

BERT 기반 End-to-end 신경망을 이용한 한국어

상호참조해결

김기훈*^o, 박천음*, 이창기*, 김현기**

강원대학교 컴퓨터학과*, 한국전자통신연구원**

{rlarlgnsu, parkce, leeck}@kangwon.ac.kr, hkk@etri.re.kr

Korean End-to-end Neural Coreference Resolution with BERT

Kihun Kim*, Cheonum Park*, Changki Lee*, Hyunki Kim*

Kangwon National University*, Electronics and Telecommunications Research Institute**

요약

상호참조해결은 주어진 문서에서 상호참조해결 대상이 되는 멘션(mention)을 식별하고, 같은 개체(entity)를 의미하는 멘션을 찾아 그룹화하는 자연어처리 태스크이다. 한국어 상호참조해결에서는 멘션 탐지와 상호참조해결을 동시에 진행하는 end-to-end 모델과 포인터 네트워크 모델을 이용한 방법이 연구되었다. 구글에서 공개한 BERT 모델은 자연어처리 태스크에 적용되어 많은 성능 향상을 보였다. 본 논문에서는 한국어 상호참조해결을 위한 BERT 기반 end-to-end 신경망 모델을 제안하고, 한국어 데이터로 사전 학습된 KorBERT를 이용하고, 한국어의 구조적, 의미적 특징을 반영하기 위하여 의존구문분석 자질과 개체명 자질을 적용한다. 실험 결과, ETRI 질의응답 도메인 상호참조해결 데이터 셋에서 CoNLL F1 (DEV) 71.00%, (TEST) 69.01%의 성능을 보여 기존 연구들에 비하여 높은 성능을 보였다.

주제어: 딥 러닝, 상호참조해결, BERT

1. 서론

상호참조해결(coreference resolution)은 명사, 대명사, 명사구 등 멘션 후보를 식별하고, 동일한 개체(Entity)로 사용되는 멘션들을 찾아 그룹화(Clustering)하는 자연어처리 태스크이다. 최근 한국어 상호참조해결 연구는 포인터 네트워크(pointer network)를 사용하는 연구[1-4]와 LSTM 기반 멘션 랭킹 모델을 사용하는 연구[5]가 진행되었다. [4, 5]는 멘션 탐지 태스크와 상호참조해결 태스크를 동시에 진행하는 end-to-end 상호참조해결 모델을 제안하였으며, 기존 연구들[1-3]에 비하여 좋은 성능을 보였다.

구글에서 공개한 BERT (Bidirectional Encoder Representations from Transformer) [6]는 다양한 자연어처리 태스크에 적용되어 각 태스크의 성능을 향상시켰다 [4,7,8]. BERT는 트랜스포머(transformer)의 인코더(encoder) 부분으로 구성되며, 주어진 입력에 대한 문맥 정보를 양방향으로 확인할 수 있다. BERT는 BPE (byte-pair-encoding)[9]를 사용하여 미등록어(unknown word) 문제를 해결하였으며, BPE를 적용한 대용량 말뭉치를 기반으로 언어 모델(language modeling)을 학습한다. 언어 모델을 학습할 때 masked language modeling (masked LM)과 다음 문장 예측을 함께 학습하게 된다. 사전 학습된 BERT 모델은 fine-tuning하는 방법으로 다른 자연어처리 태스크에 알맞게 모델링되어 사용된다.

본 논문에서는 멘션 랭킹 모델[5]에 BERT를 추가한 모델[10]을 기반으로, 한국어의 구조적, 의미적 특징을 반영하기 위한 의존 구문 분석 자질과 개체명 자질을 적용하고, bi-LSTM 레이어를 추가하여 BERT표현과 문맥정보에 대한 히든 스테이트를 만들어 한국어 상호참조해결을 수

행하는 모델을 제안한다. 또한 성능 향상을 위하여 한국어 대용량 말뭉치로 사전 학습한 KorBERT를 사용한다.

2. 관련 연구

최근 한국어 상호참조해결 연구는 기계 학습을 이용한 방법 중 포인터 네트워크를 사용하는 방법[1-4]과 멘션 랭킹을 사용하는 end-to-end 방법[5]이 연구되었다.

