J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr
Copyright ⓒ 2021 KSII

자기 지도 학습 기반의 언어 모델을 활용한 다출처 정보 통합 프레임워크☆

Multi-source information integration framework using self-supervised learning-based language model

김 한 민[1] 이 정 빈[1] 박 규 동[2] 손 미 애[1*]
Hanmin Kim Jeongbin Lee Gyudong Park Mye Sohn

요 약

인공지능(Artificial Intelligence) 기술을 활용하여 인공지능 기반의 전쟁 (AI-enabled warfare)가 미래전의 핵심이 될 것으로 예상한다. 자연어 처리 기술은 이러한 AI 기술의 핵심 기술로 지휘관 및 참모들이 자연어로 작성된 보고서, 정보 및 첩보를 일일이 열어 확인하는 부담을 줄이는데 획기적으로 기여할 수 있다. 본 논문에서는 지휘관 및 참모의 정보 처리 부담을 줄이고 신속한 지휘결심을 지원하기 위해 언어 모델 기반의 다출처 정보 통합 (Language model-based Multi-source Information Integration, LAMII) 프레임워크를 제안한다. 제안된 LAMII 프레임워크는 자기지도 학습법을 활용한 언어 모델에 기반한 표현학습과 오토인코더를 활용한 문서 통합의 핵심 단계로 구성되어 있다. 첫 번째 단계에서는, 자기지도 학습 기법을 활용하여 구조적으로 이질적인 두 문장간의 유사 관계를 식별할 수 있는 표현학습을 수행한다. 두 번째 단계에서는, 앞서 학습된 모델을 활용하여 다출처로부터 비슷한 내용 혹은 토픽을 함양하는 문서들을 발견하고 이들을 통합한다. 이 때, 중복되는 문장을 제거하기 위해 오토인코더를 활용하여 문장의 중복성을 측정한다. 본 논문의 우수성을 입증하기 위해, 우리는 언어모델들과 이의 성능을 평가할 때 활용되는 대표적인 벤치마크 셋들을 함께 활용하여 이질적인 문장간의 유사 관계를 예측의 비교 실험하였다. 실험 결과, 제안된 LAMII 프레임워크가 다른 언어 모델에 비하여 이질적인 문장 구조간의 유사 관계를 효과적으로 예측할 수 있음을 입증하였다.

☞ 주제어 : 자기지도 학습, 언어 모델, 문장간의 유사 관계, 다출처 정보 통합

ABSTRACT

Based on Artificial Intelligence technology, AI-enabled warfare is expected to become the main issue in the future warfare. Natural language processing technology is a core technology of AI technology, and it can significantly contribute to reducing the information burden of underrstanidng reports, information objects and intelligences written in natural language by commanders and staff. In this paper, we propose a Language model-based Multi-source Information Integration (LAMII) framework to reduce the information overload of commanders and support rapid decision-making. The proposed LAMII framework consists of the key steps of representation learning based on language models in self-supervsied way and document integration using autoencoders. In the first step, representation learning that can identify the similar relationship between two heterogeneous sentences is performed using the self-supervised learning technique. In the second step, using the learned model, documents that implies similar contents or topics from multiple sources are found and integrated. At this time, the autoencoder is used to measure the information redundancy of the sentences in order to remove the duplicate sentences. In order to prove the superiority of this paper, we conducted comparison experiments using the language models and the benchmark sets used to evaluate their performance. As a result of the experiment, it was demonstrated that the proposed LAMII framework can effectively predict the similar relationship between heterogeneous sentence compared to other language models.

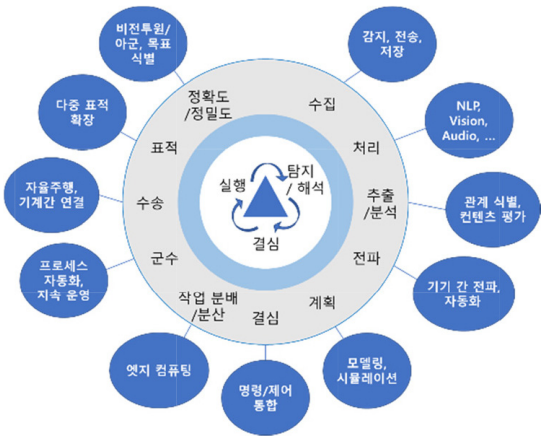
☞ keyword : Self-supervised learning, Language model, similar relationship between sentences, Multi-source information integration

1 Dept. of Industrial Engineering, Sungkyunkwan University, Gyeonggi (Suwon), 16419, Korea
2 Agency for Defense Development, Seoul, Korea
* Corresponding author: myesohn@skku.edu
[Received 12 October 2021, Reviewed 20 October 2021, Accepted 5 November 2021]

☆ 본 연구는 국방과학연구소의 국방 지휘통제 통합·연동 기반 기술 특화연구실 과제의 지원을 받아 수행되었습니다 (UE201114ED).

