Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.2, Apr. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.2.305

딥러닝을 활용한 개인정보 처리방침 분석 기법 연구

조 용 현,[1†] 차 영 균[2‡]
[1,2]고려대학교 (대학원생, 교수)

Privacy Policy Analysis Techniques Using Deep Learning

Yong-Hyun Jo,[1†] Young-Kyun Cha[2‡]
[1,2]Korea University (Graduate student, Professor)

요 약

개인정보보호법에서는 정보 주체의 권리보장을 위해 개인정보보호 정책문서인 개인정보 처리방침을 공개하도록 규정하고 있고 공정거래위원회에서는 개인정보 처리방침을 약관으로 보고 약관규제법에 따라 불공정약관심사를 하고 있다. 그러나, 정보 주체는 개인정보 처리방침이 복잡하고 이해하기 어려워 읽지 않는 경향이 있다. 개인정보 처리방침의 내용을 간단하고 읽기 쉽게 한다면 온라인 거래에 참여할 확률이 증가하여 기업의 매출 증가에 기여하고, 사업자와 정보주체간의 정보 비대칭성 문제 해결에 기여할 것이다. 본 연구에서는 복잡한 개인정보 처리방침을 딥러닝을 이용하여 분석하여 정보주체로 하여금 가독성 높은 단순화된 개인정보처리 방침을 구현하기 위한 모델을 제시한다. 모델을 제시하기 위해 국내 258개 기업의 개인정보 처리방침을 데이터셋으로 구축하고 딥러닝 기술을 활용하여 분석하는 방안을 제안하였다.

ABSTRACT

The Privacy Act stipulates that the privacy policy document, which is a privacy statement, should be disclosed in order to guarantee the rights of the information subjects, and the Fair Trade Commission considers the privacy policy as a condition and conducts an unfair review of the terms and conditions under the Terms and Conditions Control Act. However, the information subjects tend not to read personal information because it is complicated and difficult to understand. Simple and legible information processing policies will increase the probability of participating in online transactions, contributing to the increase in corporate sales and resolving the problem of information asymmetry between operators and information entities. In this study, complex personal information processing policies are analyzed using deep learning, and models are presented for acquiring simplified personal information processing policies that are highly readable by the information subjects. To present the model, the personal information processing policies of 258 domestic companies were established as data sets and analyzed using deep learning technology.

Keywords: Privacy Policy, Text Mining, Decision Tree Model, Privacy Protect, Subject Right

Ⅰ. 서 론

ICT 기반으로 사회가 급격히 발전하면서 개인정보를 이용한 서비스가 증가하고 개인정보 수집과 이용, 제공이 활발하게 이뤄지고 있다. 개인정보보호법과 E.U GDPR(General Data Protection Regulation)에서는 이러한 사회적 변화 환경에서 정보 주체의 권리보장과 자기 결정권 보호를 위해 개인정보 처리과정의 명시적 동의 절차를 마련하고 있고, 사용자는 동의 과정 이전에 서비스 제공자가 공개한 개인정보 처리방침을 확인하고 있다. 그러나,

Received(02. 21. 2020), Modified(1st: 03. 20. 2020, 2nd: 04. 06. 2020), Accepted(04. 06. 2020)
† 주저자, yhjo13@korea.ac.kr
‡ 교신저자, ykcha@korea.ac.kr(Corresponding author)

수많은 서비스를 이용하고 있는 현대인이 본인이 열람, 동의한 개인정보 처리방침을 모두 인지하고 정보주체의 권리를 행사하기 위한 활동을 하기 어렵다.

본 연구에서는 국내 주요 기업과 개인정보보호 관리체계 인증기업들이 공개하고 있는 개인정보 처리방침 수집하여 분석할 수 있도록 벡터화하여 학습하고, 학습된 모델을 이용하여 개인정보보호 정책을 분류하고 체계화 할 수 있는 방안을 제시한다.

II. 관련 연구

본 장에서는 한글로 표현된 개인정보 처리방침에 관한 연구와 기계학습, 지도학습 알고리즘의 특징 분석, 인공지능을 이용한 유사한 사례, 보안 정책을 딥러닝을 이용하여 활용한 사례 등을 정리하였다.

