# 불확실성 정량화 기반 OOD 검출을 통한 대화 의도 분류 모델의 성능 향상

신종훈○, 이요한, 권오욱, 김영길

한국전자통신연구원

{jhshin82, carep, ohwoog, kimyk}@etri.re.kr

# Improving Dialogue Intent Classification Performance with Uncertainty Quantification based OOD Detection

Jong-Hun Shin○, Yohan Lee, Oh-Woog Kwon, Young-Kil Kim

Electronics and Telecommunications Research Institute

**요 약**

지능형 대화 시스템은 줄곧 서비스의 목표와 무관한 사용자 입력을 전달받아, 그 처리 성능을 의심받는다. 특히 종단간 대화 이해 생성 모델이나, 기계학습 기반 대화 이해 모델은 학습 시간대에 한정된 범위의 도메인 입력에만 노출됨으로, 사용자 발화를 자신이 처리 가능한 도메인으로 과신하는 경향이 있다. 본 연구에서는 대화 생성 모델이 처리할 수 없는 입력과 신뢰도가 낮은 생성 결과를 배제하기 위해 불확실성 정량화 기법을 대화 의도 분류 모델에 적용한다. 여러 번의 추론 샘플링이 필요 없는 실용적인 예측 신뢰도 획득 방법과 함께, 평가 시간대와 또다른 도메인으로 구성된 분포 외 입력 데이터를 학습에 노출시키는 것이 분포 외 입력을 구분하는데 도움이 되는지를 실험으로 확인한다.

주제어: 대화 의도 분류, 분포 외 입력 검출, 대화 이해, 불확실성 정량화

## 1. 서론

대화 처리 시스템은 서비스의 목표 대상에 따라 목적 지향 대화 시스템(goal/task-oriented dialog system)과 챗봇(chatbot)과 같은 자유 대화 시스템으로 분류된다. 하지만 사용자는 시스템에 두 시스템의 경계를 넘나드는 발화를 시도하며, 이에 대응하기 위해 대화 이해와 응답에 필요한 도메인을 세세하게 나누어 따로 또 같이 처리하는 범용 지능형 대화 서비스를 구성하게 된다. 이들 도메인은 데이터 수집과 모델링 생명 주기가 제각기 다른 문제가 있어, 효율성을 위해 사용자 발화 입력 처리 파이프라인 상에 도메인 특화된 여러 개의 대화 이해/생성 모듈을 배치하게 된다. 대화 의도 분류 모델은 입력된 사용자 발화에 가장 적합한 응답을 생성하기 위해 적합한 하위 대화 이해/생성 모듈을 선택하거나, 각 하위 모듈 스스로 해당 입력이 자신이 처리할 수 있는지를 구별하기 위해 사용되기도 한다.

다른 기계학습 기반 모델과 마찬가지로, 현대의 딥러닝 기반 대화 의도 분류 모델은 닫힌 집합 가정(closed-set assumption)에 기초한다. 학습 단계에서 관측되지 못한 입력이 들어오는 실제 환경을 위해서는 자신이 처리할 수 있는 영역의 주제에 해당하는지를 구별함과 동시에, 처리할 수 없는 영역의 주제가 입력으로 들어오는지를 구별할 필요가 있다. 더불어 신뢰도가 낮은 예측 결과를 스스로 배제하여 다른 대화 처리 대화 처리 시스템의 처리 신뢰도를 향상할 수 있다.

이처럼 학습 시간대에 관측되지 못하는 입력을 구별하는 문제와 함께 추론 결과의 신뢰도를 다루는 문제는, 종종 불확실성 추론 및 정량화(Uncertainty Estimation or Quantification), 신규성 검출(Novelty detection), 열린 집합 인식(Open Set Recognition) 또는 분포 외(Out-Of-Distribution; 이후 OOD로 표기)데이터 검출 문제로 다루어 지기도 한다[1].

