

전문용어 탐지와 해석 모델: 한국어 의학용어 중심으로

염하람^o, 김재훈

한국해양대학교, 한국해양대학교 컴퓨터공학과 및 해양인공지능융합전공
rami7878@naver.com, jhoon@kmou.ac.kr

Detecting and Interpreting Terms: Focusing Korean Medical Terms

Haram-Yeom^o, Jae-Hoon Kim

Korea Maritime & Ocean University, Dept. of Computer Engineering and Interdisciplinary Major of
Maritime AI Convergence,
Korea Maritime & Ocean University

요 약

최근 COVID-19로 인해 대중의 의학 분야 관심이 증가하고 있다. 대부분의 의학문서는 전문용어인 의학 용어로 구성되어 있어 대중이 이를 보고 이해하기가 어려움이 있다. 의학용어를 쉬운 뜻으로 풀이하는 모 델을 이용한다면 대중이 의학 문서를 쉽게 이해할 수 있을 것이다. 이런 문제를 완화하기 위해서 본 논문 에서는 Transformer 기반 번역 모델을 이용한 의학용어 탐지 및 해석 모델을 제안한다. 번역 모델에 적용 하기 위해 병렬말뭉치가 필요하다. 본 논문에서는 다음과 같은 방법으로 병렬말뭉치를 구축한다: 1) 의학 용어 사전을 구축한다. 2) 의학 드라마의 자막으로부터 의학용어를 찾아서 그 뜻풀이로 대체한다. 3) 원자 막과 뜻풀이가 포함된 자막을 나란히 배열한다. 구축된 병렬말뭉치를 이용해서 Transformer 번역모델에 적용하여 전문용어를 찾아서 해석하는 모델을 구축한다. 각 문장은 음절 단위로 나뉘어 사전학습 된 KoCharELECTRA를 이용해서 임베딩한다. 제안된 모델은 약 69.3%의 어절단위 BLEU 점수를 보였다. 제 안된 의학용어 해석기를 통해 대중이 의학문서를 좀 더 쉽게 접근할 수 있을 것이다.

주제어: 번역모델, 전문용어 탐색, 전문용어 해석, 병렬말뭉치

1. 서론

세계보건기구(WHO)의 Coronavirus Disease-19 (COVID-19) 팬데믹 선언으로 대중들의 의학 분야에 대한 관심이 증가하고 있다. COVID-19의 증상 및 백신 관련 의학 문서를 개인이 열람하는 경우가 많아지면서 의학용어가 대중에게 노출되는 빈도가 늘어나고 있다. 그러나 의학용어가 많이 포함된 문서는 비전문가가 읽고 이해하기가 많은 어려움이 있다. 의학용어는 주로 외국어에서 유래된 것이 많으며, 같은 대상을 여러 명칭으로 부르는 특징을 가진다. 또한 긴 의학용어는 주로 약어로 사용되 기 때문에 의학용어가 생소한 비전문가의 의학 문서 접근성은 떨어질 수밖에 없다. 비전문가인 대중들이 의학 용어를 알아야 하는 가장 큰 이유는 소통이다. 의학 지식을 가진 전문가들이 증상과 치료법에 대해서 이야기하 면 환자는 이를 이해할 수 있어야 한다. 또한 COVID-19 와 같은 전 세계적인 질병이 발생했을 때 쏟아져 나오는 의학용어를 보고 이해할 수 있어야 질병을 예방하고 치 료할 수 있을 것이다. 일상생활에서 주로 사용되지 않으며, 소통에 비효율적인 현재 의학용어를 개선하기 위해 오래전부터 의학 전문가들의 많은 노력이 있었지만 이미 많은 문서가 어려운 의학용어로 작성되어 있으며 오래도 록 사용해온 의학용어를 한 번에 바꾸기엔 어려움이 있 다¹⁾. 이와 같은 소통의 문제를 해결하기 위해 기계번역 을 사용하여 의학용어로 쓰인 문서를 해석하는 모델을

제안하고자 한다.

기계번역은 입력으로 받은 원시문장을 자동으로 컴퓨 터가 목적문장으로 번역하는 소프트웨어를 말한다. 인공 신경망 기계번역(Neural Machine Translation; NMT)은 신경망 모델의 가중치들을 학습하는 End-to-End 방법론 으로 기존 규칙 및 통계기반 기계번역을 뛰어넘는 성능 을 보여준다[1]. NMT는 Sequence-to-Sequence 모델을 기 반으로 인코더-디코더 구조를 가진다. 인코더를 통해 원 시문장을 벡터화한 정보를 문맥벡터로 압축하고 압축된 정보를 기반으로 디코더가 목적문장으로 번역한다[2].

