Transdisciplinary Engineering for Resilience: Responding to System Disruptions
L. Newnes et al. (Eds.)

doi:10.3233/ATDE210090

Technology Mining for Intelligent Chatbot Development

Min-Hua CHAO[1], Amy J.C. TRAPPEY, Chun-Ting WU and Yi-An SU
Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan

Abstract. Natural language processing (NLP) is an indispensable part of advancing the AI era, especially in the realm of the human-computer interface/interaction (HCI) for all state-of-the-art software applications. NLP enables interfaces between machines and humans allowing machines/computers/systems to understand human languages and engaging in dialogues. An intelligent chatbot development must incorporate NLP technologies to allow the understanding of users' utterance and responding in understandable sentences in versatile scenarios. This research investigates the emerging technological trend of intelligent chatbot development. The systematic trend analysis is described in the research. First, patents related to intelligent chatbot domain are retrieved using a well-defined search query. The queries are derived from the knowledge ontology, which is extracted using text-mining algorithms - key term frequency analysis, clustering for sub-domain identification, and Latent Dirichlet Allocation (LDA) for topic modelling. Afterwards, the management and technology maps of a patent portfolio, such as patenting trends and technology function matrix, are extracted and drawn. The technology trend analysis also investigated the distributions of the relevant patent claims for specific industries.

Keywords: Natural language processing (NLP), chatbot; patent analysis, Latent Dirichlet Allocation (LDA), ontology, text mining, transdisciplinary

Introduction

In the context of the global fight against epidemics, when the communication between people is restricted, artificial intelligence (AI) has got more expectations and important tasks. It has demonstrated its skills in the fields of information collection, data aggregation and real-time updates, epidemiological investigations, vaccine drug development, and new infrastructure construction. At the same time, with the continuous emergence of new technologies and new formats, the power of AI to condense global wisdom and help global economic recovery has become more prominent. AI is extremely practical and a very representative multidisciplinary subject. At present, AI has been applied to various fields such as machinery, electronics, economy and even philosophy. The global industry is digitizing. As the first window of customer service, chatbots can handle simple customer questions in real time. However, the development of AI has made chatbots more intelligent, so chatbots can replace junior employees. Intelligent chatbots mean that they can understand/process natural language. It is a transdisciplinary engineering technology that integrates multiple disciplines to enable chatbots to simulate human behavior. Therefore, NLP is the key to intelligent chatbots. By collecting a large number of conversations with customers and obtaining key information, the company

[1] Corresponding author, Mail: s106034803@m106.nthu.edu.tw.

can enhance customer relationship management and make customer preference predictions. There are countless examples of applying NLP in various fields, such as the application in the medical industry. Using NLP technology to analyze the elderly who narrate specific events, which can diagnose the level of cognitive impairment in patients with Alzheimer's disease [1]. In some applications of administrative affairs automation, NLP is also a helpful tool. Google incorporates NLP technology and optical character recognition to identify key parameters of documents, thereby saving time-consuming manpower reading work [2]. As mentioned previously, NLP is the most widely used in customer service. Customer's emotions and intentions can be extracted by analyzing a large number of customer feedback of NLP, and used to improve products or as a strategy for developing new products [3].

Chatbot is a virtual system composed of functional programs. It is known as a conversation agent that simulates human thinking and responds. To determine whether a chatbot is intelligent, the Turing test can be used, which is an experiment to judge whether a machine has the ability to think [4]. In the beginning, chatbots could only respond according to the rules defined by the developer. With the advent of deep learning, the process design of making chatbots respond is no longer just based on rules, but to learn human communication logic by reading a lot of conversations. Since then, chatbots have begun to incorporate voice recognition functions to make them more intelligent. Nowadays, people can operate machines or make automatic changes through chatbots, and this technology is widely used in car assistant and robots.

This research uses the Derwent Innovation patent index (DWPI) to retrieve patents related to intelligent chatbot technology and provides an ontology map and patent portfolio analysis of the subject to understand the development trend of intelligent chatbots and the current market layout. In addition, this research provides strategies for formulating technological development in related fields. First, an overall ontology map is required, which is followed by a well patent analysis strategy. Also, several text mining techniques are adopted in the ontology construction process, such as k-means clustering [5], LDA [6] and term frequency-inverse document frequency (TF-IDF) [7].

