

KoBART 기반의 Seq2Seq 모델을 활용한 생도생활 예규 챗봇

박성혁, 타이, 권현*

육군사관학교 AI·데이터과학과

austin021022@gmail.com, tranthai.army.kr@gmail.com, *hkwon.cs@gmail.com(교신저자)

SOP Chatbot Using a KoBART-based Seq2Seq Model

Park Seong Hyeok, Thai, Hyun Kwon

Department of AI Data Science at Korea Military Academy

요 약

최근 군인에게 적법한 명령의 판단 여부가 요구되나, 예하 지휘관·부사관은 법적 자문 부재로 규정 적용 및 의사 결정에 어려움을 겪고 있다. 이를 해결하기 위해 본 연구는 생도생활 예규 50페이지를 JSONL 데이터셋으로 구조화하고, test.json을 통해 QnA 쌍을 생성하여 RAG 기반 외부 지식 저장소와 학습용 QnA 데이터를 구축했다. KoBART를 HuggingFace Seq2SeqTrainer로 30epoch 미세조정하며 batch size 4, learning rate 2e-5, FP16 혼합 정밀도 적용으로 효율적 학습을 수행했다. 평가 결과 BLEU 0.8192, Exact Match 0.7143, ROUGE-1 0.0952를 기록하여 표현 유사성과 정확성은 우수하나 핵심 정보 재현에 제약이 있음을 확인했다. 이 챗봇은 생도 및 학교 구성원이 훈련·복장·훈육 규정에 대해 실시간 해석·절차 안내를 제공하여 규정 준수와 의사결정 속도 향상에 기여점이 있다.

I. 서론

최근 대한민국이 정치적, 사회적 혼란을 겪으며 군에 대한 신뢰도가 저하되었다. 사회는 군인에게 무조건 복종이 아닌 정당한 명령에 대해서만 복종할 것을 요구하고 있다. 그렇다면 정당한 명령의 기준은 무엇인가? 명령의 정당성을 평가하는 요소는 여러 가지가 있지만, 가장 우선시 되어야 하는 것은 적법성이다. 아무리 다른 요소가 정당했다고 하여도 법에 저촉되는 명령은 잘못된 명령이고 복종할 필요가 없다. 그렇다면 현재 사회가 군인에게 요구하는 것은 상관의 명령이 적법한 명령인지 판단할 수 있는 능력일 것이다. 대대장 이상의 상급 지휘관은 법적 참모를 가지지만 예하 지휘관·부사관은 법적 참모가 없다. 이에, 군 관련 법률 지식을 내장한 챗봇을 통해 개별 군인도 법적 자문을 받을 수 있도록 한다면 적법한 명령인지에 대한 판단을 도울 수 있다는 생각에 생도생활의 규칙과도 같은 생도생활 예규를 학습한 챗봇을 만들었다. 이를 기반으로 헌법, 군사법, 군형법, 군인 복무에 관한 기본법 등을 학습시킨 챗봇으로 발전시켜 궁극적으로 국군 전체가 올바른 판단을 할 수 있도록 돕고자 한다.

II. 본론

1. 데이터 구축 및 전처리

1.1 데이터 구축

생도생활 예규라는 데이터의 엄격한 보안 요건을 준수하기 위해 원문 규정 및 표를 수작업으로 구조화된 디지털 형식으로 변환하였다. 본 연구는 구현 가능 여부 파악에 목적이 있으므로 생도생활 예규 중 약 50페이지를 데이터로 변환하였다. 구체적으로, 먼저 원문 규정 및 표를 수작업으로 옮겨 적은 후 각 문서는 개별 조항 또는 표의 행(row)에 상응하는 의미론적으로 일관된 청크(chunk)로 분할되어 JSON 형식으로 직렬화하였다. 본래성(provenance) 보장 및 감사 가능성(auditability) 확보를 위해, 각 텍스트 단위에는 조문번호(article), 항목(clause) 등의 구조화된 메타데이터 필드를 추가하였으며, 표 항목에는 table 보조 필드를 부여하였다. 메타데이터 보강으로 법령 조문과 표 구성 요소 간의 정확한 참조 관계를 유지함으로써, 이후 단계에서의 정밀한 검색 및 결과 검증에 용이하게 하였다. 생성된 모든 텍스트 단편은 로컬 환경에서 실행된 Python 스크립트를 통해 필드 완전성, 라벨링 적합성, 전사 오류 부재 여부를 다단계로 검증하였으며, 특히 자동화 도구를 활용하여 중복 또는 누락 사항을 신속히 식별하고 수정하였다.

