문장 임베딩 모델 기반 112 신고데이터 학습데이터 레이블링

권은정, 변성원, 박현호, 장동만, 정의석
한국전자통신연구원

ejkwon@etri.re.kr, swbyon@etri.re.kr, hyunhopark@etri.re.kr, dmjang@etri.re.kr, esjung@etri.re.kr

Labeling of 112 report data training data based on sentence embedding model

Eun-Jung Kwon, Sungwon Byon, Hyunho Park, Dongman Jang, Euisuk Jung
Electronic and Telecommunications Research Institute (ETRI)

요 약

본 논문은 112 신고접수 과정에서 사건 종별 유형에 따라 생성되는 다양한 텍스트 데이터의 레이블링 처리에 관한 내용이다. 인공지능 기반 사건 유형별 현장 상황에 대응할 수 있는 모델 개발을 목적으로 치안현장의 업무처리 기준에 따라 생성,기록되는 데이터에 대한 이해가 선행되어야 한다. 이를 바탕으로 다양한 형태의 112 신고 사건에 대한 의사결정 지원 모델의 훈련에 필요한 데이터 레이블 작업을 효과적으로 처리하기 위해 본 논문은 112 신고데이터 사건 종별 문장 단위의 효과적인 레이블링 방법을 제안하고, 실험을 통해 제안 방법의 향후 연구 결과를 도출한다.

Ⅰ. 서 론

문장 임베딩 모델은 벡터 공간 상에 자연어 형태의 문장을 컴퓨터가 이해하도록 숫자 또는 벡터형태로 변환하는 과정을 말한다. 대개 띄어쓰기 단위와 일치되는 어절을 기준으로 단어에 대해 벡터형태로 변환한다. 어절단위로 구분된 단어는 한국어 형태소 분석에 따라 품사 형태별로 구분하여 세밀하게 표현할 수 있고, 추가적으로 의미별로 벡터와의 관련성을 표현하여 좀더 정교한 문장 임베딩 처리가 가능하다.

단어 임베딩의 경우 개별 단어에 한하여 벡터로 표현하기 때문에 동음 이이어의 경우 의미가 다르더라도 단어의 형태가 동일한 값으로 표현되기 때문에 문장에 대한 의미를 정확하게 구분하기 어렵다. 이에 반해, 문장 임베딩은 전체 문장의 흐름을 파악해 개별 단어를 벡터로 표현할 수 있기 때문에 문맥적 의미를 표현할 수 있다는 장점이 존재한다.

본 논문에서는 치안 도메인에서 사용되는 말뭉치를 문장 임베딩 모델을 통해 사건종별 구분에 따라 레이블링할 수 있는 반자동 처리하는 과정을 제안하고 실험한다. 한국어 문장 임베딩 사전학습 언어 모델을 활용하여 문장간 유사도 값 비교를 통해 문장이 내포한 실질적 의미에 더 적합한 값을 레이블 값으로 조정하여 모델 학습용 데이터에 활용하고자 한다.

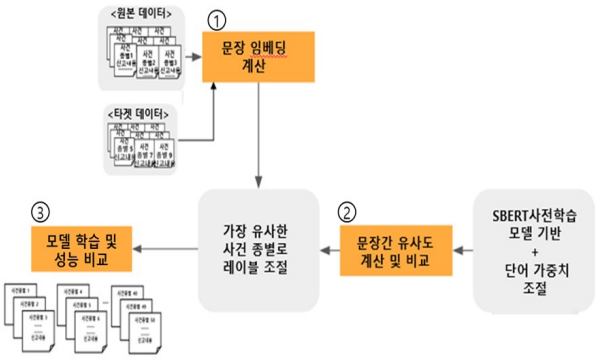
Ⅱ. 관련연구

문장들의 의미에 대한 유사성은 자연어 추론 이란 주제로 연구되어 왔다. [1]은 이를 자연어 이해(Natural Language Understanding)를 기반으로 문장 쌍 분류(Sentence-Pair Classification)의 일종으로 두 문장 사이의 의미 관계를 함의, 모순, 중립으로 구분하여 처리하는 분류 모델로 설명된 바 있다.

