사전 학습 모델과 Specific-Abstraction 인코더를 사용한 한국어 의존 구문 분석

김봉수○, 황태선, 김정욱, 이새벽

와이즈넛

{usgnob, taesunwhang, jwkim, saebyeok}@wisenut.co.kr

Korean Dependency Parsing using Pretrained Language Model and Specific-Abstraction Encoder

Bongsu Kim○, Taesun Whang, Jungwook Kim, Saebyeok Lee

Wisenut Inc.

요 약

의존 구문 분석은 입력된 문장 내의 어절 간의 의존 관계를 예측하기 위한 자연어처리 태스크이다. 최근에는 BERT와 같은 사전학습 모델기반의 의존 구문 분석 모델이 높은 성능을 보이고 있다. 본 논문에서는 추가적인 성능 개선을 위해 ALBERT, ELECTRA 언어 모델을 형태소 분석과 BPE를 적용해 학습한 후, 인코딩 과정에 사용하였다. 또한 의존소 어절과 지배소 어절의 특징을 specific하게 추상화 하기 위해 두 개의 트랜스포머 인코더 스택을 추가한 의존 구문 분석 모델을 제안한다. 실험결과 제안한 모델이 세종 코퍼스에 대해 UAS 94.77 LAS 94.06의 성능을 보였다.

주제어: ALBERT, ELECTRA, 의존구문분석

1. 서론

의존 구문 분석이란 자연어 문장내의 어절인 의존소(modifier)가 가지는 지배소(head)와 문법적 역할(label)을 찾는 문제로, 문장의 구조적, 의미적 중의성을 해결하기 위한 자연어처리 태스크이다. 문장의 모든 구 구조를 분석하는 구 구조 분석 방법 보다 각 어절의 지배소와 문법적 역할을 표현함으로써 상호참조해결, 의미역 인식, 개체명 인식 등의 의미 분석 모델에 응용 될 수 있어서 많이 사용되고 있다.

한국어의 의존 구문 분석은 외국어와 다른 세가지의 특징을 가지고 있다. 첫째로 의존 구문 분석의 기본 단위는 어절이다. 교착어의 특징을 갖는 한국어는 한 어절에 한 개 이상의 형태소를 가지고 있다. 하지만, 문장의 의존 구조를 분석할 때, 어절 안의 후위에 위치하는 형태소는 주로 접사, 조사, 어미 등으로 선행하는 형태소에 의존적이다. 따라서, 문장 내에서 어절 단위의 구문적 역할만 고려한다. 둘째는, 지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석하고, 각 어절 및 지배소의 쌍은 서로 교차하지 않도록 한다. 마지막으로 의존소 어절은 하나의 지배소 어절 만을 가지게 된다. 각 문장내의 어절들은 모두 의존소이며 자기 자신의 후위에 위치한 하나의 지배소를 갖고 문장의 가장 후위에 위치한 어절의 경우 자기 자신을 지배소로 갖는다. 각각의 어절은 구문 태그와 기능 태그의 조합으로 이루어져 있는 문법적 역할인 레이블을 갖는다.

전통적인 의존 구문 분석 방법은 그래프기반 방법과 전이기반 방법이 있다. 최근 딥 러닝을 적용한 연구에서는 [1-2]와 같은 주의집중 모델을 적용해 좋은 성능을 보이고 있다. 이를 활용하여 의존 구문 분석 문제를 풀기 위해 그래프기반의 Deep Biaffine Attention[3]이나 전이기반 모델을 활용한 Stack Pointer Network[7] 등의 구조가 주로 사용된다. 또한, 최근에는 BERT[12], ALBERT[16], XLNET[18], ELECTRA[19] 등 대용량 코퍼스로 사전학습된 언어모델이 다양한 자연어처리 태스크에 적용되어 큰 성능 향상을 보이고 있다. 기존 의존 구문 분석 모델 또한 이런 사전 학습을 적용하는 것으로 성능 향상을 보이고 있다[14].