[1]은 RNN encoder-decoder를 기반으로 문맥 정보를 인코딩(encoding)하고 포인터 네트워크를 이용하여 참조 관계를 밝히는 시도를 하였다. [2]는 서로 유사한 단어 간의 얼라인먼트 스코어가 높게 계산되는 셀프 매칭 어텐션(self-matching attention)을 기반으로 포인터 네트워크를 적용하여 상호참조해결에 도움을 주는 방법을 제안하였다. [3]은 양방향 언어 모델을 사전 학습한 ELMo[11] 표현과 음절 표현을 학습하는 convolution neural network(CNN)을 적용하여 서브 단어(sub-word)에 대한 학습을 가능하게 한 모델을 제안하였다. [4]는 포인터 네트워크에 BERT와 Deep Biaffine을 이용하여 멘션 탐지 태스크와 상호참조해결 태스크를 한번에 해결하는 모델을 제안하였으며, 한국어 상호참조해결에 많은 성능 향상을 보였다.

[5,10,12]의 연구는 모든 스펠을 멘션 후보로 간주하여 멘션 탐지 태스크와 상호참조해결 태스크를 동시에 해결하는 end-to-end 방식이다. [10]은 [12]에 BERT를 추가하여 영어권에서 많은 성능 향상을 보였다.

3. BERT를 이용한 End-to-end 상호참조해결 모델

본 논문에서는 사전 학습된 BERT를 기반으로 하는 end-to-end 상호참조해결 모델을 제안하며, 한국어 특성

하여 cores-to-fine을 진행 후, 상위 c 개에 대하여 $s_a(i, j)$ 선행사 스코어를 계산한다. 마지막으로, 수식(14)와 같이 멘션 스코어와 선행사 스코어를 합하여 상호참조해결 스코어를 계산하며, 엔티티 한 개에 멘션 한 개만 존재하는 싱글톤(singleton)의 경우에는 더미 선행사(ϵ)로 취급하여 상호참조해결 스코어는 0이 된다. 또한 [5,12]는 문장 수준에서는 상호참조해결이 되는 것처럼 보이지만, 전체 문서로 보았을 때 상호참조해결이 안되는 일관성 문제(consistency errors)를 해결하기 위하여 고차 추론을 적용하며, 수식은 아래와 같다.

$$P_n(y_i) = \frac{e^{s(g_i^n, g_{y_i}^n)}}{\sum_{y \in Y(i)} e^{s(g_i^n, g_y^n)}} \quad (15)$$

$$\alpha_i^n = \sum_{y \in Y(i)} P_n(y_i) \cdot g_{y_i}^n \quad (16)$$

$$f_i^n = \sigma(W_f[g_i^n, \alpha_i^n]) \quad (17)$$

$$g_i^{n+1} = f_i^n \circ g_i^n + (1 - f_i^n) \circ \alpha_i^n \quad (18)$$

고차 추론[5,12]은 n 번 반복하며, 현재 멘션의 선행사들에 대한 얼라인먼트 스코어 α_i^n 와 현재 멘션 정보에 대한 게이트 벡터 f_i^n 를 사용하여 스캔 표현을 업데이트 해준다. 모델의 출력은 상호참조해결 스코어가 가장 높은 k 개의 (멘션, 선행사) 쌍이다.

4. 실험 및 결과

본 논문에서 제안한 상호참조해결 모델을 실험하기 위하여 ETRI 질의응답 도메인 상호참조해결 데이터 셋을 이용하였으며, 성능 측정을 위하여 중심어 경계(head boundary)를 기준으로 MUC, B³, CEAF-e, CoNLL F1[13]을 사용하였다. 데이터 셋은 학습 데이터(Train) 2,819 문서, 개발 데이터(Dev) 645 문서, 평가 데이터(Test) 571 문서로 구성된다. 그 중, 학습 데이터에서 문서 길이가 512 이상인 데이터는 제외하였다.