1. 서 론

미국은 "모자이크 전쟁 (mosaic warfare)"을 수행하기 위해 군 사령부들이 운용하는 정보수집 센서와 전술통제망을 단일화하기 위한 지휘통제 연결망 구축사업인 합동전영역지휘통제(Joint All-Domain Command and Control, JADC2) 체계를 구축하고 있고 중국은 "지능화 전쟁(intelligentized war)"이라 명명한 새로운 전쟁에서 승리하기 위한 투자를 아끼지 않고 있다 [1]. 여기서 주목해야 할 사실은 이들 패러다임을 관통하는 핵심 기술이 인공지능(Artificial Intelligence, AI)라는 것이다. 우리 군도 2017년 국방개혁 2.0을 통해 AI의 도입을 명문화했고, 정부는 2019년 AI 국가전략 발표를 통해 지능형 플랫폼과 지휘체계 지원기능 개발을 담은 '국방 AI'를 공개한 바 있으며 국방부의 AI 장기 계획을 담은 '국방 인공지능(AI) 발전계획 수립연구'를 수행하는 등 AI 도입에 박차를 가하고 있다 [2, 3]. 이러한 인공지능이 기반이 되는 전쟁 (AI-enabled warfare)은 AI 기술을 특정 무기체계, 기술 또는 작전에 국한해 적용하는 것이 아니라 전투의 모든 측면에 AI 기술이 적용·통합되는 것을 의미한다[4]. 예를 들어, 그림 1은 지휘통제의 과정에서 활용가능한 AI 기술을 요약한 것이다.



(그림 1) 지휘통제에서의 활용가능한 인공지능
(Figure 1) AI technologies for command and control

그림 1에 도식화한 바와 같이 지휘통제 분야에서 적용가능한 AI 기술은 매우 다양하다. 이러한 기술 중 기기간 통신을 통한 자동화 기술은 지휘통제 절차의 자동화를 통해 신속·정확한 작전의 수행에 기여할 수 있으며 [5], 자연어 처리기술은 지휘관 및 참모들이 정보 처리 부담, 즉 자연어 작성되거나 구술된 정보나 첩보의 내용을 일일이 열어 확인하는 부담을 줄이는 데 기여할 수 있다 [6].

이에 본 논문에서는 지휘관 및 참모의 정보 처리 부담을 줄이고 신속한 지휘결심을 지원하는 데 필요한 언어 모델 기반의 다출처 정보 통합 (LAnguage model-based Multi-source Information Integration, LAMII) 프레임워크를 제안한다. LAMII 프레임워크를 이용한 통합의 대상은 다출처에서 수집된 형식과 내용이 다양하고 이질적인 자연어 문서이다. 그러나 수집·전달·저장하고 있는 문서의 양이 방대하기 때문에 통합의 대상이 되는 자연어 문서를 선별하는 과정이 필요하다. 이를 위해, 본 논문에서는 자연어 문서들이 의미적으로 유사한 문장을 내포하고 있는지 여부를 기준으로 1차 선별을 수행한 후, 지휘결심을 수행해야 하는 전장 상황을 고려해 2차 선별을 수행한다. 최종적으로 선별된 자연어 문서들을 하나의 통합 문서로 생성해 지휘관과 참모에게 전달함으로써 정보처리의 부담을 줄이는 데 기여하고자 한다.

1차 선별, 즉 자연어 문장의 의미적 유사도를 도출하기 위해 가장 널리 사용되는 방법은 표현학습(representation learning)이다 [7]. 표현학습은 자연어에 내재된 특성을 학습을 통해 추출하는 방법으로 주로 발견된 특성을 이용해 일반적인 연산이 불가능한 자연어를 연산이 가능한 공간으로 임베딩하여 나타낸다. 그러나 자연어 문서의 범위는 교범과 같이 구조나 형식이 정형화된 문서에서 음성이나 속기로 작성된 회의록과 같이 불완전한 구조를 갖는 비정형화된 문서까지 매우 다양하다. 이는 기존의 풍부한 정보를 지닌 정형화된 문서에 대해 이루어지는 표현학습으로는 정보량이 부족하고 비구조화된 단문이나 음성에 대한 의미적 유사도를 계산하기 어려움을 의미한다.

1차 선별을 통해 발견한 의미적으로 유사한 문장을 내포하고 있는 문서들에 대해 2차 선별을 수행한다. 유사한 문장을 내포하고 있는 문서라 할지라도 해당 문서 모두가 지휘 결심을 수행하는 전장 상황에 적합한 정보를 포함하는 것은 아니다. 따라서 발견된 문서들에 대해 현재의 전장 상황을 고려하여 동일한 맥락을 가진 정보를 식별하는 과정을 수행한다. 마지막으로 문서로부터 핵심 정보를 담은 요약문을 생성한다. 요약문을 생성하기 위해 주로 사용되는 방법은 사전에 사람이 만들어낸 요약문을 활용하는 지도학습이다. 그러나 새로운 무기체계의 등장이나 주변 환경의 변화로 전장 상황은 매번 상이할 수밖

에 없다. 이는 지휘결심에 필요한 요약 정보를 생성하는 데 지도학습이 부적합함을 의미한다. 이러한 한계를 극복하기 위해, LAMII 프레임워크는 오토인코더(Autoencoder) 기반의 비지도학습 기법을 활용하여 문서로부터 중복되는 문장을 제거해 핵심 정보만을 포함하는 요약문 생성 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 연구들을 살펴보고, 3장에서 본 논문에서 제안한 방법의 프레임워크와 각 모듈에 대해 설명한다. 4장에서 실험과 그 결과에 대해 설명하고 5장에서 결론 및 추후 연구에 대해 논의한다.