본 논문에서는 NMT에서 높은 성능을 보이는 Transformer 모델을 이용하여 의학용어를 탐지하고 해석 하는 모델을 제안한다. 제안된 모델은 입력문장의 의학 용어를 찾고 이를 뜻풀이로 치환한 문장을 출력하는 역 할을 한다. 번역모델을 학습하기 위해서는 대량의 병렬 말뭉치가 필요하다. 하지만 의학용어를 탐지하고 해석하 기 위한 한국어 말뭉치는 현재 존재하지 않는다. 따라서 본 논문에서는 웹에서 의학용어를 크롤링하여 의학용어 사전을 제작하고, 의학 드라마의 자막을 이용해 두 종류 의 병렬 말뭉치를 구축한다. 하나는 의학용어를 의학용 어 뜻풀이로 치환하여 원문장과 나란히 배열한 말뭉치이 고, 다른 하나는 의학용어/뜻풀이 앞뒤에 의학용어 경계 토큰을 추가한 말뭉치이다. 의학용어를 포함한 문장을 원시문장으로, 용어의 뜻이 풀어진 문장을 목적문장으로 정의한다.

모델의 성능 평가는 기계번역의 성능을 평가할 때 주로 사용되는 BLEU 점수를 이용한다. 두 종류의 말뭉치

1) <http://www.jjan.kr/2055264>

를 각 어절, 음절단위로 성능을 평가했으며 경계 토큰이 없는 말뭉치로 학습한 모델은 어절단위 0.693, 음절단위 0.770의 성능을 보인다. 의학용어 경계 토큰이 포함된 말뭉치로 학습한 모델은 어절단위 0.586, 음절단위 0.661의 성능을 보인다.

의학용어를 찾아 뜻풀이로 해석하는 모델을 만듦으로 비전문가는 전문가의 도움 없이 의학 문서를 읽고 이해할 수 있을 것이다. 이를 통해 비전문가의 의학 분야 접근성이 높아질 것을 기대한다.

본 논문의 구성은 다음과 같다. 2장에서 기계번역과 관련된 연구를 조사하고, 3장에서 모델 훈련을 위한 데이터 구축 방법에 대해 자세히 기술한다. 4장에서는 모델을 평가하고, 5장에서 결론을 맺는다.

2. 관련 연구

기계번역은 입력으로 받은 원문장을 자동으로 컴퓨터가 목적문장으로 번역하는 소프트웨어를 말한다. 기계번역은 규칙 기반, 통계 기반, 딥러닝 기반 순으로 발전해왔다. 규칙 기반 기계번역은 언어학자가 정해놓은 규칙에 기반하여 번역하는 방법이다. 이 방법의 장점은 규칙에 맞는 문장은 정확하게 번역하는 것이다. 하지만 많은 양의 언어적 지식이 필요하고, 모든 언어의 규칙을 아는 것은 불가능하기 때문에 번역 언어 확장이 어렵다는 약점이 있다[3].

통계 기반 기계번역은 대용량 병렬 말뭉치로부터 학습된 통계 정보를 이용하여 번역을 진행하는 방식이다[4]. 통계 기반 기계번역은 언어학적 지식이 불필요하며 데이터의 양이 많아질수록 번역의 품질이 높아진다는 장점이 있다. 그러나 대량의 데이터를 구축하기는 어려우며 단어나 구 단위로 번역이 진행되기 때문에 문맥을 이해하기 어렵다는 단점을 가진다[5].

딥러닝 기반 기계번역인 인공신경망 기계번역은 신경망 모델의 가중치들을 학습하는 end-to-end 방법론으로 인코더-디코더 구조를 가지는 Sequence-to-Sequence 모델을 기반으로 한다. 원문장을 인코더의 입력으로 넣어 문맥벡터를 계산하고 이 정보를 이용해 디코더가 목적문장을 만들어 낸다[6-7]. Transformer 모델[8]은 기계번역 분야에서 최고의 성능과 학습 효율성을 보여준다.

