The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.3, pp.148~155

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.3.148

Korean Ironic Expression Detector

Seung Ju Bang† · Yo-Han Park†† · Jee Eun Kim††† · Kong Joo Lee††††

ABSTRACT

Despite the increasing importance of irony and sarcasm detection in the field of natural language processing, research on the Korean language is relatively scarce compared to other languages. This study aims to experiment with various models for irony detection in Korean text. The study conducted irony detection experiments using KoBERT, a BERT-based model, and ChatGPT. For KoBERT, two methods of additional training on sentiment data were applied (Transfer Learning and MultiTask Learning). Additionally, for ChatGPT, the Few-Shot Learning technique was applied by increasing the number of example sentences entered as prompts. The results of the experiments showed that the Transfer Learning and MultiTask Learning models, which were trained with additional sentiment data, outperformed the baseline model without additional sentiment data. On the other hand, ChatGPT exhibited significantly lower performance compared to KoBERT, and increasing the number of example sentences did not lead to a noticeable improvement in performance. In conclusion, this study suggests that a model based on KoBERT is more suitable for irony detection than ChatGPT, and it highlights the potential contribution of additional training on sentiment data to improve irony detection performance.

Keywords : Irony Detection, KoBERT, ChatGPT, Transfer Learning, MultiTask Learning

한국어 반어 표현 탐지기

방 승 주† · 박 요 한†† · 김 지 은††† · 이 공 주††††

요 약

자연어 처리 분야에서 반어 및 비꼼 탐지의 중요성이 커지고 있음에도 불구하고, 한국어에 관한 연구는 다른 언어들에 비해 상대적으로 많이 부족한 편이다. 본 연구는 한국어 텍스트에서의 반어 탐지를 위해 다양한 모델을 실험하는 것을 목적으로 한다. 본 연구는 BERT기반 모델인 KoBERT와 ChatGPT를 사용하여 반어 탐지 실험을 수행하였다. KoBERT의 경우, 감성 데이터를 추가 학습하는 두 가지 방법(전이 학습, 멀티태스크 학습)을 적용하였다. 또한 ChatGPT의 경우, Few-Shot Learning기법을 적용하여 프롬프트에 입력되는 예시 문장의 개수를 증가시켜 실험하였다. 실험을 수행한 결과, 감성 데이터를 추가학습한 전이 학습 모델과 멀티태스크 학습 모델이 감성 데이터를 추가 학습하지 않은 기본 모델보다 우수한 성능을 보였다. 한편, ChatGPT는 KoBERT에 비해 현저히 낮은 성능을 나타내었으며, 입력 예시 문장의 개수를 증가시켜도 뚜렷한 성능 향상이 이루어지지 않았다. 종합적으로, 본 연구는 KoBERT를 기반으로 한 모델이 ChatGPT보다 반어 탐지에 더 적합하다는 결론을 도출했으며, 감성 데이터의 추가 학습이 반어 탐지 성능 향상에 기여할 수 있는 가능성을 제시하였다.

키워드 : 반어 탐지, KoBERT, ChatGPT, 전이 학습, 멀티태스크 학습

1. 서 론

반어법(反語法)은 발화의 내용과 그 말의 실제 의미 사이에 반대 관계나 모순 관계를 보이는 표현법이다[1]. 반어는 주로 상황에 대한 불만이나 부조리함 또는 유머감각을 드러내기 위해 사용한다. Table 1은 반어에 대한 예시이다.

Table 1의 (1)은 비가 와서 소풍에 가기 어려운 부정적인 상황이지만 긍정적인 표현을 통해 유머 감각을 전달하거나 상황의 부조리함을 지적하는 데 사용되었고 (2)는 시험에서 좋은 성적을 받지 못했으나 그와 반대되는 단어를 반복함으로써 자신의 불만을 드러냈다. 이처럼 반어는 상황에 따라 의미가 반전되는 특성을 지니고 있으며, 반어를 탐지하는 것은 챗봇 기반 서비스와 같은 사용자와의 자연스러운 의사소통이 필요한 작업에서 그 중요성이 부각되고 있다.

최근, 인공지능을 활용한 다양한 서비스의 활성화로 인해 자연어 처리에 대한 관심이 급증하고 있다. 그 중에서도 반어 탐지는 언어의 복잡성과 다의성에 대한 이해를 높이기 위한

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00241142).
† 준 회 원 : 충남대학교 전파정보통신공학과 학사과정
†† 비 회 원 : 충남대학교 전파정보통신공학과 석 · 박사통합과정
††† 비 회 원 : 한국외국어대학교 ELLT학과 교수
†††† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : January 24, 2024
Accepted : February 29, 2024
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

Table 1. Examples of Korean Ironic Sentences

(1) 소풍가려니 비가 오네? 정말 소풍하기 딱 좋은 날씨야.
(2) 또 시험에서 0점을 받았구나? 잘했다 잘했어!

핵심적인 과제로 여겨진다. 서구권에서는 영어를 중심으로 한 반어 탐지 연구[2]가 활발히 이루어진 반면, 한국어에 대한 반어 탐지 연구는 아직 부족한 편이다. 이는 언어의 특성과 문화적 맥락의 차이로 인해 더욱 복잡한 과제로 작용하고 있다.

