공 학 석 사 학 위 논 문

AI-In-The-Loop for NER tagging: 대용량 언어모델과 Few-Shot 학습을 이용한 개체명 인식 데이터셋 구축 프로세스

2023년 8월

부 경 대 학 교 대 학 원

산업및데이터공학과

박 은 빈

공 학 석 사 학 위 논 문

AI-In-The-Loop for NER tagging: 대용량 언어모델과 Few-Shot 학습을 이용한 개체명 인식 데이터셋 구축 프로세스

지도교수 최 성 철

이 논문을 공학석사 학위논문으로 제출함.

2023년 8월

부 경 대 학 교 대 학 원

산업및데이터공학과

박 은 빈

박은빈의 공학석사 학위논문을 인준함

2023년 8월 18월

위 원 장 공학박사 서 원 철 (인)

위 원 공학박사 이 지 환 (인)

위 원 공학박사 최 성 철 (인)

목 차

표 목차 ····· i

그림 목차 ····· ii

Abstract ·····

Ⅰ. 서론 ····· 1

1. 연구 배경 ····· 1

2. 연구 목표 및 내용 ····· 3

Ⅱ. 선행 연구 ····· 4

1. Named Entity Recognition Annotation Process ····· 4

2. LLM Annotations NLP Datasets ····· 5

Ⅲ. 데이터셋 태깅 프로세스 ····· 7

1. 연구 방법 소개 ····· 7

2. 데이터 설명 ····· 8

2.1 CoNLL 2003 ····· 8

2.2 Broad Twitter Corpus ····· 9

2.3 WNUT 2017 ····· 9

2.4 WikiAnn ····· 9

3. Effective Prompt Engineering and Entity Annotation ····· 10

3.1 Generative Pre-trained Transformer ····· 10

3.2 Prompt Message ····· 12

3.3 Task Description ····· 13

3.4 Few-Shot Description ····· 13

3.5 Entity Annotation in Context with Large Language Models ····· 15

4. Training Details ····· 16
4.1 Using Match Rate for Evaluation Consistency among Models ····· 18
4.2 AI-In-The-Loop ····· 19
4.3 반복 ····· 19

Ⅳ. 실험결과 ····· 21
1. Performance F1 Score ····· 21
1.1 Fine-tuning with Original Dataset ····· 21
1.2 Active Learning with Large-scale Language Model ····· 21
1.3 CoNLL2003 Entities ····· 23
2. Labelling Cost Comparison ····· 24
Ⅴ. 결론 ····· 26
Ⅵ. 부록 ····· 28
참고문헌 ····· 34

표 목차

Table 1 실험에 사용하는 데이터셋 ····· 8
Table 2 모델 하이퍼파라미터 ····· 17
Table 3 벤치마크 데이터셋 기본 성능 ····· 22
Table 4 능동적 학습과 대규모 언어 모델의 보정 결과 ····· 23
Table 5 대규모 언어 모델의 태깅 성능 ····· 24
Table 6 보정된 데이터셋과 추산 금액 ····· 25

그림 목차

Figure 1 Learning from Human Preference ····· 12
Figure 2 Few-shot Task Description ····· 14
Figure 3 전체 프로세스 흐름 ····· 21

AI-In-The-Loop for NER tagging: A Novel Process for building Named Entity Recognition Dataset using a Large-scale Language Model and Few-Shot Learning

Eun Bin Park

Department of Department of Industrial and Data Engineering, The Graduate School,
Pukyong National University

Abstract

This study proposes a novel Named Entity Recognition labelling process using large-scale language models, and demonstrates its potential to compensate for incomplete datasets and improve labelling. We show that the process can be used to proactively label and calibrate datasets, despite the fact that only 10% of the responses in the training dataset are correct.

Furthermore, this work provides important insights into the feasibility of integrating active learning and model community-based tagging. In doing so, we propose a perspective on entity name recognition in large-scale language models and suggest strategies for correctly tagging unlabelled data in a cost-effective manner, thereby improving the model's understanding of the data. Furthermore, we show that our method plays an important role in minimising the cost of data labelling and maintaining a certain level of model performance even with insufficient data. Based on these findings and proofs, the focus of this thesis is to provide practical suggestions for the effective use of large-scale language models and the improvement of labelling tasks.

Ⅰ. 서 론

1.1 연구 배경

자연어 처리 (Natural Language Processing)의 개체명 인식(Named entity recognition, NER)은 텍스트에서 특정한 개체를 식별하고 분류하는 과정이다. 이는 정보 검색, 질문응답, 지식 추출 등 다양한 자연어 처리 분야에서 핵심적인 역할을 한다. 이러한 작업은 텍스트로부터 유용한 정보를 추출하고, 사용자 질문에 대한 정확한 답변을 제공하며, 기존의 지식 베이스를 확장하는 데에 개체명 인식을 진행한다. 따라서 개체명 인식의 정확도와 효율성은 전반적인 성능에 큰 영향을 미친다.

최근 딥러닝의 발전으로 개체명 인식 모델 성능이 크게 향상되었지만, 모델 성능은 여전히 데이터셋 품질에 의존한다. 데이터셋 품질이 모델의 성능을 결정짓는 중요한 요소이며 정확한 라벨의 데이터셋은 지도 학습(Supervised-Learning)의 성능을 결정하는 핵심이다. 그러나 좋은 품질의 데이터셋을 제작하기 위해서는 많은 시간이 필요하며 소요되는 비용도 높음에도 불구하고 대부분의 데이터 라벨링 방법은 여전히 시간과 비용이 많이 드는 노동집약적 작업에 의존하고 있다.

크라우드 소싱은 대량의 인적자원을 활용해 작업을 수행하는 방식으로, 대규모 데이터셋에 빠르고 경제적으로 데이터셋 구축이 가능하다. 다양한 작업에서 크라우드 소싱을 통해 데이터셋을 구축하기도 한다. 기술자나 전문가와 같은 고비용의 전문가 라벨링을 대체하기 위해, 크라우드 워커를 고용하고 그들에게 구체적인 지시사항을 제공하는 방법을 활용한다. 다수의 인력을 활용하여 빠른 시간 내 주석을 달 수 있기에 데이터셋 구축에 많이 사용되었고, 크라우드소싱 플랫폼이나 크라우드소싱으로 라벨을 단

데이터 자체가 연구의 대상이 되기도 한다[1]. 다만 크라우드소싱 플랫폼인 MTurk 데이터 품질 저하에 관한 연구도 제기될 만큼 일관되지 않은 품질, 인력과 관련된 비용 및 시간 등의 문제가 있다[2].

크라우드소싱을 활용한 라벨링은 라벨 일관성, 주석의 품질, 작업 지시사항의 명확성 등 여러 문제가 품질에 영향을 미친다. 이를 극복하기 위해 베이지안 접근방식이나[3] Adversarial Learning[4] 등의 여러 방법론을 데이터 라벨링에 활용하려는 시도가 있다. 능동적 학습 (Active Learning)을 적용해 정해진 예산 안에서 라벨을 갱신하고 학습에 사용될 데이터를 선별하는 방법을 사용하기도 한다[5]. 비용과 인력 문제로 최근 주목받는 방법론 중 하나는 대규모 언어 모델 (Large Language Model, LLM)을 이용한 데이터셋 구축 방법이다.

대규모 언어 모델은 대용량의 텍스트 데이터를 통해 학습된 인공지능 모델로 특히 자연어 처리에서 주로 활용되고 있다. 대규모 언어 모델은 일반적으로 수십억 혹은 그 이상의 파라미터를 가지며 모델의 규모와 학습에 사용되는 데이터셋의 크기에 따라 성능이 결정된다. 주로 트랜스포머 (Transformer) 아키텍처를 기반으로 하며 대표적으로 GPT(Generative Pretrained Transformer), T5(Text-to-Text Transfer Transformer) 등이 있다. 문맥을 이해하고 문장을 생성하는 능력 덕분에 챗봇이나 대화형 AI, 문서 생성 등의 작업에도 활용된다. 이런 특성을 활용하여, 대규모 언어 모델을 크라우드 워커와 비교하거나[6], 함께 사용하여 라벨을 다는 연구가 진행되고 있다[7]. 대규모 언어 모델로 주석을 달 때는 실제 사람에게 제공하는 가이드라인과 비슷하게 작업에 대한 구체적 의미를 알려주고, 참고할 수 있는 예시를 몇 가지 제공한다. 이를 통해, 대규모 언어 모델은 주어진 문맥에 대한 이해를 바탕으로 더 효과적인 라벨을 제공할 수 있게 된다.

