BERT를 이용한 한국어 특허상담 기계독해

민재옥*, 박진우, 조유정, 이봉건, 황광수, 박소희
한국특허정보원
{okauto, znu808, zzoyou12, bglee, hjordan, ssoe1006}@kipi.or.kr

Korean Machine Reading Comprehension for Patent Consultation using BERT

Jae-Ok Min*, Jin-Woo Park, Yu-Jeong Jo, Bong-Gun Lee,
Kwang-Su Hwang, So-Hee Park
Korea Institute of Patent Information

요 약

기계독해는(Machine reading comprehension) 사용자 질의에 대한 답변이 될 수 있는 내용을 기계가 문서를 이해하여 추론하는 것을 말하며 기계독해를 이용해서 챗봇과 같은 자동상담 서비스에 활용할 수 있다. 최근 자연어처리 분야에서 많은 성능 향상을 보이고 있는 BERT모델을 기계독해 분야에 적용 할 수 있다. 본 논문에서는 특허상담 분야에서 기계독해 task 성능 향상을 위해 특허상담 코퍼스를 사용하여 사전학습(Pre-training)한 BERT모델과 특허상담 기계학습에 적합한 언어처리 기법을 추가하여 성능을 올릴 수 있는 방안을 제안하였고, 본 논문에서 제안한 방법을 사용하여 특허상담 질의에 대한 답변 결정에서 성능이 향상됨을 보였다.

1. 서론

기계독해는(Machine reading comprehension) 사용자 질의에 대한 답변이 될 수 있는 내용을 기계가 문서를 이해하여 추론하는 것을 말하며 기계독해를 이용해서 챗봇과 같은 자동상담 서비스에 활용 할 수 있다. 기계독해에 대한 인공지능 연구를 위해 사용하는 대표적인 학습 데이터로는 영문으로 된 SQuAD(Stanford Question Answering Dataset)[1][1)]가 대표적이며, 한국어 데이터 셋으로는 SQuAD를 벤치마킹한 KorQuAD[2][2)]가 대표적인 표준 데이터 셋이다. 본 논문에서는 특허분야에 대한 기계독해 연구를 위해 KorQuAD를 벤치마킹하여 자체적으로 구축한 특허상담 데이터 셋을 대상으로 실험을 진행한다.

최근 자연어처리 분야에서는 구글에서 공개한 사전학습(Pre-training)된 언어모델인 BERT(Pre-training of Deep Bidirectional Transformers for Language Understanding)[3][3)]모델을 이용해서 높은 성능 향상을 이루었다. BERT모델은 Fine-tuning을 통해 KorQuAD 기계독해 task에서는 좋은 성능을 보였지만 동일한 BERT모델로 특허상담 데이터에 대한 기계독해 task 실험에서는 다소 낮은 성능을 보였다. 본 논문에서는 특허상담 코퍼스를 마련하여 사전학습(Pre-training)을 진행하고 사전학습 된 모델로 특허상담 언어처리 기법을 포함한 Fine-tuning을 통해 최적의 기계학습 모델을 찾아서 BERT basic 모델을 사용한 것과 특허상담 코퍼스를 이용한 BERT모델이 기계독해 task에서 성능이 향상되는지를 실험하고자 한다.

1) https://rajpurkar.github.io/SQuAD-explorer/
2) https://korquad.github.io/
3) https://github.com/google-research/bert

2. 관련 연구

2.1 한국어 질의응답 데이터 셋

KorQuAD[2]는 한국어 기계독해 task를 위한 질의응답 데이터 셋으로 Context-Question-Answer 셋으로 이루어져 있으며 KorQuAD Leaderboard에서 데이터 셋을 이용한 언어모델 평가가 활발하게 진행되고 있다. SQuAD를 벤치마킹하여 표준성이 확보되었으며, Training set 60,407건, Dev set 5,774건으로 한국어 기계독해 task를 위한 다량의 학습데이터를 쉽게 확보할 수 있다. 하지만 KorQuAD의 데이터 셋은 Wikipedia article 문단의 일부 하위 영역으로 구축된 일반상식분야에 대한 내용이므로 특허상담분야에 대한 기계독해 학습 데이터로는 적합하지 않다. 따라서 특허상담분야에 대한 기계독해 task 실험을 위해서 특허상담분야의 데이터 셋 마련이 필요하다.