반어를 감지한다는 것은 겉으로 드러난 표현에서 내재된 의미를 이해하는 과정에 해당한다. 이 내재된 의미는 문맥을 고려한 정보들을 기반으로 하며, 주어진 상황을 파악함으로써 비로소 드러난다. 다시 말해, 반어를 감지함에 있어서 문맥 정보의 적절한 반영이 매우 중요한 역할을 한다. 이러한 문맥 정보를 효과적으로 반영하는 모델로는 대표적으로 ELMo[3], BERT[4], GPT[5]가 있다. 본 연구에서는 이러한 언어 모델들을 활용하여 한국어에서의 반어 표현을 탐지하는 방법을 탐구한다. 더불어, 감성 분석이 반어 탐지 성능에 긍정적인 영향을 미치는지 분석하기 위해 모델을 학습시키는 과정에서 감성 데이터를 활용하여 모델의 성능을 평가한다.

2. 관련 연구

2.1 BERT

BERT는 트랜스포머(Transformer)[6] 모델의 인코더 부분을 사용한 모델이다. BERT는 양방향을 고려한 단어 예측과 문장 수준의 이해를 얻는 것을 목표로 하므로 일반적인 언어 모델과 다르게 언어 모델의 마스킹(Mask Language Model)과 다음 문장 여부를 예측(Next Sentence Prediction)하는 작업을 통해 사전 학습된다. BERT는 대규모 코퍼스에 대해 사전 학습이 되어있기 때문에 미세 조정(fine-tuning)을 활용하면 상대적으로 적은 데이터와 학습시간으로도 좋은 성능을 얻을 수 있다. BERT를 활용한 학습 방법 중에는 전이 학습(Transfer Learning)과 멀티태스크 학습(MultiTask Learning)이 포함되어있다.

전이 학습은 한 도메인에서 학습된 모델을 다른 관련 도메인으로 이전하여 학습시키는 방법이다. 전이 학습은 기존에 학습된 특징이나 패턴을 유지한 채 새로운 태스크에 특화된 특징을 학습하기 때문에 한정된 데이터에 대한 모델 학습에 유용하다. [7]은 기본 BERT를 방글라데시 코퍼스 데이터로 사전 학습한 후, 해당 언어에 대한 정보를 전이시킨 상태에서 방글라데시어로 구성된 텍스트에 대한 이진분류를 수행하였다. 그 결과, 방글라데시어에 대한 전이 학습을 수행한 모델은 mBERT 모델보다 2%p 높은 F-1 Score인 0.94 기록하여 state-of-the-art 수준의 성능을 달성하였다. 또한, [8]은 생물의학 도메인에서의 전이 학습 모델의 효용성을 연구하였다. 해당 연구에서는 생물의학 논문의 초록, 캡션 및 관련 텍스트를 사전 학습한 후 단백질 개체명 인식 태스크의 데이터를 추가 학습하는 방식으로 전이 학습을 진행하였다.

멀티태스크 학습은 하나의 모델이 여러 가지 관련된 태스크들을 동시에 수행하도록 학습하는 방법이다. 주로 하나의 태스크에 대해 훈련되었던 전통적인 기계 학습 방법과 달리, 멀티태스크 학습을 사용하면 훈련 중에 여러 태스크 사이의 정보를 공유함으로써 태스크 전체에 걸쳐 모델을 일반화할 수 있다. [9]는 BERT를 활용한 멀티태스크 학습의 효과를 입증하는 연구를 수행하였다. 해당 연구에서는 인터넷 여성 의류 리뷰에 대한 감정 분류 태스크와 추천 여부에 대한 데이터를 동시에 학습시켰다. 그 결과, 멀티태스크 학습을 수행한 모델은 단일 태스크 학습을 수행한 모델 및 LSTM모델을 기반으로 한 멀티태스크 학습에 비해 평균적으로 5%p ~ 6%p 정도 더 높은 AUC를 기록하여 더 뛰어난 성능을 보여주었다.

2.2 GPT

GPT는 트랜스포머 모델의 디코더 부분을 사용한 생성형 언어 모델이다. GPT와 BERT는 구조적으로 유사하나, 입력 문장의 임베딩에 대해 Masked Multi-Head Attention을 사용하는 차이가 있다. 이는 GPT가 다음 단어를 생성할 때 미래 시점의 단어에 대한 정보를 사용하지 않고, 현재까지 생성된 부분에만 집중하도록 제한하기 위함이다. 이를 통해 GPT는 자기 회귀적(Auto-regressive)인 특성을 유지하면서 단어를 순차적으로 생성할 수 있다. GPT는 레이블이 지정되지 않은 대규모 텍스트 말뭉치에 대한 로그 우도(likelihood)를 최대화하는 방식으로 사전 학습되어있다. GPT도 BERT와 마찬가지로 특정 태스크에 맞게 미세 조정함으로써 원하는 태스크에 적합한 모델로 조정할 수 있다. 그러나 미세 조정을 하는 과정에서 특정 도메인에 대한 충분한 라벨링 데이터를 확보하기 어렵다는 한계도 존재한다.

GPT-3[10]는 2020년에 OpenAI에서 공개한 거대 생성형 언어 모델로, 1750억개의 파라미터와 40TB 크기의 데이터를 기반으로 사전 학습되었다. 앞서 언급한 한계를 지닌 이전의 GPT 모델과는 달리, GPT-3는 파라미터의 미세 조정 없이도 Zero-Shot 및 Few-Shot Learning이 가능하며, 특정 도메인에 대한 추가적인 학습 없이 범용적인 문제에 만족할만한 수준의 성능을 보여준다. 그러나 GPT-3 또한 몇 가지 한계점이 존재하는데, 거짓된 정보를 담은 텍스트 또는 유해한 텍스트를 생성하거나 사용자의 의도에 부합하지 않게 답변하는 것이 대표적인 예이다.