2.1 자연어 처리와 임베딩(Embedding) 기법

문서에 포함된 단어의 빈도수나 TF-IDF 등을 이용한 통계적인 방법 또는 SVM(Support Vector Machine), Naive Bayesian과 같은 지도/비지도 학습을 이용하여 문서를 요약하는 방법을 사용하여 왔다. 자연어를 처리하여 요약하기 위해서는 임베딩 과정을 통해 단어나 문장을 수치화하고 벡터로 변환하여야 한다. 단어를 벡터로 변환하는 대표적인 방법으로는 Word2Vec, 문장을 벡터로 변환하는 Doc2Vec이 사용되고 있다.

Kim Yoon[1]은 단어를 벡터로 표현하는 벡터(Vector)화하는 방법으로 Word2Vec과 CNN(Convolutional Neural Network)을 이용한 문장 분류 모델을 Fig. 1과 같이 설명하였다.

Word2Vec 방법으로 CBOW(Continuous Bag of Words)는 주변의 단어들을 통해서 중간에 있는 단어들을 예측하는 방법이고, Skip-Gram은 주어진 단어로 근처에 올 단어들을 예측하는 방법으로 그 차이는 Fig. 2와 같이 설명될 수 있다.

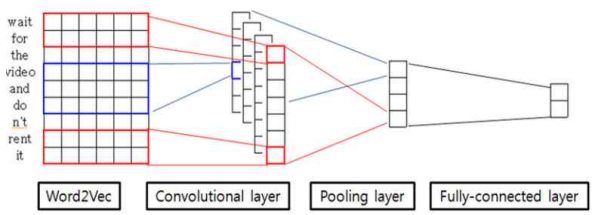

Fig. 1. CNN model on top of word2vec

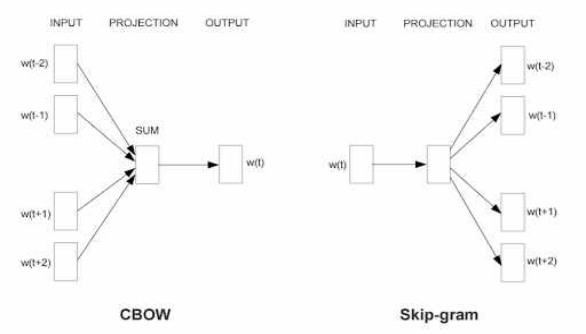

Fig. 2. CBOW and Skip-gram model

허지욱[2]은 워드 임베딩을 이용하여 질의 기반 한국어 문서요약 기법으로 Word2vec과 FastText를 이용하여 문서를 요약하였는데, 한글의 특성을 고려하지 않는 Word2vec 보다는 FastText를 이용한 n-gram의 문자 단위의 워드 임베딩이 더 우수하다고 설명하였다. FastText는 2016년 facebook에서 제안한 단어 임베딩 방법론으로 Word2Vec의 Skip-gram 모델과 거의 유사하나 단어를 n-gram Bag-of words(nBoW)로 나타낸다.

김도우, 구명완[3]은 Word2Vec에서 활용한 CNN모델은 입력으로 고정 길이를 요구하기 때문에 데이터셋 내에서 CNN의 입력 길이로 맞추기 위해서 Padding을 추가해야 하므로 자원 사용량을 낭비하고 훈련 시간이 증가하는 비효율이 발생한다고 설명하였다.

따라서 본 연구에서는 Word2Vec과 유사하지만, 부분단어의 벡터들로 표현 가능하고 한글의 특성을 고려하여 워드임베딩이 우수한 FastText를 이용하였다.

2.2 개인정보 처리방침

개인정보 처리방침을 정의하는 규정은 별도로 존재하지 않으나 개인정보보호법과 정보통신망법에 의거하여 정보 주체 및 정보통신 서비스 이용자에게 개인정보 자기결정권을 보장하기 위해 개인정보처리자 또는 사업자가 공개하는 문서이다. 행정안전부에서는 “개인정보를 처리하는 자가, 자신의 개인정보 처리 방식이나 내부 정책 등에 관하여 만든 일종의 설명 또는 진술”로 설명되고 있다[4].