2. 관련 연구

2.1 언어모델

언어모델은 전통적인 통계 기반 모델에서부터 인공신경망 기반 모델, 최근에는 Elmo, GPT, BERT 등의 Transformer 기반의 지도 학습법에서 시작하여 최근에는 비지도 학습의 언어모델로 발전해왔다[8.9.10]. 비지도 학습 또는 자기지도 학습의 언어모델은 한 언어의 방대한 데이터를 사용해 사전에 부분적으로 학습한 후 기계번역, 감정분석, 문서요약, 질문응답 등의 다양한 Task에 따라 점차적으로 모델이 이해할 수 있는 데이터를 확장하는 방식이다. 또한, 특정 Task에 따라 몇 가지 일반화된 지식을 활용하여 자기지도 학습법은 자동화된 fine-tuning이 가능하다. 이처럼 목적에 맞게 fine-tuning된 언어모델들은 양질의 레이블 정보가 없어도 사람의 performance를 능가할 정도로 성능상의 이점이 많기 때문에 최근 다양한 분야에서 크게 관심받고 있다 [11].

2.2 문서요약

문서 요약은 기존 문서의 의미 혹은 정보를 최대한 유지한 채 규모가 더 작은 문서를 생성하는 것을 목표로 한다. 문서요약 기법은 크게 추출요약(extractive summarization)과 생성요약(abstractive summarization)으로 구분할 수 있다.

추출요약은 기존 문서에서 중요 문장을 그대로 가져와서 요약문을 생성하는 기법이다. 이 때, 중요 문장을 선택하기 위한 기법에 따라 빈도수 기반 기법, 유사도 기반 기법, 그래프 기반 기법 등의 여러 방법이 존재하며, 특히 최근에는 문서요약을 일종의 문장분류 task로 보는 딥러닝 기법들이 연구되고 있다[12].

생성요약은 기존 문서를 통해 언어 모델을 활용해 새로운 문장을 생성하여 요약문을 전체를 모델에 기반하여 생성하는 기법이다. 기존 문서의 몇몇 문장을 그대로 가져올 경우 문서의 주요한 의미는 유지될 수 있지만 맥락이 맞지 않는 문서가 생성될 수 있는 문제를 해결하기 위해 고안되었다. 하지만 추출요약에 비해 낮은 성능을 보여 최근에는 추출요약과 생성요약을 결합한 기법 등 새로운 연구들이 등장하고 있다[13].

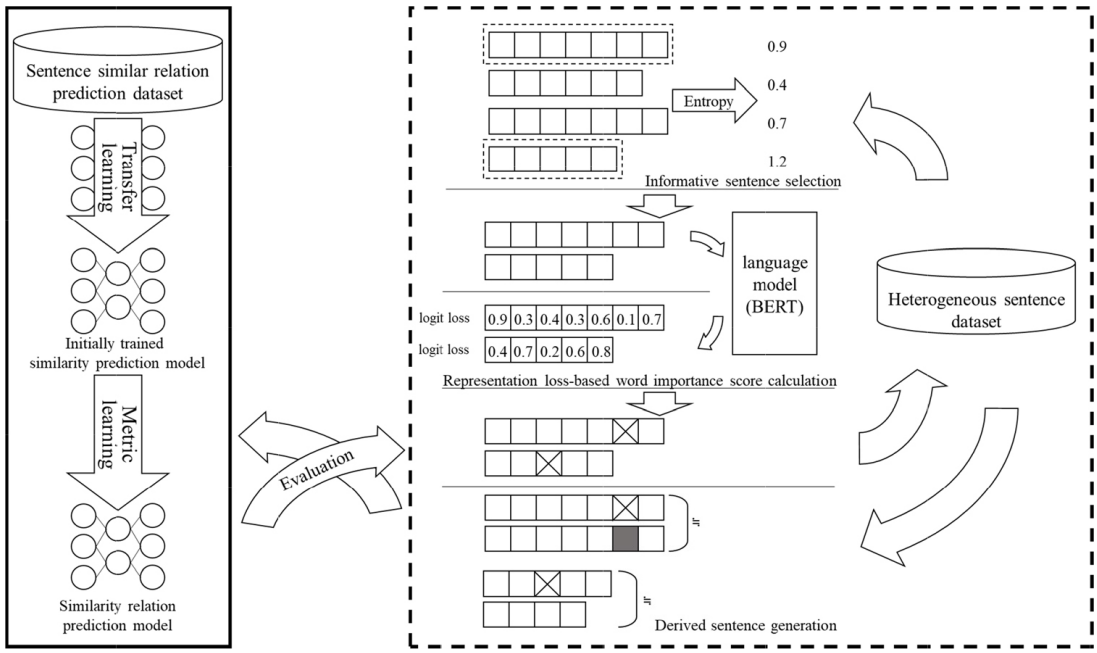
3. 언어 모델 기반 다출처 정보 통합 프레임워크

지휘결심에 필요한 다출처 문서를 통합하는 LAMII 프레임워크는 3개의 모듈로 구성되어 있다. 3개의 모듈은 이질적인 문장들 간의 의미적 유사 관계를 예측하고 이를 기반으로 의미적으로 유사한 문서를 식별하는 데 필요한 언어모델 학습 모듈, 식별된 유사 문서들 중에서 지휘 결심을 수행해야 하는 전장 상황과 연관된 문서를 발견한 모듈 그리고 발견한 문서를 통합하는 모듈이다.

