ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 September: 28(3): 45~55 http://dx.doi.org/10.13088/jiis.2022.28.3.045

# 저성능 자원에서 멀티 에이전트 운영을 위한 의도 분류 모델 경량화

**윤용선**
(주) 자이냅스
*(ysyoon@xinapse.ai)*

**강진범**
(주) 자이냅스
*(jb.kang@xinapse.ai)*

최근 자연어 처리 분야에서 대규모 사전학습 언어모델(Large-scale pretrained language model, LPLM)이 발전함에 따라 이를 미세조정(Fine-tuning)한 의도 분류 모델의 성능도 개선되었다. 하지만 실시간 응답을 요하는 대화 시스템에서 대규모 모델을 미세조정하는 방법은 많은 운영 비용을 필요로 한다. 이를 해결하기 위해 본 연구는 저성능 자원에서도 멀티 에이전트 운영이 가능한 의도 분류 모델 경량화 방법을 제안한다. 제안 방법은 경량화된 문장 인코더를 학습하는 과제 독립적(Task-agnostic) 단계와 경량화된 문장 인코더에 어답터(Adapter)를 부착하여 의도 분류 모델을 학습하는 과제 특화적(Task-specific) 단계로 구성된다. 다양한 도메인의 의도 분류 데이터셋으로 진행한 실험을 통해 제안 방법의 효과성을 입증하였다.

**주제어** : 목적 지향 대화 시스템, 의도 분류, 모델 경량화

논문접수일 : 2022년 6월 22일 논문수정일 : 2022년 7월 10일 게재확정일 : 2022년 7월 24일
원고유형 : 학술대회용 Fast Track 교신저자 : 강진범

## 1. 서론

의도 분류(Intent classification)는 목적 지향 대화 시스템(Task-oriented dialog system)의 주요 단계 중 하나로, 사용자 발화의 의도를 예측하는 작업이다(Zhang et al., 2020A). 부정확한 의도 분류는 사용자가 원하지 않는 대화 흐름으로 이어질 수 있기 때문에 정확한 의도 분류 모델을 개발하는 것이 중요하다.

최근 딥러닝을 활용한 자연어 처리 분야는 대규모 사전학습 언어모델(Large-scale pretrained language model, LPLM)을 중심으로 빠르게 발전하고 있다. BERT, RoBERTa, GPT-3와 같은 트랜스포머(Transformer) 기반의 모델들은 대규모 코퍼스를 학습하여 다양한 언어 처리 문제에서 높은 성능을 보였다(Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020). 최근에는 이러한 LPLM을 가짜뉴스 탐지, 추천 시스템, 도메인 특화 분류 등 다양한 산업 분야에서 활용한 연구들이 제안되었다(Jeong, I., & Ahn, H., 2022; Park, H.-y., & Kim, K.-j., 2021; Kim et al., 2022).

의도 분류 분야에서도 많은 연구들이 LPLM을 미세조정(Fine-tuning)하여 문제를 해결하고자 했다. 특히 의도 분류 문제는 데이터가 적고 클래스가 많은 특징(*few-shot, highly-multiclass*) 때문에 좋은 사전학습 모델을 사용하는 것이 중요하다(Wei et al., 2021). Mehri, Eric, and Hakkani-Tur (2020)은 오픈 도메인 대화 데이터를 학습한

BERT를 의도 분류 문제에 적용했고, Zhang et al. (2020A)은 자연어 추론 데이터(Natural language inference)를 학습한 RoBERTa 모델을 사용해 성능을 개선했다.

LPLM을 미세조정하는 방법이 의도 분류 문제에서 좋은 성능을 보이지만, 많은 운영 비용을 필요로 한다는 단점이 있다. 빠른 응답을 요하는 대화 시스템에서 대규모 모델을 사용할 경우 GPU와 같은 고성능 자원을 사용해야 한다. 고성능 자원에 대한 필요성은 다량의 대화 시스템을 동시에 운영하는 멀티 에이전트 상황에서 비용을 더욱 가중시킨다. 미세조정은 모델의 모든 파라미터를 업데이트하기 때문에 운영하는 에이전트 수가 증가할수록 필요한 자원도 비례해서 증가하기 때문이다.

