNet, biaffine

1. 서론

최근 XLNet[1]이 자연어처리 분야의 다양한 문제를 성공적으로 해결하였다. XLNet은 BERT가 가지는 단점을 보완하기 위해 제안된 사전학습(pretraining) 문맥 기반 언어모델이다. XLNet은 학습의 대상이 되는 단어나 구의 모든 문맥 가능성에 대하여 학습을 진행한다. 트랜스포머[2]를 기반으로 문장 전체 문맥에 대해서 학습을 진행한다.

구문분석은 문장 안에서 성분들의 관계를 찾는 과정이다. 구문분석을 통하여 문장의 구조를 확인할 수 있고, 애매성을 해소하여 주요 정보를 추출할 수 있다. 구문분석은 문장 전체를 분석하는 것이므로 구조가 복잡할수록 오류가 많아지고 구문분석 정보를 사용하는 상위 시스템에 영향을 미치게 된다.

최근의 의존 구문분석 연구는 딥러닝 기반에 어텐션(attention)을 이용하는 방법이 주류를 이루고 있다[3-6]. 최근에는 문맥 기반 단어 임베딩을 이용하는 방법이 제안되었다. 본 논문에서는 OOV(Out Of Vocabulary) 문제를 해결하기 위하여 형태소 분석을 수행하고 대용량 데이터로 학습한 XLNet 위에 딥바이어파인 어텐션(Deep Biaffine Attention)을 적용한 한국어 의존구문분석 방법을 제안한다.

2. XLNet을 이용한 한국어 사전 학습

기존 언어 모델링을 위한 구조로 auto encoding 방법을 이용한 BERT[2]가 있었다. BERT[2]는 양방향 문맥 정보를 학습시키기 위해 입력 문장을 임의로 masking하는데, 이 때 사용하는 masking 토큰은 실제 문제에 적용할 때 나타나지 않으므로 사전학습(pretraining)과 미세조정(fine-tuning) 사이에 불일치가 발생한다. XLNet은 BERT[2]의 이 제약을 극복하기 위해 permutation language modeling과 two stream self attention을 채용하였다.

$$\max_{\theta} E_{z\sim Z_t}\left[\sum_{t=1}^{T}\log p_\theta(x_{z_t}|\mathrm{X}_{z_{<t}})\right] \quad (1)$$

식 1은 permutation language modeling의 목적함수를 설명하는 수식이다. 식 1에서 Z_t는 입력된 문장으로 만들 수 있는 모든 순열이고, x_{z_t}는 순열 z의 t번째 단어, $\mathrm{X}_{z_{<t}}$는 순열 z에서 t보다 앞선 단어열이다. 위 permutation language modeling을 학습하기 위해서는 지금 추론하고자 하는 단어의 위치 정보가 필요하다. 저자는 이를 위해 기존의 self attention을 개선한 two-stream self attention을 제안한다.

$$g_{z_t}^{(m)} \leftarrow \mathrm{Attention}(\mathrm{Q}=g_{z_t}^{(m-1)}, \mathrm{KV}=\mathrm{h}_{\mathrm{z<t}}^{(\mathrm{m}-1)};\theta) \quad (2)$$

$$h_{z_t}^{(m)} \leftarrow \mathrm{Attention}(\mathrm{Q}=h_{z_t}^{(m-1)}, \mathrm{KV}=\mathrm{h}_{\mathrm{z\leq t}}^{(\mathrm{m}-1)};\theta) \quad (3)$$

식 2는 query stream이라 명명된 attention이다. 목적 단어의 표현 $g_{z_t}^{(m-1)}$은 Q에는 부여하되, KV에는 사용하지 않게 하여 정보를 제한한다. 식 3은 content stream이라 명명된 attention 연산의 수식인데, 여기에서는 KV로 목적 단어의 정보를 부여한다. 이렇게 함으로써 permutation이 수행된 문장열로 학습을 수행하면서도 위치정보를 고려할 수 있게 된다.

본 논문에서는 한국어로 사전학습된 XLNet을 이용해 구문 분석을 수행한다. 학습에는 백과사전 4.9GB와 신문기사 102.3GB를 합하여 총 107.2GB의 텍스트 데이터를 이용하였다. 학습에 사용한 XLNet의 주요 파라메터 설정값은 표1과 같다.

