Contents lists available at ScienceDirect

Technovation

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

Inventor–licensee matchmaking for university technology licensing: A fastText approach

Gyumin Lee [a], Sungjun Lee [b], Changyong Lee [c,*]

[a] *School of Business Administration, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulsan, 44919, Republic of Korea*
[b] *Korea Intellectual Property Strategy Agency, 145 Teheran-ro, Gangam-gu, Seoul, 06132, Republic of Korea*
[c] *Department of Public Administration, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea*

ARTICLE INFO

Keywords:
University technology licensing
Inventor–licensee matchmaking
Technological function
Business requirement
FastText

ABSTRACT

Although many previous studies have explored university technology licensing, few have examined the value of quantitative data and scientific methods in improving operational efficiency. Focusing on the marketing phase of university technology licensing processes, this study proposes an analytical framework for inventor–licensee matchmaking by linking technological functions and business requirements. The proposed framework utilises fastText to construct a technological function-business requirement landscape, where similar technological functions and business requirements are located in close proximity. Potential pairs of inventors and licensees for technology licensing are identified through similarity analysis based on the constructed landscape. To validate the framework's effectiveness, an inventor-licensee matching rate is calculated by comparing the matchmaking results to actual technology licensing contracts. A case study covering 16,517 disclosed inventions and 565 licensed technologies from Sogang University confirms that the proposed analytical framework is useful in identifying potential inventor–licensee pairs. It can serve as a valuable complementary tool for university technology licensing in the era of open innovation.

1. Introduction

University technology licensing has attracted much attention from academic and industrial practitioners in pursuing their mutual and complementary interests concerning economic benefits and business efficiency (Kotha et al., 2018). Previous studies have examined various topics related to university technology licensing activities, such as business models of technology licensing offices (TLOs) (Baglieri et al., 2018; Sengupta and Ray, 2017), productivity and efficiency of TLOs (Chapple et al., 2005; Chau et al., 2017; Hsu et al., 2015), organisational structures of TLOs (Battaglia et al., 2017), and patterns of technology licensing outcomes (Caviggioli et al., 2020; Grimpe and Fier, 2010; Shane, 2002; Thursby et al., 2001; Wu et al., 2015). Although the results of these studies have deepened our understanding of the objectives, characteristics, and outcomes of university technology licensing, a major question remains in the literature regarding the use of quantitative data and scientific methods to provide operational support for university technology licensing.

A key component underpinning the success of university technology licensing processes is inventor–licensee matchmaking (Dong and Pourmohamadi, 2014). While searching for potential partners for technology licensing, universities and firms rely on intermediaries to reconcile their motivations and bridge the cognitive and organisational gaps between academia and industry (Clayton et al., 2018). TLOs, as intermediaries, find potential suitors for disclosed university inventions by linking the technological functions that inventors can provide and the business requirements that licensees pursue, based on a comprehensive understanding of the objectives, operations, and capabilities of both parties for university technology licensing (Debackere and Veugelers, 2005). In this regard, TLO officers engage in a range of activities within inventor–licensee matchmaking, from recognising licensing opportunities to building external relationships, using such capabilities as domain knowledge, personal contacts, and commercialisation skills (Mom et al., 2012; Soares and Torkomian, 2021).

However, this human-centric approach is time-consuming and labour-intensive, becoming less effective as the number of technologies and complexity of technological knowledge increase. Moreover, inventor–licensee matchmaking based on personal relationships is often limited to small networks, complicating the ability to find prospective licensees for inventions in a broader search space. This is mainly because

[*] Corresponding author.
E-mail addresses: optimist@unist.ac.kr (G. Lee), lsj@kista.re.kr (S. Lee), changyonglee@korea.ac.kr (C. Lee).

https://doi.org/10.1016/j.technovation.2023.102765
Received 31 May 2022; Received in revised form 6 April 2023; Accepted 10 April 2023
Available online 23 April 2023

of psychological inertia in human thinking, which puts more weight on exploiting the known than on exploring the unknown. Although many previous studies have examined the determinants of improving the productivity and efficiency of university technology licensing in terms of human resources, institutional resources, and financial resources (Chapple et al., 2005; Debackere and Veugelers, 2005; Hsu et al., 2015; Soares and Torkomian, 2021), little attention has been paid to the use of data-driven approaches that can lower the opportunity costs and overcome the cognitive barriers. This necessitates the development of an analytical framework for inventor–licensee matchmaking based on scientific methods and quantitative data, which can facilitate a more efficient and extensive search for potential partners for technology licensing.

A review of previous studies on university technology licensing identifies two main issues that are central to developing an analytical framework for inventor–licensee matchmaking. First, although university technologies are often codified in patents, patented research represents only a small proportion of all research work being conducted within academia (Agrawal and Henderson, 2002). University researchers might not disclose their valuable technological knowledge through patenting for many reasons (e.g. inventors' incentives to disclose as little as possible and the timing of patent disclosures). Instead, they transfer their technological knowledge through various channels other than patenting, such as scientific publications, conference presentations, and research project proposals (Agrawal and Henderson, 2002). Hence, any proposed analytical framework for inventor–licensee matchmaking should use various types of research outcomes to model the technological functions involved in university inventions. Second, inventor–licensee matchmaking corresponds to the problem of linking two disparate knowledge sources: the technological functions and solutions provided by inventors, and the business requirements demanded by potential licensees. Given the nascent nature of university technologies, it is challenging to identify linkages between technological functions and business requirements because of the information asymmetries that may arise in inventor–licensee interactions (Bradley et al., 2013; Soares and Torkomian, 2021). In this context, understanding the semantic information embedded in each of the technological functions and business requirements is necessary to translate the knowledge involved in university technologies into more applicable projects and to resolve the heterogeneity in the two knowledge sources (Lee, 2021). Therefore, any proposed analytical framework for inventor–licensee matchmaking should consider the semantic relationships between technological functions and business requirements to link the two disparate knowledge sources associated with inventors and licensees.

Considering these issues, we propose the following analytical framework for inventor–licensee matchmaking in university technology licensing contexts. Firstly, we construct technological function and business requirement databases from various research outcomes produced by university researchers and from contract documents for university technology licensing, respectively. We apply text cleaning techniques to represent the technological functions and business requirements as sentences comprising a series of key technical or business terms. Secondly, we use fastText (Bojanowski et al., 2017) to construct a technological function–business requirement landscape by embedding technical and business terms into a high-dimensional vector space, where words sharing common contexts are located in close proximity. fastText is a word embedding technique based on the skip-gram model, which uses a shallow neural network architecture to learn distributed representations of words. Notably, fastText embeds not only words but also character n-grams, generated by dividing words into multiple subwords, to a continuous vector space, and therefore, can obtain the vector embeddings of words that are not in the training data (i.e. out-of-vocabulary words). This enables us to capture the semantic relationships between technological functions and business requirements considering the futuristic concepts inherent in university technologies, which are represented by composing and arranging multiple existing technical terms (e.g. 3D bioprinting (Thayer et al., 2020) and neurorobotics (Knoll et al., 2017)). After the technological function–business requirement landscape is constructed, we represent each technological function and business requirement by averaging the vector embeddings of the technical or business terms that constitute it. Thirdly, we conduct a similarity analysis using the cosine similarity index to identify potential inventor–licensee pairs by measuring the semantic similarity between the technological functions that inventors can provide and the business requirements that licensees pursue. Finally, we evaluate the reliability and validity of the inventor–licensee matchmaking process by defining an inventor–licensee matching rate, which is computed by comparing the matchmaking results with actual university technology licensing contracts concluded in the field.

We applied the proposed analytical framework to a complete sample of inventions disclosed to Sogang University and the technology licensing contracts concluded by Sogang University's TLO from 2015 to 2020. Our case study, which covers 16,517 inventions and 565 licensed technologies, confirmed that the proposed analytical framework is reliable for conducting inventor–licensee matchmaking based on the links between technological functions and business requirements. Furthermore, comparative studies showed that our framework outperforms an alternative method of sentence embedding and pre-trained word embedding models in inventor–licensee matchmaking. Thus, the proposed analytical framework can serve as a complementary tool to support expert decision-making in identifying potential inventor–licensee pairs for university technology licensing.