현재 전문용어를 해석하는 연구는 활발히 진행되고 있지 않으며, 한국어의 경우 전문용어 인식에 관한 연구[9]는 존재하지만, 전문용어를 뜻풀이로 해석하는 모델은 존재하지 않는다. 본 논문은 NMT를 이용해 전문용어를 해석하는 모델을 제안하고자 한다.

3. 병렬말뭉치 구축

병렬말뭉치 구축은 다음과 같은 세 단계로 구성된다: 1) 의학용어 사전 제작 2) 의학 문서 전처리 3) 의학 병렬말뭉치 구축

3.1 의학용어 사전 제작

본 논문에서 의학용어라 함은 일반적인 의학용어뿐 아니라 병원에서 주로 사용되는 용어나 약어를 총칭한다.

의학용어 웹사이트²⁾에서 크롤링한 용어와 뜻풀이를 CSV 파일로 저장한다. 이때 저장된 뜻풀이는 일반적으로 여러 문장으로 구성되어 용어의 정의를 일컫는 부분 이외의 것을 제외하는 과정이 필요하다. 크롤링한 대부분 용어의 핵심적인 뜻은 첫 문장에 위치하고 있어[10] 휴리스틱을 사용하여 사전의 뜻은 각 용어 정의 문서의 첫 문장을 이용해 정의한다. 크롤링한 데이터에는 하나의 대상을 의미하는 여러 용어가 같은 정의로 함께 저장되어 있다. 각 용어를 추출하여 정의와 짝을 맞춘다. 의학용어는 한국어, 영어, 약어가 서로 다른 단어로 저장된다.

표 1. 의학용어 사건의 예

| Type | Term | Meaning |
|------|----------------|--------------------|
| 한국어 | 노르에피네프린 | 혈압을 올리기 위해 사용하는 약물 |
| 영어 | Norepinephrine | 혈압을 올리기 위해 사용하는 약물 |
| 약어 | 노르에피 | 혈압을 올리기 위해 사용하는 약물 |

표 1은 의학용어 사건의 예이다. 용어[“노르에피네프린”, “Norepinephrine”, “노르에피”]와 용어의 뜻풀이인 “혈압을 올리기 위해 사용하는 약물”이 병렬로 저장된 형태를 각각 같은 뜻풀이를 가지는 세 단어로 저장하여 사전을 구성한다. 만들어진 의학용어 사전은 5,240개의 단어로 이루어져 있다.

3.2 의학 문서 전처리

의학 문서는 일상생활에서 의학용어를 쉽게 접할 수 있는 의학 드라마의 자막을 이용한다. 의학 드라마의 자막 파일은 시간, 글꼴 등 다양한 정보를 포함하고 있으므로 대화 정보만을 가지고 온다. 드라마의 자막은 구어체를 그대로 문서화하였기 때문에 말을 더듬는 것이 그대로 표현되어있고, 의미 없이 나열되는 감탄사가 많이 포함되어 있다. 또한, 인물의 감정을 나타내는 지문이 포함되어 있다. 의미 없이 반복되는 단어와 지문을 제거하여 의학 문서를 수집한다.

3.3 병렬말뭉치 구축

의학 문서 문장에 있는 의학용어가 자체 제작 사전에 포함되어 있다면 이를 의학용어의 뜻으로 치환하여 병렬말뭉치를 만든다. 병렬말뭉치는 의학용어가 포함된 문장과 포함되지 않은 문장으로 이루어져 있다. 만들어진 병렬말뭉치는 20,248문장이다.

2) <http://www.kmle.co.kr/>

표 2. 병렬말뭉치의 예

| | |
|----------|-------------------------|
| 원시문장 | 목적문장 |
| 드랩하면 되죠? | 소독 후 멸균 천으로 덮기하면 되죠? |

표 2는 제작한 병렬말뭉치의 예이다. 원시문장의 “드랩”을 의학용어 사전의 뜻 “소독 후 멸균 천으로 덮기”로 치환하여 목적문장을 구성한 것을 볼 수 있다.

