Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing

Hyejin Jang, Yujin Jeong, Byungun Yoon*

*Department of Industrial & Systems Engineering, School of Engineering, Dongguk University, Seoul, Republic of Korea*

ARTICLE INFO

*Keywords:*
Patent mining
Natural language processing
Text mining
Lexical analysis
WordNet

ABSTRACT

The role of text mining based on technological documents such as patents is important in the research field of technology intelligence for technology R&D planning. In addition, WordNet, an English-based lexical database, is widely used for pre-processing text data such as word lemmatization and synonym search. However, technological vocabulary information is complex and specific, and WordNet's ability to analyze technological information is limited in its reflecting technological features. Thus, to improve the text mining performance of technological information, this study proposes a methodology for designing a TechWord-based lexical database that is based on the lexical characteristics of technological words that are differentiated from general words. To do this, we define TechWord, a technology lexical information, and construct a TechSynset, a synonym set between TechWords. First, through dependency parsing between words, TechWord, a unit word that describes a technology, is structured and identifies nouns and verbs. The importance of connectivity is investigated by a network centrality index analysis based on the dependency relations of words. Subsequently, to search for synonyms suitable for the target technology domain, a TechSynset is constructed through synset information, with an additional analysis that calculates cosine similarity based on a word embedding vector. Applying the proposed methodology to the actual technology-related information analysis, we collect patent data on the technological fields of the automotive field, and present the results of the TechWord and TechSynset. This study improves technological information-based text mining by structuring the word-to-word link information in technological documents based on an automated process.

## 1. Introduction

With the rapid changes in the technology development environment, the role of technology intelligence has been emphasized for technology management. An abundance of patent data has accumulated and is widely utilized as a qualified source for technology intelligence in analyzing technological trends, strategic technology planning, and so forth (Abbas, Zhang, & Khan, 2014; Bonino, Ciaramella, & Corno, 2010). In particular, textual data in patent databases provide worthwhile technological information that is frequently adopted for various text mining approaches. Text mining-based methodologies have enhanced previous bibliographic information analysis based on class code and citations that are limited in interpreting specific technical contents. In the early stages of text mining for technology intelligence, a keyword-based analysis was conducted to derive main words that appear frequently in documents based on Term Frequency-Inverse Document Frequency (TF-IDF) (Lee, Yoon, & Park, 2009; Yoon & Park, 2005). Subject-Action-Object (SAO) analysis has developed keyword analysis by considering the technical properties and syntactic relations between the subject and verb clause (verb and object) in a sentence (Moehrle, 2010; Yoon & Kim, 2012b). The SAO structure, in which Action and Object (AO) indicate the problem and subject (S) represents the solution, is analyzed (Wang et al., 2017). In addition, in another text mining approach, topic modeling shows topic clusters consisting of a keyword list by considering the distributions of words and documents (Jang, Roh, & Yoon, 2017; Kwon, Kim, & Park, 2017). All of the text mining-based studies have constructed a lexical vocabulary for their analysis. For that reason, structuring lexical information is directly related to the performance and quality of patent-based text mining.

WordNet is the lexical database system of English developed by the Cognitive Science Laboratory of Princeton University (Miller, 1995),

* Corresponding author at: Department of Industrial & Systems Engineering, School of Engineering, Dongguk University, 3-26, Pil-dong 3ga, Chung-gu, Seoul 100-715, Republic of Korea.
*E-mail addresses:* jhj_9055@naver.com (H. Jang), withss501@naver.com (Y. Jeong), postman3@dongguk.edu (B. Yoon).

https://doi.org/10.1016/j.eswa.2020.114042
Received 15 April 2020; Received in revised form 22 August 2020; Accepted 22 September 2020
Available online 25 September 2020

and its lexical information is actively used in the text mining process. In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (called synsets), each expressing a distinct concept. This technique is widely used to enhance the quality of textual semantic analysis. Previous studies have expanded the content of synsets and their semantic relations by combining knowledge sources such as an existing domain ontology or Wikipedia (Barbu, 2015; Esuli & Sebastiani, 2006; Suchanek, Kasneci, & Weikum, 2008). WordNet is also used to measure semantic similarity by calculating the distance between words based on the hierarchical structure of the database (Wei, Lu, Chang, Zhou, & Bao, 2015; Wu & Palmer, 1994).

Prior researches on technology intelligence have utilized the WordNet database. Most of the existing studies have applied WordNet synset and lexical relation information to preprocess synonym words, or to calculate the semantic similarity between words (Choi, Yoon, Kim, Lee, & Kim, 2011; Joung & Kim, 2017; Park, Kim, Choi, & Yoon, 2013; Park, Yoon, & Kim, 2013; Wang et al., 2019). In addition, a methodology for morphology analysis was developed through the specific relationship information in WordNet, where the meronym–holonym relationship was composed of the morphology's dimension, and the value of each dimension was generated as the hypernyms–hyponyms (Geum & Park, 2016). However, WordNet itself has limitations in terms of applying text mining to technology intelligence because of the heterogeneous nature of technology-related textual information that is unlike the general one. WordNet does not cover all the lexical information of technology-related documents. For instance, in the particular case of nouns, a technological concept is expressed in the form of a phrase, such as a multi-word. Furthermore, WordNet organizes the "entity" into the highest level of nouns in terms of general vocabulary, including "physical entity" and "abstract entity" as sub-levels. WordNet's hierarchical structure system is disparate when applied to technology-related textual information.

The main differences between general natural language processing (NLP) words and technological words can be summarized in two points based on the characteristics of the relations between words. First, technological objects mainly expressed as nouns are written in multi-word form rather than as single words and contain the hierarchical semantic information of the description. For example, the word "system," which frequently appears in a patent document, is written as a compound noun such as "control system" and "brake control system" to describe a detailed technological object. In the previous study on structuring sentences in patent documents, this point was defined as a major problem in analyzing complex multi-word domain terms (Yang & Soo, 2012). The second difference is that the relations between a noun and verb in a clause, where the technological function is written through verbs that are syntactically linked to nouns describing the technical subject or object, contains important information. Due to these structural features, existing studies on SAO analysis are actively being conducted. Developing a systematic framework is necessary and increasingly more important as high-accuracy NLP techniques are continuously being developed, along with the description method's features that are based on the syntactic features of technical words.

This study aims to structure textual technology information as a cornerstone of a lexical database for technology-related information. A methodology is suggested to extract unit lexical items named as a TechWord, and construct synonym sets for each TechWord as a TechSynset based on the textual data of the patent database. First, textual technology information is structured through dependency parsing, which extracts a grammatical structure to derive phrase-based technology information according to both noun phrases and SAO-based verb structures. Structured technology-related lexical information is the explicated as a TechWord by calculating the network degree centrality that constructs networks between lexicon (nodes) and dependency relations (arcs). A TechSynset, meaning a group of TechWords that are semantically equivalent or analogous, is constructed on the basis of the synset information of WordNet; and if WordNet does not cover the synset of a TechWord, it is additionally explored through the word embedding model of Bidirectional Encoder Representations from Transformer (BERT). A final lexical database of a TechWord is constructed by considering the word-to-word relations and synonym sets that result from the previous step.

The remainder of this paper is structured as follows. Section 2 delineates backgrounds, and reviews the previous studies on structuring technology lexical information, which is the main research theme in this paper, as well as the two main methodologies related to NLP and the BERT model. Section 3 proposes a research framework to develop a technology lexical database of TechWord and Section 4 includes a case study in the automotive field to illustrate the proposed framework, followed by a discussion section regarding the results of the case study in depth. Finally, Section 6 concludes the paper with the contribution of the research, its limitations, and further study.

## 2. Background

### 2.1. Structuring technology lexical information

Technology intelligence based on text mining aims to systematically interpret the complex and vast text-based information on technology. As words used by humans are divided into parts of speech (POS) according to their grammatical functions, forms, and meanings, and as words form relationships between words, technological attributes are described in words that describe the technology and are expressed by the relations between each attribute. Structuring lexical information that forms technology-related textual data helps to interpret the content of technology more systematically, and to provide high-level insights (Yoon & Park, 2004). In organizing technology information, various technological attributes have been considered a base concept by analyzing grammatical patterns in a sentence tagged with POS information, representatively using abstract and claim documents in the patent database.

Property–function analysis was introduced to extract technological attributes (Dewulf, 2011; Verhaegen, D'hondt, Vertommen, Dewulf, & Duflou, 2009). Property is defined as "what a product is or has" and its attributes that are mainly expressed in adjectives; function is described as "what does or undergoes" its useful action as based on a specific purpose that is mainly expressed in verbs. Based on the property–function analysis, patent networks were analyzed for technology forecasting and strategic planning (Yoon & Kim, 2012a). SAO analysis became a widely used approach to form technology-related lexical information. Based on SAO structures and by referring specific action word sets that link the subject to the object, a framework for extracting keyword phrases from patents was suggested by dividing them into four types: product, technology, material, and technology attribute (Choi, Park, Kang, Lee, & Kim, 2012). For determining the direction of technological change, SAO chains were analyzed to construct technology morphological structure as having a product, components (attributives), and material (Guo, Wang, Li, & Zhu, 2016). Furthermore, technology opportunities were investigated through SAO-based methodology, where core technological attributes were considered as elements, field, and purposes/effects (Kim, Park, & Lee, 2019). In addition, meaningful keyword sets of technological information were structured based on technological information types, composed of function, object, component, and method of operation (Roh, Jeong, & Yoon, 2017). POS-tagged grammatical rules then extracted these types of technological information.

