표층형을 이용한 BERT 기반 한국어 상호참조해결

허철훈○, 김건태, 최기선
한국과학기술원
{fairy_of_9, kuntaek, kschoi}@kaist.ac.kr

Korean Co-reference Resolution using BERT with Surfaceform

Cheolhun Heo○, Kuntae Kim, Key-sun Choi
KAIST

요 약

상호참조해결은 자연언어 문서 내에서 같은 개체를 나타내는 언급들을 연결하는 문제다. 대명사, 지시 관형사, 축약어, 동음이의어와 같은 언급들의 상호참조를 해결함으로써, 다양한 자연언어 처리 문제의 성능 향상에 기여할 수 있다. 본 논문에서는 현재 영어권 상호참조해결에서 좋은 성능을 내고 있는 BERT 기반 상호참조해결 모델에 한국어 데이터 셋를 적용시키고 표층형을 이용한 규칙을 추가했다. 본 논문의 모델과 기존의 모델들을 실험하여 성능을 비교하였다. 기존의 연구들과는 다르게 적은 특질로 정밀도 73.59%, 재현율 71.1%, CoNLL F1-score 72.31%의 성능을 보였다. 모델들의 결과를 분석하여 BERT 기반의 모델이 다양한 특질을 사용한 기존 딥러닝 모델에 비해 문맥적 요소를 잘 파악하는 것을 확인했다.

주제어: 상호참조해결, BERT, 표층형(surfaceform)

1. 서론

상호참조해결은 문서 내에서 같은 개체를 나타내는 언급들을 연결하는 문제를 말한다. 자연언어로 이루어진 문서에서 개체는 대명사, 지시 관형사, 축약어 등과 같이 다양한 표층형을 가질 수 있고, 동음이의어와 같이 동일한 표층형을 가지는 다른 개체들이 존재할 수 있다. 따라서 언급들의 상호참조 관계를 해결함으로써, 다양한 자연언어처리 문제의 성능 향상에 기여할 수 있다.[1]

BERT(Bidirectional Encoder Representations from Transformers)는 트랜스포머 구조를 기반으로 한 사전학습된 언어 모델이다.[2] 사전학습은 주어지는 자연언어 문서에서 임의의 단어에 마스크를 씌우고, 해당 마스크를 예측하는 문제와 주어진 문장들이 연결된 문장인지를 예측하는 문제에 대해 동시에 학습한다. 사전학습을 통해 모델은 문맥을 잘 반영하는 단어 임베딩을 만들어 내고 여러 문제들에 적합하게 fine-tuning 할 수 있다. 하지만 사전학습 당시 최대 512개의 토큰을 학습하기 때문에, BERT는 최대 512개 토큰만 입력받을 수 있다. 토큰은 BERT가 자체적으로 가지는 토크나이저에 의해 만들어진다. 그럼에도, 자연언어처리 분야의 상호참조해결을 포함한 다양한 문제에서 BERT를 기반으로 한 모델들이 최고 성능을 내고 있다.[3]

기존의 영어권 상호참조해결 모델은 종단 간 학습을 통해 문서내에 모든 스팬에서 언급들을 찾아내고 찾아낸 언급들의 상호참조 관계를 밝힌다.[3,4,5] 모델들은 모든 스팬에 대해 언급 점수를 내고, 모든 스팬 쌍에 대한 선행사 점수를 내어 최종적으로 스팬 쌍의 언급 점수와 선행사 점수를 통해 해당 스팬 쌍의 상호참조 관계를 결정한다. 이 중 BERT를 기반으로 한 모델[3]이 OntoNotes와 GAP 벤치마크에서 각각 F1-score 76.9%, 85%의 성능을 보여주었고 OntoNotes 벤치마크에서 최고 성능을 냈다.

본 논문에서는 BERT를 기반으로 한 영어권 상호참조해결 모델[3]에 한국어를 적용하고, 표층형을 사용한 규칙을 추가적으로 적용하여 CoNLL F1-score 72.31%의 성능을 보였다.

