자연어처리 기법을 적용한 무기체계의 상호운용성 평가방법

Evaluation method for interoperability of weapon systems applying natural language processing techniques

김용균*,1) . 이동현2)

Yong-Gyun Kim*,1) . Dong-Hyen Lee2)

[초 록]

현재의 무기체계는 다양한 표준과 프로토콜이 적용된 복합무기체계가 운용되어서 전장에서 연합 및 합동작전시 원활한 정보교환 실패의 위험이 있다. 무기체계간 신속한 상황판단으로 핵심표적에 대한 정밀타격을 수행하기 위한 무기체계들의 상호운용성은 전쟁수행의 핵심요소이다. 한국군은 전력화 이후 다수의 소프트웨어 및 하드웨어의 형상변경과 성능개선 소요가 발생하고 있으나, 상호운용성에 미치는 영향에 대한 검증제도가 없으며, 관련 시험 도구 및 시설도 전무한 실정이다. 또한 연합 및 합동훈련시 무기 / 전력지원체계의 세부 운용방식과 소프트웨어를 임의로 변경한 후 이에 따른 사용자 간 오류가 빈번히 발생하고 있다. 그래서 주기적인 무기체계간 상호운용성 검증이 필요하다. 이러한 문제를 해결하기 위하여 사람이 평가기간을 잡아서 1번 평가를 진행하는것이 아니라, AI가 24시간 무기 / 전력지원 체계간 상호운용성을 지속적으로 평가하여 전쟁수행 능력을 고도화해야 한다, 이러한 문제점을 해결하기 위하여 자연어 처리기법(①Word2Vec 모델 ②FastText 모델 ③Swivel 모델)을 적용(공개된 알고리즘과 소스코드 사용)하여 국방상호운용성 능력향상을 위한 사전연구를 수행하였다. 이 실험의 결과를 바탕으로 사람에 의존하지 않고, 자동화된 국방상호운용성 평가도구를 구현하기 위한 방법론(자연어 처리 모델을 통한 상호운용성 소요평가 / 수준측정의 자동화된 평가)을 향후 제시하고자 한다.

[ABSTRACT]

The current weapon system is operated as a complex weapon system with various standards and protocols applied, so there is a risk of failure in smooth information exchange during combined and joint operations on the battlefield. The interoperability of weapon systems to carry out precise strikes on key targets through rapid situational judgment between weapon systems is a key element in the conduct of war. Since the Korean military went into service, there has been a need to change the configuration and improve performance of a large number of software and hardware, but there is no verification system for the impact on interoperability, and there are no related test tools and facilities. In addition, during combined and joint training, errors frequently occur during use after arbitrarily changing the detailed operation method and software of the weapon/power support system. Therefore, periodic verification of interoperability between weapon systems is necessary. To solve this problem, rather than having people schedule an evaluation period and conduct the evaluation once, AI should continuously evaluate the interoperability between weapons and power support systems 24 hours a day to advance warfighting capabilities. To solve these problems, To this end, preliminary research was conducted to improve defense interoperability capabilities by applying natural language processing techniques (①Word2Vec model, ②FastText model, ③Swivel model) (using published algorithms and source code). Based on the results of this experiment, we would like to present a methodology (automated evaluation of interoperability requirements evaluation / level measurement through natural language processing model) to implement an automated defense interoperability evaluation tool without relying on humans.

Key Words : Natural language processing(자연어 처리), Deep learning(딥러닝), Interoperability test(상호운용성 시험평가), Weapon system(무기체계)

1) 국군지휘통신사령부 합동상호운용성기술센터 (Joint Interoperability Technology Center, Armed Forces Command and Communications Command, Department of Defense, Korea) 2) 메디컬아이피 ITC 소프트웨어 (Medical IP ITC Software)

* Corresponding author, E-mail: ygzard@naver.com
Received : August 29, 2023 Revised : Accepted : September 22, 2023

1. 서 론

현재의 무기체계는 다양한 표준과 프로토콜이 적용된 복합 무기체계의 운용으로 전장에서 연합 및 합동작전시 원활한 정보교환 실패의 위험이 있으며, 무기체계 운용개념 변경 시 기존체계(Legacy System) 간 연동을 위해서는 성능개량이 필요하다. 또한 전술데이터링크(TDL)를 탑재한 첨단 무기체계의 운용 증가에 따라 실시간 정보교환 및 전장상황인식, 상·하급부대 간 정보공유가 필수적이며, 敵의 핵 및 미사일, 대량살상무기(WMD)의 위협증가에 따라 신속한 상황판단으로 핵심표적에 대한 정밀타격을 수행하기 위한 무기체계 간 상호운용성이 전쟁수행의 핵심요소이다. 무기체계와 전력지원체계의 상호운용성을 확보하기 위한 국방상호운용성 관리지시(국방부지시 제14-2001)의 상호운용성(Interoperability)의 정의는 다음과 같다.『각각 다른 운용목적을 가진 2개 이상의 체계 간에 잘 정의된 인터페이스를 통해 의미 있는 정보를 교환하고 이용할 수 있는 능력. 또는 서로 다른 군, 부대 또는 체계 간 특정 서비스, 정보 또는 데이터를 막힘없이 공유, 교환 및 운용할 수 있는 능력』이다. 상호운용성의 기본원칙은, 상호운용성 관리지침(방위사업청지침)(제2012-13호)에 다음과 같이 정의 되어 있다. ①네트워크 중심전(NCW) 및 합동개념에 입각한 각 체계간 상호운용성을 확보하여 연합 및 합동작전의 효율성 제고 ②서로 다른 체계간 시간과 장소에 구애받지 않는 정보의 공유 ③사업관련 기관(부서)은 네트워크 중심의 국방정보화환경에서 전투원이 정보를 원활하게 사용할 수 있도록 정보유통이 보장된 무기체계를 구축하여야 한다.

한국군은 무기체계 획득 시 기존의 플랫폼 중심(Platform-Centric)과 연동 위주의 상호운용성 개념을 벗어나서, 현대전의 개념인 네트워크 중심전(NCW) 수행에 필수요소인 '센서-투-슈터' 실현을 위한 '네트워크 중심(Net-Centric)의 무기체계 상호운용성 구현'을 목표로 상호운용성 개념을 일대 전환하여야 한다. 상호운용성의 빠른 기술 및 표준 변화를 고려하여 무기체계 / 전력지원 체계간 주기적인 상호운용성 능력을 측정하기 위한 Process의 도입이 필요하다. 그러나 그것을 실행할수 있는 자동화된 평가 소프트웨어가 없고, 전력화 이후 별도의 관리 또한 없어서, 후발 체계와의 상호운용성 통신오류(메시지 교환) 발생 가능성도 높다.