이정운[5]은 미국 공정거래위원회(FTC)는 FTC Act. 15 U.S.C. §45에 따라 개인정보 처리방침을

위반한 경우 행정제재를 하고 있으며, 한국은 개인정보처리자가 개인정보 처리방침에 반하여 개인정보를 처리하고, 손해 발생시에는 민법 제750조에 의거하여 손해배상을 청구할 수 있고, 허위의 고지가 이루어졌다면 대법원 1993.8.13. 선고92다 52665 판례를 참조하여 위법성이 인정된다고 하였다. 또한, 개인정보 처리방침이 거짓 또는 기만 표시라고 주장하는 경우 표시·광고의 공정화에 관한 법률 제10조 제2항에 따라 무과실책임을 주장할 수 있다고 하였다.

정보보호 및 개인정보보호 관리체계 인증(ISMS-P, Personal Information & Information Security Management System) 대상 사업자는 인증기준(3.5.1.개인정보 처리방침 공개)에 따라 다음 주요 사항을 확인하여야 한다.

- 개인정보 처리방침을 정보 주체가 쉽게 확인할 수 있도록 인터넷 홈페이지등에 지속적으로 현행화하여 공개하고 있는가?
- 개인정보 처리방침에는 법령에서 요구하는 내용을 모두 포함하고 있는가?

또한, 개인정보보호법 제3조(개인정보 보호 원칙) ⑤항에서는 "개인정보처리자는 개인정보 처리방침 등 개인정보의 처리에 관한 사항을 공개하여야 하며, 열람 청구권 등 정보 주체의 권리를 보장하여야 한다"고 규정하고 있다. 동법 제30조(개인정보 처리방침의 수립 및 공개) ①항에서는 "개인정보처리자는 다음 각 호의 사항이 포함된 개인정보의 처리방침을 정하여야 한다"고 규정하고 있고 구조는 Fig.3과 같다. 각호는 1.개인정보의 처리목적, 2.개인정보의 처리 및 보유 기간, 3.개인정보의 제3자 제공에 관한 사항, 4.개인정보처리의 위탁에 관한 사항, 5.정보 주체와 법정대리인의 권리·의무 및 그 행사방법에 관한 사항, 6.개인정보 보호책임자의 성명 또는 개인정보보호 업무 및 관련 고충사항을 처리하는 부서의 명칭과 전화번호 등 연락처, 7.개인정보를 자동으로 수집하는 장치의 설치·운영 및 그 거부에 관한 사항, 8. 그밖에 대통령령으로 정한 1.처리하는 개인정보의 항목, 2.개인정보 파기에 관한 사항, 3.개인정보 안전성 확보조치에 관한 사항으로 규정하고 있다.

김형건, 최석환 외 2인[6]은 인터넷 사이트에서 제공하는 회원가입 약관은 90%의 사용자가 내용이 어렵거나 읽기 번거롭다는 이유로 자세히 확인하지 않고 동의하고 있어 회원가입 약관의 개인정보 영향

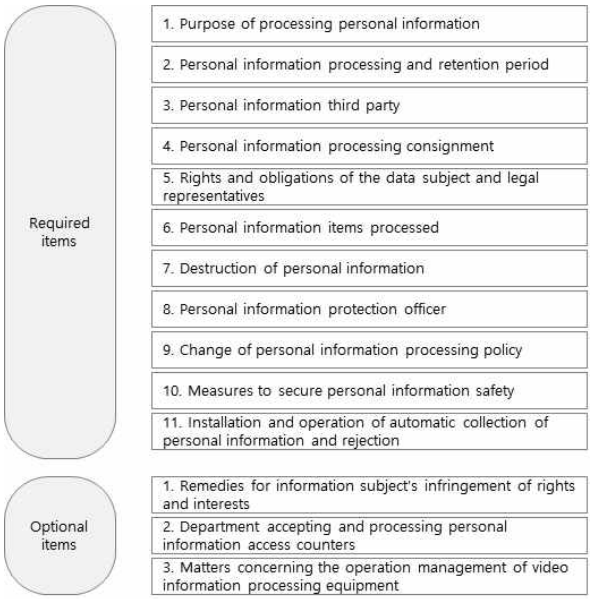

Fig. 3. Privacy Policy Structure

도 분석을 통해 정보 주체가 전문적 지식이 없어도 시각화된 정보를 통해 자기결정권 보장에 이바지 할 수 있어야 한다고 하였다

장원창, 신일순[7]은 3,843명을 대상으로 개인정보 처리방침 인지 여부를 설문조사 하였는데 그중 72.1%는 개인정보 처리방침 내용이 많아 읽어보기 어렵다고 응답하였고, 개인정보 처리방침을 간단하고 읽기 쉽게 만든다면 이용자가 온라인 거래에 참여할 확률이 6.05% 증가하는 것으로 분석하였다.