2. 관련 연구

상호참조해결은 규칙 기반, 머신러닝을 이용한 모델이 오랫동안 연구되어 왔고,[6] 최근 딥러닝을 이용한 종단 간 학습 모델이 좋은 성능을 보여주고 있다.[3,4,5]

처음으로 종단 간 학습을 통한 상호참조해결을 보인 [4]는 Bi-LSTM을 기반으로 GloVe[7]와 character 임베딩을 활용하여 문서 내의 모든 스팬에 대한 스팬 표현(벡터 임베딩)을 만든다. 언급 점수 함수와 상호참조 점수 함수가 스팬 표현을 통해 모든 스팬 쌍의 상호참조 여부를 결정하여 F1- score 67.2%를 달성했다. 이를 기반으로 한 [5]는 문맥을 더 잘 반영하는 ElMo[8]를 워드 임베딩으로 사용한다. 또한 선행사 분포를 사용하여 스팬 표현을 상호참조해결에 적합하도록 수정하여 F1-score 73%를 달성했다. 이를 기반으로 한 [3]은 워드 임베딩을 BERT 임베딩으로 그리고 Bi-LSTM을 트랜스포머로 대체하여 CoNLL F1-score 76.9%를 달성했다.

한국어 상호참조해결 또한 딥러닝을 통한 모델이 활발하게 연구 중이다. 포인터 네트워크를 이용한 모델[9]이 있었으며, [10]는 [5]에 형태소, 개체 유형, 언급 후보 목록의 경계 정보를 특질로 사용하였고, [11]는 BERT를 기반으로 형태소, 의존구문분석, 개체 유형을 특질로 이용한 연구이다.

3. 접근 방법

기존 논문의 상호참조해결 모델은 다음과 같은 분포를 학습하게 된다.[5]

$$P_n = \frac{e^{s(x,y)}}{\sum_{y' \in Y} e^{s(x,y')}}$$

Y는 스팬 x의 선행사가 될수 있는 스팬들의 집합이다. $s(x,y)$는 스팬 x와 스팬 y의 상호참조 점수이다. 본 논문의 모델은 그림 1과 같은 흐름도를 가진다.

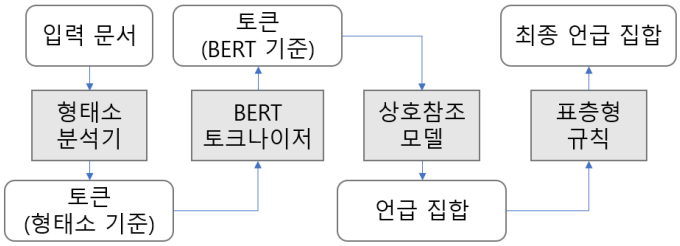

그림 1. 상호참조모델의 흐름도

주어진 문장을 형태소 단위로 분절한 후 BERT 토크나이저를 사용해 토큰화한다. 형태소 분석기는 BERT는 토큰 길이 제한이 있기 때문에 최대 토큰 길이를 넘어가는 문장은 뒷부분의 토큰을 잘라낸다. 이 길이 하이퍼-파라미터로 조정된다. 여기에 첫 번째 토큰과 마지막 토큰에 각각 [CLS]와 [SEP]를 추가한다. 그림 2와 같이 사전학습된 BERT에 해당 토큰들이 입력으로 들어가고, 토큰에 대한 벡터 임베딩을 출력한다. 이를 기반으로 스팬 표현을 생성한다.

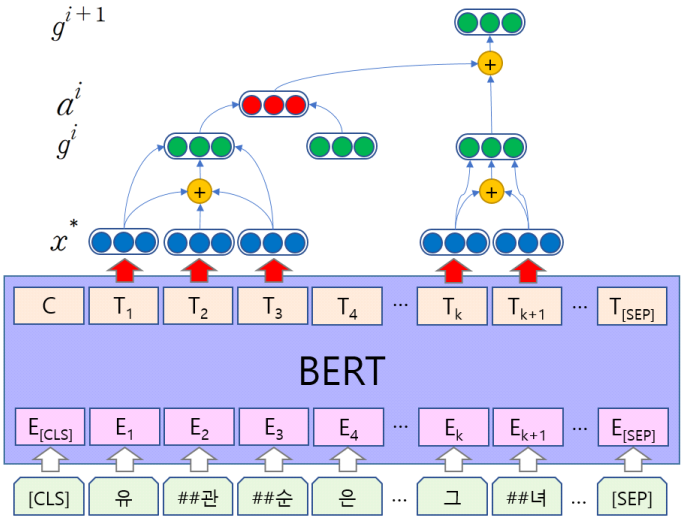

그림2. BERT 기반 스팬 표현 생성[3]

