Vol.12, No.11 pp.481~492 https://doi.org/10.3745/KTSDE.2023.12.11.481
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

A Study on Dataset Generation Method for Korean Language Information Extraction from Generative Large Language Model and Prompt Engineering

Jeong Young Sang† · Ji Seung Hyun†† · Kwon Da Rong Sae†††

ABSTRACT

This study explores how to build a Korean dataset to extract information from text using generative large language models. In modern society, mixed information circulates rapidly, and effectively categorizing and extracting it is crucial to the decision-making process. However, there is still a lack of Korean datasets for training. To overcome this, this study attempts to extract information using text-based zero-shot learning using a generative large language model to build a purposeful Korean dataset. In this study, the language model is instructed to output the desired result through prompt engineering in the form of "system"-"instruction"-"source input"-"output format", and the dataset is built by utilizing the in-context learning characteristics of the language model through input sentences. We validate our approach by comparing the generated dataset with the existing benchmark dataset, and achieve 25.47% higher performance compared to the KLUE-RoBERTa-large model for the relation information extraction task. The results of this study are expected to contribute to AI research by showing the feasibility of extracting knowledge elements from Korean text. Furthermore, this methodology can be utilized for various fields and purposes, and has potential for building various Korean datasets.

Keywords : Large Language Model, Prompt Egineering, Zero-shot Learning, Dataset Generation, Information Extraction

생성형 대규모 언어 모델과 프롬프트 엔지니어링을 통한 한국어 텍스트 기반 정보 추출 데이터셋 구축 방법

정 영 상† · 지 승 현†† · 권 다 롱 새†††

요 약

본 연구는 생성형 대규모 언어 모델을 활용하여 텍스트에서 정보를 추출하기 위한 한글 데이터셋 구축 방법을 탐구한다. 현대 사회에서는 혼합된 정보가 빠르게 유포되며, 이를 효과적으로 분류하고 추출하는 것은 의사결정 과정에 중요하다. 그러나 이에 대한 학습용 한국어 데이터셋은 아직 부족하다. 이를 극복하기 위해, 본 연구는 생성형 대규모 언어 모델을 사용하여 텍스트 기반 제로샷 학습(zero-shot learning)을 이용한 정보 추출을 시도하며, 이를 통해 목적에 맞는 한국어 데이터셋을 구축한다. 본 연구에서는 시스템-지침-소스입력-출력형식의 프롬프트 엔지니어링을 통해 언어 모델이 원하는 결과를 출력하도록 지시하며, 입력 문장을 통해 언어 모델의 In-Context Learning 특성을 활용하여 데이터셋을 구축한다. 생성된 데이터셋을 기존 데이터셋과 비교하여 본 연구 방법론을 검증하며, 관계 정보 추출 작업의 경우 KLUE-RoBERTa-large 모델 대비 25.47% 더 높은 성능을 달성했다. 이 연구 결과는 한국어 텍스트에서 지식 요소를 추출하는 가능성을 제시함으로써 인공지능 연구에 도움을 줄 것으로 기대된다. 더욱이, 이 방법론은 다양한 분야나 목적에 맞게 활용될 수 있어, 다양한 한국어 데이터셋 구축에 잠재력을 가진다고 볼 수 있다.

키워드 : 대규모 언어 모델, 프롬프트 엔지니어링, 제로샷 학습, 데이터셋 구축, 정보 추출

1. 서 론

비정형적인 정보가 매우 빠르게 발생하는 현대 사회에서는 선택적으로 정보를 추출하는 것이 비용 효율적인 측면에서 매우 중요하다. 특히, 정치, 경제, 사회, 과학 등 다양한 분야에서는 데이터 분석을 통한 인사이트 도출이 필수적으로 요구되고 있다. 정보의 중요성은 개개인이나 단체마다 상대적으로 다르게 평가할 수 있으므로, 특정 도메인 온톨로지(Ontology)로 나타낸 클래스 체계에 맞춰 정보를 추출한다. 이를 위해 정보 추출(Information Extraction) 작업을 수행할 수 있으며, 구체적으로는 개체명 추출(Named Entity Recognition, NER), 관계 추출(Relation Extraction, RE), 사건 추출(Event Extraction, EE) 등으로 세부 작업(Task)을 구분할 수 있다. 정보 추출의 목표는 텍스트의 의미 구조를 명시적으로 만들어 이를

※ 이 연구는 2023년 정부(방위사업청)의 재원으로 국방과학연구소의 지원을 받아 수행된 미래도전국방기술 연구개발사업임(No.915026201).
※ 본 논문은 2023년 한국군사과학기술학회 종합학술대회에서 발표된 "ChatGPT와 프롬프트 엔지니어링을 통한 국방 및 안보 텍스트 기반 지식요소 추출 한글 데이터셋 구축 방법" 연구를 확장한 것임.
† 정 회 원 : 텔레픽스 주식회사 주임연구원
†† 비 회 원 : 텔레픽스 주식회사 연구원
††† 비 회 원 : 텔레픽스 주식회사 데이터사이언스부문장
Manuscript Received : October 4, 2023
Accepted : November 1, 2023
* Corresponding Author : Kwon Da Rong Sae(darong.kwon@telepix.net)

활용할 수 있도록 하는 것이다[1].

정보 추출에는 규칙적으로 나타나는 언어적 패턴 특징을 이용한 방법론부터 딥러닝을 이용한 지도학습 방식까지 다양하게 연구되고 있다[2]. 그중 지도학습 방식은 규칙 기반 방식보다 성능이 높지만, 학습에 레이블링 된 데이터가 필요하다. 사전학습(Pre-trained) 모델을 사용하더라도, 이를 미세 조정 학습(Fine-tuning)하기 위해서는 마찬가지로 해당 작업에 대한 데이터가 필요하다. 그러나 정답이 레이블링 된 데이터셋을 구축하는 데에는 많은 시간과 비용이 든다는 단점이 있고, 온톨로지 지식 체계가 달라지면 다시 처음부터 학습해야 한다는 점에 있어 비효율적이다.

이를 해결하기 위해 소수의 데이터셋만을 이용하여 학습한 뒤 정보를 추출하는 준지도학습 연구[3], 특정 도메인(Domain)의 소수 데이터를 이용해 데이터를 증강하여 정보 추출 모델을 학습시키는 연구[4] 등이 이어졌으나 이 또한 일정 수준의 데이터셋이 필요하며 전문적인 도메인의 경우 데이터셋을 구축하는 비용 또한 증가한다.