Existing studies on the structure of technical information have mostly categorized technology attributes through grammatical patterns using POS. In SAO approaches, major technical keywords that are considered as "subject" or "object" were extracted based on nouns (or noun phrases), "function" and "operation methods" of technical keywords were based on verbs, and "properties" and "attributes" were analyzed based on adjectives. In addition, further attributes such as components and material were based on relational information linked to the specific verbs. In this study, we focus on the nouns and verbs of the parts of words to define the vocabulary of technology, and structure the technical information by analyzing the relationship between noun–nouns and noun–verbs. Table 1 summarizes the comparison between existing approaches and the TechWord presented in this study.

### 2.2. Dependency parser

A dependency parser is one of the NLP tasks that aims to analyze the grammatical structure of a sentence. Relationships between "head" words and words that modify those heads are established as dependency parsing results. Parsing information describes a sentence's syntactic structure with regard to the words (or lemmas) in a sentence, and an associated set of directed binary grammatical relations that hold among the words (Martin & Jurafsky, 2009). The result of a dependency-parsed sentence is expressed in the form of a directed graph with a dependency relation (dobj; direct objective, nsubj; nominal subjective, etc.) between the head and dependent word, as shown in Fig. 1. The parsed data is employed as a feature of the deep learning architectures of the word or sentence embedding model (Ma, Huang, Xiang, & Zhou, 2015), document summarization (Moawad & Aref, 2012; Rachabathuni, 2017), and other NLP tasks and text mining related models. In this paper, the latest version of the CoreNLP toolkit (version 3.9.2), which was developed by the Stanford NLP Group, is utilized as a dependency parser (Manning et al., 2014). CoreNLP is one of the most stabilized open sources for natural language parsing, which was initially released at the end of 2002, and has been updated. CoreNLP is available as an online demo version (http://corenlp.run/), and in this study, a Java-based library was installed and analyzed through Python (https://stanfordnlp.github.io/CoreNLP/index.html).

In addition, the technology intelligence field has adopted dependency parsing, especially for the preprocessing step. Lexical chains in patent claims were identified by using a dependency parser to develop a patent summarizer that identifies the idiosyncrasies of the patent genre (Brügmann et al., 2015). A framework for content-oriented patent document processing was developed by incorporating a series of linguistic processing tools that included dependency parsing (Wanner et al., 2008). A methodology of proposition-based semantic analysis was suggested to develop a patent network by dependency parsing (An, Kim, Mortara, & Lee, 2018). Since grammatical patterns exist wherein prepositions are used in the texts to link technological keywords, the relationships between tech-keywords were identified. Yang & Soo (2012) have extracted conceptual graphs from a patent claim using syntactic information (POS, and dependency tree), and semantic information (background ontology) (Yang & Soo, 2012). In summary, the parsed result itself extracts textual information, such as a noun or verb phrase, or interpreted as a network structure to expand to an additional model. This study corresponds to the latter, and attempts to structure textual information based on lexical relationships as a network.

**Table 1**
Comparison with Previous researches and the proposed approach.

| References | Pros and cons |
|---|---|
| Property–function analysis (Dewulf, 2011; Verhaegen et al., 2009) | Analyzing product-view technological information based on the part of speech<br>Using only adjective(property) or verb (function) type of single word form textual information |
| SAO structure analysis (Yoon & Kim, 2012a; Choi, Park, Kang, Lee, & Kim, 2012; Guo, Wang, Li, & Zhu, 2016; Kim, Park, & Lee, 2019) | Analyzing technological problem and solution based on SAO structures<br>Necessary to manually predefine verb word list related to technological action |
| Others (POS-tagged grammatical rules) (Roh, Jeong, & Yoon, 2017) | Extracting function, object, component and method of operation based on structures of patent document<br>Analyzed based on detailed rules defined manually |
| Proposed approach (TechWord) | Automatically extracting TechWords by considering both multi-word nouns and SAO structures<br>Reflecting importance of words based on network-based relations between words |

### 2.3. Bidirectional Encoder representations from Transformer (BERT)

Word embedding is one of the most useful techniques in NLP, where words or phrases from the vocabulary are mapped to numerical vectors. Recently, deep learning-based word embedding models have been utilized, including Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014). The result of embedding is utilized for various text mining fields to calculate the similarity between words or phrases by considering their contextual information. In previous studies, term vectorization has been conducted, which aimed to build a TechNet, a semantic network of technological concepts based on semantically relevant terms (Sarica, Luo, & Wood, 2020), keyphrase extraction, and technology opportunity discovery by using both user opinion and technological information (Roh, Jeong, Jang, & Yoon, 2019). In the patent document, TechNet did not reflect the syntactic characteristics of the vocabulary expressing the technology as it applied text preprocessing through NLP and vectorization through the embedding learning model.

BERT, a state-of-the-art language representation model introduced by Google, is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers, as shown in Fig. 2 (Devlin, Chang, Lee, & Toutanova, 2018). BERT is a model for transfer learning to labeled data with specific tasks after pre-training the model with the big unlabeled data, including Wikipedia and Book Corpus (Zhu et al., 2015). Pre-training of the language model is an effective way to improve the performance of many NLP tasks (Dai & Le, 2015; Radford, Narasimhan, Salimans, & Sutskever, 2018). The BERT model is a fine-tuning approach, where the objective function minimizes task-specific parameters, and fine-tunes pre-trained parameters by learning downstream tasks. While the previous models, ELMo and OpenAI GPT, train with the identical objective function during pre-training, BERT learns pre-trained language representation in a new way, which outperforms the previous models. In experiments that evaluated the grounded common sense inference using the Situations with Adversarial Generations (SWAG) dataset, the BERT model outperformed the ELMo and GloVe systems by +27.1% and 33.6%, respectively.

## 3. Research framework

We propose a framework that develops a technology-related lexical database for structuring textual technology information based on NLP. A complex syntactical configuration is considered as the technological relationship, linking noun phrases or verbs to technological subjects, objects, or action. The overall framework consists of three steps in which the data is collected and processed, the TechWord is extracted, and the TechSynset is constructed. First, in the data collection and preprocessing step, the dependency relationship is extracted from the text-based bibliographic information collected from the patent database. Based on the dependency relations, the TechWord is defined as a keyword that describes a technology by analyzing grammatical dependencies focusing on noun phrases and the clauses, including subject and verb, and scoring by network-based degree centrality. Finally, to define the TechSynset, a synonym set for TechWord, we employ the synset information in the WordNet database, and calculate the cosine similarity based on the word embedding vector for each TechWord. Fig. 3 illustrates the dataflow-based processes. In this study, the code for the analysis was implemented by python, and the result data, represented as data storage within a data flow chart, was stored in a json format.

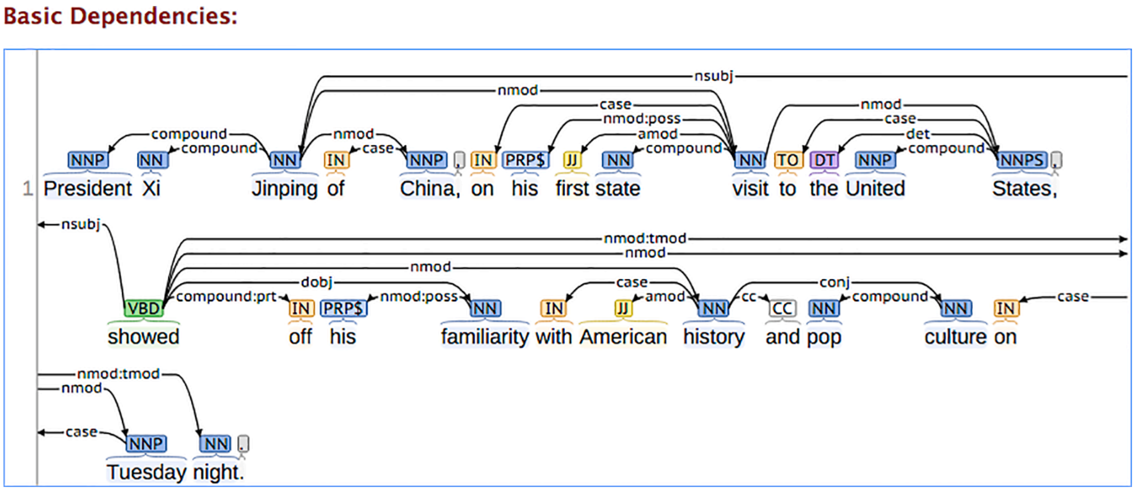
Basic Dependencies:
1 President Xi Jinping of China, on his first state visit to the United States,
showed off his familiarity with American history and pop culture on
Tuesday night.
NNP NN NN IN NNP , IN PRP$ JJ NN NN TO DT NNP NNPS ,
VBD IN PRP$ NN IN JJ NN CC NN NN IN
NNP NN .
compound
nsubj
nmod
case
nmod:poss
amod
det
compound:prt
dobj
nmod:tmod
cc
conj

Fig. 1. An example of dependency parser (CoreNLP).

