T5-기반 문장임베딩과 템퍼러처 스케일링 기법을 사용한 범위 외 의도 탐지 기법

이명훈°, 송은영, 이현영, 임지희*
KT
{myunghoon.lee, eunyoung.song, lee.hyunyoung, jihui.im}@kt.com

Out-of-Scope Intent Detection Method using T5-based Sentence Embedding and Temperature Scaling

Myunghoon Lee°, Eunyoung Song, Hyunyoung Lee, Jihui Im*
KT

요 약

사용자와 상호작용하는 대화시스템에서 사용자의 의도를 이해하기 위한 의도 분류는 중요한 역할을 한다. 하지만, 실제 대화시스템에서는 범위 내의 의도를 가진 발화 뿐만 아니라 범위 외의 의도를 가진 발화에 대한 인식도 중요하다. 본 논문에서는 기존에 사용되던 인코더 기반의 모델이 아닌 인코더-디코더 구조를 가지는 T5 모델을 활용하여 의도 분류 실험을 진행하였다. 또한, (K+1)-way 의도 탐지 방식이 아닌 K-way의 방식에 템퍼러처 스케일링 기법을 적용하여 범위 외 의도 발화 데이터 구축과 재학습이 필요 없는 확장성 있는 범위 외 의도 탐지 방법을 제안하였다. 범위 내 의도 분류 실험 결과 인코더-디코더 구조의 T5 모델이 인코더 구조의 모델에 비해 높은 성능을 보이며, 흔히 생성 태스크에서 활용되던 모델의 분류 태스크로의 확장 가능성을 확인하였다. 또한, 범위 외 의도 탐지 실험 결과에서는 T5 모델이 인코더 구조의 모델인 RoBERTa 보다 범위 외 탐지 재현율이 14.2%p 이상의 높은 성능을 기록하여 인코더-디코더 구조를 활용한 모델이 인코더 구조를 활용한 모델보다 범위 외 의도 탐지에 강건함을 확인하였다.

주제어: 의도 분류, 인코더-디코더, Out-of-Scope, 템퍼러처 스케일링

1. 서론

오늘날 기가 지니, 구글 홈(Google Home), 아마존 에코(Amazon Echo)와 같은 인공지능 스피커를 활용하여 사용자의 음성을 기반으로 상호작용하는 인공지능(AI) 대화시스템(Dialogue System)에서 사용자 발화의 의도 파악은 사용자의 요구에 적절하게 반응하기 위한 시작 단계라는 점에서 중요하다[1]. 예를 들어, AI 콜센터와 같은 대화시스템에서 고객이 “오늘 날씨 어때?”라는 시스템이 답변할 수 없는 질문을 했을 때, 시스템은 “죄송합니다. 저는 AI 콜센터 상담 관련 응답만 할 수 있습니다.”라는 방식으로 답변해야 한다. 그러나, 일반적인 대화시스템은 학습 데이터를 기반으로 사전에 정의된 범위 내(In-Scope) 의도에서만 판단하기 때문에 사전에 정의된 의도를 벗어난 범위 외(Out-of-Scope) 의도를 탐지하지 못해, 의도와는 관계없는 답변을 하게 된다. 이와 같은 대화시스템의 오류를 방지하고자 기존 연구에서는 범위 내 K개의 의도에 범위 외 의도를 가진 발화 데이터를 추가로 구축하여 K+1개의 의도로 분류 모델의 재학습을 통해 범위 외 의도를 식별하였다[1]. 하지만, 이러한 지도학습 방법은 데이터를 구축하는데 많은 비용과 새로 구축한 데이터로 모델을 재학습해야 한다는 번거로움을 야기한다. 본 논문에서는 범위 외 의도 데이터를 추가적으로 구축 및 재학습을 수행하지 않고 템퍼러처 스케일링(Temperature Scaling) 기법으로 소프트맥스(Softmax)의 확률분포를 조절하여 범위 외 의도 탐지에 간단하고 효율적인 방안을 제시하였다.