The remainder of this paper is organised as follows. Section 2 presents the research background, and Section 3 explains the research methodology, which is then illustrated by a case study in Section 4. Section 5 discusses the research implications for theory and practice while offering guidelines for implementing and customising the proposed analytical framework. Finally, Section 6 concludes and discusses the limitations of this study and suggestions for further research.

2. Background

2.1. Inventor–licensee matchmaking in university technology licensing contexts

University technologies are inventions created by university researchers as a result of their research work, encompassing various types of innovative discoveries and creations, such as methods, devices, and processes (Kim et al., 2019). With the development of knowledge-based economies, universities have become a major source of knowledge creation for technological innovation. Thus, university technologies are considered to have a significant potential impact on economic growth (Bradley et al., 2013; Tseng et al., 2020). Industrial firms are often drawn to the potential of university technologies, seeing them as a means to improve their technological competitiveness and drive technological breakthroughs (Wu et al., 2020). Academic and industrial practitioners have become increasingly interested in university technology licensing to pursue mutual and complementary interests in terms of economic benefits and business efficiency (Kotha et al., 2018).

The success of university technology licensing processes often depends on the results of inventor–licensee matchmaking (Dong and Pourmohamadi, 2014). Universities and firms have different objectives, operations, and capabilities; hence, TLOs serve as intermediaries to align their different motivations and bridge the cognitive and organisational gaps between academia and industry during the search for potential partners for university technology licensing (Clayton et al., 2018). TLOs evaluate inventions disclosed by university researchers with consideration of manufacturing feasibility, patentability, novelty, and possible potential markets, and begin marketing based on the evaluation results to find potential licensees that may be interested in obtaining university technologies (Siegel et al., 2004). Therefore, TLOs require human experts with a variety of capabilities in terms of research and marketing (Soares and Torkomian, 2021). Specifically, TLO officers with research-oriented capabilities use their domain knowledge to understand the knowledge embedded in university technologies and examine the value of inventions as commercialisation opportunities. TLO officers with marketing-oriented capabilities build external networks with industries and other research organisations based on their personal contacts to find potential licensees (Mom et al., 2012). Moreover, inventors are often highly involved in identifying potential licensees for their inventions because their specialised knowledge and personal relationships positively impact university technology licensing (Siegel et al., 2004; Thursby and Thursby, 2004).

Although human intelligence is crucial in the process of university technology licensing, relying solely on human-centric approach can be time-consuming and labour-intensive, especially with the increasing number of technologies and growing complexity of technological knowledge. Moreover, TLOs highly depend on the personal relationships of individual TLO officers with researchers and firms while searching for potential inventor–licensee pairs, which often restricts the search space to small networks. This is because human thinking tends to place more weight on exploiting the known than on exploring the unknown, creating psychological inertia. Many previous studies have investigated the determinants for improving the productivity and efficiency of TLOs' technology licensing processes from various perspectives, such as human resources, institutional resources, and financial resources (Chapple et al., 2005; Debackere and Veugelers, 2005; Hsu et al., 2015; Soares and Torkomian, 2021). However, most have focused on qualitative analyses and have not considered the use of quantitative data and scientific methods that can facilitate a more productive market search and lower search and opportunity costs. These research gaps provide our primary motivation and are fully addressed in this study.

2.2. Integration of disparate data sources

Integrating multiple data sources disparate in content and structure is beneficial for discovering new knowledge across many research areas. Such integration enables linking knowledge scattered across different data sources (Brody et al., 1999). There are two main approaches to integrating disparate data sources: identifier and similarity approaches.

First, the identifier approach links disparate data sources by identifying data instances that represent the same entity using a common identifier shared across data sources. This approach provides a global view of all attributes associated with each entity, distributed across multiple disparate data sources, given the existence of a common identifier. Accordingly, the identifier approach has been applied in many previous studies in the field of technology management aiming to understand the relationships between technology and business (Kim et al., 2019; Kim et al., 2022a; Kim et al., 2022b; Lee et al., 2019). For instance, Kim et al. (2019) integrated technology transaction, patent, and publication databases using the name of inventors of patents and journal papers as a common identifier, for technologies transacted through the TLO of Stanford University. Based on the integrated database, they developed a random forest-based approach to assess university-originated technologies by capturing the relationships between the technological characteristics and economic value of the technologies. Lee et al. (2019) used the four-digit standard industrial classification code as a common identifier for linking business segment data of firms in the COMPUSTAT database with financial characteristics data of industries where the firms have entered or exited. They identified potential areas for business diversification using a sequential pattern mining technique, and assessed the feasibility of the business areas identified based on a business segment–financial characteristics integrated dataset. Despite its reliability when a common identifier exists, the identifier approach is often of limited use in practice because most databases do not share common identifiers.

Second, the similarity approach integrates disparate data sources by discerning the data instances corresponding to the same entity based on their similarity. This approach has advantages in its applicability to heterogeneous data sources, even in the absence of a common identifier. Many researchers have presented various algorithms for measuring the similarity between data instances from disparate sources, considering the characteristics of the data sources to be integrated, such as the statistical technique-based (Copas and Hilton, 1990), rule-based (Ganesh et al., 1996), distance-based (Dey et al., 2002), and attribute entropy-based (Qiang et al., 2008) entity matching algorithms. With the development of machine learning algorithms for text analysis in recent years, several approaches for integrating disparate data sources based on similarity measured using text mining or natural language processing (NLP) techniques have been suggested. For example, Kim and Lee (2017) connected patent and online futuristic document data based on the similarity between their keyword distributions by constructing keyword–document matrices using text mining techniques. They assessed the paradigm unrelatedness of signals using the similarities between patents and futuristic documents to detect weak signals of future innovation. Ko et al. (2020) presented a framework to understand business competition among technology-based firms by linking their business areas and technologies using trademark and patent databases. They adopted fastText to measure the similarities between the products and services in trademarks and the technologies described in patents, and identified the relationships between the technologies and business areas of firms based on the similarities. Motohashi et al. (2021) examined the relationship between science and technology by measuring the similarities between patents and academic papers collected from the Web of Science and PATSTAT databases using fastText. While the similarity approach is more broadly applicable than the identifier approach due to its advantages in terms of data availability, establishing criteria to determine the degree of similarity can be challenging, leading to limitations in validating its effectiveness and reliability.

Inventor–licensee matchmaking is the process of matching the technological functions provided by inventors with the business requirements demanded by licensees. This corresponds to the problem of linking two disparate knowledge sources. Considering that most disclosed university inventions are embryonic in nature, information asymmetries may exist in inventor–licensee interactions; thus, it is challenging to identify linkages between technological functions and business requirements. Accordingly, the similarity approach is more appropriate for integrating disparate knowledge sources for inventor–licensee matchmaking, compared to the identifier approach, which requires a common identifier between knowledge sources.

2.3. Word embedding techniques

Word embedding is a method for capturing the semantic and syntactic information of words by mapping the words in text documents to a continuously valued vector space. In recent years, word embedding techniques based on machine learning algorithms have been actively developed in the field of NLP. For example, word2vec, a representative word embedding technique, was designed to reconstruct linguistic contexts of words using a shallow neural network with two approaches: one predicts the target word based on the surrounding context words (i.e. CBOW), and the other predicts the surrounding context words based on the target word (i.e. skip-gram) (Mikolov et al., 2013). Doc2vec was introduced based on insights from word2vec to obtain fixed-length feature representations from variable-length paragraphs using a neural network architecture (Le and Mikolov, 2014). fastText was developed to improve word2vec by considering not only the context but also the subword information of words (Bojanowski et al., 2017). Such word embedding techniques have been widely used as feature extraction methods in many studies based on machine learning models because the latent feature of words can be represented as a numerical vector. Several previous studies aimed to extract insights from technological knowledge by employing word embedding techniques to capture the semantic relationships among technological terms. For example, Lee et al. (2020) proposed an analytical framework for technology opportunity analysis using word2vec, which identifies product areas across multiple domains in which firms may enter based on the technological capabilities embodied in the firms' existing products. They used word2vec to construct a product landscape in which products with similar technological bases are located in close proximity, by regarding the relationships between patents and products to which the patented inventions can be applied as those between documents and words. Hong et al. (2021) developed an idea screening method that considers the technological value of ideas inferred from their technical descriptions. The authors utilised word2vec to obtain vector representations of words considering the semantic relationships among words and to construct matrices that represent the technical content of ideas. A convolutional neural network was then used to model the relationships between the technical content and the technological value of ideas. Jeon et al. (2022) suggested a method to identify novel patents expected to have a high technological impact using textual information from patent documents. They employed doc2vec to obtain vector representations of patent documents, and the local outlier factor technique to measure the novelty of patents on a numerical scale.