1997년부터 W3C 주도로 미국 IT 업계와 시민단체가 참여하여 AT&T가 무료로 제공한 Privacy Bird는 P3P(Platform for Privacy Preferences) 표준에 의한 프로그램으로 인터넷 사이트에 접속할 때마다 개인정보 처리방침을 자동으로 수집하여 해당 사이트가 프라이버시 보호수준을 알려주었다. P3P는 xml 포맷으로 생성되어 표준화된 개인정보보호정책을 구현하고 2002년에 W3C에 의해 국제표준으로 승인되었다.

P3P에 관한 연구결과는 많지 않은데 2008년 Lorrie Faith Cranor, Serge Egelman 등의 연구결과에 따르면 P3P 표준을 채택하고 있는 사이트는 10% 미만이나 표준을 채택한 사이트는 개인정보 처리방침(이 연구에서는 Privacy Policies라 표기하였다.)이 모호하지 않고 명확한 이해가 가능하였다고 설명하였다[8].

이와 같이 개인정보 처리방침은 개인정보보호법과

개인정보보관리체계인증기준에 따라 지속적이고 최신화하여 관리가 필요하고, 특히 정확도가 떨어질 경우 허위, 기망의 고지로 판단되어 손해배상 또는 무과실 책임주장 등의 문제로 야기될 수 있다. 해외에는 개인정보 처리방침의 표준화를 통해 정보 주체의 가독성과 인지성을 높이려는 일환으로 추진된 P3P는 채택율이 저조하였다. 다만, P3P는 표준화된 개인정보 처리방침으로 사용자의 명확한 정책 이해에 도움이 되는 것을 알 수 있었다.

R Ramanath, F Liu[9]는 상기와 같이 정렬되지 않거나 제각기 다른 개인정보정책을 해결하기 위한 방안으로 1,010개의 개인정보 처리방침을 수집하고 새로운 말뭉치를 수집한 후, 은닉 마코프 모델을 이용하여 정책을 분리하는 모델을 제안하였다.

2.3 인공지능을 이용한 컴플라이언스 분석

인공지능 기술이 다양한 산업에서 개발되면서 법률 서비스 분야에 적용되기 시작하였는데 이 분야는 Lawtech으로 분류되어 기존의 법률정보 제공과 분석이 자동화되는 시도가 이뤄지고 있다. 관련 연구로는 딥러닝을 이용하여 계약서의 독소조항, 법률 위반 조항을 검출하기 위해 Ilias Chalkidis, Ion Androutsopoulos는 계약서의 주요 법률 요소들을 추출하는 기법을 제안하였다. 이 연구에서는 3,500개의 영문 계약서 데이터셋을 계약서 조항 11개로 구분하고, 각 조항을 Element Type으로 분류하였다. 법률 요소 추출을 위해 개체명 인식에서 모두 가장 높은 성능을 보이는 BilSTM을 기반으로 CRF 방식을 결합한 BilSTM-CRF을 사용하는 것이 효과적이라고 설명하였다[10].

2.4 개인정보보호 정책 데이터셋

국내에서는 개인정보보호 정책의 데이터셋에 관한 자료는 찾을 수 없었고, 해외에서는 OPP-115 코퍼스가 2016년도에 115개 온라인 사이트의 개인정보보호 정책을 데이터화 하여 제공하고 있고, APP-350 코퍼스는 2019년도에 350개의 안드로이드 모바일앱의 개인정보보호 정책을 데이터화 하여 제공하고 있다[11][12]. 웹사이트의 개인정보보호정책을 게시하고 있는 URL을 집합화한 자료로는 MAPS 데이터세트에서 441,626개의 개인정보보호 정책 URL을 구성하고 있다[13].