보안 규정상 데이터를 공개할 수 없으므로 그림 1과 같은 방식으로 헌법에 적용해 이해를 돕고자 한다.

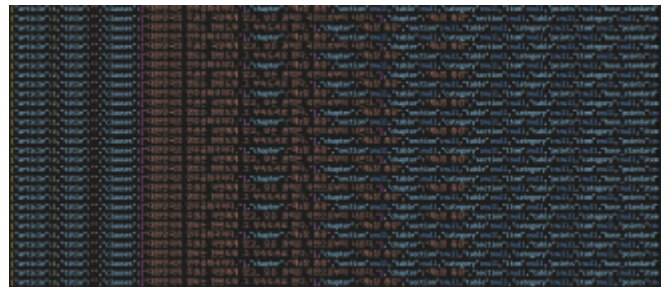


그림 1. 헌법에 적용한 전처리 예시

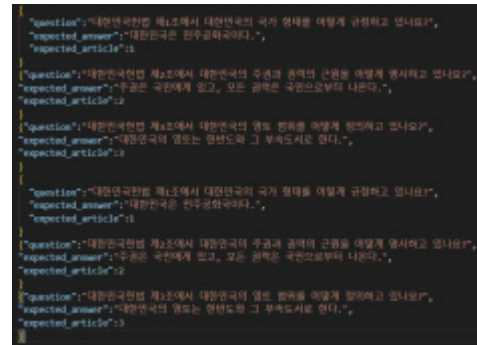


그림 2. 헌법에 적용한 전처리 예시

또한, 지도형 파인튜닝(supervised fine-tuning)을 위해 QA(question-answer) 데이터셋을 체계적으로 구축하였다. 각 조항 및 표 항목에 대하여 형식적인 한국어 표현으로 2~3종의 문법적 변형 질문을 작성함으로써 모델이 다양한 구문 구조에 대응할 수 있도록 하였으며, 정답(answer)은 해당 텍스트 단편에서 직접 추출하여 초기 모델 성능 평가 결과를 기반으로 반복적으로 보완·정제하였다. 그림 1과 동일하게 이해를 돕기 위해 헌법 조항에 동일한 방식을 적용한 결과는 그림 2와 같다.

1.2 데이터 전처리

데이터 전처리로 두 가지 방법을 수행하였다. 첫째, 챗봇의 외부 지식 저장소로 활용하기 위해 rag_dataset.jsonl 파일의 조

항, 제목, 세부항목, 표 등의 데이터를 구조적으로 분해하였다. 조항 번호와 제목을 결합한 접두어를 생성하고, 텍스트 및 표 형식의 세부항목을 간결한 문장으로 재구성하여 RAG 시스템의 문서 저장소용 문장 리스트로 만들었다.

둘째, test.json 파일의 질의-응답 쌍을 HuggingFace의 Dataset 형식으로 변환하여 모델 학습 데이터로 활용하였다. 입력과 출력에 각각 max_length=128 및 truncation을 적용하고, label을 명시해 teacher forcing 기반의 지도 학습 환경을 구성하였다.

2. 모델

생도생활 예규의 질의응답(Seq2Seq QA) 시나리오를 구현하기 위해 한국어 전용 인코더-디코더 모델인 KoBART를 사용했다. 일반적으로 국내 법령·행정 문체는 긴 복합 문장, 숫자·기호 기반 조항 표기(예: “제7조 ①항”), 그리고 격식적 전문용어가 빈번하다는 특징이 있다. 따라서 질의와 조항을 동시에 해석하면서 문맥을 압축·요약할 수 있는 구조, 한국어 형태소 변화에 최적화된 어휘 사전, 그리고 제한된 학습 자원을 고려한 효율성을 동시에 충족하는 모델이 필요했다. KoBART는 약 1.5 억 문장 규모의 한국어 대규모 코퍼스로 사전학습되어 다층적 어순과 경어법·한자어를 포함한 문체적 변이를 충분히 학습했으며, 이는 공문서와 유사한 생도생활 예규 텍스트를 처리하는 데 유리하다고 판단하였다. BART의 구조는 그림 3과 같다.