ChatGPT-4는 GPT-4 기반 모델에 RLHF(Reinforcement Learning with Human Feedback) 알고리즘[11]을 적용한 응용 어플리케이션이다. RLHF는 강화 학습과 인간의 피드백을 결합하여 모델의 성능을 향상시키는 방법으로, 사전 훈련된 모델은 라벨러들의 피드백을 통해 보완된다. ChatGPT-4는 이러한 RLHF 기법을 활용하여 사용자에게 더 안전하고 신뢰할 만한 텍스트를 제공한다. ChatGPT-4를 활용한 텍스트 분류의 선행 연구 중, [12]는 은행 업무와 관련된 고객들의 질의로

이루어진 텍스트 데이터에서 ChatGPT-4의 분류 성능을 평가하였다. 해당 연구는 Few-Shot Learning기법을 적용하여 ChatGPT-4가 적은 예시로도 고객의 질문의 의도를 이해하고 분류하는 것에 얼마나 효과적인지를 탐구하였으며, 3-shot 환경에서 F-1 Score 82.7을 기록하였다.

2.3 반어 탐지 (Irony Detection)

반어 탐지 및 분류를 위한 연구는 과거부터 현재까지 지속적으로 이루어졌으며, 감정 분석의 중요성으로 인해 최근 몇 년 동안 해당 분야에 대한 관심이 매우 높아졌다. 텍스트 안에 미묘하게 담긴 비꼼, 유머, 그리고 반어적인 의미를 감지하고 세밀한 차이를 이해하는 것은 자연어 처리의 중요하면서도 도전적인 과제 중 하나로 인정받고 있다. 이러한 특성을 이해하고 모델링 하는 것은 언어의 다양한 측면을 파악하는 데 큰 도움이 된다. 특히, 트랜스포머 아키텍처가 등장함으로써 이러한 작업에 대한 성능은 다른 언어 모델과 비교하여 상당히 향상되었다 [13]. 트랜스포머는 어텐션 매커니즘을 활용하여 문맥을 파악하고 단어 간에 상호작용을 강화함으로써 텍스트의 복잡한 의미를 더 효과적으로 학습할 수 있기 때문에 반어 탐지와 같은 작업에서 모델의 정확성을 향상시키는 데 기여한다.

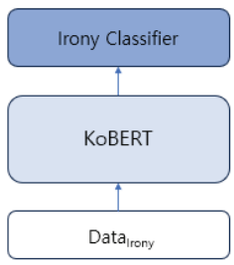
뿐만 아니라, 선행 연구들은 다양한 언어에 대한 반어 탐지에도 중점을 두고 있다. [2]는 BERT 기반 모델을 활용하여 영어 반어 탐지 연구를 수행하였으며, 해당 연구에서는 F-1 Score 0.80에 가까운 성능을 보여주었다. 또한, [14]는 Bi-LSTM 모델을 활용하여 페르시아어에 대한 반어 탐지 연구를 수행하였고, 해당 연구에서는 0.83에 이르는 정확도를 기록하였다. [15]는 BERT에 러시아 위키피디아 코퍼스 등으로 학습된 RuBERT[16]를 활용하여 러시아어 반어 탐지 연구를 수행하였으며, F-1 Score 0.74를 기록하였다.

이렇듯 영어를 넘어 페르시아어, 러시아어와 같은 다양한 언어에 대한 연구들이 진행되어왔으나, 현재까지는 한국어에 대한 연구는 제한적이다. 따라서 본 연구에서는 다양한 실험과 접근 방식을 통해 모델의 성능을 향상시키고 한국어의 언어적 특성을 잘 반영한 반어 탐지 모델을 개발하고자 한다.

3. 실험 모델

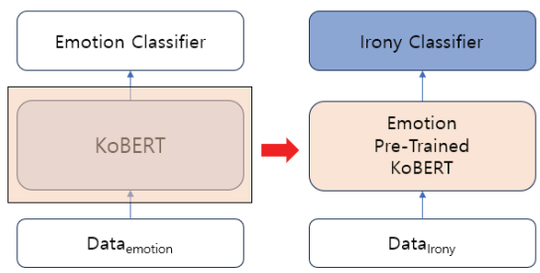
모든 실험은 두세 문장으로 구성된 하나의 문서 단위에 대한 반어 분류를 수행한다. 실험에 사용된 모델은 베이스 모델, 전이 학습 모델, 멀티태스크 학습 모델 그리고 OpenAI의 ChatGPT-4로 총 4가지이고 KoBERT를 활용한 3가지 모델의 분류기는 모두 선형 분류기로 이루어져있다.

Fig. 1은 베이스 모델의 구조를 보여주는 그림이다. 베이스 모델은 기본 KoBERT에 반어 데이터를 학습시킨 모델이며, 다른 실험 결과와의 비교를 위한 기준이 된다. 해당 모델의 입력 단에는 두세 문장으로 구성된 하나의 문서 단위로 입력이 들어온다. 입력으로 들어온 문서는 KoBERT 층에서 각 토큰



Baseline Model

Fig. 1. Baseline Model



Transfer Learning Model

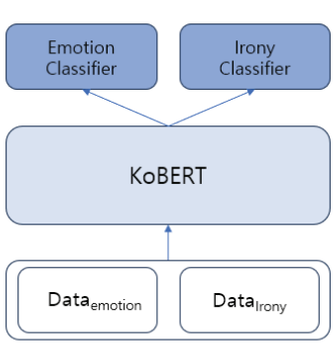
Fig. 2. Transfer Learning Model

에 대한 self-attention을 수행하여 토큰 간의 문맥 정보를 고려한 후 분류기 층에 도달한다. 최종적으로, 분류기 층에서는 입력된 문서가 'Irony' 또는 'Literal' 라벨 중 어느 것에 해당하는지 분류하는 역할을 한다.