본 논문에서는 사전 학습 모델로 임베딩 파라메터 감소와 레이어간의 파라메터 공유 전략을 활용해 학습 효율을 높인 ALBERT 모델과 마스크 언어 모델(MLM: Masked Language Model) 대신 대체 토큰 탐지(RTD: Replaced Token Detection)로 BERT 보다 빠른 학습 속도로 높은 성능에 수렴하는 ELECTRA 모델을 사용하고, 구조적인 성능 개선을 위해 의존 구문 분석의 의존소 및 지배소 어절 표현 문제를 독립적으로 추상화 하여 해결하는 Specific Abstraction 인코더 모델을 제안한다. 최종적으로 실험을 통해 기존의 연구와 제안한 의존 구문 분석 모델의 성능 차이를 확인하며, 유의미한 성능 개선 결과를 제공한다.

2. 관련 연구

의존 구문 분석 연구는 Deep Biaffine Attention[3], Pointer Network[7] 등의 모델이 주로 연구되어 왔다. 최

근에는 BERT[12]와 같이 대용량 코퍼스로 사전학습된 언어모델을 통해 의존 구문 분석 모델의 성능을 개선시키는 연구가 진행되고 있다.

Pointer Network와 비교해 Deep Biaffine Attention 기반 의존 구문 분석 모델은 디코더를 필요로 하지 않는다. 따라서 신경망을 통해 인코딩 된 토큰들을 각각 의존소와 지배소로 추상화하고, 의존소와 지배소 조합간의 어텐션 스코어를 계산한다. [3]은 LSTM신경망과 Attention을 적용한 의존구문 분석 모델[4]과 기계 번역을 위한 Bilinear Attention[5] 을 확장해 어텐션 레이어를 추가한 Deep Biaffine Attention 함수를 사용하였다. [6]는 한국어 의존 구문 분석에 맞게 문자나 형태소 등의 자질을 활용해 신경망을 통하여 단어를 추상화 시켜 모델의 입력으로 사용한다.

Pointer Network 모델[7]은 입력된 토큰 임베딩에 대하여 문장 정보를 함축해서 표현하는 인코더 부분과 Attention Mechanism을 사용하여 의존소에 따른 지배소와 레이블을 찾는 디코더 부분으로 나뉘어져 있다. [8]은 한국어 자연어처리에 맞게 인코더의 입력단위로 형태소 단위로 임베딩된 토큰을 사용하고 신경망을 통해 추상화 시킨다. Pointer Network를 확장한 Stack-Pointer Network 모델[9]은 하향식(Top-down) 방식으로 마지막 어절인 서술어에서 가장 가까이 있는 의존소를 자식으로 하고 그 다음 의존소를 자식으로 하는 의존 트리를 구성하고 이를 디코더의 입력으로 사용한다. 이때 인코더와 디코더 각각의 임베딩 결과를 신경망에 적용시켜 추상화 시킨 후 Biaffine Attention 을 사용해 지배소를 찾는다. 마찬가지로 [10]에서는 [9]를 기반으로 한국어의 특성에 맞게 그리고 형태소 분석결과를 활용한 토큰 임베딩을 인코더의 입력으로 사용했다.

[14]은 Biaffine Attention과 Bilinear Attention 기반 의존 구문 분석 모델에 사전학습 모델(BERT[12])을 적용하여 state of the art를 달성 함으로써 한국어 의존 구문 분석에도 사전학습된 언어모델이 유용하게 쓰일 수 있음을 확인했다.

3. 사전학습 모델

BERT[12] 이후 다양한 후속 연구들에서 성능과 학습에 소모되는 자원적인 측면에서 더 효율적인 사전학습을 위한 시도가 있었다. 그 중 본 논문에서 사용하는 ALBERT[16]와 ELECTRA[19]는 기존 사전학습 모델의 메모리 소비를 줄이고 학습속도와 성능을 개선하기 위한 파라메터 감소 방법과 새로운 사전학습 태스크를 제안한다.