1.2 연구 목표 및 내용

본 논문에서는 능동적 학습의 Query by Committee와 Human In The Loop 컨셉을 바탕으로 모델 커뮤니티가 존재하며 데이터 재보정에서 사람이 아닌 대규모 언어 모델을 활용하는 새로운 방식을 제안한다. 이 접근법에서는 '모델 커뮤니티'라는 개념을 도입한다. 이는 한 개 이상의 트랜스포머 기반 사전학습 언어 모델 (Pre-trained Language Model, PLM)을 의미하며, 모델 커뮤니티는 점차 보정되는 데이터로 능동적 학습을 진행하고 학습할 데이터를 미리 평가하며 보정할 데이터를 선별하는 역할이다. 선별된 데이터는 대규모 언어 모델에 의해 새 라벨을 부여받는다. 모델 커뮤니티가 선정한 불확실성이 높은 데이터는 모델이 이견을 내놓은, 아직 잘 이해하지 못하는 부분을 나타내므로 이 데이터에 대한 라벨 갱신은 모델의 성능을 향상시키는 데 도움이 될 수 있다. 또한 라벨이 갱신된 데이터셋은 전체적인 데이터셋 구축에 기여한다.

본 방법을 통해 적은 비용으로 라벨이 없는 데이터를 올바르게 교정할 수 있으며, 충분하지 않은 데이터로도 모델이 어느 정도의 성능을 보여줌을 확인한다. 본 연구는 능동적 학습과 Few-shot 그리고 커뮤니티 기반 태깅을 개체명 인식 태깅 프로세스에 통합할 수 있는 가능성에 대한 인사이트를 제공하며, 개체명 인식 작업에 대한 대규모 언어 모델의 적용 가능성에 대한 새로운 관점을 제시하고 이 분야의 향후 연구를 위한 기반을 구축하는 것을 목표로 한다.

Ⅱ. 선행 연구

2.1 Named Entity Recognition Annotation Process

크라우드 소싱은 데이터 수집에 효과적인 방법이지만, 주석자에 의해 크게 좌우되기도 한다. Tim Finin 등 (2010)은 Amazon Mechanical Turk와 CrowdFlower를 사용하여 트위터 데이터에 대한 Named Entity(NE) 라벨을 수집하는 방법에 대한 연구를 진행한다[8]. 이 연구의 목표는 페이스북이나 트위터와 같은 소셜 미디어 플랫폼에서의 개체명 인식에 대한 전반적인 연구로 3년 동안 수집된 1.5백만명의 사용자로부터의 1.5억개 이상의 트윗을 분석한다. 트위터 데이터는 비공식적이고 줄임말이 많은 특성과 같은 애로사항을 가지고 있다는 점을 강조한다.

MTurk와 CrowdFlower 모두 상대적으로 사용하기 쉽고 매우 비용 효율적인 장점이 있지만 CrowdFlower의 “골드 스탠다드” 평가 도구에 문제가 있다고 지적하지만 플랫폼 지원이 빠른 피드백으로 일부 문제를 완화할 수 있었다고 한다. 하지만 작업의 잘못된 보정, 완료된 작업에서 결과 다운로드 오류, MTurk에서 설정한 것과 다른 가격 표시, CrowdFlower 시스템에 골드 스탠다드 데이터가 저장되지 않는 문제 등이 있다고 지적하며, 시스템 상에서의 토큰 제한에 대한 문제를 지적한다.

Hege Fromreide 등 (2014)은 트위터에서 개체명 인식에 대한 두 가지 새로운 데이터셋을 제시한다. 저자는 언어의 편향성과 은어 등의 사용으로 트위터 데이터의 개체명 인식의 어려움을 강조하였다[9]. 수동으로 주석을 단 데이터셋과 크라우드소싱 데이터셋을 소개하며 크라우드소싱 데이터셋을 통해 제안한 모델이 좋은 성능을 달성하였다고 언급한다. 다만 트위터의 언어 변화가 상당히 빠르며, 이로 인해 최신 트윗에 대한 모델을 훈련

시키기 위해 몇 달 전의 트위터 훈련 데이터는 거의 쓸모가 없다는 점을 주장한다. 크라우드소싱의 한계도 지적하며 저자들은 무작위로 주석을 단 스패머와 작업을 충분히 이해하지 못한 주석 작성자의 사례를 발견했다. 또한 시스템 훈련에 사용된 데이터와 다르게 샘플링된 트윗에 대해 트위터 시스템용 개체명 인식 모델을 평가할 때 상당한 성능 저하를 관찰했다.

2.2 LLM Annotates NLP Datasets

Shouhang Wang 등 (2021)의 연구에서는 GPT-3를 저렴한 데이터 라벨링 도구로 활용하는 방법을 연구한다[6]. 저자들은 GPT-3에서 생성된 라벨과 인간이 제공한 라벨을 결합하는 프레임워크를 제안하였다. 이 프레임워크를 통해 제한된 라벨링 예산 내에서 더 나은 성능을 달성할 수 있음을 주장한다. 다양한 자연어 이해(Natural Language Understanding, NLU) 및 자연어 생성(Natural Language Generation, NLG) 작업에서 동일한 성능을 달성하기 위해 GPT-3에서 라벨을 사용하는 것이 인간으로부터 라벨을 사용하는 것보다 50%에서 96% 더 저렴하다는 것을 얘기한다. 또한, GPT-3로부터 얻은 라벨과 인간 라벨을 결합하면 제한된 라벨링 예산 내에서 더 나은 성능을 달성할 수 있음을 보여주었다. 이러한 결과는 많은 실용적인 응용 분야에 일반화 가능한 비용 효율적인 데이터 라벨링 방법론을 제시한다. 그러나 GPT-3의 라벨링이 인간의 라벨링보다 훨씬 저렴하지만 라벨의 품질의 경우 사람의 라벨링이 품질에서 성능에서 우세하다는 점이다. 예산이 제한된 경우, GPT-3 라벨링이 더 비용 효율적인 선택이 될 수 있다는 점이다.

Gilardi, F 등 (2023)은 대규모 언어 모델의 텍스트 주석 작업에 대한 잠재

력을 연구하며 특히 2022년 11월에 출시된 ChatGPT에 초점을 맞춘다[7]. 2,382개의 트위터 텍스트를 사용하여 ChatGPT가 관련성, 입장, 주제, 프레임 감지 등의 작업에 대해 크라우드 워커를 능가함을 보여준다. ChatGPT의 제로샷 정확도는 5개 작업 중 4개에서 크라우드 워커를 압도하며, ChatGPT의 인터코더 합의는 모든 작업에 대해 크라우드 워커와 훈련된 주석자를 모두 능가한다. 또한, ChatGPT의 주석 당 비용은 $0.003 미만으로, MTurk보다 약 20배 저렴하다고 언급한다. 이러한 결과는 대규모 언어 모델이 텍스트 분류의 효율성을 극적으로 향상시킬 수 있는 잠재력을 보여주었다.

Ⅲ. 데이터셋 태깅 프로세스

3.1 연구 방법 소개

본 논문에서는 네 가지의 벤치마크 데이터셋으로 실험을 진행한다. 전체 학습 데이터셋의 10%를 초기 학습 데이터셋으로 두고 이외에는 라벨을 모두 제거해 라벨이 없는 데이터셋으로 데이터풀을 구성한다. 마지막으로 검증과 테스트 데이터셋으로 학습을 마친 모든 커뮤니티의 모델을 평가한다. 모든 모델은 라벨이 정확한 학습용 데이터셋으로 첫 학습을 진행한다. 이후 데이터풀에서 100개의 샘플 데이터를 랜덤으로 추출해 학습된 모델로 각 문장의 예측 개체명을 확인한다. 커뮤니티의 각 모델이 예측한 값을 상호간에 비교하며 라벨의 일치 비율을 계산한다. 미리 지정한 임계값(Threshold) 보다 이 일치 비율이 낮은 경우 이를 불확실한 데이터로 분류한다. 불확실하다고 판단된 데이터는 미리 작성해둔 프롬프트 메시지와 함께 대규모 언어 모델로 전달되고, 대규모 언어 모델은 해당 문장의 정답을 보정한다. 모든 불확실한 문장에 대한 답변을 얻으면 데이터풀에 있던 기존 값을 갱신하며, 라벨의 일치 비율이 설정한 임계값보다 높았던 예측 데이터와 합쳐 다음 학습 라운드의 데이터로 추가한다.