2.2 BERT기반 기계독해

본 논문에서 실험으로 사용한 모델은 BERT-base Multilingual Cased 모델[3]로 한국어 포함 103 languages에서 Pre-trained BERT Model을 이용하여 Fine-tuning만으로 기계독해 task 실험을 할 수 있다. 자연어처리에서 일반적으로 사용하는 RNN과 달리 Recurrent 하지 않아서 장기 의존성 문제점(Long Term Dependency)을 해결할 수 있고 더 좋은 성능을 낼 수 있는 양방향성을 가진 Transformer[4]의 인코더를 사용한 모델이다. BERT의 학습 방법 중 Masked Language Model는 문장 내 랜덤하게 단어를 마스킹하고 이를 예측하도록 하는 학습방식이고 Next Sentence Prediction 방식은 두 문장에 대해서 두 번째 문장이 코퍼스 내에서 첫 번째 문장의 바로 다음에 오

는지 여부를 예측하는 학습방법이다. 위 2가지 학습방식을 이용해 BERT 마지막의 Transformer layer의 추가만으로 기계독해를 이용한 질의응답 문제를 해결하는데 사용한다. [5]에서 제안한 ETRI에서 공개한 KorBERT모델 이용하여 SRU(simple recurrent unit)와 질의응답에 적합한 기법을 추가한 모델을 통해 기계독해 성능 향상을 이루었고, 긴 문자열의 경우에는 Max sequence length 파라미터에 의해 학습에 사용하지 않는 문자열이 발생하는 문제가 있는데 길이가 512보다 긴 시퀀스 경우 한 번에 처리할 수 있도록 개선한 Multi-level Attention에 Co-Attention과 Fusion 함수를 결합하여 적용한 모델을 제안[6]하여 문제를 해결하고 기계독해 성능 향상을 이루었다. 본 논문에서는 [5], [6] 성능에 준하는 실험을 Baseline으로 하여 특허상담 데이터로 기계독해 실험을 진행한다.

3. 특허상담 학습 데이터

특허분야의 학습 데이터 셋 구축을 위해 공개된 특허상담사례집과 특허상담고객센터의 상담 이력 데이터를 제공받아 단순 질의 및 많은 비중을 차지하는 질의를 기준으로 데이터를 구성하였다. KorQuAD에서 구성한 Context, Question, Answer에 대한 문자 길이를 분석 및 벤치마킹하여 300자 이상 3,000자 미만의 문단으로 정제했으며, 질문은 5자 이상 150자 미만으로, 정답은 5자 이상 50자 미만으로 구성 하였다. 각각의 문단에 대해 최소 10개 이상의 Question-Answer 셋을 생성하였다. 질문의 유형은 낮은 수준의 빈칸 채우기 수준의 유형부터 문단 내에 없는 단어를 활용하고 단서가 부족한 추론을 요구하는 어려운 유형의 질문 등 다양한 질문을 생성하도록 하였다. 최종 193개의 Context, 5,871건의 Question-Answer 셋이 마련되었고 KorQuAD와 마찬가지로 기계독해 실험을 위해 9대1의 비율로 Training set과 Dev set으로 분할하였다.

또한, BERT모델에 사전학습하기 위해 특허상담분야 데이터를 BERT Pre-training[4] 방식처럼 가공하여 코퍼스를 마련하였다.