개인정보 처리방침을 인공지능 기술을 활용하여 분석하기 위해서 필요한 자연어 처리 기술분야에서는 EMNLP 2017에서 개인정보보호정책 중 옵트아웃 정책을 머신러닝과 자연어 처리 모델을 훈련하기 위해 115개의 웹사이트 개인정보보호정책을 기반으로 (이 정책들은 상기에서 언급한 OPP-115와 같다.) 분류하고 문장화 하였다. 115개 웹사이트 개인정보보호 정책을 단락 단위로 세그먼트화 하고, 세그먼트는 정책의 문장 단위별로 세분화하고, 정책에서 포함하고 있는 하이퍼링크는 사전으로 구분하였다[14].

III. 본 론

본 장에서는 기업들의 개인정보보호 정책인 개인정보 처리방침을 수집하여 분석하는 과정을 설명한다. 이를 위해 data collect 부분에서는 기업의 개인정보 처리방침을 수집하여 파서를 통해 저장하고, Mining 부분에서는 KonlPy 형태소 분석을 통해 불용어 처리를 통한 불필요 과정을 줄이고, 빈도를 기반으로 핵심단어를 추출한다. Machine Learning 부분에서는 Fast2Text를 참고로 핵심단어와 처리방침을 벡터화하여 개인정보 처리방침 기반의 데이터셋을 구성하여 분석한다. 이 과정은 Fig.4와 같다.

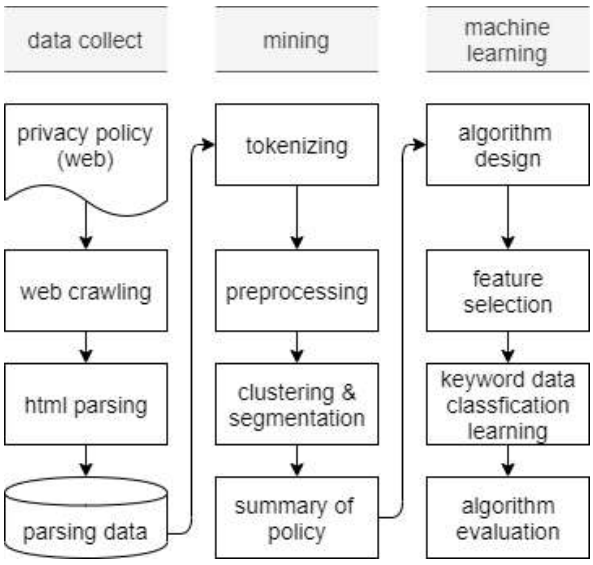

Fig. 4. Process

3.1 Data Collect

본 항에서는 국내 100대 주요기업과 158개 개인정보보호관리체계 인증기업 개인정보 처리방침을 수집하여 방침의 항목별 분류를 통해 3,279개의 개인정보보호 처리방침을 데이터셋으로 구축하였다. 개인정보 처리방침은 기업별로 HTML 구조와 형태가 모두 상이하였고 지침의 항목 이름도 기업별로 달랐다. 개인정보 제3자 제공 및 위탁현황의 경우 일부 기업들은 현황정보를 별도의 URL LINK로 처리하거나 첨부파일 형태로 분리하고 있었다. 개인정보 처리방침은 주기적으로 검토하여 개정이 이루어 지고 있으므로 web crawler는 개정여부를 확인하여 이전 방침에서 개정이 이루어 졌을 때 개정정보를 수집하도록 하였다. 이와 같이 수집된 방침은 Fig.5와 같이 개인정보 처리방침의 14개 구성 항목으로 분류하였다.

분류된 항목들은 방침의 필수적 기재사항 항목제목에 따라 Fig.6과 같이 총 3,119개가 라벨링(Labeling) 되었다.

또한, 향후 연구를 위해 개인정보 처리방침은 정기적으로 검토되고 개정되기 때문에 web crawling

Privacy Policy												
Purpose of processing personal information	Personal information processing and retention period	Personal information third party	Personal information processing consignment	Rights and obligations of the data subject and legal representatives	Personal information items processed	Destruction of personal information	Personal information protection officer	. Change of personal information processing policy	Measures to secure personal information safety	. Installation and operation of automatic collection of personal information and rejection	Remedies for information subject's infringement of rights and interests	Department accepting and processing personal information access counters	Matters concerning the operation management of video information processing equipment

