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


# 국방 빅데이터/인공지능 활성화를 위한 다중메타데이터 저장소 관리시스템(MRMM) 기술 연구☆

# A Research in Applying Big Data and Artificial Intelligence on Defense Metadata using Multi Repository Meta-Data Management (MRMM)

신 우 택[1*] 이 진 희[1] 김 정 우[1] 신 동 선[1] 이 영 상[1] 황 승 호[1*]
Philip Wootaek Shin Jinhee Lee Jeongwoo Kim, Dongsun Shin Youngsang Lee Seung Ho Hwang

## 요 약

국방부는 감소되는 부대 및 병력자원의 문제해결과 전투력 향상을 위해 4차 산업혁명 기술(빅데이터, AI)의 적극적인 도입을 추진하고 있다. 국방 정보시스템은 업무 영역 및 각군의 특수성에 맞춰 다양하게 개발되어 왔으며, 4차 산업혁명 기술을 적극 활용하기 위해서는 현재 폐쇄적으로 운용하고 있는 국방 데이터 관리체계의 개선이 필요하다. 그러나, 국방 빅데이터 및 인공지능 도입을 위해 전 정보시스템에 데이터 표준을 제정하여 활용하는 것은 보안문제, 각군 업무특성 및 대규모 체계의 표준화 어려움 등으로 제한사항이 있고, 현 국방 데이터 공유체계 제도적으로도 각 체계 상호간 연동 소요를 기반으로 체계간 연동합의를 통해 직접 연동을 통하여 데이터를 제한적으로 공유하고 있는 실정이다. 4차 산업혁명 기술을 적용한 스마트 국방을 구현하기 위해서는 국방 데이터를 공유하여 잘 활용할 수 있는 제도마련이 시급하고, 이를 기술적으로 뒷받침하기 위해 국방상호운용성 관리지침 규정에 따라 도메인 및 코드사전을 생성된 국방 전사 표준과 각 체계별 표준 매핑을 관리하고 표준간 연계를 통하여 데이터 상호 운용성 증진을 지원하는 국방 데이터의 체계적인 표준 관리를 지원하는 다중 데이터 저장소 관리(MRMM) 기술개발이 필요하다. 본 연구에서는 스마트 국방 구현을 위해 가장 기본이 되는 국방 데이터의 도메인 및 코드사전을 생성된 국방 전사 표준과 각 체계별 표준 매핑을 관리하고, 표준간 연계를 통하여 데이터 상호 운용성 증진을 지원하는 다중 데이터 저장소 관리 (MRMM) 기술을 제시하고, 단어의 유사도를 통해 MRMM의 실현 방향성을 구현하였다. MRMM을 바탕으로 전군 DB의 표준화 통합을 좀 더 간편하게 하여 실효성 있는 국방 빅데이터 및 인공지능 데이터 구현환경을 제공하여, 스마트 국방 구현을 위한 막대한 국방예산 절감과 전투력 향상을 위한 전력화 소요기간의 감소를 기대할 수 있다.

☞ 주제어: 국방,메타데이터, 자연어 처리, 표준사전 ,메타데이터 관리, 빅데이터, 인공지능, 머신러닝

## ABSTRACT

The reductions of troops/human resources, and improvement in combat power have made Korean Department of Defense actively adapt 4th Industrial Revolution technology (Artificial Intelligence, Big Data). The defense information system has been developed in various ways according to the task and the uniqueness of each military. In order to take full advantage of the 4th Industrial Revolution technology, it is necessary to improve the closed defense datamanagement system.However, the establishment and usage of data standards in all information systems for the utilization of defense big data and artificial intelligence has limitations due to security issues, business characteristics of each military, anddifficulty in standardizing large-scale systems. Based on the interworking requirements of each system, data sharing is limited through direct linkage through interoperability agreement between systems. In order to implement smart defense using the 4th Industrial Revolution technology, it is urgent to prepare a system that can share defense data and make good use of it. To technically support the defense, it is critical to develop Multi Repository Meta-Data Management (MRMM) that supports systematic standard management of defense data that manages enterprise standard and standard mapping for each system and promotes data interoperability through linkage between standards which obeys the Defense Interoperability Management Development Guidelines. We introduced MRMM, and implemented by using vocabulary similarity using machine learning and statistical approach. Based on MRMM, We expect to simplify the standardization integration of all military databases using artificial intelligence and bigdata. This will lead to huge reduction of defense budget while increasing combat power for implementing smart defense.

☞ keyword: National Defense, Metadata, Natural Language Processing, Standardized Dictionary, Metadata Management, Big Data, Artificial Intelligence, Machine Learning

[1] Datastreams Corp., Seoul, 06651, South Korea
* Corresponding author:
(wtshin@datastreams.co.kr/sshwang@datastreams.co.kr)

[Received 14 November 2019, Reviewed 21 November 2019, Accepted 8 Decebmer 2019]
☆ 국방 데이터 관리정책 방향에 대한 자문을 해주신 아주대학교 정찬기 교수님께 감사드립니다.