본 연구는 이러한 문제를 해결하고, 저성능 자원에서도 다량의 에이전트를 운영할 수 있는 의도 분류 모델 경량화 방법을 제안한다. 제안 방법은 세 단계로 구성된다: 1) 문장 인코더 학습, 2) 문장 인코더 경량화, 3) 어답터 기반의 의도 분류 모델 학습. 단계 1)과 2)는 의도 분류 학습 전에 한번만 진행하는 과제 독립적(Task-agnostic) 단계이며, 3)은 의도 분류 데이터마다 진행하는 과제 특화적(Task-specific) 단계이다. 제안 방법의 유효성을 검증하기 위해 3개 도메인의 의도 분류 데이터셋으로 실험을 진행했고, LPLM의 21%의 크기만으로 98%의 정확도를 보임을 확인하였다.

## 2. 관련 연구

### 2.1. 의도 분류

목적 지향 대화 시스템은 주어진 도메인에서 사용자가 원하는 업무를 수행하는 것을 돕는 시스템으로 다양한 분야의 산업에서 활용하고 있다(Zhang et al., 2020A). 목적 지향 대화 시스템은 주로 의도 분류, 개체명 인식, 대화 관리, 응답 생성 등의 모듈로 구성된다. 그 중 의도 분류는 사용자 발화가 담고 있는 사용자의 목적을 예측하는 기능을 한다.

의도 분류는 일반적으로 텍스트 분류 문제에 속하지만 다음과 같은 특징이 있다. 1) 입력 텍스트의 길이가 짧다. 의도 분류에서 다루는 텍스트는 30자 내외로 이루어지기 때문에, 모델은 제한된 정보량만으로 문장 간의 의미 차이를 파악해야 한다. 2) 분류해야 할 클래스 수가 많다. 한 개의 대화 시스템에서 분류해야 할 의도의 수는 적게는 수십 개에서 많게는 100개가 넘는다. 성능 비교를 위한 벤치마크 데이터셋들도 이러한 특징을 반영해 60~150개의 클래스로 이루어졌다 (Casanueva et al., 2020; Larson et al., 2019; Liu et al., 2021). 3) 클래스 당 학습할 수 있는 데이터 수가 적다. 클래스가 많은 데이터셋의 학습 데이터 수를 늘리는 일은 많은 비용을 초래하기 때문에 많은 연구들이 데이터가 적은 상황(Few-shot)을 가정해 성능을 평가하고 비교한다 (Zhang et al., 2020A; Zhang et al., 2021).

위와 같은 의도 분류의 특징들로 인해 사전학습된 언어 모델을 미세조정하는 방법들이 제안되었다. Casanueva et al. (2020)은 사전학습된 문장 인코더인 USE와 ConvRT에 다층 퍼셉트론(Multi-layer perceptron)을 쌓은 모델을 제안했다. Mehri, Eric, and Hakkani-Tur (2020)은 BERT를 오픈 도메인 대화 코퍼스(Open-domain dialogue corpus)로 추가 학습한 ConvBERT를 제안했고, ConvBERT를 의도 분류 데이터로 미세조정하여 높은 성능을 보였다. Zhang et al. (2020A)은

RoBERTa를 NLI 데이터셋과 의도 분류 데이터셋으로 대조 학습(Contrastive learning)시킨 모델을 제안했다. Zhang et al. (2021)은 더 나아가 영어 의도 분류 데이터셋을 종합한 뒤 RoBERTa를 대조 학습시켰고, 지도 기반 미세조정 단계를 추가하여 성능을 개선하였다.

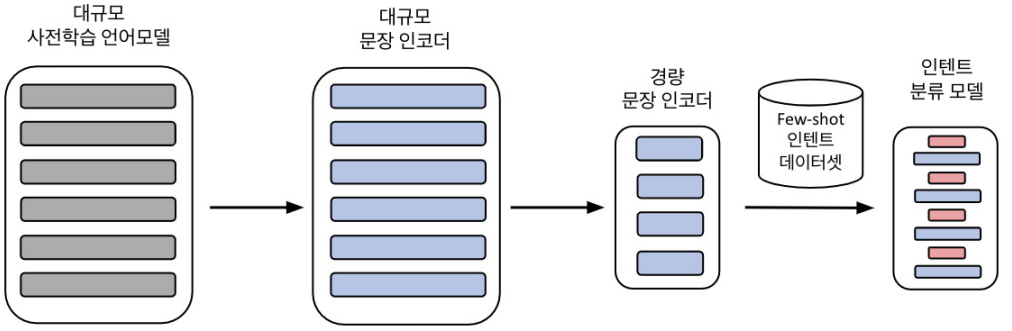
### 2.2. 어답터 (Adapter)