본 논문에서는 불확실성 정량화 기법을 적용, 분포 외 입력과, 신뢰도가 낮은 예측 결과를 모두 배제하는 대화 의도 분류 모델을 제안한다. 성능 평가를 위해 사전학습 언어모델 기반의 입력 시퀀스 분류 모델을 기반으로, Temperature Scaling[2]을 통한 신뢰도 보정(Confidence Calibration)을 통한 예측 결과 배제 방법 및 입력 자질 간의 거리를 고려해 분포 외 입력을 검출할 수 있도록 SNGP(Spectral-Normalized Gaussian Process)[3]를 사용한 추론 방법을 통해 OOD 분류 성능 개선 능력을 확인하고, 도메인 내 발화 의도 분류 성능에 어떤 도움이 되는지를 살펴본다. 또한, 레이블 된 분포 내(In-Distribution; 이후 ID로 표기) 데이터만으로 학습한 모델과, 평가 시점과 다른 도메인으로 구성된 소량의 OOD 데이터를 학습 시간대에 포함한 모델을 비교, OOD 분류 성능 향상을 위해 직접적으로 OOD 데이터를 학습 시간대에 노출시키는 것이 테스트 환경에서의 대화 의도 분류 수행에 효과적인 방법인지를 실험으로 확인한다.

## 2. 관련 연구

기계학습 모델의 예측 결과의 신뢰도 문제는 의료분야

나 자율주행 자동차에서 주로 다루었으나, 모든 열린 집합을 다루는 언어처리 영역에서도 중요한 문제이기도 하다. 예측 불확실성 정량화(Predictive uncertainty quantification)는 도메인 변화와 같이 모델 학습 단계에 노출되지 않은 OOD 데이터가 유발하는 불확실성과 함께, 관측방법과 관련된 입력 데이터 자체 및 모델링 과정에서 발생하는 불확실성을 예측 결과의 신뢰도 값으로 나타낸다. 본 논문에서는 딥러닝 모델을 위한 예측 불확실성 정량화와 관련된 대규모 연구[4]에서 사용된 방법 중 Deep Ensemble, Monte-Carlo Dropout과 같이 명시적인 샘플링 또는 복수의 모델을 통한 추론 등 많은 연산이 요구되어 응답 시간에 악영향을 끼치는 방법을 적용하지 않는다. BERT와 같은 Transformer 기반 사전학습 모델이 다른 신경망 구조의 모델 및 Scratch부터 학습된 모델 대비 OOD에 강건함을 보인 연구[5,6]들을 바탕으로, 단일 추론이 가능하고 적용이 편리한 방법을 사용한다.

### 2.1 Temperature Scaling을 통한 신뢰도 보정

Guo 외[2]의 연구에서 딥러닝 모델의 신뢰도를 보정하는 방법으로, 아래의 수식과 같이 BERT 신경망 모델 출력 logit을 temperature라고 부르는 단일 스칼라 하이퍼파라미터 $\tau$로 나누는 단순한 접근으로 실현한다:

$$p(y|x) = softmax(\frac{logit(x)}{\tau})$$

예측 결과의 변경 없이 신뢰도를 보정하는 방법이며, [4]의 연구에서는 i.i.d 가정을 만족하는 데이터에 한정해 좋은 신뢰도 보정 능력을 보였다.

### 2.2 SNGP를 통한 불확실성 정량화

SNGP[3]는 테스트 시점에서 입력 $x$에 대한 자질을 산출하고, 이를 바탕으로 사후확률 분포의 평균과 분산을 구해, 이를 기반하여 다음의 수식으로 예측 사후확률 분포를 획득한다:

$$p(y|x) = \int_{g(x)\sim N(logit(x),var(x))} sigmoid(g)dg \quad (1)$$

수식 (1)과 같이 가우시안 분포에서 샘플링을 통해 몬테카를로 평균을 취하게 되어 있으나, Mean-field 근사를 통해 수식 (2)와 같이 softmax 가우시안 사후확률 분포를 생성, 한번의 샘플링으로 예측 사후확률 분포를 반환한다:

$$p(y|x) = softmax\left(\frac{logit(x)}{\sqrt{1+\lambda*var(x)}}\right) \quad (2)$$

수식 (2)의 하이퍼파라미터 $\lambda$ 는 $\pi/8$ 또는 $3/\pi^2$를 주로 사용하나, [3]에서는 소량의 분포 내 검증 데이터셋을 사용하여 추정하기도 한다. 앞서 입력에 의해 생성된 분산을 사용해 $logit(x)$의 자릿수를 rescaling 한다는 점에서 temperature scaling과 흡사하다. 또한 이 역시 신뢰도, 즉 불확실성의 정도를 바꿀 수는 있어도 클래스별 순위를 바꾸지 못함을 알 수 있다.