이러한 누적된 문제점의 대안으로 검증된 자연어 처리기법을 적용(공개된 알고리즘과 소스코드 사용)하여 무기체계간 상호운용성 능력향상을 위한 사전연구를 수행하였다. 이를 통하여 사람에 의존하지 않고, 자동화된 국방 상호운용성 평가도구를 구현하기 위한 방법론(자연어 처리 모델을 통한 상호운용성 소요평가 / 수준측정의 자동화된 평가)을 향후 제시하고자 한다.

2. 상호운용성에 평가에 관한 업무 프로세스

2.1 상호운용성의 정의

국방전력발전업무훈령(국방부훈령 / 제2749호)의 제15조(작전운용성능 결정)에서『작전운용성능은 주요 작전운용성능, 합동성 및 상호운용성, 보안대책, 기술적·부수적 성능으로 구성된다.』상호운용성의 중요성을 명시하고 있다. 국방전력발전업무훈령(국방부훈령 / 제2749호)의 제64조(개발시험평가계획 수립)의 5항『 5. 합동성 및 상호운용성시험 가. 운용개념 및 체계 특성 나. 연동성 및 정보교환』으로 시험평가 항목을 정의하고 있다.

2.2 상호운용성을 보장하기 위한 단계별 평가 활동

국방상호운용성 관리지시(국방부지시 제 14-2001)의 제98조(상호운용성 평가 업무)는『상호운용성 소요평가, 수준평가, 상호운용성 시험평가, 적합성 검토·시험 등 상호운용성을 평가하는 일련의 활동』으로 명시되어 있다. 획득단계별로 수행하는 상호운용성 평가는 다음 각 호와 같다.

1. 소요기획: 상호운용성 소요평가
2. 정보화전략계획수립개념연구: 수준측정
3. 탐색개발: 수준측정, 운용성 확인
4. 체계개발: 수준측정
5. 시험평가: 개발시험평가, 운용시험평가, 구매시험평가

으로 명시되어 있다.

2.3 상호운용성을 보장하기 위한 문서의 작성(연동합의서 / 연동통제문서 / 상호운용성 확보계획서)

국방상호운용성 관리지시(국방부지시 제 14-2001)의 제78조(연동합의서 작성) 1항에서『연동합의서는 연동에 대한 각 책임부서와 개괄적인 연동방식 및 항목 등을 약속하는 문서로서, 향후 연동통제문서의 작성 및 연동기능 구현을 위한 제반사항들을 지원하는 근거문서로 활용한다』으로 명시하고 있다.

국방상호운용성 관리지시(국방부지시 제 14-2001)의 제79조(연동통제문서 작성) 1항에서『연동통제문서는 연동에 대한 책임 및 범위, 연동을 구현하기 위한 기술과 데이터 포맷을 상세하게 정의함으로써 상호운용성 시험평가의 기준자료로 활용』으로 명시하고 있다. 상호운용성 관리지침(방위사업청지침)(제2012-13호)의 제8조(상호운용성 확보계획 작성)에서『상호운용성 적용 항목에 대하여 확보계획을 획득단계별로 구체화하여 작성한다』으로 명시하고 있다.

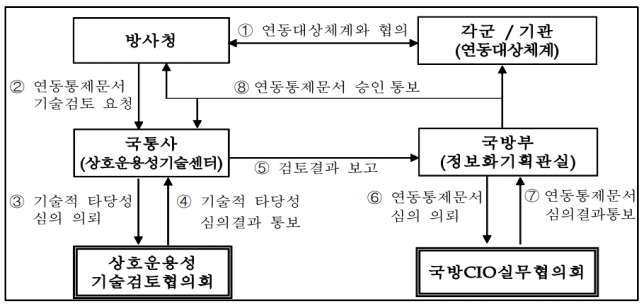

그림 1. 국방상호운용성 관리지시 별표 3 /
전장관리 · M&S 체계 연동통제문서 검토 절차도
Defense Interoperability Management Directive asterisk 3 /
Battlefield Management · M&S System interlocking
control document review procedure diagram

2.4 상호운용성을 보장하기 위한 소요평가의 평가 활동

국방상호운용성 관리지시(국방부지시 제 14-2001)의 102조(소요평가 항목 및 기준) 1항에서 『상호운용성 소요평가 항목은 1. 운용개념 및 체계특성 2. 연동성 및 정보교환(연동성 : 가) 연동대상체계, 연동개념도, 연동 방법 또는 기능을 작성한다. 나) 정보교환 능력 : 정보교환내역/목록을 작성한다. 다) 상호운용성 수준 : 일반 상호운용성 수준, 특정 상호운용성 수준을 작성한다』으로 명시하고 있다.

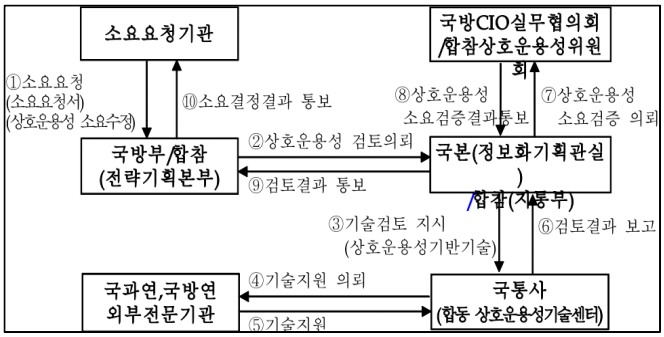

그림 2. 국방상호운용성 관리지시 별표 6 / 상호운용성 소요평가 및 검증절차

Defense Interoperability Management Directive asterisk 6 / Interoperability evaluation and verification procedures

2.5 상호운용성을 보장하기 위한 수준측정의 평가 활동

상호운용성 관리지침(방위사업청지침 / 제2012-13호)의 제27조(상호운용성 수준측정 목적) 『무기체계 상호운용성 수준측정의 목적은 4항 연동대상 무기체계와의 상호운용성 확보 전략 수립 5항 목표 상호운용성 수준 달성여부 판정』으로 명시하고 있다.