3.1 이질적인 문장의 유사 관계 학습 모듈

이 모듈의 목적은 이질적인 문장들 간의 의미적 유사 관계를 예측하고 이를 기반으로 의미적으로 유사한 문서를 식별하는 데 필요한 언어모델을 학습시키는 것이다. 언어 모델의 학습을 위해서는 학습 데이터가 필요하다. 본 논문에서 사용하는 학습 데이터는 문장에서 영향력이 낮은 어휘를 대체 혹은 제거하는 것으로 문장의 어휘적 이질성과 형태의 이질성을 고려하여 생성한다. 또한 직접 생성한 레이블 된 데이터를 사용하기 이전에 유사 관계를 학습하기 위한 소량의 데이터셋을 사용해 전이학습을 수행하여 모델의 재현율과 학습 성능을 향상시킨다. 이 때, 유사 관계를 예측한 소량의 데이터 셋(이후 유사 관계 데이터 셋이라 함)과 사전 학습된 언어모델은 가지고 있다고 가정한다.

첫 모듈에서 수행되는 다음의 단계들을 도식화하면 그림 2와 같다.



(그림 2) 문장의 유사 관계 학습 절차
(Figure 2) Learning procedure of sentence similarity relation

3.1.1 언어 모델의 표현 손실을 이용한 파생 문장 생성

본 논문에서는 레이블이 존재하지 않는 이질적인 문장들 간의 유사 관계를 학습하기 위해 자기지도학습(self-supervised learning)을 수행한다. 자기지도학습이란 레이블이 없는 원 데이터(문장)로부터 문장들의 관계를 통해 레이블을 자동으로 생성하는 방법으로[14], 지도학습 방법을 레이블이 없는 데이터(문장)에 적용해 높은 예측 성능을 도출할 수 있다는 장점이 있다. 자기지도학습에서 레이블을 자동 생성하기 위해 필요한 문장 셋은 원 문장 셋과 원 문장과의 의미적 유사도 비교를 수행하기 위한 파생 문장 셋이다. 원 문장(original sentence, OS)과 원 문장으로부터 선별된 선별 문장(selected sentence, SS) 그로부터 파생된 파생 문장 (derived sentence, DS)은 다음과 같이 정의된다.

정의 1 원 문장 셋(original sentence set, OS) OS는 일련의 어휘로 구성된 문서를 구성하는 문장들의 집합으로 다음과 같이 표현된다.

$$OS = \{os_{1,} os_{2,} \cdots, os_{i,} \cdots\}, i \in \mathbb{N}$$

OS를 구성하는 i번째 원 문장 os_i는 일련의 어휘로 구성된 벡터로서 다음과 같이 표현된다.

$$os_i = (ow_{i1,} ow_{i2,} \cdots, ow_{ih,} \cdots), h \in \mathbb{N}$$

이때 ow_{ih}는 os_i의 h번째 어휘이다.

정의 2 선별 문장 셋(selected sentence set, SS) SS는 OS로부터 정보량을 기준으로 선별된 문장들의 집합으로 다음과 같이 표현된다.

$$SS = \{ss_{1,} ss_{2,} \cdots, ss_{j,} \cdots\}, j \in \mathbb{N}...$$

*SS*를 구성하는 j번째 선별 문장 ss_j는 일련의 어휘로 구성된 벡터로서 다음과 같이 표현된다.

$$ss_j = (w_{j1,} w_{j2,} \cdots, w_{jk,} \cdots), k \in \mathbb{N}$$

이때 w_{jk}는 ss_j의 k번째 어휘이다.

정의 3 파생 문장 셋 (derived sentence set, DS) DS는

SS를 구성하는 문장으로부터 임의의 어휘를 제거하여 생성된 문장들의 집합으로 다음과 같이 표현된다.

$$DS = \{ds_1, ds_2, \cdots, ds_j, \cdots\}, j \in \mathbb{N}$$

DS를 구성하는 j번째 파생 문장 ds_j는 ss_j를 구성하는 일련의 어휘로부터 임의의 어휘를 대체한 벡터이며 다음과 같이 표현된다.

$$ds_j = (w_{j1}, \cdots, w_{(jk-1)}, w_{(jk+1)}, \cdots)$$

DS는 다음과 같은 절차를 통해 생성된다.

Step 1 문장 엔트로피를 이용한 SS 식별

전술한 바와 같이, SS는 OS 중에서 정보량이 낮아 학습 데이터로는 부적합하다고 판명된 문장이 삭제된 문장의 집합이다. 이때 정보량이 낮은 문장이란 불완전한 문장이나 등장 빈도가 높은 어휘들로 구성된 문장으로서 학습 데이터로는 부적합한 문장을 의미한다. 특히 불완전한 문장은 문장을 구성하는 일부 어휘가 달라지면 문장의 의미가 쉽게 변질될 수 있기 때문에 파생 문장과의 유사도 비교 성능에 부정적인 영향을 미칠 수 있다.

문장의 정보량은 문장을 구성하는 어휘들의 출현 확률 기반의 Shannon 엔트로피를 이용해 계산한다. 임의의 i번째 원 문장 os_i로부터 도출된 문장 엔트로피 E_i의 계산 방법은 다음과 같다.

$$E_i = -\sum_h p_{ow_{ih}} \log_2 p_{ow_{ih}}$$

이때, $p_{ow_{ih}}$는 os_i의 h번째 어휘 ow_{ih}의 등장 확률이다. 모든 원 문장에 대해 문장 엔트로피를 계산하며 문장 엔트로피가 주어진 임계치 η 이상인 문장들이 SS의 원소가 된다.