i번째 스팬 표현 g_i^n 를 생성하기 위해 먼저 g_i^1를 생성한다.[4]

$$g_i^1 = [x^*_{START(i)}, x^*_{END(i)}, \hat{x}_i, \phi(i)]$$

$x^*_{START(i)}$와 $x^*_{END(i)}$는 각각 i번째 스팬의 시작과 마지막 토큰의 임베딩 값이고 x'는 해당 스팬 내에 모든 토큰 임베딩에 어텐션 메커니즘을 적용한 값이며, $\phi(i)$는 해당 언급의 길이에 대한 특질값이다.[4]

반복마다 현재 선행사 분포 $P_n(y_i)$를 사용하여 i번째 스팬에 대한 선행사 표현 a_i^n을 계산한다.[5]

$$a_i^n = \sum_{y_i \in Y(i)} P_n(y_i) \cdot g^n_{y_i}$$

현재 i번째 스팬 표현 g_i^n는 선행사 표현 a_i^n를 이용하여 갱신된다. 선행사 표현과의 갱신 횟수는 하이퍼 파라미터로 조정된다.[5]

$$f_i^n = \sigma(W_f[g_i^n, a_i^n])$$
$$g_i^{n+1} = f_i^n \circ g_i^n + (1 - f_i^n) \circ a_i^n$$

앞에서 생성한 스팬 표현을 사용하여 각 스팬에 대한 언급 점수 s_m를 다음과 같이 정의한다.[5]

$$s_m(i) = w_m \cdot FFNN_m(g_i^n)$$

언급 점수를 기반으로 스팬 쌍의 선행사 점수 s_a를 다음과 같이 정의한다.[5]

$$s_a(i,j) = s_k(i,j) + s_l(i,j)$$
$$s_k(i,j) = (g_i^n)^T w_k\, g_j^n$$
$$s_l(i,j) = w_l \cdot FFNN_M([g_i^n, g_j^n, g_i^n \circ g_j^n, \Phi(i,j)])$$

여기서 $\Phi(i,j)$는 장르, 화자 정보와 두 스팬 사이의 거리에 대한 특질 벡터이다.[4] 본 논문에서 장르, 화자 정보는 사용하지 않는다.

위의 점수들을 이용하여 스팬 쌍에 대한 상호참조 점수 $s(i,j)$를 다음과 같이 정의한다.[5]

$$s(i,j) = s_m(i) + s_m(j) + s_a(i,j)$$

선행사를 찾고자 하는 스팬의 앞쪽에 위치한 스팬에 대해 상호참조 점수를 내어 점수를 바탕으로 선행사를 선택한다. 상호참조 점수들이 모두 음수라면 스팬의 선행사가 존재하지 않다고 결정한다.[4] 스팬들 간의 선행사를 모아 같은 개체들에 대한 언급들의 집합을 만들다.

기존의 상호참조모델들[3,4,5]은 계산 효율 및 성능 향상을 위해 선행사 후보 Y의 개수를 한정시킨다. 이로 인해 멀리 떨어진 스팬 간의 상호참조 여부는 고려하지 않는다. 따라서 본 논문에서는 표층형을 사용하는 규칙을 추가하여 성능을 향상시킨다. 규칙은 다음과 같다.

모든 집합 쌍에 대해서 집합 내부의 모든 스팬의 표층형을 비교한다. 이때 '그', '그녀' 와 같은 대명사를 제외한 표층형이 겹치는 스팬이 존재한다면 해당 집합 쌍을 하나로 합친다.