구분	
계층 수	24
multi-head attention의 head 수	16
hidden의 차원	1024
embedding 차원	1024
최대 토큰 길이	256
batch size	24

표 1. XLNet의 파라미터 설정값

사전학습은 위 데이터셋을 총 100만 스텝 학습하였다. 학습에는 4개의 Titan RTX가 부착된 단일 서버를 이용하였고, 약 720시간 가량이 소요되었다.

3. 제안 모델

본 논문에서는 그림 1과 사전 학습된 XLNet을 기반으로 하여 biaffine을 이용하여 구문분석을 수행한다. 입력된 어절은 subword(BPE)[8]로 나누어져 학습된다. XLNet을 통해 나온 각 subword 임베딩을 어절을 임베딩(w_i)을 위해 연결한다. 또한 품사, 형태소와 같은 자질 정보(p_i)를 추가로 연결한다. 여기에 사용한 정보는 표2과 같다.

발생/NNG+하/XSV+였/EP+지만/EC+,/SP	
왼쪽 첫 번째 품사	NNG
왼쪽 두 번째 품사	XSV
오른쪽 두 번째 품사	였/EP
오른쪽 첫 번째 품사	지만/EC
콤마 정보	SP

표 2. 품사, 형태소 자질 상세 정보

연결된 임베딩은 bi-GRU 층을 통과한 후에 MLP를 통해서 구문태그를 추정한다 (그림1).

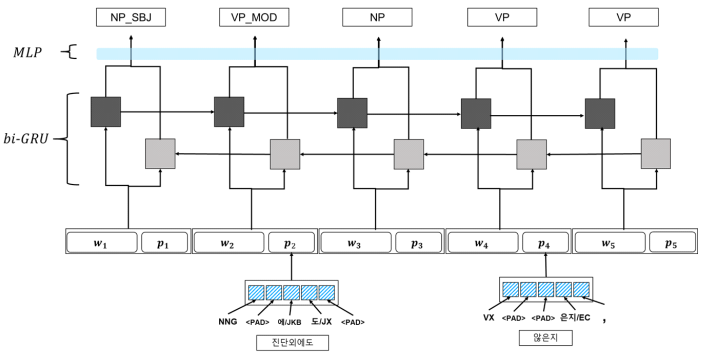

그림 1. 구문관계를 추정하는 구조

구문관계를 추정한 정보는 의존관계를 추정할 때 사용한다 (그림 2). 이렇게 만들어진 자질 정보는 bi-GRU를 통과한 후 [6]에 따라 의존구문분석을 수행하기 위해 i

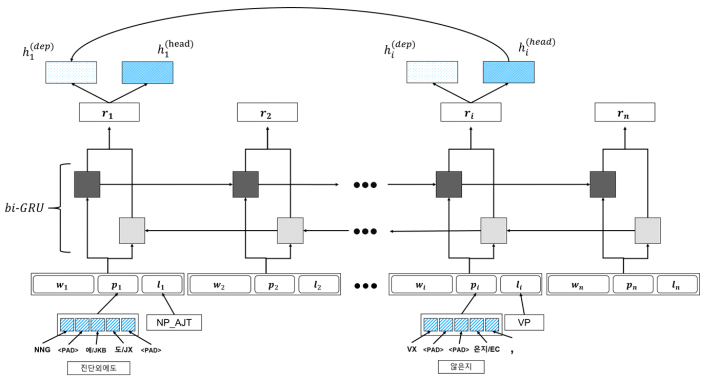

그림 2. biaffine attention을 이용하여 의존관계를 추정하는 구조

번째 어절에 대하여 식2와 같이 비선형 연산을 하여 히든 스테이트를 만든다.

$$\boldsymbol{h_i^{(dep)}} = \mathrm{relu}(\mathrm{MLP}^{(dep)}(\boldsymbol{r_i})) \qquad (2)$$
$$\boldsymbol{h_i^{(head)}} = \mathrm{relu}(\mathrm{MLP}^{(head)}(\boldsymbol{r_i}))$$

이때 사용한 활성화 함수는 relu이다. 의존관계를 구하기 위해서 식3과 같은 biaffine 식을 사용한다.

$$s(j,i) = \boldsymbol{h_i^{(dep)T}} \mathbf{U} \boldsymbol{h_j^{(head)}} + \mathbf{w_1^T} \boldsymbol{h_j^{(head)}} + \mathbf{w_2^T} \boldsymbol{h_i^{(dep)}} + \mathbf{b} \qquad (3)$$

식3에서 $\mathbf{U}$는 bi-linear 가중치 행렬, $\boldsymbol{w_1}$, $\boldsymbol{w_2}$는 각각 디코더 은닉 상태와 인코더 은닉 상태에 대한 가중치 벡터이고 b는 편향을 나타낸다. 이들은 학습되는 파라미터들이다. 계산된 $s(j,i)$ 값은 softmax가 적용되어 의존구조 결과를 출력한다.