자연어 처리에서 의미적으로 유사한 문장들은 가까운 벡터 스페이스(vector space) 상에 표현하여 문장 임베딩을 수행하게 된다. 최근 BERT (Bidirectional Encoder Representations from Transformers)의 문장 임베딩을 응용하여 문장 임베딩 값을 얻을 수 있는 Sentence BERT(SBERT)는 개별 문장에 대한 고정크기의 임베딩 벡터값을 획득한 이후, 유사도 측정 함수를 이용하여 의미적으로 문장들 간의 유사도에 대한 점수를 계산할 수 가 있다[2]. 이러한 SBERT 의 특징을 이용해 다국어로 표현되는 문장에 대한 의미론적 유사성 비교에 대한 연구가 수행된 바 있다. [3]은 유사한 단일 언어 문장을 벡터 공간 상에 가깝게 배치한 다음 다른 언어에서 동일한 의미의 문장을 이와 가까운 벡터 공간에 정렬하도록 모델 훈련한 성능 결과를 제시하여 문장 유사도 기반 모델 성능에 대한 연구가 수행된 바 있다..

Ⅲ. 문장 임베딩 기반 학습데이터 레이블링

본 장에서는 112 신고로 접수된 데이터를 범죄 사건 종별로 문장 임베딩 기반 사전학습 모델을 활용한 레이블링 하기 위한 방법을 기술한다. 112 긴급번호로 접수되는 각종 신고 중 신고센터에 정식으로 접수되는 사건 내용에 대한 구분은 '사건종별' 기분으로 50 여개로 구분된다. 그 중 '보호조치'로 신고 접수된 내용은 주취자, 소음 또는 행패소란 등과 같이 '소음','행패소란', 또는 '청소년비행' 항목과도 구분이 모호할 수 있다. 이러한 상황을 고려하여 본 논문에서는 문장의 의미 유사성을 비교할 수 있도록 BERT 의 문장 임베딩 성능을 개선한 SBERT(Sentence-BERT)를 사용하였다. SBERT 는 두개의 문장으로부터 의미의 유사도를 계산하여 유사도 값의 크기에 따라 레이블을 정의할 수 있다. 따라서, 원시 데이터의 레이블 값이 명확하지 않더라도 재조정을 통해 모델 학습용 데이터로 품질을 높일 수 있다는 전제로 실험을 수행하였다. 다만, 문장 임베딩 사전학습 모델이 치안 도메인의 말뭉치를 학습하지 않은 상태이기 때문에, 논 논문에서는 유사도 점수 10 점 기준 8 점 이상인 경우에 유사성의 정확성을 신뢰기준으로 하여 최종 레이블 값을 선택하였다.



(그림 1) 문장 임베딩 모델 기반 학습용 데이터 레이블 조절을 위한 워크플로우 흐름도

그림 1 은 본 논문에서 제안하는 레이블링이 명확하기 않은 원시 데이터를 문장 간 의미적 유사도 계산을 통하여 레이블링 값을 재조정하기 위한 처리 순서이다. 그림 1 의 (1)을 학습용 원본 데이터세트와 레이블을 조정할 타겟 데이터를 선별하여 SBERT 를 이용하여 문장 임베딩 값을 계산한다. (2) 레이블 값 조정 대상이 되는 타켓 데이터세트의 문장 임베딩 값과 (1)에서 계산된 학습용 데이터세트의 문장 임베딩 값을 비교한다. (3) 유사도 값이 가장 큰 사건종별의 레이블 값으로 타켓 데이터세트의 레이블 값을 조정한다.

