프롬프트 튜닝기법을 적용한 한국어 속성기반 감정분석

김봉수○, 전현규, 최승호, 김지윤, 장정훈

와이즈넛

{usgnob, eddie14, csh1019, jiyoonkim, jhjang}@wisenut.co.kr

Prompt Tuning For Korean Aspect-Based Sentiment Analysis

Bong-Su Kim○, Hyun-Kyu Jeon, Seung-Ho Choi, Ji-Yoon Kim, Jung-Hoon Jang

Wisenut Inc.

요 약

속성 기반 감정 분석은 텍스트 내에서 감정과 해당 감정이 특정 속성, 예를 들어 제품의 특성이나 서비스의 특징에 어떻게 연결되는지를 분석하는 태스크이다. 본 논문에서는 속성 기반 감정 분석 데이터를 사용한 다중 작업-토큰 레이블링 문제에 프롬프트 튜닝 기법을 적용하기 위한 포괄적인 방법론을 소개한다. 이러한 방법론에는 토큰 레이블링 문제를 시퀀스 레이블링 문제로 일반화하기 위한 감정 표현 영역 검출 파이프라인이 포함된다. 또한 분리된 시퀀스들을 속성과 감정에 대해 분류 하기 위한 템플릿을 선정하고, 데이터셋 특성에 맞는 레이블 워드를 확장하는 방법을 제안함으써 모델의 성능을 최적화한다. 최종적으로, 퓨샷 세팅에서의 속성 기반 감정 분석 태스크에 대한 몇 가지 실험 결과와 분석을 제공한다. 구축된 데이터와 베이스라인 모델은 AIHUB(www.aihub.or.kr)에 공개되어 있다.

주제어: 속성 기반 감정 분석, 프롬프트 튜닝

1. 서론

속성 기반 감정 분석(Aspect Based Sentiment Analysis)은 입력 텍스트 내에서 감정(Sentiment)과 해당 감정이 특정 속성(Aspect), 예를 들어 제품의 특성이나 서비스의 특징에 어떻게 연결되는지를 분석하는 태스크이다. 그리고 입력 텍스트 내에 개체(Entity) 또는 구절의 형태로 존재하는 속성 영역을 찾아내고, 이를 속성-감정 레이블과 함께 분류하는 태스크로 정의된다. 이러한 분석은 고객의 의견이나 선호를 더욱 구체적으로 이해하게 해주며, 제품 개선, 서비스 개선, 마케팅 전략 수립 등에 근거를 제공하는 중요한 역할을 한다.

이러한 중요성에도 불구하고, 속성 기반 감정 분석은 연구하는데 있어 어려운 도전과제를 안고 있다. 주된 어려움 중 하나는 데이터 구성이며, 속성 영역과 함께 세밀하게 연결된 속성-감정이 레이블링된 데이터가 필요하다는 점이다. 또한 모델링 측면에서 여러 토큰 단위의 속성영역에 대해 다양한 속성과 감정 레이블을 동시에 예측해야 하는 다중 작업-토큰 레이블링 문제(Multi Task-Token Labeling)이기 때문에, 단순 분류로서 일반적인 감성분석 태스크보다 더 복잡하다.

이러한 도전 과제에 대해 본 논문에서는 적은 양의 학습 데이터로도 효과적으로 태스크를 수행할 수 있게 해주는 퓨샷러닝(Few-Shot Learning)이라는 패러다임에 주목하며, 사전 훈련된 언어 모델(Pre-trained Language Model)을 구조화된 입력 형식을 사용해 조정하는 프롬프트 튜닝(Prompt-Tuning)기법에 중점을 둔다. 일반적으로 프롬프트 튜닝에서는 레이블 셋과 레이블 워드 셋을 효과적으로 맵핑할 수 있는 버벌라이저(Verbalizer)를 구축하는 것이 중요하다. 또한 적절한 템플릿(Template)선정 과정을 통해 모델이 수행해야 할 태스크의 범위를 정의하고 지시해야 한다.

속성 기반 감정 분석 데이터는 텍스트 내 개별 토큰에 연결된 속성-감정을 반영하기 위해 라벨링된다. 프롬프트 튜닝의 범위를 이러한 토큰 라벨링 문제에 확장, 적용할 방법을 고안할 필요가 있다. 이때 모델에 대한 버벌라이저-레이블 워드의 적합성은 데이터의 특성에 따라 크게 달라질 수 있고, 템플릿은 속성분류와 감정분류을 동시에 처리할 수 있을 만큼 유연해야 하면서도 각각에 필요한 특수성을 유지해야 한다.