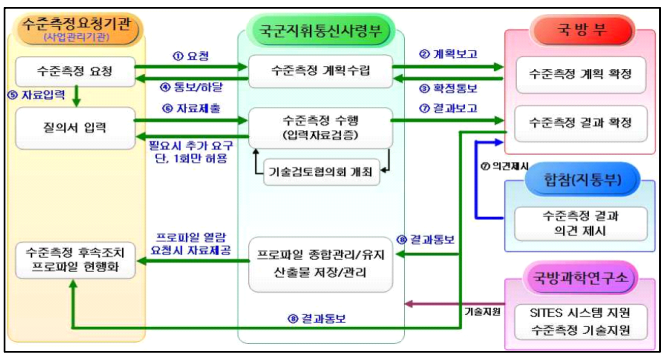

그림 3. 국방상호운용성 관리지시 별표 8 / 상호운용성 수준측정 절차도

Defense Interoperability Management Directive asterisk 8 / Interoperability level measurement procedure diagram

2.6 한국군 무기체계 / 전력지원체계의 상호운용성 확보時 개선 제안사항

한국군은 전력화 이후 다수의 소프트웨어 및 하드웨어의 형상변경과 성능개선 소요가 발생하고 있으나, 상호운용성에 미치는 영향에 대한 검증제도가 없으며, 관련 시험 도구 및 시설도 전무한 실정이다. 즉 주기적인 상호운용성 검증이 필요하다. 연합 및 합동 훈련시 무기 / 전력지원체계의 세부 운용방식과 소프트웨어를 임의로 변경한 후 이에 따른 사용 간 오류가 빈번히 발생하고 있다. 그리고 훈련 목적상 변경사항에 대한 상호운용성 영향여부에 대한 사전 검증 능력이 부재한 실정이다.

방위력개선분야의 시험평가제도상 상호운용성에 대한 중요성 인식이 부족하다. 현재는 “군 운용 적합성” 분야의 하위요소로 상호운용성 평가를 취급함에 따라 시험평가 결과에 대한 반영이 미흡하고, 체계간 상호운용성 오류비율이 과다하게 나타나고 있다.

2.7 한국군 상호운용성 수준측정의 현황분석

미군의 정보체계 중심의 상호운용성 수준을 측정하기 위해 최초 성숙도 모델개념을 도입한 LISI(Level of information system interoperability) 모델을 개발하였다. 한국군은 2006년 미군의 LISI 모델을 부분적으로 적용한 수준측정 평가도구(SITE)를 개발하여 상호운용성 수준측정 성숙도 수준을 측정하고 있다. 연동대상체계를 고려한 기술표준의 적용이 상호운용성의 확보 가능성을 높일수 있다. 그러나 상호 운용성수준이라는 개념이 고도화된 상호운용성의 요구능력에 대한 정도를 표현하기 위하여 정의되었기 때문에 체계간의 상호운용성을 확보하기 위해서 연동성 및 정보교환 관점에서 대상 무기체계의 운용개념, 작전 요구성능, 연동성 확보를 위하여 다음과 같이 Integration readiness assessment를 수행한다.

1. 선행연구단계 : 체계통합기술과의 호환성 검증
2. 탐색개발단계 : 연동 인터페이스 검증
3. 체계개발단계 : 작전운용성능환경에서 연동 인터페이스의 상호운용성을 검증
4. 운용유지단계 : 임무수행을 위한 연동 인터페이스의 운용성 검증

현재 미군은 NR-KPP(Net Ready - Key Performance Parameter)와 같이 측정 가능하고, 시험 가능한 체계의 연동성 및 정보교환 요구능력을 개발 요청서에 구체적으로 기술하도록 하고 있다. 미군은 복합체계의 NR-KPP와 고도화된 상호운용성을 확보하기 위하여 NCOIC 규격을 제정하였다. NCOIC은 산업체와 정부가 사용하게 할수 있는 Net-centric에 초점을 맞춘것이며, 네트워크 중심의 상호운용성을 확보위해 SCOPE 모델, NIF 및 아키텍처 패턴모델을 운용하고 있다. SCOPE 모델은 기존 정보체계 중심의 LISI 모델과 플랫폼 중심의 DODAF ver 1.0을 통한 작전정보 교환성능 및 보장능력을 보장한다.

2.8 합동상호운용성(Joint interoperability test) 시험방법에 대한 국방표준

MND-STD-0028(합동상호운용성 시험을 위한 요구사항 및 시험방법)은 각군의 전력이 효과적으로 통합된 합동차원의 상호운용성을 보장하기 위하여 체계의 운용시험평가 단계에서 수행되는 상호운용성 시험에 대한 요구사항과 그에 대한 시험방법을 표준화한 것이다. 이 표준은 무기체계의 상호운용성 시험을 위한 표준으로 획득 단계별로 국방정보화 표준을 준수하도록 설계된 산출물을 가지는 것이 매우 중요하다. 아래는 시험대상체계와 연동체계간의 상호운용성 시험기간의 테스트 절차이다.

1. 아키텍처 산출물(6종)과 실 체계 구현이 일치하는지 확인
2. 체계데이터 교환목록 간 데이터 교환 항목/속성을 점검 (산출물 문서 확인)
3. 두 개 이상의 시험대상 소프트웨어 간 데이터를 교환하고, 교환된 데이터를 상호간에 불일치 없이 정확하게 처리하는지 시험(구현된 2개 이상의 시스템이 정보를 교환하는데 문제가 없는지 측정하는 것이다) 일반적으로 상호운용성 시험에는 제약이 따르는데, 제한사항은 ①시스템들이 시험/평가 당시 상호운용성이 보장되지 않을수 있다(연동 실패) ②시간 / 메모리의 제한된 자원으로 상호운용성이 보장되지 않음

3. 자연어 처리에 관한 연구

3.1 상호운용성 평가도구 개발을 위한 자연어 처리 기법연구

사람이 사용하는 글자를 100 퍼센트 이해하는 인공지능을 개발하기 위해서는 컴퓨터가 이해할 수 있게, 자연어를 계산 가능한 형식으로 변형해야 한다. Embedding은 자연어를 숫자의 나열인 Vector로 바꾼 결과이다. 단어와 문장을 벡터로 변환하여 『Vector space에 Embed(배치한다)』 의 과정이다. 임베딩에는 Corpus(말뭉치)의 의미, 문법 정보가 녹아있다. 임베딩은 사칙연산이 가능하기 때문에 단어 / 문서 Relevance(관련도)를 계산하여 추정한다. Transfer learning(전이학습)은 특정문제를 풀기 위하여 사전에 학습한 모델을 다른 문제를 푸는데 재사용하는 기법이다. 대규모 말뭉치를 사전에 Pertain(학습)한 임베딩을 포함한 Model을 문서분류 모델의 Input으로 사용하고, 학습된 Model을 다른 문서에 적용하여 분류를 잘 할수 있게 Fine tunning을 진행한다. 전이학습은 사람학습과 비슷한 점이 있다. 사람이 새로운 사실을 알아차리는 것은 사전에 비슷한 경험지식이 있기 때문이다.