거대 언어 모델(Large Language Model, LLM)은 언어 모델을 확장한 개념으로 특정 작업 수행에 국한된 기존 모델과 달리 다양한 작업을 수행할 수 있는 역량을 보유하고 있으며 일반적인 자연어 작업에 대해 뛰어난 성능을 자랑하여 학계와 산업계 모두에서 상당한 관심을 불러일으키고 있다[5]. 특히, 모델의 크기가 커지면 커질수록 모델의 일반 성능이 계속해서 증가한다는 것이 실험으로 증명되었고[6], LLM의 성능을 측정하고자 하는 노력이 계속되고 있으며[7], 여러 작업에서 제로샷 학습(Zero-shot learning)으로도 높은 성능을 달성할 수 있다. 하지만 LLM의 특성상 추론을 할 때 많은 비용이 발생한다는 단점이 있다. 이를 해결하는 방안으로는 LLM으로 레이블링된 데이터셋을 구축하고, 소규모의 언어 모델로 지식 증류(Knowledge Distillation) 방법을 사용해 학습하는 방안이 있다[8]. 이를 통해 LLM이 주어진 온톨로지나 클래스 체계를 잘 이해한다면, 목적 작업을 수행하는 것에 있어 필요한 일반적인 지식이 레이블링된 데이터셋을 만들 수 있을 것이다.

이에 본 연구에서는 LLM의 상황 내 학습(In-context Learning) 특성을 이용한 제로샷 학습(Zero-shot learning)과 프롬프트 엔지니어링(Prompt Engineering)을 통해 한글 텍스트로부터 정보 추출 데이터셋을 구축하는 방법을 제안한다. 본 연구가 추구하는 목표는 크게 두 가지다. 첫째, 한글 텍스트에서의 정보 추출 기술의 적용 가능성을 확대하고자 한다. 둘째, 제로샷 학습과 프롬프트 엔지니어링을 통한 효율적인 데이터셋 구축 방법론을 제시하여, 정보 추출 작업에 대한 연구 및 적용을 촉진하고자 한다. 이를 통해 얻을 수 있는 연구의 기여점은 다음과 같다. 첫째, 데이터셋의 일관성을 결정하는 도메인 온톨로지에 상관없이 목적에 맞는 한국어 데이터셋을 생성할 수 있다. 둘째, 적절한 프롬프트 엔지니어링을 통해 이를 자동화할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 먼저, 관련 연구를 리뷰하고, 그 후에 제안하는 방법론을 전체적인 파이프라인과 함께 상세히 설명한다. 이어서 실험 과정 및 결과를 검증용 데이터셋과 함께 제시하고, 마지막으로 결론과 향후 연구 방향을 논의한다.

2. 관련 연구

2.1 정보 추출

정보 추출은 크게 명명된 개체명 인식(NER), 관계 추출(RE), 사건 추출(EE) 세 가지 작업(Task)으로 나눌 수 있다. NER은 엔티티를 식별할 뿐만 아니라 사람, 조직, 위치, 날짜 등과 같은 사전 정의된 클래스로 분류함으로써 정보 검색, 질문 답변, 지식 그래프 구축과 같은 수많은 NLP 애플리케이션에서 중요한 역할을 한다. RE는 문장 내 또는 여러 문장에 걸쳐 추출된 개체 간의 의미 관계를 식별하고 분류하는 작업으로서 온톨로지 구축, 소셜 네트워크 분석, 자동 요약과 같은 작업에 이용될 수 있다. EE는 트리거로 구분되는 사건 인스턴스를 식별하여 텍스트의 문맥적 이해를 명시적으로 나타낸다. 또한, 사건과 관련된 인자(Argument)를 감지하여 에이전트, 객체, 시간, 위치 등의 역할(Argument Role)을 할당한다.

각각의 작업은 서로 다른 지식을 요구하므로, 규칙 기반(Rule-based) 방법의 경우 이들을 별개의 작업으로 분리하여 수행한다. 해당 작업의 방법론으로는 문법 규칙에 따른 개체명 인식 방법[9], DBpedia 데이터 및 구문 구조를 활용한 관계 추출 방법[10], 특정 사건 패턴을 사전 정의하여, 문장에 해당 패턴이 존재하는지 검사하는 방법으로 사건 추출 방법[11] 등이 제안되었다. 규칙 기반 정보 추출 방법은 규칙을 갱신하거나 삭제함으로써 간단하게 관리할 수 있다는 장점이 있으나, 비정형적인 텍스트나 사전에 고려하지 않은 패턴에 대해 취약하다는 단점이 있다[12].

Transformer, BERT와 같은 입력과 출력의 복잡한 관계를 자동으로 학습하는 신경망 언어 모델이 나옴에 따라, 정보 추출 모델에도 규칙 기반 방법 대신 지도학습(Supervised Learning) 방법이 적용되기 시작했다[13,14]. 이를테면 주의집중(Attention)을 활용하여 사건을 추출하거나[15], 토큰 분류(Token Classification)를 통한 개체명 인식을 수행하는 방법이 제안된 바 있다[16]. 특히 한국어로도 KLUE와 같은 정보 추출 관련 작업에 대한 지도학습 데이터셋이 공개된 바 있어, 사전 학습(Pretrained) 언어 모델을 미세 조정 학습(Fine-tuning) 하는 것으로 일정 수준의 성능을 달성할 수 있다[17]. 하지만 지도학습의 특성상 목표 작업 혹은 목표 도메인에 부합하는 데이터셋이 없을 경우, 좋은 성능을 기대하기 어렵다. 이를 보완하기 위해 특정 도메인의 소수 데이터셋을 증강

(Augmentation) 하는 연구의 경우, 정보 추출 모델의 성능을 성공적으로 개선시킨 것으로 나타났다[4].

2.2 생성형 대규모 언어 모델

생성형 LLM은 별도의 학습 데이터셋이 없더라도 다양한 작업을 수행할 수 있는 것으로 알려져 있다[18]. 특히 작업에 대한 응답 예시를 여러 개 제시하거나, 복잡한 작업에 대해서는 단계적으로 추론하도록 지시하는 명령문을 통해 작업 성능을 크게 향상시킬 수 있다[19]. 이는 언어 모델이 연속되는 텍스트 패턴이 아닐 경우 사전 학습된 지식(Concept)을 활용한 조건부 확률(Conditional Probability)을 계산하거나, 지시사항에 해당하는 패턴임을 인지할 수 있기 때문으로 추정된다[20]. 또한, 정보 추출 작업에 해당하는 개체명 인식, 관계 추출, 사건 추출 등도 생성형 거대 언어 모델로 수행할 수 있는 것으로 여겨진다[21]. 하지만 생성형 거대 언어 모델은 매개변수(Parameter)가 매우 많으므로 특정 목표 작업 및 목표 도메인에 최적화되도록 학습하기 어렵다[22].

