J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

자연어 처리 및 협업 필터링 기반의 전장상황 관련 문서 자동탐색 및 요약 기법연구☆

A Study on Automatic Discovery and Summarization Method of Battlefield Situation Related Documents using Natural Language Processing and Collaborative Filtering

김 건 영[1] 이 정 빈[1] 손 미 애[1*]
Kunyoung Kim Jeongbin Lee Mye Sohn

요 약

정보통신기술이 발달함에 따라 전투공간에서 생산·공유되는 정보 및 체계 내 저장·관리되는 정보의 양이 폭발적으로 증가하였다. 이는 지휘관이 전장상황 인식 및 지휘결심을 수행하는 데에 활용할 수 있는 정보의 양이 증가하였음을 의미하지만, 한편으로는 지휘관의 정보 부담을 증가시킴으로써 신속한 지휘결심을 저해하는 요인이 되기도 한다. 이러한 한계를 극복하기 위해, 본 연구에서는 지휘관이 전장상황 보고 문서를 수신하였을 때, 체계 내 보유 문서 중에서 이를 해석하는 데에 도움을 줄 수 있는 문서들을 자동적으로 탐색 및 선별하고 요약하는 기법을 제안하였다. 첫째로, 개체명 인식 방법을 활용하여 수신된 전장상황 보고 문서로부터 개체들을 식별한다. 둘째로, 각 개체와 관련된 체계 내 보유 문서들을 탐색한다. 셋째로, 언어모델과 협업 필터링을 활용하여 이러한 문서들을 선별한다. 이때 언어모델은 수신된 보고 문서와 탐색된 문서 간의 유사도를 산출하기 위해 활용되고, 협업 필터링은 지휘관의 문서 열람 히스토리를 반영하기 위해 활용된다. 마지막으로, 선별된 문서들로부터 각 개체가 포함된 문장을 선별하고 이를 정렬한다. 실험은 군 문서와 비슷한 특성을 지니는 학술논문들을 활용하여 수행하였고, 제안된 방법의 타당성을 검증하였다.

☞ 주제어: 전장상황 인식, 지휘결심, 자연어 처리, 언어 모델, 협업 필터링, 개체명 인식

ABSTRACT

With the development of information and communication technology, the amount of information produced and shared in the battlefield and stored and managed in the system dramatically increased. This means that the amount of information which cansupport situational awareness and decision making of the commanders has increased, but on the other hand, it is also a factor that hinders rapid decision making by increasing the information overload on the commanders. To overcome this limitation, this study proposes a method to automatically search, select, and summarize documents that can help the commanders to understand the battlefield situation reports that he or she received. First, named entities are discovered from the battlefield situation report using a named entity recognition method. Second, the documents related to each named entity are discovered. Third, a language model and collaborative filtering are used to select the documents. At this time, the language model is used to calculate the similarity between the received report and the discovered documents, and collaborative filtering is used to reflect the commander's document reading history. Finally, sentences containing each named entity are selected from the documents and sorted. The experiment was carried out using academic papers since their characteristics are similar to military documents, and the validity of the proposed method was verified.

☞ keyword: Battlefield situation awareness, Decision making, Natural language processing, Language model, Collaborative filtering, Named entity recognition

1. 서 론

NCO(Network-Centric Operation)[1]는 센서에서 슈터에 이르는 모든 전장 요소의 솔기 없는 연결과 통합을 통해 성공적인 군사 작전의 수행을 달성한다. 이 때, 지휘관들은 다양한 전장 요소들이 수집・전송한 데이터를 활용해 신속・정확하게 전장상황을 인식한 후 최적의 방책을 수립하는 등의 지휘 결심을 수행한다[2]. 즉, 전장 상황에

1 Dept. of Industrial Engineering, Sungkyunkwan University, Suwon, 16419, Korea
* Corresponding author: myesohn@skku.edu
[Received 13 October 2023, Reviewed 24 October 2023(R2 9 November 2023), Accepted 15 November 2023]
☆ 이 논문 또는 저서는 2023년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019R1A2C1004102)

대응하기 위한 최적의 방책을 수립하기 위해서는 전장상황에 대한 명확한 인식이 우선되어야 한다[3]. 지휘관들은 전장상황을 인식하기 위해 다양한 정보와 첩보를 활용한다. 이러한 정보 중의 하나가 트랙보고이다[4]. 트랙보고는 탐지 대상의 위치, 경로, 방향, 속도 등을 포함하는 보고문서이다[5]. 트랙보고는 특정 시점의 전장 상황을 기술하는 사실 데이터로 전장 상황을 인식하는 기본 정보로 활용한다.

이후 지휘관들은 트랙보고를 근거로 관련 첩보나 교리, 작전 계획 등과 같은 관련 문서를 탐색・열람한 후 지휘결심을 수행한다. 그러나 정보의 출처가 다양해짐에 따라 지휘관들이 탐색・열람해야 할 문서의 양이 폭발적으로 증가하고 있으며, 이는 지휘관의 지휘결심을 어렵게 하는 장애 요소가 될 수 있다[6]. 이에 본 연구에서는 지휘관의 정보 처리 부담을 줄여 줌으로써 신속・정확한 지휘결심을 지원하는 인공지능(Artificial Intelligence, AI) 기술인 언어모델(Language Model)과 협업 필터링(Collaborative Filtering) 기반의 정보 선별 및 요약 기법을 제안한다.