임베딩 품질이 좋으면 단순한 모델로도 원하는 성능을 낼수 있다. 모델구조가 동일하다면 Converge(수렴) 속도가 빠르다. Embedding from learning model(ELMo), Bidirectional encoder representations from transformer(BERT), Generative pre-training(GPT) 등 최고 성능의 자연어 처리 소프트웨어는 모두 Transfer learning 이후 Fine tunning 기법을 적용하였다.

임베딩을 활용하면 컴퓨터가 자연어를 계산하는 것이 가능해진다. 임베딩은 자연어를 컴퓨터가 처리할수 있는 숫자들의 나열인 벡터로 바꾼 결과이기 때문이다. 컴퓨터는 임베딩을 계산 / 처리하여 사람이 알아들을 수 있는 형태의 자연어로 출력한다. 사람의 말을 100% 이해하는 인공지능이 등장하더라도 그 동작의 근본 방식은 계산이다. 임베딩에 자연어 의미를 함축시킬 수 있는 원리는 자연어의 통계적 패턴정보를 100% 임베딩에 넣는 것이다. 자연어의 의미는 해당 언어 사용자들이 실제 사용하는 일상 언어에서 드러나기 때문이다. 임베딩을 만드는 과정에서 사용하는 통계 정보는 2가지가 있다.

표 1. 임베딩을 만들때 쓰는 통계 정보
Table 1. Statistical information used when creating embeddings

가정	단어가 어떤 순서로 쓰였는가 ?	어떤 단어가 같이 쓰였는가 ?
대표 모델	ELMo, BERT, GPT	Word2Vec, FastText
분류 기준	단어의 등장순서	단어가 어떻게 분포 하나?

3.1.1 예측기반의 임베딩 : Word2Vec, FastText

단어 유사도 평가는 사람이 말하는 언어와 언어모델의 코사인 유사도 값이 얼마나 일치하는가의 상관관계를 계산하여 임베딩의 품질을 계산하는 방식이다. 『Word2Vec, FastText』은 단어 유사도 평가에서 높은 점수를 받고있다. 그리고 조사나 어미가 발달한 한국어에 좋은 성능을 낼수 있는 FastText을 분석하였다. FastText는 단어를 n-gram으로 표현하기 때문에 한글과 궁합이 잘 맞는 편이다. 한글은 자소단위로 분해할수 있고, 이 자소 각각을 하나의 문자로 보고 임베딩을 실시한다. 또한 말뭉치에 미등록된 단어에 대해서도 임베딩 값을 도출하기 때문에 상당한 경쟁력이 있다.

3.1.2 행렬분해 방법 : Swivel

단어 유추평가는 『아니땐 굴뚝에』 ⇒ 『언어모델을 사용한 임베딩 작업』⇒ 『연기 나랴 ?』와 같이 평가하는 방법이다. 임베딩 되어진 단어벡터 사이의 유사도 측정이 수월하고, 말뭉치 전체의 통계 정보가 잘 반영된 Swivel 모델의 장점이다.

3.2 상호운용성 평가에 적용하기 위한 자연어 처리 기법의 실험 계획안

1차 단계로서 『국방전력발전업무훈령(국방부훈령 / 제2749호)』을 말뭉치로서 사용 ⇒ 임베딩 과정을 거침 ⇒ ① Word2Vec 모델 ②FastText 모델 ③Swivel 모델의 코사인 유사도를 계산 한다.

위의 과정을 거쳐 국방상호운용성 평가에 사용되는 문서도 자연어 임베딩을 거쳐 언어의 유사도 평가가 가능하다는 것을 입증하고자 한다. 국방전력발전업무훈령을 사용한 이유는 『소

요평가 / 수준측정』에 사용되는 『연동합의서 / 연동통제문서 / 상호운용성 확보계획서)』 4종류의 문서는 군사기밀이라 유출이 불가한데, 전력발전업무훈령은 인터넷에 공개된 문서이기 때문이다.

3.3 단어 - 문서 행렬을 통한 단어간 유사도 계산

아래와 같이 3가지의 문서가 있다고 가정한다.

표 2. 단어간 유사도-1
Table 2. Similarity between words-1

구분	문장
문서_1	나는 아침보다 저녁 좋다
문서_2	사과는 아침보다 저녁 좋다
문서_3	사과는 아침보다 저녁 간식으로 좋다

띄어쓰기를 기준으로 토큰화를 진행하여서, 문서-단어 행렬(Document-Term Matrix, DTM)으로 변경하면 아래 표이다.

표 3. 단어간 유사도-2
Table 3. Similarity between words-2

	나는	아침 보다	저녁	좋다	사과 는	점심	간식 으로
문서_1	1	1	1	1	0	0	0
문서_2	0	1	1	1	1	0	0
문서_3	0	1	1	1	1	1	1

문서_2과 문서_3 다른 문서보다 상대적으로 많이 겹치고 있다. 『단어를 많이 공유하기 때문에 유사할수 있다』라고 추정할 수 있다. 문서내 단어의 연관성 유추과정을 그림으로 표현하면 다음과 같다.

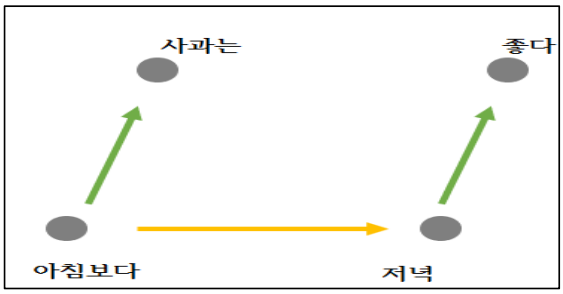

그림 4. 문서내 단어의 연관성 유추과정
Fig. 4. Process of inferring the relevance of words in a document

3.3 코사인 유사도(Cosine Similarity) 계산

문장/단어의 유사도를 구하는 일은 자연어 처리의 주요한 주제 중 하나이다. 문장의 유사도는 주로 문장들 간에 동일한 단어 또는 비슷한 단어가 얼마나 공통적으로 많이 사용되었는지를 계산하여 평가한다. 기계가 계산하는 문장의 유사도의 성능은 각 문장의 단어들을 어떤 방법으로 수치화하여 표현했는지(DTM, Word2Vec 등), 문장 간의 단어들의 차이를 계산 했는지에 달려있다. 코사인 유사도를 이용하여 문장의 유사도를 구하는 게 가능하다. 코사인 유사도는 두 벡터 간의 코사인 각도를 이용하여 구할 수 있는 두 벡터의 유사도를 의미한다. 두 벡터의 방향이 동일한 경우에는 1의 값을 가지며, 90°의 각을 이루면 0, 180°로 반대의 방향을 가지면 -1의 값을 갖게 된다. 즉, 결국 코사인 유사도는 -1이상 1이하의 값을 가지며 값이 1에 가까울수록 유사도가 높다고 판단할 수 있다. 이를 직관적으로 이해하면 두 벡터가 가리키는 방향이 얼마나 유사한가를 의미한다.