4. 실험

4.1 데이터 셋

본 실험에서 사용한 데이터 셋은 크라우드소싱을 통하여 위키피디아 2660개 문서를 대상으로 상호참조해결 데이터 셋을 구축했다. 데이터 셋 중 2407개 문서를 학습 데이터, 253개 문서를 개발 데이터로 사용했다. 테스트 데이터는 크라우드소싱을 통하여 위키피디아 140개

표 1. 상호참조 모델별 성능

단위 (%)	MUC F1	B^3 F1	CEAF-e F1	평균		
				정밀도	재현율	F1-score
Bi-LSTM 모델[5]	54.86	49.1	48.55	48.62	53.27	50.84
+ 정답 언급 후보 경계, 개체 유형 특질[10]	67.93	64.18	66.24	83.41	54.81	66.12
+ 표층형을 사용한 규칙(ours)	74.23	70.26	69.41	**85.71**	61.05	71.3
BERT 모델[3]	56.71	51.34	51.5	49.04	58.15	53.18
+ 표층형을 사용한 규칙(ours)	79.56	72.47	64.9	73.59	**71.1**	**72.31**

표 2. 테스트 데이터의 일부 문서. 예시문장 내에 굵은 글씨는 같은 개체를 나타내는 언급들이다.

예시 1	**마르코 혼파이 파비안 데 라 모라**는 멕시코의 축구 선수로 … **파비안**은 과달라하라 출신의 선수로 지역, 클럽 치바스의 유소년팀을 거쳤다.
예시 2	**오범석**은 대한민국의 축구 선수로서 … 반칙을 많이 범하여 '**반칙왕**'이라는 별명을 얻기도 하였다.
예시 3	**이영도**의 작품에서 등장하는 인공어들은 **작가 자신**이 댓글을 통해서 "언어에 공을 들이지 않았다"고 언급한적이 있으나, …

문서를 대상으로 개체 후보 목록을 만들고 전문가 4명이 수동으로 개체 후보들 간의 상호참조관계를 주석하였다. 이들 중 긴 문서는 반으로 잘라 총 207개의 문서가 되었다. 데이터 셋에 대한 문서 당 평균 구성은 표 3을 통해 확인 할 수 있다.

표 3. 실험 데이터 셋에 대한 문서당 평균 구성

	평균 문장 갯수	평균 언급 개수	평균 상호참조대상 언급 개수
학습 데이터	20.9	96.3	41.4
개발 데이터	19.8	92.4	38.4
테스트 데이터	12	52.5	18.5

4.2 상호참조해결 실험

본 실험에서는 형태소 분절을 위해 ETRI 형태소 분석기[12]를 사용한다. multilingual-BERT-base 모델[2]을 사전학습된 모델로 사용한다. 해당 모델은 12개의 트랜스포머 블록, 768개의 히든 레이어 차원 수, 12개의 self-attention 헤드를 가지는 사전학습된 모델이다. 4.1에서 설명한 훈련 데이터 셋을 20번 학습하였고 드롭아웃은 0.3이다. BERT 파라미터의 학습률은 1×10^{-5}이고 상호참조 파라미터의 학습률은 2×10^{-4}이다. 최대 토큰 길이는 128 이고, 스팬 표현 갱신 횟수는 1이다. 인공 신경망의 히든 레이어 차원 수는 1000이다.

4.3 실험 결과

표 1은 상호참조해결 연구들의 성능을 비교한 것이다. Bi-LSTM을 기반으로 한 연구[5], 이에 정답 언급 후보 경계와 개체 유형을 특질로 추가한 연구[10], 이에 표층형을 사용한 규칙을 추가하였고, BERT를 기반으로 한 연구[3], 이에 표층형을 사용한 규칙을 추가하여 성능을 비교하였다.

5. 분석

기존 두 모델[3,5]은 재현율 보다 정밀도가 떨어지는 모습을 볼 수 있다. 상대적으로 BERT 기반 모델[3]이 Bi-LSTM 기반 모델[5] 보다 정밀도, 재현율, F1-score에서 앞서는 걸 확인할 수 있다. Bi-LSTM 모델에 여러 특질을 추가한 연구[10]에서 해당 특질들을 통해 정밀도를 크게 향상시켰다. 이를 통해 본 논문의 모델에도 해당 특질을 추가한다면 정밀도가 크게 오를 것을 기대하고 있다.