## 3. 모델, 학습, 평가 방법 및 데이터 구성

### 3.1. 모델 및 학습 파라미터 구성

KorBERT모델[7]을 기반으로, [CLS] 임베딩을 출력 레이어로 전달해 산출된 logit에 Softmax를 적용하여 대화 의도 분류 모델을 학습한다. BERT와 같은 사전학습 모델에 SNGP를 적용하고자, 학습 시점에는 gradient의 업데이트 직후 공분산 행렬을 초기화를 수행하도록 변경하였으며, [CLS] 임베딩을 받는 출력 레이어를 SNGP 레이어로 변경하였다. 학습 최적화는 베이스라인 모델과 SNGP 모델 모두 AdamW(eps=1e-8)를 사용하였으며, learning rate는 초기 값을 3e-5로 설정 후, 선형적으로 감소시켰다. 배치 사이즈 32, 최대 8 epoch까지 학습하였다.

### 3.2. 평가 시나리오의 구성

대화 의도 분류 레이블 된 분포 내(ID) 데이터만을 사용하여 학습하는 방법을 시나리오 1로, 평가셋의 OOD 도메인과 무관한 샘플을 사용, OOD 레이블을 부여해 명시적으로 학습하는 OOD 노출(Exposure) 세팅을 시나리오 2로 설정하여 실험을 진행하였다. 실제 환경에서 입력되는 OOD 데이터의 분포를 알 수 없으므로, 평가셋과는 관련이 없는 도메인의 OOD 샘플을 학습에 사용하더라도 OOD 구별 성능에 도움이 되는지를 확인하기 위함이다.

한편, 시나리오 1, 2 공통 사항으로, 종합 성능을 평가하기 위해 신뢰도의 하한선에 해당하는 Threshold를 ID/OOD 가 혼합된 검증 데이터셋으로 결정하고, 이 값을 사용해 평가 데이터셋으로 성능을 확인한다. Threshold 탐색을 위해, 각 모델 별로 Scaling에 영향을 주는 하이퍼파라미터 값을 고정 후 최고의 매크로 평균 F1을 갖는 값을 완전 탐색(exhaustive search)을 통해 결정하였다.

### 3.3. 학습 및 검증, 평가 데이터의 구성

시나리오 1에서는, ID 데이터 2,789 문장을 사용하여 학습하였으며, 시나리오 2를 위해서는 시나리오 1에 사용한 학습데이터에 추가로 8천 문장의 OOD 데이터를 포함한다. 대화 의도 분류를 위한 ID 학습 데이터는 대화 시스템에서 자주 들어오는 FAQ 문장의 응답 생성을 위해 구축된 자체 데이터로, 세부 발화 의도는 162개, 각 발화 의도는 평균 12.3 문장으로 구성되어 있다. 세분화된 발화 의도를 직접 분류하는 대신 세부 발화 의도를 7개의 대분류로 취합하였으며, 해당 모델의 사용 목적 상 음성인식 결과를 사용하며 잘못된 인식 시퀀스를 검출할 수 있도록 음성인식기 출력으로 획득한 비문 샘플을 포함, 총 8개 레이블로 분류하도록 설정하였다. 시나리오 2를 위해 학습에 사용된 OOD 데이터는, 같은 대화 시스템의 파이프라인에 포함된 종단간 페르소나 기반 대화 모델의 학습 말뭉치[8]를 사용하였다.