3.4 말뭉치

자연어 임베딩을 시행할때 모델이 사용한 가능한 형태로 변환하는 과정을 전처리라고 한다. Corpus라고 명칭하며, 말뭉치에 오류가 없을수록 자연어 처리 모델의 정확도가 높아진다. 자연어 처리에서 전처리 과정은 말뭉치 수집 ▶ 정제 ▶ 문장 단위분절 ▶ 분절 ▶ 병렬 말뭉치 정렬 ▶ Subword 분절의 과정으로 이루어진다.

3.4.1 말뭉치 수집

공개된 데이터를 사용하거나, 인터넷을 Crawling 하여서 수집한다. 본 논문에서는 「국방전력발전업무훈령」을 사용한다.

3.4.2 정제

텍스트를 사용하기에 앞서, 언어모델이 사용할 수 있게끔 필요한 형태를 얻어내는 과정이다. ① 전각문자 제거 : 한국어 문서의 숫자, 영어, 기호 등이 전각문자인 경우 일반적으로 사용되는 반각문자로 바꾸어 주어야 한다. 전각문자는 하나의 글자가 정사각형을 이루는 문자를 의미하고, 반각문자는 전각문자의 가로 폭을 반으로 줄인 문자이다. ② 대소문자 통일 : 영어 단어를 대문자 또는 소문자로 통일하여 표현한다. ③ 정규표현식 : Crawling 하여서 만든 말뭉치에는 기호에 의해 노이즈가 발생한다. 이를 제거하고 자연어 임베딩에 사용하여야 한다.

3.5 문장단위 분절

정제를 거친 말뭉치는 많은 경우 문장들이 연속해서 등장하기 때문에, 한 라인당 한 문장만 존재하도록 문장단위로 분절을 해야한다. 마침표만 사용하여 분절하면, 문장단위 이외의 곳에서 분절이 될수 있으므로, 직접 분절 알고리즘을 만들기보다는 많이 활용되고 있는 Toolkit인 NLTK(3.2.5 버전)을 추천한다.

3.5.1 분절

한국어의 경우, Mecab 또는 KoNLPy를 이용하여 분절이

가능하다. 먼저 Mecab은 한국어 분절에 가장 많이 사용되는 프로그램이며, 속도는 빠르나 설치가 어렵다. KoNLPy는 여러 한국어 형태소 분석기를 모아놓은 Wrapping library를 제공하는데, 설치와 사용이 쉬우나 속도가 느리다.

3.6 병렬 말뭉치 정렬

인터넷을 Crawling 해서 얻은 문서들은 문서와 문서사이의 Mapping일뿐 문장대 문장에 대한 정렬이 없는 경우가 다반사이다. 각 문장에 대해서 정렬을 해주어야 한다.

3.7 Subword 분절

단어는 의미를 가진 더 작은 Subword들의 조합으로 이루어진다는 가정하에 적용되는 알고리즘이다. 작업을 해서 얻어지는 효과는 ①희소성 감소 ②UNK(unkown) 토근에 대한 효율적인 대처가 가능하다(임베딩의 품질이 향상 / UNK가 등장하면 코사인유사도 계산값이 작아진다).

3.8 형태소 분석기

한국어는 영어처럼 띄어쓰기만으로 단어를 분리하면 제대로 되질 않는다. 한국어는 어미와 조사 등이 발달되어 있기 때문이다. 그래서 한국어는 형태소에 따라 단어 분리를 하게 된다. 자주 쓰이는 파이썬 한국어 형태소 분석 패키지로는 바로 KoNLPy가 있다. 이 안에는 여러 종류의 한국어 형태소분석기(Okt, Komoran, Kkma, 은전한닢(Mecab))가 있다. 예를 들어 실행하면 txt = '전력발전업무훈령의상호운용성평가를기준으로한다' ①Okt(Open Korean text): [‘전력/Noun’, ‘발전/Noun’, ‘업무/Noun’, ‘훈령/Noun’, ‘의/Josa’, ‘상호/Noun’, ‘운용/Noun’, ‘성/Noun’, ‘평가/Noun’, ‘를/Josa’, ‘기준/Noun’, ‘으로/Josa’, ‘한다/Josa’]와 같은 결과가 출력.

4. 자연어 처리 기법연구

4.1 분포가정 원리를 적용한 Word2Vec

Distribution hypothesis(분포 가정)란 특정 Window내에 동시에 등장하는 단어의 집합이다. 단어의 분포는 그 단어가 문장 내부에 어떤 위치에 등장하는지, 근처에 어떤 단어가 자주 나타나는지에 따라서 달라진다. 어떤 단어의 쌍이 문맥환경에서 자주 등장한다면 그 의미 또한 「유사할 것」이라고 유추한 것이 분포가정이다. word2vec는 인간 언어의 표현하는 방법을 어떻게 학습하는 것일까? word2vec의 핵심적인 아이디어는 이것이다. 단어의 주변을 보면 그 단어를 안다. You shall know a word by the company it keeps. - 언어학자 J.R. Firth. 단어의 주위만 보았는데도 어떤 단어가 적합하고 어떤 단어가 부적합한지가 어느 정도 드러난다. 빈칸에 들어갈 수 있는 단어들은 서로 비슷한 맥락을 갖는 단어들, 즉 서로 비슷한 단어들이다. 단어의 주변을 보면 그 단어를 알 수 있기 때문에, 단어의 주변이 비슷하면 비슷한 단어라는 말이 된다.

word2vec 안에도 두가지 방식이 있다. 하나는 맥락으로 단어를 예측하는 CBOW(continuous bag of words) 모델이다. 또다른 하나는 단어로 맥락을 예측하는 skip-gram 모델이다. 그 중에서 여기서는 CBOW 하나만 살펴보자. CBOW 모델을 반대로 뒤집으면 skip-gram 모델이 되므로, 하나만 이해하면 다른 하나는 쉽다. CBOW 모델은 주변 단어, 다른 말로 맥락(context)으로 타겟 단어(target word)를 예측하는 문제를 푼다. 주변 단어란 보통 타겟 단어의 직전 몇 단어와 직후 몇 단어를 뜻한다. 타겟 단어의 앞 뒤에 있는 단어들을 타겟 단어의 친구들이라고 보는 것이다. 이 주변 단어의 범위를 window라고 부른다.