# 1. 서 론

4차 산업혁명 추진의 핵심동력인 빅데이터, 인공지능 기술을 국방에 적용하기 위해서는 국방 데이터의 전략적 활용이 필요하다[1,2]. 미국은 데이터를 전략 자산으로 인식하고 있으며, 데이터 기반 의사결정 지원체계를 구축하기 위해 데이터의 표준화, 품질관리, 시각화 기술 등이 상당한 수준으로 발전되었다[3]. 그러나, 현 국방 정보시스템은 폐쇄적인 문화와 보안관련 정책/제도, 데이터 관리 환경 미비 등으로 데이터를 전략적으로 활용하기 위한 기술적인 뒷받침이 부족한 실정이며, 국방 메타 데이터 관리 체계(MDR)[4]가 구축되어 있으나 데이터 표준화를 위한 활용성이 저조한 실정으며, 선진국과의 기술격차, 스마트 국방 구현[5,6]의 시급성을 고려할 때 이를 해결할 수 있는 기술개발이 절실한 상태이다.

국방부문에서 정보체계 개발은 업무 영역별로, 각군의 특수성에 맞춰어 다양하게 개발되어 왔다. 각 시스템간에 발생하는 데이터의 공유체계의 개선이 필요하며, 이는 발전하고 있는 ICT 기술 기반의 네트워크 중심전 (NCW: Network Centric War) 환경을 구축하는데 있어서 필수적이다. 그동안 국방부는 국방 데이터 사전 시스템(DDDS: Defense Data Dictionary System), 국방 부대 코드 관리 시스템 (UCMS: Unit Code Management System), 군대 부호 표준 관리 시스템(SCMS: Symbol Code Management System), 국방 메타데이터 관리 시스템(MDR) 등을 구축하여 데이터 표준을 보급하는데 노력하여 왔다[4].

하지만, 국방정보시템 전체에 데이터 표준을 제정하여 적용하는 것은 보안문제, 각군 업무 특수성, 대규모 체계의 표준화 어려움 등으로 제한사항이 있고, 현 국방 데이터 공유체계 제도적으로도 각 체계 상호간 연동 소요를 기반으로 체계간 연동합의를 통해 직접 연동을 통하여 데이터를 제한적으로 공유하고 있는 한계점이 있는 실정이다[7]. 그러나, 이러한 연동방식은 정보시스템의 숫자와 규모가 증가하면서 국방 데이터 활용 소요의 대폭적인 증가와 연동구조가 복잡해짐에 따라 신기술 활용에 대해 효율적으로 대응하지 못하고 있다. 증가하는 국방 데이터의 전략적 활용을 위해서는 데이터 통합의 요구가 해결되어야 하고 통합을 위해서는 데이터 공유체계가 필수 요소이다[8,9]. 이를 기반으로 국방에 빅데이터 기반 인공지능 기술 적용이 가능하여 실효성 있는 스마트 국방 구현이 가시화 될 것이다.

이에 국방 데이터의 전략적 활용성을 향상시키기 위한 기반기술로 국방 도메인 및 코드사전을 생성된 국방 전사 표준과 각 계계별 표준 매핑을 관리하고 표준간 연계를 통하여 데이터 상호 운용성 증진을 지원하는 국방을 위한 다중 데이터 저장소 관리 (MRMM) 기술의 개발이 필요하다. 본 연구에서는 MRMM을 위한 단어의 표준화를 유사도로 제공하고 개별적으로 표준화 되어있는 표준화 사전을 통합하여 단일 표준사전을 배포하는 시스템을 구축하는데 도움을 주고자 한다.

# 2. 배 경

국방정보시스템의 데이터는 폐쇄적인 군대문화와 보안환경으로 데이터가 어디에 어떻게 존재하고 활용되는지 식별과 접근이 어려운 상태이며, 빅데이터 및 인공지능에 활용하기 위한 데이터 품질도 저조한 실정이다. 가장 중요한 시발점인 데이터 표준화를 위해 국방 데이터 용어에 대한 표준화가 필요한데, 수많은 용어의 표준화를 지원하는 기술이 자연어 처리 기술이다.

자연어 처리란 자연어를 처리하는 분야, 즉 쉽게 말해 인간의 언어를 컴퓨터에게 이해시키기 위한 분야이다. 자연어 처리에서 단어의 의미를 컴퓨터에게 잘 표현하는 방법으로 시소러스를 활용한 기법, 통계 기반 기법, 추론 기반 기법이 있다[10].

## 2.1 자연어 처리 - 통계적 관점

통계적인 방법은 말뭉치(Corpus)를 통하여 자연어 처리를 한다. 말뭉치에서 핵심을 추출하는 방법으로 다음과 같은 절차로 진행된다. 먼저, 말뭉치에서 단어를 분리한다. 이 경우 말뭉치는 용어로 볼 수 있다. 단어를 ID리스트로 이용 할 수 있도록 손질 후 단어의 벡터화(vectorization)를 통해 분산표현(Distributed representation)을 한다.

자연어 처리에서 분포 가설(Distributional Hypothesis)은 단어의 의미는 주변의 단어로부터 형성되며, 즉 단어의 의미는 주변의 문맥(context)로부터 생성된다[11]. 주변의 단어를 확인하는 방법으로 말뭉치로부터 동시발생 행렬(Co-Occurrence Matrix)을 생성하고 주변의 문맥 판별 기준에 따라 윈도우 크기(window size)를 정의할 수 있다. 예를들어 윈도우 크기가 1이면 문맥을 알고 싶은 단어의 양옆 한단어로부터 문맥을 유추한다는 것이다. 윈도우 크기에 따라 추천하는 것이 완전히 달라질 수 있지만 표준사전의 경우 용어가 길어도 10단어 이내이고 대부분의 용어가 5단어 미만으로 구성되고 있다.