또한, 본 연구에서는 감성 데이터를 활용하여 반어 탐지 성능을 개선하기 위해 전이 학습과 멀티태스크 학습을 수행하였다. 다음은 본 연구에서 제시한 전이 학습 모델과 멀티태스크 모델의 구조이다.

Fig. 2는 전이 학습 모델의 구조를 보여주는 그림이다. 전이 학습 모델은 기본 KoBERT에 감성 데이터를 사전 학습시킨 후 기존의 감성 분류기를 제거하고 새로운 분류기인 반어 분류기를 쌓아 반어 데이터로 학습시킨 모델이다. 해당 모델을 통해 감성 데이터에 대한 특징을 유지하면서 반어 데이터를 학습시킨 것이 반어 탐지 성능 개선에 도움이 되는지 확인한다.

Fig. 3은 멀티태스크 학습 모델의 구조를 보여주는 그림이



MultiTask Learning Model

Fig. 3. MultiTask Learning Model

```
You are a helpful assistant that can classify a document into 'Irony' or 'Literal'.
'Irony' often involves saying something but meaning the opposite, usually to highlight
the absurdity or incongruity of a situation.
It frequently employs understatement, exaggeration, or sarcasm, and may be used for
comedic effect or social critique.
I will give you some examples with sentences and answers which is 'Irony' or 'Literal'.

To which category does the text belong based on given examples?
Reply among in 'Irony' or 'Literal' without any reason:

Q: 전달 잘 되자나. 이해가 안가네 코치존 벗어나면 더 잘들리자나.
A: Literal

Q: 먼저 배풀어라! 고객은 똑똑한 장사꾼을 싫어하는 법이다. 먼저 받을려고 계산하는 사람을 누가
좋아하겠는가? 인사도 칭찬도 웃음을 포함한 다른 모든것을 먼저 배풀어라! 분명 더 큰 선물이
기다릴것이다.
A: Irony

.
.
.

Q: input text
A:
```

Fig. 4. Input prompt of Few-Shot Learning

다. 멀티태스크 학습 모델은 기본 KoBERT에 반어 데이터와 감성 데이터를 동시에 입력으로 넣어 학습시킨다. 해당 모델의 마지막 층에서는 두 가지의 분류기를 사용하여 각 데이터 종류에 해당하는 태스크로 분류하도록 설계하였다. 해당 모델을 통해 서로 다른 태스크의 특징을 공유하며 동시에 학습시킨 것이 반어 탐지 성능 개선에 도움이 되는지 확인한다.

멀티태스크 모델은 두 가지 데이터셋을 동시에 입력으로 받아 학습하므로 모델에 사용되는 손실 함수는 각 태스크에 대한 손실 함수의 합으로 나타낼 수 있다. 그러나 본 연구에서 실험에 사용한 데이터셋은 반어 데이터 2,689개와 감성 데이터 74,106개로 그 개수의 차이가 매우 크다. 그렇기 때문에, 단순히 두 데이터셋에 대한 손실 함수를 더하여 모델을 학습시키면 감성 데이터에 지배적인 특징만을 학습하는 문제가 발생할 수 있다. 이 문제를 해결하기 위해 각 데이터셋의 손실 함수에 가중치(weight)를 곱하여 더한 값을 모델의 최종 손실 함수로 사용한다. Equation (1)은 멀티태스크 학습에 사용되는 손실 함수를 나타낸 것이다. $L_{emotion}$과 L_{irony}는 각각 감성 데이터에 대한 손실 함수와 반어 데이터에 대한 손실함수이며, $w_{emotion}$과 w_{irony}는 각각 감성 손실함수의 가중치와 반어 손실함수의 가중치이다. L_{multi}는 최종 손실 함수이다.

$$L_{multi} = w_{emotion} * L_{emotion} + w_{irony} * L_{irony} \quad (1)$$

또한, 모델 학습 시 클래스 개수에 따른 Weighted Sampler를 사용하여 하나의 배치마다 적어도 하나의 반어 데이터가 포함될 확률을 높임으로써 앞서 언급한 문제점을 해결하였다.

마지막으로, 본 연구에서는 반어 탐지를 위해 비교적 최신의 ChatGPT-4 모델도 도입하여 이를 사용한 분류 실험을 해보았다. 실험을 위해 OpenAI의 'gpt-4-1106-preview'를 사용하여 Few-Shot Learning을 수행하였다. 입력 프롬프트의 예시 문장 개수에 따른 성능을 비교하기 위해 120, 200, 500개의 예시 문장을 넣어 실험한 120-shot, 200-shot, 500-shot의 결과를 비교한다. 예시 문장 라벨의 적절한 균형을 위해 모든 Few-Shot 실험에서 각 라벨의 개수를 비슷하게 유지하였다. Fig. 4는 ChatGPT-4를 활용한 실험인 Few-Shot Learning의 입력 프롬프트이다.