하이퍼 파라미터 최적화 실험은 개발 셋을 이용하여 진행하였다. KorBERT 사전 학습을 위한 파라미터는 BERT-base(L = 12, H = 768, A = 12)와 같으며, 히든 레이어 활성화 함수로 gelu, 드랍아웃(dropout)은 0.1, 최대 문장 길이는 512가 사용되었다. 상호참조해결 모델 학습을 위한 파라미터는 BERT fine-tuning 학습률(earning rate)은 1e-5, end-to-end 모델 학습률은 2e-4, FFNN은 드랍아웃 0.2, 히든 레이어 차원수 100, 스택수는 2, LSTM은 드랍아웃 0.2, 히든 레이어 차원수 800, 스택수는 1, 고차 추론 반복 횟수는 2로 설정하였다. 계산량을 위한 하이퍼파라미터는 선행사 후보 수 50, 최대 멘션 후보 수는 50, 최대 스캔 길이 70, 멘션 스코어 pruning 비율은 0.4로 설정하였다. 학습을 위하여 Adam을 사용하였고, 미니 배치 크기는 1이다.

표 1. 모델에 따른 상호참조해결 성능 비교(%)

DEV				
Model	MUC	B ³	CEAF _e	Dev F1
Ptr-net 기반 [3]	60.87	57.12	55.64	57.88

Ptr-net + BERT + deep biaffine [4]	70.23	68.01	65.66	67.97
Higher-order e2e-coref [5]	70.33	67.55	68.18	68.69
Higher-order e2e-coref + BERT [10]	70.37	67.89	68.8	69.02
Our Model	72.41	69.87	70.71	71.00
TEST				
Model	MUC	B ³	CEAF _e	Test F1
Higher-order e2e-coref [5]	64.14	61.55	62.90	62.86
Ptr-net + BERT + deep biaffine [4]	65.98	64.03	63.64	64.55
Our Model	70.07	67.67	69.29	69.01

[표 1]은 기존의 한국어 상호참조해결 모델과 본 논문에서 제안한 BERT를 이용한 end-to-end 상호참조해결 모델에 대한 성능 비교 결과이다. 개발 셋(dev set)에 대하여 본 논문에서 제안한 모델이 기존 end-to-end 한국어 상호참조해결 모델[5]보다 CoNLL F1 2.31% 더 좋은 성능을 보였으며, BERT를 이용한 포인터 네트워크 상호참조해결 모델[4]보다 3.03% 좋은 성능을 보였다. 평가 셋(test set)에서는 [5]보다 6.15%, [4]보다 4.46% 높은 성능을 보여 개발 셋보다 평가 셋에서 더 큰 향상 폭을 보였다.

표 2. bi-LSTM 히든 레이어 크기에 따른 성능 비교(%, dev)

Dimension size	MUC	B ³	CEAF _e	Dev F1
100	71.38	64.66	69.81	70.09
200	72.03	69.63	70.1	70.59
400	71.97	69.35	69.72	70.33
800	72.41	69.87	70.71	71.00

[표 2]는 개체명과 의존 구문 분석 자질을 추가하기 위해 BERT모델 위에 추가한 bi-LSTM의 히든 레이어 차원수에 따른 성능 비교이다. 실험은 100, 200, 400, 800 차원으로 수행하였으며, 개체명 자질과 의존 구문 분석 자질 모두 추가한 모델에 대하여 실험을 진행하였다. 실험 결과는 개발 셋에 대하여 히든 레이어 차원수 800일 때 CoNLL F1 71.00%로 가장 높은 성능을 보였다.

표 3. 자질 추가 성능 비교(%, dev)

Feature	MUC	B ³	CEAF _e	Dev F1
Higher-order e2e-coref + BERT [10]	70.37	67.89	68.80	69.02
+ bi-LSTM	71.77	69.25	70.45	70.49
+ 개체명 자질	71.99	69.6	70.17	70.59
+ 의존 구문 분석 자질	72.41	69.87	70.71	71.00

[표 3]은 최적화된 파라미터를 사용하여 bi-LSTM, 개체명 자질, 의존 구문 분석 자질을 추가한 성능을 비교한 결과이다. bi-LSTM 레이어는 기존 end-to-end 상호참조해결 모델에 개체명 자질과 의존 구문 분석 자질을 결

합하기 위하여 BERT 모델 위에 추가하였으며, bi-LSTM 레이어를 추가하지 않은 모델[10]에 비해 개발 셋 기준 1.47% 높은 성능을 보였다. 그 후, 개체명 자질을 추가하여 0.1% 성능 향상이 있었고, 의존 구문 분석을 추가하여 0.41% 성능이 향상됨을 보였다. 최종적으로 bi-LSTM, 개체명 자질, 의존 구문 분석 자질을 추가한 모델이 기존 BERT 사전 학습 모델만 사용한 결과보다 1.98% 높은 성능을 보였으며, 위 두 자질이 문서의 문맥을 파악하고 단어와 어구의 특정한 의미를 파악하여 한국어 상호참조해결에 도움이 되는 것으로 사료된다.