검증과 평가 데이터는 시나리오 1, 2가 동일하게 사용된다. 검증 및 평가를 위해 대화 의도 분류 레이블 된 데이터의 일부를 분리하여 155 문장을 ID 데이터로 사용

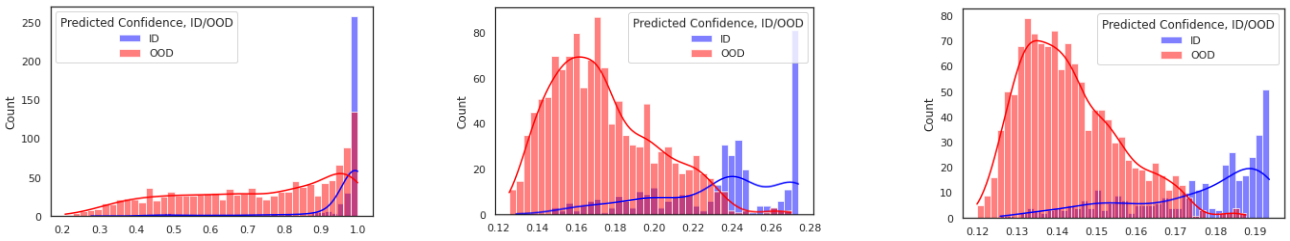


그림 1. 왼쪽부터 각각 베이스라인, Temperature Scaling($\tau = 8.0$), SNGP($\lambda = 3/\pi^2$)를 적용시, ID 클래스 및 OOD 클래스의 Softmax 예측 신뢰도 분포 히스토그램

한다. 또한, 분류 신뢰도의 하한선을 결정하기 위한 Threshold 탐색을 위해, 시나리오 2의 학습에 사용된 OOD 데이터와 다른 도메인을 갖는 말뭉치를 사용하였다. 이를 위해, NIA에서 구축한 AI HUB의 일상생활/구어체 한-영 병렬 말뭉치에서 한국어 문장 5,000개를 추출하였다.

평가 데이터는 대화 의도 분류를 위한 ID 평가셋의 샘플 수를 362 문장(이중 비문 입력 예문의 수는 40문장), OOD 레이블의 샘플 수는 1,239로, 1:3.44 비율로 구성하였다. OOD 샘플은 학습, 검증 데이터셋과 또 다른 도메인으로 분류된 데이터로, 엑소브레인 공개 말뭉치[7] 중 단문질의 900문장과 엑소브레인 한국어 SQuAD 문장을 OOD 데이터 339문장을 포함하였다.

### 3.4. 평가 지표의 구성

제안된 방법론을 적용한 모델의 OOD 검출 성능 평가를 위해 3가지 자동 평가 지표를 사용한다. One-vs-Rest로 OOD-ID 레이블을 구분한 뒤 AUROC, AUPR과 FPR95로 평가를 실시한다. FPR95는 95%의 ID가 양성(True-positive)로 분류되는 지점의 Softmax 확률을 넘어서는 위양성(False-Positive)의 비율로 측정하며, 적을 수록 좋다.

대화 의도 분류 성능 평가의 경우, OOD 데이터를 포함하지 않는 ID-only accuracy로 모델의 기본 분류 성능을 확인하고, ID/OOD가 혼합된 평가셋을 통해 전체 종합 성능을 평가한다. 평가셋의 클래스 불균형 및 실제 환경에서 각 대화 의도 클래스의 출현 빈도가 독립적인 것을 고려하여, 정확도(Accuracy)와 클래스별 샘플 수에 둔감한 매크로 평균(Macro-averaged) F1을 함께 사용하여 최종 성능을 확인한다. Threshold 탐색을 위한 평가 지표 역시 매크로 평균 된 F1을 사용했는데, 마이크로 및 가중 평균 된 F1을 사용하여 탐색할 경우 검증 데이터의 막대한 비중을 차지하는 OOD 분류 성능을 극대화하는 방향으로 움직이게 되기 때문이다.

## 4. 실험 결과 및 논의

테스트 집합을 대상으로 ID와 OOD 데이터의 Softmax 예측 신뢰도의 분포를 시나리오 1의 모델로 생성하여, 이를 히스토그램으로 시각화 하였다. [그림 1]의 각 그래프에서 X축은 신뢰도 점수를, Y축은 해당 신뢰도 점수에 해당하는 샘플의 수를 표기한 것이다. 베이스라인의 결과에서는 ID/OOD가 겹친 영역이 거의 모든 부분에 발생하는 것을 볼 수 있다. 또한, 같은 모델에 Temperature Scaling을 적용하는 것으로, ID/OOD의 분포 영역의 겹침이 완화됨을 확인할 수 있다. SNGP를 사용한 가장 우측의 히스토그램에서 ID 클래스의 좀 더 고른 분포 형태를 관측할 수 있다. 이제 시나리오 별 OOD 검출 성능과 함께 대화 의도 분류 성능을 확인한다.