Step 2 파생 문장 생성을 위한 어휘 별 영향력 계산

파생문장인 DS는 문장을 구성하는 어휘들 중에서 문장에 대해 영향력이 낮은 어휘를 삭제해 생성한다. 문장에서 개별 어휘의 영향력은 사전 학습된 언어모델을 이용해 해당 어휘를 삭제한 문장과 해당 어휘를 포함한 문장의 임베딩 결과 값의 차이로 계산한다.

$$I_{(w_{jk})} = \| o(ss_j) - o(ss_{j/w_{jk}}) \|_2, \text{for } all\ j \qquad (2)$$

$$ss_{j/w_{jk}} = (w_{j1}, \cdots, w_{jk-1}, [\text{MASK}], w_{jk+1}, \cdots)$$

이때 $o(ss_j)$는 ss_j의 임베딩 결과 값이고 $o(ss_{j/w_{jk}})$는 ss_j의 k^{th} 어휘를 [MASK]로 대체한 문장의 임베딩 결과 값이다. 이 과정은 문장을 구성하는 모든 어휘에 대해 반복적으로 수행한다.

ss_j에서 w_{jk}의 영향력과 $I_{w_{jk}}$값은 정비례의 관계가 있다. 즉, 특정 어휘의 영향력이 클수록 $I_{w_{jk}}$값은 크며 $I_{w_{jk}}$가 작을수록 해당 어휘는 ss_j의 의미를 표현하는데 영향을 주지 못한다. ss_j의 모든 어휘들의 영향력 점수를 비교하여 문장의 의미에 미치는 영향이 가장 낮은 어휘를 제거함으로써 ss_j의 의미를 가장 잘 보존할 수 있는 파생 문장 ds_j를 생성한다. 이때 어휘의 제거를 통해 생성된 불완전한 파생 문장은 학습 과정에서 이질적인 구조를 지닌 문장과의 유사 관계를 학습하는 데 기여한다. 이 과정을 SS의 모든 원소에 대해 반복하여 DS를 생성한다.

이상의 과정은 그림 2의 점선 박스내에 도식화되어 있다.

3.1.2 메트릭 러닝을 사용한 문장의 유사 관계 학습

DS를 활용하여 문장의 유사 관계 예측 모델을 학습하기에 앞서 기존에 유사 관계가 판명된 유사 관계 데이터셋을 활용해 전이 학습(transfer learning)을 수행한다. 이는 학습이 되지 않은 초기 단계에서 모델에 의해 생성된 데이터만을 이용해 학습할 경우 모델의 재현율과 초기 성능을 보장할 수 없기 때문이다. 전이 학습을 통해 초기 학습을 완료하면 SS와 DS로부터 생성한 유사 문장 데이터를 이용하여 문장의 유사 관계를 모델에 학습시킨다.

문장의 유사 관계 학습은 문장 간의 관계를 거리의 형태로 나타내 학습하는 메트릭 러닝을 사용해 수행한다. 메트릭 러닝은 벡터 공간으로 사영된 문장들의 거리를 활용하는 학습 기법으로 각각의 문장 하나의 특징이 아닌 문장들의 관계를 학습하는데 효과적인 학습 방법이다 [15]. 일반적으로 메트릭 러닝을 적용하기 위해 삼중항 손실(triplet loss)가 사용된다. 이는 유사한 두 데이터의 거리를 줄이고, 관련 없는 데이터와의 거리를 증가시키도록 설계되며 본 연구에서는 모델의 손실 값을 거리 대신 사용하여 유사한 두 문장 사이에서 발생하는 손실은 줄이고 관련 없는 두 문장 사이에서 발생하는 손실을 증가시키는 방향으로 모델을 학습시킨다. 이를 위해 다음과 같은 목표 함수 J를 최소화한다.

$$J = \sum_j o_l(ss_{j,ds_j}) - o_l(ss_j - ss_p),\ p \neq j \qquad (3)$$

이때, $o_l(ss_j, ds_j)$는 초기 학습을 완료한 판별 모델에 ss_j와 ds_j를 입력했을 때 얻어지는 로짓 결과값을 의미하며 ss_p는 SS에서 임의로 선택된 ss_j와 관련 없는 문장이다. 모델 평가를 위해 SS와 DS로부터 일부 데이터를 임의로 추출해 사용하며 그 평가 결과에 따라 모델의 추가 학습 여부를 결정한다. 추가 학습이 결정되면 이전 단계로 돌아가 SS로부터 DS를 새롭게 생성한다. 모든 학습이 완료되면 문장의 유사 관계 예측 모델을 그 결과물로 저장한다.

이상의 과정은 그림 2의 실선 박스내에 도식화되어 있다.

3.2 전장 상황을 고려한 관련 문서 식별 모듈

전장에서 발생한 이벤트는 상이한 시공간과 오브젝트(무기체계, 부대, 병사 등)로 구성되기 때문에 이벤트별 대응 방책 또한 상이한 정보를 이용해 상이하게 수립되어야 한다. 이때, 지휘관과 참모의 신속정확한 지휘 결심을 지원하기 위해서는 이벤트 방책 수립에 필요한 핵심 정보만을 선별해 제공하는 것이 필요하다. 이 모듈의 목적은 특정 이벤트와 관련된 방책 수립에 필요한 정보를 식별하는 것이다. 이를 위해, 지휘관과 참모는 이벤트 발생에 대한 보고는 문서나 유무선 통신을 통한 녹취 정보나 첩보를 수신한 것으로 가정한다.

이 모듈이 구동되는 절차는 다음과 같다.