ALBERT[16]는 vocab(V)을 임베딩 하는 과정에서 바로 embedding size(E)에 매핑하지 않고 embedding size(E)를 hidden size(H)보다 낮게 설정하여 프로젝션 레이어를 추가했다. 이러한 Factorized Embedding Parameterization 기법을 통해 기존의 $O(V*H)$만큼의 파라메터 개수를 $O(V*H+E*H)$로 파라메터 크기를 줄일 수 있다. 또한 각 레이어의 Attention 파라메터와 Feed-Forward Network의 파라메터를 공유하는 Cross Layer Parameter Sharing 기법을 사용하여 모델의 효율을 향상시켰다.

ELECTRA[19]는 일부 토큰을 마스킹하는 대신 replaced 토큰으로 바꾸고 original 토큰을 예측하는 Replaced Token Detection(RTD) 태스크를 사용해 효율적으로 학습을 수행할 수 있다. RTD 태스크 학습을 위해 생성기(Generator)와 판별기(Discriminator), 두 개의 트랜스포머 인코더 기반 네트워크를 필요로 한다. 하나의 example에 대해 15%만 학습하는 것이 아닌 생성기로 token을 생성하고 전체 token에 대해 판별함으로서 학습 효율을 향상시켰다.

본 논문에서는 ALBERT와 ELECTRA 사전학습 모델을 학습시키기 위해 대용량 말뭉치를 수집하고 한국어에 적합한 형태소 단위 모델을 구축했다.

위키 백과 및 신문기사 등 23개 종류의 말뭉치에서 총 939GB의 raw데이터를 수집했고, 전처리 및 형태소 분석 후 42GB의 데이터를 학습에 사용했다. 기본 token인 [PAD], [UNK], [CLS], [SEP], [MASK]를 직접 입력한 후, BPE(Byte Pair Encoding)를 적용해서 byte 단위로 subword tokenize를 수행했다. 형태소 분석에는 자체 보유중인 형태소 분석기를 사용했고, vocab size는 32001로 설정되었다.

[표1 사전학습 파라메터]

model	embedding size	hidden size	layer	attention heads
ALBERT-Base	128	768	12	12
ALBERT-large	128	1024	24	16
ELECTRA-small	128	256	12	4
ELECTRA-base	768	768	12	12

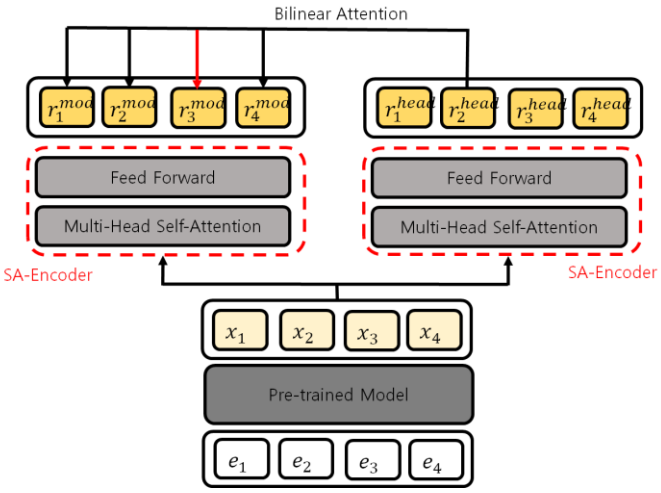
본 논문에서 실험을 위해 적용된 사전학습 모델은 ALBERT-Base, ALBERT-large, ELECTRA-small, ELECTRA-base이며 설정된 파라메터는 [표1]과 같다. 모델은 BERT와 같이 gelu 트랜스포머 인코더를 사용하였으며, 큰 배치 사이즈에서 학습할 수 있도록 LAMB Optimizer[20]를 사용한다.