벡터의 유사도를 판별하는데에는 유클리드 거리 유사도(Euclidean Distance Similarity), 코사인 유사도(Cosine Similarity)등 다양한 방법들이 존재한다. 단어 벡터의 유사도를 나타낼때에는 코사인 유사도를 많이 쓴다. 다른 방법의 경우 고차원 벡터(high dimensionalty Vector)를 사용할 경우 정확도가 떨어진다[12]. 코사인 유사도는 다음과 같이 정의한다.

$$CosSimilarity(x, y) = \frac{x \bullet y}{\| x \| \| y \|}$$

이 통계적 기법을 개선 하기 위하여 상호정보량(PMI:Pointwise Mutual Information)을 사용한다. 동시발생 행렬의 한계는 단어의 발생빈도 및 문맥과 상관 없이 윈도우 크기이내의 단어들을 수집한다. 그리하여 예를 들어, 자연어 처리에서 중요하지 않은 단어(a, the)의 경우 다른 단어와 연관관계가 깊다고 판단할 수 있어 이 PMI라는 지표로 그러한 요소들을 제거 할 수 있다[13]. 또한 고차원의 벡터를 줄이기 위하여 특이값 분해(SVD:Singlular Value Decomposition)를 사용하는데, SVD로 차원은 감소 시키지만, 특이값 즉 특징은 그대로 유지하는 효과를 볼 수 있다[14].

### 2.2 자연어처리 - 추측적 접근

추론기반 즉, 머신러닝 기반의 자연어 처리 방법에는 여러 가지 선행 모델들 (Word2Vec[15], GloVe[16] 등)이 존재하며, 새로운 네트워크와 모델들을 구성하여 컴퓨터에게 인간의 언어를 가르치려는 끊임없는 노력이 있다.

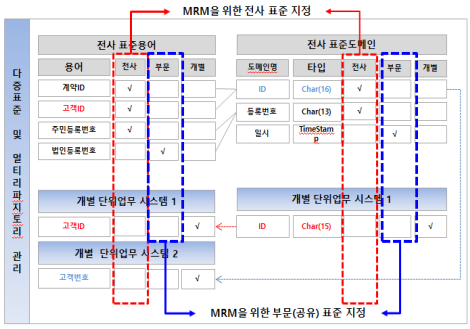
Word2Vec에서 크게 skip-gram 모델과 CBOW(Continous Bag of Words) 모델이 존재한다[15]. CBOW 모델은 주변 단어(맥락)을 사용하여 타겟단어를 추측한다. Skip-gram 모델은 타겟단어로부터 주변 단어(맥락)을 추측한다. GloVe는 통계적인 기법과 머신러닝의 기법을 혼합한 형태의 모델이다.

## 3. 다중 데이터 저장소 관리(MRMM)

군에서도 민간에서 쓰는 기법을 사용하여 개별 체계별 상호간 연동을 하고 있다. 군 자원 자원관리정보체계는 표준화되어 있

지 않고 대상 체계별로 군수 연동 모듈, 전사적 애플리케이션 통합(EAI: Enterprize Application Integration) 등 다양한 연동 방식을 사용하고 있다[4]. 따라서 본 논문에서는 기업의 예를 대상으로 통합기법을 제시하여 이를 군에서 활용할 수 있는 방식을 제시하겠다.

기업의 형태와 관계 없이 이질적 업종으로 조직을 확장할 때, 개별법인 또는 독립적인 사업체로 운영하는 것을 전제로 하는데, 그에 따라 정보 및 DB관리도 개별적으로 추진하는 경향이 있다. 업무 특성상 발생하는 다중 표준과, 그룹간의 표준 공유를 위하여 각 조직별 표준을 기준으로 분산되어 있는 메타시스템간의 데이터 연관성 확보가 필요하며 한 조직 내 다양한 업종을 보유하고 있거나 물리적으로 분리된 시스템 또는 그룹간 표준의 공유 및 데이터 공동 활용을 위해 각각의 분산된 메타 데이터 공유대상에 대한 관리가 필요하다. 또한 각각의 메타데이터에 존재하는 표준간 상호 관계를 정하여 조직간 또는 표준이 다른 시스템간 데이터 공유 기반 시스템의 마련이 필요하다. 이러한 필요성이 있어, 본 논문에서는 다중 데이터 저장소 관리시스템(MRMM: Multi Repository Meta-Data Management)을 제안한다. MRMM이란 업무의 특성상 발생하는 업종별 메터데이터와 개별표준을 기반으로 그룹 및 계열사 간의 표준 공유를 위한 그룹 표준을 정의하고, 그룹 표준을 기준으로 각 표준간의 매핑 또는 포함 관계를 관리하는 시스템이다.