Step 1. 이벤트 관련 보고서를 문장 단위로 파싱

자연어 문서인 이벤트 관련 보고서를 문장 단위로 파싱한 후, 문장 별 키워드를 추출해 저장한다. 이때 이벤트 관련 q 번째 보고서(Ev_q)는 다음과 같이 표현된다.

$$Ev_q = \{ts_{q1}, ts_{q2}, \cdots, ts_{qr}, \cdots\}$$

ts_{qr}은 Ev_q를 구성하는 r^{th}문장을 의미하며 다음과 같이 키워드를 원소로 갖는 집합이다.

$$ts_{qr} = \{kw_{(qr_1)}, kw_{(qr_2)}, \cdots\}$$

이 과정을 통해 Ev_q를 구성하는 문장별 키워드 집합을 생성한다.

Step 2. 학습된 언어모델을 이용한 Ev_q의 관련 문서 발견

Ev_q와 관련 문서를 발견하기 위해, 첫 번째 모듈에서 학습된 언어모델을 활용하여 Ev_q의 임의의 문장과 유사한 문장을 갖는 관련 문서를 발견한다. 이를 모든 Ev_q를 구성하는 문장(ts_{qr}, for *all* r)에 대하여 반복적으로 수행한다. 다시 말해, 학습된 언어모델을 활용한 문장 대 문장의 유사 관계 예측을 통해 비슷한 내용의 문장을 포함하는 문서를 Ev_q의 관련 문서로 식별한다.

Step 3 LDA기반 이벤트 관련 토픽 발견

전장 상황에서 발생한 이벤트에 대해 적합한 방책 수립을 지원하기 위해 앞서 찾아진 관련 문서로부터 Ev_q와 보다 밀접하게 연관된 문서를 식별해야 한다. 식별된 관련 문서가 문장의 유사 관계에 기반하였기 때문에 전체 문서의 내용이 Ev_q와 밀접한 관련이 있다고 말하기 어렵기 때문이다. 이를 위해, 본 논문에서는 토픽 모델링 기법을 활용하여 Ev_q와 관련 문서를 구성하는 토픽간 유사 관계를 분석하여 보다 밀접하게 연관된 문서를 식별하기 위해 Latent Dirichlet Allocation(LDA) 기반 토픽모델링을 수행한다. 토픽 모델링은 문서에 내재되어 있는 유의미한 단어의 집합을 토픽 형식으로 찾아 이를 활용해 유사 토픽을 공유하는 문서를 발견한다.

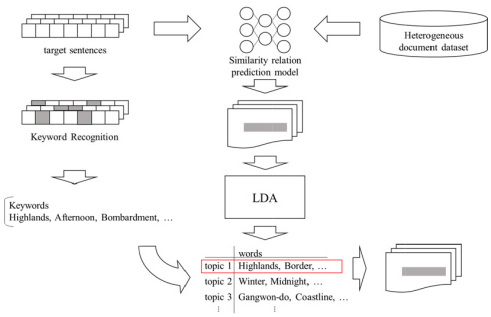
Step 4 키워드 비교를 통한 밀접한 관련 문서 식별

앞서 생성한 Ev_q의 키워드 집합과 관련 문서의 토픽-단어 행렬을 모두 이용해 Ev_q와 가장 밀접한 관련 문서를 식별한다. 문서에서 가장 중요한 토픽을 결정하기 위해 Ev_q의 키워드들이 각 토픽에서 차지하는 비중을 계산하고 그 비중이 가장 높은 토픽을 Ev_q의 대표 토픽으로 결정한다. 결정된 Ev_q의 대표 토픽에 대해 LDA에 의해 계산된 각 관련 문서의 토픽 분포를 확인하여 Ev_q의 토픽과 같은 토픽의 문서를 밀접한 관련 문서로 식별한다.

이 과정을 도식화하면 그림 3과 같다.

3.3 의미가 중복된 문장의 제거를 통한 통합 문서 생성 모듈

이전 모듈을 통해 식별된 Ev_q와 밀접하게 관련된 문서 집합으로부터 정보를 요약해 하나의 통합된 문서를 생성하는 모듈이다. 이는 지휘결심에 필요한 문서를 중복



(그림 3) 전장 상황을 고려한 관련 문서 식별 절차
(Figure 3) Identification procedure of relevant document considering the battlefield context

정보가 포함된 여러 개가 아니라 핵심 정보가 요약된 하나의 문서로 제공하기 위해 수행한다. 문서들을 요약하고 정보를 통합하기 위해 본 논문에서는 오토인코더(autoencoder) 모델의 손실 값을 활용하여 문장의 중복성을 식별한다. 중복 문장으로 판단된 문장을 삭제하여 관련 문서로부터 새로운 정보를 갖는 문장만을 남긴다. 그 다음, 맥락적으로 자연스러운 통합 문서를 생성하기 위해 문서에서의 문장의 위치를 고려하여 남겨진 문장을 통합한다. 모듈의 수행 절차는 다음과 같다.

Step 1 각 문장의 이벤트 관련 문장과의 상대 위치 식별

먼저 이전 모듈에서 식별한 관련 문서에 대해 문서 내에 포함된 Ev_q와 공유하는 문장을 기준으로 Ev_q의 앞에 등장하는 문장들과 뒤에 등장하는 문장들을 구분한다. 이는 각 문장들의 상대적 위치 값을 구하기 위함으로 식별된 모든 관련 문서에 대해 동일한 작업을 수행해 관련 문서 집합을 선행 문장 집합 FS_q와 후속 문장 집합 BS_q로 나누어 저장한다.