Table 1 실험에 사용하는 데이터셋

데이터셋	학습	데이터풀	검증	테스트
CoNLL2003	1,401	12,637	3,250	3,453
BTC	633	5,705	1,001	2,000
WNUT 2017	339	3,055	1,009	1,287
WikiAnn	2,000	18,000	10,000	10,000

3.2 데이터 설명

실험에 사용하는 데이터셋은 특정 도메인에 국한되지 않은 데이터셋으로만 선정했다. 다만 비일반적인 엔티티의 인식률을 확인하기 위해 WNUT 2017을 포함하였다. 모든 데이터셋은 BIO 형식을 따르며 B (개체명의 시작), I (개체명의 중간 또는 끝) 접두사가 붙으며 실제로 사용되는 태그는 B-PER, I-LOC 등으로 표현된다. 단어가 어떤 개체명에도 속하지 않는다면 O (Non-entity token) 태그가 부여된다.

성능 평가는 Precision, Recall과 F1 Score를 사용하여 진행된다. 실험에서는 모델이 실세계에서 수집된 라벨이 없는 데이터셋이나, 노이즈 데이터를 처리하는 능력을 더 정확하게 평가할 수 있도록 데이터셋의 10%만을 온전한 정답이 있는 데이터셋으로 분리하고, 이외의 학습 데이터셋을 모두 O 태그를 부여해 데이터풀로 관리한다. 데이터풀 이외의 검증과 테스트용 데이터셋은 모델의 정확한 성능을 평가할 수 있도록 원본 데이터셋 그대로 둔다.

3.2.1 CoNLL 2003

The Conference on Natural Language Learning (CoNLL) 2003은 개체명 인식 작업을 위한 대표적인 벤치마크 데이터셋이다[10]. 영어와 독일어 두 가지 버전으로 제공되며 본 논문에서는 영어만을 사용한다. 뉴스 기사에서 추출된 문장으로 구성되어 있으며, 사람(PER), 위치(LOC), 조직(ORG) 및 이 세 가지 범주에 속하지 않는 기타(MISC) 유형이 있다. MISC는 광범위한 카테고리이며 특정 기준이나 명확한 정의에 의해 분류되지 않는 경우다.

3.2.2 Broad Twitter Corpus

트위터에서 수집된 텍스트로 구성된 BTC 데이터셋은 다양한 주제를 함양한다[11]. 개체명은 사람, 위치, 조직인 3가지 엔티티로 이루어져 있다. 이 데이터셋은 트위터 텍스트의 특징인 문자의 간결성, 비공식 언어 사용, 은어 및 약어가 포함되어 있다. 기존의 문법과 어휘가 적용되지 않을 수 있는 소셜 미디어 데이터에서 일반적인 엔티티를 검출할 수 있는지 확인할 수 있는 특징이 있다.

3.2.3 WNUT 2017

The Workshop on Noisy User-generated Text (WNUT) 2017은 BTC와 같이 소셜 미디어 게시물이나 웹 포럼, 뉴스 댓글과 같은 사용자 생성 텍스트로 이루어져있다[12]. 이 데이터셋은 비표준적인 언어 사용, 문법 오류, 오타, 축약어 등을 포함하고 있다. 목표 개체명 또한 일반적인 개체명 범주 이외에도 특정 이벤트나 상품과 같이 비표준적인 엔티티가 포함된다. 이러한 엔티티는 표준 범주에 잘 들어맞지 않고 많이 출현하지 않으며 예측하기 어려운 경우가 많기에 비일반적인 엔티티를 인식하고 처리하는 능력을 확인한다.

3.2.4 WikiAnn/en

WikiAnn은 Wikipedia에서 추출된 데이터로 구축된 대규모 개체명 인식 데이터셋이다[13]. Wikipedia에서 약 40만 개의 페이지로부터 추출된 데이터

셋이며 282가지 언어에 대한 개체명 인식에 대응한다. 본 논문에서는 영어 데이터만을 사용한다. 다양한 주제와 분야에 걸쳐 사람, 위치, 조직 개체명을 포함한다.

3.3 Effective Prompt Engineering and Entity Annotation

프롬프트 메시지는 불완전한 데이터셋에 대한 라벨을 얻기 위해 대규모 언어 모델에 사용된다. 대규모 언어 모델이 어떤 작업을 처리해야 하는지 알려주고, 더 좋은 답변을 도출할 수 있게 유도하며 각 개체명의 범주를 설명하고, 문장 전체 태깅 예시를 주어 출력 형식을 제어한다.

프롬프트 메시지의 전체적인 흐름은 OpenAI에서 게시한 글을 바탕으로 구성되었다[16].

3.3.1 Generative Pre-trained Transformer

GPT는 OpenAI에 의해 개발된 대규모 트랜스포머 기반의 언어 생성 모델이다[14]. 이 모델은 주어진 문맥에 따라 다음 단어를 예측하는 방식으로 동작한다. GPT는 지속적으로 업데이트 되어왔으며 그 중 GPT-3는 1750억 개의 파라미터를 가진 것으로 알려져 있다.

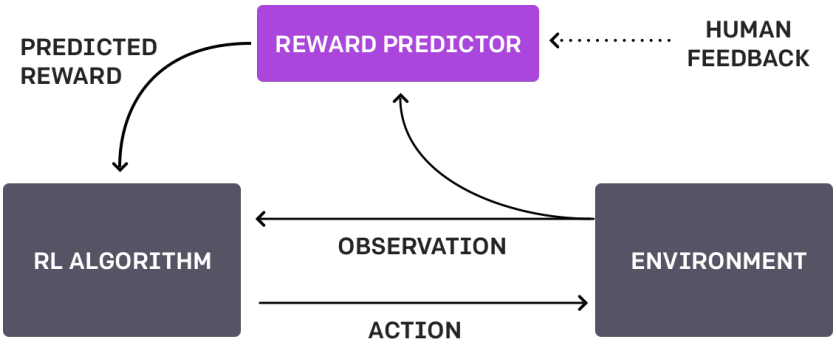

Figure 1. Learning from Human Preference [19]

InstructGPT는 OpenAI가 개발한 GPT3의 미세 조정 버전이다. 기본적으로 GPT3의 구조를 활용하지만 인간의 피드백에서 학습하는 강화학습 방법인 RLHF (Reinforcement Learning from Human Feedback)이 적용되었다. 기존의 GPT3은 다음에 위치할 단어를 예측하도록 학습되었지만, InstructGPT는 사람이 작성한 데이터로 지도학습을 받고 추가로 강화학습이 적용되었다[15]. InstructGPT는 사용자의 질문에 답하거나, 특정 주제에 대한 요약하거나, 문제를 해결하는 방법을 설명하는 등의 사용자로부터 주어진 명령에 따라 텍스트를 생성하는 데 특화되어 있다. 현재 이 모델은 GPT-3.5로 알려져있다.

GPT와 같은 생성형 AI가 환각 현상에 빠진다는 비판이 제기되고 있다. 이는 모델이 사실 관계와 무관하게 논리적인 구조를 만들어 내어 잘못된 정보를 생성하거나, 부적절하게 편향된 문장을 생성하는 경향이 있다는 것을 의미한다. 이러한 한계가 명확한 만큼 생성형 AI에게 어떻게 지시하는지에 대한 프롬프트 엔지니어링 (Prompt Engineering)이라고 방법론이 부상하고 있다. 기존 언어 모델은 모델 자체를 미세 조정해하여 원하는 답변을 얻게끔 학습했다면, 생성형 모델은 원하는 답변을 얻기 위해 모델에게 질문하는 방식이 취해진다.