### 4.1 시나리오 1(도메인 내 데이터만 사용) 평가 결과

Temperature Scaling은 아래 표에서 TS로 표기하였으며, 하이퍼파라미터는 SNGP와 유사한 스케일을 갖도록 8로 설정하였다. SNGP 논문[3]에서 사용된 하이퍼파라미터 $\lambda$를 $3/\pi^2$(~=0.3048)으로 설정한 결과와, 공분산을 스케일링 하지 않도록 1로 설정한 결과를 모두 포함하였다.

표 1. 시나리오 1, OOD 검출 성능 평가 결과

| Model | AUROC | AUPR | FPR95 |
|---|---|---|---|
| MSP(Baseline) | 0.8914 | 0.9537 | 0.3343 |
| TS( τ =8.0) | 0.8965 | 0.959 | 0.3343 |
| SNGP($\lambda$=1.0) | **0.9086** | **0.9631** | **0.2928** |
| SNGP($\lambda=3/\pi^2$) | 0.9083 | 0.9629 | 0.2956 |

ID 데이터만 학습 시간대에 노출시킨 경우, SNGP를 사용한 모델에서 더 나은 OOD 탐지 성능을 보였다. Temperature Scaling은 더 나은 AUROC/AUPR 성능을 보임으로 OOD를 구별하는데 도움이 됨을 보였으나, 변화하지 않는 FPR95 성능에서 알 수 있듯 위양성 오류를 감소시키지 못함을 알 수 있다. 즉, ID 분류 성능의 개선을 위해 Threshold 값을 높이게 되면 결국 OOD 입력을 ID로 분류할 가능성이 함께 올라가게 된다.

### 4.2 시나리오 2(도메인 외 데이터를 포함) 평가 결과

표 2. 시나리오 2, OOD 검출 성능 평가 결과

| Model | AUROC | AUPR | FPR95 |
|---|---|---|---|
| MSP(Baseline) | 0.9105 | 0.9689 | 0.3674 |
| TS( τ =8.0) | 0.9268 | 0.9771 | 0.3646 |
| SNGP($\lambda$=1.0) | **0.9391** | **0.9785** | **0.2514** |
| SNGP($\lambda=3/\pi^2$) | 0.9387 | 0.9783 | 0.2514 |

[표 2]와 같이, OOD를 학습에 포함했을 때, 검증 데이터셋 내 OOD 도메인과 평가셋에 사용된 OOD 샘플이 제각기 다른 도메인을 다루고 있음에도 불구하고, AUROC/AUPR 평가에서 시나리오 1 대비 모든 모델에서 더 나은 OOD 구별 성능을 보였다. 다만 학습 불안정으로 FPR95의 비율도 상승하였다. SNGP 모델에서는 베이스라인과 시나리오 1의 결과보다 더 나은 FPR95 성능을 확인할 수 있어, OOD의 직접 노출이 도움이 된 것으로 확인된다. 특히, OOD를 직접 노출시킬 경우 Temperature Scaling으로도 위양성 오류가 감소(0.3674 vs 0.3646)되는 것을 관측하였다. 이런 결과는 사전학습 모델이 대화 의도 분류에 사용되는 학습 데이터보다 훨씬 다양한 데이터들로 학습되기 때문에, 대화 의도 분류 모델 학습 데이터의 제약에도 불구하고 신뢰도 보정에 도움이 되는 것으로 보인다.

### 4.3. Threshold 적용을 통한 종합 분류 성능 평가

SNGP 모델의 $\lambda$ 값은 $3/\pi^2$ 의 결과를 사용하였으며, $\lambda$=1.0의 실험 결과가 다소 더 높게 관측되었으나, 유의미하지 않은 차이를 보여 평가결과에서 생략하였다.