딥러닝과 인공지능 기술의 발달로 대용량 말뭉치로부터 습득한 지식을 활용하는 사전 학습 언어모델(Pre-trained Language Model)이 공개되면서 질의응답, 감정 분석, 문장 분류, 의도 분류와 같은 다운스트림 태스크(Downstream Task)에서 높은 성능 향상을 보였다[2]. 특히 인코더-디코더(Encoder-Decoder) 구조를 가지는 트랜스포머[3]의 인코더만을 활용한 BERT[4], RoBERTa[5]와 같은 모델은 분류 태스크에서 우수한 성능을 보여주고 있다. [6]에서는 BERT를 활용하여 범위 외 의도 탐지를 수행하고, [7]에서도 오픈 의도 탐지(Open Intent Detection)를 위해 BERT를 사용한다. 하지만, BERT 모델의 경우 범위 내 분류에서 높은 정확도를 보였지만 범위 외 의도 재현율(Recall)에서는 낮은 성능을 기록하였다[6].

* 교신저자(Corresponding author)

최근 연구는 트랜스포머의 인코더-디코더 구조를 생성이 아닌 분류 태스크에 활용되고 있다[8]. [8]은 문장 분류 태스크에서 BERT보다 우수한 성능을 보여주고 있다. 이에 본 논문에서는 주로 생성 태스크에 사용되던 인코더-디코더 구조의 T5[9] 모델을 활용하여 한국어 문장에 적합한 문장임베딩 기법을 제안하고, [6]에서의 인코더 모델인 BERT의 한계를 극복하고 범위 외 의도 탐지에 강건함을 실험적으로 비교 분석하였다.

2. 관련 연구

대화시스템에서 사용자의 발화 문장은 두 개의 이상의 단어로 이루어져 대화시스템에 입력된다[10]. 신경망 기반의 시스템에서 문장을 하나의 고정된 길이의 벡터로 표현하는 것은 사용자의 의도를 파악하는데 중요한 요소이다[11]. [12]는 자연어 추론 (Natural Language Inference, NLI) 태스크와 텍스트 의미적 유사성 (Semantic Textual Similarity, STS) 태스크에서 BiLSTM을 활용하여 문장을 고정된 길이의 벡터로 표현하였다. 표현한 벡터에서 Max 또는 Mean 풀링하여 추출된 두 벡터로 문장임베딩 실험을 진행하였다. [4]의 연구를 토대로 한 [13]에서는 샴(siamese) 네트워크 구조를 BERT의 문장임베딩에 활용하여 텍스트 의미적 유사성 태스크에 적용하였다. [8]에서는 T5의 인코더만 활용한 Encoder-only 방식과 인코더와 디코더를 모두 활용한 EncDec 방식으로 문장임베딩에 대해 실험적으로 비교 및 분석하였다. 본 논문에서는 [8]의 방식을 한국어에 적용하고 확장하여 T5 기반의 문장임베딩 기법을 제안한다.

대화시스템에서 신경망 기반의 시스템을 현실 서비스에 도입한 경우 학습 데이터와 무관한 범위 외의 입력들을 인식하지 못해 높은 확률값으로 잘못된 예측을 한다. 이와 같은 문제를 해결하기 위해 [14]는 컴퓨터 비전, 자연어 처리 및 음성인식 분야에서 소프트맥스 함수의 확률 분포를 조절하는 방식으로 범위 외 도메인 문제에 적용하였다. [6]은 대화시스템에서 10개 도메인의 150개 의도 클래스를 SVM, FastText, BERT, Google's DialogFlow 등 다양한 분류 모델을 적용하여 대화시스템이 인식할 수 없는 범위 외 발화를 탐지하는 실험을 수행하였다.

3. 범위 외 의도 탐지 기법

본 논문에서는 T5를 활용하여 한국어 문장을 하나의 고정된 길이의 벡터로 표현하는 기법을 3.1절에서 설명하며, 범위 외 의도 데이터를 추가적으로 구축하지 않고 K-way 분류 모델의 재학습 없이 템퍼러처 스케일링을 통한 범위 외 의도 탐지 기법을 3.2절에 설명한다.