본 논문에서는 속성 기반 데이터를 사용한 다중 작업-토큰 레이블링 문제에 몇 가지 학습에 대한 포괄적인 방법론을 소개한다. 이러한 방법론에는 토큰 레이블링 문제에 프롬프트 튜닝을 적용하기 위한 속성 영역 검출 파이프라인이 포함된다. 또한 검출된 시퀀스의 속성-감정에 대해 다중 작업 기반 분류를 하기 위한 템플릿을 제안한다. 또한 데이터 특성에 맞게 레이블 워드를 확장, 추출하는 방법을 제안함으써 모델의 성능을 최적화한다.

이와 같이 여러 전략을 제시하고 결과에 대한 자세한 분석을 제공하기 위해 다양한 실험 설정을 통해, 여러 종류의 버벌라이저와 템플릿을 조사한다. 최종적으로, 우리는 퓨샷 세팅에서의 속성 기반 감정 분석에 대한 몇 가지 벤치마크를 공개한다. 본 논문의 기여는 아래와 같다.

1) 토큰 레이블링에 퓨샷러닝 적용을 위한 파이프라인 제안
2) 다중 작업을 위한 Templete 세팅 선정 과정 제시
3) 데이터 특성에 맞는 Verbalizer 확장 방법, 정제 방법 제안
4) 후속연구를 위한 다양한 실험결과 및 벤치마크 성능 제시

2. 관련 연구

본 절에서는 속성기반 감정분석 데이터셋에 대해 벤치마크 성능을 제시했던 기존 연구를 살펴보고, 프롬프트 튜닝을 적용하기 위해 연구했던 기존 사례를 살펴본다

먼저 국내의 경우 속성 기반 감정 분석 연구가 활발히 수행되고 있지는 않다. 먼저 데이터셋의 경우 [1]의 연구에서 17582개의 감정 표현을 레이블링하고, SVM 모델을 사용해 모델 적합성 검증을 수행한 바 있다. [2-3]은 BERT 기반의 사전학습 모델을 사용하여, 텍스트에 존재하는 감정 표현 영역을 추출하는 태스크를 수행하였다. 이와 같은 연구는 감정 표현 영역 추출(AE : Aspect term Extraction)을 후에 감성 분류(AS : Aspect-level Sentiment classification)를 수행하는데 활용 될 수 있다.

해외의 경우도 이러한 2단계 방법론이 전통적으로 연구되어 왔으며, 특히 [4-5]는 AE단계에서 입력 텍스트에 대해 의존 구문 분석을 수행하고, 감정 표현과 그 수식하는 대상을 분석하여, 구문론적으로 예측 가능한 감정 표현 영역을 검출하였다.

[6-8]와 같이 다른 한편에서는 다중 작업 기반의 종단간 학습(E2E : End to End Learning) 기반의 방법 또한 활발히 연구되고 있었으며, 비교적 최근의 연구인 [9]는 사전학습 모델을 사용하여, 주목할만한 벤치마크 성능을 거두었다.

한편 데이터의 한계를 해결하고 파인튜닝의 효율성을 향상하기 위해 프롬프트 기반 방법이 제안되고 있다. 이러한 방법은 파인튜닝 단계의 목표를 자연스러운 인풋 형태인 Close-Style 작업으로 재구성함으로써, 사전 학습 훈련 태스크와 파인튜닝 태스크와의 간격을 줄인다.

대표적으로 GPT-3 [10]는 이러한 방식의 퓨샷 세팅을 활용하여 다양한 작업에서 인상적인 성능을 보였다.

이를 필두로 많은 연구들 [11-12]등의 후속연구로 이어져, 사전학습 모델을 효율적으로 활용하기 위한 프롬프트 탐구에 큰 진전이 있었다. [13]에서는 프롬프트 기반 방법을 NLG 태스크가 아닌 분류을 포함한 NLU 태스크로의 적용을 목표로 하였다. 그 외에[14]는 외부 지식을 사용하여 프롬프트를 구축하여, NLU 벤치마크에서 효율적인 성능 개선을 달성했다. [15-16]과 같이 프롬프트 설계에 드는 자원을 절약하기 위한 자동화된 학습기반 프롬프트 구축에 대한 연구도 진행되고 있다.