본 연구에서는 API를 통해 사용이 가능한 OpenAI사의 “gpt-3.5-turbo-0613” 모델을 사용한다. ChatGPT는 GPT 아키텍처 기반으로 학습된 LLM으로, 지침(Instruction)이 프롬프트로 주어졌을 때, 출력 답변(Completion)을 생성하는 작업에 특화되어 있고, 모델 답변에 대한 사람의 피드백을 강화학습을 통해 모델에 적용했다[23]. 따라서 맥락에 맞는 대답을 생성할 수 있으므로 비정형 텍스트로부터 정형화된 정보를 추출하는 것과 같은 다운스트림 작업(Downstream Task)이 가능하다[24].

2.3 정보 추출 작업 검증 데이터셋

본 연구에서는 KLUE-NER, KLUE-RE, ACE2005 데이터셋을 사용하여 LLM을 통한 정보 추출 데이터셋 구축 방법론에 대한 검증을 한다.

KLUE 벤치마크 데이터셋 GLUE, SuperGLUE와 같이 한국어 자연어 이해(Natural Language Understand) 언어 모델을 평가할 수 있는 표준 데이터셋 구축을 목표로 제작되었으며, NER, RE 등 8가지의 작업을 평가할 수 있다. 이중 KLUE-NER은 PS(Person), LC(Location), DT(Date), TI(Time), QT(Quantity), OG(Organization)의 총 6개의 엔티티 타입으로 구성되어 있으며, 문자 수준의 BIO(Begin-Inside-Outside) 태깅 체계를 통해 태그가 지정되어 있다[17].

KLUE-RE 데이터는 문장과 함께 주어진 두 엔티티를 통해 관계를 식별하는 {esubj-relation-eobj} 구조의 적절한 트리플렛(Triplet)을 찾아내는 작업이며, 사람 관련 관계 18개, 조직 관련 관계 11개와 “no_relation”으로 구성된 30개의 관계 타입으로 구성되어 있다[17].

ACE2005 데이터셋은 Linguistic Data Consortium(LDC)에서 제작한 데이터셋으로, 영어, 중국어, 아랍어 등의 정보 추출을 지원하기 위해 제작되었다[25]. 특히 사건 추출 작업에 사용되는 데이터는 총 33개의 사건 타입과 35개의 사건 인자 역할 타입으로 이루어져 있다.

3. 연구 방법

3.1 한국어 정보 추출 데이터셋 구축 파이프라인

본 논문에서는 LLM을 이용한 데이터셋 구축 방법론을 제시하며, 이에 대한 파이프라인은 Fig. 1에서 확인할 수 있다. Fig. 1은 ACE2005 데이터셋의 온톨로지 정보를 사용해 예시로 작성한 파이프라인으로, 여기에 표현된 각 작업에 대한 레이블들은 데이터셋을 구축할 때 목적에 맞게 만들어진 온톨로지로 충분히 수정될 수 있다. 예시에서 사용한 문장은 한국어 문장으로부터 정보가 추출되는 것을 보여주며, 영어 등의 타 국어 또한 한국어로 번역한다면 충분히 사용이 가능하다.

먼저, 정보를 추출하고자 하는 대상 문장이 있을 때, 해당 문장으로부터 엔티티의 후보가 될 수 있는 멘션을 추출한다. 이 작업은 단순히 엔티티를 추출할 때 조사나 수식어 등의 불필요한 언어적 요소들을 걸러주는 역할을 한다. 그리고 전처리를 통해 의미를 가질 수 있는 멘션의 조합에 대한 리스트를 추출하도록 전처리를 한다.

도메인-레인지 리스트(Domain-Range List)는 관계 타입을 표현할 때 등장할 수 있는 주격 명사와 목적격 명사의 구간을 제한해주는 역할을 한다. 자세한 설명은 3.2절에서 설명한다. 이후, 관계 추출 작업에서 도메인-레인지 리스트를 통해 필터링된 트리플렛 조합을 추출한다.

사건 정보 추출에서는 트리거(Trigger)와 사건 타입, 사건 인자 역할(Argument Role)을 추출한다. 트리거는 사건에 영향을 미치는 용언이나 체언이 될 수 있으며, 사건 인자 역할은 사건 발생에 참여하는 엔티티의 역할이라 간주하여 레이블을 부여한다.

정보 추출 시스템 전체 알고리즘에 대한 의사 코드는 Algorithm 1에 나타나 있다.

3.2 프롬프트 구조

생성형 LLM을 통해 데이터셋을 생성하기 위해서는 구조화된 프롬프트를 입력 데이터로 사용해야 한다. 여기서 구조화된 프롬프트란 LLM에게 역할 부여, 지침 제시, 추출 타겟 데이터 등을 명확히 구분해서 입력하는 것을 말한다. 본 연구에서는 시스템(System), 지침(Instruction), 입력 데이터(Input), 출력 형식(Output Format)으로 나누어 프롬프트를 작성한다.

시스템은 LLM에게 역할을 부여하고 대략적인 작업에 대한 설명해 대해 작성하며, 각 정보 추출 작업별로 다른 역할을 부여하고 작업 설명을 추가했다. 지침으로는 LLM이 작업을 수행하면서 따라야 하는 지시사항을 작업 특성에 맞게 작성했다. 예를 들어, 개체명 인식의 경우 추출되어야 하는 온톨로지