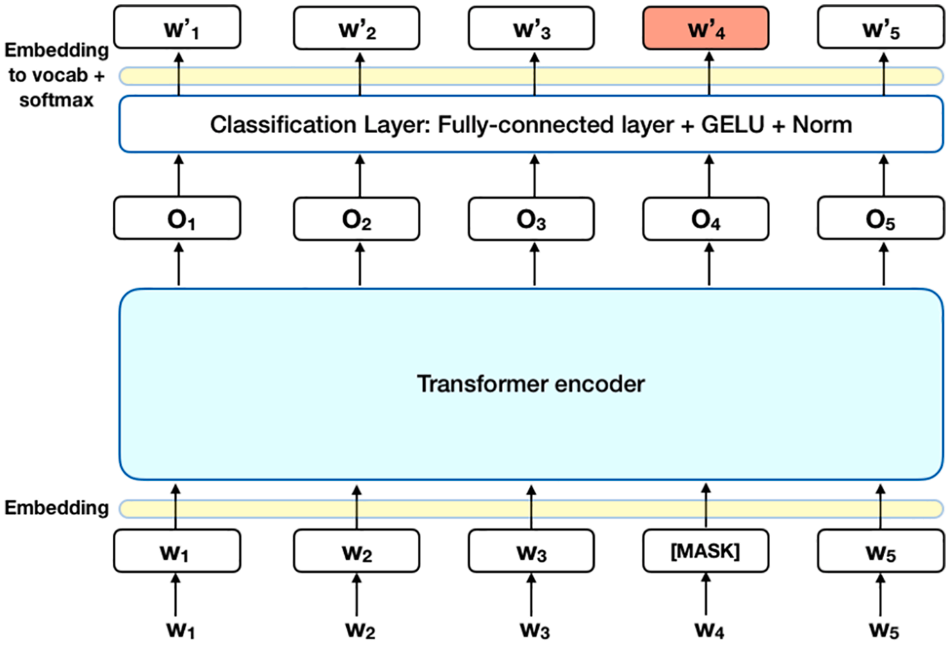
w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Fig. 2. The BERT model architecture.

### 3.1. Data collection & preprocessing

In this study, we analyze text-based abstracts and claims in the bibliography field of patents. Since this study proposes a methodology for technology management and intelligence, patent data, which is a vast and high-quality collection of refined technological information, is used as the analytical data. Fig. 4 shows the patent collection and preprocessing processes. After collecting patent documents related to the technology to be analyzed, textual information is extracted from patent abstracts and claims. In the sentence tokenizing, all documents are split into sentences to analyze the dependency parsing for each sentence. Claims in patent documents have long sentences, which can cause crash problems in dependency parsing. To solve this problem, a sentence is divided into sub-sentences by applying the claim sentence splitting process proposed by Yang and Soo (2012). The target token is then defined, which includes the transition phrase (comprising, including, etc.), the conjunction-word (wherein, etc.), the list item ("(a)". "(i)", etc.), and punctuation (";") in consideration of the patent claim writing style. After this, the heuristic-based split manual for a specific target token is defined and divided into sub-sentences using syntactic information. A grammatically complete sentence is an input of dependency parsing without eliminating stop words and punctuation. For each word in a sentence, the head word and the dependency relation type associated with the head word are extracted through dependency parsing. NLTK (https://www.nltk.org/) and CoreNLP Dependency Parser of Python packages are employed for the NLP.

### 3.2. TechWord extraction

The TechWord candidate is extracted from the noun phrase and the SAO structure based on the dependency relation information derived in the previous section, and the importance of the TechWord candidate is evaluated through the centrality index by interpreting the connection relationship between each word as a network. First, TechWord candidates are divided into two types: nouns and verbs, which are extracted through the grammatical patterns of dependency relation information, as shown in Table 2. The noun phrase words are structured based on the concept of hierarchy from the technology tree perspective. TechWord

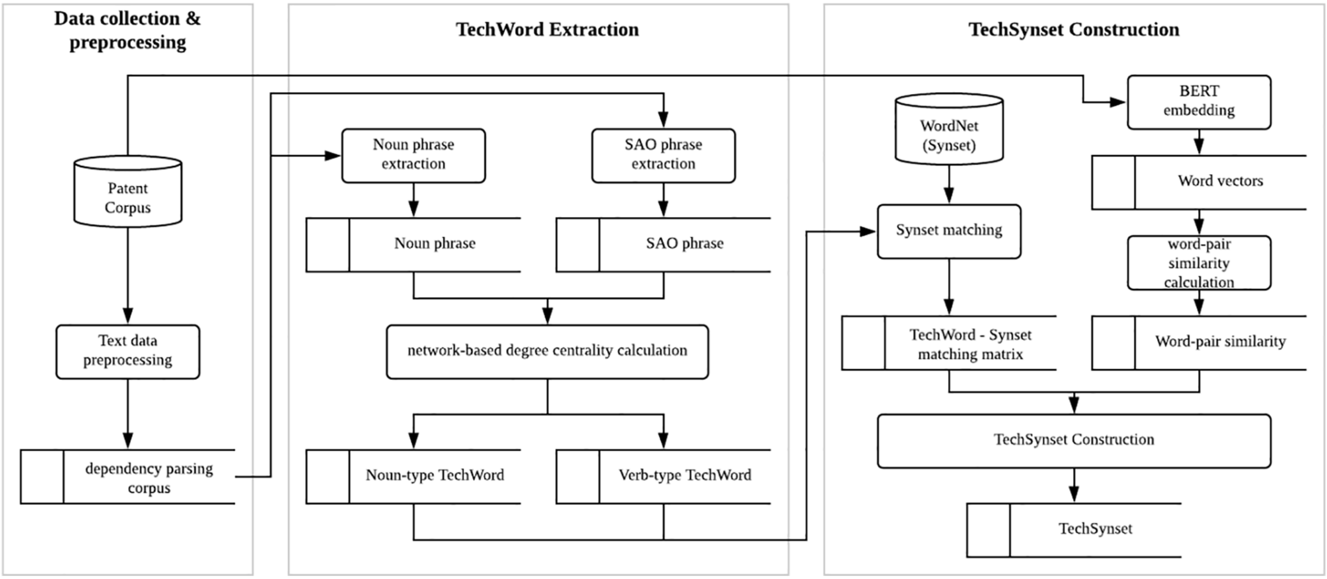

**Fig. 3.** Data flow chart.

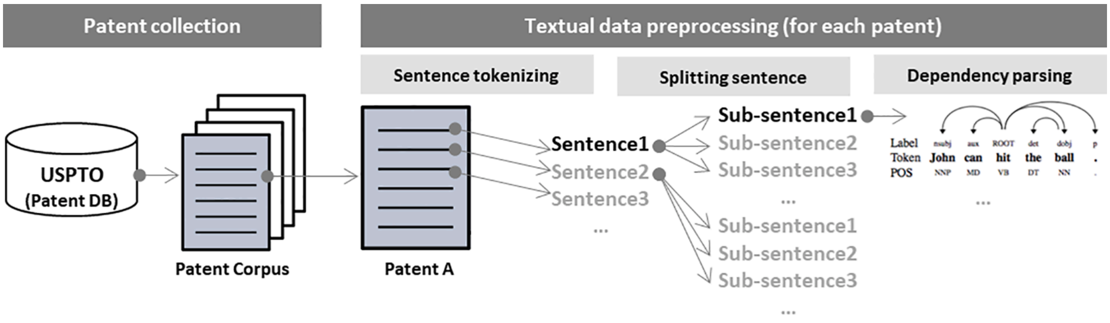

**Fig. 4.** Processes of data collection and preprocessing.

**Table 2**
Description of TechWord type.

| TechWord Type | | Dependency Relation | | Example |
|---|---|---|---|---|
| Noun (phrase) | Hierarchical structure | compound / amod | | (a) **vehicular control system**…<br>- compound: control ← system<br>- amod: vehicular ← system |
| Verb | SAO structure | active | nsubj + dobj | (b) A vehicular control **system includes** a forward viewing **camera** that views…<br>- nsubj: system ← include<br>- dobj: include → camera |
| | | | acl + dobj | (c) a central computer system **having** a network interface<br>- acl : system → having<br>- dobj: having → interface |
| | | passive | nsubjpass + nmod + case(by) | (d) The magnetic **attraction** is **provided***by***incorporation** of magnetic particles…<br>- attraction ← provided (pp)<br>- provided(pp) → incorportation |
| | | | nmod + case(by) + acl | (e) … image **data captured***by* the **camera**.<br>- nmod: captured(pp) → camera<br>- acl: data → captured(pp) |

candidates in the form of multi-word and noun phrases are extracted through the dependency relation types of a compound word and adjective modifier tagged as "compound" and "adjective modification (amod)", respectively. Verb types are extracted based on the SAO structure, where subject and object are connected in the form of extracted noun phrases. In this case, the rules of dependency relation chains are defined by four types: active and passive type, and clause and modifying phrase type.