4. 실험

본 논문에서는 제안한 방법을 평가하기 위해서 2019 한국어 정보처리 경진대회에서 제공한 코퍼스를 사용하였다. 학습 데이터는 53,842 문장을 사용하였고, 테스트 데이터로는 5,817 문장을 사용하였다.

앞서 소개한 사전학습된 XLNet 모델을 세종 구문분석 코퍼스로 전이학습하고, 제안 모델에 소개한 구조를 XLNet의 위에 추가한다. 여기서 새롭게 추가된 biaffine 구조의 파라메터와 학습 파라메터는 표3과 같다.

구분	
learning rate	2e-6
batch size	8
MLP hidden의 차원	1024
biGRU의 hidden의 차원	512

표 3. 제안 구조의 파라미터 설정값

수행된 실험에서, 제안 모델은 UAS 91.93%, LAS 89.30%의 성능을 보였다. 모델의 결과를 분석하였을 때, 다음 두 가지 오류가 주로 발생하였고, 명사 나열 구문(NP_CNJ)과 용언 선택 문제였다.

명사 나열 구문은 명사가 나열될 때 앞의 명사가 뒤따르는 명사 중 어떤 명사를 지배소로 가질 것인가에 대한 문제로써, 제안하는 biaffine 구조는 입력된 어절 사이의 위치 정보를 면밀히 사용하지 않고, 두 어절의 정보만으로 관계를 모델링하는 데에 집중하기에 오류가 많이 발생하였다고 판단된다.

용언 선택 문제의 경우, 앞선 체언이 지배소로 어떤 용언을 가질 것인가를 선택하는 중 발생하는 오류를 말한다. 문장에 용언이 많이 발생할수록 용언 지배소를 선택하는 데 오류가 발생하였다. 이 문제는 문장 내 위치 관계를 더욱 잘 파악할 수 있도록 구조를 개선하거나, 공기 정보 등과 같은 추가 정보를 활용한다면 개선될 수 있으리라 판단된다.

5. 결론

본 논문에서는 XLNet과 어텐션 기법을 이용한 한국어 의존구문분석 방법을 제안하였다. 대용량 문서에서 사전 학습된 XLNet은 BPE를 통하여 학습을 진행하였다. 어절 단위 자질, 형태소 단위 자질, 품사 자질을 통합하여 사용하였다. 2019 정보처리 경진대회 코퍼스로 실험한 결과 UAS 91.93%, LAS 89.30%의 성능을 보였다.

제안 방법은 나열된 명사들 사이 관계를 포착하거나 앞선 체언이 뒤따르는 복수의 용언 중 하나의 용언을 선택하는 문제에 오류를 발생시켰다. 이는 제안 방법이 비교적 문맥 선후 관계를 모델링하는 것 보다 두 어절의 정보로 관계를 모델링 하는 데에 집중하기 때문이라 분석하였다. 이 문제를 해결하기 위해, 체언과 용언 사이의 공기 정보를 활용하거나, 모델 구조를 변경하는 연구를 향후 연구로 남겨 둔다.

참고문헌

[1] Zhilin Yang el al., “XLNet: Generalized Autogressive Pretraining for Language Understanding”, axXiv:1906.08237, 2019

[2] J. Devlin, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv preprint arXiv:1810.04805, 2018.

[3] A. Vaswani, et al. “Attention Is All You Need.” Neural Information Processing Systems (NIPS), pp. 5998-6008, 2017.

[4] D. Bahdanau, et al. Neural machine translation by jointly learning to align and translate. Proc. of ICLR’ 15, arXiv:1409.0473, 2015.

[5] 박천음,etal.멀티레이어포인터네트워크를이용한한국어의존구문분 석, HCLT, pp. 92-95, 2017.

[6] 나승훈, et al. Deep BiaffineAttention을 이용한 한국어 의존 파싱, KCC, pp. 584-586, 2017.

[7] Timothy Dozat et al., “Deep Biaffine Attention for neural Dependency Parsing”, in ICLR, 2017

[8] R. Sennrich, et al. Neural Machine Translation of Rare Words with Subword Units. In Proc. of ACL, pp.1715-1725, 2016.