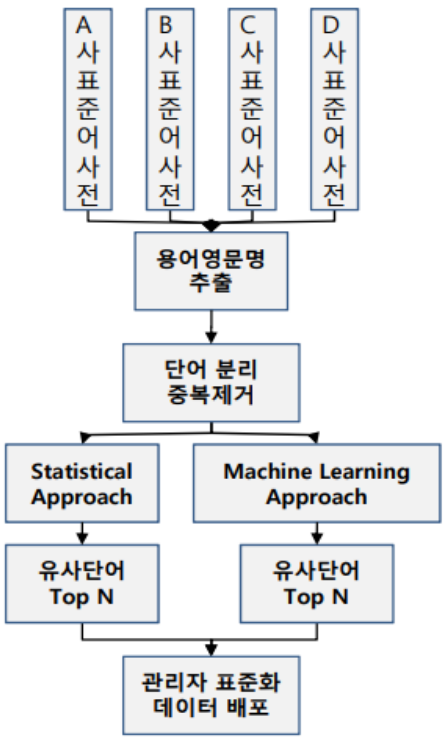

(그림 1) 표준통제형의 예시
(Figure 1) Top to bottom approach example

MRMM은 두가지 형태로 구현이 가능하다. 첫째 표준통제형과 둘째 활용 매핑형이다. 표준 통제형(그림 1)이란 전사 표준을 우선적으로 각 시스템의 표준으로 반영하고, 개별 표준으로 지정된 속성은 개별로 따로 관리한다. 활용 매핑형이란 각 시스템별 개별 메타 데이터를 수집 통합 매핑한 후, 통합표준으로 정제하여 타기관에 배포 공유하는 시스템이다. 표준 통제형의 경우 개별법인을

시작하는 경우에는 전사표준을 적용하여 시스템의 반영에 용이하겠지만, 이미 여러 계열사 및 개별 표준을 가지고 있는 경우 효율적일 수 없다. 활용 매핑형의 경우 각 시스템별 개별 메타데이터를 수집하여 통합 매핑하여 통합표준을 만드는데 인력과 비용이 들지만 한번 통합표준을 만들면 사용이 용이해 진다.

### 3.1 표준용어사전

표준화란 코드, 용어, 데이터 도메인 등의 표준을 수립하여 공공 데이터베이스(DB:Database)에 일관되게 적용하는 일련의 활동을 말하며, 표준용어란 데이터 사용자간의 명확한 의사소통을 지원하기 위해 각 공공기관 에서 업무적으로 사용하고 있는 단어 및 단어의 조합으로 구성되는 용어의 표준을 정의한 것을 말한다. 표준용어사전이란 표준용어와 공통표준용어의 집합을 말한다[18]. 각각 계열사 및 그룹에는 각각 개별적으로 표준화 된 표준용어사전이 존재한다.




(그림 2) 활용 매핑형 구현 flowchart
(Figure 2) Implementation flowchart of bottom to top

이 연구의 중점은 표준 통제형 보다는 기존에 존재하는 데이터베이스에 정의되어있는 각 표준용어사전을 통합하고 정제하여 타기관에 배포하는 활용 매핑형을 더 중점적으로 다룰 것이다. 물론 여기서 실험한 컨셉을 바탕으로 표준 통제형 모델 또한 구축될 수 있을 것이라고 판단된다.

그림 2는 구현된 시스템의 구성도이다. 여러가지 표준사전을 하나로 통합하기 위해서는 일단 각 용어의 유사도의 분석이 필요하다. 용어는 여러 개의 표준화 된 단어로 구성되어 있는데, 용어의 유사도 판단을 하기 위해서는 일단 각 계열사별 단어의 유사 단어 그룹핑을 하고, 동일한 의미의 단어가 여러 개로 정의 되어있다면 단일화 해주어야 한다. 이러한 표준화를 위하여 많은 기업들은 컨설턴트를 고용하여 비용과 시간을 소비하여 단일표준화 사전에 대한 표준화 검토를 한다. 하지만 여러 개의 표준화 사전을 단일 표준화 하려면 더 막대한 자본과 시간이 소모될 것인데, 그것을 해소하고자 표준사전에 자연어처리 적용을 제시한다.

## 4. 표준사전에 자연어처리 적용

표준사전에서 용어들 역시 단어의 집합이므로, 자연어처리의 방식을 활용하여 단어의 유사도를 판단할 수 있을 것이라 사료되어 통계 기반 기법과 추론 기반 기법을 활용하여 단어의 유사도를 판단하였다. 시소러스를 활용한 기법의 경우, 사람의 언어는 여러 연구들이 진행되어 수 많은  유의어 사전이 존재하지만, 표준화 사전 데이터는 대개의 기업들이 데이터를 따로 보존하기 때문에 유의어 사전 방법의 시소러스를 활용한 기법은 실험하기 어렵다고 판단하여 통계 기반 기법과 추론 기반 기법(Machine Learning based Approach)으로 실험을 진행하였다. 표준화 용어 사전의 추론적인 방법은 자연어 처리에서 가장 많이 쓰이는 신경망 중 word2vec에서 제안하는 CBOW모델을 사용하였다. 왜냐하면 CBOW 모델이 단어 임베딩(word embedding)의 연관성(relatedness)과 유추(analogy) 실험에서 가장 효과적으로 보여지기 때문이다[19].

통계적인 기법은 코사인 유사도를 계산한 후 데이터를 분석하여 결과만 뽑아 내면 된다. 하지만 추론적인 기법, 즉 머신러닝의 기법의 경우 훈련(training)을 필요로 하여 하이퍼파라미터(hyperparameter)의 튜닝을 필요로 한다.

통계적인 기법과 추론적인 기법 모두 용어의 크기가 작기 때문에 윈도우 크기를 크게 하지 않고 1로 고정하여

실험을 진행하였다. 하지만 필요의 경우 윈도우 크기를 사용자가 조절할 수 있게 구현하였다. 또한 단어 유사도를 코사인 유사도를 구하여 유사한 결과를 기준으로 유사단어 랭킹을 표시할 수 있으며, 필요에 따라 유사도 랭킹의 단어를 원하는 만큼 확인할 수 있게 구현하였다.