1. Literature review

In this section, the recent patent review workflow is discussed in the first paragraph, which aims to develop a reliable patent review process for this research. In order to be able to analyze massive related patents, some powerful patent analysis methods are discussed in the second paragraph. Last, an ontology construction is discussed, which can organize patent review findings into a systematic knowledge framework.

Patent contains an ample amount of textual content and attribute data. How to effectively integrate these data into useful information is the focus of a patent review. However, the development of information technology has made patent review easy and reliable [8]. For instance, text mining is a technique widely adopted in patent review, which provides patent reviewers with specific information, such as key terms and topic distribution. Nowadays, many powerful information technologies are used to assist in a patent review. However, if a proper patent review process is not established, it may lead to the possibility of deviation from the research subject. Abbas, Zhang [9] proposed an overview of the patent review workflow, which contains three main sections, pre-processing, processing and post-processing. The key to preprocessing is to search for appropriate patents and convert unstructured data into usable information. Afterwards,

extracting specific quantitative or qualitative data is the purpose of the processing section, such as topic modeling and outputting some statistical information. Finally, patent analytics methods are related to the past-processing section, which is classified into two patterns. Text mining-based methods are for quantitative data, and visualization approaches are for qualitative data [10]. Kim and Bae [11] proposed an approach that can predict emerging medical technologies through patent analytics. Their patent review workflow can be divided into four processes: retrieving domain patents, technology clustering, defining the output of the technology clustering, and patent clustering evaluation. In the third process, the results may vary from analyst to analyst. To avoid this problem, Cooperative Patent Classification (CPC) is used to define the output of the technology clusters. Both patent and non-patent literature are available resources for exploring emerging technologies. Thilakaratne, Falkner [12] proposed a literature-based research review workflow, which formulates a literature retrieval process in detail to avoid missing any representative literature. The main research purpose and keywords are the criteria for judging whether the article is suitable for being placed into the database. After the resources are collected, all literature needs to be further filtered through three processes. The first is to analyze the title and abstract to check the relevance, the second is to analyze the introduction and conclusions of the article, and finally, to fill in the quality checklist through a complete reading. Afterwards, some graphs were constructed through visualization techniques to present the findings. In summary, the knowledge document review workflow can be divided into three stages: resource acquisition, knowledge document analysis in a database, and output presentation.

Since patent contains information that has characteristics of volume and variety, therefore, how to filter information becomes a big issue of patent analysis, which is a method that attempts to gather information from a structured patent document. The realm of patent analysis has two major methods, which are text mining [13] and visualization [14]. When it comes to the text mining approach for text mining, the main ways of representing a patent are semantic information identification and semantic similarity comparison [15]. By way of illustration, first, Hu, Li [16] made use of the characteristic of the frequency of appearance to extract key terms from patents and compare them with TF-IDF method. Second, Li, Hu [17] utilized a deep learning model called DeepPatent to classify patents, the ensemble model merges CNN model and the word embedding model. Third, Lee and Hsiang [18] wanted to gain a better result, they fine-tuned a BERT model and compared it with DeepPatent consequently. The final result had higher precision by 9 percent. As the above three ways mentioned, they all use the common text mining wat by selecting keywords and converting them into machine-readable vectors. Compared with text mining, visualization approaches for patent analysis have progressed early. Take three major development of visualization approaches for example, ontology map focuses on domain-related knowledge discerption, K-means focuses on topic modeling [19], and technology function matrix focuses on tracking status [20].

An ontology graph is a tool of knowledge engineering, which center is the research theme and links related fields from the center outward, provides a clear domain knowledge classification. In this research, ontology is used to construct a logical criterion for classifying a technology and show are core techniques of the domain. Ontology construction must rely on sufficient data and some text mining for extracting key terms and topic modeling. Weng, Tsai [21] purposed a lexicon-based ontology construction approach that utilizes term frequency and weighted factor to define the relationship between key terms and research theme. Trappey, Trappey [22] presented an ontology construction approach which is knowledge-based and utilized an unsupervised machine

learning technique to extract the information in the smart retailing industry for chasing emerging technologies and trend. For constructing a complete ontology map, some algorithms are applied to continuously refine the ontology, such as LDA and clustering. Tsatsou, Davis [23] presented an automatically constructing ontology method, which utilized TF-IDF to determine key terms that may be branches or nodes of the ontology. Subhashini and Akilandeswari [24] mentioned that constructing an ontology is required to follow the six key steps: determining the scope of the ontology, capturing related data, encoding those useful data to machine-usable, integrating the results, evaluating the results, and documenting the ontology. In summary, constructing an ontology can mainly be divided into three parts: data source, determining the relationship between terms and effectiveness evaluation.