In the context of inventor–licensee matchmaking, although every word embedding model has its own characteristics that can be useful for representing technological functions and business requirements, fastText is particularly well-suited for the proposed analytical framework because of its ability to incorporate subword information. As fastText is trained using character *n*-grams of words, it can capture not only the semantic features of words but also the meaning of subwords included in the words. Considering that the technological knowledge inherent in new university technologies can involve a futuristic concept, fastText enables us to generate more reliable vector representations of new technological terms that are a composite of existing words. For example, the meaning of the technical term '3D bioprinting', which is a compound word of 'Biology' and '3D printing', can be inferred by assembling vector representations of the prefix 'bio' and the word '3D printing', in the vector space constructed using fastText.

3. Data and methodology

Overall, the proposed analytical framework is designed to be executed in four discrete steps, as illustrated in Fig. 1: (1) data collection and pre-processing, (2) construction of a technological function–-business requirement landscape using fastText, (3) identification of potential inventor–licensee pairs via similarity analysis, and (4) validation of inventor–licensee matchmaking.

3.1. Data collection and pre-processing

The proposed analytical framework for inventor–licensee matchmaking uses technological function and business requirement databases to construct a technological function–business requirement landscape. The technological function database is established by collecting various forms of research outcomes produced by university researchers (e.g. journal articles, conference presentations, and non-confidential research proposals) for a comprehensive understanding of the knowledge inherent in university technologies. The business requirement database is constructed using contract documents for already licensed technologies, which involve the business requirements licensees pursued through technology licensing activities.

Generally, the title is considered the most informative section in a document containing technological knowledge (e.g. patents, academic papers, and invention disclosures), because the core contents of the technologies are encapsulated in the semantic composition of words in the title. Therefore, this study uses the titles of research outcomes and technology licensing contracts to capture the semantic relationships between technological functions and business requirements. The titles are extracted from the established databases and pre-processed to build an integrated vocabulary set in the following manner: (1) special symbols are removed; (2) meaningless or irrelevant words, such as standalone numbers, punctuation, and white-spaces, are removed; (3) stop words, such as 'the', 'is', and 'of' are eliminated; and (4) words that are too frequent or rare are eliminated. The resulting vocabulary set contains every word in the two databases and is used to construct a technological function–business requirement landscape based on a word embedding technique.

3.2. Construction of a technological function–business requirement landscape

Using fastText, a technological function–business requirement landscape is constructed, where words related to university technologies that share similar meanings are located in close proximity. Similar to word2vec, fastText uses the skip-gram model, which predicts surrounding context words based on the current word to capture the semantic meanings of words (Bojanowski et al., 2017). However, fastText considers not only the context of words, but also subword information.

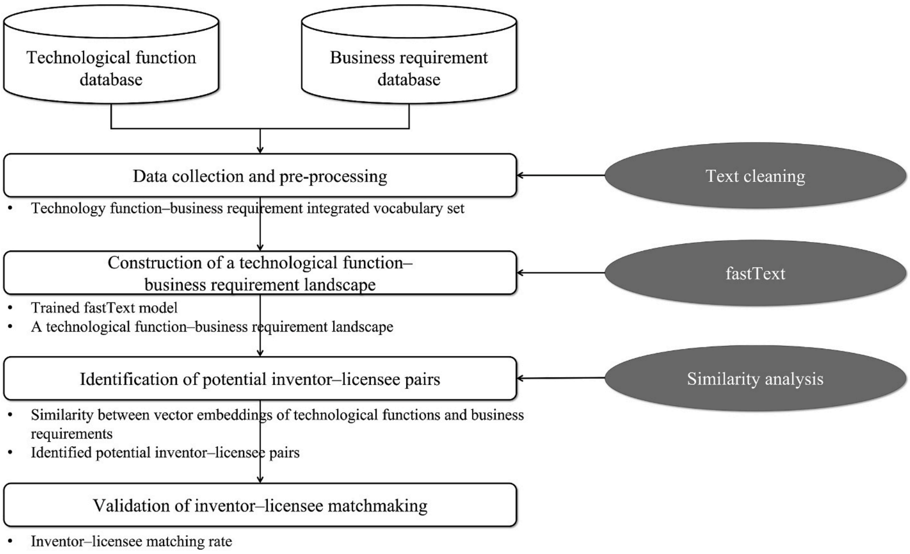

Technological function database
Business requirement database
Data collection and pre-processing
Text cleaning
• Technology function–business requirement integrated vocabulary set
Construction of a technological function–business requirement landscape
fastText
• Trained fastText model
• A technological function–business requirement landscape
Identification of potential inventor–licensee pairs
Similarity analysis
• Similarity between vector embeddings of technological functions and business requirements
• Identified potential inventor–licensee pairs
Validation of inventor–licensee matchmaking
• Inventor–licensee matching rate

Fig. 1. Overall process of the proposed analytical framework.

This enables us to obtain even the vector embeddings of words not in the training data. Given that the technological knowledge to be transferred through university technology licensing may involve futuristic concepts, fastText is used in this study to effectively represent the technological functions and business requirements, which may include new technical or business terms not in the current text corpus.

The skip-gram model is a neural network with one hidden layer, as shown in Fig. 2. The input layer takes a word encoded as a one-hot vector, where the value of the position corresponding to the word is 1, and those of the remaining positions are 0s. For each word in the vocabulary, the output layer independently produces the probability of the word appearing in the context of the input word. The skip-gram model is then trained by regarding the problem of predicting context words as a set of independent binary classification tasks to predict the presence (or absence) of context words. In particular, negative sampling, which updates the weights in the model using only a small fraction of training data, is applied to reduce the computational cost of using a large corpus. For training each input word, all its actual context words are considered positive examples, whereas negative examples are randomly drawn from the vocabulary. Given the words $w_1, w_2, \dots, w_T$ in the text corpus, the objective of the skip-gram model is to maximise the following negative log-likelihood:

$$\log\left(1+e^{-s(w_t,w_c)}\right)+\sum_{w_n\in\mathcal{N}_{t,c}}\log\left(1+e^{s(w_t,w_n)}\right), \tag{1}$$

where w_t and w_c are the input word and its context words, respectively; $\mathcal{N}_{t,c}$ is a set of negative examples; and s is a scoring function that computes the probability of the presence of the context words. In general, the skip-gram model defines the scoring function as a scalar product of the word and context vectors. During the model training process, the prediction error is calculated based on the difference between the actual presence of the context words and the estimated scores. This error is propagated backward through the network, and the weights in the hidden layer are updated using the stochastic gradient descent algorithm to maximise the objective function of the model. The weights between the input and hidden layers are a matrix of size $T \times D$, where T is the length of the input vector and D is the dimension of the hidden layer. Thus, each row of the weight matrix becomes a D-dimensional vector representation of each word in the text corpus fastText extends the skip-gram model to capture the internal structure of words by considering the subword information. For this, a bag of character n-gram is employed to represent each word. For example, for $n = 3$, the word ‘where’ is represented by the following character n-grams: <wh, whe, her, ere, re>, and <where>. Notably, the word ‘where’ itself is included in the set of n-grams for the model for learning a representation of the complete word. The scoring function for fastText is then designed to reflect not only the word but also all the n-grams that comprise the word. Given a dictionary of n-grams of size G, the scoring function between word w and its context word c is obtained as follows:

$$s(w,c)=\sum_{g\in\mathcal{G}_w} z_g^{\top} v_c, \tag{2}$$

where $\mathcal{G}_w \subset \{1, \dots, G\}$ is a set of n-grams appearing in word w, z_g is a vector embedding of the n-gram g and v_c is a vector embedding of the context word c. fastText uses this scoring function to represent a word by the sum of the vector embeddings of its n-grams, and can learn reliable representations for rare words.