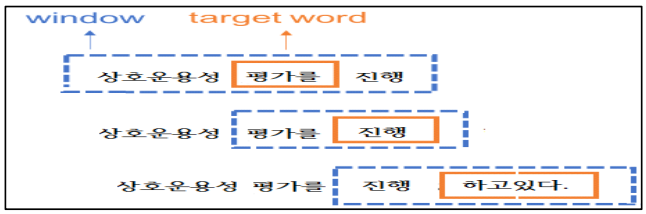

그림 5. 타겟 단어(target word)를 예측
Fig. 5. Target word prediction

이 원리를 적용한 대표적 모델이 Word2Vec이다. Mikolov et al. (2013a)에서 제안한 Skip-gram 모델의 구조는 다음과 같다. Skip-gram 모델은 타깃단어를 가지고 주변 문맥단어가 무엇인가 ? 예측하는 과정에서 학습된다. Skip-gram의 학습 데이터는 [타깃단어, 타깃단어 다음단어], [타깃단어, 타깃단어 다음 두 번째 단어] 이렇게 4개의 쌍이 된다. 전체 말뭉치를 단어별로 슬라이딩해 가면서 학습 데이터를 만든다. 정답 문맥단어가 나타날 확률을 높이고 나머지 단어들은 그에 맞게 낮추어야한다. 이럴 경우 계산량이 너무 많아져서 Mikolov et al. (2013b)에서 제안한 Skip-gram 모델은 타깃단어, 문맥단어 쌍이 주어졌을 때 해당 쌍이 포지티브 샘플(+) 인지, 네거티브 샘플인지(-) Binary classification(이진분류)하는 과정에서 학습된다.

◉ 이 드라마 재미있다 + Positive(긍정)
◉ 이 드라마 재미없다 + Negative(부정)

이 샘플링 방식으로 학습하게 되면 1개의 포지티브 샘플과 k개의 네거티브 샘플만 계산하게 되어, 계산량이 많이 줄게 된다. 네거티브 샘플을 선정하는 방법은 Mikolov et al. (2013b)에서 말뭉치에서 자주 등장하지 않는 단어가 네거티브 샘플로 선정되게 되도록 수식을 만들었다.

$$P_{negative}(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=0}^{n} f(w_j)^{3/4}}$$

수식 1. 단어가 네거티브 샘플로 선정되는 확률

4.2 Word2Vec의 모델학습 방식

Skip-gram 모델은 타깃단어와 문맥단어 쌍이 주어졌을 때 해당 쌍이 포지티브 샘플인지(+) 인지, 네거티브 샘플인지(-)

예측하는 과정에서 학습된다. 타깃단어와 문맥단어 쌍이 실제 포지티브 샘플이라면 아래의 조건부 확률을 최대화해야 한다.

$$P(+|t, c) = \frac{1}{1+\exp(-u_t v_c)}$$

수식 2. t, c가 포지티브 샘플일 확률(t 주변에 c가 존재

Skip-gram 모델의 학습 파라미터는 U와 V 행렬 2개뿐이다. 크기는 어휘집합 크기 X, 임베딩 차원수도 동일하다. U는 타깃단어, V는 문맥단어에 대응한다.

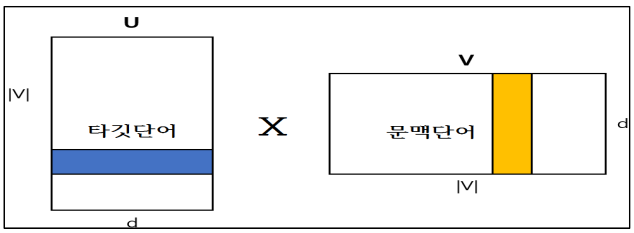

그림 6. Skip-gram 모델의 자연어 학습 파라미터
Fig. 6. Natural language learning parameters of Skip-gram model

수식 2. 의 확률값을 최대화 하려면 포지티브 샘플에 해당하는 단어벡터인 U와 V의 내적값을 키워야한다(두 벡터의 내적은 코사인 유사도에 비례) 즉 벡터 공간상으로는 가깝다는 의미이다(내적값이 작아지면 벡터공간 거리가 멀어지게 된다). Skip-gram 모델이 Likelihood function을 최대화 하여야 한다. θ를 한번 업데이트 할 때 1개 쌍의 포지티브샘플과 k개 쌍의 네거티브 샘플이 학습된다는 것이다. 아래의 수식을 최대화하는 과정에서 말뭉치의 분포정보를 단어 임베딩에 입력 시키게 된다.

$$L(\theta) = \log P(+|t_p, c_p) + \sum_{i=1}^{k} \log P(-|t_{n_i}, c_{n_i})$$

수식 3. Skip-gram 모델의 Likelihood function

4.3 FastText의 모델학습 방식

페이스북 연구팀 2017년 공개한 FastText는 진화된 임베딩 기법이다. 『A. Bojanowski P .. Grave . E (2017) “Enriching word vectors with subword information. Transactions of the association for computational linguistics”』 FastText 임베딩기법은 단어를 n-gram으로 표현하고, 네거티브 샘플기법을 사용한다. 아래의 수식의 조건부 확률을 최대화 하는 과정에서 학습된다.
입력단어 쌍이 실제로 포지티브 샘플이라면 모델은 해당 쌍이 포지티브라고 맞추어야 한다. FastText는 타깃단어, 문맥단어 쌍을 학습할 때 타깃단어에 속한 문자 단위 n-gram 벡터들을 모두 업데이트 한다.

$$P(+|t, c) = \frac{1}{1+\exp(-u_t v_c)} = \frac{1}{1+\exp(-\sum_{g\in G_t} z_g^T v_c)}$$

수식 4. t, c가 포지티브 샘플일 확률(t 주변에 c가 존재)

포지티브 샘플이 주어졌을 때 위의 수식을 최대화 하려면 분모를 최소화해야 한다. 즉 코사인 유사도를 높여야(벡터 공간상 거리가 가깝게)한다. FastText 모델은 네거티브 샘플 단어 쌍에 대해서 아래의 수식의 조건부 확률을 최대화 하여야 한다. 네거티브 샘플을 입력하면 네거티브 샘플이라고 판정해야 한다는 것이다.

$$L(\theta) = \log P(+|t_p, c_p) + \sum_{i=1}^{k} \log P(-|t_{n_i}, c_{n_i})$$

수식 5. FastText 모델의 Likelihood function

위 공식의 의미는 한번 업데이트 할 때 마다 1개의 포지티브 샘플과 k개의 네거티브 샘플이 학습된다는 의미이다. 네거티브 샘플이 주어졌을 때 위 수식의 확률값을 최대화 하려면 단어 벡터의 내적값이 작아야한다(벡터 공간상 거리가 멀게). FastText 모델의 장점은 조사나 어미가 발달한 한국어에 좋은 성능을 낼수있다. 문자 단위 n-gram을 쓰기 때문에 한글을 자소단위로 분해할수 있고, 이 자소각각을 문자로 보고 임베딩을 실행하기 때문이다.