프롬프트 튜닝을 시퀀스 태깅에 적용하는 것은 아직 열린 도전이라고 할 수 있다. 이 도전을 해결할 몇 가지 단서는 개체명 인식(Named Entity Recognition) 모델에게서 찾을 수 있다. 예를 들어, 하나의 접근법은 개체명 인식을 스팬에 대한 분류 작업으로 모델링하는 것이며[17], 이를 통해 중첩된 엔터티의 식별을 가능하게 한다. 스팬 분류를 기반으로 한 경계 인식하는 연구도 수행된 바가 있다[18-19].

[20]의 작업은 각 레이블 영역에 대한 별도의 프롬프트를 구성함으로써 개체명인식 태스크에 프롬프트 튜닝을 적용했다. 마찬가지로, [21]은 같은 개체 유형의 단어들 사이의 교차예측을 활용한 Templete-Free 프롬프트 튜닝 방법을 제안했다.

본 논문에서는 [1]의 연구처럼 구축된 데이터셋에 대해 새로운 벤치마크 성능을 제시하는 것을 목표로 하지만, 단순한 속성 기반 감정 분석 태스크와 달리 프롬프트 튜닝을 적용을 위해 여러가지 독창적인 방법을 제안하려한다. 먼저 [4-5]와 의존구문 분석을 통해 같이 감정 표현 위치하는 영역을 추출한 후, [20]과 같이 각각의 레이블 영역에 대한 별도의 프롬프트를 구성하려고 한다.

하지만 레이블 태그가 하나인 개체명 인식과는 달리 다중 작업-토큰 레이블링 태스크로서, 속성 및 감정 분류를 위한 템플릿 또한 필요하다. 본 논문에서는 감정 표현 영역 추출 및 시퀀스 분류 모델을 각각 파이프라인으로 연결하여 온라인 추론을 가능하도록 하는 것을 목표로 한다. 그 외에 버벌라이저-레이블 워드의 확장 및 선정 등 최종적으로 다양한 프롬프트 튜닝 전략에 대해 실험평가 수행을 통해 목표으로 하는 데이터의 특성을 조명하고, 벤치마크 성능을 제시한다.

3. 속성 기반 감정 분석 데이터

본 절에서는 사용되는 속성 기반 감정 분석 데이터셋의 전반적인 특징을 분석한다. 또한 토큰 레이블링 태스크를 시퀀스 레이블링으로 일반화하기 위해, 데이터 레벨에서의 전처리 및 선별 과정도 함께 소개한다.

본 연구에서 사용하는 속성 기반 감정 분석 데이터셋의 총 규모는 약 400만건이며 아래의 [표1-2]와 같다.

[표1 속성별 데이터 규모]

데이터	Test(%)	train(%)	valid(%)
O	269,601 (34.37)	378,493 (13.42)	72,519 (23.07)
효과/성능/기능	182,991 (23.33)	165,606 (5.87)	42,005 (13.36)
품질/디자인/구성	105,568 (13.46)	191,780 (6.80)	21,506 (6.84)
사이즈/무게/개수	61,209 (7.80)	653,162 (23.16)	23,807 (7.57)
사용감/착용감	53,350 (6.80)	972,541 (34.49)	20,751 (6.60)
가격	51,367 (6.55)	183,796 (6.51)	108,989 (34.67)
편의성/활용성	45,674 (5.82)	219,965 (7.80)	18,455 (5.87)
제조/유통/서비스	14,484 (1.84)	54,131 (1.92)	6,278 (1.99)
전체	784,244	2,819,474	314,310

[표2 속성별 데이터 규모]

데이터	test	train	Valid
긍정	342,692(43.70)	1,231,806(43.69)	137,023(43.59)
O	269,601(34.38)	972,541(34.49)	108,989(34.68)
부정	145,451(14.55)	519,648(18.43)	57,886(18.42)
중립	26,500(3.38)	95,479(3.39)	10,412(3.31)
전체	784,244	2,819,474	314,310

데이터는 Train, Test, Valid 세트로 나누어져 있지만,

사용되는 Train의 크기는 퓨샷 세팅에 따라 일부만 사용되게 된다. 데이터셋에는 두 가지 주요 유형의 레이블인 감정과 속성이 있다. 감정 레이블은 "긍정", "부정", "중립"과 같은 전달되는 감정을 나타낸다. 속성 레이블은 감정이 연결된 대상이나 주제, 예를 들면 "서비스", "가격", "품질"를 지정한다.