3.3.2 Prompt Message

프롬프트 엔지니어링은 대규모 언어 모델로부터 최상의 결과물을 얻기 위해 적합한 프롬프트를 디자인하는 과정이다. 대규모 언어 모델이 입력 프롬프트에 대해 어떻게 반응하는지를 조정하고 최적화하는 과정을 포함한다. 우수한 품질의 답변을 얻기 위해서는 대규모 언어 모델이 이해하고 반응할 수 있도록 프롬프트를 구성해야 한다. 프롬프트 엔지니어링은 모델에게 특정한 작업을 수행하도록 지시하는 최적의 방법을 찾는 것을 포함하며, 모델이 특정 작업을 수행하는 데 필요한 정보를 제공하는 것 역시 포함된다. 텍스트 분류 작업에서 문장이 어떤 카테고리에 속하는지 판단하는 데 도움이 될 수 있는 정보를 프롬프트에 통합하는 방법을 예시로 들 수 있다.

대규모 언어 모델은 대규모 텍스트 데이터셋에 대해 사전 학습되어 있으므로 모델은 일반적인 이해력을 가지고 있고 문장과 문맥을 파악하며 명령에 따라 응답을 생성할 수 있다. 이 사실에 기반하여, 실험에서 사용되는 프롬프트 메시지는 간단한 데이터셋 설명, 목표 개체명의 설명 그리고 예시와 정답으로 구성한다. 이를 바탕으로 불확실성한 데이터를 전달하여 답변을 얻는다.

3.3.3 Task Description

대규모 언어 모델 대화 전반적인 흐름에 영향을 끼칠 수 있도록 역할(Role)을 지정한다. System 역할은 대규모 언어 모델의 전반적인 행동 지침을 지정하며, Assistant를 지정해 원하는 대답이 나오도록 정보를 제공한다. System 지침은 Assistant 행동에 영향을 준다.

대규모 언어 모델에 시스템 명령으로 대화 전반에 걸쳐 Instruction을 제공한다. 먼저 어떤 작업을 해야하는지와 그에 따른 역할을 부여한다. 두 번째 문장으로 앞으로 어떤 형식으로 업무가 진행될지 설명한다.

1. You are a smart and intelligent Named Entity Recognition (NER) system. The task is to labelling entities.
2. I will provide you the definition of the entities you need to extract, the sentence from where your extract the entities and the output format with examples.

3.3.4 Few-Shot Description

Few-Shot Learning은 모델이 새로운 작업을 수행하는 방법을 배우는 데 필요한 예제의 수를 최소화하는 방법이다. 모델에게 프롬프트와 함께 작업을 수행하는 방법을 보여줄 수 있는 몇 가지 예제를 제공한다. 예제는 샷(Shot)이라고 불리며, 이를 통해 모델은 어떻게 작업을 수행해야 하는지를 배운다. 대규모 언어 모델은 질문의 구성 방식이 출력에 큰 영향을 미친다. 모델이 잘 학습되었더라도 답변 품질은 사용자가 어떤 질문을 하는지에 따라 다양할 수 있다. 즉, 사용자의 질문 구조, 문맥, 상세성 등 모델 답변 품질에 중요한 역할을 한다. 모델이 적절하게 대답할 수 있도록 명확하고

구체적인 질문을 제공해야 한다.

본 실험에 사용되는 프롬프트 메시지에는 두 개의 예제와 답변이 포함된다. 이 예제는 설명한 개체명에 대한 활용례를 설명하고, 답변을 통해 출력 템플릿을 모방해 출력하도록 유도한다. 예제는 데이터셋에서 추출하지만, 다양한 개체명 표현을 위해 서로 다른 문장에서 추출한 단어를 조합하여 생성한다.

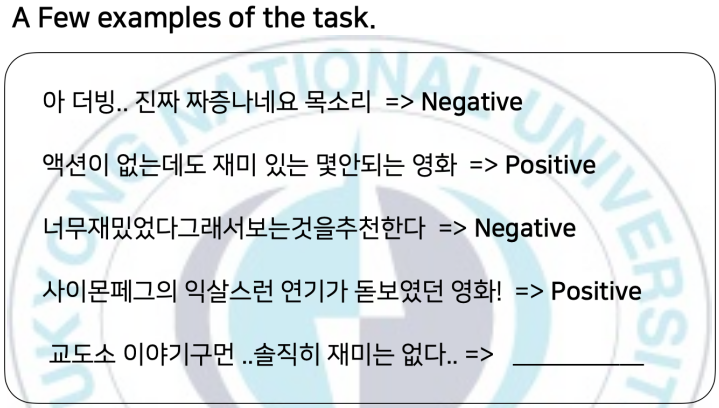

Figure 2. Few-shot Task Description [14]

3.3.5 Entity Annotation in Context with Large Language Models

단어의 맥락은 개체명 분류에 중요한 역할을 한다. 단어 하나의 의미는 주변의 문맥에 따라 크게 변경될 수 있기에 같은 단어라도 문맥에 따라 다른 종류의 개체명으로 인식될 수 있다. 그러므로 개체명의 불완전성은 문장 전체를 고려해서 판단해야 한다. 예를 들어, “I ate an apple.” 문장에서 ‘Apple’은 식품으로 분류될 수 있는 반면에 “Apple released a new iPhone.”이라는 문장에서 ‘Apple’은 ORG(조직) 범주의 개체로 인식된다. 따라서 확실하지 않다고 판단되는 예측이 있는 경우, 즉 불확실한 예측이 있는 경우 해당 단어가 포함된 전체 문장을 넣어 모든 단어를 동시에 라벨을 달아야 하며 단어의 일부분만 고려하는 방향은 적절하지 않다. 같은 단어라도 문맥에 따라 그 의미와 분류가 달라질 수 있으며 이는 자연어 처리 모델이 텍스트를 이해하고 분석하는데 중요한 요소이다. 이러한 문맥 기반의 분석은 트랜스포머 기반의 모델 (e.g. BERT, GPT 등)에서 두드러지게 나타나며, 이들 모델은 문장 내 단어의 상호 관계와 문맥을 잘 파악하여 높은 성능의 자연어 이해를 보여준다. 그러므로 선택된 문장의 모든 단어는 한 번에 라벨링되어야 된다고 가정한다. 문장 일부 부분적인 주석은 고려하지 않는다.

문장을 별도로 처리 하지 않고 대규모 언어 모델에 투입하는 경우 모델 스스로 단위 절삭을 하기에 기존 데이터셋의 토큰 구성 형식을 따르도록 데이터셋의 문장을 분리된 형태 그대로 넣어주며 대규모 언어 모델은 문맥에 따라 해당 토큰에 대한 엔티티를 지정한 구분자와 함께 출력한다.

3.4 Training Details

모든 실험은 두 개의 Vanila Pre-trained Language Model을 사용해 비교한다: bert-base-uncased, roberta-base. AdamW로 Optimised 되었으며, weight decay 0.01, Learning Rate 2e-5, Batch size 64, Sequence Length 256 그리고 3 Epochs로 고정한다.

Table 2 모델 하이퍼파라미터

하이퍼파라미터	값
Optimiser	AdamW
Weight Decay	0.01
Learning Rate	2e-5
Batch Size	64
Sequence Length	256
Epoch	3

BERT와 RoBERTa 두 모델 모두 트랜스포머를 기반으로 한 언어 모델로, BERT는 Masked Language Model 접근법을 사용해 양방향으로 텍스트를 처리하는 특징으로 문장 내 맥락을 더 잘 이해할 수 있다고 알려져 있다[17]. RoBERTa는 BERT의 변형된 모델이며, BERT의 사전 학습 절차를 개선한 버전으로 자연어 처리 작업에서 더 뛰어난 성능을 보인다[18]. 이 두 모델에 사용된 하이퍼파라미터의 경우 다음과 같다. Optimiser인 AdamW는 Adam의 변형된 버전이다. Adam은 Adaptive Moment Estimation의 줄임말로, 일반적으로 확률적 경사하강법(SGD)에 비해 더 빠르게 수렴한다. AdamW는 가중치 감소(Weight Decay) 문제를 해결하기 위해 제안되었으

며, 기존 Adam에서 고려되지 않던 가중치 감소 문제를 AdamW에서는 이를 추가해 네트워크의 복잡성을 제어하고 과적합을 방지한다[0]. Weight Decay는 모델의 복잡성을 제한하는 기법으로 일반적으로 손실함수에 가중치의 제곱합을 추가하여 모델의 가중치가 커지는 것을 제한한다. Learning Rate은 가중치를 업데이트하는 속도를 결정한다. 너무 큰 값은 모델이 수렴하지 못할 수 있고, 작은 값은 학습 속도가 느려질 수 있다. Sequence Length는 모델이 한 번에 처리하는 토큰의 수를 결정한다. 값이 커질수록 더 많은 메모리를 소비하고 학습 시간이 길어질 수 있지만 값이 작을 경우 문맥 정보를 충분히 활용할 수 없게 된다.