IV. 실험 결과

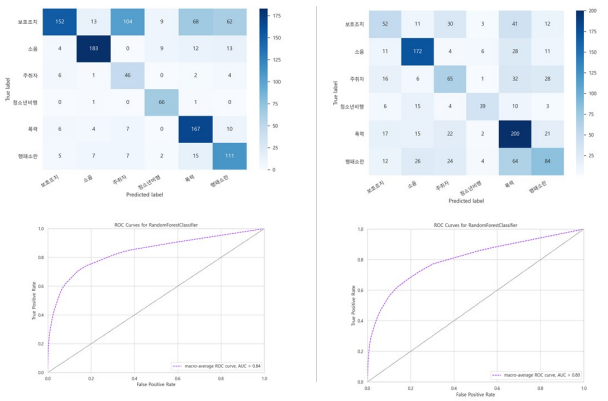
본 논문의 실험을 위해 사용되는 데이터세트 일부 예시는 표 1 과 같다. 실험에 사용된 학습용 데이터는 '사건종별'로 신고내용을 구분하여 생성된 가상 데이터이다. '보호조치'의 신고내용 중 일부는 '주취자', '소음', '청소년비행' 등의 항목과 유사한 의미를 포함할 가능성이 높기때문에 실험을 통해 '보호조치' 사건종별 대한 신고내용을 '주취자', '소음', '청소년비행', 소란행패', '폭력'으로 재분류하고, 학습용 데이터 레이블링 조정 이전과 이후에 대한 머신 러닝 기반 분류기 Random Forest Classifier 와 SVM Classifier 를 각각 비교하였다.

(표 **1**) 실험에 사용된 학습용 데이터세트 샘플 예시

	신고내용	사건종별
0	취객 남자가 쓰러져있다고	주취자
1	남자가 쓰러져 있다고//물무늬 옷// 난간에 가려져 있다고	보호조치
2	주취 남자가 거리에서 자고 있다며	보호조치
3	술취한 아줌마가 자고있다	주취자
4	정문앞에 취객이 자고있다고	주취자
...	...	...
1378	술취한 분이 정신을 못차리고 있다고..	주취자
1379	마트 5층 사우나/술 마신 손님이 자다가 바닥에 떨어졌다/자고 있는데 아무리 깨...	보호조치
1380	남자가 쓰러져잇고주취상태로 추정	주취자
1381	버스정류장에서 어떤아저씨가 도로 가운데서 계속 ...	보호조치
1382	남학생이 따라왔는데정상이 아닌 학생 같다고	청소년비행

(표 2) 실험 결과

모델	구분	Accuracy	F1-Score
Random Forest (+ Random Oversampling)	원본	66% (79%)	66% (78%)
	레이블조절	56% (73%)	57% (74%)
SVM Classifier (+ Random Oversampling)	원본	78% (95%)	77% (95%)
	레이블조절	63% (81%)	63% (82%)



<a. 원본 데이터에 대한 모델 성능 확인>　　<b 레이블값 조정 후 모델 성능 확인>

(그림 1) 문장 임베딩 사전학습 모델을 통한 학습용데이터 레이블 조정을 통한 Random Forest Classifier 적용 실험결과

사용된 학습 데이터는 전체 3,636 건으로 1,383 건의 '보호조치'에 대한 신고 내용이 소음(726), 주취자(496), 청소년비행(286), 폭력(916), 행패소란(751), 그리고 보호조치(481)로 조정되었다. 표 1 과 같이 실험을 통해 레이블 조절후의 모델 Accuracy 및 F1-Score 가 레이블 조절 전 보다 성능이 개선되지는 않았다. 하지만, 모델 Accuracy 대비 F1-Score 값이 레이블 조절 결과 상대적으로 개선됨을 알 수 있다.

V. 결 론

본 논문은 실험에 사용된 데이터 세트 크기가 작고, 사전 학습모델을 치안 도메인 언어로 재학습하지 않은 베이스모델을 사용하여 성능에 큰 영향을 미치지 못함을 알 수 있었다. 향후, 문장 임베딩 모델을 재학습하고, 사건종별 주요 키워드를 가중치 값으로 적용하여 문장의 의미 기반 레이블링 처리 방법을 개선하고자 한다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: 112 긴급출동 의사결정 지원 시스템/연구개발과제번호: PR08-03-000-21]

참 고 문 헌

[1] 김슬기, 김홍진, 김학수, "의존 구문 분석을 활용한 자연어 추론", 제 33 회 한글 및 한국어 정보처리 학술대회 논문집, pp 189~194, 2021

[2] Nils Reimers, Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", EMNLP 2019, arXiv:1908.10084v1

[3] Jiyeon Ham, Eun-Sol Kim , "Semantic Alignment with Calibrated Similarity for Multilingual Sentence Embedding", EMNLP 2021, pages 1781- 1791, November 7- 11, 2021