원본 데이터셋은 토큰 레이블링을 목표로 하면서도 속성과 감정의 다중 작업을 목표로 설계되었다. 따라서 속성과 감정 레이블이 위치한 속성 영역과 감정 표현 영역은 정확히 일치하도록 데이터셋이 구축되었다. 그 후 데이터 적합성 검증을 위한 모델은 다중 작업 학습을 수행하여 속성과 감정과 관련된 토큰을 동시에 태깅해야 한다. 여기서 구축된 데이터와 검증에 사용된 베이스라인 모델은 AIHUB(www.aihub.or.kr)에 공개되어 있다.

원본 데이터에서 속성 레이블은 상위 카테고리와 하위 카테고리로 분류되어 있지만, 본 연구에서는 속성 레이블간의 불균형을 해소하기 위해 상위 카테고리 레이블만을 사용했고, 하위 카테고리는 5절에서 버벌라이저의 레이블 워드를 확장하기 위한 인풋으로 사용된다.

기존에 사용되던 O 태그는 "속성-중립" 및 "감정-중립" 레이블로 대체되었다. 또한, B, I 태깅 방식은 본 연구에서 제거되었다. 토큰 레이블링 문제를 시퀀스 레이블링 문제로 일반화 하기 위해 레이블이 없는 토큰의 공백을 제거해야하고, 시퀀스 레이블로서의 O 태그는 감정과 속성이 레이블링 되지 않은 중립 태그와 다름 없다고 간주한다.

4절에서 제안할 2단계 파이프라인에서는 감정 표현 영역 예측에서의 오류를 최소화하는 것을 목표로 한다. 그렇기 때문에 기존에 구축된 데이터셋의 품질을 보정하기 위해, 데이터 선별 과정을 통해 데이터셋의 특징을 강제하고, 전형적이지 않은 데이터 예제를 제거한다.

먼저 의존 구문 모델[22]을 통해 구문분석을 수행 후, 최소 동사구 단위로 인풋 텍스트를 분할한다. 그 후 최소 동사구 범위내에 두 가지 이상의 감정 표현 영역이 있는 경우 예제 전체를 제거한다. 즉 최소 동사구 내에 의존소 중 감정표현이 독립적으로 있는 경우를 제거한다.

이는 인간 주석자의 일관되지 않은 레이블링을 제거함으로써 고품질의 속성 기반 감정 분석 데이터를 활용하기 위함이며, 추가적인 데이터셋 구축시 새로운 레이블링 가이드로 활용 될 수 있다.

본 절에서는 다양한 전처리 및 데이터 구조 변경을 통해 효율적인 모델 학습과 정확한 평가가 가능하도록 데이터셋을 선별하였다. 4절과 5절에서는 이러한 선별된 데이터셋을 활용하여, 속성 기반 감정 분석에 프롬프트 튜닝 및 퓨샷 세팅을 적용하는 방법에 대해 소개한다.

4. 2단계 파이프라인 아키텍처

본 절에서는 의존 구문 분석 과정에서 의해 선별된 데이터의 특성을 고려하여, 속성 기반 감정 분석 태스크에 프롬프트 튜닝을 적용하기 위한 2단계 파이프라인을 제안한다.

4-1. 1단계 : 감정 표현 영역 추출 및 분류 에제 생성

1단계에서는 전체 훈련 데이터 세트에서 훈련된 Albert-Base 구조를 기반으로 한 의존 구문 분석모델을 사용하여, 입력 텍스트를 감정 표현 영역 단위로 분할 한다

구체적으로 모델은 MLP, Bi-Affined Attention 레이어와 함께 의존 관계 분석을 수행하게 되며, 후처리로서 입력을 토큰 레이블링보다 일반화된 구-절 단위의 시퀀스 레이블링을 위한 감정 표현 영역으로 분할한다. 1단계의 감정표현 영역 검출에 대한 성능은 데이터 선별 과정에서 사용했던 모델을 그대로 사용함으로써 견고함이 보장된다.

분할된 감정표현 영역 단위의 입력 텍스트는 후처리를 통해 서브배치를 생성하게 된다. 최소 동사구 단위를 기반으로 인접한 동사구의 감정표현 영역은 합쳐진다. 이러한 과정을 통해 시퀀스 레이블링 태스크는 다시한번 시퀀스 클래시피케이션 문제로 치환되어 2단계의 모델로 전달되고, 분할된 감정표현 영역 단위의 입력 텍스트는 서브배치로서 사용된다.