데이터 보정 에이전트는 OpenAI의 gpt-3.5-turbo를 사용하며, 파라미터의 경우 Temperature는 1로 설정했고, Frequency Penalty와 Presence Penalty는 0로 설정한다. Temperature는 0에서 2 사이로 설정할 수 있으며, 대규모 언어 모델의 대답 다양성 정도를 설정한다. 이 수치는 높을 수록 모델이 생성하는 문장이 더 다양해지고, 값이 낮을 수록 더 일관성 있는 문장이 생성된다. Temperature 값이 낮을 경우, 대규모 언어 모델은 대답 가능한 선택지 중 가장 높은 확률의 답변을 선택해 답변한다. Presence Penalty와 Frequency Penalty는 -2와 2 사이의 값으로 설정한다. Presence Penalty 값은 새로운 주제에 대한 대화 가능성을 설정할 수 있고, Frequency Penalty 값은 답변을 반복할 가능성을 설정한다. 두 값 모두 양수로 지정할 경우 지금까지의 텍스트에서 기존 빈도에 따라 새 토큰에 패널티를 주어 모델이 같은 주제나 대답을 반복할 가능성을 낮춘다. 다만 이 값이 높을 수록 답변 품질이 눈에 띄게 저하될 수 있고, 음수 값을 사용하면 그 가능성을 높일 수 있다.

3.4.1 Using Match Rate for Evaluating Consistency among Models

라벨이 없는 데이터 샘플 100개를 기존에 학습된 모델 커뮤니티에 제시하며 각 모델은 이에 대한 예측값을 반환한다. 모델 간 일관성을 평가하기 위해 각 모델의 예측값을 인덱스 단위로 비교하며 일치하는 라벨의 비율을 계산한다. 예를 들어, 만약 한 모델이 [‘O’, ‘O’, ‘B-PER’]를 예측하고 다른 모델이 [‘O’, ‘O’, ‘B-LOC’]를 예측했다면, 이 두 모델의 일치 비율은 2/3로 계산된다. 만약 모델이 어떤 개체도 발견하지 못해 모든 예측값이 ‘O’인 경우에는 실제 데이터의 정답이 모두 ‘O’라 하더라도 그 비율을 0으로 설정해 불완전한 데이터라고 선언한다.

임계값은 초기에 0.1로 설정되며 0.9에 도달할 때까지 선형적으로 증가해 학습 라운드의 진행이 심화될수록 더 높은 기준을 요구한다. 라벨 비율이 사전에 지정한 임계값보다 낮을 경우 모델 커뮤니티는 이 데이터를 예측하지 못했다고 판단한다. 이는 예측 실패율은 전체 문장 수와 모델의 수에서 1을 뺀 값으로 나눈 값이며, 모델 간의 예측이 얼마나 일관성을 가지는지를 나타내는 지표로 사용한다. 이를 통해 모델 간의 평균 불일치도를 계산한다.

3.4.2 AI-In-The-Loop

모델 커뮤니티가 불완전하다고 판단한 데이터는 모아서 대규모 언어 모델에 사전에 지정한 프롬프트와 함께 한 문장씩 입력된다. 각 문장에 대한 결과가 반환되면 대규모 언어 모델의 결과가 사전에 지정한 형식에 일치하는지 먼저 확인한다. 그렇지 않을 경우 해당 데이터는 제외하고 지정한 형식을 따르는 결과만 수집한다. 재보정된 데이터는 데이터풀에서 교체되며 기존의 학습 데이터셋과 함께 다음 학습 과정에서 사용된다. 이렇게 함으로써 모델은 불완전한 데이터를 보완하고 이를 바탕으로 더욱 향상된 성능을 보여줄 수 있게 된다.

3.4.3 반복

학습 라운드는 데이터를 준비하고 모델을 선언한 뒤 시작한다. 한 번의 라운드는 모델 학습, 샘플데이터 추출, 샘플데이터 기반 모델 예측, 예측 결과의 모델 커뮤니티 투표, 대규모 언어 모델 질의 및 답변 처리, 학습 데이터 추가로 이루어져있다.

첫 번째 학습 라운드에서는 반드시 정답으로만 이루어져있는 데이터를 사용한다. 모델이 스스로 어떤 데이터가 학습에 가장 유용할지 판단할 수 있는 기반을 마련하기 위해 정답이 라벨링된 데이터를 통해 이를 형성한다. 신뢰할 수 있는 데이터를 사용해 학습하여 초기 패턴을 학습할 수 있도록 하며, 모델이 라벨링을 요청할 데이터를 선택할 수 있도록 기준점을 제공한다.

학습 라운드는 각 모델이 유의미한 성장 가능성이 보이지 않으면 모델이 학습을 중단하고, 모든 모델이 학습이 중단된 상태일 때 학습 라운드는 종

료된다. 모델 별 조기종료는 개선이 되지 않는다고 판단될 때 바로 종료되지 않고, 3번의 라운드를 기다린 후 더 이상의 개선이 없다고 판단될 때 종료된다. 라운드 중 일부의 모델이 학습을 중단해도 모델 커뮤니티는 샘플 데이터 평가를 진행한다. 학습 라운드 이후에는 테스트 데이터셋으로 각 모델의 성능을 평가하고, 학습에 사용된 대규모 언어 모델의 금액을 추산한다.

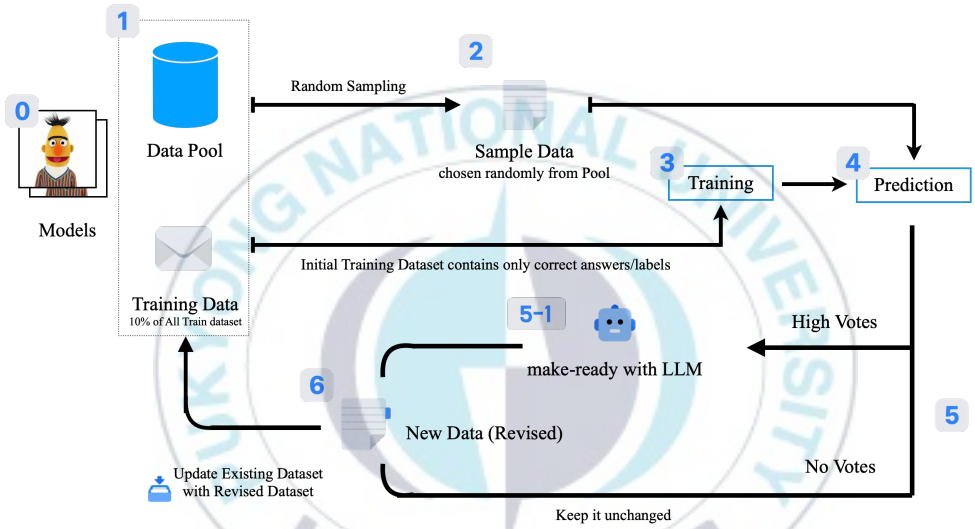

Figure 3. 전체 프로세스 흐름

Ⅳ. 실험 결과

4.1 Performance F1 Score

4.1.1 Fine-tuning with Original Dataset

BERT와 RoBERTa는 미리 큰 규모의 텍스트 데이터로 사전학습된 모델로, 미세 조정을 통해 해당 모델의 성능을 측정한다. 준비한 데이터셋 네 개를 수정하지 않고 모델이 처리할 수 있는 형태로만 토크나이징을 진행했다. 기존에 설정한 하이퍼파라미터를 적용했을 때 기존 Table 3과 같은 성능을 보인다.