3.1 T5를 이용한 문장임베딩

T5를 활용하여 문장을 하나의 고정된 길이의 벡터로 표현하는 다양한 기법을 설명한다. T5를 활용하여 문장이 내포하는 의미를 벡터로 표현하는 전략은 크게 두 가지로 구분될 수 있다. 그림 1에서 볼 수 있듯이 첫번째는 모델 구조 측면에서 인코더-디코더 구조인 T5의 인코더 부분만을 사용한 Encoder-only 방식과 인코더와 디코더 모두 사용한 EncDec 방식으로 구분할 수 있다[8]. 두번째로는 문장이 나타내는 의미를 대표하는 벡터를 표현하기 위해 그림 1과 같이 입력 $X = \{x_1, x_2, x_3, \cdots x_n\}$를 T5 Encoder-only와 T5 EncDec를 통해 문맥을 고려한 출력 벡터 $V = \{v_1, v_2, v_3, \cdots v_n\}$을 활용하는 측면에서 첫번째 입력 x_1에 대응하는 출력 벡터 v_1이 문장을 대표하는 벡터로 표현하는 Start 토큰 임베딩, 마지막 입력 x_n에 상응하는 출력 벡터 v_n이 문장을 대표하는 벡터로 표현하는 End 토큰 임베딩, 문맥을 고려한 출력 벡터 V에 각 벡터의 차원별 평균값을 취하는 Mean 풀링과 최댓값을 취하는 Max 풀링으로 구분하여 T5를 활용하여 범위 내 및 범위 외 발화 의도 분류 실험을 진행하였다. 실험 시 총 두 가지 방식의 모델 구조 측면 및 출력 벡터 V로 표현하는 네 가지 방식을 조합하여 T5를 활용하여 문장을 하나의 고정된 길이로 표현하는 다양한 문장임베딩 기법의 성능을 비교 및 분석하였다.

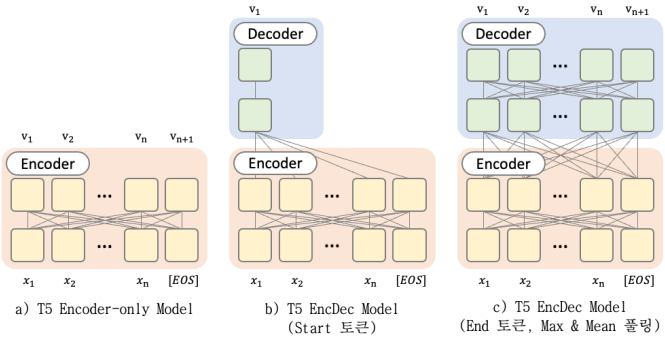

그림 1. T5를 이용한 문장임베딩

T5 Encoder-only 모델과 T5 EncDec 모델의 입력 문장에 토큰화를 수행한 후 스페셜 토큰을 사용하였다. 그림 1은 입력 문장의 단어를 토큰화 한 상태를 나타낸다. T5를 활용한 문장임베딩을 하기 위해 입력 패턴은 T5 Encoder-only 모델과 T5 EncDec 모델의 인코더에 토큰화된 문장에 끝에 스페셜 토큰 [EOS] 토큰을 붙여 입력하고, T5 EncDec 모델의 디코더의 입력 패턴은 토큰화 된 문장 앞에 스페셜 토큰 [PAD]를 붙여 입력하였다[1]. 최종적으로 T5 Encoder-only 모델과 T5 EncDec 모델의 각 출력 벡터를 활용하여 Start 토큰 임베딩, End 토큰 임베딩, Max 풀링, Mean 풀링 기법을 적용하여 의도 분류의 정확도를 비교 분석하였다. T5 EncDec 모델의 Start 토큰 임베딩의 경우에는 디코더의 입력을 스페셜 토큰 [PAD] 토큰으로만 하여 디코더의 [PAD] 토큰 입력에 상응하는 출력 벡터로 미세조정(Fine-tuning)을 진행하여 의도 분류를 수행하였다. T5 EncDec 모델을 이용하여 문장임베딩

[1] https://huggingface.co/docs/transformers/model_doc/t5

경우에는 T5 EncDec의 디코더에 입력 패턴은 토큰화 수행한 입력 문장 앞에 스페셜 토큰 [PAD] 토큰을 덧붙여 입력하여 출력 벡터 End 토큰 임베딩, Max 풀링, Mean 풀링을 수행하여 입력 문장을 대표하는 문장임베딩을 수행하여 의도 분류의 정확도를 측정하였다.