4. 실험 데이터

4.1 반어 데이터 수집 및 가공

모든 실험에서 학습, 평가에 사용된 반어 데이터는 [18]의 연구에서 수집한 데이터를 활용하였다. 이 데이터는 네이버 영화 리뷰, 트위터 트윗, 그리고 스포츠 뉴스 댓글의 짧은 텍스트 문서로부터 수집되었다. 해당 문서들은 제한된 글자 수 안에 작성자의 기분을 반영한 메시지를 효과적으로 전달하기 위해 작성된 글들로, 반어적이거나 비꼬는 어조가 강하다는 특징이 있다. 네이버 영화 리뷰 데이터[1]는 네이버[2]에서 수집한 영화 리뷰에 기반한 감성 분류 데이터셋으로, [17]에서 소개한 방법을 활용하여 생성되었다. 트위터에서 수집한 실험 데이터는 2011년 10월의 트윗들을 크롤링한 데이터이다. 스포츠 뉴스 댓글은 올림픽과 같이 많은 독자들의 의견이 오갈 것 같은 대중적인 스포츠 이벤트에 대한 뉴스를 선택하여, 한국의 웹 포털 사이트인 네이버와 다음[3]에서 수집하였다. 수집된 원시 데이터의 총 문서 수는 2,166,934개이며, 네이버 영화 리뷰 712,404개, 트위터의 트윗 1,118,647개 그리고 스포츠 뉴스 댓글 335,913개로 구성되어있다.

수집된 2,166,934개의 원시 데이터에서 형태소 분석 절차를 거쳐 반어적인 표현을 나타내는 40가지 언어적 특징을 통해 자동으로 3,564개의 문서가 추출되었다. 문서에 포함된 각 문장들은 그것이 진짜 반어적이거나 비아냥조를 표현하고 있는지 평가되며, 만약 참이라면 해당 문장이 수사의문문 형태

1) https://github.com/e9t/nsmc/tree/master/raw
2) https://sports.news.naver.com/
3) https://sports.daum.net/

Table 2. Irony Dataset Statistics

Data Source	Num. of Docs	Num. of Irony Docs	Num. of Literal Docs	Avg. Num. of Sents in Docs	Avg. Num. of words in Docs	Avg. Num. of words in Sents
Movie Review	873	412	461	2.66	16.54	6.58
Twitter	880	404	476	3.26	19.29	6.33
Sports News	936	569	367	3.78	25.36	6.98
Total	2,689	1,385	1,304			

Table 3. Redefined Emotion Data Labels

Label	Num. of Label
Negative	33,609
Happy	10,014
Neutral	30,483
Total	74,106

Table 4. Hyper Parameters Used to Train Models

Hyper Parameters	Values
Max Sequence Length	256
Batch Size	16
Warmup Ratio	0.1
Epoch	30
Max Grad Norm	10
Dropout Rate	0.1
Learning Rate	1e-5
Loss Function	CrossEntropy
w_{irony}	1
$w_{emotion}$	0.07

를 갖고 있는지도 평가된다. 이 평가 작업은 한국어를 모국어로 하는 자연어 처리 전공자가 수행하였으며, 크로스 체킹을 통해 결과가 도출되었다. 도출된 결과를 기반으로 여러 문장으로 구성된 문서에서 하나라도 반어적인 문장이 포함된다면 해당 문서는 'Irony'로 라벨링 되었다. 3,564개의 문서 중 1,385개의 문서가 'Irony'로 라벨링 되었으며, 데이터 라벨 개수의 균형을 맞추기 위해 반어가 사용되지 않은 문서 2,179개 중 1,304개의 문서를 추출하여 'Literal' 라벨로 정의한 후, 이를 1,385개의 'Irony' 문서에 추가하였다. 이로써 총 2,689개의 문서가 최종 데이터셋으로 구축되었다 [18]. Table 2는 반어 데이터셋의 통계이다.

2,689개의 문서는 8,734개의 문장으로 구성되어 있고, 1,762개의 반어적 의미를 담고 있는 문장들과 그렇지 않은 6,972개의 문장들로 구성되어있다. 모든 종류의 데이터에서 각 문서는 1개~4개 사이의 문장으로 구성되어 있고, 각 문장은 다시 6개~7개의 단어로 구성되어있다. 이는 수집한 데이터들이 제한된 공간 안에서 작성된 글이기 때문이다. 세 가지의 서로 다른 데이터 종류 중 스포츠 뉴스 댓글이 가장 많은 문장과 단어로 이루어져 있다.

4.2 감성 데이터 수집 및 가공

멀티태스크 학습에 사용되는 감성 학습용 데이터로는 AIHub의 멀티모달 영상 데이터[4]를 활용하였다. 데이터 뭉치들 사이에서 발화 스크립트에 대한 감정 정보(79,980개)만을 추출하였으며, 감정 정보는 8가지의 라벨('화남', '경멸', '혐오', '공포', '행복', '중립', '슬픔', '놀람')로 이루어져 있다. 본 연구에서는 라벨 별 빈도수와 라벨 분류에 대한 단순화를 고려하여 '화남', '경멸', '혐오', '공포', '슬픔'을 '부정(Negative)'이라는 하나의 라벨로 통합하였으며, '부정(Negative)', '행복(Happy)', '중립(Neutral)' 라벨 중 어디에도 속하지 않는 애매한 라벨인 '놀람'을 감성 학습에서 배제하고 실험을 진행하였다. Table 3은 새로 정의된 라벨 별 데이터 개수를 보여주는 표이며, 감성 학습에 사용되는 최종 데이터셋이다.

5. 하이퍼파라미터 및 평가지표

본 연구에서 사용된 데이터는 모두 한국어로 이루어져 있으므로, BERT는 대규모 한국어 코퍼스에 대해 사전 학습된 KoBERT[5]를 사용하였다. 모델 학습에 사용된 하이퍼파라미터들은 Table 4와 같다. 최적화 도구는 AdamW[19]를 사용한다. 마지막으로, 멀티태스크 학습의 손실 함수에 사용되는 하이퍼파라미터인 w_{irony}와 $w_{emotion}$는 각각 1과 0.07이다. 모델 간의 성능 비교를 위해 평가 지표로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F-1 Score를 사용하였다. 본 연구에서는 정확도(Accuracy)와 각 클래스에 대한 F-1 Score의 평균값인 Macro F-1 Score를 평가 지표로 사용한다.