4. 의존구문 분석 모델

기존 연구 [14]에서는 사전학습 모델을 기반으로 Biaffine Attention과 Bilinear Attention을 통해 의존소 어절에 대한 지배소 어절과 문법적 역할을 찾는다. 사전학습 모델로는 BERT-base를 사용하였으며, BERT의 최종 레이어 결과 x_i를 bi-LSTM 레이어의 은닉상태 h_i로 표현한다. 최종적으로 추상화된 은닉상태 h_i를 지배소와 의존소에 대해 각각 elu로 한번 더 추상화 해서 Attention Score

를 계산한다.

본 논문에서 제안하는 의존 구문 분석 모델은 의존소 어절에 대한 지배소 어절과 문법적 역할을 찾기 위해 Bilinear Attention을 사용한다. 사전학습 모델로는 ALBERT-base, ALBERT-large, ELECTRA-small, ELECTRA-base 총 4개의 모델을 적용했으며, bi-LSTM 을 사용하지 않고 BERT의 마지막 레이어 결과 x_i를 토큰의 최종 표현으로 사용한다. 또한 기존의 방법보다 의존소의 표현과 지배소 표현의 차이를 specific하게 추상화 하기 위해 트랜스포머[11] 인코더 구조를 통해 한번 더 임베딩한다. 최종적으로 입력된 토큰에 대해 의존소와 지배소 표현 $R^{(modifier)}$, $R^{(header)}$ 각각을 얻는다.



[그림1 의존 구문 분석 모델]

[그림1]는 제안하는 의존구문 분석 모델의 전체 구조도이다. 먼저 입력된 문장을 형태소 분석과 BPE 토큰화을 통해 토큰 임베딩을 얻고, 세그먼트 임베딩과 포지션 임베딩을 더하여 사전학습 모델의 입력 표현 $E = \{e_1, \dots, e_1\}$를 만든다. 그 후 입력 표현을 통해 사전학습 모델의 최종 레이어 결과 $X = \{x_1, \dots, x_1\}$를 얻는다.

토큰이 의존소로 사용될 경우 아래의 식과 같이 최종 레이어 결과 x_i 를 의존소 인코더 모델을 통해 $R^{(modifier)} = \{r_1^{(modifier)}, \dots, r_2^{(modifier)}\}$로 추상화 시키고, 지배소로 사용될 경우 지배소 인코더 모델을 통해 $R^{(header)} = \{r_1^{(header)}, \dots, r_2^{(header)}\}$를 얻는다.

$$X = PretrainedModel(E)$$
$$R^{(modifier)} = ENCODER^{(modifier)}(X)$$
$$R^{(header)} = ENCODER^{(header)}(X)$$

최종 의존소와 지배소의 표현은 Bilinear Attention의 입력으로 쓰이고, 이때 지배소(head)와 문법적 역할(label)을 각각 예측한다. [4]의 연구와 다르게 Attention 레이어에 적용하기 전에 어절의 표현을 MLP 레이어에 적용하여 추상화시키지 않는다. 대신 Parameter Sharing을 하지 않는 트랜스포머 기반 의존소와 지배소 인코더를 사용한다. 이때 트랜스포머 인코더의 파라메터는 hidden size : 256, layer 6, attention head : 8 이다.

지배소와 문법적 역할을 예측하는데 Bilinear Attention[5]을 사용한다. 아래의 식과 같이 최종적으로, Bilinear Attention함수를 적용하여 점수를 계산한다. 각 어절의 의존 관계 결정을 위해 $S_{ij}^{(head)}$를 활용하여 head를 결정한다.

$$S_{ij}^{(head)} = r_i^{T(modifier)} U r_j^{(header)}$$

위의 식과 같이 Bilinear Attention는 $r_i^{(modifier)}$ 와 $r_j^{(header)}$의 행렬곱으로 $S_{ij}^{(head)}$을 계산한다. 문법적 역할 $S_{ij}^{(label)}$은 앞서 결정된 head의 위치 r_{ij}표현에 대해 어텐션 스코어를 계산한다. 행렬은 label 집합의 수 만큼 늘려 계산하며 소프트 맥스를 취해 head의 위치에 대응되는 문법적 역할을 찾는다.