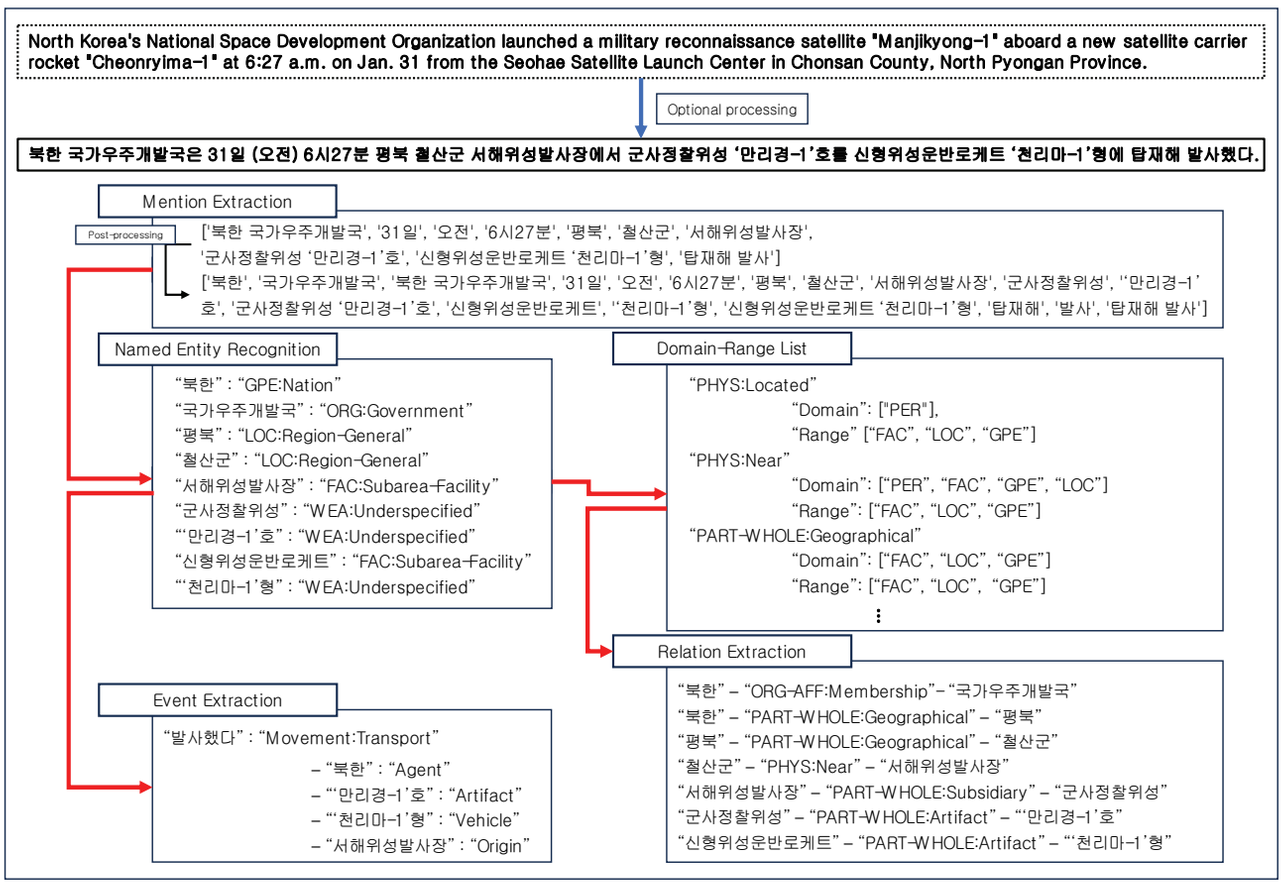

Fig. 1. Dataset Generation Pipeline from Large Language Model(ACE2005 Ontology Example)

Algorithm 1. Korean Information Extraction Dataset Construction Pipeline

```
function ExtractMentions(sentence)
    // Extract mention candidates from the sentence
    return mentionCandidates
end function

function PreprocessMentions(mentionCandidates)
    // Create a list of mention combinations
    return mentionCombinations
end function

function ExtractTriplets(mentionCombinations, DomainRangeList)
    // Extract triplets filtered by the Domain-Range List
    return tripletList
end function

function ExtractEventInfo(sentence)
    // Extract triggers, event types, and argument roles
    return eventInfo
end function

function GeneratePrompt(systemInfo, instructions, inputData, output-
Format)
    // Define the prompt structure
    return prompt
end function

procedure BuildDataset(sentences)
    finalDataset ← new map
    finalDataset[' MentionList' ] ← empty list
    finalDataset[' TripletList' ] ← empty list
    finalDataset[' EventInfo' ] ← empty list
    for each sentence in sentences do
        mentionCandidates ← ExtractMentions(sentence)
        mentionCombinations ← PreprocessMentions(mentionCandidates)
        DomainRangeList ← GetDomainRangeList(...)
        tripletList ← ExtractTriplets(mentionCombinations, Domain-RangeList)
        eventInfo ← ExtractEventInfo(sentence)
        prompt ← GeneratePrompt(systemInfo, instructions, sentence, "JSON" )
        finalDataset[' MentionList' ].append( mentionCombinations)
        finalDataset[' TripletList' ].append(tripletList)
        finalDataset[' EventInfo' ].append(eventInfo)
    end for
    Save finalDataset to file(prompt[ 'OutputFormat' ])
    return finalDataset
end procedure
```

레이블을 리스트 형태로 제시한 후 해당 리스트 내에 존재하는 레이블로 추출하도록 지시한다. 입력 데이터에는 정보 추출의 대상이 되는 문장을 입력한다. 출력 형식은 추출하고자 하는 정보의 형태를 명시한다. 예를 들어, JSON 형식의 출력을 원한다면 JSON 확장자로 저장될 수 있도록 키(Key)와 밸류(Value)로 이루어진 자료형을 문자열로 제시한다.

한국어 정보 추출 데이터셋 구축에 사용되는 프롬프트는 Table 1, 2, 3, 4에 각각 명시했다. 프롬프트에서 ***sentence***와 같이 이탤릭체로 작성된 부분은 문장 및 레이블 리스트 등 미리 데이터를 저장해둔 변수를 의미한다. Table 1의 ***mention_combinations***는 멘션 추출 이후 스팬(Span)으로 이루어진 멘션들을 후처리를 통해 여러 조합으로 나타낸 리스트를 말하며, 해당 리스트 내에서 엔티티를 추출하도록 한다.

Table 3의 ***valid_relation_types***는 Fig. 1에 있는 도메인-레인지 리스트 내에서 문장에 등장한 엔티티들의 타입만 고려했을 때 가능한 관계 타입들을 필터링한 리스트를 말한다. 이것은 LLM에 입력되는 토큰의 수를 줄이기 위한 것으로, 예를 들어, “PHYS:Near” 관계 타입은 도메인에 “PER”, “FAC”, “GPE”, “LOC”만 등장할 수 있으며, 관계의 대상이 되는 레인지에는 “FAC”, “LOC”, “GPE”만 등장할 수 있다. 따라서, “LOC”에 속하는 “철산군”과 “FAC”에 속하는 “서해위성발사장”의 관계 타입이 “PHYS:Near”으로 결정된다.

Table 1. Mention Extraction Prompt

System:
You are an expert at extracting mentionsfrom sentences. Generate a mention list, making sure to follow the instructions between the bullets below:

Instruction:
- Extract mentions from a given sentence.
- Consider all spans of extracted mentions.
- Extract as mention, including the pronouns in the sentence.
- A mention is a noun or noun phrase in a sentence excluding any mention of time, date, or day of the week.
- An extracted mention contains all the words that modify that mention.