4-1. 2단계 : 프롬프트 튜닝 기법을 사용한 감정-속성 분류

2단계 모델 역시 Albert 구조를 기반으로 하지만, 프롬프트 튜닝을 통한 시퀀스 분류를 수행하도록 설계되었다. 이 모델은 퓨샷 설정을 사용하여 훈련되며, 첫 번째 단계에서 생성된 감정표현 영역의 텍스트를 입력으로 받는다. 이 시점에서 첫 번째 단계 모델은 훈련 중에 두 번째 단계 모델의 매개 변수만 업데이트되도록 Freezing 된다.

문제의 다중 작업 성격을 고려하여, 속성 및 감정 레이블링을 모두 용이하게 하는 입력 템플릿을 정의한다. 레이블 영역 입력 앞에 속성을 나타내는 템플릿이 접두사로 추가되고, 그 뒤에 감정을 나타내는 템플릿이 접미사로 추가된다. 4절에서 구체적으로 설명할 템플릿은 입력에 두 개의 마스크 토큰이 자연스럽게 삽입 될 수 있도록 하는 구조를 정의함으로써 다중 작업 학습을 용이하게 한다.

프롬프트 기반 학습 전략과 일치하게, 시퀀스 분류 작업은 마스크된 언어 모델링(MLM : Masked Language Modeling) 작업으로 변환되고, 이는 전통적인 MLP 헤더 대신 MLM 헤더를 사용함으로써 달성된다. MLM 작업이 실행된 후에는 정의된 레이블 워드와 매핑 되는 레이블 클래스를 예측한다. 그 후 이러한 예측을 기반으로 현재 모델 파라미터의 loss가 계산된다.

4절에서 구체적으로 설명할 버벌라이저는 레이블 예측을 레이블 워드 예측으로 대신함으로써 사전학습과 미세조정간의 간격을 줄일 수 있도록 함으로써 프롬프트 튜닝 기법의 접근을 보여준다. 레이블 워드가 여러 개라면 각 레이블 워드의 logit 값을 평균내어 최종 레이블의 확률을 계산한다.

기존의 프롬프트 튜닝 연구들이 지적하듯 모든 레이블 워드가 vocab에 온전하게 존재하지 않을 수 있다. 이를 해결하기 위해, 사용 가능한 토큰을 조합하여 레이블 단어를 구성한다. 각 레이블 단어의 확률값은 이러한 구성 요소 토큰과 연관된 logit 값을 평균내어 계산된다.

본 연구의 2단계 모델은 퓨샷, 속성 기반 감정 분석의 도전 과제를 종합적으로 해결하려고 한다. 1단계는 감정 표현 영역 분리에 중점을 두고, 전체 데이터불일치를 제거하기 위해 세심하게 전처리된 데이터 세트를 활용해, 레이블 영역을 예측한다. 2단계 모델은 속성과 감정 레이블을 모두 처리할 수 있도록 맞춤화된 프롬프트 튜닝 시퀀스 레이블링 접근법을 사용한다.

5. 버벌라이저 및 템플릿 선정

본 절에서는 프롬프튜 튜닝을 위해 구조화된 인풋을 만들기 위한 버벌라이저와 템플릿의 확장 및 선정 방법에 대해 더 구체적으로 살펴보고, 실험 결과를 구분할 수 있도록 실험 세팅명을 분류하고 정리한다

1) Baseline Setting

2단계 아키텍쳐를 통해 감정 표현 영역 단위로 분할된 예제들에 대해 단순히 분류를 수행한다. 이때 벡터 표상을 위한 임베딩 모델은 공유하며, 감정과 속성을 각각 분류하기 위한 MLP 헤드를 사용한다.

2) Manual Setting

초기 레이블 워드를 미리 정의한 후, 이를 기반으로 레이블 워드를 확장한 버벌라이저를 사용한다. 감정 레이블의 초기 레이블 워드를 선정하기 위해, 학습 데이터셋에 대해 형태소 분석을 수행하고 가장 빈번하게 등장한 형용사 형태소와 해당 형태소가 속한 단어를 참고한다. 예를 들어 감정 레이블이 긍정인 데이터 예제에서 가장 많이 등장한 형태소는 편하/VA 이며, 부정에서는 아쉽/VA 이다. 형태소 분석에는 자사 솔루션인 KMA-Black이 사용되었다.

속성 레이블의 경우는 데이터 수집 및 구축 당시에 사용된 온라인 쇼핑몰의 세부 카테고리를 사용한다. 예를 들어 제조/유통/서비스 대분류는 레이블로 사용되고 있으며, 세부 중분류인 제조사, 유통기한 등은 초기 레이블 워드로 사용 한다.