3.4 경계 토큰이 추가된 병렬말뭉치

3.3에서 만든 병렬말뭉치를 이용하여 모델을 학습한 결과, 의학용어를 잘 인식하지 못하는 경우가 다수 발견되어 입력에 위치 토큰을 추가하여 위치 정보를 이용한다[10]. 의학용어와 뜻풀이의 앞과 뒤에 새로운 토큰을 추가하여 의학용어의 위치 정보를 입력문장에 포함한다. 이는 의학용어의 경계를 인식하는 의도를 반영하는 것이다. 의학용어에 경계 토큰을 추가하여 트랜스포머의 입력으로 넣었을 때 의학용어를 잘 인식하고, 해석을 더 잘 수행하는지 확인하는 새로운 말뭉치를 구축한다. 의학용어의 앞과 뒤와, 의학용어 뜻풀이의 앞과 뒤를 표시하는 토큰을 추가한다. <m>은 용어와 뜻이 시작되는 지점에 추가하고, </m>은 용어와 뜻이 끝나는 지점에 추가한다.

표 3. 위치 토큰이 추가된 병렬말뭉치의 예

| | |
|--------------------|--------------------------------|
| 원시문장 | 목적문장 |
| <m>드랩</m>하면 되죠? | <m>소독 후 멸균 천으로 덮기</m>하면 되죠? |

표 3은 의학용어 경계 토큰을 추가하여 제작한 말뭉치의 예이다. 원시문장의 의학용어 “드랩” 앞에 <m>토큰을 추가하였고, 끝에 </m>토큰을 추가하였다. 목적문장에도 마찬가지로 뜻풀이 “소독 후 멸균 천으로 덮기” 앞에 <m>토큰을 추가하고, 끝에 </m>토큰을 추가한 것을 볼 수 있다.

4. 실험 및 평가

4.1 실험환경

3장에서 구축한 두 종류의 병렬말뭉치 20,248문장을 실험 데이터 셋으로 사용한다. 전체 데이터의 훈련 및 평가 데이터는 표 4와 같이 분리한다.

표 4 훈련 데이터와 평가 데이터

| | |
|--------|--------|
| 훈련 데이터 | 평가 데이터 |
| 19,248 | 1,000 |

학습에 사용된 Transformer 모델의 하이퍼파라미터는 다음과 같다. 3개의 Transformer 블록 계층을 사용하였고, 8개의 attention head를 사용하며, 임베딩 size는 768로 지정한다. 옵티마이저는 Adam을 사용한다.

GAN에서 영감을 얻은 ELECTRA[12] 모델을 기반으로 하며 한국어 음절 단위로 사전학습된 KoCharELECTRA[13] 모델을 이용하여 문장의 단어표현을 얻는다.

4.2 실험평가

모델의 번역 정확도를 평가하기 위해 BLEU 점수를 이용한다. BLEU 점수는 0과 1사이의 값을 가지며 1에 가까울수록 높은 정확도를 가진 번역기를 의미한다. 평가를 위한 토큰나이즈 단위는 어절 단위(word), 음절 단위(syllable)로 나누어 측정한다.

표 5. 모델의 성능

| Model | Metric | Unit | |
|-----------------------|--------|-------|----------|
| | | Word | Syllable |
| w/o term boundary | BLEU1 | 0.693 | 0.770 |
| | BLEU2 | 0.608 | 0.748 |
| | BLEU3 | 0.514 | 0.727 |
| | BLEU4 | 0.426 | 0.699 |
| with term boundary | BLEU1 | 0.586 | 0.661 |
| | BLEU2 | 0.446 | 0.606 |
| | BLEU3 | 0.341 | 0.562 |
| | BLEU4 | 0.251 | 0.516 |

표 5는 경계 토큰이 추가되지 않은 말뭉치로 학습한 모델(w/o term boundary)과 경계 토큰을 추가한 말뭉치로 학습한 모델(with term boundary)의 BLEU 점수를 보여준다. BLEU 점수 뒤에 있는 숫자는 BLEU 점수의 N-그램을 나타낸다. 원시문장과 목적문장에 동일한 단어가 많기 때문에 다른 두 언어 번역모델에 비해 BLEU 점수가 높다. 어절 단위보다 음절 단위 BLEU 점수가 더 높은 점수를 보여주었다.

표 6은 각 모델 출력의 예이다. 원시문장의 의학용어 “CPR”을 인식하여 그 뜻인 “인공적으로 심장과 폐의 기능을 대신하여 주는 방법”으로 두 모델 모두가 잘 해석한 것을 볼 수 있다.