Input:
sentence = ***{sentence}***

Output Format:
{**“mentions”**: [{Extract noun or noun phrase mentions in a sentence and represent them as a list.}]}

Table 2. Named Entity Recognition Prompt

System:
entity_types: ***{entity_list}***
mentions: ***{mention_combinations}***
You are a NER task expert. Given a {sentence} and {mentions}, generate a named entity recognition example using only the classes in the entity type list, making sure to follow the instructions between the bullets below:

Instruction:
- The "entity_type" is one of elements in the {entity_types} list. But, if no reasonable entity type is in the list, assign "None".
- The "text" must be one of mentions in the {mentions} list.
- Perform the NER task on all elements in the {mentions} list.
- Use common sense in your judgment.
- Always perform NER based on the entire sentence.
- Don't make up other entity types. The entity type must be one of the elements in the entity_typeslist.
- Perform a NER on the entity mention combination and include it.
- Do not answer in code format.

Input:
sentence = ***{sentence}***

Output Format:
{**“entity_mentions”**= [
“text”: {The text of the entity mention in sentence. The text is one of the mentions list.},
“entity_type”: {The entity type of the entity mention in sentence.}]}

Table 3. Relation Extraction Prompt

System:
relation_types: ***{valid_relation_types}***
entity_mentions: ***{entity_mentions}***
You are a Relation Extraction task expert. Given a sentence, generate a relation extraction example using only the classes in the relation type list, making sure to follow the instructions between the bullets below:

Instruction:
- The text of arguments must be in the form of a triple with one relationship, andhas meaning in the form of "Arg-1"-"relation_type"-"Arg-2" order.
- Use common sense in your judgment and create as many relationships as possible.
- The {relation_type} must be one of the given {relation type list}. Don't modify the relation type that exists in the following {relation_types}. If the relationship type is not on the {relation_types}, exclude it.
- Relation "text" of "subject" and "object" must be one of the give {NER results}.

Input:
sentence = ***{sentence}***

Output Format:
{**"relation_mentions"**: [
"Arg-1": {The entity mention text of the relation triple in the given sentence. The mention text must be one of the {entity_result}. The values of Arg-1 in {relation_types} are the possible entity types. For example, 'PHYS:Located' can only get PER for Arg-1.},
"relation_type": {The relation type of the relation triple in sentence. You must pick one from the {relation_list}.},
"Arg-2": {The entity mention text of the relation triple in the given sentence. The mention text must be one of the {entity_result}. The values of Arg-2 in {relation_types} are the possible entity types. For example, 'PHYS:Located' can only get FAC, LOC, and GPE for Arg-2.}]}

Table 4. Event and Argument Role Extraction Prompt

System:
event_types: ***{event_type_list}***
argument_roles: ***{argument_role_list}***
mentions: ***{entity_mentions}***
I want you to act aEvent Extraction and Argument Role Extraction task expert. Given a sentence, you generate aevent and argument role extraction example using only the classes in the event type list and argument role list, making sure to follow the instructions between the bullets below:

Instruction:
- In event extraction, an event trigger is a word or multi-word that depicts the occurrence of an event in a text.
- In event extraction, an argument role the relationship between an argument and the event in which it participates.
- The event_typeJSON data structure looks like this Use this to extract the event type, trigger, and argument role.: {{"event type"}}: {{{{"argument role"}}: [{{"The type of entity that can be in the location"}}]}}
- "event_type" must exist in the given {{event_types}} key list.
- The "role" in "arguments" must exist in the given {{argument_role_list}} list.
- There can be multiple argument roles for an event.
- The "role" is one of the {{mentions}} list.
- The "trigger" is the main word that most clearly expresses an event occurrence, typically a verb or a noun. "trigger" must be existed in the given sentence.
- **Extract as many event triggers and argument roles as you can.**

Input:
sentence = ***{sentence}***

Output Format:
{**"event_mentions"**= [
"event_type": {The word that most clearly expresses the mention of an event, most often for a single verb or noun. The event type must be one of the {valid_event_types}},
"trigger": {The text of the event trigger in sentence. The event text must be one of the word in sentence.},
"arguments": {"properties": {
"text": {The entity mention text of the event argument role in sentence. The mention text must be one of the {entity_result}.},
"role": {An entity mention, temporal expression or value (e.g.Job-Title) that serves as a participant or attribute with a specific role in an event mention. The arguementrole type must be one of the {argument_roles}.}]}

4. 실 험

실험 세션에서는 LLM을 통한 정보 추출의 성능을 확인하며, F1 점수 메트릭을 사용한다. NER, RE, EE와 같은 작업들은 대체로 데이터에 불균형이 있고, 예측의 정밀성과 실제 대상을 놓치지 않는 능력(즉, 정밀도와 재현율) 간의 균형을 맞춰야 하는 특성을 가진다. F1 점수는 이 두 요소를 조화롭게 평가하여 모델의 성능을 단일 지표로 요약할 수 있게 해줌으로써, 모델이 너무 많은 대상 토큰을 예측하는 경향이 있는지, 또는 실제 목표로 하는 대상 토큰을 놓치고 있는지에 대한 이해를 확인할 수 있고 다양한 모델과 설정을 공정하게 비교할 수 있는 기준을 제공한다. 따라서, 이러한 정보 추출 태스크에서 F1 점수는 모델의 균형 잡힌 성능 평가를 위한 적절한 메트릭으로 간주 되므로 본 연구에서의 주된 메트릭으로 사용한다.

4.1 실험 데이터셋 및 실험 설계

본 논문에서는 LLM으로 구축 가능한 정보 추출 데이터셋의 질을 확인하기 위해, 기존 벤치마크 데이터셋에 대한 비교 평가를 수행하였다. NER과 RE는 한국어 벤치마크 데이터셋인 KLUE-NER과 KLUE-RE 벤치마크로 성능을 확인하고 EE는 한국어 벤치마크 데이터셋이 존재하지 않기 때문에 ACE2005 영어 데이터셋을 한국어로 번역한 결과를 기준으로 실험하였다.

NER 작업은 KLUE-NER의 엔티티 수준 매크로(Macro) F1 점수와 문자 수준 Macro F1 점수로 성능을 측정한다[17]. 이는 BIO 태깅을 통해 엔티티 타입 또는 문자의 위치가 정답과 동일한지 확인한다.

RE 작업에 대해서는 KLUE-RE 전체 레이블 중 “no_relation”을 포함시킨 상태로 추론한 뒤, 이를 제외시킨 후 Micro F1 점수를 통해 성능을 측정한다.