[표 1.] 특허상담 학습 데이터 셋 예시

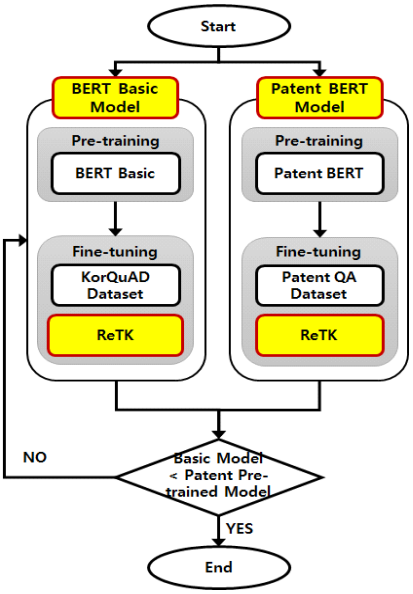
구 분	내 용
Context	신규설정등록이란 ···(중략)··· 등록번호는 납부서를 제출한 날로부터 2일에서 3일 후에 알 수 있으며, 등록증은 약 7일에서 8일 정도 지나면 ···(중략)···
Question	납부서 제출일로부터 언제 정도에 등록증을 받아 볼 수 있나?
Answer	7일에서 8일
Index	156

4) https://github.com/google-research/bert

4. BERT기반 기계독해

본 논문에서는 구글에서 공개한 BERT basic[5] 모델과 특허상담 코퍼스를 사전학습 후 기계학습 최적화를 위한 특허상담 언어처리 기법을 추가로 Fine-tuning한 모델을 나누어 실험한다. [그림 1.]은 본 논문에서 실험하기 위한 BERT모델 기계학습 실험 흐름도이다. 실험 Model은 사전학습(Pre-trained)된 BERT Model에서 KorQuAD 데이터 셋을 Fine-tuning한 BERT basic Model과 BERT basic 모델에서 전이학습(Transfer Learning)을 통해 특허상담 코퍼스를 추가 학습한 Patent BERT Model, 특허상담 언어처리 기법(ReTK : The Optimization Method)로 정의하였다.

[그림 1.] 특허분야 기계학습 흐름도



4.1 특허상담 언어처리 기법

한국어 특허상담 언어처리 기법을 적용하기 전에 KorQuAD 데이터 셋을 사용하여 BERT basic Model로 Fine-tuning을 진행하여 Baseline을 설정하였다.

Encoder Block 12개, Self-attention Head 12개, Hidden-size 768, Dropout 0.1, Activate function은 GELU로 설정하였다. 학습 데이터 구성에 맞는 하이퍼 파라미터는 최대 문장 길이는 128, Context stride는 64, Adam Optimizer를 적용하여 학습을 하면서 최적의 Epoch와 Learning Rate를 얻었다. [표 3.]은 KorQuAD와 특허상담 데이터를 BERT basic 모델로 평가한 Base-line결과이다.

[표 3.] BERT basic으로 성능 평가

Model	평가 데이터	EM (%)	F1 (%)
BERT Basic Model	KorQuAD	82.42	87.43
	Patent	44.03	74.71

5) https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

다음은 특허상담 분야 자질을 추가하였을 때의 성능 비교를 하였다. 특허상담 코퍼스를 구성하는 용어들은 KorQuAD에 등장하는 용어와 상이한 부분이 많기 때문에 본 실험에서 사용한 한국어 형태소 분석기인 Mecab Tokenizer에서 잦은 OOV(Out of Vocabulary)가 발생하는 문제와 기계학습 하는 과정에서 토큰화 된 용어가 Real Answer와 불일치가 되어 학습률을 떨어트리는 문제가 발생한다. 이 문제를 해결하기 위하여 WPM(Word piece Model)[7] 방식으로 특허상담 Vocabulary를 만들어 적용하였고, 불일치 된 Token과 Real Answer을 맞추기 위해 불일치한 Token에 대한 Left word, Right word를 구분하고 대조하여 전체 Token을 재구성 하는 ReTK[6] (Re-Tokenization)기법을 적용하고 Fine-tuning한 결과 KorQuAD 데이터 기준으로 불일치된 Token을 37,125개에서 472건으로 줄었다. [표 4.]는 ReTK기법을 추가하여 Fine-tuning한 결과로 KorQuAD 데이터 셋으로 평가한 결과 EM score은 82.42%에서 84.55%로 2.13%p가 올랐고, F1 score는 87.43%에서 92.90%로 5.47%p가 올랐다. 특허상담 데이터 셋으로 평가한 결과는 EM score는 44.03%에서 63.16%로 19.13%p가 올랐고, F1 score는 74.71%에서 82.27%로 7.56%p가 올랐다.