We apply fastText to embed the technical and business terms in the integrated vocabulary set into a vector space where words with similar semantic meanings are assigned to a similar position. In our databases, individual technological functions and business requirements are represented as sentences consisting of a series of words. Thus, each sentence representing a technological function or business requirement is mapped into the vector space by averaging the vector embeddings corresponding to its constituent words. The technological function–business requirement landscape is finally constructed using the organised vector space and used to represent the technological functions and business requirements for identifying potential inventor–licensee pairs.

3.3. Identification of potential inventor–licensee pairs

In university technology licensing, the probability of licensing an invention to a specific licensee depends on the semantic proximity between the technological function provided by the inventor and the business requirement demanded by the prospective licensee. In this study, as both technological functions and business requirements are embedded in the same vector space (i.e. the technological function–-business requirement landscape), potential matches between inventors and licensees are identified by measuring the spatial closeness (i.e. semantic similarity) of the technological functions and business requirements in the constructed landscape. The cosine similarity index is employed in this context, as formulated in Eq. (3).

$$\mathrm{Cos}(\boldsymbol{v}_I,\boldsymbol{v}_L)=\frac{\boldsymbol{v}_I\cdot\boldsymbol{v}_L}{|\boldsymbol{v}_I||\boldsymbol{v}_L|}, \tag{3}$$

where $\boldsymbol{v}_I$ and $\boldsymbol{v}_L$ are vector embeddings representing the technological function provided by an inventor and the business requirement demanded by a prospective licensee, respectively. The cosine similarity index ranges from 0 to 1, which enables us to regard the measured semantic similarity as the probability that an inventor and licensee are matched. Hence, based on the semantic similarity of the technological function and business requirement measured for each pair of inventor and licensee, we identify inventor–licensee pairs that may potentially match with one another in university technology licensing.

3.4. Validation of inventor–licensee matchmaking

The practicality of the proposed analytical framework for inventor–licensee matchmaking lies in its ability to identify potential pairs that can successfully conclude technology licensing contracts. In this context, we use university technology licensing contracts that have been previously concluded and inventions that have been disclosed, as references for validation of the developed inventor–licensee matchmaking

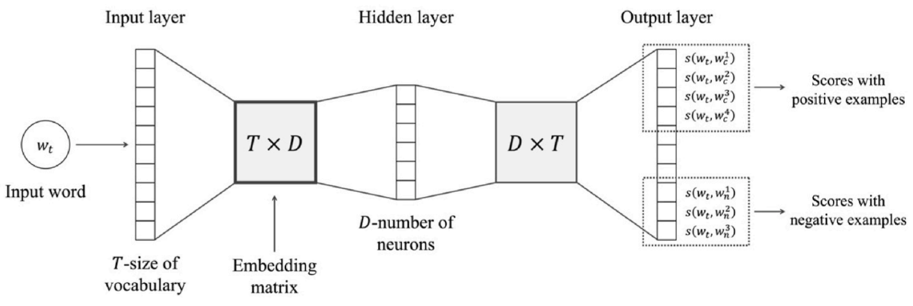


Fig. 2. Skip-gram model architecture.

method. For each licensed technology, the semantic similarities between the business requirement involved in the licensing contract and the technological functions provided by the disclosed inventions are calculated using the cosine similarity index. Based on the calculated similarities, the Top K disclosed inventions that are most semantically similar to the licensed technology are selected, and their inventors are considered as potential inventor candidates. If the actual inventor of the licensed technology is included in the candidate list, the identified inventor–licensee pairs are considered matched. From this, an inventor–licensee matching rate is defined as the ratio of the matched inventor–licensee pairs to the whole licensed technologies, as formulated in Eq. (4).

$$matching\ rate\ @\ K = \frac{\sum_{i \in I_{ref}} H\left(i, I^{i}_{TopK}\right)}{\left|I_{ref}\right|} \tag{4}$$

where I_{ref} is a set of actual inventors of the licensed technologies, I^{i}_{TopK} is a set of inventors of the Top K most semantically similar inventions selected from the inventor–licensee matchmaking for the actual inventor i, and H is a function that determines whether the actual inventor i is included in the potential inventor candidates, as defined in Eq. (5).

$$H\left(i, I^{i}_{TopK}\right) = \begin{cases} 1, if\ i \in I^{i}_{TopK} \\ 0, otherwise \end{cases} \tag{5}$$

4. Case study

4.1. Data collection and pre-processing

To demonstrate the proposed analytical framework, we present a case study on university technology licensing contracts concluded by Sogang University. Our data comprise a complete sample of 16,517 inventions disclosed to the TLO of Sogang University over 6 years (2015–2020), including patents, academic papers, conference proceedings, and research project outcomes. Only a small proportion of these inventions (approximately 3%) were successfully licensed, and some of the licensed inventions (approximately 6%) were the results of university–industry collaboration projects. Accordingly, we collected data on 565 contracts for the licensed inventions, excluding the results of university–industry collaboration projects demanded and funded by firms. We used these two types of obtained data (i.e. inventions and technology licensing contracts) as the technology function and business requirement databases, respectively.

The databases consist of both structured and unstructured data, which provide detailed information about each invention or contract, as presented in Table 1. For example, each invention in the technological function database involves its submission date, intellectual property type, name of its inventor, department to which the inventor belongs, and the title of the invention. Each technology licensing contract in the business requirement database involves the contract date, type of technology, name of inventor and licensee, and the title of the licensed technology. We anonymised the name and department information to preserve the privacy of the inventors. The resulting databases contain 553 individual inventors and 56 departments indexed by order of appearance. We built an integrated vocabulary set by extracting the titles of the inventions and licensed technologies from the databases and using them as technological functions and business requirements. The titles representing the technological functions and business requirements were then pre-processed by eliminating special symbols, meaningless words, stop words, and words that are too frequent or rare. The integrated vocabulary set was constructed by collecting unique words included in the titles. The resulting vocabulary set contains 26,080 unique words relevant to university technologies.

4.2. Construction of a technological function–business requirement landscape

A technological function–business requirement landscape that renders the semantic relationships between technological functions and business requirements was constructed in three steps using fastText. First, we trained a fastText model on the collected data to predict the context words that appear together with an input word in the title of each invention or licensed technology. Second, we obtained vector embeddings for the words in the vocabulary set from the hidden layer of the trained model. Finally, we computed the sentence embedding vector for each technological function or business requirement by taking the average of the vector embeddings corresponding to its constituent words. Given that words with similar meanings are placed in close proximity in the vector space, the distance of sentence embedding vectors between a technological function and business requirement indicates their semantic closeness. As a result, the technological function–business requirement landscape was constructed by the organised vector space, from which the relationship between inventors and licensees can be inferred, using the sentence embeddings for the technological functions and business requirements.