이를 위해, 군용 자연어 문서 데이터(Natural Language Data, NLD)[7]의 특징을 분석하였다, 첫째, 군용 문서는 작성 목적이 명확하고 분명하다. 즉, 문서를 통해 전달하고자 하는 내용이 해당 문서에 명시적으로 표현되어 있다. 둘째, 군사 용어는 그 개념 및 정의 등이 체계적으로 정립되어 있다. 즉 지칭하는 각 대상이 명확한 명칭 및 식별자로 표현되어 있다. 셋째, 공적인 어휘, 표현 및 문장을 활용한다. 비속어, 은어 등은 거의 활용되지 않는다. 넷째, 보안이 강조되는 국방 도메인의 특성으로 인해 자연어 처리의 대상이 되는 타도메인(소셜 미디어 포스트, 전자상거래 리뷰 등)에 비하여 자연어 문서 데이터셋을 구축하고 활용하기가 용이하지 않다.

이러한 군 NLD의 특징을 고려하여, 본 연구에서 제안한 전장상황 관련 문서 자동 탐색 및 요약 프레임워크는 다음과 같은 기능을 탑재한다. 첫째, 개체명 인식(Named Entity Recognition, NER)[8] 기법을 활용해 수신된 전장상황 보고 문서(트랙보고 등)로부터 주요 개체를 식별한다. 그러나 군 NLD의 양이 충분하지 않기 때문에 이것만으로 NER 모델을 학습시켜 활용하는 것은 현실적으로 어렵다. 이에 본 연구에서는 사전 학습된 NER 모델을 활용하여 주요 개체를 식별한다. 군 NLD는 비속어, 은어 등이 활용되지 않으며, 개념 및 용어가 체계적으로 정리되어 있기 때문에 사전 학습된 NER을 활용해 전장상황 보고 문서에 포함된 개체들을 발견하는 것은 어렵지 않다, 두 번째, 식별된 주요 개체와 관련된 문서 리스트를 일차적으로 도출한다. 이때, 문서는 체계와 연동된 모든 데이터베이스에 대한 쿼리를 통해 수행한다. 군사 용어의 경우, 의미와 정의가 명확하기 때문에 키워드 검색만으로도 관련 문서를 찾을 수 있다. 후보 문서 리스트를 대상으로 사전 학습된 언어모델[9]과 협업 필터링[10]을 적용하여 최종 추천 문서 리스트와 그들의 추천 랭킹을 도출한다. 이때 언어모델은 상황 보고 문서와 후보 문서 간의 유사도를 산출하기 위해 활용되며, 협업 필터링은 정보이용자들의 정보 이용 히스토리를 반영하기 위해 활용한다. 마지막으로, 최종 리스트에 포함된 문서들에 대해 개체명을 포함하고 있는 문장들을 선별하여 요약문을 생성한다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구들을 다루고, 3장에서는 간단한 예시 시나리오와 함께 본 논문에서 제안한 방법에 대해 상세히 설명한다. 4장에서는 실험과 그 결과에 대해 설명하고, 5장에서는 결론과 추후 연구에 대해 다룬다.

2. 관련 연구

2.1 군 문서 자연어 처리

자연어 처리 기술이 최근 획기적으로 발전함에 따라 군 분야에서 자연어 처리 기술을 적용하기 위한 연구들이 활발히 수행되고 있다. [7]에서는 말뭉치 기반(corpus-based) 접근법 및 규칙 기반(rule-based) 접근법을 모두 활용하여 군 도메인 전문지식에 대한 용어를 식별하고 지식을 추출하기 위한 연구를 수행하였다. [11]은 군 도메인에 적합한 사전학습 언어 표현 모델을 제안하였는데, 이때 지식 통합 및 엔터티 교체 방법을 통해 외부 사전 지식이 모델에 반영될 수 있도록 하였다. [12]는 공개되어 있는 군 NLD를 수집하였고, 다양한 NLP(Natural Language Processing) 알고리즘을 이에 적용하여 이들의 군 NLD에서의 효용성을 검증하였다. [13]은 군에서 Chat GPT의 적용가능한 영역들과 향후 역할에 대해 분석하였다.

이처럼 자연어 처리를 군에 적용하기 위한 연구들은 수행되고 있지만, 구체적으로 이를 활용하여 전장상황 인식 및 지휘결심을 어떻게 지원할지에 대한 연구는 아직 미비한 실정이다. 이에 본 연구에서는 실제 전장 보고문서를 수신하였을 때 이로부터 필요한 추가정보를 기존 보유문서로부터 탐색하기 위한 방법을 제시하였다.