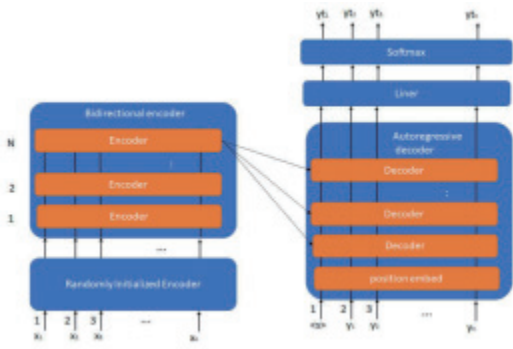


그림 3. BART 구조

KoBART의 인코더-디코더 구조는 질문과 규정 전문을 양방향 인코더로 통합적으로 이해한 뒤, 디코더에서 순차적으로 핵심 정보를 생성함으로써 “장문의 규정 → 응답 요약”이라는 흐름에 자연스럽게 부합한다. 그리고 생도생활 예규 데이터는 조·항·호 계층 구조가 뚜렷하고, 한 조항이 60 ~ 120 토큰으로 길어 정보 밀도가 높다. KoBART의 양방향 인코더는 조항 전체에서 흩어진 단서를 통합해 ‘명예의 정의’처럼 조문 전반에 분산된 정보를 결합할 수 있으며, 고정된 공식 문체를 기반으로 정제된 자연어 응답을 생성할 수 있다.

3. 학습

모델 학습은 HuggingFace의 Seq2SeqTrainer를 활용하여 KoBART 모델을 미세조정(fine-tuning)하는 방식으로 진행되었다. 학습 초기 단계에서는 KoBART 사전학습(pre-trained)가중치를 불러와 효율적인 학습과 빠른 수렴을 도모하였다. 학습 과정은 GPU 자원 및 효율적 학습 관리를 고려하여 배치 크기(batch size)는 장치당 4로 설정하고, 학습률(learning rate)은 2e-5로 고정하였다. 총 학습 에폭(epoch)은 30회로 설정하여, 데이터셋에 포함된 질의응답 패턴을 충분히 학습할 수 있도록 하였다. GPU 자원 활용의 최적화를 위해 FP16 혼합정밀도(floating-point 16) 설정을 적용하여 메모리 사용을 감소시키고 학습 속도를 향상시켰다. 또한, 학습 진행 상태는 매 에폭 종료 시 손실(loss)을 기록하여 모니터링했으며, 과적합을 방지하고 최적의 checkpoint를 관리하기 위해 최신 epoch

의 모델 checkpoint만 유지하도록 하였다.

학습 결과를 평가하기 위해 train data에서 사용한 QnA의 질문을 유사한 형태의 여러 질문으로 변환하여 그림 4와 같이 제작하였다.

기존 질문:	"question": "제7조(사관생도의 명예)에서 '명예'의 정의와 그 중요성은 무엇입니까?"
변형한 질문:	"question": "제7조(사관생도의 명예) 조항에서 '명예'의 의미가 무엇인지 설명해 주세요."
	"question": "제7조에 명시된 '명예'의 개념과 그 핵심 가치는 무엇인지 알려주세요."
	"question": "사관생도의 명예(제7조)에 대해 정의와 중요성을 구체적으로 기술해 주십시오."
기존 질문:	"question": "제7조에서 제시하는 사관생도 신조 제 3항(국가 헌신·명예·절의)의 구체적 내용은 무엇입니까?"
변형한 질문:	"question": "제7조에 나오는 '사관생도 신조'의 세 항목(국가 헌신·명예·절의)의 구체적인 문구를 설명해 주세요."
	"question": "제7조에서 규정한 국가 헌신, 명예, 절의 세 조항의 구체적 내용을 알려주세요."
	"question": "'사관생도 신조(제7조)'를 구성하는 세 조항의 문장별 의미를 설명해 주세요."

그림 4. 훈련데이터에서 여러 질문을 변형한 예시

기존의 QnA 데이터 20개 중 6개를 랜덤으로 추출해 위와 같이 3개의 비슷한 질문으로 성능 평가를 진행하였다. 그 결과는 그림 5와 같다.