Step 2 중복 문장 제거를 위한 오토인코더 적용

관련 문서의 집합으로부터 요약된 정보를 추출하기 위해 같은 의미의 중복되어 나타나는 문장들을 제거하는 것이 필요하다. 이를 위해 두 문장의 중복 여부를 손실 값을 이용해 판단하는 오토인코더 모델의 학습이 필요하다. 먼저 관련 문서를 구성하는 모든 문장에 대해 입력 받은 문장을 다시 동일한 문장으로 출력하도록 오토인코더를 학습시킨다. 오토인코더의 학습이 완료되면 중복될 것으로 예상되는 서로 다른 문장을 입력과 출력으로 설정하고 그 때 발생하는 오토인코더 모델의 손실이 설정된 임계치보다 낮을 경우 중복문장으로 결정해 문장을 삭제한다. 이 과정에서 FS_q와 BS_q에서 중복된 문장이 나온 횟수를 측정해 문장의 통합 문서에서의 위치를 설정한다. 이를 모든 문장에 대해 반복하고 남은 문장을 결정된 위치에 따라 배열해 최종적으로 하나의 통합 문서를 생성한다.

4. 실험 및 평가

본 논문에서 제안한 프레임워크의 우수성을 실험 및 평가하기 위해, 우리는 실험을 위한 이질적인 문서로 구성된 다출처 문서 데이터 셋을 준비하였다. 데이터 셋은 사전에 학습된 언어 모델 및 평가 데이터를 활용하기 위해 영어 문장으로 구성하였으며 충분한 정보를 담고 있는 잘 구성된 문서인 학술 논문 데이터와 불완전한 문장이 포함되어 다소 부족한 정보를 지닌 소셜 미디어 포스트 셋으로 구성하였다. 두 데이터는 이질적인 어휘와 문장 구성을 지니며 학술 논문은 arxiv에 게재된 논문들을 활용하였으며, 소셜 미디어 포스트 셋은 트위터로부터 직접 수집하였다. 각 실험 데이터의 특징은 다음 표 1과 같다.

(표 1) 실험 데이터 셋에 대한 요약
(Table 1) The summary of experimental data sets

Properties	Arxiv Dataset	Tweeter Post
#Documents	10,000	100,000
#Sentences	110,484	225,503
#Vocabs	52,769	174,866

다출처의 이질적인 문서 데이터로부터 학습한 결과를 비교하기 위해 다른 언어 모델과 그 성능을 비교하였다. 본 논문에서는 제안하는 프레임워크에서의 언어 모델을 학습시키기 위해 문서 데이터로부터 각 문장의 엔트로피를 계산해 상위 10%의 문장을 사용하였다. 또한 모델의 비교 및 평가를 위해, 대표적인 언어 모델인 Embeddings from Language Model(ELMo)와 문장 임베딩을 위한 모델인 Universal Sentence Encoder(USE)를 사용하였고, 유사관계 식별 결과에 대한 레이블을 가지고 있는 MRPC와 SICK-E 데이터 셋을 활용하였다. MRPC 데이터 셋은 뉴스 기사의 문서 쌍이 의미적 유사성 여부를 식별하며,

SICK-E 데이터 셋은 사진과 영상에 대한 설명문으로부터 추출한 문장 쌍의 관계를 의미적으로 유사함, 대조됨, 관련 없음으로 나누어 식별한다. 실험에는 모델의 정확도와 F1-score를 비교하였으며, 3개의 레이블로 분류되는 SICK-E 데이터 셋에 대해서는 정확도만을 사용해 평가하였다. 실험 결과는 표 2와 같다.

(표 2) 문장의 의미적 유사 관계 식별 실험 결과 요약
(Table 2) The summary of semantic textual similarity identification experiment

Models		MRPC	SICK-E
ELMo	ACC	72.7	80.76
	F1	79.74	-
USE	ACC	66.72	82.44
	F1	73.38	-
LAMII	ACC	80.05	85.06
	F1	86.22	-

다른 모델에 비해 본 논문에서 제안한 LAMII 프레임워크가 전반적으로 우수한 성능을 보여주었으며 특히 문장의 유사 관계를 판단하는 MRPC 데이터 셋에서 높은 성능을 보였다. 이는 본 연구에서 추가한 메트릭 러닝이 실제로 문장의 유사 관계를 식별하는데 도움을 주었음을 나타낸다. SICK-E 데이터 셋 역시 다른 모델과 비교하여 근소하게 좋은 성능을 보였다. 이는 MRPC 데이터 셋과 비교하여 학습 데이터가 보다 풍부한 데이터로 구성되어 모든 모델이 보편적인 수준의 학습이 보장되며, SICK-E 데이터 셋에 존재하는 문장을 구성하는 어휘가 비슷한 의미가 대조되는 문장을 분류하기 어려워 다른 모델과 유사한 성능을 기록하였다.

5. 결론 및 추후 연구

본 연구에서는 전장 상황에 따른 지휘관의 적합한 지휘결심을 지원하기 위한 언어모델 기반의 다출처 정보 통합 프레임워크를 제안하였다. 이질성을 지닌 다출처 정보의 통합을 위해 기존의 언어모델과 다르게 풍부한 정보를 지닌 문장으로부터 이질성을 지닌 파생문장을 생성하여 그 관계를 메트릭 러닝으로 학습하였다. 또한, 전장 상황에 맞는 통합 문서 생성을 위해 전장 상황에서 발생한 이벤트 정보와 같은 토픽을 지닌 문서를 식별하여 전장 상황을 보다 잘 반영한 통합 문서를 생성하였다.