2.2 관련 문서 탐색

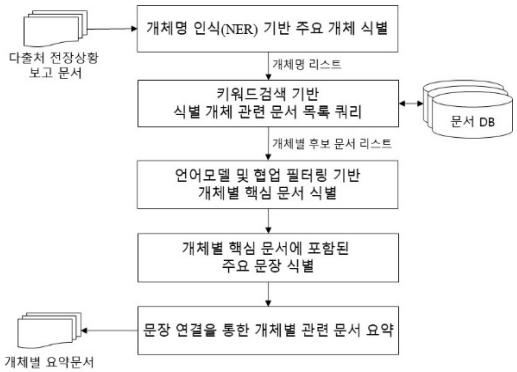
언어모델및 텍스트 마이닝 방법을 활용하여 문서 간의 유사도를 산출하고 관련 문서를 탐색하기 위한 연구는 꾸준히 수행되고 있다. [14]는 제조현장에서 조립에 대한 문제해결 기록 문서들에 대해 개량된 LDA(Latent Dirichlet Allocation) 토픽 모델, TF-IDF(Term Frequency-Inverse Document Frequency) 및 개량된 DBSCAN(Density-based spatial clustering of applications with noise)을 적용하여 텍스트 문서를 클러스터링하고 이로부터 문제와 원인 및 해결책을 발견하기 위한 연구를 수행하였다. [15]는 문서 분류 및 유사도 측정을 위해 활용되는 토픽을 발견하는 데에 기존의 LDA 기반 토픽 모델링이 아닌, 사전 학습된 언어모델을 활용함으로써 문서 외부의 언어적 지식이 문서 토픽 발견에 반영될 수 있도록 하였다. [16]은 자연어 문서를 분할하고 이들 각각에 하위 토픽(sub-topic)을 할당함으로써 문서 클러스터링을 보다 효과적으로 수행하기 위한 방법을 제시하였다.

이처럼 문서 간의 유사도를 산출하고 관련 문서를 탐색하기 위한 연구들이 수행되고 있다. 그러나 본 연구의 목적은 이들과는 달리 단순히 수신된 전장상황 보고 문서와 유사한 체계 내 보유 문서를 탐색하는 것이 아니라, 수신된 문서를 이해하고 지휘결심을 수행하는 데에 필요한 문서를 발견하는 것을 목표로 한다. 이에 본 연구에서는 수신된 문서로부터 우선적으로 추가 정보 탐색이 필요한 개체들을 식별하고, 이들에 대한 정보를 포함하는 문서를 발견하는 방법을 제시하였다.

3. 제안 프레임워크

전투 상황에서 활용할 수 있는 작전 교리, 작전 계획 및 첩보의 형식은 자연어이다. 본 논문에서는 이들 NLD를 두가지 유형으로 분류활용한다. 첫 번째 NLD는 전장상황을 전파하기 위해 실시간으로 생성공유되는 보고 문서나 메시지를 의미하는 '전장상황 보고 문서'이다. 대표적인 예가 트랙보고이다. 이 유형의 NLD는 지휘관 및 참모들이 현재의 전장 상황(situational awareness)을 인식하기 위한 핵심 데이터로서, 특정 상황이 종료된 이후에는 가치가 없어지는 시한성을 지닌다. 두 번째 NLD는 지휘통제체계내 다중의 데이터베이스에 저장된 문서로, 병력이나 무기체계 현황 데이터나 작전 교리, 작전 계획 및 과거에 발생했던 사건사고나 그에 대한 조치 등에 대한 보고 문서 등을 의미한다. 이러한 유형의 데이터는 버전관리가 중요하며 시간에 따른 정보 가치의 변동 폭이 매우 낮다.

자신의 임무 영역에서 발생한 특정 사건사고와 관련된 트랙보고가 수신되면, 지휘관은 이를 종결하기 위한 지휘결심을 수행한다. 효과적인 지휘결심을 수행하기 위해 지휘관은 자신의 역량을 총동원할 뿐만 아니라 다양한 자원, 즉 교리, 작전 계획 및 첩보 문서를 활용한다. 이 때, 정보의 출처가 다양해짐에 따라 지휘관들이 탐색·열람해야 할 문서 양의 폭발적 증가로 신속정확한 지휘결심이 어렵게 한다. 이러한 문제를 해결하기 위해, 본 논문은 '지휘결심을 지원하는 문서 자동 탐색 및 요약 프레임워크'를 아래 그림 1과 같이 제안한다.

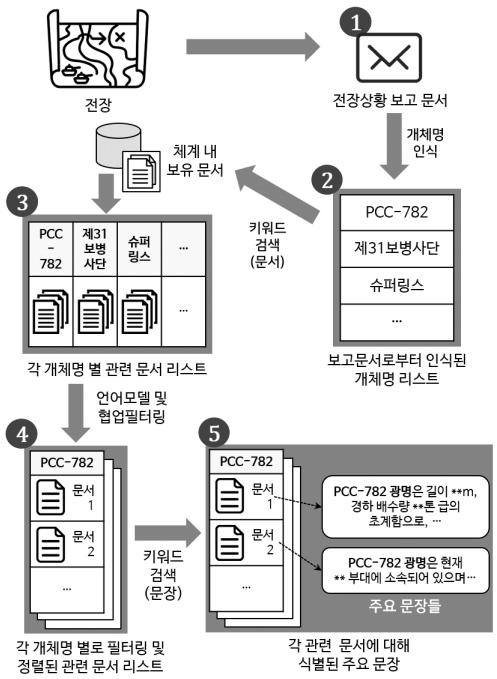


(그림 1) 지휘결심에 필요한 관련 문서 탐색 및 요약 프레임워크
(Figure 1) Relevant documents discovery and summarization framework for decision support of commanders

3.1 예제 시나리오

가상의 시나리오를 이용해 본 논문에서 제안한 프레임워크의 필요성을 입증한다. 시나리오는 1998년 발생한 여수 반잠수정 격침사건을 기반으로 작성하였다. 해군 지휘관 A는 현재 도주 중인 북한 반잠수정에 대한 대응 방책을 수립하기 위해, 인접 부대로부터 보유 병력에 대한 보고 문서를 수신하였다. 이때 지휘관 A가 효과적인 방책을 수립하기 위해서는 인접 부대 병력에 대한 상세 정보(규모, 가용 무기체계 등)와 관련 첩보 등을 열람참고해야 한다. 이를 위해, 제안 프레임워크는 보고 문서로부터 'PCC-782,' '제31보병사단,' '슈퍼 링스' 등과 같은 열람 문서가 포함하고 있어야 할 개념들을 자동으로 식별한 후, 이를 포함하고 있는 문서를 탐색한다. 탐색된 문서들

