기계 독해 기술을 이용한 한국어 대명사 참조해결

이동헌○, 김기훈, 이창기, 류지희, 임준호

강원대학교, 한국전자통신 연구원
[skh1578, rlarlgnsu, leeck]@kangwon.ac.kr, [chrisjihee, joonho.lim]@etri.re.kr

Korean Coreference Resolution using Machine Reading Comprehension

Dong-heon Lee○, Ki-hun Kim, Chang-ki Lee, Ji-hee Ryu, Joon-ho Lim
Kangwon University, Electronics and Telecommunications Research Insitute

요 약

대명사 참조해결은 문서 내에 등장하는 대명사와 이에 대응되는 선행사를 찾는 자연어처리 태스크이다. 기계 독해는 문단과 질문을 입력 받아 질문에 해당하는 알맞은 정답을 문단 내에서 찾아내는 태스크이며, 최근에는 주로 BERT 기반의 모델이 가장 좋은 성능을 보이고 있다. 이러한 BERT 기반 모델의 성공에 따라, 최근 여러 연구에서 자연어처리 태스크를 기계 독해 문제로 변환하여 해결하는 연구들이 진행되고 있다. 본 논문에서는 최근 여러 자연어처리에서 뛰어난 성능을 보이고 있는 BERT 기반 기계 독해 모델을 이용하여 한국어 대명사 참조해결 연구를 진행하였다. 사전 학습 된 기계 독해 모델을 사용하여 한국어 대명사 참조해결 데이터로 fine-tuning하여 실험한 결과, 개발셋에서 EM 78.51%, F1 84.79%의 성능을 보였고, 평가셋에서 EM 70.78%, F1 80.19%의 성능을 보였다.

주제어: 대명사 참조해결, 상호참조해결, 기계독해, BERT

1. 서론

상호참조해결(coreference resolution)은 멘션 후보를 식별하고, 동일한 개체의 서로 다른 멘션들을 찾아 같은 개체로 그룹화 하는 자연어처리 태스크이다. 상호참조해결은 입력 받은 문서에서 명사(구) 혹은 대명사(구)와 같은 멘션 후보들을 식별하는 멘션 탐지 단계와 탐지한 멘션에 대한 선행사를 찾아 연결하는 상호참조해결 단계로 나뉘어지며, 그 예는 다음과 같다.

문단: "**[감자역병균]**은 **[[난균]의 일종]**으로, **[감자]**에 **[[감자마름병]이라는 심한 질병]**을 일으킨다. **[이]**는 **[1845년 아일랜드와 1846년 스코틀랜드 고지에서 일어난 [[감자] 대기근의 주요 원인 가운데 하나이다.]**"

위의 문단에서 "[]"로 표시된 단어들이 멘션 탐지 단계에서 추출한 멘션들(**[감자역병균]**, **[난균]**, **[난균의 일종]**, **[감자]**, **[감자마름병]**, …)에 해당한다. 멘션 탐지 단계 이후, 상호참조해결 단계에서 탐지한 멘션들 사이의 동일한 개체들을 그룹화한다(**[[감자역병균]**, **[난균]**, **[이]**, **[1845년 … 하나이다.]]**).

대명사 참조해결은 문서 내에 등장하는 대명사와 이에 대응되는 선행사를 찾는 자연어처리 태스크로, 상호참조해결과 유사하지만 대명사와 이 대명사의 선행사 만의 관계를 찾는다는 차이가 있다. 위의 예제에서 대명사 "[이]"의 선행사는 "[감자역병균]"이다(본 논문에서는 선행사가 여러 개일 경우 제일 먼저 나온 선행사를 찾는다고 가정함).

기계 독해는 입력 받은 문단과 질문을 이해하여 문단 내에서 알맞은 정답의 위치를 출력하는 자연어처리 태스크이며, 최근에는 주로 BERT 기반의 모델이 가장 좋은 성능을 보이고 있다[1]. 이러한 BERT 기반 모델의 성공에 따라, 최근 여러 연구에서 개체명 인식, 관계 추출, 상호참조해결과 같은 자연어처리 태스크를 기계 독해 문제로 변환하여 해결하는 연구가 진행되고 있다[2,3,4].