2. Ontology construction using patent

A patent-based ontology, which presents the knowledge connection that contains sub-domains or key terms for a specific topic, provides a way for non-experts to can quickly understand the topic. In this research, the ontology construction process contains four-stage. Stage 1 is patent retrieval, which aims to find the patents related to an intelligent chatbot. This research selects DWPI as the source for searching patents. At stage 1, search some keywords related to the topic and the most relevant 50 patents of the search result be manually checked whether related to the topic. If not, adjust the keywords and re-query until the results are related to the topic. Table 1 lists the final query condition and the amount of the relevant patent.

Table 1 Query condition for ontology construction

Derwent innovation query keywords	Result
"natural language processing “&”natural language understanding" & "NLP" & "NLU" & "chatbot" & "VIRTUAL ASSISTANT"&"INTELLIGENT ASSISTANT" & "automated conversational interface"	508 patent families

Stage 2 is patent clustering and main sub-topic selection, which aims to find the main domains related to the topic. At stage 2, the patents’ abstract and claim are used for k-means clustering. Using k-means clustering requires inputting the number of clusters. In order to have the best result, this research utilizes the Silhouette score for determining the number of clusters. On the other hand, TF-IDF is used to identify keywords or key terms and the results of TF-IDF are utilized to check the result of k-means clustering whether related to the topic. If not, back to stage 1 and re-query.

Stage 3 is topic modelling. In order to understand more detail in each sub-topic, the process of stage 3 is further classifying the result of stage 2 into several subdomains. The previous stage has identified several sub-topics related to the topic. At stage 3, LDA is used to do topic modelling for each sub-topic. For each sub-topic, first is setting respective search conditions which similar to stage 1, and then doing LDA topic modelling. After that, define each topic-word and if can’t clearly define the topic-word then ignore the topic. The distribution of the ontology has been almost completed at this stage. The last stage is keywords and key phrases finding. Representative keywords and key phrases will be selected from stage 2 and stage 3 to strengthen the description of

topic modelling results.

Figure 1 shows the ontology of this research. Based on the query condition listed in Table 1, 508 most relevant patents are found and used to construct the ontology. After doing patent clustering which is stage 2, 13 clusters are classified and summarized into three sub-topics, which are natural language techniques, model and system. Level 3 information connected with the sub-topics are the results of stage 3. For instance, in Nature language techniques sub-topic, 4 main sub-domains are found, voice control, linguistics, dialogue and knowledge. This research figures out the keywords or key phrases for each category classified by LDA. After that, generate a topic-words for each category based on the keywords found. Last, level 4 information which are the leaves of the ontology are set based on the frequency of appearance and the importance.

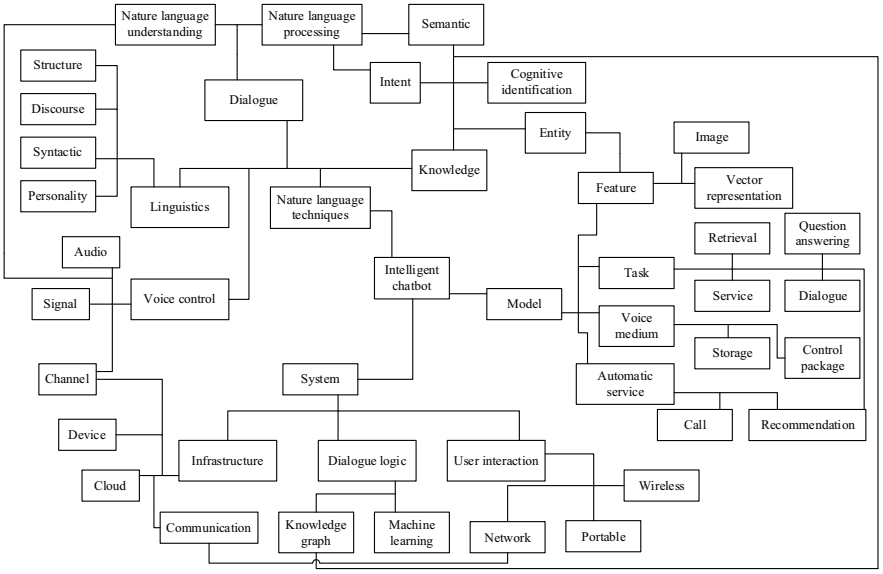

Figure 1. Intelligent chatbot ontology.