3.2 K-way 의도 분류에서 템퍼러처 스케일링을 이용한 범위 외 의도 탐지

본 절에서는 K-way 의도 분류에서 템퍼러처 스케일링을 이용한 범위 외 의도 탐지 기법을 설명한다. 템퍼러처 스케일링은 수식 1의 소프트맥스 함수에서 템퍼러처 T 값을 사용하여 소프트맥스 함수의 확률 분포를 교정한다. 템퍼러처 T 값을 1 보다 큰 수를 적용하고 확률분포는 점차 고르게 분포하게 된다. 수식 1에서 z_i는 언어모델의 출력인 문장 벡터를 한 층의 완전연결계층(Fully-Connected Layer)의 입력으로 하여 출력한 i번째 클래스에 해당하는 값을 의미한다.

$$P(x) = \frac{exp(z_i/T)}{\sum_{j \in C} exp(z_j/T)} \quad (1)$$

범위 외 의도 탐지를 위해 템퍼러처 값 T를 변경하여 확률분포를 교정하고, 최대 확률값을 가지는 의도 클래스의 확률값 p가 수식 2와 같이 일정 임곗값 τ보다 크거나 같으면 최대 확률값에 해당하는 의도로 분류하고, τ 미만의 확률값을 가지면 범위 외 의도 클래스로 분류를 수행하여 범위 외 의도를 가지는 발화를 탐지하는 실험을 수행하였다[15].

$$g(x;\tau,T,\theta) = \begin{cases} Predicted\ intent & (p \geq \tau) \\ Out\text{-}of\text{-}scope\ intent & (p < \tau) \end{cases} \quad (2)$$

범위 외 의도 탐지를 수행하기 전에 T5 Encoder-only 모델과 T5 EncDec 모델은 K-way 방식으로 범위 내 의도를 가진 발화 데이터만을 사용하여 수식 3의 크로스엔트로피(Cross-entropy)를 손실 함수로 하여 학습을 수행하였다. 그 후, 실험에서 범위 외 의도 데이터에 대한 재학습 없이 템퍼러처 T값과 임곗값 τ을 통해 범위 외 의도 탐지를 수행하였다. 4장에서 K-way 분류 모델에 대한 범위 외 의도 탐지 성능을 비교 분석하였다.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log P(x) \quad (3)$$

4. 실험 및 결과

4.1 데이터 셋과 모델 학습 방법

실험에 사용된 자체적으로 구축한 데이터 셋은 305개의 의도와 범위 외 의도 1개를 추가하였다. 범위 외 분류 실험을 위해 의도를 가지는 전체 191,784개의 발화 데이터를 각 의도별로 8:1:1의 비율로 나누어 학습(Train) 153,304개, 검증(Validation) 19,042개, 평가(Test) 데이터 19,438개에 범위 외 데이터 1,000건을 포함하여 20,438개로 구성하였다. 학습 시 305개의 의도를 가진 학습 데이터로 모델을 학습하였고 검증 데이터로 의도가 있는 경우만의 정확도를 측정하여 성능이 가장 높은 경우 모델을 저장하였다.

학습에 사용된 언어모델은 한국어 말뭉치로 사전 학습한 RoBERTa 모델[16]을 베이스 라인으로 하여 한국어 말뭉치로 사전 학습한 T5 모델을 사용하였다. RoBERTa 모델은 은닉층 크기가 768인 12개의 층인 인코더로 이루어져 있고 T5 모델은 은닉층 크기가 768인 인코더, 디코더 각 12층으로 구성하였다. RoBERTa는 문장의 시작과 끝에 스페셜 토큰 [SOS]와 [EOS]를 붙여 입력으로 하였다.

학습에 사용된 하이퍼 파라미터는 모두 동일하게 설정하였다. 옵티마이저는 AdamW[17]를 사용하였고 학습률은 0.00005로 하며 그 외의 값들은 기본값으로 하였다. 미세조정 시에는 템퍼러처 T 값을 1로 고정하였고 임곗값을 0으로 하여 100번의 학습 횟수로 학습하였다. 추론 시에는 T 값을 0.1부터 10까지, 임곗값을 0에서 0.9까지 실험을 수행하였다.