LPLM을 미세조정하는 방법은 다양한 자연어 처리 문제를 해결하는 기본적인 전략이 되었다. 하지만 미세조정은 모델의 모든 파라미터를 업데이트 하기 때문에 두 가지의 단점을 지닌다. 1) 미세조정 과정에서 많은 양의 자원이 필요하다. BERT-base 모델은 약 1억개의 파라미터로 구성되며, 순전파와 역전파 과정에서 모델 크기의 두 배의 메모리를 차지한다. 2) 과제가 증가할수록 학습과 운영에 요구되는 자원이 많아진다. 예를 들어 BERT-base 모델을 사용해 10개의 의도 분류 시스템을 운영할 경우, 미세조정된 모델의 파라미터가 모두 달라지기 때문에 총 10억개의 파라미터에 해당하는 자원을 사용해야 한다.

이러한 미세조정의 비효율성을 해결하기 위해 Houlsby et al. (2019)은 어답터 모듈을 제안했다. 어답터는 모델의 파라미터를 고정한 뒤, 레이어 사이에 작은 어답터 모듈을 추가하여 어답터 모듈만을 업데이트해 과제에 특화된 지식을 학습하는 방법이다. 어답터의 파라미터 수는 전체 모델의 3% 수준이기 때문에 적은 자원으로도 빠르게 학습할 수 있고, 에이전트가 늘어남에 따라 증가하는 파라미터 수가 적다는 장점이 있다.

## 3. 제안 방법

본 연구에서 제안하는 방법의 구조는 <Figure 1>과 같다. 먼저 LPLM을 대조 학습으로 미세조정해 대규모 문장 인코더로 만든다. 다음으로 대규모 문장 인코더를 지식 증류 기법(Knowledge distillation)을 사용해 경량화 한다. 마지막으로 경량화된 문장 인코더에 어답터와 메트릭 학습(Metric learning)을 적용해 의도 분류 모델을 학습한다. 문장 인코더 학습과 경량화 단계는 초기 한 번만 진행하는 과제 독립적 단계이고, 어답터를 적용한 의도 분류 학습은 의도 분류 데이터셋마다 진행하는 과제 특화적 단계이다.




〈Figure 1〉 의도 분류 모델 경량화 구조

### 3.1. 문장 인코더

의도 분류 데이터는 클래스는 많지만 클래스당 학습 가능한 데이터 수가 적다는 특징이 있다. 따라서 적은 데이터로도 문장 간의 의미를 잘 구분할 수 있는 문장 인코더 모델(Sentence encoder)이 좋은 성능을 보인다(Mehri, S., Eric, M., & Hakkani-Tur, D., 2020; Zhang et al., 2020). 문장 인코더는 문장을 정해진 크기의 벡터로 변환하는 모델로, 유사한 의미의 문장을 가깝게, 다른 의미의 문장을 멀리 배치하는 것을 목표로 한다. 본 연구에서는 문장 인코더 학습에서 주로 사용하는 대조 학습을 사용하여 LPLM을 문장 인코더로 만들었다(Gao, T., Yao, X., & Chen, D., 2021).

$$L = -log \frac{e^{sim(h_i, h_i^+)}}{\sum_{j=1}^{N} e^{sim(h_i, h_j^+)}} \quad (1)$$

$$sim(h_i, h_j) = \frac{h_i \cdot h_j}{||h_i|| \, ||h_j||} \cdot \tau \quad (2)$$

$h_i$, $h_j$는 각각 i, j번째 문장의 임베딩 벡터를 의미하며, 문장의 모든 토큰의 은닉 상태(Hidden state)를 평균하여 계산한다. $h_i^+$는 $i$번째 문장과 같은 의미를 지닌 긍정 샘플(Positive sample)의 임베딩 벡터이다.