표 2는 테스트 데이터의 일부를 예시를 가져온 것이다. 예시 1에서 '마르코 혼파이 파비안 데 라 모라' 와 '파비안' 은 같은 개체를 가지는 언급들이다. 예시 2에서 '오범석' 과 '반칙왕' 은 같은 개체를 가지는 언급들이고, 예시 3에서 '이영도' 와 '작가 자신' 또한 같은 개체를 가지는 언급들이다. 예시 1에서의 언급들은 비록 다른 표층형을 가지지만 두 번째 언급의 표층형이 첫 번째 언급의 표층형 내에 포함되어 있다. 예시 2와 3에서의 언급들은 완전히 다른 표층형을 가지기 때문에 문맥적 요소를 충분히 이해하여야 상호참조를 할 수 있는 예시이다. 예시 1-3 모두 본 논문의 모델에서는 맞췄지만 [10]의 모델에서는 틀린 예시이다. 이를 통해 BERT를 이용한 fine-tuning 모델이 기존 Bi-LSTM 기반 모델[10] 보다 문맥적 특징을 잘 잡아낸다고 볼 수 있다.

6. 결론

본 논문에서는 영어권 상호참조 모델[3]을 기반으로 multilingual-BERT-base 모델에 한국어 데이터를 적용해 보았다. 또한 상호참조 관계에 있는 언급들의 집합 간 표층형을 비교하는 규칙을 추가하여 성능을 높였다. 실험을 통해 기존의 상호참조 연구들[3,5,10]과 성능을 비교하였다. 실험 결과, CoNLL 정밀도 73.6% 재현율 71.1%, F1-score 72.3%의 성능을 보였으며 이는 정답 언급 후보 경계, 개체 유형을 특질로 사용한 기존 한국어 상호참조해결 연구[10]보다 더 좋은 성능이다. 또한 실험 결과를 분석하여 기존 연구[10] 보다 BERT를 사용한 본 논문의 모델이 문맥적인 요소를 잘 이해하는 것을 확인했다. 향후 연구로는 본 논문의 모델에서 정답 언급 후보 경계, 개체 유형과 언급 집합 내의 스팬 표현을 특질로 추가할 예정이다.

사사

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2013-0-00109, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] 김지호, et al., “지식베이스 확장을 위한 행렬 분해 모델”, 제 29회 한글 및 한국어 정보처리 학술대회 논문집, pp. 3-7, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformersfor language understanding,” NAACL-HLT, 2018.

[3] M. Joshi, O. Levy, D. S. Weld, and L. Zettlemoyer, “Bert for coreference resolution: Baselines and analy-sis,” ArXiv, Vol. abs/1908.09091, 2019.

[4] Lee, K., et al., “End-to-End Nerual Coreferece Resolution”, Proceedings of the 2017 Coreferece on Empirical Methods in Natural Language Processing, pp. 188-197, 2017.

[5] Lee, K., He, L., & Zettlemoyer, L., “Higher-order Coreference Resolution with Coarse-to-fine Inference”, Proceedings of NAACL-HLT 2018, pp. 687-692, 2018.

[6] H. Lee, A. X. Chang, Y. Peirsman, N. Chambers,M. Surdeanu, and D. Jurafsky, “Deterministic corefer-ence resolution based on entity-centric, precision-rankedrules,” Computational Linguistics, Vol. 39, pp. 885-916, 2013.

[7] J. Pennington, R. Socher and C. Manning, “Glove: Global vectors for word representation”, Proceedings of the 2014 conference on empirical methods in natural language processing, pp. 1532-1543, 2014.

[8] M.E. Peters, M. Neumann, M. Ivver, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, “Deep contextualized word representations”, arXiv:1802.05365, 2018.

[9] 박천음, 이창기, “포지션 인코딩 기반 스택 포인터 네트워크를 이용한 한국어 상호참조해결”, 정보과학회 컴퓨팅의 실제 논문지, 24.(3), pp.113-121, 2018.

[10] 신기연, et al., “언급 특질을 이용한 Bi-LSTM 기반 한국어 상호참조해결 종단간 학습”, 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp. 247-251, 2018.

[11] 박천음, et al., “BERT기반 Deep Biaffine을 이용한 한국어 상호참조해결” 한국정보과학회 학술발표 논문집, pp. 488-490, 2019.

[12] http://aiopen.etri.re.kr/