Fig. 5. Privacy Policy Dataset

__label__001__01__	The homepage collects only the minimum personal information necessary to use the service when registering for membership.
__label__001__02__	The specific purpose of collecting and using each personal information item is as follows.
__label__001__03__	Name, ID, password, date of birth, i-pin DI (duplicate registration confirmation information)
__label__001__04__	The retention period of personal information collected on the homepage is as follows. Member information is immediately destroyed upon termination (such as withdrawal application),
__label__001__05__	In principle, the personal information of the data subject is processed within the scope specified for the purpose of collection and use.
__label__001__06__	In handling your personal information, personal information is lost, stolen, leaked, altered or damaged.
__label__001__07__	The website uses a 'cookie' that stores and finds your information from time to time.
__label__001__08__	Protect your personal information and handle complaints related to personal information
__label__001__09__	When personal information becomes unnecessary, such as the elapse of the retention period of personal information or achievement of the purpose of processing, the personal information is destroyed without delay.
__label__001__10__	Announcement of the reason for the change and contents through the website

Fig. 6. Privacy Policy Labeling

단계에서는 상기 대기업과 정보보호 관리체계 인증기업을 포함하여 개인정보 처리방침 개정 때 수집하고, 그 외 기업들의 방침을 수집하기 위해 인터넷에 공개된 기업의 한국어와 영어 개인정보 처리방침을 검색엔진 API를 통해 일 1,000개가 수집하도록 하였다. 이 단계는 다음과 같이 진행하였다. 1) 수집된 데이터에서 HTML, CSS 등의 코드를 제거하여 순수 텍스트(Text)로 정제한다. 2)정제된 텍스트(Text)에서 빈칸, 공백행을 제거한다. 3)한글을 표현하기 위해 UTF-8로 저장한다. 4)텍스트(Text)를 글자 집합으로 저장한다. 5)글자 집합에 인덱스를 부여한다.

3.2 Mining

개인정보처리방침의 문자를 인식하고 최소한의 의미가 있는 데이터로 분리하기 위해 KoNLPy를 이용하여 한글 데이터의 형태소 분석을 시행하였고 전체 방침에서 사용된 형태소 출현 빈도는 Fig.7과 같았고, 개인정보 처리방침 Dataset의 필수적 기재사항 항목별로 형태소 분석과 출현 빈도를 산출하였다.

Fig. 7. Nouns Extracted from the Privacy Policy

분석된 결과 데이터에서는 유의미한 단어 토큰을 선별하고 큰 의미가 없는 단어 토큰을 제거하기 위해 불용어(stopwords)를 제거하였고 품사 태깅을 통해 명사 단어만을 활용하였다. Fig.8는 명사 단어와 출현 빈도수를 나타내고 있는데 명사 외에 '후, 위, 내, 관, 그, 자, 함'과 같은 명사의 오탐글자는 제외하였다.

managem	253	law	144	managem
hu	243	next	140	cost
customer	242	as	136	terminatio
wie	234	alliance	134	card
trade	225	nej	125	delivery
offer	223	collect	108	korea
month	223	gwan	107	phone nu
address	221	Arbitrarily	106	Foreigner
consumer	217	relation	105	역
Applicable	214	step	102	alliance
secession	211	notice	97	date
Item	207	geu	96	name
use	198	credit	93	contract
e-commer	193	member	91	cost
personal	193	Credit info	88	ja
user	184	rule	84	ham

Fig. 8. Nouns Extracted from the Privacy Policy

3.3 임베딩

컴퓨터가 문자를 인식하고 분석하기 위해서는 최근에 Word Embedding이 많이 사용 되는데 이는 '벡터'를 통해서 단어의 관계를 추론하거나 의미적인 유사성으로 분류하는 것이 가능하기 때문이다. 학습 모듈은 방침의 내용들과 중심단어, 주변 단어들을 벡터로 산출하고 조항간의 관계를 도출하는 학습을 진행한다. 출력 모듈은 벡터의 유사도를 계산하여 조항에 대응하는 연관단어를 출력한다. 생성된 모델에 입력값을 넣어 학습시키고 학습이 완료되면 각 방침의 단어벡터의 코사인 유사도가 가장 높은 방침 정보를 제시할 수 있다. 훈련 데이터를 학습하기 위해 Fast2Text를 활용하여 개인정보 처리방침을 학습시키기 위한 모델은 skip-gram 모델을 사용하여 단어 v_c를 기준으로 조항 $u_o..u_3$을 예측하는 모형을 구축한다.