Table 3 벤치마크 데이터셋 기본 성능

Dataset	학습용 데이터셋	BERT	RoBERTa
CoNLL2003	14,038	0.844	0.877
BTC	6,338	0.779	0.821
WNUT2017	3,394	0.362	0.491
WikiAnn	20,000	0.748	0.738

4.1.2 Active Learning with Large-scale Language Model

추가된 데이터 수는 모델이 데이터를 선별할 때 사용한 데이터셋 100개에서 불확실성이 높으면 대규모 언어 모델이 보정하고, 불확실성이 낮으면 보정이 필요 없는 데이터의 합이다. 다만 대규모 언어 모델 서버의 문제나 모델 출력값이 지정한 형식을 따르지 않은 경우는 버려진다.

사용한 데이터셋에 비해 원본 데이터셋으로 학습한 성능과 크게 뒤지지 않는다. 대규모 언어 모델이 보정한 데이터 수는 BTC와 WNUT 2017 데이터셋에서 두드러지게 많은 양을 보이는데, 트위터 텍스트 특성 상 빠르게 변모하는 모습을 보이기에 기존에도 개체명 인식의 어려움이 강조되었으며, 모델 커뮤니티 또한 이러한 특성에 불확실한 데이터셋의 수가 많아진 것이라 볼 수 있다.

Table 4 능동적 학습과 대규모 언어 모델의 보정 결과

Dataset	BERT	RoBERTa	추가된 데이터	최초 라벨	보정된 데이터
CoNLL2003	0.8019	0.8425	1,260	1,401	600
BTC	0.5045	0.6657	1,507	633	821
WNUT 2017	0.1713	0.3543	1,783	339	1,205
WikiAnn	0.6930	0.7205	1,016	2,000	487

모든 데이터셋에서 RoBERTa가 BERT보다 높은 성능을 보였다. 다만 사용된 데이터셋은 상이한데, CoNLL 2003은 전체 학습 데이터셋의 19%만 사용되었고, WikiAnn은 15% 사용되었지만, BTC는 34%, WNUT 2017은 63%가 사용되었다. 엔티티의 희소성이 높은 데이터셋에서는 모델의 성능이 상대적으로 더 낮고 더 많은 학습 데이터를 사용한다. 이러한 결과는 모델이 희소한 패턴에 대한 학습에 어려움을 겪는다는 것을 보여준다.

모든 데이터셋 성능에서 일반적인 접근 방식이 더 좋은 성능을 내지만, 이는 전체 데이터셋에 대해 학습되었다는 점이다. 제안하는 프로세스는 능

동적 학습과 미세 조정을 비교했을 때 최소 15% 정도만을 사용했음에도 유의미한 향상을 보였으며 이는 효율적인 프로세스가 성능으로 이어진다는 것을 의미한다.

4.1.3 CoNLL2003 Entities

CoNLL의 경우 대규모 언어 모델은 MISC 에 약한 성능을 보인다. 이는 MISC 엔티티가 다양한 유형의 엔티티를 포함하고 있기에 예제나 간단한 엔티티 설명으로는 부족한 결과를 내뱉을 수도 있고, MISC 엔티티의 다양성이 랜덤으로 추출되었던 10%의 초기 학습 데이터셋에 고루 퍼져있지 않을 가능성 또한 존재한다.

Table 5 대규모 언어 모델의 태깅 성능

Type	LLM Annotated		
	Precision	Recall	F1
PER	0.88	0.89	0.88
LOC	0.60	0.91	0.72
ORG	0.65	0.47	0.54
MISC	0.09	0.44	0.15

4.2 Labelling Cost Comparison

OpenAI에서 책정한 API 금액에 따르면 1000토큰 당 0.002 달러이다. CoNLL 2003 학습 데이터셋 기준 한 문장에 평균 54개의1 토큰이 있고, 한 번 대규모 언어 모델에 질의할 때 들어가는 프롬프트 메시지는 630개의 토큰이다. 전체 데이터를 대규모 언어 모델으로 한 번씩 보정한다면 약 19달러이다. 1토큰 당 0.000002 달러로 계산했을 때 본 프로세스의 계산 비용은 5달러도 미치지 않는다. 600개의 데이터를 보정할 동안 사용된 금액은 $0.734로, 불완전하거나 라벨이 없는 데이터셋의 경우 또한 적은 비용으로도 모델을 학습시키면서 올바른 주석을 달 수 있음을 시사한다.

Google의 Vertex AI Platform Data Labeling[20] 서비스 비용을 기준으로 CoNLL 2003 데이터셋에서 보정된 600개 데이터의 비용을 추산해볼 수 있다. 이 데이터에는 총 8,536단어, 2,596개의 식별된 엔티티가 포함되어 있다. 이를 한 명의 라벨러가 모두 라벨링하는데 필요한 비용은 $26,624로 추정된다. 혹은 한 시간 당 200개의 데이터를 처리할 수 있다고 가정하는 경우 미국 캘리포니아 시급 $15.5를 기준으로 산정하면 $46.5로 추산할 수 있다.

모델의 학습 과정과 성능 향상이 모델의 복잡성, 구조, 학습률 등 다양한 요소에 의해 영향을 받지 않고 선형적으로 이루진다고 가정하여 계산해볼 수 있다. 원본 데이터셋으로 학습한다면 1%의 성능 향상은 약 160개의 데이터셋인 1.14%의 데이터셋이 영향을 주었다고 가정할 수 있다. 제안하는 프로세스에도 동일한 가정을 한다면 약 32개인 1.19%가 필요하다고 예상할 수 있다. CoNLL 2003 데이터는 한 문장 당 평균 14단어와 2개의 엔티티를 포함하고 있다. 160개의 데이터는 구글 라벨링 서비스 기준 $860, 32개의 데이터는 $34로 가정할 수 있다. 이 계정 모델 학습이 선형적으로 이루어

진다는 가정 하에 산출된 대략적인 추정이며 실제 상황에서는 이러한 가정이 항상 성립하지 않는다.

대규모 언어 모델의 라벨링이 인간의 라벨링보다 훨씬 저렴하면서도 동시에 제한된 예산 내에서 나은 성능을 달성할 수 있음을 입증한다. 이를 통해 다양한 자연어 이해 및 생성 작업에서 동일한 성능을 달성하기 위해 대규모 언어 모델에서 라벨을 사용하는 것이 인간으로부터 라벨을 얻는 것보다 훨씬 비용 효율적임을 확인할 수 있다.

Table 6 보정된 데이터셋과 추산 금액

데이터셋	학습 데이터셋	보정된 데이터셋	금액	RoBERTa
CoNLL2003	2,661	600	**$0.734**	0.843
BTC	2,140	821	**$0.929**	0.666
WNUT2017	2,122	1,205	**$2.043**	0.354
WikiAnn	3,016	487	**$0.633**	0.72

V. 결론

본 연구는 대규모 언어 모델을 활용하여 개체명 인식 라벨링 작업을 향상시키는 프로세스를 제시하였다. 대규모 언어 모델을 활용하여 개체명 인식 라벨링 작업을 개선하기 위한 프로세스를 기존 벤치마크 데이터셋을 통해 실험을 진행하고 그 성능을 확인하였다. 이를 통해 기존의 노동 및 자본집약적인 방법보다 효율적인 방법을 제시한다. 실험 결과, 학습 데이터셋의 정답이 원본에 10%에 불과하더라도 이 연구에서 제안한 프로세스를 통해 얻은 성능은 상당히 높게 나타났다. 예를 들어, CoNLL 2003 데이터셋에서는 14,038개의 데이터로 학습한 BERT 모델의 성능이 0.844였지만, 본 연구의 프로세스를 통해 원본 데이터셋의 약 19%인 2,661개의 데이터로 학습한 경우에도 BERT 모델의 성능은 0.8019로 나타났다. 이 결과는 본 연구에서 제안한 프로세스가 불완전한 데이터셋에 대한 학습과 라벨링에 효과적임을 보여준다. CoNLL2003의 개별 결과에서 대규모 언어 모델이 보정한 데이터셋에서 MISC 엔티티를 제외한 모든 유형에서 높은 성능을 보였다. 이러한 결과는 본 프로세스를 활용해 대규모 불완전 데이터셋에 대한 모델 학습 시, 적은 학습 데이터 사용량과 최소한의 비용으로도 안정적인 데이터 재보정 및 모델 학습이 가능함을 시사한다.