본 연구에서는 사건 정보 추출 작업을 위해 ACE2005 데이터셋의 사건 트리거 및 사건 인자 역할에 대한 주석 데이터를 ChatGPT를 통해 한국어로 번역하였으며, 엔티티 정보와 관계 정보 및 사건 정보가 모두 존재하는 데이터 중 문장 내에 존재하는 요소들이 모두 한국어로 번역이 가능한 영문 데이터를 샘플링하였다. 평가를 위한 EE 번역 데이터는 LLM과 프롬프트만으로 추출한 결과물과 직접 비교하기 위해 별도의 샘플링 작업이 필요하다. 왜냐하면, 한국어에는 조사의 존재와 문장 순서의 자유도, 문법의 차이 등으로 인해 영어에서 한국어로 번역할 경우 생략되거나 역할이 바뀌는 경우가 있기 때문이다. 엔티티나 사건에 영향을 미치는 인자 멘션 등으로 레이블링 되어있는 경우, 번역되는 과정에서 해당 멘션이 문장으로부터 생략된다면 기존 데이터셋이 변형된다. 따라서 레이블과 문장 내에 존재하는 멘션이 매칭되는 문장을 샘플링하였다. 그 후, 프롬프트 엔지니어링을 통해 영문 데이터셋의 문장과 레이블링 된 텍스트를 한국어로 번역함과 동시에 JSON 형태로 추출하도록 출력 형태를 고정하였다.

EE 작업은 ACE2005 데이터셋의 사건 멘션 레이블 데이터를 ChatGPT를 통해 번역하여 사용했다. 평가 메트릭은 사건 트리거에 대한 F1-score와 사건 인자 역할에 대한 F1-score를 구분하여 추출하였다. 이는 각각 탐지(Identification)와 식별(Classification)로 구분하는데, 탐지는 트리거 단어 또는 인자 역할의 단어가 일치하는 경우를 나타내고, 식별은 단어와 타입 모두 일치하는 경우를 나타낸다[2]. 하지만, 생성형 모델 특성상 정답이 여러 단어가 합쳐진 합성 단어이거나 구분이 애매한 단어일 경우 이를 정확하게 추출하기 힘들다. 따라서 본 연구에서는 chrF 성능 지표를 사용하여 단어의 일치 여부를 나타냈다. chrF는 번역 작업의 성능을 나타내는 지표로써 번역된 텍스트의 문자 단위 N-그램(Character N-gram)에 대한 문자 단위 정밀도(Character Precision, chrP)와 재현성(Character Recall, chrR)을 이용한다[26]. Equation (1)에 나와 있듯이 β값에 따라 정밀도와 재현성의 비중이 달라지는데, 연구에서는 β값을 1로 고정했다[26]. 정답 텍스트와 생성된 결과 텍스트에 대해 각각 지표를 도출한 뒤 chrF의 값이 5 이상일 경우 동일한 단어로 간주했다.

$$\mathrm{chrF}\beta = (1+\beta^2)\frac{\mathrm{chrP}\times\mathrm{chrR}}{\beta^2\times\mathrm{chrP}+\mathrm{chrR}} \tag{1}$$

4.2 실험 결과

4.2절에서는 연구에서 제시한 LLM 데이터셋 구축 방법론을 통해 생성한 결과를 기존 벤치마크 데이터셋과 비교함으로써 방법론에 대한 성능을 확인하며, 자세한 점수는 Table 6에 작성되어 있다.

NER 작업은 공개된 KLUE-NER 평가용 데이터셋(Dev) 5,000개를 사용했으며, 엔티티 F1 점수와 문자 F1 점수를 각각 평가했다. 두 점수 모두 벤치마크에서 제시한 BERT 기반 모델의 성능보다는 낮았지만, 본 실험에서 F1 점수가 낮은 이유는 다음과 같이 분석할 수 있다. 먼저, 생성형 모델의 특성상 양방향 문맥을 고려하고 마스킹 된 토큰을 정해진 범위에서 추론하는 BERT 기반 모델과는 달리, 앞서 나열된 문맥에 따라 이후 토큰이 확률적으로 생성되기 때문에 클래스 범위 내에서 정답이 추출되지 않을 수 있다. 하지만, 추가 학습 없이 제로샷 러닝을 통해 추출한 결과인 점을 감안하면 기본적

Table 5. Evaluation Results(ACE2005-EE)

Metrics \ Tasks	ACE2005 (Event Trigger)	ACE2005 (Event Argument Role)
F1-score (Identification)	63.81	48.88
F1-score (Classification)	49.11	37.00

Table 6. Evaluation Results(KLUE-NER, KLUE-RE)

Tasks / Models	KLUE-NER(Dev)		KLUE-RE(Dev)	
	F1-score(Macro/Entity)	F1-score(Macro/Character)	F1-score(Micro)	AUC
KLUE-BERT-base	83.97	91.39	66.44	66.17
KLUE-RoBERTa-large	85.00	91.86	71.13	72.98
KLUE-RoBERTa-base	84.60	91.44	67.65	68.55
KLUE-RoBERTa-small	83.65	91.14	60.89	58.96
Ours(Zero-Shot)	67.60	72.03	89.25	-

인 엔티티 추출 능력을 가지고 있다고 볼 수 있었다.

RE의 경우, 벤치마크 모델 중 성능이 가장 높은 KLUE-RoBERTa-large 모델보다 25.47% 더 높은 성능을 달성했다. 공개된 KLUE-RE 평가용 데이터셋에서 “no_relation” 레이블을 제외한 3,143개 데이터를 평가용 데이터로 사용했다. Open AI의 ChatGPT API의 경우, 각 생성한 토큰의 신뢰도 점수(Confidence Score)를 확인할 수 없기에 AUC 점수는 제외하였다. 마찬가지로 추가 학습 없는 제로샷 러닝 성능을 확인하였고, 본 실험을 통해 충분히 거대한 대규모 생성 모델은 BERT 기반 모델보다 두 엔티티 간의 관계 추론 능력이 뛰어나다는 것을 확인할 수 있었다.

EE는 4.1에서 설명한 바와 같이, 생성형 모델의 특성과 번역된 데이터셋인 점을 고려하여 chrF를 이용한 F1 점수를 통해 평가했다. 엔티티, 관계, 사건 정보가 모두 포함된 ACE2005 데이터셋 1,970개 중 레이블 데이터 손실 없이 한국어로 번역이 가능한 데이터 560개를 샘플링한 데이터로 평가했으며, Table 5에 해당 점수가 나타나 있다. EE는 문장의 트리거를 파악하고 사건 유형에 영향을 미치는 인자들을 식별 및 분류하는 복합적인 작업이다. 따라서, F1 점수가 다른 모델들보다 [2] 상대적으로 낮을 수는 있지만, 본 실험에 사용된 데이터는 번역된 데이터이므로 다른 모델과 직접적으로 비교가 불가능하다. 또한, 생성형 모델 특성상 여러 인자 역할이 등장할 수 있고 데이터셋이 온톨로지에 따라 다른 기준으로 작성될 수 있음을 감안하여 봤을 때, 기본적인 추론 성능은 어느 정도 달성했음을 확인할 수 있었다. 이러한 결과로 봤을 때, EE의 경우 퓨샷 러닝 내지는 작업의 스텝을 나누어 여러 번 추론하는 등의 추가적인 작업이 더해진다면 성능을 높일 수 있을 것으로 분석할 수 있었다.