표 7은 말뭉치를 다르게 학습한 모델의 차이점을 보여준다. 경계 토큰을 추가하지 않은 모델은 “혈압”, “맥박” 의학용어를 인식하지 못하고 원시문장의 단어 그대로를 출력하였지만, 경계 토큰을 추가한 데이터로 학습한 모델은 의학용어를 잘 인식하여 두 의학용어 모두를 “혈관에 걸리는 압력”, “심장의 박동”으로 해석하였음을 알 수 있다. 이러한 차이점에도 경계 토큰을 추가한 모델의 BLEU 점수가 낮은 이유는 정성평가 표를 통해 알 수 있다.

정성평가에 대한 데이터는 평가데이터의 10%인 100문장을 무작위로 추출하여 진행하였다. 평가문장 100문장 중 해석을 잘 수행하여 정답문장과 동일하게 출력한 결

표 6. 두 모델이 동일하게 예측한 예

| | | |
|--------------------|----|---|
| | 정답 | 혹시 인공적으로 심장과 폐의 기능을 대신하여 주는 방법 하셨어요? |
| w/o term boundary | 입력 | 혹시CPR 하셨어요? |
| | 출력 | 혹시 인공적으로 심장과 폐의 기능을 대신하여 주는 방법 하셨어요? |
| with term boundary | 입력 | 혹시<m>CPR</m> 하셨어요? |
| | 출력 | 혹시 <m>인공적으로 심장과 폐의 기능을 대신하여 주는 방법</m> 하셨어요? |

표 7. 두 모델이 서로 다르게 예측한 예

| | | |
|--------------------|----|--|
| | 정답 | 혈관에 걸리는 압력, 심장의 박동도 괜찮아졌네요 |
| w/o term boundary | 입력 | 혈압, 맥박도 괜찮아졌네요 |
| | 출력 | 혈압, 맥박도 괜찮아졌네요?? |
| with term boundary | 입력 | <m>혈압</m>, <m>맥박</m>도 괜찮아졌네요 |
| | 출력 | <m>혈관에 걸리는 압력</m>, <m>심장의 박동</m>도 괜찮아졌네요 |

과, 경계 토큰을 추가한 모델이 31문장, 경계 토큰을 추가하지 않은 모델이 11문장을 정답과 동일하게 출력했다. 정성평가 분류기준은 총 5개로 나뉘며 다음과 같다.

탐지 오류(No traslation) 입력 문장의 의학용어가 탐지되지 않아 해석되지 않고 그대로 출력된 문장을 의미한다.

해석 오류(Wrong terminology) 입력 문장의 의학용어는 탐지했으나 목적문장에 포함된 용어의 뜻풀이가 아닌 다른 용어의 뜻풀이가 출력된 문장을 의미한다.

탐지 및 해석 오류(Wrong sentence) 전혀 다른 용어가 생성되었거나, 정답문장과 다르게 생성한 문장을 의미한다.

문장 미완성 오류(Incomplete sentence) 문장의 종결까지 생성되지 못하고 도중에 중단된 문장을 의미한다.

토큰 생성 오류(Wrong word) [unknown] 토큰이 생성되거나 문맥과 상관없이 의미 없는 음절 또는 특수기호(?!.,)가 반복적으로 생성된 문장을 제외하고 정답과 일치하는 문장을 의미한다.

표 8은 정성평가 결과를 보여준다. 탐지 오류 문장이 5문장인 경계 토큰이 추가되지 않은 모델에 비해 경계 토큰이 추가된 모델은 입력 문장의 의학용어 탐지를 잘 수행한 것을 볼 수 있다. 또한 정답과 동일하게 출력한 경우도 많았다. 경계 토큰을 이용하여 용어의 시작과 끝을

명확하게 구분하였기 때문에 용어의 뜻풀이가 중간에 끊기지 않고 끝까지 출력이 되었다는 특징이 있었다. 하지만 인식한 의학용어의 뜻이 정답이 아닌 다른 용어의 뜻이 출력된 경우가 많았고, 경계 토큰이 추가되지 않은 모델이 의미 없는 음절을 소수 생성하는 것과 다르게 경계 토큰을 추가한 모델은 의미 없이 생성된 음절이 <m> 토큰인 경우 다른 용어의 뜻이 출력되어 전체 BLEU 점수가 낮게 측정되었다. 하지만 정성적 결과로 보았을 때 토큰을 추가하여 학습시킨 모델이 의학용어 해석 능력이 더 좋았음을 알 수 있다.