4.4 Swivel의 모델학습 방식

Swivel(Submatrix vector embedding learner)은 구글 연구팀(Shazeer et al.. 2016)이 발표한 행렬분해 기반의 단어 임베딩 기법이다. PMI 행렬을 U와 V 행렬으로 분해하고, 학습이 종료되면 U를 단어 임베딩으로 사용할 수 있다. Swivel은 PMI 행렬을 분해한다는 점에서 단어-문맥 행렬을 분해하는 점에서 학습 효율이 높다. Swivel의 목적함수는 PMI 행렬의 단점을 극복할 수 있도록 설계되었다는 점에서, 타깃단어와 문맥단어가 사용자가 정한 윈도우 내에서, 단 한건이라도 동시에 등장한 적이 있는 경우의 적용에 매우 유리하다.

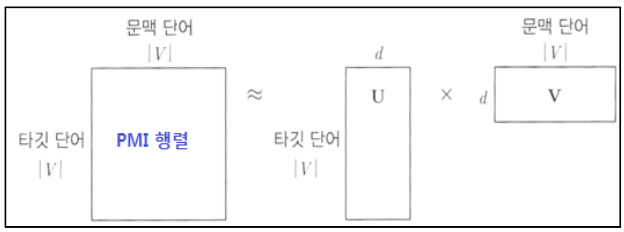

그림 7. 행렬분해 기반의 단어 임베딩 기법
Fig. 7. Matrix decomposition-based word embedding technique

4.4.1 말뭉치에 동시 등장한 케이스가 있는 경우
타깃단어 해당하는 I 벡터와 문맥단어에 해당하는 J 벡터의 내적이 두 단어의 PMI 값과 일치하도록 두 벡터를 순차적으로 업데이트한다. 함수 f(x)의 의미는 타깃단어와 문맥단어의 동시 등장 빈도를 의미한다. 함수 값이 클수록 I 벡터 / J 벡터의

내적이 두 단어의 PMI값과 좀더 비슷해야 학습 손실이 줄어든다. 부연 설명하면 타깃과 문맥단어가 동시에 자주 등장 할수록 두 단어에 해당하는 벡터의 내적이 PMI값과 일치하도록 더욱 강제화 한다는 의미이다. 타깃단어와 문맥단어가 말뭉치의 특정 윈도우 내에서 동시에 등장한 적이 한번도 없는 경우에는 PMI 값이 음의 무한대로 발산하기 때문에 Shazeer et al.. 2016은 이 같은 케이스에 대해 목적함수를 별도로 설정했다. 동시 등장 횟수를 0대신 1로 가정하고 계산한 PMI 값이다.

$$J = \frac{1}{2} f(x_{ij})(U_i \cdot V_j - PMI(i,j))^2$$

수식 6. 목적함수(한건이라도 동시에 등장한 적이 있는 경우)

4.4.2 말뭉치에 동시 등장한 케이스가 없는 경우

Shazeer et al.. 2016은 두 개의 단어(I / J)가 각각 고 빈도 단어인데 두 단어의 동시 등장 빈도가 0이라면, 주 단어는 정말로 같이 등장하지 않는, 의미상 무관계한 단어일 것이라고 가정했다. 이런 경우, 두 단어의 벡터 내적 값이 PMI 값 보다 작게 되도록 학습한다. 반대로 저 빈도 단어 인데 두개 단어의 동시 등장 빈도가 0라면 두 단어는 의미상 관계가 일부 있을 수 있다고 보았다.

우리가 가지고 있는 말뭉치 크기가 작아서, 우연히 해당 쌍의 동시 등장 빈도가 전혀 없는 걸로 나타났을 수도 있는 것이다. 『상호운용』이라는 단어와 『무기체계』라는 단어는 인터넷에서 검색되는 문서에는 흔하지는 않지만, 국방관련 문서(예 : 소요 제기서 / 수준측정 문서)에는 동시에 자주 등장하는 편이다. 제작한 말뭉치에서 두 개의 단어가 동시에 등장하는 빈도가 0 이라면, 이런 경우에는 두 단어에 해당한 벡터의 내적 값이 PMI 값보다 약간 크게 학습된다. Swivel 자연어 학습 엠베딩은 U 행렬과 V 행렬을 램덤방식으로 초기화한 뒤 목적함수를 최소화 하는 방향으로 행렬 값들을 조금씩 업데이트 하는 방식으로 학습한다.

$$J = \log[1 + exp(U_i \cdot V_j - PMI^*(i,j))]$$

수식 7. 목적함수(한건이라도 동시에 등장한 적이 없는 경우)

5. 자연어 처리 실험

5.1 처리 절차

5.1.1 국방전력발전업무훈령(국방부훈령 / 제2749호) 다운로드한다.

5.1.2 국방전력발전업무훈령을 파이썬으로 전처리 한다.

5.1.3 전력발전업무훈령을 파이썬으로 토큰화 처리한다.

5.1.4 형태소분석 완료된 데이터 다운로드

5.1.5 Word2Vec skip-gram 모델로 학습을 진행한다. (파이썬 사용).

```
-----------------------------------------
corpus_fname =
"/notebooks/embedding/data/tokenized/corpus_mecab.t
xt"
model_fname = "/notebooks/embedding/data/
word-embeddings/word2vec/word2vec"

from gensim.models import Word2Vec
corpus = [sent.strip().split(" ") for sent in open
(corpus_fname, 'r').readlines()]
model = Word2Vec(corpus, size=100, workers=4, sg=1)
model.save(model_fname)
-----------------------------------------
```

5.2 실험 결과(분석)

5.2.1 Word2Vec skip-gram 모델로 학습한 결과의 코사인 유사도 상위분석

```
-----------------------------------------------
from model.word_eval import WordEmbeddingEvaluator
model = WordEmbeddingEvaluator(
"/notebooks/embedding/data/word-embeddings/word2ve
c/word2vec",
method="word2vec", dim=100,
tokenizer_name="mecab")
model.most_similar("상호", topn=4)
--------------------------------------
```

5.2.2 FastText N-gram 모델로 학습을 진행한다.

```
--------------------------------------
mkdir -p data/word-embeddings/fasttext
models/fastTExt/fasttext skipgram -input data
/tokenized/corpus_mecab.txt -output data
/word-embedding/fasttext/fasttext
--------------------------------------
```