### 4.1 훈련 환경(Training Setting)

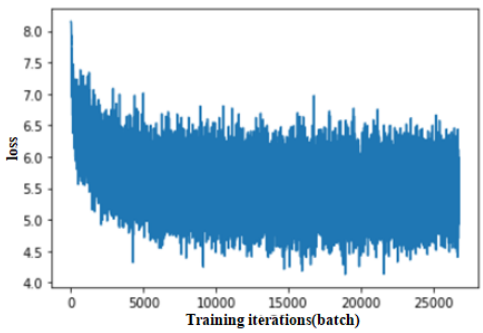
사용한 데이터셋의 경우 자사에서 보유한 고객의 데이터 A사와 B사의 데이터를 사용하였다. 데이터의 이질성을 최대한 배제하기 위하여 A사와 B사의 데이터는 같은 업종(증권)으로 선택하였다. 하지만 두 데이터는 각사의 표준에 맞추어 표준화 되어 있고 같은 업종이여서 유사한 용어가 많을 것으로 판단된다. A사의 경우 28987건의 표준화된 용어가 등록되어 있으며, 단어를 분리할 경우 3461개의 중복되지 않는 단어를 확인할 수 있었고 B사의 경우 14835개의 표준화된 용어가 등록되어 있었고 2218개의 단어로 분리 할 수 있었다. 두 단어를 총 합치면 5679개이지만 두 사에서 동일하게 사용하는 단어는 518개이다.

메터데이터의 많은 데이터들 중 용어영문명을 사용하였고 필요시에는 용어영문명을 용어한글명과 매핑하여 활용할 계획이다. 영어단어의 벡터화기법(vectorization method)들은 많지만 한글의 경우 형태소 분석을 해야할 뿐만 아니라 한글로 된 자연어처리 연구가 영어에 비해 활성화 되어 있지 않아 용어영문명을 활용하여 채택하였다.

기존의 머신러닝을 훈련할때에는 훈련셋(Training Set), 검증셋(Validation Set), 평가셋(Test Set)으로 분리하여 진행하는 것이 옳지만 이 데이터의 경우 사전에 용어/단어의 표준화 과정이 존재하여 평가셋과 검증셋이 무의미한 것이라고 판단된다. 왜냐하면, 표준화 되지 않은 용어 자체가 등록이 안되고, 또 표준화 작업이 선행되어야 하기 때문에 유사한 단어를 찾을 때, 찾으려는 단어가 등록 되어있지 않으면 유사한 용어를 검색할 수 없다. 그리하여 등록되어있지 않은 단어는 사용자에게 표준화 등록이 선행되어야 한다고 사용자에게 알려준다.

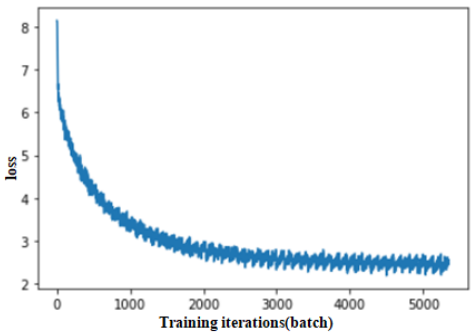
### 4.2 하이퍼파라미터 튜닝

그림 3에서 보면 하이퍼파라미터를 튜닝하지 않는 경우 손실(loss)이 수렴하지 않는 것을 볼 수 있고 25000 계산(iteration)이 된 이후에도 손실이 6.5와 4.5 사이에 분포되어 있는 것을 확인할 수 있다. 여기서 loss는 cross entropy error를 의미하며, iteration은 이폭(epoch)을 배치의 크기(batchsize)로 나눈것이다.



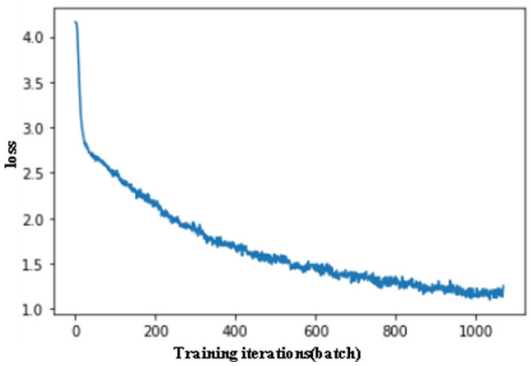

(그림 3) Fine-tuning하기 전 loss-graph
(Figure 3) Loss-graph before fine-tunning

그림 4를 보면, Fine-tuning을 한 후 손실을 보면 5000 iteration까지 손실이 수렴하는 것을 확인할 수 있으며 수렴범위 또한2와 3 사이로 수렴하는 것을 확인 할 수 있다. 여기서 그림 3과 비교하여 변형한 하이퍼파라미터는 배치의 크기, 은닉층의 크기(hidden layer size)이다. 또한 수렴을 확인하였을 때 이폭(epoch)의 수도 줄여 트레이닝 시간을 더 용이하게 했다.




(그림 4) Fine-tuning 한 후의 loss graph
(Figure 4) Loss graph after Fine-tuning

자연어 처리 분야에서 Word2vec을 개선하기 위한 많은 연구가 진행되었다. 은닉층(Embedding layer)은 단어의 밀집벡터 표현을 저장하고 그 효과로 메모리 사용량과 불필요한 계산을 줄이는 효과를 준다[20]. 또한 부정적 샘플링(Negative Sampling)을 통해 소프트맥스층(softmax)의 계산량을 제거하고 은닉층 이후의 계산량을 제거할 수 있다[17].