6. 실험 결과

Table 5는 ChatGPT-4 모델에 대한 Few-Shot Learning 실험 결과이다. 각 실험 결과는 입력 프롬프트에 사용되는 예시 문서의 개수에 따라 다르다. 120-shot, 200-shot, 500-shot은 각각 0.64, 0.64, 0.62의 정확도를 보여주었으며, F-1 Score는 각각 0.59, 0.60, 0.61의 성능을 나타내었다. 주어진 예시 문

4) https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=58

5) https://github.com/sktbrain/kobert

Table 5. Results for ChatGPT by the Number of Shots

Model	Setting	Accuracy	Macro F-1 Score
ChatGPT-4	120-shot	0.64	0.59
	200-shot	0.64	0.60
	500-shot	0.62	0.61

Table 6. Results for each Model

Model	Setting	Accuracy	Macro F-1 Score
Baseline	Full Data	0.81	0.81
Transfer Learning	Full Data	0.84	0.84
MultiTask Learning	Full Data	0.84	0.84
ChatGPT-4	500-shot	0.62	0.61

장의 개수가 증가함에 따라 정확도는 살짝 낮아졌고, Macro F-1 Score는 약간의 성능의 향상이 있었지만 극적인 변화는 보이지 않았다. 실험 결과를 통해 ChatGPT-4의 Few-Shot Learning은 주어진 예시 문장 수가 일정 수준에 도달하면 성능이 상대적으로 더 이상 향상되지 않는 경향이 있음을 확인하였다.

한편, Table 6은 모델 별 반어 평가 데이터 분류 결과를 보여주는 표이다. 2,689개의 반어 데이터 중, 2,016개를 학습용으로, 673개를 평가용으로 나누어 사용하였다. 실험에 사용한 3개 모델은 모두 KoBERT를 활용한 실험 결과이며, 학습을 위해 반어 학습용 데이터 전체를 사용하였다. 베이스 모델은 반어 데이터에 대한 기본적인 학습을 수행한 모델로서 정확도와 Macro F-1 Score가 각각 0.81로 나타났다. 전이 학습 모델은 감성 데이터를 먼저 학습한 후 반어 데이터를 추가 학습한 모델로서 정확도와 Macro F-1 Score가 0.84로 측정되었으며, 베이스 모델에 비해 성능이 향상된 모습을 보여주었다. 멀티태스크 학습 모델은 감성 데이터와 반어 데이터를 동시에 학습하였으며, 전이 학습 모델과 동일한 성능인 0.84를 보였다. ChatGPT-4는 반어 훈련 데이터에서 예시 문장 500개를 활용한 Few-Shot Learning의 결과를 함께 비교해 보았다. 실험 결과를 통해 전이 학습 및 멀티태스크 학습을 통한 감성 데이터의 추가 학습이 모델의 반어 탐지 성능 향상에 기여했음을 확인하였다. 또한, ChatGPT-4는 KoBERT 기반 모델들에 비해 현저히 낮은 성능을 보여주어 반어 탐지에 있어서 효과적이지 않은 것으로 나타났다.

Fig. 5는 ChatGPT-4를 활용한 세 가지 Few-Shot Learning 실험 결과를 보여주는 Confusion Matrix이다. 모든 경우에서 ChatGPT-4는 입력된 문장에 대해 'Irony'라벨에 편향되게 예측하는 경향을 보인다. 이 경향은 입력되는 예시 문장의 개수가 많아짐에 따라 약화되었으나, 여전히 한쪽으로 편향되는 경향을 유지하기 때문에 'Literal'라벨에 대한 예측에서 어려움이 있음을 확인하였다.