표 8. 정성평가 결과

| Category | w/o term boundary | with term boundary |
|---------------------|-------------------|--------------------|
| No translation | 5 | - |
| Wrong terminology | 1 | 32 |
| Wrong sentence | 9 | 1 |
| Incomplete sentence | 9 | 6 |
| Wrong word | 66 | 21 |

5. 결론

본 논문에서는 COVID-19로 대중의 의학 분야 관심이 높아져 의학용어로 된 문서를 많이 접하게 되고, 이를 이해하기에 많은 어려움이 있다는 것을 언급한다. 또한 전문용어를 해석하는 연구가 활발히 진행되지 않는 점에 집중하였다. 이에 기계번역을 이용한 의학용어 해석 모델을 제안하였다. 의학용어 해석 모델을 위한 병렬말뭉치를 2개의 방법으로 구축하여 모델을 학습시켰으며, 모델의 성능을 어절 단위, 음절 단위로 비교하였다. 경계 토큰이 없는 말뭉치로 학습한 모델은 어절 단위 0.693, 음절단위 0.770의 성능을 보였으며, 경계 토큰을 추가한 말뭉치로 학습한 모델은 어절 단위 0.586, 음절단위 0.661의 성능을 보였다. BLEU 점수뿐만 아니라 표 7과 같이 의학용어를 더 잘 인식하고 해석하였음을 알 수 있었다.

본 논문이 제안한 의학용어 해석기를 통해 비전문가인 대중들이 의학 문서를 쉽게 읽고 이해할 수 있게 될 것이고, 의학 분야의 접근성이 높아질 것을 기대한다.

참고문헌

- [1] 어수경, 박찬준, 문현석, 임희석, “한국어 인공신경망 기계번역의 서브 워드 분절 연구 및 음절 기반 종성 분리 토큰화 제안”, 한국융합학회논문지, 제12권, 제3호, pp. 1-7, 2021.
- [2] 박찬준, 김경희, 박기남, 임희석, “Coronavirus Disease-19(COVID-19)에 특화된 인공신경망 번역기”, 한국융합학회논문지, 제11권, 제9호, pp. 7-13, 2020.
- [3] P. Charoenpornasawat, V. Sornlenetamvanich, and T. Charoenporn, "Improving translation quality of rule-based machine translation", Proceeding of the 19th International Conference on Computational Linguistics: Machine Translation in Asia. Taipei, 2002.
- [4] P. Koehn, F. J. Och and D. Marcu, "Statistical Phrase-Based Translation", University of Southern California Marina Del Rey Information Sciences Institute, 2003.
- [5] 박찬준, 임희석, “공공 한영 병렬 말뭉치를 이용한 기계번역 성능 향상 연구”, 디지털융복합연구, 제18권, 제6호, pp. 271-277, 2020
- [6] D. Datta, P. E. David, D. Mittal and A. Jain, "Neural machine translation using recurrent neural network", International Journal of Engineering and Advanced Technology, 9(4):1395-1400, 2020.
- [7] T. Hiroki, T. Akihiro, N. Takashi and K. Kideki, "Multimodal Neural Machine Translition Using CNN and Transformer Encoder", Proceeding of COLING and the 20th International Conference on Computational Linguistics and Intelligent Text Processing, 2019.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", Proceeding of Advances in Neural Information Processing System, 2017.
- [9] 오종훈, 최기선, “기계학습에 기반한 생의학분야 전문용어의 자동 인식”, 정보과학회논문지:소프트웨어 및 응용, 제33권, 제8호, pp. 718-729, 2006.
- [10] K. Ishikawa, S. Ando, and A. Okumura, “Hybrid Text Summarization Method based on the TF Method and the LEAD Method”, Proceedings of the 2nd National Institute of Informatics Text Collection Information Retrieval Workshop, pp. 325-330, 2001.
- [11] L. B. Soares, N. Fitz, G. J. Ling and T. Kwiatkowski, "Maching the blanks: Distributional similarity for relation learning", arXiv preprint arXiv:1096.03158, 2019.
- [12] K. Clark, M. T. Luong, Q. V. Le and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", arXiv preprint arXiv:2003.10555, 2020.
- [13] J. Park, "KoELECTRA: Pretrained ELECTRA Model for Korean", 2020. <https://github.com/monologg/KoCharELECTRA>