We adopted the fastText module, provided by the Gensim Python library,[1] to embed words in the integrated vocabulary set into a vector space. The hyperparameters associated with the fastText model should be carefully selected to obtain adequate vector representations of words. The number of nodes in the hidden layer (i.e. dimensionality for vector representations) primarily affects the performance of the word embedding model and the computational cost of model training. Word embedding models, such as fastText, in general, have a trade-off between performance and training time depending on the number of nodes in the hidden layer: the larger the number of nodes, the better the model performs, but more training time is needed. Previous studies have empirically found that the quality of word embedding is not significantly improved when the number of hidden nodes is greater than 300 (Pennington et al., 2014), and this is one of the reasons why the pre-trained fastText model provided by Facebook AI Research uses 300 as the dimensionality for its vector representations. Considering these issues, we selected 300 as the number of nodes in the hidden layer of our fastText model to preserve the quality of word embedding and to allow a fair comparison with the pre-trained model. We additionally conducted a sensitivity analysis on the number of nodes in the hidden layer (i.e. 100, 200, and 300) to investigate its robustness against the inventor–licensee matchmaking results. We confirmed that there is no significant difference in matching rate at every K (i.e. the number of inventor candidates) among the dimensionalities, as reported in Appendix A.1. Other important hyperparameters of the fastText model are the window size which determines the number of context words to be considered, the number of negative examples to be used for negative sampling in the model training, and the number of training epochs for the model. We examined how the inventor–licensee matchmaking results change with the choice of these hyperparameters, to identify the most suitable set of hyperparameters for representing the technological functions and business requirements based on word embedding vectors. The results of the sensitivity analysis, which are reported in Appendix A.2, showed that there are only slight differences in the performance of matchmaking when using different sets of hyperparameters. Accordingly, we selected the best set of hyperparameters, as follows: the window size of 10, 5 negative examples, and 2500 training epochs. The resulting technological function–business requirement landscape constructed using the fastText model, which is trained with the selected hyperparameters, is not reported here in its entirety owing to a lack of space, but parts of the sentence embedding vectors are presented in Table 2.

[1] https://radimrehurek.com/gensim/.

Table 1
Parts of the technological function and business requirement databases.

(a) Part of the technological function database

Invention No.	Type	Date	Title	Department ID	Name ID
20646	Academic paper	05/30/2015	YouTube acceptance by university educators and students: a cross-cultural perspective	$\mathit{Department}_{8}$	Name_{398}
21054	Academic paper	10/01/2015	Emotional information processing based on feature vector enhancement and selection for human-computer interaction via speech	$\mathit{Department}_{60}$	Name_{150}
…	…	…	…	…	…
32179	Academic paper	01/10/2021	Bottom-up solutions in a time of crisis: the case of Covid-19 in South Korea	$\mathit{Department}_{15}$	Name_{236}
…	…	…	…	…	…
4000000259	Patent	10/23/2015	Optional signal processing devices and methods on distributed antenna systems	$\mathit{Department}_{53}$	Name_{264}
4000000751	Patent	10/30/2015	The method of data transmission in mobile communication systems using relay stations and their systems	$\mathit{Department}_{63}$	Name_{197}
…	…	…	…	…	…
4000012414	Patent	01/21/2021	Electrolytes for supercapacitors and supercapacitors containing them	$\mathit{Department}_{53}$	Name_{344}

(b) Part of the business requirement database

Technology No.	Type	Date	Title	Department ID	Name ID
5000000737	Know-How	01/30/2015	Know-how to develop high-speed signal interface circuits for future intelligent healthcare platforms	$\mathit{Department}_{53}$	Name_{246}
5000000740	Know-How	03/27/2015	Improving image sensor design capabilities and transferring know-how to improve image quality	$\mathit{Department}_{53}$	Name_{591}
…	…	…	…	…	…
5000002151	Know-How	06/01/2018	Knowledge-How technology transfer in vitro protein activity using mobile phone imaging devices	$\mathit{Department}_{53}$	Name_{514}
…	…	…	…	…	…
5000000813	Patent	06/08/2015	Amalgam electrodes, their manufacturing methods, and electrochemical reduction methods of carbon dioxide using them	$\mathit{Department}_{63}$	Name_{286}
5000000815	Patent	08/04/2015	New Organic Light Emitting Transistor Registration Patent Transfer Using DMDCNQI as an N Type Electron Transport Layer for P3HT Polymer Transistors	$\mathit{Department}_{62}$	Name_{317}
…	…	…	…	…	s…
5000002148	Patent	10/05/2016	Quantitative Analysis of Oligomers in Polymers Using MALDI-TOF Mass Analysis Method	$\mathit{Department}_{63}$	Name_{323}

Table 2
Parts of technological function–business requirement landscape.

Technological functions	Vector						
	v_1	v_2	v_3	…	v_{298}	v_{299}	v_{300}
YouTube acceptance by university educators and students: a cross-cultural perspective	−0.1221	−0.1219	0.8319	…	−0.3707	−0.2819	0.2260
Emotional information processing based on feature vector enhancement and selection for human-computer interaction via speech	−0.4557	−0.0495	0.4961	…	−0.4990	−0.0973	−0.2009
…	…	…	…	…	…	…	…
Quantitative Analysis of Oligomers in Polymers Using MALDI-TOF Mass Analysis Method	−0.4279	−0.5558	0.7802	…	−0.1017	0.1606	0.4277
Business requirements	**Vector**						
	v_1	v_2	v_3	…	v_{298}	v_{299}	v_{300}
Know-how to develop high-speed signal interface circuits for future intelligent healthcare platforms	−0.6423	−0.3844	0.6056	…	−0.7626	0.1014	0.2054
Improving image sensor design capabilities and transferring know-how to improve image quality	−0.6307	−0.0481	0.5203	…	−0.5800	0.1236	0.3463
…	…	…	…	…	…	…	…
Transfer of patent rights in addition to beamforming method using likelihood maximisation	−0.5673	−0.0753	0.5016	…	−0.5993	−0.2266	0.3699

4.3. *Identification of potential inventor–licensee pairs*

From the technological function–business requirement landscape, in which the titles of inventions and licensed technologies are mapped as sentence embedding vectors, the semantic similarity between the technological functions and business requirements was identified by calculating the cosine similarity between their vector representations. For each business requirement demanded by each licensee through university technology licensing contracts, we ranked the technological functions that could be matched in order of the similarity index, as presented in Table 3. For instance, for the licensed technology titled 'Distributed agreement protocol for matching keys in a blockchain-based dynamic encryption key generation environment in an internal vehicle network', the top five technological functions that semantically correlate with the business requirement were identified as follows: 'How to generate dynamic link keys in a sensor network system based on blockchain and above sensor network systems', 'Optimise Ceph Messenger performance with multiple connectivity structures and load balancing algorithms', 'Beacon-based non-adjacent using personal information hidden authentication (personal replacement key) technology,' 'Game-based stochastic routing protocol techniques in vehicle ad hoc network environments', and 'Mobile blockchain-based intelligent healthcare solution training for disaster/emergency sites'. For another licensed technology titled 'Technology transfer of know-how to develop and analyse ultrasonic contrast medium combining nanoparticles', which was invented by a researcher in $\mathit{Department}_{62}$, the top five inventions that semantically

Table 3
Examples of the potential inventor–licensee pairs.