[표 4.] 특허상담 언어처리 기법 추가에 따른 평가

Model	평가 데이터	EM (%)	F1 (%)
BERT Basic Model + ReTK	KorQuAD	84.55	92.90
	Patent	63.16	82.27

4.2 사전학습(Pre-training)

BERT모델을 사용하여 구글에서 제안[7]한 방식에 따라 특허상담 코퍼스를 만들었다. 특허상담 코퍼스의 양이 적기 때문에 BERT모델 위에 코퍼스를 추가하여 학습하였다. 시퀀스 길이는 128로 하였고, Learning Rate는 3e-5로 하였다. 300,000 step에서 최적의 학습률에 도달하여 Patent BERT Model을 생성하였다. 생성한 모델로 4.1에서 도출한 최적의 하이퍼 파라미터로 특허상담 언어처리 기법을 추가하여 Fine-tuning을 하였다. [표 5.]에서 BERT Basic Model + ReTK의 결과보다 Patent BERT Model + ReTK이 더 성능이 향상된 결과를 보였다.

[표 5.] 사전학습에 따른 성능 평가

Model	Fine-tuning	평가 데이터	EM (%)	F1 (%)
Patent BERT Model + ReTK	KorQuAD	Patent	68.57	89.93
	Patent	Patent	71.43	86.34
	KorQuAD +Patent	Patent	80.00	91.30

6) 특허상담 언어처리 기법(The Optimization Method)
7) https://github.com/google-research/bert

5. 결론 및 향후 연구

본 논문에서는 특허상담 기계독해를 위해 KorQuAD 데이터 셋을 벤치마킹한 특허상담 코퍼스와 학습 데이터를 마련하여, KorQuAD+특허상담 코퍼스 사전학습 및 특허상담 언어처리 기법을 추가한 Fine-tuning 최적화 방법을 제안하였다.

특허상담 언어처리 기법(ReTK)을 추가한 실험에서 특허상담 데이터 셋의 평가 결과가 향상되어 진 것으로 보아 특허분야에 적합한 자질의 추가가 특허상담 평가에서 더 높은 성능 개선의 효과를 보였다.

또한, 특허상담 코퍼스를 사전학습하고 KorQuAD 데이터 셋과 합쳐서 학습하여 특허상담 데이터를 평가한 결과 EM score는 80.00%, F1 score는 91.30%로 성능이 가장 높게 나왔다.

향후 연구로는 특허상담 코퍼스를 더 확보하여 사전학습을 하고 다양한 Transformer 모델 적용 및 특허분야 언어처리 알고리즘 개선을 통해 기존보다 성능 향상을 기대할 수 있는 방안에 대해서 연구 할 예정이다.

Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(2013-2-00131, (엑소브레인-1세부) 휴먼 지식 증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발)을 받아 수행된 연구임.

참고문헌

[1] P. Rajpurkar, et al. Squad: 100,000+ questions for machine comprehension of text. arxiv.org/abs/1606.05250, 2016.

[2] 임승영 외, KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋, 한국정보과학회 학술발표논문집, 539-541, 2018

[3] J. Devlin, et al. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, arxiv.org/abs/1810.04805, 2018

[4] Vaswani et al, Attention Is All You Need, https://arxiv.org/abs/1706.03762, 2017

[5] 이동헌 외, BERT를 이용한 한국어 기계 독해. 한국정보과학회 학술발표논문집, 557-559, 2019

[6] 박광현 외, BERT와 Multi-level Co-Attention Fusion을 이용한 한국어 기계독해. 한국정보과학회 학술발표논문집, 643-645, 2019

[7] Rico Sennrich et al, neural machine translation of rare words with subword units, arxiv.org/abs/1508.07909, 2015