본 논문에서 제안한 의존 구문 분석 모델의 경우 두 개의 트랜스포머 인코더를 통해 특정 어절이 의존소로 사용될 때와 지배소로 사용될 때의 특징을 각각 specific하게 추상화 할 수 있다. 그리고 Bilinear Attention을 통해 주변의 의존소를 고려해 지배소의 특징을 학습 할 수 있을 것으로 기대된다.

5. 실험

본 논문에서는 ALBERT와 ELECTRA 사전학습 모델을 학습시키기 위해 대용량 말뭉치를 수집하고 전처리 및 형태소 분석 후 42GB의 데이터를 학습에 사용했다. 형태소 분석 결과에 BPE(Byte Pair Encoding)를 적용해서 사전을 구축하였으며 [표1]과 같이 파라메터를 구성했다.
또한 성능 검증을 위해 같은 데이터로 학습시킨 ALBERT-base, ALBERT-large, ELECTRA-small, ELECTRA-base 모델을 기반으로 의존구문 분석기를 구현하였다.

의존 구문 분석의 학습을 위하여 사용된 데이터는 의존 구문 형태로 변환된 세종 데이터 셋[18]으로 약 60,000 문장이고, 이 중 90%의 문장을 학습셋으로 사용했고, 10%의 문장을 평가셋으로 사용하였다. 모델의 평가 척도는 Unlabeled Attachment Score(UAS)와 Labeled Attachment Score(LAS)를 사용하였다.

본 논문에서는 [표2]의 3~10과 같이 실험군을 8개로 나눠 실험을 진행하였으며, 사전학습 모델로 실험군 3, 4, 5, 6에서는 ALBERT-base와 ALBERT-large를 사용했고 7, 8, 9, 10에서는 ELECTRA-small과 ELECTRA-large를 사용했다.

실험군 3, 5, 7, 9의 경우 사전학습 모델의 아웃풋 레이어 결과 X에 대해 비선형 연산을 하여 $r_i^{(header)}$, $r_i^{(modifier)}$ 각각을 얻는다. 이때 사용되는 활성화 함수는 elu 이며 어텐션 스코어를 계산하고 head와 label을 결정하기 위해 Biaffine Attention 모델을 사용한다.

실험군 4, 6, 8, 10은 제안하는 모델인 의존소 인코더

와 지배소 인코더가 추가된 Specific Abstraction 인코더 모델(SA-Encoder)을 사용하였다.

[표2 실험결과]

no	Model	UAS	LAS
1	BERT + LSTM deep bilinear[14]	93.85	91.78
2	BERT + LSTM deep biaffine[14]	94.06	92.00
3	ALBERT-base + biaffine	93.92	92.41
4	ALBERT-base + SA-Encoder + bilinear	94.31	93.22
5	ALBERT-large + biaffine	94.68	93.52
6	ALBERT-large + SA-Encoder + bilinear	**94.77**	**94.06**
7	ELECTRA-small + biaffine	92.83	91.91
8	ELECTRA-small + SA-Encoder + bilinear	93.84	93.39
9	ELECTRA-base + biaffine	93.88	93.23
10	ELECTRA-base + SA-Encoder + bilinear	94.69	94.01

실험 결과 ALBERT-large 모델을 사용한 실험 모델 5와 6의 경우 ALBERT-base 사전 학습 모델 3,4 보다 더 개선된 성능을 보인다. 또 구현된 실험군 중에서 ALBERT 기반의 3~6 모델이 ELECTRA 기반의 7~10 모델보다 전반적으로 좋은 성능을 보인다. 모델 3, 5, 7, 9의 성능보다 각각 4, 6, 8, 10의 성능이 개선되었고, 제안하는 의존구문 분석 모델이 유의미한 성능을 보임을 확인할 수 있다.