본 논문에서는 한국어 대명사 참조해결을 위해 기존의 상호참조해결 과정의 하나인 멘션 탐지 단계를 생략하고 기계 독해 기술을 이용하여 대명사의 선행사 위치(시작 및 끝)를 찾아주는 모델을 제안한다. 이를 위해, 대명사가 포함된 문장으로부터 대명사의 앞 뒤에 마크업 태그를 붙여 질문을 생성하고 기계 독해 기술을 이용하여 주어진 문서에서 대명사의 선행사 위치(시작 및 끝)를 찾는다.

2. 관련 연구

최근 한국어 상호참조해결 연구는 BERT 기반 end-to-end 방법과 포인터 네트워크를 사용한 기계 학습 방법이 연구되었다.

[2]에서는 멘션 랭킹 모델에 BERT를 추가한 모델[3]을 사용하였으며, 의존 구문 분석 자질과 개체명 자질을 적용하여 한국어의 의미적, 구조적 특징을 반영하는 모델을 제안하였다.

[4]에서는 포인터 네트워크를 사용하여 RNN encoder-decoder의 고정된 출력 사전 길이를 갖는 문제를 해결하여 대명사 참조해결에 많은 성능 향상을 보였다.

[5,6,7]의 연구는 개체명 인식, 관계 추출, 상호참조해결 태스크를 기계 독해 문제로 변환하여 해결하였다. [5]은 특정 토큰에 단일 레이블에만 할당 할 수 있는 시퀀스 레이블링 모델의 문제를 해결하여 단일 및 중첩 개체명 인식 문제를 모두 처리하는 모델을 제안하였다. [6]은 문맥과 질문의 관계를 파악하여 기존에 없던 새로운

사실 관계를 유추하여 관계 추출 문제를 해결하였다. [7]은 주변 문맥을 사용하여 후보 멘션에 대한 질문을 생성하고 이를 사용하여 문단 내에서 상호 참조 멘션의 위치를 출력하였다.

3.기계 독해 기술을 이용한 대명사 참조해결 모델

본 논문에서는 한국어 대명사 참조해결 문제를 기계 독해 문제로 변환하여 해결한다. 이를 위해, 참조해결이 필요한 대명사가 포함된 문장으로부터 질문을 생성하고 대명사가 포함된 문서를 문단으로 사용하여 기계 독해 모델을 통해 대명사의 선행사 위치(시작 및 끝)를 찾는다.

표 1은 대명사 참조해결을 위한 기계 독해 모델의 입력인 질문과 문단의 예이다. 문단은 기존 상호참조해결 데이터의 문서 전체를 사용하였고, 질문은 참조해결이 필요한 대명사가 포함되어 있는 문장으로 지정하고 <S>, <T> 태그를 이용하여 대명사 위치를 표시하였다. 정답은 해당 대명사의 선행사 중 대명사(구)가 아니면서 첫 번째로 등장하는 선행사(예: 감자영병균)로 설정하였다. 이러한 기계 독해 모델의 학습을 위해서 ETRI 상호참조해결 데이터를 SQuAD (Stanford Question Answering Dataset)[8]형식(정답, 질문, 문단)으로 변경하는 전처리 작업을 수행하였는데, 대명사 참조해결 만을 위한 데이터의 양이 부족하기 때문에 표 2와 같은 일반 명사의 상호참조해결 데이터도 함께 이용하여 SQuAD 형식으로 변경하였다. 이렇게 변경된 SQuAD 형식의 학습데이터의 양은 학습셋 6771문서, 개발셋 242문서, 평가셋 332문서이다.