3. Patent portfolio analysis

A patent portfolio is a patent combination that related to a specific subject, used to analyze the market outlook and investigate the value of a potential market. In terms of chasing emerging technologies, a patent portfolio is an efficient and valuable tool for knowledge mining. In this section, the patent portfolio of intelligent chatbot is discussed, which presents technology hotspots through a layout of emerging technologies. First, use DWPI smart search to retrieve patents related to an intelligent chatbot. **Table 2** lists the query condition and 5,638 patents are found.

Figure 2 shows the trend of intelligent chatbot publishing patents, which presents the publishing patents are concentrated to 2020 and illustrates the topic is recently developed technology. Since avoiding selecting outdated patents, this research narrows down the publishing date to 2020 (5,390 patents) for follow-up analysis.

Table 2. Query condition for patent portfolio analytics.

Query keywords	Publication year
"intelligent" & "Natural language processing" & "Natural language" & "Natural language understanding" & "NLP" & "NLU" & "chatbot"	2010 to 2020

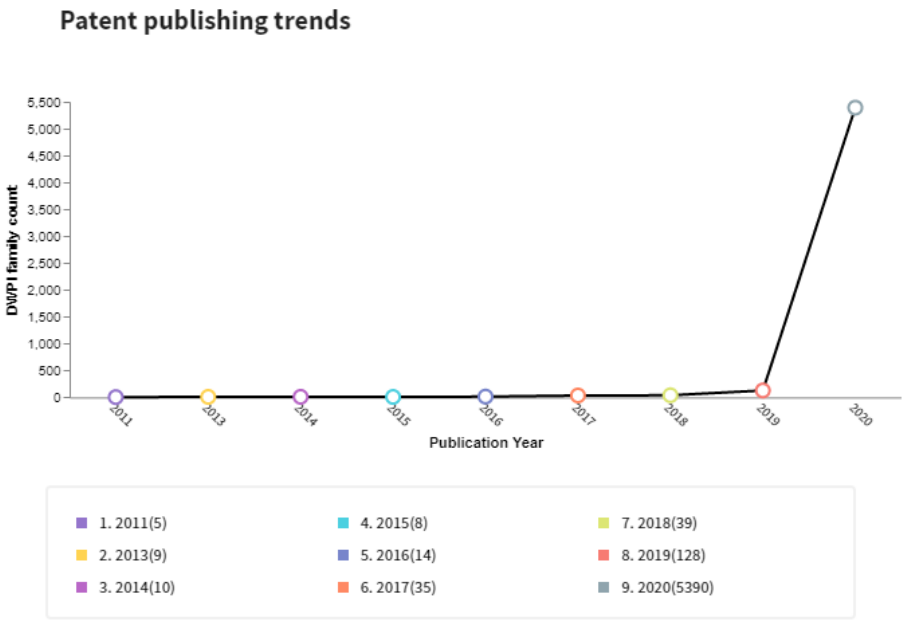

Figure 2. Patent publishing trends.

A technology function matrix (TFM) is a 2-dimensional matrix, which presents the distribution of patents by placing patents in the cells corresponding to defined technology and function. Technologies and functions of TFM can be defined according to the clustering result found in k-means before. After defining technologies and functions, each patent is reviewed iteratively to determine the belonging technology and function. According the patent distribution of TFM, analysts can quickly understand which technology is applied to achieve which functions in a specific field. Furthermore, It represents the current patent layout of technology in a specific industry, by which companies can avoid hot areas of technology or deploy areas where technology is not yet mature.

Technology definition. International Patent Classification (IPC) is a standard taxonomy developed and administered by the World Intellectual Property Organization (WIPO) for classifying patents and patent applications, which covers all areas of technology and is currently used by the industrial property offices around the world. Table 3 lists Top 10 IPCs of 5,390 patents found in DWPI. In terms of IPCs, technology related to an intelligent chatbot can be classified into five categories, which are the medium of communication, Natural language processing, intention recognition, language model, Interface.

Table 3. Top 10 IPCs.