표 3. Threshold 적용 전/후
대화 의도 분류 성능 평가 결과

| | | ID Only | ID+OOD | ID+OOD(after) | |
|---|---|---|---|---|---|
| 시나리오 | Model | Acc. | Acc. | Acc. | Macro Avg. F1 |
| 1 | MSP | 0.9254 | 0.2092 | 0.8576 | 0.6156 |
| | TS | | | 0.8957 | 0.5416 |
| | SNGP | 0.9254 | 0.2092 | **0.9063** | **0.6698** |
| 2 | MSP | 0.8536 | **0.3529** | 0.8688 | 0.7156 |
| | TS | | | 0.8620 | **0.7210** |
| | SNGP | **0.8734** | 0.2873 | **0.8932** | 0.7200 |

SNGP 모델은 시나리오 1, 2에서 개선된 OOD 구별 성능을 바탕으로 Threshold 적용 후 종합 성능 평가에서도 좋은 성능을 보였다. Temperature scaling은 OOD를 학습에 포함시켰을 때, 베이스란 성능보다 더 좋은 성능을 보였다. 시나리오 1에서 TS가 MSP 대비 낮은 매크로 평균 F1을 보이는 것은, Threshold 적용으로 일부 클래스를 모두 OOD로 분류했기 때문인 것으로 확인하였다.

## 5. 결론

본 논문을 통해 발화 의도 분류 모델을 위해 분포 외 입력(OOD) 구분 성능을 향상하는 방법을 적용한 모델을 제안하고, 각 방법론의 기여를 살펴보았다. 다양한 도메인의 말뭉치로 학습된 Transformer 기반의 사전학습 모델은 그 자체로도 분포 외 입력에 대해 비교적 좋은 성능을 보였으나, 다른 도메인의 말뭉치라도 소량의 OOD 데이터를 학습에 포함하고, 간단히 Temperature Scaling을 적용하는 것으로 개선될 수 있음을 보였다. 또한 SNGP를 사전학습 언어모델에 적용 시 OOD 데이터의 노출 여부와 상관없이 분포 외 입력에 더 강건한 추론 모델을 생성할 수 있음을 확인하였다. 다만, 본 논문에서는 온전히 출력 신뢰도를 중심으로 살펴보았기 때문에, 각 불확실성 속성을 분리하여 분포 외 입력의 구분과 추론 결과의 신뢰도를 구분하지 못한다는 한계를 지니고 있다. 따라서, 향후에는 입력 자질의 밀도를 중심으로 OOD를 탐지하고, ID 입력에 대한 추론 신뢰도의 정확성을 분리한 접근 방법으로 확장될 필요가 있으며, 각 부분에서 요구되는 문제점을 개선해 Active Learning과 같은 학습 문제로의 확장을 고민하고자 한다.

## 감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(2019-0-00004, 준지도 학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)과 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. RS-2022-00187238, 효율적 사전학습이 가능한 한국어 대형 언어모델 사전학습 기술 개발).

## 참고문헌

[1] Salehi et al., “A Unified Survey on Anomaly, Novelty, Open-Set and Out-of-Distribution Detection: Solutions and Future Challenges”, arXiv preprint:2110.14051, 2021.

[2] Guo et al., “On Calibration of Modern Neural Networks”, 34th International Conference on Machine Learning (ICML 2017). 2017.

[3] Liu et al., “Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness”, in Advances in NeurIPS, Vol. 33, pp.7498-7512, 2020.

[4] Ovadia et al., “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift.”, in Advances in NeurIPS, Vol.32, pp.14003-14014, 2019.

[5] Hendrycks et al., "Pretrained Transformers Improve Out-of-Distribution Robustness". In Procs. of the 58th Annual Meeting of the ACL, pp. 2744-2751, 2020.

[6] Fort, Ren and Lakshminarayanan, “Exploring the Limits of Out-of-Distribution Detection”, in Advances in NeurIPS, Vol.34, pp.7068-7081, 2021.

[7] ETRI, KorBERT 모델 및 엑소브레인 공개 말뭉치 데이터. https://aiopen.etri.re.kr/service_dataset.php

[8] 이요한 외, “페르소나 기반 한국어 대화 모델링을 위한 데이터셋”, 제34회 한글 및 한국어 정보처리 학술대회 논문집. 2022.