5.2.3 FastText N-gram 모델로 학습한 결과의 코사인 유사도 상위분석

```
-----------------------------------------------
from model.word_eval
import WordEmbeddingEvaluator
model = WordEmbeddingEvaluator(
vecs_txt_fname="data/word-embeddings/fasttext/fasttext
.vec",
vecs_bin_fname="data/word-embeddings/fasttext/fasttex
t.bin",
method="fasttext", dim=100,
tokenizer_name="mecab")
model.most_similar("상호", topn=4)
-----------------------------------------------
```

5.2.4 Swivel-PMI 행렬분해 모델로 학습을 진행한다

```
------------------------------------------------
cd /notebooks/embedding
mkdir -p data/word-embeddings/swivel
models/swivel/fastprerp --input
data/tokenized/corpus_mecab.txt--output_dir
data/word-embeddings/swivel/swivel.data
python models/swivel/swivel.py --input_base_path
data/word-embeddings/swivel/swivel.data--output
_base_path/word-embeddings/swivel -dim 100
-------------------------------
```

5.2.5 Swivel-PMI 행렬분해 모델로 학습한 결과의 코사인 유사도 상위분석

```
------------------------------------------------
from models.word_eval
import WordEmbeddingEvaluator
model = WordEmbeddingEvaluator
("data/word-embeddings/swivel/row_embedding.tsv",
method="swivel", dim=100, tokenizer_name="mecab")
model.most_similar("상호", topn=4)
------------------------------------------------
```

5.3 자연어 처리 실험 결과(코사인 유사도 상위분석)

5.3.1 코사인 유사도 상위분석(Word2Vec)

표 4. Word2Vec의 실험결과

Table 4. Experimental results of Word2Vec

Word2Vec의 자연어 처리결과			
검색단어	상호운용성	방사청	운영개념
1	합동성	각군	단위전력
2	시험평가	기관	무기체계
3	기술센터	소요제기기관	정립
4	위원회	합참	함참

5.3.2 코사인 유사도 상위분석(FastText)

표 5. FastText의 실험결과

Table 5. Experimental results of FastText

FastText의 자연어 처리결과			
검색단어	상호운용성	방사청	운영개념
1	합동성	각군	단위전력
2	기술센터	기관	무기체계
3	관리	청장	과
4	위원회	소요제기	함참

5.3.3 코사인 유사도 상위분석(Swivel)

표 6. Swivel의 실험결과

Table 6. Experimental results of Swivel

Swivel의 자연어 처리결과			
검색단어	상호운용성	방사청	운영개념
1	합동성	각군	단위전력
2	시험평가	기관	무기체계
3	기술센터	소요제기	정립
4	위원회	청장	함참

6. 결론 및 향후 과제

6.1 연구결과

본 논문에서는 분포가정의 원리를 적용한 『Word2Vec, FastText, Swivel』 모델을 적용하여 단어적 임베딩 기법을 분석 및 실험하였다. 분포가정은 문장에서 『어떤 단어가 같이 쓰였는지』 와 『어떤 단어 쌍이 얼마나 자주 나타나는지』의 정보인 단어시퀀스에 확률을 부여하는 모델이다. 실험 과정은 다음과 같다. 『국방전력발전업무훈령(국방부훈령 / 제2749호)』을 말뭉치로서 사용 ⇒ 임베딩 과정을 거침 ⇒ ①Word2Vec 모델 ②FastText 모델 ③Swivel 모델의 코사인 유사도를 계산한다.

코사인 유사도 상위분석(Word2Vec / FastText / Swivel 모델)의 결과를 보면 사람이 인지하는 것과 유사하다. 1,2,3,4는 계산값(확률값)이 높은 순서이다(1이 가장 큰 확률값). 이러한 자연어 처리 모델은 국방관련 문서에서 일반인이 이해하기 생소한 단어들인, 예를 들면 “합동성”, “상호운용성”, “운영개념”이다. 그러나 이러한 단어는 군내부 문서에서 많이 사용된다. 이러한 단어의 코사인 유사도 계산값이 사람이 인지하는 수준과 많이 유사하여서 자연어 처리 모델이 소요평가 / 수준측정을 진행 하여도, 필요한 단어 집합과 연결이 되어서 향후 군에 고도화된 AI 기반 평가 모델이 적용이 가능하다는 것을 입증 하였다.

이러한 결과는 자연어 처리 모델을 현업에 적용하여도 군인/군무원이 작성한 문서를 컴퓨터가 충분히 이해한다고, 결론을 내려도 무방하다. 이번의 실험을 통하여 군에 자연어 처리 기법을 적용한 AI 기반 평가 모델의 개발 예산책정(중기계획에 개발예산 반영)을 위한 기본 데이터로 사용가능하다.

6.2 향후연구에서 진행할 방향

지난 20년간 합동상호운용성기술센터에서 실시한 무기체계와 전력지원체계의 상호운용성을 확보하기 위하여 사람이 평가한 소요평가 / 수준측정의 문서 산출물인 『연동합의서 / 연동통제문서 / 상호운용성 확보계획서』를 Word2Vec, FastText, Swivel의 언어모델로 임베딩 작업을 실시한후 ①단어 유사도 평가 ②단어 유추평가를 실시한다. 그리고 그 데이터의 파인튜닝 작업을 실시하여 사람에 의존하지 않고 언어모델에 의한 상호운용성 평가의 목표를 달성하기 위하여 노력할 예정입니다.

References

[1] Mc-Cormick C. “Word2Vec tutorial - The Skip-Gram model”, pp. 33-90, 2016.
[2] Jurafasky. D & Martin . J “Speech and language processing”, pp. 73-85, 2019.
[3] Mikolov et al. “Efficient estimation of word representations in vector space“, pp. 10-50, 2013.
[4] Mikolov et al. “Distributed representation of words and phrase and their compositionality“, pp. 20-50, 2013.
[5] Bojanowski P .. Grave . E (2017) “Enriching word vectors with subword information. Transactions of the association for computational linguistics”, pp. 20-23, 2017.
[6] Shazeer, N.. Doherty, R Evans . C & Waterson, (2016) Swivel : Improving embedding by noticing what's missing, pp. 21-29, 2016.
[7] Cho. K. Natural language understanding with distributed representation, pp. 10-25. 2015.
[8] Socher. R "Deep learning for natural language processing", pp. 10-70, 2016.
[9] 조경현 (2018) "딥러닝을 이용한 자연어 처리" 박문각, pp, 10-150, 2018.
[10] 김현중 (2019) “한국어 자연어 처리를 위한 파이썬 라이브러리”,: 랜덤하우스코리아, pp. 10-60, 2019.
[11] 박은정 (2014) “간결한 한국어 정보처리 파이썬 패키지”, 자유아카데미 , pp. 10-35, 2014.
[12] 잘라지 트하니키(이승준 옮김 / 2018) “파이썬 자연어 처리의 이론”, 선진문화사, pp. 20-365, 2018.