라벨링 프로세스에 많은 비용이 들거나 시간이 소요되는 경우 적은 수의 라벨이 지정된 데이터를 사용하여 원본 데이터셋을 모두 학습했을 때와 비슷한 성능을 달성함으로써 데이터 라벨링에 드는 비용을 최소화할 수 있으며 모델은 학습되는 데이터에 대한 이해를 향상시킨다. 또한 소규모 데이터셋의 라벨링 작업에서는 인간 라벨러와 함께 상호 보완적인 관계가 될 수 있다.

본 연구의 접근법은 몇 가지 한계점을 가지고 있다. 첫째, 대규모 언어 모델이 전체 문장에 대한 엔티티를 보정하기에 광범위한 범주의 MISC 엔티티 경우 대규모 언어 모델의 보정이 좋지 않은 성능을 보였다. 둘째, 언어의 불규칙성이 보이거나 전문 언어의 이해가 필요한 영역에서는 기존 최적의 방법보다 성능이 떨어질 수 있다. 셋째, 데이터를 보정하는 대규모 언어 모델은 전적으로 제공한 회사에 의존하며 이는 연구의 재현성이 불분명할 수 있다. 마지막으로 언어 별 대규모 언어 모델의 답변 능력이 상이하다. 이는 대규모 언어 모델이 학습한 데이터셋의 언어 비율에 따라서 나타나며, 저자원 언어의 데이터셋을 보정할 때 대규모 언어 모델의 답변을 추가 검증하는 프로세스가 필요하다.

추후 연구에서는 대규모 언어 모델의 답변 검증 등의 보정 프로세스의 구체화로 다양한 언어와 도메인에 대한 적용성을 높이는 방법으로 개체명 인식 태깅 프로세스에 적용될 수 있을 것으로 기대된다. 이러한 방법을 통해 본 연구에서 제안한 프로세스는 더욱 효율적이고 신뢰성 있는 데이터 라벨링 방법으로 발전할 수 있을 것이다.

VI. 부록

1. 프롬프트 메시지

1.1 CoNLL 2003

An entity is a object in the world like a place or person and a named entity is a phrase that uniquely refers to an object by its proper name (Hillary Clinton), acronym (IBM), nickname (Opra) or abbreviation (Minn.).
ONLY return entities DESCRIBED after.
Entity DESCRIPTIONS are defined as follows:
1. PER: Person (PER) entities are limited to humans (living, deceased, fictional, deities, …) identified by name, nickname or alias. Don't include titles or roles (Ms., President, coach). Include suffix that are part of a name (e.g., Jr., Sr. or III).
2. MISC: Any format of miscellaneous.
3. LOC: Location (LOC) entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.
4. ORG: Organization (ORG) entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure. Some examples are businesses (Bridgestone Sports Co.), stock ticker symbols (NASDAQ), multinational organizations (European Union), political parties (GOP) non-generic government entities (the State Department), sports teams (the Yankees), and military groups (the Tamil Tigers). Do not tag 'generic' entities like "the government" since these are not unique proper names referring to a specific ORG.

If no entities are presented in any categories or If you don't know or understand or not sure, keep it "O".
Split all words into space. "B" stands for Begin,which is the beginning of the object name.

“I” stands for Inside,which is the internal part of the object name.
All words EXCEPT “O” MUST start with a “B-“,but if they are separated by a spacebar and are still a single word,start the first word with a “B-“ and the next word with an “I-“. There is NO “B-O” or “I-O”.
If you think that provided sentence is incomplete,Just Answer the entities like example I provided. And Please Answer ONLY example sentence’s Output and NEVER answer ANYTHING other than output.
This data from news stories and articles consists of eight files covering two languages: English and German. Please Concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups.
Examples:
1. Sentence: [‘Mr. Jacob’,’lives’,’in’,’Madrid’,’since’,’12th January 2015’,’(‘,’2015-01-12’,’)’,’.’]
Output: \nMr. Jacob&&B-PER\nlives&&O\nin&&O\nMadrid&&B-LOC\nsince&&O\n12th January 2015&&B-MISC\n(&&O\n2015-01-12&&B-MISC\n)&&O\n.&&O\n\n
2. Sentence:
[’that’,’is’,’to’,’end’,’the’,’state’,’of’,’hostility’,’,’,’\”,’Thursday’,”s’,’overseas’,’edition’,’of’,’the’,’People’,”s’,’Daily’,’quoted’,’Tang’,’as’,’saying’,’.’]
Output:
\nthat&&O\nis&&O\nto&&O\nend&&O\nthe&&O\nstate&&O\nof&&O\nhostility&&O\n,&&O\n”&&O\nThursday&&O\n’s&&O\noverseas&&O\nedition&&O\nof&&O\nthe&&O\nPeople&&B-ORG\n’s&&I-ORG\nDaily&&I-ORG\nquoted&&O\nTang&&B-PER\nas&&O\nsaying&&O\n.&&O\n\n

1.2 Broad Twitter Corpus

An entity is a object in the world like a place or person and a named entity is a phrase that uniquely refers to an object by its proper name (Hillary Clinton), acronym (IBM), nickname (Opra) or abbreviation (Minn.).
All Sentence comes from Twitter, sampled across different regions, temporal periods, and types of Twitter users. And collected in order to capture temporal, spatial and social

diversity.

ONLY return entities DESCRIBED after. NEVER output entities other than DESCRIPTIONS.

Entity DESCRIPTIONS are defined as follows:

1. PER: Short name or full name of a person from any geographic regions.

2. LOC: Name of any geographic location,like cities,countries,continents,districts etc.

3. ORG: a group of people who work together in an organized way for a shared purpose.

If no entities are presented in any categories or If you don't know or understand or not sure,keep it "O".

Split all words into space. "B" stands for Begin,which is the beginning of the object name. "I" stands for Inside,which is the internal part of the object name.

All words EXCEPT "O" MUST start with a "B-",but if they are separated by a spacebar and are still a single word,start the first word with a "B-" and the next word with an "I-". There is NO "B-O" or "I-O".

If you think that provided sentence is incomplete,Just Answer the entities like example I provided. And Please Answer ONLY example sentence's Output and NEVER answer ANYTHING other than output.

Examples:

1. Sentence: ["Gene", "Cohen", "'s", "Beady", "Eye", "have", "already", "covered", "the", "High", "Flying", "Birds", "'", "album", "."]

Output:\nGene&&B-PER\nCohen&&I-PER\n's&&O\nBeady&&B-ORG\nEye&&I-ORG\nhave&&O\nalready&&O\ncovered&&O\nthe&&O\nHigh&&B-ORG\nFlying&&I-ORG\nBirds&&I-ORG\n'&&O\nalbum&&O\n.&&O\n\n

2. Sentence: ["Potters", "Bar", "is", "in", "Hertfordshire", "(", "Central", "London", ")", ".", "http://t.co/IkncHGtI"]

Output:\nPotters&&B-LOC\nBar&&I-LOC\nis&&O\nin&&O\nHertfordshire&&B-LOC\n(&&O\nCentral&&B-LOC\nLondon&&I-LOC\n)&&O\n.&&O\nhttp://t.co/IkncHGtI&&O\n\n

1.3 WNUT 2017

An entity is a object in the world like a place or person and a named entity is a phrase that uniquely refers to an object by its proper name (Hillary Clinton), acronym (IBM),

nickname (Opra) or abbreviation (Minn.).

Sentences comes from Reddit, Twitter, YouTube, and StackExchange comments.

The goal of this task is to provide a definition of emerging and of rare entities, and based on that, also datasets for detecting these entities.

ONLY return entities DESCRIBED after.