[표3 확장된 레이블 워드 예시]

태스크	레이블	레이블 워드
aspect	중립	일반,없음,기타,해당사항없음,미정,미분류,그냥,평소,보통,대다수,정규,이외,평범 …
	효과/성능/기능	음량,음질,화질,세정력,속도,보습력,흡수력,내구성,유지력,발림성,발색력,보습력,절삭력 …
	품질/디자인/구성	색상,제품구성,용기,마감,포장,굽,수납,소재,재질,질감,외관,유니크함,스타일,디테일 …
	사이즈/무게/개수	무게,용량,사이즈,두께,치수,모델,판매단위,부피,총중량 …
	사용감/착용감	착용감,향,냄새,소음,향,착용감,자극성,착화감,핏, 촉감,감촉,촉감,사용감,조립성,호흡성 …
	가격	가격대,할인,프로모션,가성비,저렴,고가,중가,할인율,부가세 …
	편의성/활용성	활용도,수납,공간,연마용이성,세척용이성,정리성,수납력,다용도,접이식,취급성,휴대성,조작…
	제조/유통/서비스	제조일,제조사,서비스,유통기한,제조국,배송,AS,제조공정,리콜,유통경로,보증서,인증마크,,,
sentiment	중립	안정된,평소의,중립적인,평소와같은,멍한,관심없는,생각없는,그저그런,그냥그런,평범한,보통의,심심한 …
	긍정	편한, 좋은, ,사랑스러운,놀라운,매력적인,굉장한,아름다운,유익한,풍성한,훌륭한,깨끗한,멋진 …
	부정	아쉬운, 나쁜,화난,슬픈,짜증난,불안한,끔찍한,진부한,, 소름끼치는,두려운,역겨운,혐오스러운,불만족스러운…

그 후 레이블 워드를 확장하기 위해서는 단순하게 Open API의 CHATGPT[23]와 같은 LLM 서비스를 사용한다. 레이블 마스크의 위치에 치환이 될만한 동의어, 유사어, 반의어, 연관어 등을 추천해주는 식으로 프롬프트를 구성하여 레이블 워드를 확장한다.

3) Refinement Setting

레이블 워드의 정제 단계에서, 감정-속성 레이블별로 100개의 샘플 예시를 램덤 샘플링한 후 사전 학습 모델에 마스크 토큰과 함께 입력한다. 그 후 예측된 vocab내 토큰들의 모든 logit값을 얻고, 확장된 레이블 워드와 비교하여, 레이블과 레이블워드의 연관성을 미리 살펴본다. 레이블과 상관없이 일관되게 낮은 로짓 값을 보이는 단어들은 레이블 워드 목록에서 제거된다.

4) Recycle Setting

버려진 레이블들은 감정과 측면 카테고리에 대한 중립 태그 (O 태그)로 재활한다. 중립 태그에 대해 logit값 기반 정제 단계가 한번 더 수행되고, 이 반복적인 과정을 마친 후, 중립 태그에 사용될 레이블 워드를 확정한다.

[표4 속성 기반 감정분석 템플릿 예시]

구분	템플릿
1인칭	>제품 특성 중 {"mask"} 관점에서, {"text"} 그래서 {"mask"} 느낌. >{"mask"} 관점에서, {"text"} 그래서 {"mask"}것 같아. …
3인칭	>헤딩 리뷰는 {"mask"}에 대한 글이야 {"text"} 그래서 글쓴이는 {"mask"} 기분이야. >다음은 {"mask"}에 관한 리뷰야. {"text_a"} 그래서 작성자는 {"mask"} 것 같아. …..

템플릿의 경우에는 속성 기반 감정 분석 데이터의 출처인 리뷰 데이터 도메인을 고려해, 두가지 실험 세팅을 제안한다.

5) 1-Veiw Setting

첫번째는 1인칭 주인공 시점의 템플릿이다. 이는 리뷰 작성자의 관점에서, 감정과 속성 레이블이 예측될 [MASK]를 인풋 텍스트의 앞뒤에 삽입한다. 미세조정 데이터의 구어체를 타겟으로 마스크 토큰이 포함된 자연스러운 인풋 데이터를 구성한다.

6) 3-Veiw Setting

두번째는 3인칭 관찰자 시점의 템플릿으로서, 사전학습 당시 대용량 코퍼스 중 많은 비율을 차지했던 문어체 데이터와 비슷한 형태로 인풋을 구성한다.