중에서 반드시 열람참고해야 하는 문서만을 선별하기 위해 지휘관 A의 과거 문서 열람 행태(이력)와, 보고문서와 탐색된 체계 내 보유 문서 간 유사도를 활용한다. 유사도가 높을수록 지휘관의 선호도가 높은, 즉 반드시 열람해야 할 문서로 판명한다. 이를 통해, 지휘관 A는 별도의 정보요구 없이 인접 부대의 병력 정보(PCC-782 등)뿐 아니라, 이의 상세정보(탑재 76mm 함포 등)를 전달받는다. 이를 활용해, 지휘관 A는 PCC-782 광명함에 북한 반잠수정 격퇴하라는 명령을 하달하고 광명함에 탑재된 76mm 함포로 반잠수정을 사격하여 반잠수정을 격침시켰다. 위 시나리오를 기반으로 전장상황 관련 문서 자동 탐색 및 요약 과정을 도식화하면 아래 그림 2와 같다.



(그림 2) 시나리오 기반 관련 문서 자동 탐색 및 요약 과정 도식화
(Figure 2) An illustration of the automatic discovery and summarization process of the battlefield situation-related documents based on a scenario

3.2 개체명 인식 기반 주요 개체 식별

수신한 전장상황 보고 문서에 사전 학습된(pre-trained) NER 모델을 활용하여 추가 정보를 탐색하고자 하는 개체명들을 발견한다. 사전학습 NER 모델의 개발은 본 연구의 범위가 아니므로 더 이상의 언급은 하지 않는다. 수신된 전장상황 보고 문서는 다음과 같이 정의된다.

정의 1. 전장상황 보고 문서 (sr) sr은 트랙보고 등 전장상황 정보를 담고 있는 문서이다. 이는 KMTF 포맷의 문서뿐만 아니라 실시간으로 전송되는 음성이나 단문 보고도 포함한다. sr 보고 문서에 대해 NER을 적용하여 다음과 같이 개체명의 집합 $(NE=(sr))$을 얻는다.

$$NE(sr)=\{ne_1,ne_2,\cdots,ne_m,\cdots,ne_N\} \quad (1)$$

이때, ne_m은 $NE=(sr)$에 포함된 m번째 개체명을 의미한다 $(m=1,2,...,N)$.

3.3 식별 개체 관련 문서 목록 쿼리

개체명을 활용해 체계 보유 문서에 대한 키워드 기반의 검색을 수행하면 해당 개체를 포함하고 있는 관련 문서의 목록을 획득한다. 이 때, 정보 검색의 부담을 줄이기 위해 체계 내 보유 문서(D)를 기반으로 개체명에 대한 필터링을 수행한다. 이때 D는 sr과는 달리 체계 내에서 저장·관리하는 정보 문서들(무기체계에 대한 기능명세서 등)을 의미한다. 개체명 필터링 수행 방법은 다음과 같다. 각 $ne_m(\text{for}\,\forall m)$에 대해, D를 대상으로 역문서 빈도(Inverse Document Frequency, IDF)를 계산해 사전에 지정한 역치보다 작으면 필터링한다. IDF는 ne_m의 정보량을 평가하는 지표로 큰 값을 가질수록 중요한 개체, 즉 정보량이 많은 개체로 판명한다. 필터링된 개체명 $(NEF(sr))$ 집합은 다음과 같이 표현된다.

$$NE(sr)=\{ne_1,ne_2,\cdots,ne_n,\cdots,ne_{NF}\} \quad (3)$$

이때, ne_n은 IDF 기반 필터링된 n번째 개체를 의미한다 $(N \geq NF)$.

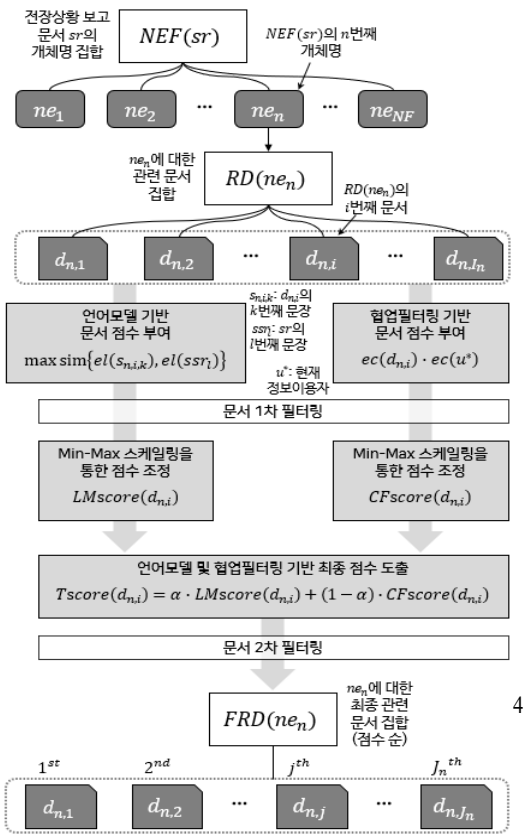
$ne_n(\text{for}\,\forall n)$를 이용해 D에 대한 키워드 검색을 수행하면 ne_n을 포함하고 있는 문서를 발견할 수 있다. ne_n 관련 문서 목록$(RD(ne_n))$은 다음과 같이 표현된다.

$$RD(ne_n)=\{d_{n,1},d_{n,2},\cdots,d_{n,i},\cdots,d_{n,I_n}\} \quad (4)$$

이때, $d_{n,i}$는 이의 i번째 문서를 의미한다($RD(ne_n) \subset D$).