5. 결 론

본 연구에서는 LLM을 활용하여 한글 텍스트로부터 정보추출 데이터셋을 구축하는 방법론을 제안하였다. 실험 결과, 제안한 방법론으로 생성된 정보 추출 데이터는 기존 벤치마크 데이터셋으로 성능 측정이 가능한 수준이었으며, 벤치마크의 온톨로지 체계에 상관없이 결과를 도출할 수 있는 것을 확인할 수 있었다. 또한, 국방 도메인에 본 연구의 방법론을 적용한 Table 7의 결과처럼, 직접 한국어 정보추출 데이터셋을 만들어 언어 모델을 학습할 수 있는 기반을 마련할 수 있음을 보였다. 제로샷 학습을 통해 생성한 데이터셋은 기본적인 엔티티 추출 능력과 관계 추출 능력을 갖추었으며, 사건 추출 작업에 대해서도 일정 수준의 추론 성능을 보였다. 특히, 관계 추출에서 벤치마크 모델 중 성능이 가장 높은 KLUE-RoBERTa-large 모델보다 25.47% 더 높은 성능을 달성했다. 이러한 결과는 LLM을 이용하여 한글 정보 추출 작업을 수행하는 데에 유망한 방법론이 될 수 있다는 것을 시사한다. 따라서, 본 연구에서 제안한 데이터셋 구축 방법론은 한글 정보 추출 기술의 활용 가능성을 확대하고, 효율적인 데이터셋 구축 방법을 제시함으로써 정보 추출 작업에 대한 연구와 적용을 촉진할 수 있다. 더 나아가, 본 연구는 LLM을 활용하는 다른 도메인이나 언어에 대한 정보 추출 작업의 연구 및 응용에도 도움이 될 수 있으며, 이를 자동화하여 데이터셋 구축에 들어가는 비용을 절감할 수 있다.

Table 7. Defense Domain Korean Information Extraction Dataset JSON Format Example(ACE2005 Ontology)

```
[
    {
        "sentence_kor": "북한이 지난해 11월
                         대륙간탄도미사일(ICBM) 화성-17형
                         발사 성공을 기념해 미사일공업절을
                         제정한다.",
        "entity_mentions": [
            {
                "text": "북한",
                "entity_type": "GPE:Nation"
            },
            {
                "text": "대륙간탄도미사일(ICBM)",
                "entity_type": "WEA:Projectile"
            },
            {
                "text": "화성-17형",
                "entity_type": "WEA:Projectile"
            },
            {
                "text": "미사일공업절",
                "entity_type": "ORG:Government"
            }
```

```
    ],
    "relation_mentions": [
        {
            "Arg1": {
                "text": "북한",
                "entity_type": "GPE:Nation"
            },
            "relation_type": "ART:User-Owner-
            Inventor-Manufacturer",
            "Arg2": {
                "text": "대륙간탄도미사일(ICBM)",
                "entity_type": "WEA:Projectile"
            }
        },
        {
            "Arg1": {
                "text": "북한",
                "entity_type": "GPE:Nation"
            },
            "relation_type": "PART-WHOLE:Artifact",
            "Arg2": {
                "text": "미사일공업절",
                "entity_type": "ORG:Government"
            }
        }
    ],
    "event_mentions": [
        {
            "event_type": "Business:Start-Org",
            "trigger": "제정",
            "arguments": [
                {
                    "text": "북한",
                    "role": "Agent",
                    "entity_type": "GPE:Nation"
                },
                {
                    "text": "미사일공업절",
                    "role": "Org",
                    "entity_type": "ORG:Government"
                }
            ]
        }
    ],
    "file_name": "20231105_NEWSIS_NISX20231105_00
                02509323.json"
},
{
    "sentence_kor": "북한 국가우주개발국은 31일 (오전)
                6시27분 평북 철산군
                서해위성발사장에서 군사정찰위성
                '만리경-1'호를 신형위성운반로케트
                '천리마-1'형에 탑재해 발사했다.",
    "entity_mentions": [
        {
            "text": "북한",
            "entity_type": "GPE:Nation"
        },
        {
            "text": "국가우주개발국",
            "entity_type": "ORG:Government"
        },
        {
            "text": "평북",
            "entity_type": "LOC:Region-General"
        },
        {
            "text": "철산군",
            "entity_type": "LOC:Region-General"
        },
        {
            "text": "서해위성발사장",
            "entity_type": "FAC:Subarea-Facility"
        },
        {
            "text": "군사정찰위성",
            "entity_type": "WEA:Underspecified"
        },
        {
            "text": "'만리경-1'호",
            "entity_type": "WEA:Underspecified"
        },
        {
            "text": "신형위성운반로케트",
            "entity_type": "WEA:Underspecified"
        },
        {
            "text": "'천리마-1'형",
            "entity_type": "WEA:Underspecified"
        }
    ],
    "relation_mentions": [
        {
            "Arg1": {
                "text": "북한",
                "entity_type": "GPE:Nation"
            },
            "relation_type": "ORG-AFF:Membership",
            "Arg2": {
                "text": "국가우주개발국",
                "entity_type": "ORG:Government"
            }
        },
        {
            "Arg1": {
                "text": "북한",
                "entity_type": "GPE:Nation"
            },
            "relation_type":
"PART-WHOLE:Geographical",
            "Arg2": {
                "text": "평북",
                "entity_type": "LOC:Region-General"
            }
        },
        {
            "Arg1": {
                "text": "평북",
                "entity_type": "LOC:Region-General"
            },
            "relation_type":
"PART-WHOLE:Subsidiary",
            "Arg2": {
                "text": "철산군",
```

```
                "entity_type": "LOC:Region-General"
            }
        },
        {
            "Arg1": {
                "text": "철산군",
                "entity_type": "LOC:Region-General"
            },
            "relation_type": "PART-WHOLE:Subsidiary",
            "Arg2": {
                "text": "서해위성발사장",
                "entity_type": "FAC:Subarea-Facility"
            }
        },
        {
            "Arg1": {
                "text": "서해위성발사장",
                "entity_type": "FAC:Subarea-Facility"
            },
            "relation_type": "ART:User-Owner-
             Inventor-Manufacturer",
            "Arg2": {
                "text": "군사정찰위성",
                "entity_type": "WEA:Underspecified"
            }
        },
        {
            "Arg1": {
                "text": "군사정찰위성",
                "entity_type": "WEA:Underspecified"
            },
            "relation_type": "PART-WHOLE:Artifact",
            "Arg2": {
                "text": "'만리경-1'호",
                "entity_type": "WEA:Underspecified"
            }
        },
        {
            "Arg1": {
                "text": "'만리경-1'호",
                "entity_type": "WEA:Underspecified"
            },
            "relation_type": "GEN-AFF:Org-Location-
             Origin",
            "Arg2": {
                "text": "북한",
                "entity_type": "GPE:Nation"
            }
        },
        {
            "Arg1": {
                "text": "'천리마-1'형",
                "entity_type": "WEA:Underspecified"
            },
            "relation_type": "ART:User-Owner-
              Inventor-Manufacturer",
            "Arg2": {
                "text": "신형위성운반로케트",
                "entity_type": "WEA:Underspecified"
            }
        },
        {
                "Arg1": {
                    "text": "'천리마-1'형",
                    "entity_type": "WEA:Underspecified"
                },
                "relation_type": "PART-WHOLE:Artifact",
                "Arg2": {
                    "text": "북한",
                    "entity_type": "GPE:Nation"
                }
            }
        ],
        "event_mentions": [
            {
                "event_type": "Movement:Transport",
                "trigger": "발사했다",
                "arguments": [
                    {
                        "text": "북한",
                        "role": "Agent",
                        "entity_type": "GPE:Nation"
                    },
                    {
                        "text": "군사정찰위성 '만리경-1'호",
                        "role": "Artifact",
                        "entity_type": "WEA:Underspecified"
                    },
                    {
                        "text": "신형위성운반로케트
                        '천리마-1'형",
                        "role": "Vehicle",
                        "entity_type":
 "WEA:Underspecified"
                    },
                    {
                        "text": "평북 철산군
 서해위성발사장",
                        "role": "Origin",
                        "entity_type":
 "FAC:Subarea-Facility"
                    }
                ]
            }
        ],
        "file_name": "20230531_HANKYOREH_1094054.json"
    },
    ...
]
```


