한국어 헬스케어 개체명 인식을 위한 거대 언어 모델에서의 형태소 기반 Few-Shot 학습 기법

강수연[1], 김건우[1*]
[1]경상국립대학교 컴퓨터과학부
sillon@gnu.ac.kr, gunwoo.kim@gnu.ac.kr

Morpheme-Based Few-Shot Learning with Large Language Models for Korean Healthcare Named Entity Recognition

Su-Yeon Kang[1], Gun-Woo Kim[1*]
[1]School of Computer Science, Gyeongsang National University

요 약

개체명 인식은 자연어 처리의 핵심적인 작업으로, 특정 범주의 명칭을 문장에서 식별하고 분류한다. 이러한 기술은 헬스케어 분야에서 진단 지원 및 데이터 관리에 필수적이다. 그러나 기존의 사전 학습된 모델을 특정 도메인에 대해 전이학습하는 방법은 대량의 데이터에 크게 의존하는 한계를 가지고 있다. 본 연구는 방대한 데이터로 학습된 거대 언어 모델(LLM) 활용을 중심으로, 한국어의 교착어 특성을 반영하여 형태소 정보를 활용한 Few-Shot 프롬프트를 통해 한국어 헬스케어 도메인에서의 개체명 인식 방법을 제안한다.

1. 서론

개체명 인식(Named Entity Recognition, NER)은 문장에서 특정 카테고리의 명칭인 사람, 기관, 장소 등을 찾아내고 분류하는 작업이다.[1] 특히 헬스케어 도메인에서는 질병, 약물, 증상 등과 같은 핵심 개체들을 추출하는데 이용하며, 진단 지원, 데이터 관리 및 여러 의료 작업을 위해 사용된다.

지난 몇 년 간 딥러닝 및 기계학습의 발전에 따라, 특정 도메인의 개체명 인식 성능을 향상시키기 위한 연구가 진행되어 왔다. 특히, BERT[2]와 같은 사전 학습된 언어 모델을 학습 데이터로 전이학습하여 개체명 인식을 수행하는 연구가 진행 되었지만, 이러한 방식은 여전히 대량의 학습 데이터에 크게 의존한다는 한계를 가지고 있다. 한국어의 경우 교착어의 특성을 갖추고 있어 별도의 고려가 필요하며, 헬스케어와 같은 특정 도메인에서는 대규모 한국어 데이터셋의 구축이 어려운 상황이다.

최근에는 ChatGPT[3]와 같은 거대 언어 모델(Large Language Model, LLM)의 등장이 자연어 처리의 패러다임을 크게 바꾸었다. 이러한 거대 언어 모델은 방대한 양의 텍스트 데이터를 통해 사전 학습되어 광범위한 문맥과 지식을 포함하고 있으며, 사용자의 프롬프트에 따라 답변을 생성하는 능력을 보여주고 있다.

본 논문에서는 거대 언어 모델의 장점을 활용하여, 한국어의 교착어 특성을 고려한 형태소 정보 기반 Few-Shot 프롬프트 방식을 통한 헬스케어 도메인의 개체명 인식 방법을 제안한다.

2. 방법

2.1. 한국어 헬스케어 개체명 데이터

본 논문에서는 한국어 헬스케어 도메인에서의 개체명 인식을 목표로 진단에 관한 네이버 지식인 정보를 크롤링하여 사용하였다. 총 1830 개의 문장을 추출하였으며, 개체명의 분류는 질병(DS), 증상(ST), 신체(BD)로 나누어 BIO 태깅 방식을 적용하였다.

<표 1> 개체명 태그별 엔티티 수

질병(DS)		증상(ST)		신체(BD)	
B-DS	I-DS	B-ST	I-ST	B-BD	I-BD
818	333	1076	114	884	37

2.2. LLM 입력을 위한 데이터 구축

1830 개의 문장 중 1380 개를 학습 데이터로, 450 개를 테스트 데이터로 분할하였다. 학습 데이터는 Few-Shot 학습을 위한 데이터로 활용되었다. LLM 입력을 위해 단어와 BIO 태그 정보에서 확장하여 데이터를 구성하였으며 그림 1 은 이러한 데이터 구조의 예시를 보여준다. 먼저 문장을 공백 단위로 분리한 단어 Word 에서 Mecab 을 통해 얻은 형태소 정보를 이용하여 형태소 정보인 POS 를 구성하였다. LLM 에 개체명 인식 결과를 얻기 위하여 학습데이터에서는 BIO 정보

* 교신저자 (Corresponding Author)

를 통해 Entity 와 Though 에 대한 내용을 추가해 주었다. 이 구조는 연쇄적인 추론 과정인 CoT(Chain of Thought)[4] 을 위해 구축하였다.

원본 데이터:
Sentence: 어지러운 증상이 나타나는 듯합니다.
BIO tag: B-ST I-ST O O O

LLM 입력을 위해 구축된 데이터:
Word | POS | Entity | BIO | Thought
어지러운 |어지러운(VA+ETM)| True | B-ST | Symptom 의 개체명에 속하기 때문임
증상이 |증상(NNG)+이(JKS)| True | I-ST | Symptom 의 개체명에 속하기 때문임
나타나는 |나타나(VV)+는(ETM)| False | O | 용언이기 때문임
듯합니다 |듯(NNB)+합니다(XSA+EF)| False | O | 개체명 범주에 속하지 않은 명사
. |.(SF)| False | O | 부호이기 때문임

(그림 1) LLM 입력을 위해 구축된 데이터 예시

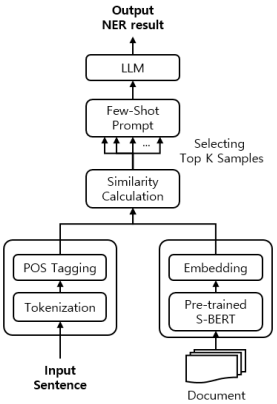
2.3. 유사도 기반 Few-Shot 프롬프트 엔지니어링

질의에 사용될 Few-Shot 은 유사도 기반 검색을 통해 구성되었다. Sentence BERT 를 활용하여 학습 데이터의 문장을 임베딩하였으며, 주어진 입력 문장에 대하여 상위 K 개의 유사 문장을 추출하여 Few-Shot 프롬프트를 구성하였다. 이때 사용된 프롬프트의 구조는 정의(Definition), 질의(Q), Few-Shot 예시(Sample), 입력(Input)로 구성되었다.