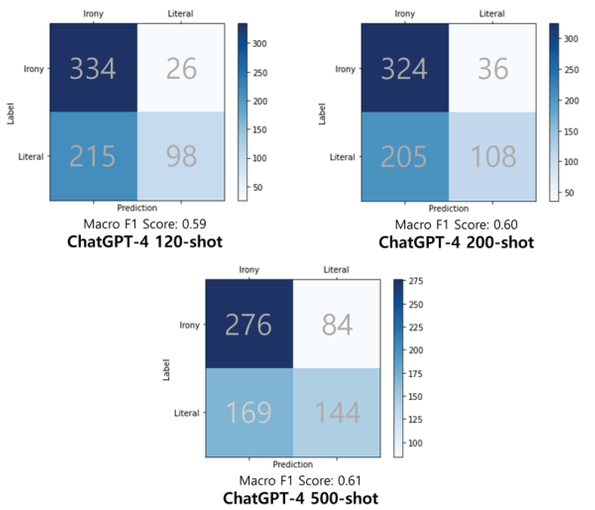


Fig. 5. Confusion Matrix for ChatGPT by the Number of Shots

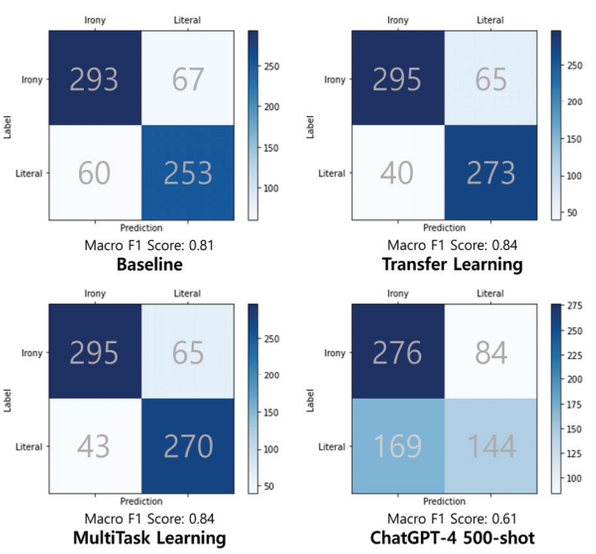


Fig. 6. Confusion Matrix for each Model

Fig. 6은 Table 6에서 제시한 모델들의 실험 결과에 대한 Confusion Matrix이다. 반어 데이터에 대한 기본적인 학습을 수행한 베이스 모델에 비해 감성 데이터로 추가 학습을 수행한 전이 학습 모델은 'Irony'라벨 예측에 대한 성능 향상은 크게 이루어지지 않은 반면, 'Literal'라벨 예측에 대한 성능 향상이 이루어졌다. 감성 데이터와 반어 데이터를 동시에 학습시킨 멀티태스크 학습 모델은 전이 학습 모델과 유사한 성능을 보이고 있으며, 여전히 베이스 모델에 비해 성능이 향상된 모습을 보여주었다.

ChatGPT-4는 KoBERT 기반 모델들과는 달리 'Irony'라벨, 'Literal'라벨 모두에서 예측 성능이 떨어지는 것으로 나타났다. 특히, 'Literal'라벨에 대한 예측에서 169건의 오분류가 발생하여 해당 모델은 'Literal'라벨에 대한 예측에서 어려움이 있음을 확인하였다.

Table 7. Examples of ChatGPT4's failed Prediction

No.	Text	Gold	Baseline	ChatGPT-4
1	마음의 피부는 어찌 가꿔요? .ㅋ 와-ㅅ- 캐비 갔다오니까 피부가 아기피부가 됐음-_-;	Literal	Literal	Irony
2	왜 0점은 줄 수 없나요?	Literal	Literal	Irony
3	인간의 복수심은 어디 까지 일까?	Literal	Literal	Irony
4	사랑과 전쟁인데 왜 신구 할아버지는 안나오나여?	Literal	Literal	Irony

Table 8. Examples Where Baseline Failed While Models with Additional Emotion Learning Succeeded in Predicting

No.	Text	Gold	Baseline	Transfer	MultiTask
1	한국대표팀!! 자신감잃지마시고 다음경기는 무슨 다음경기야 너네 밥도먹지마	Irony	Literal	Irony	Irony
2	체육계도 재앙이네여.. 여기도 재앙 저기도 재앙 ... 잘들 논다. ..	Irony	Literal	Irony	Irony
3	제발 고만들 좀 합시다~ 본인도 얼마나 괴롭고, 두렵겠어요!	Literal	Irony	Literal	Literal
4	진짜 니 경기 보는 내내 돌아버리는줄. 국대를 나오질 말았어야지. 울지마. 나머지 두명이 울고싶겠더만. 넌 울 자격도 없다	Literal	Irony	Literal	Literal

마지막으로, 실험에 사용된 모델들의 결과를 토대로 ChatGPT-4와 베이스 모델의 차이를 예제를 통해 살펴보고, 베이스 모델에 감정을 학습하는 것이 어떤 영향을 미치는지 예제와 함께 알아본다. Table 7은 정답 라벨이 'Literal'인 텍스트에 대해 베이스 모델이 'Literal'이라고 예측하였으나, ChatGPT-4는 'Irony'로 잘못 예측한 예제들에 대한 표이다. ChatGPT-4는 '?'가 포함된 텍스트에 대해서 'Irony'로 예측하는 경향이 나타났다. 이는 데이터셋에 있는 많은 반어 예제들이 문장 끝에 '?'를 포함하고 있고, 이 패턴을 클래스 예측에 사용했기 때문이다. 반면, 정답 라벨이 'Irony'인 텍스트에 대해 베이스 모델은 올바르게 예측하였으나, ChatGPT-4가 잘못 예측한 예제는 뚜렷한 특징이 나타나지 않았다. Table 8은 정답 라벨에 대해 베이스 모델은 잘못 예측하였으나, 감정 학습한 모델들이 올바르게 예측한 예제들에 대한 표이다. 해당 표의 예시들을 살펴보면, 추가적인 감정 학습을 받은 모델이 감정과 직,간접적으로 관련된 단어를 포함한 텍스트에 대해 더 나은 예측을 보인다. 이는 감정 학습을 통해 모델이 각 단어와 문장의 감정적인 의미를 이해하고 중립적인 부분과 감정적인 부분을 구분하며, 일관성 있는 감정 표현을 파악할 수 있게 되었기 때문이다.

7. 결 론

본 연구에서는 한국어 반어 탐지를 위한 다양한 모델들을 탐구하고 실험하였다. 실험 결과를 통해 다음과 같은 결론을 도출할 수 있었다. 첫째로, KoBERT를 활용한 베이스 모델, 전이 학습 모델, 그리고 멀티태스크 학습 모델의 성능을 비교하였다. 감성 데이터에 대한 추가 학습을 진행한 전이 학습 모델과 멀티태스크 학습 모델이 기존의 베이스 모델에 비해 미세하지만 향상된 성능을 보였다. 이러한 결과는 감성 학습이 반어 탐지에 긍정적인 영향을 미쳤음을 의미한다. 둘째로, ChatGPT-4를 활용한 모든 세팅의 Few-Shot Learning 실험 결과는 KoBERT를 활용한 모델의 결과에 비해 현저히 낮은 성능을 기록하였다. 특히, 'Irony'라벨에 편향되게 예측하는 경향이 있었으며, 증가하는 입력 예시 문장의 개수에도 뚜렷한 성능 향상이 나타나지 않았다. 따라서 ChatGPT-4는 반어를 탐지하기에 적합한 모델이 아님을 확인할 수 있다.