3.4 개체별 핵심 문서 식별

$RD(ne_n)\,(\text{for}\,\forall n)$에 대해, 언어모델과 협업 필터링을 활용하여 문서에 대한 필터링을 수행한 뒤 중요도 순으로 문서들을 정렬한다. 이때 언어모델은 쿼리된 문서와 sr과의 유사도를 산출하기 위해 활용되고, 협업 필터링은 현재 sr의 정보이용자의 문서 열람 히스토리를 바탕으로 각 문서를 열람할 가능성을 산출하기 위해 활용된다. 언어모델 및 협업 필터링 기반 개체별 핵심 문서 식별 과정을 도식화하면 다음 그림 3과 같다.



(그림 3) 언어모델 및 협업 필터링 기반 개체별 핵심 문서 식별 과정 도식화

(Figure 3) An illustration of the key documents recognition process of each entity using language model and collaborative filtering

3.4.1 언어모델 기반 문서 점수 부여

특정 개체명을 포함하고 있는 문서라 하더라도 전장상황 보고 문서 sr과 관련도가 높을수록 지휘결심을 수행하는 데에 핵심적인 문서일 가능성이 높다. 이때, 특정 문서 전체와 sr의 관련성은 낮지만 문서내 특정 문장과 sr과의 관련성이 높은 경우와 해당 문서가 sr 전체가 아니라 일부와 관련성이 높다고 판명되었다면 이들은 반드시 전달되어야 한다. 이를 가능하게 하기 위해, 본 연구에서는 sr과 $d_{(n,i)}\,(\text{for}\,\forall n,i)$를 문장(sentence) 단위로 파싱해 임베딩을 생성한다. 이후, 임베딩 문장 간 유사도를 계산한 후 최대값을 기반으로 문서에 대한 점수를 부여한다.

$RD(ne_i)$에 포함된 문서 $d_{n,i}$와 전장상황 보고 문서 sr의 각 문장은 다음과 같이 표현된다.

$$d_{(n,i)} = \{s_{n,i,1}, s_{n,i,2}, \cdots, s_{n,i,k}, \cdots, s_{n,i,K_{n,i}}\}$$
$$sr = \{ssr_1, ssr_2, \cdots, ssr_l, \cdots, ssr_L\} \quad (5)$$

이때 $s_{n,i,k}$는 $d_{n,i}$의 k번째 문장 ($k = 1,2,..,K_{n,i}$)을 그리고 ssr_l은 sr의 l번째 문장을 의미한다 ($l = 1,2,..,L$).

이후 언어모델을 활용하여 각 문장에 대한 임베딩 벡터를 생성한다. 언어모델로는 가장 널리 활용되는 모델 중 하나인 BERT(Bidirectional Encoder Representations from Transformers) [17]를 활용한다. 생성된 문장 임베딩 벡터들을 활용하여 $d_{n,i}$와 sr의 문장 간 코사인 유사도를 산출하고, 이의 최대값을 $d_{n,i}$의 점수($score(d_{n,i})$)로 부여한다.

$$score(d_{n,i}) = \max_{k,l} sim\{el(s_{n,i,k}), el(ssr_l)\} \quad (6)$$

이때 $el(s_{n,i,k})$과 $el(ssr_l)$은 각각 $s_{n,i,k}$과 ssr_l의 임베딩 벡터를, $sim\{el(s_{n,i,k}), el(ssr_l)\}$은 이들 간의 코사인 유사도를 의미한다. $score(d_{n,i})$ 값이 사전에 정한 임계치 보다 작은 문서는 일차적으로 필터링된다.

$d_{n,i}$에 대한 최종 언어모델 기반 점수 ($LMscore(d_{n,i})$)는 다음과 같다.

$$LMscore(d_{n,i}) = \frac{score(d_{n,i}) - \min_{(i'=1,\ldots,I_n)} score(d_{n,i'})}{\max_{(i'=1,\ldots,I_n)} score(d_{n,i'}) - \min_{(i'=1,\ldots,I_n)} score(d_{n,i'})} \quad (7)$$

3.4.2 협업 필터링 기반 문서 점수 부여

식별한 관련 문서와 지휘관의 임무 및 역할과의 연관성이 높을수록 지휘결심을 수행하는 데에 핵심적인 문서일 가능성이 높다. 본 연구에서는 협업 필터링(Collaborative Filtering, CF) 기법을 적용해 식별한 관련 문서에 대한 CF 예측치를 계산한다. 협업 필터링을 활용하면, 지휘관의 문서 열람 히스토리뿐 아니라 유사한 임무나 역할을 수행한 타지휘관들의 문서 열람 히스토리를 CF 예측값 산출에 활용할 수 있다는 장점이 있다. 협업 필터링을 활용함으로써 지휘관의 임무 및 역할을 간접적으로 문서 점수 산출에 반영할 수 있다. 이때 협업 필터링 기법으로는 행렬 분해(Matrix Factorization, MF)나 GCN(Graph Convolutional Network) 등을 활용할 수 있다[18]. 사용한 CF 기법에 상관없이 타깃 지휘관 u^*이 문서 $d_{n,i}$ 열람할 예측 값($pred(d_{n,i},u^*)$)은 다음 식과 같이 산출된다.

$$pred(d_{n,i},u^*) = ec(d_{n,i}) \bullet ec(u^*) \quad (8)$$

이때 $ec(d_{n,i})$와 $ec(u^*)$는 협업 필터링 결과 얻어지는 $d_{n,i}$와 u^*의 임베딩 벡터를 의미한다.