5. 결론

본 논문에서는 문맥 정보에 기반한 단어 표현을 사용하기 위하여 대용량의 한국어 코퍼스로 사전학습된 KorBERT를 이용하고, BERT 모델에 한국어 특성을 반영할 수 있는 의존 구문 분석 자질과 개체명 자질을 추가한 end-to-end 상호참조해결 모델을 제안하였다. 실험 결과 본 논문에서 제안한 모델은 개발 셋에서 CoNLL F1 71.00%, 테스트 셋에서 69.01%로 기존의 모델보다 좋은 성능을 보였으며, 특히 테스트 셋에서 높은 성능향상을 보였다. 향후 연구로는 512토큰 이상의 문서에 대한 학습 방법을 연구할 예정이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플랫폼 기술 개발)

참고문헌

[1] 박천음, 이창기, “포인터 네트워크를 이용한 한국어 대명사 상호참조해결”, 정보과학회논문지, 제44권, 제5호, pp.496-502, 2017.

[2] 박천음, 이창기, 김현기, “셀프 매칭 어텐션 기반 포인터 네트워크를 이용한 한국어 상호참조해결”, 한국정보과학회 학술발표논문집, 한국정보과학회 2017 한국소프트웨어종합학술대회 논문집, pp.640-642, 2017

[3] 박천음, 이창기, 류지희, 김현기, "문맥 표현과 음절 표현 기반 포인터 네트워크를 이용한 한국어 상호참조해결", 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp.239-242, 2018.

[4] 박천음, 김기훈, 이창기, 임준호, 류지희, 김현기, “BERT 기반 Deep Biaffine을 이용한 한국어 상호참조해결, 한국정보과학회 학술발표논문집, 한국정보과학회 2019 한국컴퓨터종합학술대회 논문집, pp488-490, 2019.

[5] 김기훈, 박천음, 이창기, 김현기. “고차 추론을 이용한 한국어 End-to-end 신경망 기반 상호참조해결”, 한국정보과학회 학술발표논문집, 한국정보과학회 2019 한국컴퓨터종합학술대회 논문집, 2019.

[6] J. Devlin, M.W Chang, K. Lee and K. Toutanova,

“Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.

[7] 박광현, 나승훈, 신종훈, 김영길, “BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정”, 한국정보과학회 학술발표논문집, 한국정보과학회 2019 한국컴퓨터종합학술대회 논문집, pp.584-586, 2019.

[8] 이동현, 박천음, 이창기, 박소윤, 임승영, 김명지, 이주열, “BERT를 이용한 한국어 기계 독해”, 한국정보과학회 학술발표논문집, 한국정보과학회 2019 한국컴퓨터종합학술대회 논문집, pp.557-559, 2019.

[9] R. Sennrich, B. Haddow and A. Birch, “Neural Machine Translation of Rare Words with Subword Units”, arXiv preprint arXiv:1508.07909, 2015.

[10] M. Joshi, O Levy, D. S. Weld and L. Zettlemoyer, “BERT for Coreference Resolution: Baselines and Analysis”, arXiv preprint arXiv:1908.09091, 2019.

[11] M. E. Peters, et al. “Deep contextualized word representations”, Proceedings of NAACL-HLT, pp.2227-2237, 2018.

[12] C. Park, K. H. Choi, C. Lee and S. Lim, “Korean Coreference Resolution with Guided Mention Pair Model using the Deep Learning.”, ETRIJ., vol.38, no.6, pp.1207-1217,2016

[13] H. Lee, et al. “Deterministic coreference resolution based on entity-centric, precision-ranked rules.”, Computational Linguistics, vol. 39 No. 4, pp. 107-110, 2013