4.2 실험 결과 및 성능 평가

본 논문에서 제안한 K-way 의도 분류 모델은 K개의 다중 클래스로만 분류할 수밖에 없는 한계를 극복하고자 K+1번째의 범위 외 의도 탐지를 위해 템퍼러처 스케일링 기법을 적용하여 성능 평가를 수행하였다. 여기서, K는 305개의 범위 내 의도 개수를 나타낸다. 더불어, 분류 모델에서 널리 사용되는 인코더 모델인 RoBERTa와 인코더-디코더 구조의 T5 모델을 미세조정하여 범위 내 분류 실험을 수행하였고 K+1번째의 범위 외 의도 탐지를 위해 별도의 학습 없이 템퍼러처 스케일링을 적용하여 성능을 비교 및 분석하였다.

범위 내 의도 분류 성능은 표 1과 같이 범위 내 정확도(In-Scope Acc)로 나타내고, K+1번째의 범위 외 의도를 탐지하기 위해 표 2와 같이 범위 외 재현율(Out-of-Scope Recall), 평가 데이터의 종합 정확도(Total Acc)를 측정하였다.

그 결과, 표 1과 같이 T5 EncDec 모델에 Mean 풀링 기법을 사용한 경우의 정확도가 91.29%로 가장 높은 성능을 달성하였다. 표 1의 실험 결과에서 볼 수 있듯이 트랜스포머의 인코더만을 사용한 모델인 RoBERTa와 T5 Encoder-only 모델 모두 End가 가장 높은 성능을 달성하였고, 비슷한 경향의 범위 내 정확도 성능을 보여주었다. 표 1의 실험 결과로부터 인코더-디코더 구조의 T5 모델을 생성뿐만 아니라 분류 태스크에서 충분한 활용 가능함을 확인하였다.

표 1. K-way 범위 내 의도 분류 성능

Model	Pooling	In-Scope Acc.
RoBERTa	End	91.09
	Start	90.9
	Max	90.84
	Mean	90.79
T5 Encoder-only	End	90.94
	Start	90.8
	Max	90.5
	Mean	90.12
T5 EncDec	Mean	**91.29**
	Start	91.24
	End	90.88
	Max	90.79

표 2는 305개의 범위 내 의도를 가진 발화 데이터로 학습한 K-way 의도 분류 모델에 템퍼러처 T 값과 임곗값을 적용하여 K+1번째의 범위 외 의도를 가진 발화를 탐지한 성능 결과이다. 표 1의 의도 분류 성능 결과를 바탕으로 RoBERTa, T5 Encoder-only, T5 EncDec의 각 모델에서 성능이 가장 높은 풀링 기법을 적용한 모델로 범위 외 의도 탐지 실험을 수행하였다. 표 2의 각 모델에서 첫번째 성능은 템퍼러처 스케일링을 적용하지 않은 경우의 성능이고 두번째 성능은 템퍼러처 스케일링과 임곗값을 적용하였을 때 종합 정확도가 가장 높은 경우의 성능이다. 마지막 성능은 휴리스틱 방식으로 추출한 성능이다.

표 2에서 볼 수 있듯이, T5 EncDec 모델이 템퍼러처 T가 2이고 임곗값이 0.2일 때 가장 높은 종합 정확도를 달성하였다. 이와 비교해서 T5 EncDec 모델이 템퍼러처 T가 3, 임곗값이 0.1일 때 평가 데이터의 종합 정확도가 0.84%p 성능을 하락했지만 범위 외 재현율이 10%p 향상되었다. RoBERTa 모델은 범위 내 정확도에서는 T5 Encoder-only 모델보다는 높은 성능을 나타냈지만 템퍼러처 스케일링을 적용했을 경우 범위 외 의도 재현율이 T5 Encoder-only와 T5 EncDec보다 낮은 성능을 보여주어 평가 데이터의 종합 정확도에서 낮은 성능을 보여주었다. T5 모델은 63.8%의 범위 외 재현율을 달성하여 RoBERTa 모델에 비해 범위 외 의도를 가진 발화를 탐지하는 강건함을 실험적으로 보여주었다.