### 3.2. 문장 인코더 경량화

대규모 문장 인코더는 문장 간의 의미 차이를 잘 구분할 수 있지만, 모델의 파라미터 수가많기 때문에 많은 운영 자원을 필요로 한다. 이를 해결하고자 지식 증류 기법을 사용해 문장 인코더를 경량화했다.

#### 3.2.1. 대조 경량 오차(Contrastive distillation loss)

문장 인코더를 경량화 하기 위해 대조 학습 기법을 지식 증류에 응용한 대조 경량 오차를 사용했다. 이때 학습가능한 가중치 행렬 $W \in \mathbb{R}^{d_s \times d_t}$를 학생 모델의 은닉 상태에 곱해 학생 모델과 교사 모델의 차원 차이를 해소했다.

$$L_{ct} = -log \frac{e^{sim(h_i^T, h_i^{S+}W)}}{\sum_{j=1}^{N} e^{sim(h_i^T, h_j^{S+}W)}} \quad (3)$$

$h_i^T$는 교사 모델이 만든 $i$ 번째 문장의 임베딩 벡터를 의미하며, $h_i^{S+}$는 학생 모델이 만든 $i$ 번째 긍정 샘플의 임베딩 벡터이다.

#### 3.2.2. MiniLMv2 오차

MiniLMv2은 교사 모델의 크기에 관계없이 학생 모델의 크기를 정할 수 있는 트랜스포머 지식 증류 방법이다(Wang et al., 2020). 릴레이션 헤드(Relation head)로 나눈 쿼리, 키, 벨류 벡터의 어텐션 분포 산출한 뒤, 학생 모델과 교사 모델의 어텐션 분포 차이를 줄이면서 지식 증류가 진행된다. 본 연구에서는 더욱 빠른 지식 증류를 위해 MiniLMv2 오차를 추가하였다.

$$L_{mini} = D_{KL}(R_q^T || R_q^S) \quad (4)$$

$$R_q^T = softmax\left(\frac{H_q^T H_q^{T\mathrm{T}}}{\sqrt{d_r^T}}\right) \quad (5)$$

$h_q^T$는 교사 모델의 쿼리 벡터이며, $d^T$ 는 교사

모델의 은닉 차원 수를 뜻한다. 학생 모델의 어텐션인 $R_q^S$도 같은 방법으로 계산된다. 학생 모델과 교사 모델의 어텐션 분포 차이를 계산하기 위해 쿨백-라이블러 발산(Kullback-Leibler divergence) $D_{KL}$를 사용했다.

본 연구에서 문장 인코더 경량화를 위해 사용한 최종 오차는 두 오차를 가중합하여 계산한다.

$$L = \alpha_{ct} L_{ct} + \alpha_{mini} L_{mini} \quad (6)$$

### 3.3. 의도 분류 모델 학습

본 연구는 경량화된 문장 인코더에 에이전트 특화 어답터(Agent-specific adapter)를 적용함으로써 다량의 에이전트를 효율적으로 운영할 수 있는 의도 분류 모델을 제안한다. 어답터는 딥러닝 분류 문제에서 주로 사용되는 교차 엔트로피 오차(Cross entropy loss)를 통해 학습된다. 이때 적은 학습 데이터로 인한 과적합을 방지하기 위해 메트릭 학습을 추가한다. 여러 메트릭 학습 기법 중 학습 정도에 따라 유연한 오차를 적용할 수 있는 Circle 오차를 사용했다(Sun et al., 2020).

$$L_{circle} = log\left[1 + \sum_{i=1}^{K}\sum_{j=1}^{L} \exp\left(\gamma\left(s_n^j - s_p^i + m\right)\right)\right] \quad (7)$$

$s_n^j$는 $j$ 번째 부정 샘플과의 유사도, $s_p^i$는 $i$ 번째 긍정 샘플과의 유사도, m은 마진(Margin), $\gamma$는 학습의 정도를 조절하는 스케일 팩터(Scale factor)을 뜻한다.