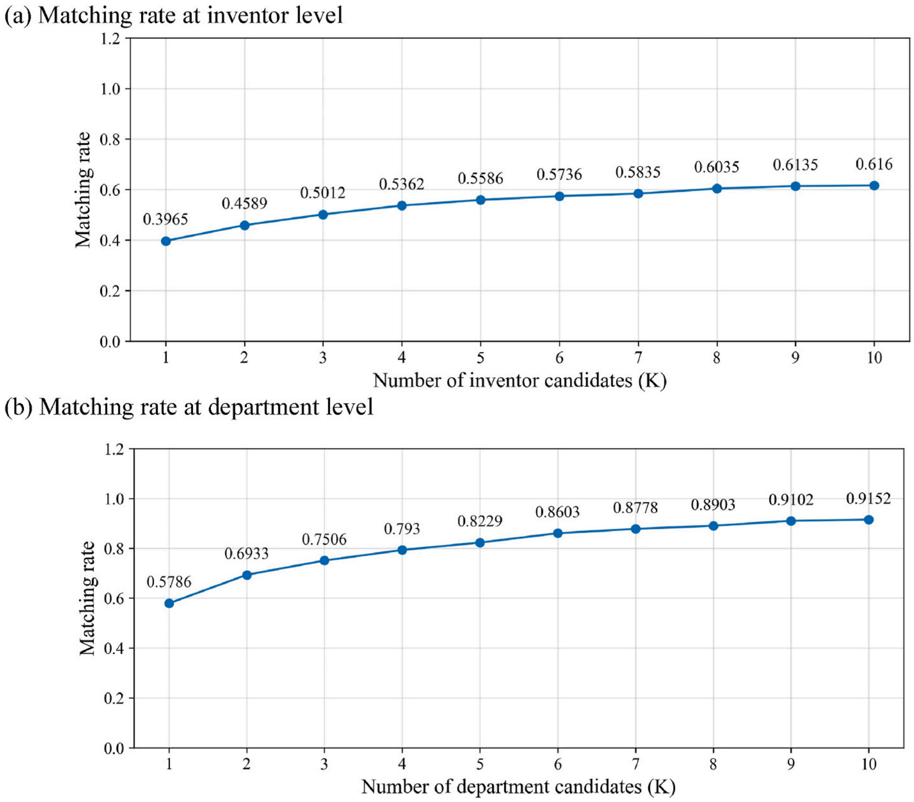
Name ID	Department ID	Business requirements	Name ID	Department ID	Technological functions	Similarity index
$Name_{215}$	$Department_{60}$	Distributed agreement protocol for matching keys in a blockchain-based dynamic encryption key generation environment in an Internal vehicle network	$Name_{215}$	$Department_{60}$	How to generate dynamic link keys in a sensor network system based on blockchain and above sensor network systems	0.9111
			$Name_{213}$	$Department_{60}$	Optimise Ceph Messenger performance with multiple connectivity structures and load balancing algorithms	0.8839
			$Name_{111}$	$Department_{53}$	Beacon-based non-adjacent using personal information hidden authentication (personal replacement key) technology	0.8800
			$Name_{107}$	$Department_{60}$	Game-based stochastic routing protocol techniques in vehicle ad hoc network environments	0.8713
			$Name_{574}$	$Department_{53}$	Mobile blockchain-based intelligent healthcare solution training for disaster/ emergency sites	0.8595
Name ID	**Department ID**	**Business requirements**	**Name ID**	**Department ID**	**Technological functions**	**Similarity index**
$Name_{176}$	$Department_{62}$	Technology transfer of know-how to develop and analyse ultrasonic contrast medium combining nanoparticles	$Name_{176}$	$Department_{62}$	Diagnostic systems and methods using photoacoustic/ultrasonic contrast medium	0.8893
			$Name_{505}$	$Department_{49}$	Interactive fogs without water droplet condensation to create augmented spaces for performances and exhibitions	0.8494
			$Name_{279}$	$Department_{53}$	Ultrasonic diagnostic devices, ultrasonic probes, and their control methods	0.8445
			$Name_{135}$	$Department_{62}$	Contact lens-based glaucoma therapy system that releases leakage enzyme-sensitive nanomedicine	0.8440
			$Name_{342}$	$Department_{63}$	Organise nanoparticles, develop high-performance molecular separators, and create devices	0.8427
Name ID	**Department ID**	**Business requirements**	**Name ID**	**Department ID**	**Technological functions**	**Similarity index**
$Name_{604}$	$Department_{14}$	Knowledge-How technology transfer of heat transfer phenomena caused by heat generation and lubricant circulation of tractor transaxle systems	$Name_{78}$	$Department_{14}$	Numerical study on flow and heat transfer characteristics of air-jet cooling system	0.8785
			$Name_{41}$	$Department_{14}$	New turbulence heat and material transfer based on high-precision flow databases and information technology	0.8748
			$Name_{604}$	$Department_{14}$	Gas-liquid mixed flow and heat transfer analysis in an exothermic agitator	0.8683
			$Name_{268}$	$Department_{14}$	Numerical analysis of convective drying of a moist object with combined internal and external heat and mass transfer	0.8667
			$Name_{309}$	$Department_{60}$	Indoor air quality measurement and trend prediction system using sensor meter and circuit type circulation unit (GRU)	0.8492

correlate with the business requirement were identified as follows: ‘Diagnostic systems and methods using photoacoustic/ultrasonic contrast medium’, ‘Interactive fogs without water droplet condensation to create augmented spaces for performances and exhibitions’, ‘Ultrasonic diagnostic devices, ultrasonic probes, and their control methods’, ‘Contact lens-based glaucoma therapy system that releases leakage enzyme-sensitive nanomedicine’, and ‘Organise nanoparticles, develop high-performance molecular separators, and create devices’.

These matchmaking results reveal that the inventor candidates are identified based on the key technological terms shared by their technological functions and the licensee’s business requirement, such as ‘blockchain’, ‘network’, ‘ultrasonic’, and ‘nanoparticles’. Moreover, in most cases, the actual inventor of each licensed technology was among the top five candidate inventors. The matchmaking results show that the proposed analytical framework can capture the semantic relationships between technological functions and business requirements, and identify inventor–licensee pairs of the actual technology licensing contracts, even without prior information, such as personal relationships.

4.4. Validation of inventor–licensee matchmaking

We quantitatively examined the results of inventor–licensee matchmaking derived from the proposed analytical framework using the inventor–licensee matching rate. For a more detailed performance evaluation of the proposed analytical framework, we obtained the inventor–licensee matching rate using different K values from 1 to 10. Moreover, considering that the colleagues of an inventor of licensed technology in the same department are more likely to be matched with the corresponding licensee than those in other departments, we also investigated the inventor–licensee matching rate using the department information of the inventor candidates. Fig. 3 illustrates the calculated inventor–licensee matching rate at different K values and different levels of analysis (i.e. at the inventor and department levels), for the results of the inventor–licensee matchmaking. At the inventor level, the inventor–licensee matching rate rose from 0.3965 to 0.6160 as K (i.e. the number of inventor candidates) increased from 1 to 10. This indicates that the proposed analytical framework identified approximately 40% of the original inventors of the licensed technologies as the most relevant potential partner and 61% of the original inventors as among the top 10 most relevant potential partners, for the collected technology licensing contracts. These results demonstrate that our framework is reliable enough to match the potential inventor–licensee pairs, given that there are 553 unique inventor candidates in the database. At the department level, the inventor–licensee matching rate approached approximately


(a) Matching rate at inventor level

(b) Matching rate at department level

Fig. 3. Matching rate for results of inventor–licensee matchmaking.

0.91 when the *K* value is set to 10 and was higher than 0.5, even when the *K* value is set to 1, indicating that our framework correctly identifies the departments of most of the actual inventors. Consequently, the results of the validation using the inventor–licensee matching rate confirm that using our proposed analytical framework, inventor–licensee matchmaking can be successfully performed based on the semantic similarity between technological functions and business requirements.

5. Discussion

5.1. Theoretical and practical implications

Our case study demonstrates that the proposed analytical framework for inventor–licensee matchmaking is practical and reliable when searching for potential partners in the university technology licensing process. Therefore, this study has several theoretical and practical implications.

From a theoretical standpoint, the proposed analytical framework extends the previous studies in the field of university technology licensing. While the previous studies have primarily focused on qualitative research to investigate the objectives, characteristics, and outcomes of technology licensing, our framework employs a systematic and scientific approach that leverages quantitative data to enhance the operational efficiency of the technology licensing process. In particular, fastText is used to develop a technological function–business requirement landscape as a latent space for representing the semantic features of technological functions and business requirements. To the best of our knowledge, this study is the first attempt to apply fastText, an emerging word embedding technique widely used in the literature regarding NLP, for inventor–licensee matchmaking in university technology licensing contexts. Because fastText enables the semantic connection of technological functions and business requirements, the proposed analytical framework can assist human experts in decision-making for inventor–licensee matchmaking which is a complex and cognitive process conducted during university technology licensing. Furthermore, the ability of fastText to obtain vector embeddings for words that are not in the training data is suitable for analysing new technological terms that represent futuristic technologies which are not present in the current text corpus. Finally, although this study focuses on inventor–licensee matchmaking, the proposed analytical framework could be helpful in addressing other research problems that should consider the semantic features of technological knowledge, such as technology forecasting, roadmapping, and planning.