References

[1] R. Grishman, "Information extraction," *in IEEE Intelligent Systems*, Vol.30, No.5, pp.8-15, 2015, doi: 10.1109/MIS.2015.68.

[2] W. Xiang and B. Wang, "A survey of event extraction from text," in *IEEE Access*, Vol.7, pp.173111-173137, 2019, doi: 10.1109/ACCESS.2019.2956831.

[3] W. Liao and S. Veeramachaneni, "A simple semi-supervised algorithm for named entity recognition," In *Pro-*

ceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pp.58-65, 2009.

[4] D. Feng and H. Chen, "A small samples training framework for deep Learning-based automatic information extraction: Case study of construction accident news reports analysis," Advanced Engineering Informatics, Vol.47, pp.101256, 2021.

[5] Y. Chang et al., "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[6] J. Wei et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[7] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. X. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," *arXiv preprint arXiv:2305.18486*, 2023.

[8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, Vol.129, No.6, pp.1789-1819, 2021.

[9] Y. Sari, M. F. Hassan, and N. Zamin, "Rule-based pattern extractor and named entity recognition: A hybrid approach," *2010 International Symposium on Information Technology*, Kuala Lumpur, Malaysia, pp.563-568, 2010, doi: 10.1109/ITSIM.2010.5561392.

[10] C. Bizer et al., "DBpedia - A crystallization point for the web of data," *Journal of Web Semantics: Science, Services and Agents on the WWW*, Vol.7, No.3, pp.154-165, 2009.

[11] Q. C. Bui, D. Campos, E. van Mulligen, and J. Kors, "A fast rule-based approach for biomedical event extraction," In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp.104-108, 2013.

[12] S. Rao, D. Marcu, K. Knight, and H. Daume III, "Biomedical event extraction using abstract meaning representation," In *BioNLP 2017*, pp.126-135, 2017.

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

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

[15] S. Liu, Y. Chen, K. Liu, and J. Zhao, "Exploiting argument information to improve event detection via supervised attention mechanisms," In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol.1: Long Papers), pp.1789-1798, 2017.

[16] L. Zhao, L. Li, X. Zheng, and J. Zhang, "A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts," *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Dalian, China, pp.1233-1238, 2021, doi: 10.1109/CSCWD49262.2021.9437616.

[17] S. Park et al., "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.

[18] T. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, Vol.33, pp.1877-1901, 2020.

[19] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, Vol.35, pp.24824-24837, 2022.

[20] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," *arXiv preprint arXiv:2111.02080*, 2021.

[21] O. Sainz, H. Qiu, O. L. de Lacalle, E. Agirre, and B. Min, "ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations," *arXiv preprint arXiv:2203.13602*, 2022.

[22] B. Sharma, Y. Gao, T. Miller, M. M. Churpek, M. Afshar, and D. Dligach, "Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning," In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, Toronto, Canada. Association for Computational Linguistics, pp.78-85, 2023.

[23] 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.

[24] X. Wei et al., "Zero-shot information extraction via chatting with chatgpt," *arXiv preprint arXiv:2302.10205*, 2023.

[25] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia, 57.

[26] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," In *Proceedings of the tenth workshop on statistical machine translation*, pp.392-395, 2015.

정 영 상

https://orcid.org/0009-0004-1931-6598
e-mail : video.jeong@telepix.net
2019년 광운대학교 산업심리학과(학사)
2021년 서울과학기술대학교
데이터사이언스학과(석사)
2021년 ~ 2022년 한국전자기술연구원(KETI) 연구원
2022년 ~ 현 재 텔레픽스 주식회사 주임연구원
관심분야 : 자연어 처리, 대규모 언어 모델, 전이 학습

지 승 현

https://orcid.org/0009-0009-0285-4265
e-mail : sorryhyun@telepix.net
2021년 숭실대학교 컴퓨터학부(학사)
2023년 숭실대학교 소프트웨어학과(석사)
2023년 ~ 현 재 텔레픽스 주식회사 연구원
관심분야 : 자연어 처리, 언어 모델 학습, 전이 학습

권 다 롱 새

https://orcid.org/0009-0006-4561-581X
e-mail : darong.kwon@telepix.net
2001년 서울대학교 통계학과(학사)
2003년 서울대학교 통계학과(석사)
2012년 시카고대학교 통계학과(박사)
2023년 ~ 현 재 텔레픽스 주식회사 데이터사이언스부문장
관심분야 : 통계모델링, 기계학습