In the active form, “subject + verb + object” is identified by linking “subject + verb” connected in the noun form of the subject relation (nsubj), and “verb + object” is linked to the identified verb in nsubj and directed object (dobj). Active forms can also be expressed in the form of the modifying phrases of “acl”, standing for the adjectival clause that modifies a nominal where head of the acl relation is the modified noun and the dependent is the head of the clause that modifies the noun, and “dobj”, which means a noun phrase that is the verb’s (accusative) object, followed by the same verb of “acl”. In the case of passive form, lexical information is structured by converting the passive-based dependency information into an active representation. That is, the lexical information is structured by converting the subject in the passive form into the object, and converting the noun phrase after “by” into the subject. The clause type in the SAO structure of the passive form is defined by the dependency relation chain connected by “nsubjpass (passive nominal subject),” “nmod (nominal modifier),” and “case (by),” which respectively stand for: a noun phrase that is the syntactic subject of a passive clause, the nominal dependents of another noun or noun phrase that functionally corresponds to an attribute, and any case-marking “by” element that is treated as a separate syntactic word. The last form of the passive modifying phrase is defined through a dependency relation chain of “nmod + case (by) + acl”. Fig. 5 shows an example of each TechWord type.

In the next step, the TechWord candidates’ importance is evaluated through node centrality by interpreting nodes as words and edges as dependency-based connections. Since both noun phrases and SAO structures are expressed as dependency relations between words, each TechWord candidate can be represented by one network graph. In this study, we analyze the degree centrality’s in-degree centrality and out-degree centrality to reflect the directed network characteristics expressing the direction of head and dependent words. Although, in the centrality index of the network node there are various indicators, such as degree, PageRank, Betweenness, and Closeness.

The main TechWords expressed in the nodes are evaluated through centrality analysis by interpreting them from the network point of view represented by the word-to-word connections (Choi et al., 2011; Park, Kim, et al., 2013; Yang, Huang, & Su, 2018). The high in-degree object node in the SAO structure is interpreted as the more useful technology in various areas, which means it is the important technological indicator for improving performance, as shown in Table 3. In addition, a high out-degree of the noun node is regarded as the existing and/or general technology to the relevant technology area, and used as a solution for achieving technology problems. The third type of high in-out degree in the inter-node connection in the AO relationship is described as a widely used function (action–object), meaning an important technology purpose. The high centrality of the in-degree node on the network graph is used to describe the high-level concept of the technology that forms the underlying structure within the noun phrases’ hierarchy. A noun word with a high out-degree within noun phrases is a concept that explains the key sub-technology of the technology analyzed. A verb type of TechWord is analyzed by dividing the SAO into subject-related actions (SA) and object-related actions (AO), respectively, to evaluate verb nodes that represent major functions.

### 3.3. *TechSynset construction*

When it comes to dealing with text data, preprocessing of synonyms in the vocabularies is one of the imperative processes because the sparsity of words has a crucial impact on the final result of text mining.

**Table 3**
Conceptual illustration of in- and out-degree centralities of TechWord.

| TechWord type | Direction of node degree | Interpretation of technology viewpoint |
|---|---|---|
| Noun word in a noun phrase | High in-degree noun | High level concept of the technology that forms the underlying structure, basis form in which technologies are embodied |
| | High out-degree noun | Sub or part of technology to be analyzed (dependent on underlying lexicon); multi-word technological lexicon (dependent on specialized lexicon) |
| Verb word in SAO structure | High in-degree verb in SA | Subject-oriented function with emphasizing the subject in problem-solving |
| | High out-degree verb in AO | Object-oriented function emphasizing the purpose of the technology |

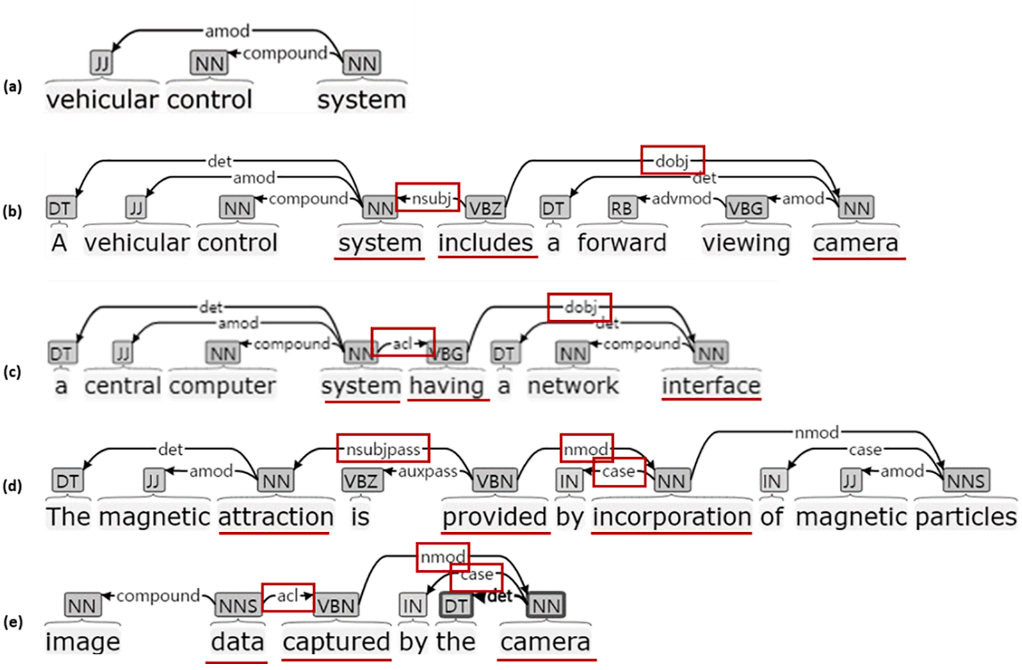

**Fig. 5.** Dependency rule of extracting TechWord candidate.

For this reason, we construct a TechSynset that constitutes a synonym set of the TechWord. Based on the existing synset information in WordNet, the contextual similarity between TechWords is calculated based on their word embedding vectors, as shown in Fig. 6. First, synonym information for each word of the TechWord is retrieved from WordNet, using the "wn" module in the Python package "nltk.corpus". A matrix between the TechWord and the synset list that each TechWord can have is created. In parallel, word embedding vectors are obtained for each TechWord, adopting BERT (Devlin et al., 2018). We calculate the cosine similarity of the embedding vectors between the TechWord pairs defined as the synset to check whether the synset information retrieved from WordNet is applicable in the technology domain. In particular, extending the synonym set by combining words is necessary, since the noun information of technological information is often composed of multi-words. A synset of the noun phrase of a TechWord is formed by substituting each word by using the synset information of each word, comparing words in one-to-one and one-to-many comparisons. At this time, the number of multi-words is limited to those composed of up to four words that occupy more than 90% of the total, which is derived from the analysis data of the following case studies.

## 4. Case study

### *4.1. Data collection and preprocessing*

The automotive sector is one of the broad and complicated fields in the hierarchical structures of technological objects, and the related technological functions embodied in the parts. For this reason, the automobile field's technological characteristics are suitable for the following analysis, which intends to assimilate technology-related information into a structural framework. Patents registered in the United States Patent and Trademark Office (USPTO) are collected to analyze English-based lexical information. Among the International Patent Classification (IPC) codes of the patent database, patents from the IPC code for automobiles, B60 (vehicles in general) are collected from a recent five-year period, followed by patent search query of "ICH = B60 AND AD >=20140701" in Wisdomain.com, a solution for patent collection and analysis. Finally, 34,823 patent data were collected for the five-year period from July 2014 to June 2019.

The abstract field in the patent document is used for the following analysis. All sentences in the patent document are split into sentence units based on the split rules. Each sub-sentence becomes the input for the dependency parsing. Dependency parsing information is then analyzed through coreNLP, extracting sentence/word id, lexicon, POS, and their relation, as shown in Table 4.

### *4.2. Structuring and extracting TechWords*

The TechWord candidates were organized based on the dependency relationship around the head word tagged with a noun, as shown in Table 5. A dependency word clarifies the technical concept of the head words. For example, the head word "ability" is extended to the "mold ability", "emergency stop ability", and so on, embodying the technical content and goal of the head word "ability." A final data frame includes sequential lexical information from a head word to dependent words, limited to three words or less. A total of 54,708 noun type of TechWords, including 1,666 unique head words, were extracted from 731,640 lexical data.