성능을 검증하기 위하여 형태소 분석을 통해 각 조항별로 핵심어를 추출하였으며 입력값과 산출된 개인정보 처리방침 조항의 관련성을 비교하여 정밀도와 재현율로 평가하였다. 정밀도는 검색된 적합 조항 Label/(검색된 적합 조항 Label+검색된 부적합 Label) * 100으로 계산하였으며 계산식은 다음과 같다.

$$Precision = \frac{tp}{tp + fp} \quad (1)$$

재현율은 검색된 적합 조항 Label/(검색된 적합 조항 Label+검색되지 않은 부적합 Label) * 100으로 계산하였으며 계산식은 다음과 같다.

$$Recall = \frac{tp}{tp + fn} \quad (2)$$

본 연구 결과 평균 정밀도는 46.98%, 재현율은 63.86%로 개인정보 국외 이전에 대한 방침상 내용과 핵심어가 부족한 경우 정밀도가 현저히 낮았으나 전반적으로 재현율은 유사한 수준이었다. 상세내용은 Table. 1과 같다.

Table 1. Precision recall rate result

Keyword	Precision	Recall
Purpose of processing personal information of KEPCO	45.3%	65.1%
Samsung Electronics Personal Information Collection Items	44.0%	63.2%
Purpose of processing personal information at Hyundai Department Store	63.9%	62.3%
Korean Air Personal Information Transfer	23.3%	44.6%
Application of power exchange privacy policy	58.4%	64.3%
Average	46.98%	59.98%

IV. 결 론

개인정보 처리방침은 사업자 측면에서는 개인정보보호법, 정보통신망법 요구사항을 이행하고 소비자 보호를 위한 투명한 개인정보 처리의무를 이행하는데 목적이 있다. 정보주체 또는 서비스 이용자 측면에서는 거래 당사자로서 사업자의 지위 남용과 부당한 개인정보처리와 정보주체 권리 침해요소를 식별할 수 있는 의미를 지닌다. 개인정보 처리방침은 개인정보 보호 정책에서 가장 소비자 노출이 높은 것으로 최신성을 유지하면서 정보 주체의 알 권리가 보장될 수 있도록 명확화 되어야 한다. 개인정보 처리방침은 국제표준으로 규격화 되었지만 전 세계적으로 표준 이행율은 저조하여 모바일, 웹, IoT 등의 서비스 증가

로 정보 주체의 피로도는 증가할 것이다. 본 연구에서는 국내 주요기업과 개인 정보관리체계 인증기업들의 개인정보 처리방침을 수집하여 분석하였다.

본 논문에서 제안한 모델을 활용하여 복잡하고 읽기 어려운 개인정보 처리방침을 정보주체가 이해하기 쉽게 제공한다면 개인정보 처리방침의 작성 주체인 사업자는 명확하고 단순화된 개인정보 처리방침을 통해 거래 매출 증가 기대가 가능하고, 정보주체와 사업자간에 존재하는 정보 비대칭성(information asymmetry)을 해결하는데 기여 할 것이다.

하지만 본 연구의 제한사항으로 처리방침을 주요기업과 ISMS 인증기업으로 한정하여 다양한 데이터셋을 구축하지 못했고, 자연어 처리와 학습모델의 영향도와 성능에 대한 고려가 추가적으로 요구될 것으로 판단된다. 향후 연구를 통해 미비점을 보강하고 충분한 국내/외 개인정보 처리방침 데이터셋을 구축하고자 한다.

본 연구는 사람(전문가)에 의해 검토하여 개인정보보호 관리체계 인증 심사를 통과한 개인정보 처리방침을 한글 자연어 처리 성능이 입증된 임베딩 기법과 임베딩을 통하여 개인정보 처리방침이 자동화된 요약과 분석이 가능한 것을 입증한 연구로써 국내 주요 기업의 개인정보 처리방침 데이터셋을 구축하고 분석을 시도한 첫 사례라는 시사점이 있다.