표 1. 대명사 참조해결을 위한 기계 독해 모델 입력 예

문단1	**감자역병균**은 난균의 일종으로, 감자에 감자마름병이라는 심한 질병을 일으킨다. **이**는 1845년 아일랜드와 1846년 스코틀랜드 고지에서 일어난 감자 대기근의 주요 원인 가운데 하나이다.
질문1	**<S> 이는 <T>** 1845년 아일랜드와 1846년 스코틀랜드 고지에서 일어난 감자 대기근의 주요 원인 가운데 하나이다.
정답1	**감자역병균**
문단2	파로나마와 유사한 이것은 **배경 위에 미니어처와 같은 축소 모형을 설치해 하나의 장면을 만드는 것이다.** 전시물의 입체감을 현장성에 충실하도록 표현해 하나의 사실 또는 주제의 시간 상황을 고정시켜 연출하는 **이것은** 무엇일까?
질문2	전시물의 입체감을 현장성에 충실하도록 표현해 하나의 사실 또는 주제의 시간 상황을 고정시켜 연출하는 **<S> 이것은 <T>** 무엇일까?
정답2	**배경 위에 미니어처와 같은 축소 모형을 설치해 하나의 장면을 만드는 것이다.**

표 2. 상호참조해결을 위한 기계 독해 모델 입력 예

문단	BC 300년경에 활약한 그리스의 수학자. **그리스 기하학**, 즉 '유클리드 기하학'의 대성자이다. 그의 저서 《기하학원론》은 기하학에 있어서의 경전적 지위(經典的地位)를 확보함으로써 유클리드라 하면 기하학과 동의어로 통용되는 정도에 이르고 있다.
질문	그의 저서 《기하학원론》은 **<S> 기하학에 <T>** 있어서의 경전적 지위(經典的地位)를 확보함으로써 유클리드라 하면 기하학과 동의어로 통용되는 정도에 이르고 있다.
정답	**그리스 기하학**

기계 독해 기술을 이용한 대명사 참조해결 모델은 기존의 BERT 기반 기계 독해 모델과 동일하다. BERT 기반 기계 독해 모델은 하나의 질문에 한 개의 정답만을 찾을 수 있기 때문에, 본 논문에서 제안하는 대명사 참조해결 모델은 한번에 한 개의 대명사 참조해결만 가능하다. 따라서 한 문서에 여러 개의 대명사가 있을 경우에는 각 대명사 마다 따로 참조해결 작업을 수행해야 한다(실제 구현 시에는 배치를 이용해 여러 대명사를 동시에 참조해결할 수 있다).

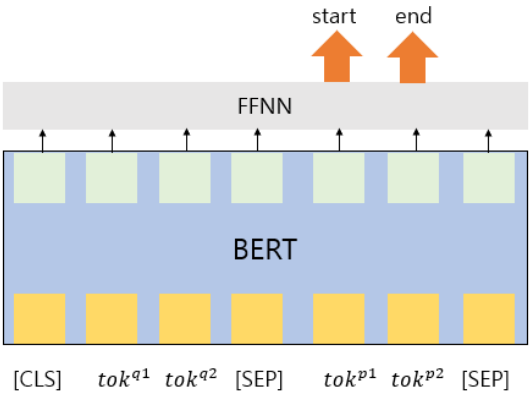
그림 1은 BERT 기반의 기계 독해 기술을 이용한 대명사 참조해결 모델을 나타낸다. 모델의 입력은 질문 $Q = \{q_1, \dots, q_n\}$와 문단 $P = \{p_1, \dots, p_m\}$를 연결한 입력열 $X = \{[CLS], tok_1^q, \dots tok_n^q, [SEP], tok_1^p, \dots, tok_m^p, [SEP]\}$이고, 입력열 X가 BERT 모델을 거쳐 히든 스테이트 T가 만들어진다. T는 시작과 끝일 확률을 출력하기 위하여 출력 차원수를 2로 맞춘 FFNN(Feed Forward Neural Network)를 거쳐 h_i를 만든다. h_i는 softmax를 적용하여 정답의 시작 및 끝 확률인 y_s와 y_e를 만든다. 이에 따른 수식은 다음과 같다.

$$T = BERT(X) \quad (1)$$

$$h_i = FFNN(T) \in \mathbb{R}^{n \times 2} \quad (2)$$

$$y_s = softmax(h_i) \in \mathbb{R}^n \quad (3)$$

$$y_e = softmax(h_i) \in \mathbb{R}^n \quad (4)$$



[그림 1] BERT 기반 기계 독해 모델

4. 실험

BERT는 양방향 트랜스포머(transformer)[9]의 인코더를 이용하며, 입력 텍스트 중 임의의 토큰에 마스크를 씌우고, 문맥을 파악하여 마스킹 된 단어를 출력하는 Masked Language Modeling(Masked LM)방법과 다음 문장을 예측하는 방법으로 사전학습한다. 본 논문에서는 ETRI에서 수집한 뉴스 및 위키피디아 데이터 약 23.5 기가바이트에 대하여 형태소 분석을 적용하여 ETRI에서 학습한 KorBERT 모델을 사용하였다. BERT 사전학습 시의 하이퍼파라미터는 BERT-base(L = 12, H = 768, A = 12)를 따르며, 히든레이어의 활성 함수는 gelu[10]이고, 드랍아웃(dropout)은 0.1이다.