Top IPC	Keywords
G06N Computing	3/08 Learning methods 3/04 Architecture 20/00 Machine learning
G06F Electric digital data processing	40/30 Unsupervised data analysis 16/33 Querying 16/332 Query formulation 16/35 Clustering; Classification
G10L Speech Recognition	15/22 Procedures used during a speech recognition process 15/18 using natural language modelling
G06K Recognition of data	9/62 Methods or arrangements for recognition using electronic means

Combining IPCs results and k-means results to arrive at technologies of TFM which lists in **Table 4**, 7 technologies are defined, speech recognition, natural language processing, feature engineering, machine learning, transformer, cloud computing and immersive technologies. Speech recognition is the technology of enabling a chatbot to process human sounds which advances the user experience of using chatbot. Feature engineering is the technology of utilizing domain knowledge to extract features which helps the process of building the knowledge base of the chatbot. Transformer is a sequence to sequence language model which is a popular language model for developing intelligent chatbot. The quality of an intelligent chatbot is based on the knowledge base and response mechanism, but a good enough knowledge base and response mechanism are a burden for the device. Through the cloud computing technology can reduce the memory and computing power of the device. Incorporating immersive technologies into chatbots can advance the user experience.

Table 4. Technologies of TFM.

#	Technology
T1	Speech recognition
T2	Natural language processing
T3	Feature engineering
T4	Machine learning
T5	Transformer
T6	Cloud computing
T7	Immersive technologies

Function definition. 6 functions listed in Table 5 are defined based on the results of ontology construction stage 4, which are natural language understanding, system efficiency, conversation, prediction, user experience and personal assistant.

Table 5. Functions of TFM.

#	Function
F1	Natural language understanding
F2	System efficiency
F3	Conversation
F4	Prediction
F5	user experience
F6	Personal assistant

Patent mapping. As mentioned previously, 5,390 patents published in 2020 are as the source for the TFM construction. First, some representative descriptions for defined technologies functions are collected from Wikipedia and other websites. Then, Organize the abstract, description and claims of all patents. Last, each patent is cross-compared and assigned to suitable cells which means the patent related to the technology and the function simultaneously. The result of TFM is listed in **Table 6**, which composed of 7 technologies and 6 functions. The area marked in red is the first third, and the blue area is the back third. From the results of TFM, it can be seen that the current patents related to an intelligent chatbot are focused on speech recognition and natural language processing. The core of an intelligent chatbot lies in the ability to understand natural language, so natural language processing, as the main technology, is also a hot spot. However, an intelligent chatbot usually provides multiple communication media, such as SIRI. Speech recognition technology plays a major role. It is a technology hotspot in the current market that transforms audio into machine readable information so that humans and machines can interact. In terms of the blue area. The current patents are relatively irrelevant in cloud computing and immersive technology, and they can be used as future development goals. Cloud computing allows a large amount of calculation to be performed, which enables some simple portable devices to have powerful functions. The immersive technology provides virtual and real image overlay, immersive environment, etc. to greatly enhance the user experience. In terms of function, user experience and personal assistant are notable potential markets. A well user experience can lay the company's image and consolidate its position in the market. Personal assistants are the future trend, such as smart assistants for cars or smart assistants for homes, to change people's lifestyles.

Table 6. The TFM result

	F1 Natural language understanding	F2 System efficiency	F3 Conversation	F4 Prediction	F5 User experience	F6 Personal assistant
T1 Speech recognition	2,233	3,036	3,967	3,566	1,934	2,051
T2 Natural language processing	1,798	1,911	1,419	1,844	627	460
T3 Feature engineering	1,519	2,033	1,059	1,527	617	407
T4 Machine learning	886	1,111	976	1,603	208	305
T5 Transformer	1,048	1,136	1,303	1,862	452	448

T6 Cloud computing	341	678	909	595	452	458
T7 Immersive technologies	992	1,220	2,517	1,504	1,442	1,270

4. Conclusion

This research proposes a macro patent analysis on the newest technologies of an intelligent chatbot. A systematic ontology construction workflow is defined, which utilizes some text mining technologies, such as k-means clustering, TF-ITF, and LDA. After that, the four-level hierarchical structure of the ontology is constructed. The ontology map can be used as the basis for strategic and sustainable R&D planning, from which researchers are able to quickly understand the related key technologies and can determine technology gaps.

This research uses TFM analysis to divide chatbots into 7 technologies and 6 functions. According to the analysis result of TFM, the main patent layout of chatbot is mostly in NLP, mainly including machine learning and information extraction. In addition, there is also the second most patent layout in speech recognition, mainly including natural language modelling and speech recognition. In the application of E-Business, there are surprisingly many patent layouts at the management level, including marketing, resources management, and office automation.