Considering the relations within the word dependency network, we conducted a network analysis to evaluate the technological words' importance. Both in-degree and out-degree centrality between nodes were calculated with their direction information. In the case of a noun with high in-degree centrality, it was widely used as a head word of various dependent words that could be interpreted as a basic form embodying the automotive field of the target technology. Superordinate concepts resulting in top words of in-degree centrality were derived in the following order: system, unit, device, assembly, portion, element, module, part, information, and apparatus. We ascertained that the technical concept in the derived word list could be interpreted as a technical basis in which automotive technology is embodied. In contrast, TechWord candidates that obtained high out-degree centrality could be interpreted as a subpart or object of function in the target technology. The vocabulary with high out-degree centrality that are valued in the automotive technology field contains vehicle, power, control, air, motor, side, battery, target, drive, wheel, seat, fuel, and tire. However, in the case of "vehicle", the centrality value of 0.1 could be interpreted as an outlier that was different more than twice from the second-highest value of 0.04, which could be classified as a representative word in the relevant technical field. Fig. 7 shows the word lists of the highest in-degree and out-degree centrality.

Tables 6 and 7 show that the vocabulary with a higher degree (both in-degree and out-degree centrality) presented the hierarchical results for the words associated with them. The top three words of the highest in-degree centrality—systems, units, and devices—are often classified as stop words in other text mining related to technology intelligence. In this study, we defined these words as the high-level of generic concepts in terms of technology, and based on this, the concept of automotive technology was embodied in their dependent words. A system resulted in the main base technology in the field of vehicle technologies, which consisted of a control system, brake system, vehicle system, management system, and so on. In all three cases, the word "control" had the highest frequency, which could be interpreted as the most important sub-technology in automotive fields. Since vocabularies with higher out-degrees can be interpreted as playing roles, such as technical functions and major parts, they are structured on the basis of noun-based vocabulary information. The main TechWords of power, control, and air among high out-degree noun vocabulary were structured through frequency-based multi-word forms. Power was deduced as the common

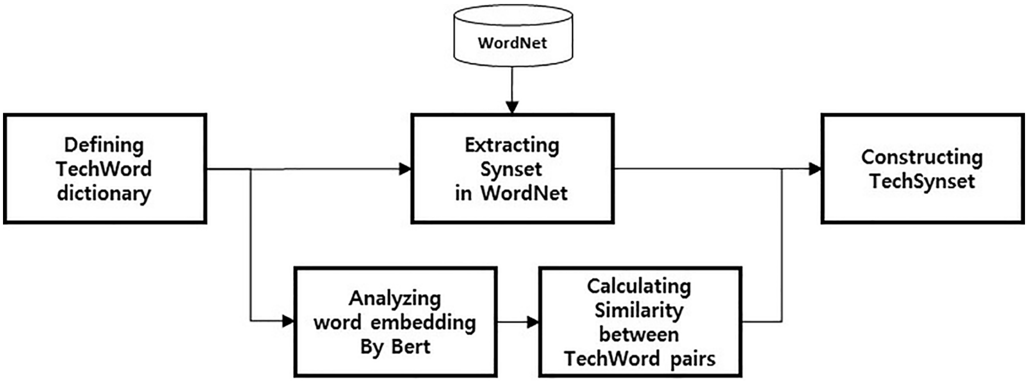
WordNet
Defining TechWord dictionary
Extracting Synset in WordNet
Constructing TechSynset
Analyzing word embedding By Bert
Calculating Similarity between TechWord pairs

**Fig. 6.** Extending process for construction TechSynset.

**Table 4**
Results of NLP and dependency parsing.

| | ID | | Word information | | | Dependency Relation | Head word information | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sentence | Word | Raw | Lemma | POS | | Id | Lemma | Raw word | POS |
| 0 | 1 | 1 | A | a | DT | det | 4 | system | system | NN |
| 1 | 1 | 2 | vehicular | vehicular | JJ | amod | 4 | system | system | NN |
| 2 | 1 | 3 | control | control | NN | compound | 4 | system | system | NN |
| 3 | 1 | 4 | system | system | NN | nsubj | 5 | include | includes | VBZ |
| 4 | 1 | 5 | includes | include | VBZ | ROOT | 0 | . | . | . |
| 5 | 1 | 6 | a | a | DT | det | 9 | camera | camera | NN |
| 6 | 1 | 7 | forward | forward | RB | advmod | 8 | view | viewing | VBG |
| 7 | 1 | 8 | viewing | view | VBG | amod | 9 | camera | camera | NN |
| 8 | 1 | 9 | camera | camera | NN | dobj | 5 | include | includes | VBZ |
| 9 | 1 | 10 | that | that | WDT | nsubj | 11 | view | views | VBZ |
| 10 | 1 | 11 | views | view | VBZ | acl:relcl | 9 | camera | camera | NN |
| 11 | 1 | 12 | forward | forward | RB | advmod | 11 | view | views | VBZ |
| 12 | 1 | 13 | through | through | IN | case | 16 | windshield | windshield | NN |
| 13 | 1 | 14 | the | the | DT | det | 16 | windshield | windshield | NN |
| 14 | 1 | 15 | vehicle | vehicle | NN | compound | 16 | windshield | windshield | NN |
| 15 | 1 | 16 | windshield | windshield | NN | nmod | 11 | view | views | VBZ |
| … | … | … | … | … | … | … | … | … | … | … |

**Table 5**
Noun type of TechWord candidate.

| Word | | | | Count | Part of speech | | |
|---|---|---|---|---|---|---|---|
| Depth-3 word | Depth-2 word | Depth-1 word | Head word | | Depth-3 | Depth-2 | Depth-1 |
| | | | ability | 58 | | | |
| | | mold | ability | 1 | | | N |
| | | storage | ability | 1 | | | N |
| | | transfer | ability | 1 | | | N |
| | emergency | stop | ability | 1 | | N | V |
| | energy | absorption | ability | 1 | | N | N |
| | road | hold | ability | 1 | | N | V |
| | | | ablation | 1 | | | |
| | | laser | ablation | 2 | | | N |
| | | | abnormality | 186 | | | |
| | | communication | abnormality | 6 | | | N |
| | | sticking | abnormality | 3 | | | N |
| … | … | … | … | … | … | | |

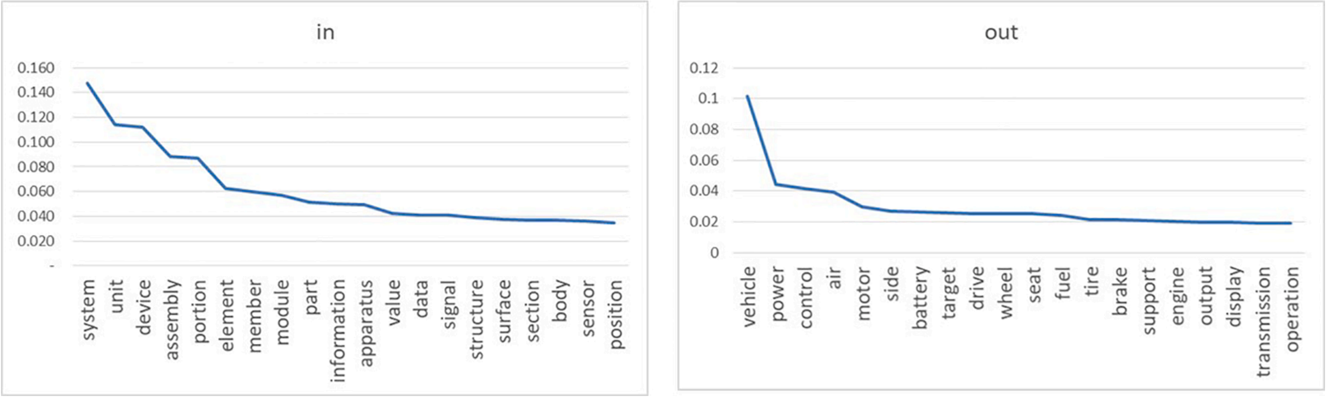

**Fig. 7.** Top 20-word list of the highest in and out-degree centrality.

part or sub-technology in the automotive field that can be linked to the concepts of supply, source, transmission, and so on.

The verb type of TechWord candidate was analyzed by SAO structures, where a phrase or sentence included verb (action) and the related noun words (subject or object), as shown in Table 8. Through the analysis process, 82,135 unique SAO structures were extracted from the patent sentences related to the automotive field, including 2,334 unique verbs. The noun type of TechWord, resulting from the previous step, was used to extend a single noun word to the phrase information based on the identical head noun of subject and object. For example, eight SAO structures were extracted from US1001545, which is entitled “vehicular control system”. A verb type of TechWord “capture” was interpreted as a technological function of “camera”, which targets “image data” through the SAO structure. Other technological relations and functions were acquired, such as “include”, “detect”, “determine”, “communicate”, “process”, and “control.” In addition, the SAO structure of “system” (S) – “include” (A) – “camera” (O) was extended to the “control system” (extended S) – “include” (A) – “camera” (O), obtaining the hierarchy-based lexical information on a single noun of subject and object.