$pred(d_{n,i},u^*)$을 활용하여 예측 값이 낮은 문서들에 대해 일차 필터링을 수행한다. 최종적인 CF 기반 점수($CFscore(d_{n,i})$)는 협업 필터링 예측 값에 대해 Min-Max 스케일링을 적용하여 산출한다.

3.4.3 최종 관련 문서 선별

최종적으로 개체명 ne_n의 관련 문서는 앞의 두 점수를 모두 고려하여 다음과 같이 산출된다.

$$Tscore(d_{n,i}) = \alpha \bullet LMscore(d_{n,i}) + (1-\alpha) \bullet CFscore(d_{n,i}) \quad (9)$$

이때 $Tscore(d_{n,i})$는 $d_{n,i}$의 최종 점수를 의미하고, α는 하이퍼 파라미터이다($0 \le \alpha \le 1$).

식 (9)의 점수를 활용하여 사전에 정한 임계치 보다 작은 문서를 이차적으로 필터링한다. 그 결과 다음 식과 같이 각 개체명에 대한 최종 관련 문서 집합을 얻을 수 있다.

$$FRD(ne_n) = \{d_{n,1},d_{n,2},\cdots,d_{n,j},\cdots,d_{n,J_n}\} \quad (10)$$

이때 $FRD(ne_n)$는 ne_n의 최종 관련 문서 집합을, J_n은 $FRD(ne_n)$의 총 문서 수를 의미한다. $FRD(ne_n)$의 문서들은 $Tscore$가 높은 순으로 배열된다.

3.5 주요 문장 식별 및 개체별 관련 문서 요약

sr로부터 발견된 각 개체명에 대해 핵심적인 관련 문서들이 탐색되었지만 이러한 문서를 모두 열람하는 것은 정보이용자에게 정보 부담을 야기한다. 이에 본 연구에서는 각 문서에서 개체와 관련된 핵심적인 부분들을 먼저 확인할 수 있도록 문장을 선별하고 이를 연결하여 요약된 형태로 제시하도록 하였다. 이를 위해, $FRD(ne_n)$에 포함된 문서들에 대해 개체명 ne_n를 포함하는 문장들을 선별하고, 선별된 문장들을 우선순위에 따라 배열한다. 우선순위는 다음과 같다. 1) 서로 다른 문서에 포함된 문장들의 경우, $Tscore$가 높은 문서에 포함된 문장들이 보다 앞에 배치된다. 2) 같은 문서 내에 포함된 문장들의 경우, 앞서 산출하였던 sr과의 언어모델 기반 유사도가 높은 문장이 보다 앞에 배치되도록 한다.

4. 실험 및 평가

보안상의이유로 실제 군 데이터에 접근하는 것은 제한되기 때문에, 본 연구에서는 군 NLD와 유사하게 개념이 명시적이면서 공적인 어휘를 사용하는 학술논문들을 활용하여 실험을 수행하였다. 데이터셋은 IT(추천시스템), 국방(국방정보체계), 의료(Covid-19), 경제(GDP), 사회(고독사) 5개의 카테고리에서 각각 20편씩 논문을 수집하여 총 100편의 논문으로 구성하였다. 이에 대해 비교적 길이가 짧고 함축적인 초록(abstract) 부분을 전장상황 보고 문서로, 보다 길고 구체적인 본문 부분을 체계 내 보유 문서로 간주하여 실험을 진행하였다.

첫째로, 전체 논문 초록에 대해 NER 기반 주요 개체 식별을 수행하였다. NER 방법으로는 파이썬 NLTK 라이브러리의 사전 학습된 모델을 활용하였다. NER의 정확도 및 식별된 개체명들은 아래 표 1과 같다.

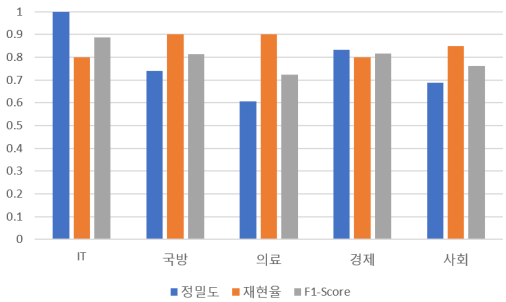
5가지 분야에서 모두 0.7이 넘는 NER 정밀도를 보여주었다. NER 성능을 저해하는 요인으로는 고유명사는 아니지만 사람들이 일상 생활에서 잘 활용하지 않는 비교적 학술적이고 전문적인 명사들(recommender, orientation, overcrowding 등)이 개체명으로 인식되었다는 점이 있었다. 이는 앞서 3장에서 작성된 바와 같이 충분한 양의 문

(표 1) 학술논문 초록에 대한 NER 적용 결과
(Table 1) NER results of the abstracts of the academic papers