In short, Knowledge is the basis; machine learning is the main method; speech-related technologies have been widely developed. Observed emerging trend focuses on speech-driven application, including automatic control for system integration and human 'object' interaction for better user experience.

Acknowledge

The research is partially supported by the research grants of Ministry of Science and Technology, Taiwan (Grant numbers: MOST-108-443 2221-E-007-075-MY3).

References

1. S. Reeves, , et al., Narrative video scene description task discriminates between levels of cognitive impairment in Alzheimer's disease. *Neuropsychology*, 2020. 34(4), p. 437-446.
2. J. Dai and Z. Ma, Automatic Identification of Bond Information Based on OCR and NLP. *Journal of Computers*, 2019, Vol. 14(6), pp. 397-403.
3. V.K. Jain and S. Kumar, Predictive analysis of emotions for improving customer services, in *Natural Language Processing: Concepts, Methodologies, Tools, and Applications*, IGI Global, Hershey, 2020.
4. A.P. Saygin, I. Cicekli, and V. Akman, Turing test: 50 years later. *Minds and machines*, 2000. Vol. 10(4), pp. 463-518.
5. S. Lloyd, Least squares quantization in PCM. *IEEE transactions on information theory*, 1982, Vol. 28(2), pp. 129-137.

6. D.M. Blei, A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. *The Journal of Machine Learning Research*, 2003, Vol. 3, pp. 993-1022.
7. S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 2004, Vol. 60, No. 5, pp. 503-520.
8. B. Yoon and Y. Park, A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 2004, Vol. 15(1), pp. 37-50.
9. A. Abbas, L. Zhang and S. Khan, A literature review on the state-of-the-art in patent analysis, *World Patent Information*, 2014, Vol. 37, pp. 3-13.
10. Y.G. Kim, J.H. Suh, and S.C. Park, Visualization of patent analysis for emerging technology, *Expert systems with applications*, 2008, Vol. 34(3), pp. 1804-1812.
11. G. Kim, and J. Bae, A novel approach to forecast promising technology through patent analysis, *Technological Forecasting and Social Change*, 2017, Vol. 117, pp. 228-237.
12. M. Thilakaratne, K. Falkner and T. Atapattu, A systematic review on literature-based discovery workflow. *PeerJ Computer Science*, 2019, Vol. 5, p. e235.
13. Y.-H. Tseng, C.-J. Lin and Y.-I. Lin, Text mining techniques for patent analysis. *Information processing & management*, 2007, Vol. 43(5), pp. 1216-1247.
14. Y.Y. Yang et al., Enhancing patent landscape analysis with visualization output. *World Patent Information*, 2010, Vol. 32(3), pp. 203-220.
15. D. Korobkin, et al. Methods of statistical and semantic patent analysis. in A. Kravets et al. (eds.) *Conference on Creativity in Intelligent Technologies and Data Science*, Springer, Cham, 2017. pp. 48-61.
16. J. Hu et al., Patent keyword extraction algorithm based on distributed representation for patent classification. *Entropy*, 2018, Vol. 20(2), 104.
17. S. Li et al., DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 2018, Vol. 117(2), pp. 721-744.
18. J.-S. Lee and J. Hsiang, Patent classification by fine-tuning BERT language model, *World Patent Information*, 2020, Vol. 61, p. 101965.
19. T. Shanie, J. Suprijadi, and Zulhanif. *Text grouping in patent analysis using adaptive K-means clustering algorithm*. in *AIP Conference Proceedings*. 2017. AIP Publishing LLC, DOI: 10.1063/1.4979457.
20. A.J. Trappey, et al., Ontology-based technology function matrix for patent analysis of additive manufacturing in the dental industry. *International Journal of Manufacturing Research*, 2017, Vol. 12(1), pp. 64-82.
21. S.-S. Weng et al., Ontology construction for information classification. *Expert Systems with Applications*, 2006, Vol. 31(1), pp. 1-12.
22. A. Trappey, C. Trappey, and A.-C. Chang, Intelligent Extraction of a Knowledge Ontology From Global Patents: The Case of Smart Retailing Technology Mining. *International Journal on Semantic Web and Information Systems*, 2020, Vol. 16, pp. 61-80.
23. D. Tsatsou, et al., *Ontology construction*, in *USPTO*. 2013, US8620964B2, Google Technology Holdings LLC: US.
24. R. Subhashini and J. Akilandeswari, A survey on ontology construction methodologies. *International Journal of Enterprise Computing and Business Systems*, 2011, Vol. 1(1), pp. 60-72.