표 2. 템퍼러처 스케일링 기법을 통한 K-way 범위 내 및 범위 외 의도 탐지 성능

Model	T	Threshold	In-Scope Acc	Out-of-Scope Recall	Total Acc
RoBERTa	1	0	91.09	0	86.63
	2	0.3	90.68	32.6	87.84
	3	0.2	88.95	49.6	87.03
T5 Encoder-only	1	0	90.94	0	86.49
	2	0.2	90.23	47.2	88.13
	3	0.1	88.47	62.1	87.18
T5 EncDec	1	0	**91.29**	0	86.82
	2	0.2	90.51	53.8	**88.71**
	3	0.1	89.11	**63.8**	87.87

5. 결론

대화시스템에서 사용되는 K-way 의도 분류 모델은 사전에 정의된 K개의 의도만 탐지하는 한계로 인해 이를 벗어나는 범위 외 의도를 탐지하지 못하는 문제를 해결하고자, 범위 외 의도 데이터를 새로 구축하고 (K+1)-way 의도 분류 모델로 학습한다. 하지만, 이는 재학습을 위한 데이터를 구축해야 하는 어려움이 존재한다. 본 논문에서는 K-way 의도 분류 모델을 K+1 의도 분류 모델로 재학습을 수행하지 않고, 범위 외 의도를 탐지할 수 있도록 소프트맥스 함수에 템퍼러처 스케일링 기법을 적용하는 간단하면서 효율적인 추론 방식을 제안하였다. 생성모델로 자주 활용되는 인코더-디코더 구조의 T5 모델을 분류 태스크에 적용하여 자연어 이해에서 우수한 성능을 보여주는 RoBERTa 모델보다 우수함을 보여주었다. 또한, RoBERTa보다 인코더-디코더 모델인 T5가 높은 범위 외 재현율을 보여주어 범위 외 의도 탐지에서도 강건함을 확인하였다. 이는 T5 모델의 분류 태스크로 확장 및 활용 가능성을 보여주었다. 본 논문에서 제안한 템퍼러처 스케일링 기법은 실제 서비스에 배치된 언어처리 시스템에 분류 모델에서 고질적인 문제가 되는 범위 외 의도 탐지를 해결하기 위한 모델 고도화 작업을 수행하는데 필요한 데이터 부족 및 구축 비용 문제에 활용될 수 있을 것으로 기대한다. 또한, 다중 도메인 문제에서 각 도메인 별 범위 외 의도 데이터 구축 시에도 템퍼러처 스케일링 기법은 활용될 수 있을 것으로 기대한다.

참고문헌

[1] L. M. Zhan, H. Liang, B. Liu, L. Fan, X. M. Wu, and A. Lam, “Out-of-scope intent detection with self-supervision and discriminative training.” arXiv preprint arXiv:2106.08616, 2021.

[2] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling." arXiv preprint arXiv:1902.10909, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need.” Advances in neural information processing systems, 30, 2017.

[4] 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.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach.” arXiv preprint arXiv:1907.11692, 2019.

[6] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzana, L. Tang and J. Mars, “An evaluation dataset for intent classification and out-of-scope prediction.” arXiv preprint arXiv:1909.02027, 2019.

[7] H. Zhang, H. Xu, and T. E. Lin, “Deep Open Intent

Classification with Adaptive Decision Boundary.” In AAAI (pp. 14374-14382), 2021.
[8] J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, “Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models.” arXiv preprint arXiv:2108.08877, 2021.
[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer.” arXiv preprint arXiv:1910.10683, 2019.
[10] C. Costello, R. Lin, V. Mruthyunjaya, B. Bolla, and C. Jankowski, "Multi-layer ensembling techniques for multilingual intent classification." arXiv preprint arXiv:1806.07914, 2018.
[11] Z. Zhang, W. Zhu, J. Zhang, P. Wang, R. Jin, and T. S. Chung, "PCEE-BERT: Accelerating BERT Inference via Patient and Confident Early Exiting." Findings of the Association for Computational Linguistics: NAACL 2022. 2022.
[12] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data.” arXiv preprint arXiv:1705.02364, 2017.
[13] N. Reimers, and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks.” arXiv preprint arXiv:1908.10084, 2019.
[14] D. Hendrycks, and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks.” arXiv preprint arXiv:1610.02136, 2016.
[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks.” International conference on machine learning. PMLR, pp. 1321-1330, 2017.
[16] 최윤수, 이혜우, 김태형, 장두성, 이영훈, and 나승훈, “RoBERTa 를 이용한 한국어 기계독해.” 정보과학회 컴퓨팅의 실제 논문지, 27(4), 198-203, 2021.
[17] I. Loshchilov, and F. Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101, 2017.