학습데이터는 3장에서 설명한 ETRI 상호참조해결 데이터를 SQuAD 형식으로 변환한 데이터(학습셋 6771문서, 개발셋 322문서, 평가셋 242문서)를 사용했으나, 학습셋의 양이 부족하기 때문에 한국어 기계 독해 데이터셋인 KorQuAD v1.0[11]을 이용하여 미리 학습된 BERT 모델을 초기 모델로 이용하였다. Fine-tuning 수행 시의 하이퍼파라미터는 다음과 같다. 학습율(earning rate)은 3e-5로 설정하였고, 입력 최대 문장 길이는 384. 문단 stride는 128로 설정하였다. 히든 레이어의 차원 수는 150으로 설정하였으며, 드랍아웃은 0.2, 스택 수는 3으로 설정하였다. 학습에는 Adam[12]을 사용하였다. 대명사 참조해결 성능 평가는 SQuAD의 평가 스크립트를 이용하여 EM(Exact Match)과 F1 점수를 측정하였다.

표 3. 대명사 참조해결 성능 비교(%)

Dev		
Model	EM	F1
BERT 기반 대명사 참조해결	78.51	84.79
BERT 기반 대명사 참조해결 (without KoQuAD pre-training)	77.27	84.71
Test		
BERT + End-to-end 상호참조해결	-	62.70*
BERT 기반 대명사 참조해결	70.78	80.19
BERT 기반 대명사 참조해결 (without KoQuAD pre-training)	67.17	78.22

표 3은 한국어 대명사 참조해결의 성능을 나타낸다. BERT 기반 대명사 참조해결 모델은 사전학습 된 기계독해 모델에 대명사 참조해결 데이터셋을 fine-tuning한 것으로, 개발셋에서 EM 78.51%, F1 84.79%의 성능을 보였고, 평가셋에서 EM 70.78%, F1 80.19%의 성능을 보였다. 사전학습된 기계 독해 모델을 사용하지 않았을 경우, 개발셋에서 EM 77.27%, F1 84.71%로 약간의 성능 하락을 보였고, 평가셋에서는 EM 3.61%, F1 1.97%의 성능 하락을 보여, 사전학습된 기계 독해 모델을 사용하는 것이 성능 향상에 도움이 됨을 알 수 있다. 또한 기존의 BERT 기반 End-to-end 상호참조해결 모델[4]의 MUC[13] 성능과 비교해 보면 대명사 참조해결의 성능이 좋음을 알 수 있다(단, 일반 명사의 상호참조해결이 포함되어 있고, head boundary 기준의 MUC 성능으로 직접적인 성능 비교는 어려움).

표 4. 대명사 참조해결 예시

문단1	파노라마와 유사한 이것은 **[배경 위에 미니어처와 같은 축소 모형을 설치해 하나의 장면을 만드는 것이다.]** 전시물의 입체감을 현장성에 충실하도록 표현해 하나의 사실 또는 주제의 시간 상황을 고정시켜 연출하는 이것은 무엇일까?
질문1	전시물의 입체감을 현장성에 충실하도록 표현해 하나의 사실 또는 주제의 시간 상황을 고정시켜 연출하는 **<S> 이것은 <T>** 무엇일까?
예측1	**배경 위에 미니어처와 같은 축소 모형을 설치해 하나의 장면을 만드는 것이다.**
정답1	**배경 위에 미니어처와 같은 축소 모형을 설치해 하나의 장면을 만드는 것이다.**
문단2	**[표유류·조류의 내이에 있는 고등 모양의 기관이다.]** 중앙에 있는 방을 달팽이 세관이라고 하는데 이것은 청각의 중심이 된다. 와우관이라고도 한다.
질문2	중앙에 있는 방을 달팽이 세관이라고 하는데 **<S> 이것은 <T>** 청각의 중심이 된다.
예측2	**표유류·조류의 내이에 있는 고등 모양의 기관이다.**
정답2	**표유류·조류의 내이에 있는 고등 모양의 기관이다.**