From a practical standpoint, this study contributes to reducing the time and cost of searching for potential inventor–licensee pairs in the technology licensing process by developing a data-driven and automated system that produces a potential partner list for technology licensing without expert knowledge. The proposed analytical framework facilitates the search stage in the university technology licensing process, enabling practitioners to focus more on the contract stage, where the economic value of university technologies is determined. This may provide strong motivation for practitioners to adopt the proposed analytical framework as a tool to support the decision-making process of university technology licensing, as the number of technologies and complexity of technological knowledge grow. Moreover, the traditional inventor–licensee matchmaking by TLOs, mainly based on personal relationships, is often limited within a narrow search space in identifying prospective licensees for disclosed inventions. The proposed analytical framework could be helpful for practitioners to expand the search space, as it identifies potential inventor–licensee pairs based on the semantic similarity between the technological functions and business requirements using only text data without any other information about the inventors and licensees. Although this study focuses on identifying potential partners for university technology licensing, the proposed analytical framework can also be used in other practical contexts. For example, university researchers could use the proposed analytical framework not only to find potential buyers of their inventions but also to identify other researchers suitable for collaboration in follow-up research.

5.2. Guidelines on implementing and customising the proposed analytical framework

Although the proposed analytical framework has shown reliability and practicality, there are many issues to consider before applying and deploying a new method in practice. First, the technological functions and business requirements are represented as sentence embeddings obtained by averaging the vector embeddings of the words constituting each of them. Although averaging word embedding vectors is useful for mapping sentences into a vector space, it has limitations in dealing with long sentences with several unimportant words, because it does not consider the weights of the words. The smooth inverse frequency (SIF) technique for sentence embedding has been suggested to circumvent such limitations (Arora et al., 2017). The basic idea of the SIF technique is to represent sentence embeddings using a weighted average of the word vectors in each sentence and to adjust the embeddings by removing the common component among the sentences (i.e. the first principal component of the sentence embeddings) using singular value decomposition for a better representation. For each sentence, the weight of a word is calculated as follows:

$$\frac{a}{\alpha + p(w)} \tag{6}$$

where α is a scalar hyperparameter and $p(w)$ is the probability that word w appears in the entire corpus. This simple technique has proven useful in improving the performance of sentence embedding in semantic textual similarity tasks. We then conducted an additional experiment using the SIF technique instead of unweighted averaging to generate sentence embeddings of the technological functions and business requirements. We chose 0.001 as the value of hyperparameter a, as its appropriate value is usually determined to be between 0.0001 and 0.001 (Arora et al., 2017). Table 4 presents the performance comparison results between the unweighted averaging and the SIF technique in inventor–licensee matchmaking.

The results demonstrate that sentence embedding using the unweighted averaging technique outperforms the SIF technique in terms of the inventor–licensee matching rate for every K value. This result is likely because the proposed analytical framework uses the technological or business terms that remained after filtering the unimportant words so that the application of word weights in the SIF technique has little impact on sentence embedding. Moreover, because the proposed analytical framework uses the titles of research outcomes and technology licensing contracts, which are relatively short sentences, the advantage of the SIF technique in dealing with long sentences could disappear. The SIF technique could be more useful for capturing the semantic features of sentences if we employ other text data, including sentences longer than the titles (e.g. the abstract of an academic paper or patent document).

Table 4
Results of the performance comparison between unweighted averaging and SIF.

Number of candidates (K)	Matching rate			
	Unweighted averaging		Smooth inverse frequency	
	Inventor level	Department level	Inventor level	Department level
1	0.3965	0.5786	0.3716	0.5037
2	0.4589	0.6933	0.4464	0.6185
3	0.5012	0.7506	0.4663	0.6758
4	0.5362	0.7930	0.5012	0.7257
5	0.5586	0.8229	0.5312	0.7681
6	0.5736	0.8603	0.5461	0.7905
7	0.5835	0.8778	0.5586	0.7955
8	0.6035	0.8903	0.5736	0.8130
9	0.6135	0.9102	0.5810	0.8279
10	0.6160	0.9152	0.5935	0.8404

Second, although the fastText architecture was trained using the technological function and business requirement databases collected in this study, several pre-trained models trained on general corpora for language representation can also be used to obtain sentence embeddings for constructing the technological function–business requirement landscape. Therefore, we conducted an additional comparative study using the following two pre-trained models to examine the applicability of the pre-trained models for the proposed analytical framework: (1) a pre-trained fastText model provided by Facebook's AI Research Lab, which was trained on 1 million words appearing in Wikipedia, and (2) Sentence-BERT (Reimers and Gurevych, 2019), which is a modification of the BERT (Devlin et al., 2019) using siamese network structures. BERT, developed by Google AI, is a transformer-based deep neural network trained on more than 3.3 billion words across several web sources. Although BERT can generate sentence embeddings by averaging its output word vectors, Sentence-BERT extends this by adding a pooling operation to the BERT output, producing a fixed-length embedding vector for each sentence. Specifically, Sentence-BERT is constructed as siamese networks of two pre-trained BERT models and is fine-tuned to find semantically similar sentences. Through fine-tuning, Sentence-BERT updates its weights so that the produced sentence embeddings are semantically meaningful. These two pre-trained models (i.e. fastText and Sentence-BERT) have been shown to be effective in improving many NLP tasks, such as semantic textual similarity and question answering tasks. In addition, we conducted incremental training on the pre-trained fastText model using our datasets. We added this fine-tuned model to the comparative study to investigate the influence of domain-specific knowledge on the sentence embedding performance of the fastText model. Table 5 presents the performance comparison results for the sentence embedding models in inventor–licensee matchmaking.

The results demonstrate that the sentence embedding models based on the fastText architecture outperform the pre-trained Sentence-BERT in every case. In particular, the custom-trained fastText, trained using the technological function and business requirement databases collected in this study, slightly outperforms the pre-trained fastText or fine-tuned fastText models. Moreover, the fine-tuned fastText outperforms the pre-trained one, indicating that the range of the data used to train the sentence embedding models influences the ability of the models to capture the semantic meanings of the words representing domain-specific knowledge, such as technological functions and business requirements. Although the pre-trained models generally perform well on various NLP tasks, the vector embeddings they produce reflect a broad context and usage of words; thus, they cannot effectively represent the technical or business terms usually used in a specific domain. Thus, using the tailored dataset covering specific domain knowledge involved in university technologies is reasonable for obtaining vector embeddings of technical or business terms used in university technology licensing contexts.

6. Conclusion

This study proposes an analytical framework for inventor–licensee matchmaking in university technology licensing contexts by linking the technological functions inventors can provide and business requirements licensees pursue. We use fastText to construct a technological function–business requirement landscape, in which the technological functions and business requirements that share common technological knowledge are located in close proximity. Similarity analysis using the cosine similarity index is then performed to match potential inventor–licensee pairs by identifying the semantic similarity between the technological functions and business requirements. Our case study, covering 16,517 inventions and 565 licensed technologies, confirmed that the proposed analytical framework for inventor–licensee matchmaking is useful as a complementary tool to support expert

Table 5
Results of the performance comparison among sentence embedding models.

Number of candidates (K)	Matching rate							
	Custom-trained fastText		Pre-trained fastText		Fine-tuned fastText		Pre-trained Sentence-BERT	
	Inventor level	Department level	Inventor level	Department level	Inventor level	Department level	Inventor level	Department level
1	0.3965	0.5786	0.3192	0.4888	0.3317	0.5436	0.1421	0.2693
2	0.4589	0.6933	0.3691	0.6110	0.3915	0.6484	0.1870	0.3741
3	0.5012	0.7506	0.3990	0.6708	0.4289	0.7157	0.2170	0.4539
4	0.5362	0.7930	0.4339	0.7207	0.4539	0.7606	0.2369	0.5012
5	0.5586	0.8229	0.4613	0.7431	0.4738	0.7706	0.2519	0.5312
6	0.5736	0.8603	0.4813	0.7656	0.5012	0.7880	0.2643	0.5711
7	0.5835	0.8778	0.5012	0.7830	0.5137	0.7980	0.2718	0.5985
8	0.6035	0.8903	0.5212	0.8005	0.5212	0.8105	0.2843	0.6409
9	0.6135	0.9102	0.5287	0.8055	0.5411	0.8279	0.2968	0.6958
10	0.6160	0.9152	0.5461	0.8130	0.5536	0.8354	0.3042	0.7107

decision-making in searching for potential technology licensing partners.