또한, 본 연구는 텍스트 분류 및 감정 분석 기술의 발전을 촉진할 것으로 예상된다. 특히, 반어의 식별은 사용자의 의도를 보다 정확하게 이해하여 챗봇과 같은 인공지능 서비스가 적절한 대화를 제공할 수 있도록 돕고, 소셜 미디어 플랫폼에서는 부적절한 콘텐츠를 식별하여 온라인 커뮤니티의 안전성과 쾌적성을 높이는데 기여할 수 있다.

종합적으로, 본 연구는 감성 학습이 반어 탐지 성능을 개선할 수 있는 가능성을 제시하였으며, 다양한 분야에 활용될 수 있음을 보여주었다. 더 나아가, 향후 연구에서는 다양한 모델의 조합이나 추가적인 실험을 통해 반어 탐지의 정확도를 높이는 방안을 모색할 필요가 있다.

References

[1] H. Kil, "How to realize rhetorical irony in Korean," *Studies in Humanities*, Vol.13, pp.1-35, 2005.

[2] C. Turban and U. Kruschwitz, "Tackling irony detection using ensemble classifiers," *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022.

[3] J. Sarzynska-Wawer et al., "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Research*, Vol.304, pp.114135, 2021.

[4] J. Devlin, M. Chang, K. Lee, and K Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*. 2018.

[5] A. Radford and K. Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018.

[6] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, Vol.30, 2017.

[7] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshoiba, "Bangla-BERT: transformer-based efficient model for transfer learning and language understanding," *IEEE Access*, Vol.10, pp.91855-91870, 2022.
[8] A. Arnold, R. Nallapati, and W. W. Cohen, "A comparative study of methods for transductive transfer learning," Seventh IEEE international conference on data mining workshops (ICDMW 2007). IEEE, 2007.
[9] O. Habimana, Y. Li, R. Li, X. Gu, and Y. Peng, "A multi-task learning approach to improve sentiment analysis with explicit recommendation," *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.
[10] T. B. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, Vol.33, pp.1877-1901, 2020.
[11] L. Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, Vol.35, pp.27730-27744, 2022.
[12] L. Loukas, I. Stogiannidis, P. Malakasiotis, and S. Vassos, "Breaking the bank with chatgpt: Few-shot text classification for finance," *arXiv preprint arXiv:2308.14634*, 2023.
[13] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Context-aware sarcasm detection using bert," *Proceedings of the Second Workshop on Figurative Language Processing*, 2020.
[14] P. Golazizian, B. Sabeti, S. A. A. Asli, Z. Majdabadi, O. Momenzadeh, and R. Fahmi, "Irony detection in Persian language: A transfer learning approach using emoji prediction," *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.
[15] M. Kosterin, I. Paramonov, and N. Lagutina, "Automatic Irony and Sarcasm Detection in Russian Sentences: Baseline Methods," *2023 33rd Conference of Open Innovations Association (FRUCT)*. IEEE, 2023.
[16] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language," *arXiv preprint arXiv:1905.07213*, 2019.
[17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.
[18] K. J. Lee, S. Bang, and J. E. Kim, "Korean irony corpus construction," *Language and Information*, Vol.27, No.1, pp.19-36, 2023.
[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

방 승 주
https://orcid.org/0009-0005-5856-2502
e-mail : robin990510@gm.gist.ac.kr
2024년 ~ 현 재 충남대학교 전파정보통신공학과 학사과정
관심분야 : 자연어처리, 기계학습, 인공지능

박 요 한
https://orcid.org/0000-0002-5023-5604
e-mail : happy115012@cnu.ac.kr
2020년 충남대학교 전파정보통신공학과 (학사)
2020년 ~ 현 재 충남대학교 전파정보통신공학과 석 · 박사통합과정
관심분야 : 자연언어처리, 기계학습, 인공지능

김 지 은
https://orcid.org/0000-0002-2886-894X
e-mail : jeeeunk@hufs.ac.kr
1985년 한국외국어대학교 영어과(학사)
1989년 Georgetown University 언어학과(석사)
1993년 Georgetown University 언어학과(박사)
1993년 ~ 1994년 한국외국어대학교 영어학부 강사
1995년 ~ 2002년 한국마이크로소프트(유) 연구원
2003년, 2006년~2007년 한국외국어대학교 영어학부 강사
2008년 ~ 현 재 한국외국어대학교 ELLT학과 교수
관심분야 : 자연언어처리, 전산 언어학, 코퍼스 언어학, 형태론

이 공 주
https://orcid.org/0000-0003-0025-4230
e-mail : kjoolee@cnu.ac.kr
1992년 서강대학교 전자계산학과(학사)
1994년 한국과학기술원 전산학과(석사)
1998년 한국과학기술원 전산학과(박사)
1998년 ~ 2003년 한국마이크로소프트(유) 연구원
2003년 이화여자대학교 컴퓨터학과 대우전임강사
2004년 경인여자대학 전산정보과 전임강사
2005년 ~ 현 재 충남대학교 전파정보통신공학과 교수
관심분야 : 자연언어처리, 기계학습, 인공지능, 정보검색