6절에서는 마지막으로 여러 실험세팅에 대해 모델 선정

과정 실험결과를 제시한후 최종적으로 벤치마크 성능을 제시한다.

6. 실험

실험에서는 5절에서 제안했던 다양한 버벌라이저와 템플릿의 선정 전략을 실험을 통해 비교 분석하고 최종적으로 제안하는 모델 세팅을 선정한다. 사용한 Metric은 F1 Score 이며, Shot 세팅에 따라 사용된 학습 데이터셋은 시행마다 랜덤으로 샘플링된다. 즉 시행마다 다른 학습셋, 동일한 테스트셋을 가지고 모델이 평가되며, 아래 [표 5-7]의 실험 결과는 총 5번의 시행을 평균낸 결과이다.

먼저 [표 5]의 실험에서는 두 가지 다른 템플릿 방식에 대한 성능을 보여준며, 이때 사용한 Verbalizer는 Manual 이다. Shot 세팅이 커질수록 두 템플릿 간의 성능차이가 줄어들긴 하지만 전반적으로 구어체 템플릿인 1-View의 성능이 더 좋은 것으로 보인다. 이는 파인튜닝 셋인 속성기반 감정분석의 데이터셋 포멧에 템플릿이 가까울수록 더 좋은 퓨샷 성능을 보인다고 해석할 수 있다.

[표5 템블릿 별 성능]

Shot	Model	Sentiment	Aspect
4	3-Veiw	0.6160	0.4899
	1-Veiw	**0.6308**	**0.5012**
32	3-View	0.6862	0.5920
	1-Veiw	**0.6940**	**0.6011**

[표 6]은 단계별로 도입한 레이블 워드의 초기화, 확장, 정제, 재활용 전략의 유용함을 확인하는 실험 결과이다. 결과적으로, Refinement 단계에서 사용되지 않은 레이블 워드를 중립 레이블의 레이블 워드로 재활용했을때, 모델 성능이 개선되는 것을 확인하였다. 이러한 결과는 데이터 전처리 선별 과정에서 0 태그를 중립태그로 치환했기 때문이라고 예상된다.

[표6 버벌라이저 별 성능]

Shot	Model	Sentiment	Aspect
4	Manual	0.6308	0.5012
	Refinement	0.6357	0.5020
	Recycle	**0.6415**	**0.5167**
32	Manual	0.6940	0.6011
	Refinement	0.7128	0.6370
	Recycle	**0.7163**	**0.6409**

[표 7]에서는 최종적으로, 두개의 MLP레이어가 적용된 베이스라인 모델의 성능과 모델 선정과정에서 채택된 각각의 방법론을 통합한 모델의 성능을 비교하였다. 제안하는 방법론(ours)은 MLP 레이어 대신 MLM 레이어가 적용되었고, 템플릿의 1-Veiw, 버벌라이저의 Recycle 세팅이 적용된 모델이다.

[표7 최종 퓨샷 세팅 성능]

Shot	Model	Sentiment	Aspect
1	Baseline	0.3964	0.1053
	ours	**0.5760**	**0.4739**
4	Baseline	0.4491	0.1901
	ours	**0.6415**	**0.5167**
8	Baseline	0.4481	0.3175
	ours	**0.6519**	**0.5546**
16	Baseline	0.5589	0.4460
	ours	**0.6997**	**0.6138**
32	Baseline	0.6397	0.5666
	ours	**0.7163**	**0.6409**
64	Baseline	0.7755	0.6971
	ours	**0.7983**	**0.6842**

Shot 세팅을 4~64까지 조절해가며 테스트한 결과, 전반적으로 제안하는 방법론의 성능이 베이스라인보다 훨씬 상회하는 결과를 보인다. 64 Shot 정도에 이르면 Baseline과 Ours는 비슷한 성능에 수렴하는 것을 확인할 수 있었다.

실험을 통해 제안하는 다양한 버벌라이저, 템플릿 세팅이 세팅이 모델 성능에 주는 영향을 분석하였다. 특히, 특히, 낮은 확률의 태그를 중립태그로 치환하는 Recycle Setting, 그리고 마스크 토큰이 포함된 구어체의 템플릿을 적용한 1-Veiw Setting이 모델 성능 개선에 긍정적인 영향을 미친다는 것을 Baseline Setting 과의 비교를 통해서 다시 한번 확인하였다.