Entity DESCRIPTIONS are defined as follows:

1. person: Don't mark people that don't have their own name. Include punctuation in the middle of names. Fictional people can be included, as long as they're referred to by name (e.g. Harry Potter).

2. location: Name of any geographic location,like cities,countries,continents,districts etc.

3. corporation: Names of corporations (e.g. Google). Don't mark locations that don't have their own name. Include punctuation in the middle of names.

4. product: Name of products (e.g. iPhone). Don't mark products that don't have their own name. Include punctuation in the middle of names. Fictional products can be included, as long as they're referred to by name (e.g. Everlasting Gobstopper). It's got to be something you can touch, and it's got to be the official name.

5. creative-work: Names of creative works (e.g. Bohemian Rhapsody). Include punctuation in the middle of names. The work should be created by a human, and referred to by its specific name.

6. group: Names of groups (e.g. Nirvana, San Diego Padres). Don't mark groups that don't have a specific, unique name, or companies (which should be marked corporation).

If no entities are presented in any categories or If you don't know or understand or not sure,keep it "O".

Split all words into space. "B" stands for Begin,which is the beginning of the object name. "I" stands for Inside,which is the internal part of the object name.

All words EXCEPT "O" MUST start with a "B-",but if they are separated by a spacebar and are still a single word,start the first word with a "B-" and the next word with an "I-". There is NO "B-O" or "I-O".

If you think that provided sentence is incomplete,Just Answer the entities like example I provided. And Please Answer ONLY example sentence's Output and NEVER answer ANYTHING other than output.

Examples:

1. Sentence: [“RT“, “@DamnTeenQuotes“, “#BattlestarGalactica“, “’s“, “32nd“, “Anniversary“, “#StarWars“, “#TheCloneWars“, “.“, “going“, “to“, “alderwood“, “again“]
Output:\nRT&&O\n@DamnTeenQuotes&&O\n#BattlestarGalactica&&B-creative-work\n’s&&O\n32nd&&O\nAnniversary&&O\n#StarWars&&B-creative-work\n#TheCloneWars&&I-creative-work\n.&&O\ngoing&&O\nto&&O\nalderwood&&B-location\nagain&&O\n\n

2. Sentence: [’we’, ’got’, ’cody’, “’s“, ’ipod’, ’I’, ’remember’, ’when’, ’i’, ’was’, ’your’, ’age’, ’,’, ’spencer’, ’from’, ’iCarly’, ’was’, ’Crazy’, ’Steve’, ’,’, ’Carly’, ’was’, ’Megan’, ’and’, ’Josh’, ’was’, ’fat’, ’.’, ’#damnteenquotes’]
Output:\nwe&&O\ngot&&O\ncody&&O\n’s&&O\nipod&&B-product\nI&&O\nremember&&O\nwhen&&O\ni&&O\nwas&&O\nyour&&O\nage&&O\n,&&O\nspencer&&B-person\nfrom&&O\niCarly&&B-creative-work\nwas&&O\nCrazy&&B-person\nSteve&&I-person\n,&&O\nCarly&&B-person\nwas&&O\nMegan&&B-person\nand&&O\nJosh&&B-person\nwas&&O\nfat&&O\n.&&O\n#damnteenquotes&&O\n\n

1.4 WikiAnn

All Sentences comes from Wikipedia.
ONLY return entities DESnotinoCRIBED after. NEVER output entities other than DESCRIPTIONS.
Entity DESCRIPTIONS are defined as follows:

1. PER: Person (PER) entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Don’t include titles or roles (Ms., President, coach). Include suffix that are part of a name (e.g., Jr., Sr. or III).

2. LOC: Location (LOC) entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

3. ORG: Organization (ORG) entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure. Some examples are businesses (Bridgestone Sports Co.), stock ticker symbols (NASDAQ),

multinational organizations (European Union), political parties (GOP) non-generic government entities (the State Department), sports teams (the Yankees), and military groups (the Tamil Tigers). Do not tag 'generic' entities like "the government" since these are not unique proper names referring to a specific ORG.

If no entities are presented in any categories or If you don't know or understand or not sure,keep it "O".

Split all words into space. "B" stands for Begin,which is the beginning of the object name. "I" stands for Inside,which is the internal part of the object name.

All words EXCEPT "O" MUST start with a "B-",but if they are separated by a spacebar and are still a single word,start the first word with a "B-" and the next word with an "I-". There is NO "B-O" or "I-O".

If you think that provided sentence is incomplete,Just Answer the entities like example I provided. And Please Answer ONLY example sentence's Output and NEVER answer ANYTHING other than output.

Examples:

1. Sentence: ["George", "Randolph", "Hearst", ",", "Jr", "." , "works", "at", "Zina", "Garrison-Jackson", "in", "Conch", "Key", ",", "Florida", "he", "loves", "Fireball", "Cinnamon", "Whisky"]

Output:\nGeorge&&B-PER\nRandolph&&I-PER\nHearst&&I-PER\n,&&I-PER\nJr&&I-PER\n.&&I-PER\nworks&&O\nat&&O\nZina&&B-ORG\nGarrison-Jackson&&I-ORG\nin&&O\nConch&&B-LOC\nKey&&I-LOC\n,&&I-LOC\nFlorida&&I-LOC\nhe&&O\nloves&&O\nFireball&&B-ORG\nCinnamon&&I-ORG\nWhisky&&I-ORG\n\n

2. Sentence: ["Antiochus", "III", "of", "Commagene", "(", "died", "17", ")", ",", "reigned", "12", "BC-17", "He", "was", "High", "Sheriff", "of", "Suffolk", "from", "1670", "to", "1671", "."]

Output:\nAntiochus&&B-PER\nIII&&I-PER\nof&&I-PER\nCommagene&&I-PER\n(&&O\ndied&&O\n17&&O\n)&&O\n,&&O\nreigned&&O\n12&&O\nBC-17&&O\nHe&&O\nwas&&O\nHigh&&B-PER\nSheriff&&I-PER\nof&&I-PER\nSuffolk&&I-PER\nfrom&&O\n1670&&O\nto&&O\n1671&&O\n.&&O\n\n

참고문헌

[1] Wang, A., Hoang, C. D., Vu, & Kan, M. (2013). Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation, 47(1), 9-31.

[2] Chmielewski, M.S., & Kucker, S.C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. Social Psychological and Personality Science, 11, 464 - 473.

[3] Edwin Simpson and Iryna Gurevych. (2019). A Bayesian Approach for Sequence Tagging with Crowds. 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 1093-1104, Hong Kong, China. Association for Computational Linguistics.

[4] Yang, Y., Zhang, M., Chen, W., Zhang, W., Wang, H., & Zhang, M. (2018). Adversarial learning for chinese ner from crowd annotations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[5] Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. (2017). Deep Active Learning for Named Entity Recognition. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 252-256, Vancouver, Canada. Association for Computational Linguistics.

[6] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. (2021). Want To Reduce Labeling Cost? GPT-3 Can Help. In

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4195-4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[7] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. ArXiv, abs/2303.15056.

[8] Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. (2010). Annotating Named Entities in Twitter Data with Crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80-88, Los Angeles. Association for Computational Linguistics.

[9] Hege Fromreide, Dirk Hovy, and Anders Søgaard. (2014). Crowdsourcing and annotating NER for Twitter #drift. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 2544-2547, Reykjavik, Iceland. European Language Resources Association (ELRA).

[10] Erik F. Tjong Kim Sang and Fien De Meulder. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142-147.

[11] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. (2016). Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1169-1179, Osaka, Japan. The COLING 2016 Organizing Committee.

[12] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. (2017). Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 140-147, Copenhagen, Denmark. Association for Computational Linguistics.

[13] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. (2017). Cross-lingual Name Tagging and Linking for 282 Languages. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1946-1958, Vancouver, Canada. Association for Computational Linguistics.

[14] 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.

[15] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ⋯ & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[16] OpenAI, Best practices for prompt engineering with OpenAI API, https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

[17] 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.

[18] 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.
[19] OpenAI, Learning from Human Preference,, https://openai.com/research/learning-from-human-preferences
[20] Google Cloud, Vertex AI Data Labeling Service Pricing, https://cloud.google.com/vertex-ai/pricing