(그림 5) Negative Sampling과 embedded layer 추가한 후의 loss graph

(Figure 5) Loss graph after applying negative sampling, and embedding layer

이 두 기법을 적용하여 word2vec을 개선한 결과(그림 5)를 보면 손실이 트레이닝 계산(training iteration)을 그림 4의 5분의 1로만 줄여도 훨씬 더 낮은 수준으로 도달하는 것을 확인할 수있고 loss 가 1.0과 1.5사이에서 수렴하는 것을 확인 할 수 있었다.

## 5. 결 과

CBOW모델과 통계적 모델의 경우 그림 6와 같이 유사한 단어 TOP N개를 뽑을 수 있다. N은 사용자가 임의로 설정하여 최고로 유사한 것부터 N개를 뽑을 수 있다.

```
[STAT]YMD                          [INFER] YMD
 YM: 0.654021775683565              RTUCH: 0.6982421875
 DT: 0.5555176661189032             EXT: 0.68212890625
 NTRADE: 0.5232153002701556         DAYNUB: 0.66943359375
 SERV: 0.5201294356033754           WANT: 0.6630859375
 CAROP: 0.49458104110733025         YM: 0.6435546875
```

(그림 6) 결과 예시, 좌측의 STAT은 통계적인 방법에서 우측의 INFER은 추론적인 결과를 중 유사한 상위 5개의 단어를 보여주는 것이다.

(Figure 6) Example of result that shows top 5 result, STAT refers to the statistical approach and Infer refers to inference approach

결과를 검증하기 위해서 5161개의 단어 모두를 검증하는 것이 가장 이상적이지만, 임의로 단어를 선별하여 주관적인 검증을 하였다. 각 사마다 고유로 쓰는 단어들이 존재하고, 또 그것의 유사도는 물론 TOP 5를 추출할 경우 통계적 관점이나 추론적관점에서 나올 수는 있지만, 그 고유의 단어에 대한 검증자체는 불필요하기 때문이다.

### 5.1 단일 표준사전 분석

단일한 표준화 사전이 있는 DB데이터에서도 같은 의미의 단어를 추출할 수 있을 것이라고 생각되었다. 왜냐하면 표준화를 하더라도 사람의 컨설팅으로 하는 것이기 때문에 분명 누락되는 것이 있을 것이라고 사료되기 때문이다. 그리하여 A사의 데이터 즉 28987건의 표준화된 용어, 3461개의 단어만 사용하여 통계적인 관점과 추론적인 관점을 분석하여 보았다. 모든 단어들을 검수하는 것이 가장 이상적이지만, 이 단어 내에는 그 고객사에 특정적으로 사용하는 단어를 확인할 수 있고 하여 단어중 임의의 표본 5%를 뽑고 그중 포괄적으로 다른 DB에서도 확인 할 수 있는 단어들을 선별하여 결과를 도출했다. 그 결과 통계적인 관점에서는 똑같은 의미의 단어를 확인할 수 없었지만 머신러닝을 활용한 추론적인 관점에서는 다음과 같은 결과가 나왔다.

```
SELL: 매도,

 [STAT]SELL                        [INFER] SELL
  BUY: 0.9642017308846954           BUY: 0.814453125
  RPY: 0.8808367937491628           SNS: 0.7880859375
  LIM: 0.8743115569846881           OPNT: 0.76708984375
  CNTC: 0.8650075771560007          BWRRT: 0.7509765625
  FACE: 0.8545202908676623          DD3: 0.75

BUY: 매수 RPY: 상환, LIM: 한도 CNTC: 계약 FACE: 액면,
BUY: 매수 SNS: 매도 OPNT: 개장, BWRRT: 신주인수권, DD3: 3일 이내,

NEW: 신규,
 [STAT]NEW                         [INFER] NEW
  TDAY: 0.7023347505512263          WEB: 0.78759765625
  BUY: 0.6847815150771549           NW: 0.75048828125
  RPY: 0.6793002214912655           ODRY: 0.72705078125
  SELL: 0.6761580260902932          BKAY: 0.71240234375
  PID: 0.6726376984053399           CGRP: 0.70654296875

TDAY: 당일, BUY: 매수, RPY: 상환 SELL: 매도 PID: 전일,
WEB: 웹, NW: 신 ODRY: 통상 BKAY: 이탈, CGRP: C그룹,
```

(그림 7) DB한 개에서 얻은 유의미한 결과

(Figure 7) Result of single database

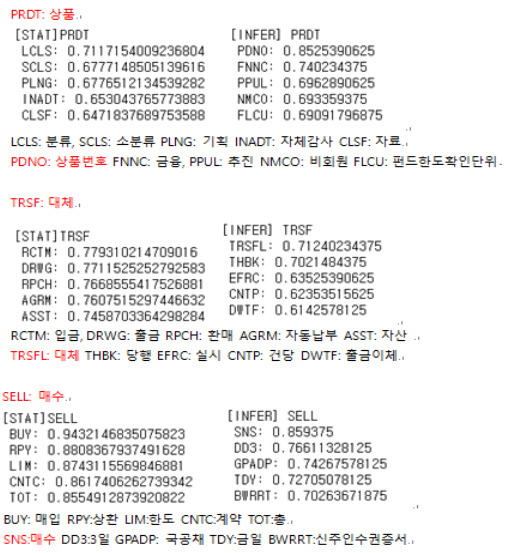
그림 7에서와 같이 [STAT]은 통계적인 기반방법의 결과를 표시한 것이고 [INFER]는 머신러닝 기반의 결과를 표시한 것이다. 단어 옆에 각 숫자들은 cosine similarity의 값이며 값이 크면 더 유사하다고 시스템에서 판단한 것이다.