Q	제7조(사관생도의 명예) 조항에서 '명예'의 의미와 그 중요성은 무엇인지 설명해 주세요.
Ref	명예는 사관생도의 기본 품성과 가치로, 위대한 신 자질을 유지하기 위한 핵심 요소입니다.
Pred	명예는 사관생도의 기본 품성과 가치로, 위대한 신 자질을 유지하기 위한 핵심 요소입니다.
Q	제7조에서 명시한 '명예'의 개념과 그 핵심 가치는 무엇인지 알려주세요.
Ref	명예는 사관생도의 기본 품성과 가치로, 위대한 신 자질을 유지하기 위한 핵심 요소입니다.
Pred	명예는 사관생도의 기본 품성과 가치로, 위대한 신 자질을 유지하기 위한 핵심 요소입니다.
Q	사관생도의 명예(제7조)에 대해 정의와 중요성을 구체적으로 기술해 주십시오.
Ref	명예는 사관생도의 기본 품성과 가치로, 위대한 신 자질을 유지하기 위한 핵심 요소입니다.
Pred	명예는 사관생도의 기본 품성과 가치로, 위대한 신 자질을 유지하기 위한 핵심 요소입니다.
Q	제7조에서 나오는 '사관생도 신조'의 세 항목(국가 헌신·명예·절의)의 구체적인 문구를 설명해 주세요.
Ref	제7조(사관생도의 명예) 조항에서 '명예'의 의미와 그 중요성은 무엇인지 설명해 주세요.
Pred	제7조(사관생도의 명예) 조항에서 '명예'의 의미와 그 중요성은 무엇인지 설명해 주세요.
Q	제7조에서 규정한 국가 헌신, 명예, 절의 세 조항의 구체적 내용을 알려주세요.
Ref	제7조(사관생도의 명예) 조항에서 '명예'의 의미와 그 중요성은 무엇인지 설명해 주세요.
Pred	제7조(사관생도의 명예) 조항에서 '명예'의 의미와 그 중요성은 무엇인지 설명해 주세요.

그림 5. 테스트 데이터로 성능평가한 예시

Eval loss는 0.0903으로 0.03755인 train loss에 비해 높게 평가되었다. 기계 번역 평가 지표로 제안된 문장과 정답 문장 간의 n-gram 정밀도(precision)를 측정하는 Eval BLEU(Bilingual Evaluation Understudy)는 0.8192로 정답과 유사한 표현으로 문장을 생성하고 있음을 알 수 있다. 모델이 생성한 출력이 “정답 문장과 1:1로 완전히 일치”한 비율을 나타내는 Exact Match는 0.7143으로 높은 성능을 내고 있음을 확인할 수 있다. 다만, 생성된 문장 안에 정답 문장의 단어(1그램)가 얼마나 많이 “포함되었는지”(재현되었는지) 측정하는 리콜(recall) 기반 점수인 ROUGE-1 (Recall-oriented Understudy for Gisting Evaluation, unigram)가 0.0952로 출력값이 정답값이 가지는 정보에서의 누락이 심한 것을 알 수 있다. 위 값들을 종합적으로 해석하면 모델은 정답과 유사한 표현을 자주 출력하지만, 실제로 정답 문장에서 중요한 단어들을 모두 포함하는 데에는 한계가 있다.

III. 결론

본 연구에서 육사 생도생활예규 전용 RAG 데이터셋을 구축하고, KoBART 기반 Seq2Seq 챗봇을 통해 생도생활 예규에 관한 실시간 법·규정 해석 및 절차 안내 시스템을 구현했다. 실험 결과, 기계번역 지표인 BLEU 0.8192, Exact Match 0.7143, ROUGE-1 0.0952의 성능을 확인하며, 표현의 유사성은 우수하나 핵심 정보 재현에 일부 한계가 있음을 보였다. 이러한 챗봇은 생도 및 학교 구성원들이 교내 혼란·훈육·복장 규정 등에 대해 즉각적이고 일관된 답변을 제공받아 의사결정 속도를 높이고 잘못된 해석으로 인한 징계나 오해를 사전 예방하는 데 기여할 수 있다. 다만, 낮은 ROUGE와 생도생활 예규에만 국한된 챗봇이라는 점에서 한계점이 있다.

참고문헌

- [1] 백인걸, 양준호, 이환희. (2024-06-26). KoSaul: 한국어 법률 언어 모델. 대한전자공학회 학술대회, 제주.
- [2] 정채연. (2024). 생성형 AI를 활용한 법률서비스의 쟁점과 과제. 법학연구, 35(3), 401-443. 10.33982/clr.2024.8.31.3.401
- [3] 박상민, 윤예진, 이재윤, 김재은. (2021-06-23). KoLegal-BERT: 법률 도메인 텍스트 마이닝을 위한 법률 언어 표상 모델. 한국정보과학회 학술발표논문집, 제주.