최종적으로 본 논문에서 제안한 ALBERT-large + SA-Encoder 모델이 UAS 94.77 LAS 94.06으로 기존 연구를 포함하여 비교했을 때 가장 높은 성능을 보였다.

6. 결론

본 논문에서는 의존구문 분석에 최신의 사전학습 모델을 적용한 후, 두 개의 인코더 스택을 추가해서 의존소 어절과 지배소 어절 각각의 특징을 추상화하는 Specific Abstraction Encoder 모델을 제안하고 실험하였다.

실험을 통해 기존에 사용하던 Biaffine Attention 기반 의존 구문 분석 모델 보다 성능이 향상되는 것을 확인 할 수 있었다. 사전 학습 모델에서는 최종적으로는 ALBERT-Large 모델이 더 좋은 성능을 보였지만, Base 모델의 경우처럼 같은 사이즈에서는 ELECTRA 모델이 더 좋은 성능을 낼 것으로 기대 할 수 있다.

향후 제안한 모델을 확장해서 지배소의 Encoder를 Non-Autoregressive Decoder로 확장하고, Sequence Level Distilation을 통해 성능을 개선시킬 수 있을 것으로 기대된다.

감사의 글

이 논문은 2020년도 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원의 지원을 받아 수행된 연구임 (No.1415166164,금융 지식 그래프를 위한 다국어 자연어 처리기술 개발)

참고문헌

[1] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, ICLR, 2015.

[2] M.-T. Luong, H. Pham, C. D. Manning, Effective Approaches to Attention-based Neural Machine Translation, EMNLP 2015.

[3] T. Dozat, C. D. Manning, Deep Biaffine Attention for Neural Dependency Parsing, ICLR, 2017.

[4] E. Kiperwasser and Y. Goldberg, Simple and accurate dependency parsing using bidirectional LSTM feature representations, TACL, 2016.

[5] M.-T. Luong, H. Pham, C. D. Manning, Effective Approaches to Attention-based Neural Machine Translation, EMNLP, 2015.

[6] 나승훈, 이건일, 신종훈, 김강일, "Deep Biaffine Attention을 이용한 한국어 의존 파싱", 한국컴퓨터종합학술대회, p0584-p0586, 2017.

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

[8] 박천음, 황현선, 이창기, 김현기. “멀티 레이어 포인터 네트워크를 이용한 한국어 의존 구문 분석”, 제29회 한글 및 한국어 정보처리 학술대회 논문집, p.92-95, 2017.

[9] Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig and Eduard Hovy. Stack-Pointer Networks for Dependency Parsing, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.

[10] 최용석, 이공주, "스택-포인터 네트워크와 어절 정보를 이용한 한국어 의존 구문 파서", 한글 및 한국어 정보처리 학술대회, p13-p18, 2018.

[11] A. Vaswani, et al, Attention Is All You Need, Neural Information Processing Systems, 2017.

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

[13] R. Sennrich, et al, Neural Machine Translation of Rare Words with Subword Units, In Proc. of ACL, 2016.

[14] 박천음, 이창기, 임준호, 김현기, "BERT를 이용한 한국어 의존 구문 분석", 한국컴퓨터종합학술대회, p530-p532, 2019.

[15] The National Institute of the Korean Language, The 21 century Sejong plan, 2012.n Processing & Management 49.1 (2013): 370-379.

[16] Zhenzhong Lan, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. CL, 2019.

[17] 임준호, 배용진, 김현기, 김윤정, 이규철, “의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치”, 제 27회 한글 및 한국어 정보처리 학술대회 논문집, 2015.
[18] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5753-5763). 2019.
[19] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555. 2020.
[20] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. arXiv preprint arXiv:1904.00962. 2019.