단일 DB를 검수해본 결과 SELL과 NEW라는 의미의 단어가 통계 기반의 방법에서는 보이지 않지만, 동일한 뜻의 단어 SNS와 NW를 머신러닝을 활용한 기법에서는 확인할수 있었다. 즉 한 DB내의 표준화가 덜 된 단어들

을 머신러닝 기법으로 확인하였고, 표준화 절차에도 도움을 줄 수 있는 것을 확인할 수 있었다.

### 5.2 다중 표준사전 분석

다중 DB의 검수를 하기 위해서는 모든 DB에서 동일하게 사용되는 단어들을 먼저 확인해야 한다. 3 장에서 언급했듯이 중복되는 단어가 표준화된 사전마다 존재할 수 있다. 이러한 경우 DB들을 통합하기 전에 중복되는 단어를 제거해 TOP N을 추출할 때 중복되는 단어의 검출을 피한다. B 회사의 데이터를 통합하면 5679개의 단어이지만, 중복되는 단어 518개를 제거하면 5161개의 표준화된 단어를 가지고 실험을 진행하였다. 5.1과 동일하게 단어중 임의의 표본 5%를 뽑고 분석하였다. 그 결과는 다음과 같다.




(그림 8)다중DB에서 얻은 유의미한 결과
(Figure 8) Result of multiple databases

그림 8와 같이 여러 DB를 통합하여 결과를 확인해 본 결과, 한 DB에서 찾을수 있었던 매수(SELL, SNS)를 DB를 통합하여도 확인할 수 있었다. 또한 똑같은 의미의 다른 DB의 단어(TRSF,TRSFL)도 확인이 가능하였고 PRDT라는 A회사의 용어의 유사어로 PDNO의 B회사의 용어가 나왔는데, 이것은 표준화 작업을 다시 거쳐 더 작은단위의 단어로 쪼개서 표준화 작업을 진행해야 하는 단어도 제시 되었다.

확인할 수 있었던 특징은, 머신러닝을 적용한 단어들은 단어의 의미가 비슷한 단어들이 많이 선별되었고 통계적인 관점의 경우, 의미보다는 형태가 비슷한 경우가 많이 보이는 것을 확인할 수 있었다. 또한 데이터가 많아지면 통계적인 관점은 같은 형태를 찾는 어려움이 있었지만, 머신러닝을 활용한 추론적인 관점은 의미적으로 유사하거나 관련성이 있는 단어를 TOP5로 잘 제시하였다.

## 6. 결 론

본 논문에서는 국방 데이터의 표준화 기법향상을 통한 빅데이터, 인공지능 기술의 전략적 활용성 향상을 위해 국방상호운용성 관리지침규정에 따라 도메인 및 코드사전을 생성된 국방 전사 표준과 각 계계별 표준 매핑을 관리하고 표준간 연계를 통하여 데이터 상호 운용성 증진을 지원하는 국방을 위한 다중 데이터 저장소 관리(MRMM) 기술을 제시하고, 단어의 유사도를 통해 MRMM의 실현방향성을 구현하였다. 이 기술을 국방정보시스템에 적용할 경우 데이터 표준화율을 높이고 좀더 실효성있는 빅데이터 및 인공지능 기술개발에 기여할 것으로 기대된다.

자연어 처리의 경우 연관성(Relatedness), 유추(Analogy), 분류(Categorization), 선택적 선호도(Selectional preference) 등 다양한 방법의 평가법이 있다[19]. 왜냐하면 언어에는 주술목 관계가 존재하며 또 문법적으로 틀린 문장들이 존재하고 또 자연어 처리에 대한 선행 연구가 컴퓨터 생성이래에 계속 연구가 되었기 때문이다. 하지만 표준화 사전의 경우, 계열사가 아닌 경우 이질적인 데이터가 존재하기 때문에, 통계적 관점 혹은 추론적 관점이던 표준화된 방향에 따라 차이가 있을 수 있다. 그래서 기업들이 몇 개월에 걸쳐 컨설턴트를 고용해 표준화된 사전이 제대로 되었는지 확인한다. 그러한 의미에서 이 연구는 여러 DB를 통합하는 표준화 작업을 도와줄 수 있을 것이라고 사료된다. 하지만, 개인정보법에 따라 여러 기업의 표준화 사전은 각 회사들의 비밀이 되어 실질적인 데이터가 부족하여 명확한 구별을 하는데 어려움이 존재하였다.

향후 연구에서는 단어의 유사도로부터 더 확장하여 용어의 유사도 측면까지 확인하여 단어의 조합이 비슷한 경우 그 것 또한 머신러닝의 방법으로 추출할 수 있을 것이라고 판단된다.

## References


[1] Seong-Woo Kim. Gak-Gyu Kim, Bong-Kyu Yoon, "A Study on a Way to Utilize Big Data Analytics in the Defense Area", Journal of the Korean Operations Research and Management Science Society, Vol. 75. No 2, pp. 1-19, June 2014. http://dx.doi.org/10.7737/JKORMS.2014.39.2.001

[2] 국경완, "4차 산업혁명 시대 인공지능을 활용한 군사적 적용방안," 합참지, 2018. 7.