분야	정밀도	개체명 예시		
IT	0.8511	GNN	KG	POI
		Yelp	Amazon	GCN
국방	0.7381	DoDAF	MODAF	INCOSE
		UML	DoD	C4ISR
의료	0.7209	COVID	Asian	California
		Vaccine	Hispanic	PDE
경제	0.7879	GDP	IMF	OECD
		European	ICT	Tabscott
사회	0.7407	PLC	Ageing	Loneliness
		Cancer	Japan	Asian

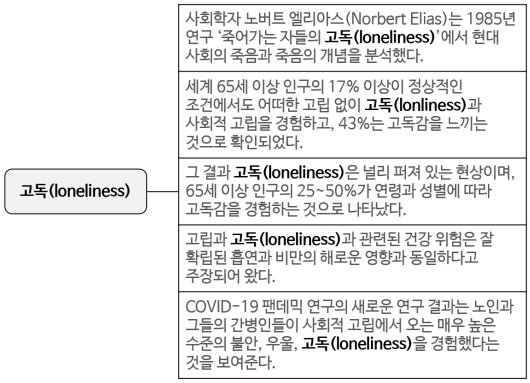
서 베이스를 갖추고 있으면 IDF 기반의 필터링으로 어느 정도 해소될 수 있으리라 기대할 수 있다.

둘째로, 각 개체명에 대해 키워드 검색 및 BERT 모델을 활용하여 문서들을 발견하고, 해당 문서들이 원래의 초록과 관련된 문서인지에 대한 평가를 수행하였다. 관련 문서 발견의 정밀도, 재현율 및 F1-Score는 다음 그림 4와 같다.



(그림 4) 관련 문서 발견 실험 결과
(Figure 4) An experimental result of the discovery of related documents

정밀도, 재현율 및 F1-Score는 데이터셋 수집 과정에서 미리 부여했던 Ground Truth를 기반으로 산출하였다. 실험 결과 대부분의 경우 정밀도, 재현율 및 F1-Score 값이 0.7을 상회했다. 다만 의료 논문의 경우 정밀도 값이 낮게 나왔는데, 의료 논문들의 경우 임상 실험 내용을 포함하는 경우가 많아 국가명이 개체명으로 인식되는 경우가 많았고, 이러한 국가명을 키워드로 가진 의학 외 논문들이 다수 탐색된 것으로 확인되었다. 실제 전장상황 보고 문서에서는 이처럼 다수의 국가명이 포함될 가능성은 많지 않기 때문에 이러한 문제가 발생할 가능성은 낮다. 그러나 본 연구 내용을 실제 군 시스템에 반영할 경우 필요시 적절한 개체명 필터링 방법을 적용해야 할 수 있다.



(그림 5) 개체명에 대해 생성된 요약문 예시
(Figure 5) An example of a summary generated for the named entity

셋째로, 인식된 개체명 중 고독(loneliness)에 대해 요약문 생성 과정을 진행하였고, 그 결과는 위 그림 5와 같다. 이는 고독과 관련된 문서들로부터 고독을 포함하는 문장들을 선별하여 생성된 것이며, 고독에 대한 정보를 압축적으로 보여준다는 것을 확인할 수 있다. 이는 본 연구에서 제시한 방법이 전장상황을 신속하게 인식하고 지휘결심을 수행할 수 있도록 필요한 정보를 압축적으로 제시할 수 있음을 시사한다.

5. 결론 및 시사점

본 연구는 지휘관의 신속한 전장상황 인식 및 지휘결심을 지원하기 위해 체계 내 저장·관리하는 문서 중 현재 지휘관이 수신한 전장상황 보고 문서와 관련된 문서를 자동적으로 탐색하고 이를 요약하기 위한 연구를 수행하였다. 이는 수신한 전장상황 보고 문서에 대해 개체명을 인식하고, 각 개체명과 관련된 체계 내 보유 문서를 탐색하고, 탐색된 문서들에 대해 언어모델 및 협업 필터링을 적용하여 문서를 선별하고, 이들 중 개체명과 관련된 문장을 추출하여 요약된 형태로 제공하는 과정으로 수행된다.

제안된 방법을 적용할 경우 지휘관이 지휘결심에 필요한 문서를 탐색하고 열람하는 시간을 줄임으로써 신속한 지휘결심을 지원할 수 있다. 그러나 제안된 방법에는 다

음과 같은 한계가 존재한다. 첫째로, 체계 내 보유 문서 간의 관계(동일한 문서의 초본과 개정본, 동일한 기능요소에 대한 기능명세서와 시험평가서 등)를 고려하지 않는다. 그렇기 때문에 제공된 문서 및 문장 간의 중복이 존재할 수 있으며, 또한 최신화되지 않은 정보가 전달될 수 있다. 둘째로, 협업 필터링 시 지휘관의 정보 이용 순서를 고려하지 않는다. 제품 구매와는 다르게, 정보 열람의 경우 선후관계가 중요할 가능성이 높다. 이를 고려하여, 추후 연구에서는 보유 문서를 단순히 개별 문서의 형태가 아니라, 그래프 형태로 이들 간의 관계를 발견 및 모델링하기 위한 방법을 개발할 것이다. 또한 협업 필터링 시 세션 기반 추천(session-based recommendation)의 개념을 차용하여 정보이용자들의 정보 열람 순서 역시 관련문서 탐색에 활용될 수 있도록 할 것이다.