에이전트 특화 어답터가 학습하는 최종 오차는 교차 엔트로피 오차와 Circle 오차를 가중합하여 계산한다.

$$L = \alpha_{ce} L_{ce} + \alpha_{circle} L_{circle} \quad (8)$$

## 4. 실험 및 결과

### 4.1. 실험 환경

문장 인코더 학습을 위해 한국어 문장쌍 데이터셋을 구축했다. 문장쌍 데이터는 (질문, 문단), (본문, 요약), (이전 문장, 다음 문장) 등 같은 의미를 공유하는 두 개의 문장들로 구성된다. 쌍이 되는 문장을 긍정 샘플, 같은 배치(Batch)에 속한 다른 문장들을 부정 샘플로 활용했다. 본 연구에서는 LPLM으로 Park et al. (2021)에서 공개한 한국어 RoBERTa-base 모델을 사용했고, 총 5만 스텝 학습했다.

문장 인코더 경량화 단계에서도 문장 인코더 학습에서 사용한 문장쌍 데이터를 사용했다. 경량 모델의 크기는 hidden layer 6, hidden size 384, feed-forward size 1536로 정했으며 대규모 문장 인코더를 교사 모델로 사용하여 경량 모델을 총 10만 스텝 학습하였다.

의도 분류 성능을 검증하기 위해 세 개의 데이터셋을 구축했다. 다양한 도메인에서의 검증을 위해 의료 도메인과 공공기관 도메인의 데이터를 자체 구축했고, Cho et al. (2020)이 공개한 데이터를 변형해 IT 도메인의 데이터셋을 구축했다. 클래스 당 학습 데이터를 2개만 사용함으로써 데이터가 적은 상황에서의 성능을 확인하였다. 더욱 엄밀한 검증을 위해 10번의 실험을 반복해 평균값을 산출했다. 자세한 데이터셋 정보는 <Table 1>과 같다.

〈Table 1〉 데이터셋 정보

| 도메인 | 클래스 수 | 학습 데이터 | 검증 데이터 |
|---|---|---|---|
| 의료 | 149 | 2 | 30 |
| 공공 | 500 | 2 | 10 |
| IT | 998 | 2 | 80 |

〈Table 2〉 의도 분류 정확도

| 모델 | 의료 | 공공 | IT |
|---|---|---|---|
| Large PLM | 94.97 | 98.88 | 96.85 |
| Large PLM + metric | 96.26 | 98.70 | 97.65 |
| Large SE | 95.15 | 98.98 | 97.10 |
| Large SE + metric | **96.53** | **99.06** | **97.86** |
| Distilled PLM | 78.70 | 86.98 | 88.79 |
| Distilled PLM + metric | 83.56 | 91.42 | 89.57 |
| Distilled SE | 91.37 | 97.48 | 92.21 |
| Distilled SE + metric | **94.16** | **98.10** | **94.46** |

〈Table 3〉 미세조정 - 어답터 비교

| 모델 | 의료 | 공공 | IT |
|---|---|---|---|
| Large SE - finetune | 94.36 | 98.92 | 96.79 |
| Large SE - adapter | **96.53** | **99.06** | **97.86** |
| Distilled SE - finetune | 93.07 | 97.88 | 94.02 |
| Distilled SE - adapter | **94.16** | **98.10** | **94.46** |

### 4.2. 의도 분류 정확도

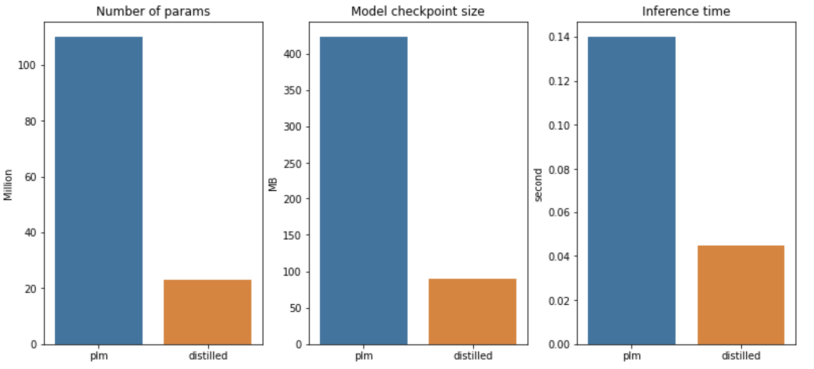
데이터셋에 따른 실험 결과를 <Table 2>에 기재했다. 먼저 문장 인코더 사용은 규모에 관계없이 좋은 효과를 보였다. 특히 경량화된 모델에서 그 차이가 뚜렷한데, 문장 인코더를 경량화할 경우 PLM을 경량화 했을 때보다 정확도가 8.48% 높았다.