Definition: Named entity tags include BodyPart, Symptom, and Disease.
For each entity, the BIO tagging scheme is used, where BodyPart is tagged as B-BD and I-BD, Symptom is tagged as B-ST and I-ST, and Disease is tagged as B-DS and I-DS.
Q: Provide an output based on the sample and fill in the blanks to complete the Output
Sample:
- sample sentence : 어지럼증이 걱정되어 질문하신 것 같습니다 .
- sample output :
Word | POS | Entity | BIO | Thought
어지럼증이 |어지럼증(NNG)+이(JKS)| True | B-ST | Symptom 의 개체명에 속하기 때문임
걱정되어 |걱정(NNG)+되(XSV)+어(EC)| False | O | 개체명 범주에 속하지 않는 명사
질문하신 |질문(NNG)+하(XSV)+신(EP+ETM)| False | O | 개체명 범주에 속하지 않는 명사
것 |것(NNB)| False | O | 개체명 범주에 속하지 않는 명사
같습니다 |같(VA)+습니다(EF)| False | O | 용언이기 때문임
. |.(SF)| False | O | 부호이기 때문임
Input:
- input sentence: 어지럼증이 1 초간 나타났다면 아찔한 느낌 정도가 아니었나 생각됩니다.

(그림 2) 제안된 프롬프트 구성 예시

제안하는 Few-Shot 프롬프트 질의 과정은 그림 3 과 같다. 이 과정은 먼저 입력 문장을 공백 단위로 분리한 단어에 형태소 정보를 태깅하는 초기 단계로 시작된다. 이후, Sentence BERT 를 통해 해당 문장과 유사한 K 개의 문서를 선정하고, 이를 LLM 입력을 위한 Few-Shot 프롬프트로 구성한다. LLM 에 최종적으로 제공되는 입력 형식은 입력 문장, 해당 문장에 대한 K-Shot, 그리고 단어들의 형태소 정보인 Word 와 POS 를 포함한다. 마지막으로, 정리된 프롬프트를 LLM 에 질의하면 Entity, BIO, 그리고 Thought 의 출력을 통해 개체명 인식 결과를 얻게 된다.



(그림 3) 제안하는 Few-Shot 프롬프트 질의 과정

3. 실험 및 결과

제안하는 프롬프트기반 개체명 인식 기법의 검증을 위해 Open AI 에서 발표한 거대 언어 모델인 'gpt-3.5-turbo-16k' 모델과 'gpt-4' 모델을 API 를 통해 호출하여 사용하였다. 비교 실험을 위해 사전 학습된 언어 모델인 'bert-base-multilingual-cased'를 활용하여 실험을 진행하였다. 평가에는 각 개체명 태그에 대한 F1-Score 를 성능에 대한 평가 지표로 사용하였다.

<표 2> 모델별 성능 비교 평가 결과

	bert-base-multilingual-cased	gpt-3.5-turbo-16k			gpt-4
		1-shot	3-shot	5-shot	5-shot
B-DS	0.75	0.67	0.77	0.77	0.81
I-DS	0.20	0.22	0.26	0.34	0.30
B-ST	0.62	0.58	0.66	0.68	0.73
I-ST	0.63	0.46	0.49	0.55	0.60
B-BD	0.79	0.64	0.71	0.74	0.80
I-BD	0.00	0.42	0.43	0.46	0.53
TOTAL	70.2%	60.6%	67.9%	70.1%	75.3%

실험을 통해, 'gpt-3.5-turbo-16k' 모델의 경우 Few-Shot 의 개수가 증가함에 따라 성능 향상이 관찰되었다. 특히, 'gpt-4' 모델에서 5-shot 프롬프트를 적용했을 때 F1-Score 는 75.3%로, 다른 모델들에 비해 가장 우수한 성능을 보였다.

4. 결론

본 논문에서는 한국어의 교착어 특성을 고려하기 위해 형태소 정보를 활용한 Few-Shot 프롬프트로 거대 언어 모델에서의 헬스케어 도메인 개체명 인식 방법을 제안하였다. 실험 결과, 'gpt-4' 모델에서 5-shot 프롬프트를 적용했을 때 가장 높은 성능을 보였으며, 이 방법은 헬스케어 도메인에서의 한국어 개체명 인식에 효과적임을 확인하였다. 본 논문에서 제안하는 방법을 통해 거대 언어 모델의 활용 가능성과 그 방법론적 적용 가이드를 제공함으로써, 헬스케어 및 다양한 분야에서의 자연어 처리 연구에 중요한 기여를 할 것으로 기대된다.

Acknowledgement

본 논문은 2023 년도 정부(교육부)의 제원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2021R1G1A1006381)

참고문헌

[1] G. Lample et al., “Neural architectures for named entity recogni-tion,” arXiv preprint arXiv:1603.01360, 2016.
[2] D. Jacob et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv:1810.04805, 2018.
[3] Openai. ChatGPT [Large language model]. https://chat.openai.com/chat
[4] Wei et al., “Chain of Thought Prompting Elicits Reasoning in Large Language Models”, CoRR abs:2201.11903, 2022.