Network centrality analysis was conducted for SAO relations to

**Table 6**
Top 3 words of in-degree centrality and their dependent word.

| Main word: “system” | | Main word: “unit” | | Main word: “device” | |
|---|---|---|---|---|---|
| Dependent word | Frequency | Dependent word | frequency | Dependent word | frequency |
| control | 1,497 | control | 2,615 | control | 1,336 |
| brake | 479 | detection | 396 | storage | 863 |
| vehicle | 424 | drive | 377 | display | 516 |
| management | 399 | storage | 303 | computing | 362 |
| drive | 341 | determination | 290 | input | 295 |
| suspension | 327 | display | 290 | communication | 286 |
| assistance | 209 | processing | 259 | lighting | 212 |
| storage | 196 | transmission | 176 | detection | 197 |
| lighting | 193 | output | 164 | processing | 175 |
| power | 187 | communication | 142 | switching | 169 |
| … | … | … | … | … | … |

**Table 7**
Top 3 words of out-degree centrality and their head word.

| Main word: “power” | | Main word: “control” | | Main word: “air” | |
|---|---|---|---|---|---|
| Head Word | Frequency | Head Word | Frequency | Head Word | Frequency |
| supply | 1,059 | unit | 2,615 | flow | 254 |
| source | 765 | system | 1,497 | conditioning | 242 |
| transmission | 509 | device | 1,336 | inlet | 183 |
| storage | 252 | signal | 598 | bag | 152 |
| system | 187 | module | 558 | pressure | 145 |
| converter | 206 | apparatus | 504 | intake | 140 |
| transfer | 194 | circuit | 338 | spring | 131 |
| generation | 152 | method | 283 | outlet | 129 |
| unit | 125 | valve | 246 | conditioner | 127 |
| control | 111 | mode | 163 | duct | 98 |
| … | … | … | … | … | … |

**Table 8**
An example of verb type TechWord candidate based on SAO structure (US10015452).

| No | Subject | Action | Object | Subject(extended) | Object(extended) | type |
|---|---|---|---|---|---|---|
| 1 | system | include | camera | control system | camera | a |
| 2 | camera | capture | data | camera | image data | s1 |
| 3 | | detect | characteristic | | road characteristic | s2 |
| 4 | | detect | vehicle | | vehicle | s2 |
| 5 | | determine | curvature | | road curvature | s2 |
| 6 | | communicate | data | | data | s2 |
| 7 | processor | process | data | image processor | image data | s1 |
| 8 | | control | speed | | speed | s2 |

ascribe the extent to which the verb type TechWord candidates were important in the network. The results are interpreted from a view of the entire technology level of automotive areas by considering every sentence in the whole document database. The in-degree and out-degree centralities of each verb word were calculated in the same way as a noun, as shown in Table 9. Verb-type words expressing the inclusion relationship, including “include”, “have”, and “comprise” ranked as the top in both in- and out-degree centralities. That is, the characteristics of a patent document, where the scope of the right should be described by clarifying the components for the subject matters of the invention. “Configure” showed high out-degree centrality, while in-degree centrality was very low. The meaning of “configure” indicates arranging something or changing the controls of a computer or other device, which could mean that the word focuses the other object that is going to be the target to arrange or change.

**Table 9**
Degree centrality of verb type TechWord.

| | Verb word | frequency | In-degree centrality | In-degree rank | Out-degree centrality | Out-degree rank |
|---|---|---|---|---|---|---|
| 0 | include | 40,971 | 0.1936416 | 1 | 0.2336594 | 1 |
| 1 | have | 11,692 | 0.1398399 | 2 | 0.1514006 | 2 |
| 2 | provide | 9,632 | 0.0538017 | 4 | 0.1462872 | 3 |
| 3 | configure | 6,010 | 0.0091152 | 114 | 0.1100489 | 4 |
| 4 | comprise | 5,381 | 0.0706981 | 3 | 0.0975989 | 5 |
| 5 | connect | 4,095 | 0.0266785 | 20 | 0.0795909 | 6 |
| 6 | form | 3,161 | 0.0424633 | 7 | 0.0695865 | 8 |
| 7 | determine | 2,928 | 0.018675 | 42 | 0.0631392 | 11 |
| 8 | detect | 2,785 | 0.0184526 | 43 | 0.0606936 | 12 |
| 9 | receive | 2,607 | 0.0346821 | 11 | 0.0455758 | 21 |

### 4.3. *TechSynset construction*

In this section, TechSynset was constructed to provide lexical information with the same and similar meanings between TechWords. First, synonym set (called synset) information from the WordNet DB was extracted for every single noun and verb type of TechWord. A Boolean matrix of TechWord by synset was created to indicate what synset contained for each TechWord, as shown in Table 10. To do this, we used the WordNet module from “nltk.corpus” in the Python package, which provided synset ID(e.g., car.n.01), concatenating word (e.g., car), POS (e.g., n; noun), and number (e.g., 01) by a dot. If a pair of synset IDs had different numbers with the same word and POS, this indicated that a homonym explaining a group of words was spelled in the same way, but

**Table 10**
Matrix between TechWord and Synset (WordNet).

| | load.n.01 | burden.n.01 | … | start.v.01 | begin.v.01 |
|---|---|---|---|---|---|
| load | 1 | 1 | … | 0 | 0 |
| burden | 1 | 1 | … | 0 | 0 |
| start | 0 | 0 | 0 | 1 | 1 |
| begin | 0 | 0 | 0 | 1 | 1 |
| battery | 0 | 0 | 0 | 0 | 0 |
| include | 0 | 0 | 0 | 0 | 0 |
| … | … | … | … | … | … |

had different meanings. We derived a matrix from a 9,501-row of TechWords and a 14,423-column of synset IDs. The related synset contained 7,593 (6,737 nouns and 856 verbs) units of unique lexical information, with an average of 1.91 synonyms per word. Among the nouns, the words "point" and "head" and the verb "clear" contained the most synsets—19 and 16, respectively.

However, the synset in WordNet was built from the view of general lexical information, lacking in its consideration of a specific technology area. Therefore, we searched for contextually similar words based on word embedding vectors using TechWord's BERT model. The original BERT model is a context-based word embedding model that takes a single sentence as input data. In this paper, all sentences in a single patent document were concatenated and used as input data, assuming that a single patent shared the same context. Pre-trained models with whole world masking, where the hyper parameters of 24-layer, 1024-hidden, 16-heads, and 340 M were applied, then adopted to build word vectors (https://github.com/google-research/bert, https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/). Hidden states are described in four dimensions: the tensor, which includes the layer number (12 layers); the batch number (1 sentence); the word/token number (number of words in every sentence); and the hidden unit (768 features).We obtained a 768-dimensional vector for each word by using an output of the last layer. From this, we derived a total of 812,319 context-based word vectors, of which the unique vocabulary was 3,848, meaning that a single word appeared on average in 211 patent documents. The word "vehicle" was the highest with 20,300 instances, followed by "one" (11,885), "first" (10,931), and "system" (10,352).

In WordNet, the relation between "word" and "synset ID" is represented by a one-to-multi relation, considering the multiple meanings of a single word. In the same vein as WordNet, all word vectors expressed in the same-spelled words were clustered when a word appeared in more than 100 patent documents, and are presentative word vector of each cluster was selected in consideration of polysemy. Since the word vector is high-dimensional data represented in 767 dimensions, we performed k-means clustering after the dimension reduction through principal component analysis (PCA), which is a general approach to high-dimensional clustering. In k-means clustering, the number of clusters was defined as the value in the range (1 to 5) at which the silhouette index, which refers to a method of interpreting and validating consistency within clusters of data, was the highest only when the index value was over 0.5. The final representative vectors were defined as vectors with the highest similarity values in all vector pairs in a vector cluster within the same word. Representative vectors were defined as those having the highest similarity values with all vector pairs in the vector cluster. In the case of additional vector operation, the deformation of the vector value could have a negative effect on the semantic analysis afterward. For this reason, a vector retaining the original value in every cluster was extracted as a representative without any further vector calculation.