의도 분류 모델 학습에 사용한 메트릭 학습 또한 긍정적인 영향을 주었다. PLM - IT의 경우를 제외한 모든 경우에서 메트릭 학습은 더 좋은 성능을 기록했다. 경량화된 문장 인코더의 경우 메트릭 학습을 적용했을 때 2.04% 향상된 정확도를 보였다.

<Table 3>은 미세조정과 어답터에 따른 성능 차이를 기록한 결과다. 대규모 문장 인코더와 경량화된 문장 인코더 모두 어답터가 미세조정보다 더 높은 성능을 보였다. 이는 에이전트 특화 어답터가 효율적인 운영 뿐만 아니라 모델 성능에도 장점이 있다는 사실을 보여준다.

### 4.3. 필요 자원 비교

<Figure 2>는 모델 크기에 따른 필요 자원의 차이를 비교한 결과다. 경량 모델은 대규모 모델에 비해 21% 수준의 파라미터 수와 체크포인트




〈Figure 2〉 모델 크기에 따른 필요 자원

〈Table 4〉 운영 개수에 따른 필요 모델 용량

| | N=1 | N=50 | N=100 |
|---|---|---|---|
| Large - finetune | 423 | 21150 | 42300 |
| Distilled - finetune | 90 | 4500 | 9000 |
| Distilled - adapter | 102 | 690 | 1290 |

용량만을 차지한다. 또한 CPU 환경(Intel(R) Xeon(R) CPU @ 2.20GHz)에서 실험한 결과, 한 문장의 의도를 예측하는데 약 0.04초만을 소요해 대규모 모델 보다 68% 더 빠른 응답속도를 보였다. 이를 통해 제안 방법이 저성능 자원에서도 활용 가능하다는 사실을 알 수 있다.

추가로 <Table 4>에서 운영하는 에이전트 수에 따른 모델의 총 체크포인트를 비교하였다. 100개의 에이전트를 운영할 경우, 경량 모델을 미세조정하는 방법은 9000MB의 체크포인트 용량을 차지한 반면, 경량 모델에 어답터를 적용할 경우 1290MB만을 차지한다. 어답터 기반의 방법이 미세조정 방법보다 다량의 에이전트를 더 효율적으로 운영할 수 있다는 점을 확인할 수 있다.

## 5. 결론

본 연구는 저성능 자원에서 다량의 에이전트 운영을 위한 의도 분류 모델 경량화 방법을 제안하였다. 먼저 대조 학습 기법으로 대규모 사전학습 언어모델을 문장 인코더로 만들었고, 이를 지식 증류 기법을 통해 경량화 하였다. 경량화된 문장 인코더에 어답터와 메트릭 학습을 적용하여 효과적이고 효율적인 의도 분류 모델을 학습했다.

의료, 공공, IT 도메인의 데이터셋으로 실험한 결과 대규모 모델의 21% 크기만으로 98% 수준의 성능을 보임을 확인했다. 또한 어답터를 사용했기 때문에 에이전트 수가 늘어남에 따라 필요한 자원의 양이 상대적으로 적다는 장점을 확인했다. 본 연구의 제안 방법을 통해 더욱 효율적

인 목적 지향 대화 시스템 운영이 가능할 것으로 기대한다.

본 연구에서는 목적 지향 대화 시스템의 구성 요소 중 의도 분류에서만 효과성을 검증했다는 한계점이 있다. 이를 보완하기 위해 후속 연구에서 개체명 인식, 대화 상태 추적 등 목적 지향 대화 시스템을 구성하는 다른 과제들에서 제안 방법의 성능을 검증할 것이다. 또한 본 연구에서 보인 문장 인코더의 효과성을 기반으로 대화 시스템에 특화된 문장 인코더를 연구할 계획이다.

## 참고문헌(References)


Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Cho, W. I., Kim, J. I., Moon, Y. K., & Kim, N. S. (2020, May). Discourse component to sentence (DC2S): An efficient human-aided construction of paraphrase and sentence similarity dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 6819-6826).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Henderson, M., Casanueva, I., Mrkšić, N., Su, P. H., Wen, T. H., & Vulić, I. (2019). ConveRT: Efficient and accurate conversational representations from transformers. arXiv preprint arXiv:1911.03688.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.