[3] 산업연구원 (KIET), "미국 신정부 국방획득정책 변화 및 대응전략 연구," 연구보고서, 2018.

[4] 김영도, 이한준, 홍진기, "국방데이터공유체제 개선방안과 추진과제," 주간국방논단, 제1631호(16-34) 2016.

[5] 국방부, "2019년 국방부 업무 보고"

[6] 국방부4차 산업혁명 스마트 국방혁신 추진단, "4차 산업 혁명과 함께 미래 국군 그린다: 4차 산업혁명 스마트 국방 혁신," 2019년 3월 15일

[7] 국방일보, "국방정보체계 유지보수 전담 '국방업무 효율화' 기여할것," URL:http://kookbang.dema.mil.kr/newsWeb/m/20181024/30/BBSMSTR_000000010026/view.do

[8] 이미영, 이상철, "국방정보체계의 상호 운용성 보장을 위한 메타데이터 표준화 구축연구", 대한산업공학회 추계학술대회 논문집, pp. 317-324, 2009.

[9] 아주대학교 장위국방연구소, "국방아키텍쳐 메타모델 표준화 구축 및 활용방안 연구", 연구보고서, 2017

[10] 齋藤 康毅(사이토 고키), "밑바닥부터 시작하는 딥러닝2", 2019 , 한빛미디어

[11] Z. Harris, "Distributional structure". Word, 10(23): pp. 146-162. 1954.

[12] D. Ayata, "Applying Machine Learning and Natural Language Processing Techniques to Twitter Sentiment Classification for Turkish and English." Thesis for M.S. degree at Bogazici University. June 2018.

[13] K. W. Church, P. Hanks, "Word association norms, mutual information, and lexicography," Compututational linguistic Vol. 16, No. 1, pp 22- 29, 1990

[14] G. H. Golub, C. Reinsch, "Singular Value Decomposition and Least Squares Solutions," , vol. 14, pp. 403-420, 1970.

[15] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space,." arXiv prepreint arXiv:1301.3781, 2013.

[16] J. Pennington, R. Socher, C. Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Quatar, pp.1532-1543, Oct. 2014.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representation of Words and Phrases and their Compositionality" Advances in neural information processing systems, Vol. 26, 2013.

[18] 행정자치부, "공공기관의 데이터베이스 표준화 지침," 행정자치부고시 제2015-26호, 2015.

[19] T. Schnabel, I. Labutov, D. Mimno, T. Joachims, "Evaluation methods for unsupervised word embeddings," Proceeding of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 298-307, 2015

## ◐ 저 자 소 개 ◑


**신 우 택(Philip Wootaek Shin)**
2018년 The Pennsylvania State University 컴퓨터공학과 학사(Honors)
Magna Cum Laude 수여
2019년 The Pennsylvania State University 컴퓨터공학과 석사
2019년 ~ 현재 ㈜ 데이터스트림즈 연구원
관심분야 : 머신러닝, 딥러닝, 빅데이터, 데이터 거버넌스
E-mail : wtshin@datatstreams.co.kr

**이 진 희(Jinhee Lee)**
2003년 경기대학교 전자계산학과 졸업
2015년 ~ 현재 ㈜ 데이터스트림즈 책임 연구원
관심분야 : 데이터 거버넌스, 데이터 품질관리
E-mail : jinhlee@datastreams.co.kr

**김 정 우(Jeongwoo Kim)**
1997년 한국외국어대학교 전자물리학과 졸업
2004년 국립부경대학교 컴퓨터공학과 석사
2009년 배제대학교 정보통신공학과 박사수료
2018년 국방부 SW정책담당
2019년 ~ 현재 ㈜ 데이터스트림즈 PPC부문 이사
관심분야 : 인공지능, 빅데이터, 데이터 거버넌스
E-mail : jwkim@datastreams.co.kr

**신 동 선(DongSun Shin)**
1985년 충북대학교 수학과 졸업
1995년 한국생산성본부 전문위원
2009년 ~ 현재 ㈜ 데이터스트림즈 거버넌스 총괄 상무
관심분야 : 데이터 품질관리, 데이터 거버넌스
E-mail : dsshin@datastreams.co.kr

## ◐ 저 자 소 개 ◑

**이 영 상(YoungSang Lee)**
1986년 경북대학교 전자공학(전산) 졸업
1989년 미시간 주립대학 전자공학 석사
2003년 한국과학기술원(KAIST) 전자공학 박사과정
2001년~현재 ㈜ 데이터스트림즈 대표이사
2010년~2012년 한국SW전문기업협회 회장
2012년~현재 한국상용SW협회 명예회장
2013년~현재 한국빅데이터학회 부회장
2014년~현재 한국PMO협회 명예회장
관심분야 : 데이터 품질관리, 데이터 거버넌스, 빅데이터
E-mail : yslee@datastreams.co.kr

**황 승 호(Seung Ho Hwang)**
1979년 서울대학교 전자공학과 졸업
1981년 한국과학기술원(KAIST) 전자공학 석사
1989년 University of California, Berkeley 전자공학 박사
1990년~1998년 한국과학기술원(KAIST) 전기 및 전자공학과 교수
1996년~2006년 Silicon Image 부사장
2006년~2013년 삼성전자 부사장
2014년~2019년 현대자동차 부사장
2019년~현재 ㈜ 데이터스트림즈 사장
관심분야 : 빅데이터 및 분석
E-mail : shhwang@datastreams.co.kr