To define the TechSynset, we extracted unique words used as a main word in the word-to-word network then calculated the cosine similarity between 2,788 word pairs, as shown in Table 11. The results of the BERT-based proposed model were 0.3137 on average, and 0.0817 on standard deviation, resembling a normal distribution, as shown in Fig. 8. A TechSynset is defined as a pair of words that are semantically similar with a cosine similarity above the threshold value. When the similarity threshold was adjusted in the range (0.6–0.9), we obtained 8,969 word pairs when the threshold was more than 0.6, 1,223 when it was more than 0.7, 174 when it was more than 0.8, and 3 when it was more than 0.9. A set of word pairs with similarity above the threshold could be defined as a TechSynset. Among the 174 pairs of words derived by the TechSynset with the 0.8 threshold, 55 word pairs excluding words with the same synset ID (manner.n.00, manner.n.01) appeared in WordNet. The results were verified based on the results from thesaurus.com, a free online thesaurus, which showed an accuracy of 74.54%, that is, 41 out of 55 word pairs. For the 41 correct synonym pairs, the semantic similarity in TechWord, WordNet, and TechNet was derived as 0.826, 0.275, and 0.601, respectively, showing a significantly higher value of TechSynset.

## 5. Discussion

This study presents a TechWord that considers the relations between lexical information that reflects the characteristics technology-describing vocabulary, and defines a TechSynset of semantically similar words. In this section, the validation and implication of our findings are presented from the following two perspectives. First, we discuss the implications by comparing the results with the existing vocabulary that reflects the lexical characteristics expressing technology in TechWord and TechSynset, which is presented as the core result of this

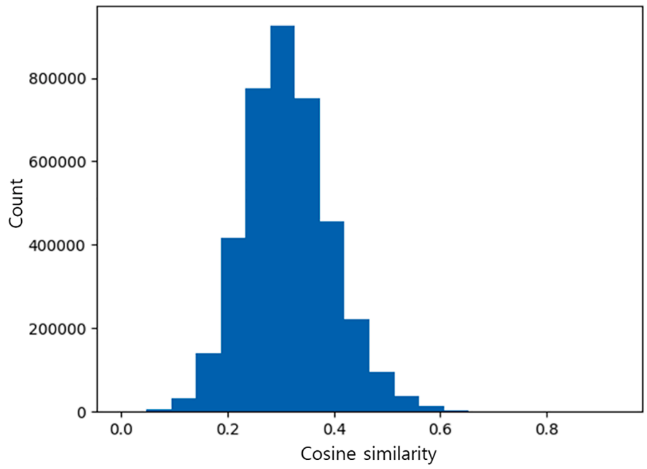

**Fig. 8.** Histogram of cosine similarity for pairs of TechWords.

**Table 11**
Cosine similarity between Techwords.

| | rope_0 (US9266490) | baseline_0 (US9809057) | du_0 (US10137733) | execution_0 (US9979238) | score_0 (US10286915) | … |
|---|---|---|---|---|---|---|
| rope_0 (US9266490) | 1 | 0.303828 | 0.220537 | 0.282456 | 0.355585 | |
| baseline_0 (US9809057) | 0.303828 | 1 | 0.209353 | 0.369852 | 0.524569 | |
| du_0 (US10137733) | 0.220537 | 0.209353 | 1 | 0.250477 | 0.230417 | |
| execution_0 (US9979238) | 0.282456 | 0.369852 | 0.250477 | 1 | 0.342545 | |
| score_0 (US10286915) | 0.355585 | 0.524569 | 0.230417 | 0.342545 | 1 | |
| … | | | | | | … |

study. Second, the performance in text mining is improved compared to the existing approach through the vocabulary list created through TechWord.

### 5.1. Considering lexical characteristics of technical text documents

We compared our findings with WordNet, a database for general vocabulary targets, to validate that the technology-related lexical characteristics are reflected. We compared the results of TechWord, which were developed in this paper by considering the characteristics of technical vocabulary, with the existing WordNet results. The coverage of technology-related words was compared by considering it from a single word to compound words. Based on the lexical data list constructed in this study, each word was checked to examine if WordNet contained the corresponding synset information. A total of 1,494 words were excluded from the synset database in WordNet, which is 24% of the entire 6,278 single lexical lists. Among them, there were 171 words used more than 10 times in the patent document, and they consisted of words related to specific parts and functions, including “airbag”, “seatback”, “powertrain”, “evaporator”, “liftgate”, and “backlight vehicle”, as well as acronyms such as “ECU”, “SOC”, and “HVAC”. The technical vocabulary characteristics were considered in terms of the noun expressed by the hierarchical description to specify the technological subject, which used the form of the phrase connected with other nouns or adjectives. In addition to covering words not included by WordNet, we were able to overcome the limitation of not being able to interpret these structural meanings. From the verb’s perspective, technology-related documents are important to express the technology’s purpose and function. This is represented by the relationship between nouns and verbs, which WordNet does not consider. In this study, we structured this through the SAO structure.

In addition, to verify the proposed process to construct the TechSynset, the second major result, we did a comparison with the existing related studies. In this study, we suggest a novel approach that calculates the similarity between TechWords through a state-of-the-art language model based on BERT. Various metrics using WordNet-based semantic similarities between synsets have been developed by using its hypernym tree structures, which are stored in the Python package “nltk”, such as “wn.path_similarity”, “wn.wup_similarity”, and “wn.lch_similarity”. However, this process derives a huge variation of values, since these metrics are dependent on the structure of WordNet. For example, if the relation information, such as hypernym and hyponym, between a pair of synsets in WordNet is stored, the similarity is quite high; otherwise, it may be very low. In addition, several existing studies analyzed semantic similarity through word vectorization from the conventional co-occurrence approach to advanced language learning models, such as word2vec. The previous approaches have mostly shown context-free static embedding, which has difficulties when utilizing cosine similarities by using the vector itself. On the other hand, BERT dynamically learns the same words by using different vectors to reflect the context of each sentence (Torres, Gutierrez, & Bucheli, 2019). TechNet, a large-scale comprehensive semantic network of technology-related data, was constructed by associating word2vec-based vectorized terms through cosine similarities (Sarica et al., 2020). We compared the similarity between TechSynsets derived from this study with WordNet-based path similarity and “relevance term,” among the functions in word2vec-based TechNet (http://www.tech-net.org/), as shown in Table 12. Seven word pairs were randomly selected with consideration of words that are common, related to cars, and that have special issues. Similarities between “unit” and “element” were 0.726, 0.563, and 0.091 in TechSynset, TechNet, and WordNet, respectively. One of the parts of the automotive “bonnet” and “hood” showed similarities 0.736, 0.436, and 0.067, respectively, which is a considerable disparity between figures. In general, the value of a BERT-based similarity better reflects the similarity between similar words. However, in the case of language model-based, there was a limit in that the similarity between the word pair “adult” and “infant,” which is an opposite language, also measured high.

**Table 12**
Examples of cosine similarity between word pairs by TechSynset, TechNet, and WordNet.

| Word pair | | Semantic similarity | | |
|---|---|---|---|---|
| Word 1 | Word 2 | TechSynset | TechNet | WordNet |
| method | technique | 0.802 | 0.675 | 0.500 |
| unit | element | 0.726 | 0.563 | 0.091 |
| join | connect | 0.731 | 0.508 | 0.333 |
| reinforce | strengthen | 0.799 | 0.634 | 0.500 |
| bonnet | hood | 0.736 | 0.436 | 0.067 |
| winding | coil | 0.801 | 0.634 | 0.063 |
| adult | infant | 0.707 | 0.723 | 0.167 |

### 5.2. Improving the performance of text mining through TechWord

In the existing text mining research, not all words of the corpus were analyzed, rather only words with the highest frequency were selected and analyzed due to limitations such as analysis time and capacity (Jang et al., 2017; Yoon & Park, 2004). The case of selecting the top 20% based on frequency, which is a general approach of the existing research, was set as the target for comparison (Roh et al., 2017). Compared with existing text mining, we analyzed the coverage of words selected for analysis by each patent document and whether the main words were selected. For the 34,823 patent corpuses collected for analysis in Section 4, we analyzed the number of words that could be analyzed for each document and the average of the TF-IDF values of the words selected as those that could be analyzed. In selecting a vocabulary set in the patent corpus to apply in this study, a total of 4,633 unique words were selected for the top 20% of frequent words as based on the existing research approach, and 13,356 words selected for TechWords as suggested in this study, were extracted based on noun phrases and the SAO structure. TechWord did not select a vocabulary dictionary with a fixed ratio and number, but expanded into noun phrases based on nouns used as head words in sentences and into verbs and nouns connected to them within the SAO structure. This TechWord-created dictionary is characterized by expanding into a vocabulary that is meaningful for text analysis of technological documents. For the two vocabulary lists to be compared, the average number of words included for analysis in one patent document (abstract) was 10.69514 for the existing method and 19.99463 for TechWord, as shown in Table 13. The coverage of the vocabulary derived through TechWord proposed in this study was higher, which can be explained as being influenced by how the number of words defined through TechWord was higher. In addition to the average number of words, TechWord demonstrated a lower variance, confirming that the number of analyzed words for a single patent document was more evenly distributed.