Jeong, I., & Ahn, H. (2022). A study on the detection of fake news – The Comparison of detection performance according to the use of social engagement networks. *Journal of Intelligence and Information Systems, 28*(1), 197-216.

Kim, D., Lee, D., Park, J., Oh, S., Kwon, S., Lee, I., & Choi, D. (2022). KB-BERT: Training and Application of Korean Pre-trained Language Model in Financial Domain. *Journal of Intelligence and Information Systems, 28*(2), 191-206.

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., ... & Mars, J. (2019). An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv: 1909.02027*.

Liu, X., Eshghi, A., Swietojanski, P., & Rieser, V. (2021). Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction* (pp. 165-183). Springer, Singapore.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta:

A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mehri, S., Eric, M., & Hakkani-Tur, D. (2020). Dialoglue: A natural language understanding benchmark for task-oriented dialogue. arXiv preprint arXiv:2009.13570.

Park, H.-y., & Kim, K.-j. (2021). Recommender System using BERT Sentiment Analysis. *Journal of Intelligence and Information Systems, 27*(2), 1-15.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., ... & Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., & Wei, Y. (2020). Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition* (pp. 6398-6407).

Wang, W., Bao, H., Huang, S., Dong, L., & Wei, F. (2020). Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*.

Wei, J., Huang, C., Vosoughi, S., Cheng, Y., & Xu, S. (2021). Few-shot text classification with triplet networks, data augmentation, and curriculum learning. *arXiv preprint arXiv: 2103.07552*.

Zhang, J., Bui, T., Yoon, S., Chen, X., Liu, Z., Xia, C., ... & Yu, P. (2021). Few-shot intent detection via contrastive pre-training and fine-tuning. arXiv preprint arXiv:2109.06349.

Zhang, J. G., Hashimoto, K., Liu, W., Wu, C. S., Wan, Y., Yu, P. S., ... & Xiong, C. (2020). Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. *arXiv preprint arXiv:2010.13009*.

Zhang, Z., Takanobu, R., Zhu, Q., Huang, M., & Zhu, X. (2020). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, *63*(10), 2011-2027.

Abstract

# Compressing intent classification model for multi-agent in low-resource devices

Yongsun Yoon* · Jinbeom Kang**

Recently, large-scale language models (LPLM) have been shown state-of-the-art performances in various tasks of natural language processing including intent classification. However, fine-tuning LPLM requires much computational cost for training and inference which is not appropriate for dialog system. In this paper, we propose compressed intent classification model for multi-agent in low-resource like CPU. Our method consists of two stages. First, we trained sentence encoder from LPLM then compressed it through knowledge distillation. Second, we trained agent-specific adapter for intent classification. The results of three intent classification datasets show that our method achieved 98% of the accuracy of LPLM with only 21% size of it.

**Key Words** : Task-oriented dialog system, Intent classification, Model compression

Received : June 22, 2022 Revised : July 10, 2022 Accepted : July 24, 2022
Corresponding Author : Jinbeom Kang

* Xinapse
** Corresponding author: Jinbeom Kang
Xinapse
An Deok Bldg 6F, 557, Nonhyeon-ro, Gangnam-gu, Seoul, Republic of Korea
Tel: +82-2-6052-5611, Fax: +82-2-6280-5612, E-mail: jb.kang@xinapse.ai

# 저 자 소 개

**윤용선**

현재 인공지능 스타트업 자이냅스(Xinapse)에서 연구원으로 재직하고 있다. 주요 연구 분야는 Knowledge Distillation, Intent Classification, Action Recognition 등이다.

**강진범**

현재 인공지능 스타트업 자이냅스(Xinapse)에서 CTO로 재직하고 있다. 한양대학교 컴퓨터공학으로 공학석사와 박사학위를 취득하였다. LG전자 MC사업본부, 신한은행 AI랩에서 인공지능 기술 도입 전략 및 기술을 연구하였다. 주요 연구 분야는 Data Mining, NLP(Natural Language Processing), STT(Speech-To-Text), TTS(Text-To-Speech), Vision 기술들의 융합 등이다.