7. 결론 및 향후 과제

본 논문에서는 속성 기반 감정 분석 데이터 프롬프트 튜닝을 적용하는 것을 목표로 한다. 그 과정에서 다중작업-토큰 레이블링 문제를 단순화하기 위한 두 단계 아키텍처를 제안하였다. 준비단계에서 먼저 감정 표현 영역 분할을 위해 데이터의 특징을 강제하기 위하여, 전형적이지 않은 예제들을 제거한다. 그 후 첫번째 단계에서 의존 구문 분석을 사용하여, 인풋 데이터는 감정 표현 영역에 따라 분할된다.

두번째 단계에서는 템플릿과 버벌라이저를 통해 구조화된 인풋 형식으로 한번 더 치환된다. 리뷰 도메인 데이터의 특성을 고려하여 템플릿과 버벌라이저의 다양한 세팅을 나열하고, 실험을 통해 최적의 실험 세팅을 선정하는 작업을 수행하였다. 실험 결과 최종적으로 선정된 모델로 실험한 결과 구축된 데이터의 특성을 다시 한번 확인할 수 있었으며, 후속 연구를 위한 유의미한 벤치마크 성능을 제시하였다

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.1711117120, 뉴럴-심볼릭(Neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)

참고문헌

[1] 김문형, 장하연, 조유미, 신효필, "KOSAC(Korean Sentiment Analysis Corpus) : 한국어 감정 및 의견 분석 코퍼스", 2013년 한국컴퓨터종합학술대회 논문집, pp. 650-

652, 2013.
[2] 박호민, 김재훈, "다중 작업 학습의 단계적 특징을 활용한 한국어 속성기반 감정 분석에서의 대상 추출", 제 34회 한글 및 한국어 정보처리 학술대회 논문, pp.630-633, 2022.
[3] 박지은, 이주상, 옥철영, "BERT+CRF를 이용한 다중 감성 표현 영역 추출", 33회 한글 및 한국어 정보처리 학술대회 논문집, pp.571-575, 2021
[4] G. Qiu, B. Liu, J. Bu and C. Chen, "Opinion word expansion and target extraction through double propagation", Computational linguistics, vol. 37, no. 1, pp. 9-27, 2011
[5] C. Zhang, Q. Li, D. Song, and B. Wang, "A multi-task learning framework for opinion triplet extraction," in Findings of EMNLP, 2020, pp. 819-828.
[6] Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In Conference on Empirical Methods in Natural Language Processing
[7] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," in AAAI, 2019, pp. 6714-6721. Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In AAAI Conference on Artificial Intelligence
[8] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018b. Exploiting document knowledge for aspect-level sentiment classification. In Annual Meeting of the Association for Computational Linguistics.
[9] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in W-NUT, 2019, pp. 34-41.
[10] Tom B. Brown, Benjamin Mann, et.al. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.
[11] Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. Language Models as Knowledge Bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3-7 November 2019; pp. 2463-2473.
[12] Davison, J.; Feldman, J.; Rush, A.M. Commonsense Knowledge Mining from Pretrained Models. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Hong Kong, China, 3-7 November 2019; pp. 1173-1178.
[13] Han, X.; Zhao, W.; Ding, N.; Liu, Z.; Sun, M. Ptr: Prompt tuning with rules for text classification. AI Open 2022, 3, 182-192
[14] Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In Proceedings of COLING, pages 5569-5578.
[15] Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1-6 August 2021; pp. 4582-4597.
[16] Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7-11 September 2021; pp. 3045-3059.
[17] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1835-1845, Online. Association for Computational Linguistics.
[18] Changmeng Zheng, Yi Cai, e. tal. 2019. A boundary-aware neural model for nested named entity recognition. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 357-366, Hong Kong, China. Association for Computational Linguis□ic
[19] Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2843-2849, Brussels, Belgium. Association for Computational Linguistics.
[20] Xue Mengge, Bowen Yu, et. al. 2020. Coarse-to-Fine Pretraining for Named Entity Recognition. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6345-6354, Online. Association for Computational Linguistics.
[21] Ma, R.; Zhou, X.; Gui, T.; Tan, Y.; Li, L.; Zhang, Q.; Huang, X.J. Template-free Prompt Tuning for Few-shot NER. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online and Seattle, WA, USA, 10-15 July 2022; pp. 5721-5732
[22] 김봉수, at al, "사전 학습 모델과 Specific-Abstraction 인코더를 사용한 한국어 의존 구문 분석" 32회 한글 및 한국어 정보처리 학술대회 논문집, 2020.
[23] ChatGPT : https://chat.openai.com/