References

[1] W. Ou, M. Chae, and D. Yeum, "Influence Factors of Effectively Executing NCW by User's Point of View," Journal of Internet Computing and Services, Vol. 11, No. 2, pp.109-127, 2010.
https://www.jics.or.kr/digital-library/757

[2] C. Han, K. Shin, S. Choi, S. Moon, C. Lee, and J. Lee, "A Methodology of Decision Making Condition-based Data Modeling for Constructing AI Staff,"Journal of Internet Computing and Services, Vol. 21, No. 1, pp. 237 - 246, 2020.
https://doi.org/10.7472/jksii.2020.21.1.237

[3] C. Lee, J. Baek, J. Son, and Y. Ha, "Deep AI military staff: Cooperative battlefield situation awareness for commander's decision making," The Journal of Supercomputing, Vol. 79, No. 6, pp.6040-6069, 2023.
https://doi.org/10.1007/s11227-022-04882-w

[4] P. Gonsalves, G. Rinkus, S. Das, and N. Ton, "A hybrid artificial intelligence architecture for battlefield information fusion," Proceedings of the Second International Conference on Information Fusion, pp. 463-468, 1999.
https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=61b08bf07d0e06910d0a16f3f9918554ac0dcfe5

[5] FM 101-5-2 U.S. Army Report and Message Formats, https://www.bits.de/NRANEU/others/amd-us-archive/fm101-5-2%28uk%29.pdf

[6] J. fuller, "Information Overload and the Operational Commander,"Naval War College, Joint Military Operations Department, 2000.
https://apps.dtic.mil/sti/citations/ADA378709

[7] L. Chen, K. Chang, and S. Yang, "Integrating corpus-based and NLP approach to extract terminology and domain-oriented information: An example of US military corpus," Acta Scientiarum. Technology, Vol. 44, e60486-e60486, 2022.
https://doi.org/10.4025/actascitechnol.v44i1.60486

[8] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Transactions on Knowledge and Data Engineering, Vol. 34, No. 1, pp. 50-70, 2020.
https://doi.org/10.1109/TKDE.2020.2981314

[9] S. Edunov, A. Baevski, and M. Auli, "Pre-trained language model representations for language generation," arXiv preprint arXiv:1903.09722, 2019.
https://doi.org/10.48550/arXiv.1903.09722

[10] Y. Koren, S. Rendle, and R. Bell, "Advances in collaborative filtering," Recommender systems handbook, pp. 91-142, 2021.
https://doi.org/10.1007/978-1-0716-2197-4_3

[11] H. Li, X. Yang, X. Zhao, L. Yu, J. Zheng, and W. Sun, "MLRIP: Pre-training a military language representation model with informative factual knowledge and professional knowledge base," arXiv preprint arXiv:2207.13929, 2022.
https://doi.org/10.48550/arXiv.2207.13929

[12] C. Gunasekara, T. Carryer, and M. Triff, "On Natural Language Processing Applications for Military Dialect Classification," In 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 211-218, 2021.
https://doi.org/10.1109/ICMLA52953.2021.00040

[13] S. Biswas, "Prospective Role of Chat GPT in the Military: According to ChatGPT," Qeios, 2023.
https://doi.org/10.32388/8WYYOD

[14] W. Ning, J. Liu, and H. Xiong, "Knowledge discovery using an enhanced latent Dirichlet allocation-based clustering method for solving on-site assembly problems," Robotics and Computer-Integrated Manufacturing, Vol. 73, 102246, 2022.

https://doi.org/10.1016/j.rcim.2021.102246

[15] Y. Meng, Y. Zhang, J. Huang, Y. Zhang, J. Han, "Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations," In Proceedings of the ACM Web Conference 2022, pp. 3143-3152, 2022. https://doi.org/10.1145/3485447.3512034

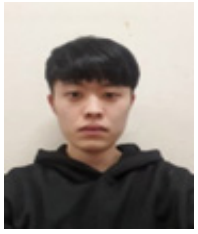
[16] M. Memon, Y. Lu, P. Chen, A. Memon, M. Pathan, and Z. Zardari, "An ensemble clustering approach for topic discovery using implicit text segmentation," Journal of Information Science, Vol. 47, No. 4, pp. 431-457, 2021. https://doi.org/10.1177/0165551520911590

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805

[18] X. Wang, X. He, M. Wang, F. Feng, and T. Chua, "Neural graph collaborative filtering," In Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval, pp. 165-174, 2019. https://doi.org/10.1145/3331184.3331267

◐ 저 자 소 개 ◑

김 건 영(Kunyoung Kim)
2017년 성균관대학교 전자전기공학부(공학사)
2017년～현재 성균관대학교 대학원 산업공학과 석박통합과정
관심분야 : 인공지능, 지식관리, 온톨로지, 추천시스템
E-mail : kimkun0@skku.edu



이 정 빈(Jeongbin Lee)
2021년 성균관대학교 시스템경영공학과(공학사)
2022년～현재 성균관대학교 산업공학과 석사과정
관심분야: 추천시스템, 지식그래프
E-mail: jim2091@skku.edu

손 미 애(Mye Sohn)
1985년 성균관대학교 산업공학과(공학사)
1988년 한국과학기술원 대학원 산업공학과(공학석사)
2002년 한국과학기술원 대학원 경영공학과(공학박사)
2004년～현재 성균관대학교 산업공학과 교수
관심분야 : 인공지능/전문가시스템, 지식그래프, 시맨틱웹, 온톨로지, IoT, 기계학습, 추천시스템
E-mail : myesohn@skku.edu