향후 연구로는 본 논문에서 수집, 분석된 데이터셋을 기반으로 개인정보 처리방침을 가시화 함으로써 가독성을 강화하고, 정보보호관리체계 인증을 통해 검증된 기업의 개인정보 처리방침 데이터를 기반으로 기업의 개인정보 처리방침 자동화 점검 도구를 제안하고자 한다.

References

[1] Yoon Kim, "Convolutional Neural Network for Sentence Classification," Proc. of the 2014 Conference on Empircal Methods in Natural Language Processing(EMNLP), pp.1746-1751, Oct.2014.

[2] Jee-Uk Heu, "Analysis and Comparison of Query focused Korean Document Summarization using Word Embedding", The Journal of The Institute of Internet, Broadcasting and Communication, vol. 19, No. 6, pp.161-167, Dec. 2019

[3] Dowoo Kim Myoung-Wan Koo , "Categorization of Korean News Articles Based on Convolutional Neural Network Using Doc2Vec and Word2Vec", Journal of KIISE, vol 44 no. 07 pp. 0742 ~ 0747, July, 2017

[4] Ministry of Public Administration and Security, Personal Information Protection Act and Guidance., Dec, 2011

[5] Jung Woon Lee, "Legal Characteristics of Personal Information Processing Policy and Its Application-Focusing on Critical Review on the Terms of Personal Information Processing Policy", Gachon Law no. 1, p.43-84, Mar. 2017.

[6] Hyung-Kun Kim, Seok-Hwan Choi and 2 others, "Web-based Privacy Impact Assessment System through Automatic Collection of Membership Terms and Conditions", Journal of Computing of Korean Information Science Society 25(9), pp. 425-435, Aug. 2019

[7] Wonchang Jang, Ilsoon Shin, "The Online Privacy Policy: Recognition, Confirmation and its Effects on Online Transaction Behavior", Korea Institutes of Information Security and Cryptology, vol. 22, Dec. 2012

[8] Lorrie Faith Cranor, Serge Egelman, Steve Sheng, Aleecia M. McDonald, and Abdur Chowdhury. P3P Deployment on Websites. Electronic Commerce Research and Applications, vol. 7, Issue 3, pp. 274-293, Apr. 2008

[9] Ramanath, R., Liu, F., Sadeh, N., and Smith, "Unsupervised alignment of privacy policies using hidden markov models", In Proceedings of the 52nd Annual Meeting of the

Association for Computational Linguistics, Vol.2, pp. 605-610, Jul 2014
〔10〕 Chalkidis, Ilias and Ion Androutsopoulos. “A Deep Learning Approach to Contract Element Extraction.”, International conference on Legal Knowledge and Information Systems, Dec. 2017
〔11〕 Shomir Wilson and Florian Schaub, “The Creation and Analysis of a Website Privacy Policy Corpus, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics”, Vol.1, aug, 2016
〔12〕 Sebastian Zimmeck, Peter Story, “MAPS: Scaling Privacy Compliance Analysis to a Million Apps”, Proceedings on Privacy Enhancing Technologies, Vol. 2019, Jul 2019
〔13〕 Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell and Norman Sadeh. “Privacy Enhancement Technology Symposium 2019
〔14〕 Kanthashree Mysore Sathyendra, “Identifying the provision of choices in privacy policy text”, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.2774-2779, Aug. 2017

〈저자소개〉

조 용 현(Yong-hyun Jo) 정회원
2004년 8월: 경희대학교 졸업
2007년 2월: 아주대학교 정보통신대학원 정보보호전공 석사
2020년 2월: 고려대학교 정보보호대학원 융합보안학과 박사 수료
2002년~2007년: 육군중앙수사단 사이버범죄수사/디지털증거분석 수사관
2009년~2014년: 비씨카드 정보보안실, 신한카드 정보보호팀
〈관심분야〉 디지털 포렌식, 사고대응, 정보보호 정책, 개인정보보호, 융합보안

차 영 균 (Young-kyun Cha) 종신회원
1989년 2월: 고려대학교 수학과 졸업
1992년 6월: 고려대학교 교육대학원 석사
2012년 8월: 고려대학교 정보보호대학원 박사
2018년~현재: 고려대학교 정보보호대학원 특임교수
〈관심분야〉 융합보안, 암호학, 물리보안, 금융보안, 정보보호 정책