표 4는 BERT 기반 한국어 대명사 참조해결 결과의 예제이다. 질문에서 <S>, <T> 태그로 표시한 단어는 참조해결 대상인 대명사를 의미하며, 문단에서 []로 묶인 단어 및 문장은 이 대명사의 선행사를 의미한다. 표 4에서 대명사('이것은', '이것은')의 선행사로 예측한 결과가 정답과 일치하는 것을 알 수 있다.

표 5. 상호참조 오류 예시

문단	탱고는 1880년경 아르헨티나 부에노스 아이레스의 동남쪽에 위치한 항구도시 보카(Boca)에서 탄생한 음악이다. 19세기 말에서 제1차 세계대전 전까지 보카를 통해 유럽으로부터 **[엄청난 수의 이민자들이]** 아르헨티나로 모여들었고 그들 속에서 탱고가 태어났다. 탱고는 유럽 계통의 무곡과 아프리카계 주민의 민속음악이 혼합된 것이라고 보는 것이 정설로 되어 있다.
질문	**19세기 말에서 제1차 세계대전 전까지 보카를 통해 유럽으로부터 엄청난 수의 이민자들이 아르헨티나로 모여들었고 <S> 그들 <T> 속에서 탱고가 태어났다.**
예측	**유럽으로부터 엄청난 수의 이민자들이**
정답	**엄청난 수의 이민자들이**

표 5는 잘못된 대명사 참조해결의 예시로 멘션 **<S> 그들 <T>**의 선행사로 예측한 **'유럽으로부터 엄청난 수의 이민자들이'** 이 정답 '엄청난 수의 이민자들이' 과 일치하지 않고 있다.

5. 결론

본 논문에서는 BERT 기반 기계 독해 모델을 이용한 한국어 대명사 참조해결 모델을 제안하였다. 사전 학습 된 기계 독해 모델을 사용하여 한국어 대명사 참조해결 데이터로 fine-tuning하여 실험한 결과, 개발셋에서 EM 78.51%, F1 84.79%의 성능을 보였고, 평가셋에서 EM 70.78%, F1 80.19%의 성능을 보였다.

향후 연구로는 개체명 태그 등의 자질을 추가하여 대명사 참조해결의 성능을 향상 시킬 예정이고, 다른 자연어처리 태스크에도 기계독해 기술을 적용할 예정이다.

감사의 글

이 논문은 2020년도 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] Devlin et al, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018

[2] 김기훈, at al. BERT 기반 End-to-end 신경망을 이용한 한국어 상호참조해결, HCLT, 2019.

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

[4] 박천음, et al. 포인터 네트워크를 이용한 대명사 상호참조해결, KCC, 2016.

[5] Xiaoya Li, at al. A Unified MRC Framework for Named Entity Recognition arXiv preprint arXiv:1910.11476, 2020.

[6] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. arXiv preprint arXiv:1706.04115. 2017.

[7] Wei Wu, et al. Corefernce Resolution as Query-based Span Prediction. arXiv preprint arXiv:1911.01746, 2019.

[8] P. Raipurkar, et al. Squad: 100,000+ questions for machine comprehension of text arXiv preprint arXiv:1606.05250, 2016.

[9] Vaswani et al, Attention Is All You Need, arXiv Preprint arXiv:1706.03762, 2017

[10] D. Hendryks, K. Gimpel. Gaussian Error Linear Units(GELUs), arXiv:1606.08415,2018.

[11] 임승영, et al. KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋. KSC, 2018.

[12] D.Kngma and J.Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[13] Marc B. Vilain, et al. A model-theoretic coreference scoring scheme, In Proc. Of ACL, 1995.

[14] R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. In Proc. Of ACL, pp.1715-1725, 2016.

[15] 이동헌, at al. BERT를 이용한 한국어 기계 독해, KCC, 2019.