본 논문의 우수성을 평가하기 위해 유사 문장 식별을 포함한 문장 표현 태스크를 수행하고 기존의 언어모델과 그 성능을 비교하였다. 실험 결과 기존 방법과 비교하여 더 뛰어난 성능을 보였으며 본 논문에서 제안한 메트릭 러닝을 활용한 유사 관계 학습이 효과적이었음을 보였다.

그러나 문장을 구성하는 어휘가 유사한 의미가 대조되는 문장을 식별하는데 다소 취약한 성능을 보였으며, 이후 연구를 통해 파생 문장을 생성하는 과정에 반의어 관계를 추가하는 것으로 더 높은 성능을 가진 모델을 생성할 수 있을 것으로 기대된다. 또한 본 논문에서는 실제 국방 데이터가 아닌 그와 유사한 성질을 지닌 데이터를 이용해 실험하였으며, 추후 연구에서 실제 국방 데이터를 활용한다면 국방 데이터의 특수성을 추가로 반영하여 국방 임무에 보다 적합한 모델을 제작할 수 있을 것으로 기대된다.

References

[1] Clark, B., Patt, D., and Schramm, H., "Mosaic warfare: exploiting artificial intelligence and autonomous systems to implement decision-centric operations", Center for Strategic and Budgetary Assessments, 2020. https://csbaonline.org/research/publications/mosaic-warfare-exploiting-artificial-intelligence-and-autonomous-systems-to-implement-decision-centric-operations

[2] Seyong Kim, Hyukjin Kwon, and Minwoo Choi., "The study of Defense Artificial Intelligence and Block-chain Convergence", Journal of Internet Computing and Services, Vol 21, No.2, pp.81-90, 2020. https://doi.org/10.7472/jksii.2020.21.2.81

[3] Changhee Han, Jong-Kwan Lee, "A Methodology for Defense AI Command & Control Platform Construction", The Journal ofKorean Institute of Communications and Information Sciences, Vol.44, No.4, pp.774-781, 2019. https://doi.org/10.7840/kics.2019.44.4.774

[4] Schmidt, E., Work, B., Catz, S., Chien, S., Darby, C., Ford, K., ... and Moore, A., "National Security Commission on Artificial Intelligence (AI)", National Security Commission on Artificial Intellegence, 2021. https://apps.dtic.mil/sti/citations/AD1124333

[5] Bastos, L., Capela, G., Koprulu, A., and Elzinga, G., "Potential of 5G technologies for military application",

International Conference on Military Communication and Information Systems, pp.1-8, 2021.
https://doi.org/10.1109/ICMCIS52405.2021.9486402

[6] Liao, F., Ma, L., Pei, J., and Tan, L., "Combined Self-Attention Mechanism for Chinese Named Entity Recognition in Military", Future Internet, Vol.11, No.8, pp.180, 2019.
https://doi.org/10.3390/fi11080180

[7] Liu, Z., Lin, Y., and Sun, M., "Representation learning for natural language processing", Springer Nature, 2020.
https://doi.org/10.1007/978-981-15-5573-2

[8] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L., "Deep contextualized word representations", 2018.
https://arxiv.org/abs/1802.05365v2

[9] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I., "Improving language understanding by generative pre-training", 2018.
https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805., 2018.
https://arxiv.org/abs/1810.04805v2

[11] Nangia, N., & Bowman, S. R., "Human vs. muppet: A conservative estimate of human performance on the glue benchmark", arXiv preprint arXiv:1905.10425, 2019.
https://arxiv.org/abs/1905.10425

[12] Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X., "Extractive summarization as text matching", arXiv preprint arXiv:2004.08795, 2020.
https://arxiv.org/abs/2004.08795

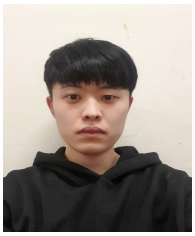
[13] Chen, Yen-Chun, and Mohit Bansal., "Fast abstractive summarization with reinforce-selected sentence rewriting", arXiv preprint arXiv:1805.11080, 2018.
10.18653/v1/P18-1063

[14] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J., "Self-supervised learning: Generative or contrastive", IEEE Transactions on Knowledge and Data Engineering, 2021.
https://doi.org/10.1109/TKDE.2021.3090866

[15] Yin, W., "Meta-learning for few-shot natural language processing: A survey", arXiv preprint arXiv: 2007.09604., 2020.
https://arxiv.org/abs/2007.09604

◐ 저 자 소 개 ◑

김 한 민(Hanmin Kim)
2020년 성균관대학교 시스템경영공학과(공학사)
2020년 ~ 현재 성균관대학교 산업공학과 석사과정
관심분야 : 자연어처리, 인공지능, 기계학습
E-mail : kimhm0705@skku.edu



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

◐ 저 자 소 개 ◑

박 규 동(Gyudong Park)
1994년: 홍익대학교 컴퓨터공학과 졸업
1996년: 홍익대학교 컴퓨터공학과 석사
2014년: 홍익대학교 컴퓨터공학과 박사
1996년 3월~현재: 국방과학연구소 연구원
관심분야: C4I, 가상화, 클라우드, 네트워크

손 미 애(Mye Sohn)
1985년 성균관대학교 산업공학과(공학사)
1988년 한국과학기술원 산업공학과(공학석사)
2002년 한국과학기술원 경영공학과(공학박사)
2004년 ~ 현재 성균관대학교 산업공학과 교수
관심분야 : 인공지능/전문가시스템, 지식그래프, 시맨틱웹, 온톨로지, IOT, 기계학습, 추천시스템
E-mail : myesohn@skku.edu
[ORCID:0000-0002-1951-3493]