By applying TF-IDF, the most basic though major methodology of text mining, the results were compared between the existing approach and the TechWord presented in this study. TF-IDF analysis results are derived as importance index values for each word for each document. In this study, for comparison at the corpus level, the TF-IDF average value of word $j$ in the vocabulary list, in which word $j$ appears for document ID$i$, was analyzed based on the $TF-IDF_{ij}$ derived for each word $j$ in the document $i$ as shown in Equation (1). The average TF-IDF value at the

**Table 13**
Comparison of top 20% words in frequency and TechWords.

| Comparison | Existing approach<br>Top 20% of frequency) | TechWords |
|---|---|---|
| Word coverage | 10.69514 words/doc. | 19.99463 words/doc. |
| (standard deviation) | (6.562973) | (6.029417) |
| Average TF-IDF | 0.000161995043032143 | 0.000608951417987657 |

corpus level used as an evaluation index in this study represents the importance of how much word information included in the vocabulary list for analysis represents a document. That is, the higher the corpus level average TF-IDF value, the more importance the words have in the document that are being used for analysis, which is linked to the performance of text mining. As a result of the analysis, the TechWord proposed in this study was 3.76 times higher than that of the conventional approach as shown in Table 13. In conclusion, the performance of text mining was improved through the vocabulary list derived through TechWord.

$$\text{average TF}-\text{IDF(corpus level)} = \frac{\sum_j TF-IDF_{ij}}{number\ of\ document\ that\ contains\ word\ j} \tag{1}$$

## 6. Conclusion

This study suggests a novel methodology for structuring technology-related terms and designing a lexical database to overcome the limitations of the existing WordNet DB for text analysis in technology intelligence. Grammatical information of dependency parsing is structured using the abstract and claim documents, which are text-based data fields of the patent for the preprocessing step. The dependency relation between words is interpreted from the network point of view, and the vocabulary with high network degree centrality is structured as the main TechWord. The noun type of the TechWord is considered through the modifier phrase's relation, and in the case of the verb, the characteristics of the technical information are considered by reflecting the relation of the noun with the subject and the object. Synset information is imported from within WordNet to configure the TechSynset, a synonym set between TechWords. To construct the TechSynset, which WordNet does not include, we derive the word vector of TechWords through the BERT model, and then extend TechSynset by calculating the similarity between vectors of TechWord.

From an academic and practical perspective, this study makes the following contributions. First, as an academic contribution, we have made the structure of lexical information from the text mining perspective of technology-related documents. By analyzing the grammatical structure using the patent's text document, we systematically derived the TechWord. Moreover, we derived a similar word set of TechSynset through the existing WordNet DB and the latest word embedding model. Developing a novel methodology for structuring technical vocabulary will provide practical implications for the text analysis of technology-related information, such as patents. When defining vocabulary during the text mining process, lexical information can be defined in a structured method based on the TechWord, and it will be possible to preprocess synonyms based on the TechSynset. As the TechWord considers the structure and relation of words and the TechSynset through the word embedding technique reflects the context of the document, it is possible to expect the different meanings of the application domain of the technology to be reflected.

However, this study has several limitations. This study does not cover all parts of speech in defining the TechWord, but extracts the TechWord from the perspective of nouns and verbs that are mainly used in technology-related document analysis. That is, additional work on the other parts of speech, such as adverbs, prepositions, and conjunctions, needs to be linked. In addition, in analyzing the technology-related document information, we do not deal with all the lexical information included in the existing WordNet, such as the relationship between words. Further research is needed to develop a methodology for analyzing the relationship information between words.

## CRediT authorship contribution statement

**Hyejin Jang:** Conceptualization, Methodology, Data curation, Writing - original draft. **Yujin Jeong:** Data curation, Validation, Writing - original draft. **Byungun Yoon:** Conceptualization, Methodology, Supervision, Writing - review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This work was supported by National Research Foundation of Korea (2019R1A2C1085388).

## References

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information, 37*, 3–13.

An, J., Kim, K., Mortara, L., & Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. *Journal of Informetrics, 12*(1), 217–236.

Barbu, E. (2015). Property type distribution in Wordnet, corpora and Wikipedia. *Expert Systems with Applications, 42*(7), 3501–3507.

Bonino, D., Ciaramella, A., & Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information, 32*(1), 30–38.

Brügmann, S., Bouayad-Agha, N., Burga, A., Carrascosa, S., Ciaramella, A., Ciaramella, M., … Mille, S. (2015). Towards content-oriented patent document processing: Intelligent patent analysis and summarization. *World Patent Information, 40*, 30–42.

Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications, 39*(13), 11443–11455.

Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C.-H. (2011). SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics, 88*(3), 863.

Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. Paper presented at the Advances in neural information processing systems.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Dewulf, S. (2011). Directed variation of properties for new or improved function product DNA–A base for connect and develop. *Procedia Engineering, 9*, 646–652.

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Paper presented at the LREC*.

Geum, Y., & Park, Y. (2016). How to generate creative ideas for innovation: A hybrid approach of WordNet and morphological analysis. *Technological forecasting and social change, 111*, 176–187.

Guo, J., Wang, X., Li, Q., & Zhu, D. (2016). Subject–action–object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change, 105*, 27–40.

Jang, H., Roh, T., & Yoon, B. (2017). User needs-based technology opportunities in heterogeneous fields using opinion mining and patent analysis. *Journal of Korean Institute of Industrial Engineers, 43*(1), 39–48.

Joung, J., & Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological forecasting and social change, 114*, 281–292.

Kim, K., Park, K., & Lee, S. (2019). Investigating technology opportunities: The use of SAOx analysis. *Scientometrics, 118*(1), 45–70.

Kwon, H., Kim, J., & Park, Y. (2017). Applying LSA text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation, 60*, 15–28.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6–7), 481–497.

Ma, M., Huang, L., Xiang, B., & Zhou, B. (2015). Dependency-based convolutional neural networks for sentence embedding. arXiv preprint arXiv:1507.01839.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Paper presented at the Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*.

Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Paper presented at the Advances in neural information processing systems*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Moawad, I. F., & Aref, M. (2012). Semantic graph reduction approach for abstractive Text Summarization. *Paper presented at the 2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*.

Moehrle, M. (2010). Measures for textual patent similarities: A guided way to select appropriate approaches. *Scientometrics, 85*(1), 95–109.

Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications, 40*(7), 2373–2390.

Park, H., Yoon, J., & Kim, K. (2013). Using function-based patent analysis to identify potential application areas of technology for technology transfer. *Expert Systems with Applications, 40*(13), 5260–5265.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Rachabathuni, P. K. (2017). A survey on abstractive summarization techniques. *Paper presented at the 2017 International Conference on Inventive Computing and Informatics (ICICI)*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.

Roh, T., Jeong, Y., Jang, H., & Yoon, B. (2019). Technology opportunity discovery by structuring user needs based on natural language processing and machine learning. *PLoS ONE, 14*(10).

Roh, T., Jeong, Y., & Yoon, B. (2017). Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing. *Sustainability, 9*(11), 2117.

Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. *Expert Systems with Applications, 142*, Article 112995.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(3), 203–217.

Torres, J. A. P., Gutierrez, R. E., & Bucheli, V. A. (2019). The performance evaluation of Multi-representation in the Deep Learning models for Relation Extraction Task. arXiv preprint arXiv:1912.08290.

Verhaegen, P.-A., D'hondt, J., Vertommen, J., Dewulf, S., & Duflou, J. R. (2009). Relating properties and functions from patents to TRIZ trends. *CIRP Journal of Manufacturing Science and Technology, 1*(3), 126–130.

Wang, X., Ren, H., Chen, Y., Liu, Y., Qiao, Y., & Huang, Y. (2019). Measuring patent similarity with SAO semantic analysis. *Scientometrics*, 1–23.

Wang, X., Wang, Z., Huang, Y., Liu, Y., Zhang, J., Heng, X., & Zhu, D. (2017). Identifying R&D partners through Subject-Action-Object semantic analysis in a problem & solution pattern. *Technology Analysis & Strategic Management, 29*(10), 1167–1180.

Wanner, L., Baeza-Yates, R., Brügmann, S., Codina, J., Diallo, B., Escorsa, E., … Pianta, E. (2008). Towards content-oriented patent document processing. *World Patent Information, 30*(1), 21–33.

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42*(4), 2264–2275.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics*.

Yang, C., Huang, C., & Su, J. (2018). An improved SAO network-based method for technology trend analysis: A case study of graphene. *Journal of Informetrics, 12*(1), 271–286.

Yang, S.-Y., & Soo, V.-W. (2012). Extract conceptual graphs from plain texts in patent claims. *Engineering Applications of Artificial Intelligence, 25*(4), 874–887.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research, 15*(1), 37–50.

Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change, 72*(2), 145–160.

Yoon, J., & Kim, K. (2012a). An analysis of property–function based patent networks for strategic R&D planning in fast-moving industries: The case of silicon-based thin film solar cells. *Expert Systems with Applications, 39*(9), 7709–7717.

Yoon, J., & Kim, K. (2012b). TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications, 39*(3), 2927–2938.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Paper presented at the Proceedings of the IEEE international conference on computer vision*.