The contributions of this study are two-fold. From a theoretical standpoint, this study contributes to prior literature by developing a systematic and scientific approach using quantitative data to facilitate the operational efficiency of university technology licensing. Moreover, to the best of our knowledge, this study is the first attempt to apply fastText for inventor–licensee matchmaking in university technology licensing contexts. The ability of fastText to obtain vector embeddings for out-of-vocabulary words is suitable for generating more reliable vector representations of technological functions and business requirements that often involve new technological terms. From a practical standpoint, the proposed analytical framework can reduce the time and cost of searching for potential inventor–licensee pairs using a data-driven and automated system that produces a potential partner list for university technology licensing without expert knowledge. Moreover, the proposed analytical framework expands the search space for inventor–licensee matchmaking, which has traditionally relied on the personal relationships of TLO officers. The proposed analytical framework is applicable not only to inventor–licensee matchmaking in university technology licensing contexts but also in various contexts of practice, such as the identification of collaboration partners for new research or technology development.

Despite its contributions, this study has some limitations that should be addressed in future research. First, although several pre-trained models for language representation were tested in Subsection 5.2, the proposed analytical framework could be improved further by employing recently advanced language models, such as ELMo (Peters et al., 2018), BERT, and RoBERTa (Liu et al., 2019). Second, this study uses the titles of research outcomes and technology licensing contracts to represent technological functions and business requirements. However, other text data exist that contain richer information about technologies, such as the abstracts of patents or academic papers. Such text data should be used in future research and tested for their usefulness in representing technological functions and business requirements. Third, although a quantitative performance evaluation was conducted on the inventor–licensee matchmaking results derived by the proposed analytical framework, the quality of vector embedding itself was not considered in this study. In this respect, the validity of vector embeddings of words and sentences generated using the proposed analytical framework should be evaluated. Finally, this study considers only the matchmaking of a single inventor and licensee. However, in practice, university researchers often conduct research by organising a team with their colleagues, and companies sometimes develop new technologies in collaboration with other companies. Hence, the subject of the analysis should be extended to multiple inventors and licensees in future research to enhance the practicality of the proposed analytical framework. Nevertheless, we argue that the systematic processes of the proposed analytical framework for inventor–licensee matchmaking offer sufficient contribution to both current research and future practice.

Data availability

The data that has been used is confidential.

Acknowledgements

This work was supported by a Korea University Grant (K2312761) and the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIP) (No. 2020R1F1A1066472 and No. 2021R1A4A1033031).

Appendix A. Results of sensitivity analyses on hyperparameters associated with the fastText model

Table A.1
Results of the sensitivity analysis on the number of nodes in the hidden layer of the fastText model

Number of candidates (K)	Matching rate		
	The number of nodes in the hidden layer		
	100	200	300
1	0.4040	0.3815	0.3965
2	0.4688	0.4514	0.4589
3	0.5162	0.5037	0.5012
4	0.5411	0.5212	0.5362
5	0.5686	0.5436	0.5586
6	0.5810	0.5561	0.5736
7	0.5910	0.5711	0.5835
8	0.6035	0.5910	0.6035
9	0.6135	0.5985	0.6135
10	0.6160	0.6085	0.6160

Table A.2
Results of the sensitivity analysis on the set of hyperparameters of the fastText model(the number of nodes in the hidden layer = 300)

Hyperparameters			Matching rate @ 1 (Inventor level)	Hyperparameters			Matching rate @ 1 (Inventor level)
Window size	Number of negative examples	Number of training epochs		Window size	Number of negative examples	Number of training epochs	
3	3	500	0.3691	3	3	5000	0.3691
3	5	500	0.3641	3	5	5000	0.3741
3	7	500	0.3791	3	7	5000	0.3666
3	10	500	0.3666	3	10	5000	0.3741
5	3	500	0.3766	5	3	5000	0.3791
5	5	500	0.3741	5	5	5000	0.3791
5	7	500	0.3741	5	7	5000	0.3716
5	10	500	0.3716	5	10	5000	0.3766
7	3	500	0.3616	7	3	5000	0.3865
7	5	500	0.3766	7	5	5000	0.3716
7	7	500	0.3716	7	7	5000	0.3766
7	10	500	0.3691	7	10	5000	0.3815
10	3	500	0.3766	10	3	5000	0.3890
10	5	500	0.3791	10	5	5000	0.3890
10	7	500	0.3716	10	7	5000	0.3890
10	10	500	0.3815	10	10	5000	0.3741
3	3	2500	0.3716	3	3	10,000	0.3815
3	5	2500	0.3791	3	5	10,000	0.3741
3	7	2500	0.3666	3	7	10,000	0.3741
3	10	2500	0.3691	3	10	10,000	0.3766
5	3	2500	0.3741	5	3	10,000	0.3815
5	5	2500	0.3766	5	5	10,000	0.3791
5	7	2500	0.3716	5	7	10,000	0.3741
5	10	2500	0.3691	5	10	10,000	0.3791
7	3	2500	0.3791	7	3	10,000	0.3840
7	5	2500	0.3791	7	5	10,000	0.3766
7	7	2500	0.3815	7	7	10,000	0.3865
7	10	2500	0.3716	7	10	10,000	0.3766
10	3	2500	0.3815	10	3	10,000	0.3865
10	**5**	**2500**	**0.3965**	10	5	10,000	0.3840
10	7	2500	0.3840	10	7	10,000	0.3890
10	10	2500	0.3865	10	10	10,000	0.3915

References

Agrawal, A.K., Henderson, R., 2002. Putting patents in context: exploring knowledge transfer from MIT. In: Advances in Strategic Management, pp. 13–37.

Arora, S., Liang, Y., Ma, T., 2017. A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations.

Baglieri, D., Baldi, F., Tucci, C.L., 2018. University technology transfer office business models: one size does not fit all. Technovation 76–77, 51–63.

Battaglia, D., Landoni, P., Rizzitelli, F., 2017. Organizational structures for external growth of University Technology Transfer Offices: an explorative analysis. Technol. Forecast. Soc. Change 123, 45–56.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Bradley, S.R., Hayter, C.S., Link, A.N., 2013. Models and methods of university technology transfer. Foundations and Trends in Entrepreneurship 9, 571–650.

Brody, A.B., Kurth, S.W., Dempski, K.L., Liongosari, E.S., Kaplan, J.E., Swaminathan, K. S., 1999. Integrating disparate knowledge sources. In: Proceedings of Second International Conference on Practical Application of Knowledge Management, pp. 77–82.

Caviggioli, F., De Marco, A., Montobbio, F., Ughetto, E., 2020. The licensing and selling of inventions by US universities. Technol. Forecast. Soc. Change 159, 120189.

Chapple, W., Lockett, A., Siegel, D., Wright, M., 2005. Assessing the relative performance of U.K. university technology transfer offices: parametric and non-parametric evidence. Res. Pol. 34, 369–384.

Chau, V.S., Gilman, M., Serbanica, C., 2017. Aligning university–industry interactions: the role of boundary spanning in intellectual capital transfer. Technol. Forecast. Soc. Change 123, 199–209.

Clayton, P., Feldman, M., Lowe, N., 2018. Behind the scenes: intermediary organizations that facilitate science commercialization through entrepreneurship. Acad. Manag. Perspect. 32, 104–124.

Copas, J.B., Hilton, F.J., 1990. Record linkage: statistical models for matching computer records. J. Roy. Stat. Soc. 153, 287.

Debackere, K., Veugelers, R., 2005. The role of academic technology transfer organizations in improving industry science links. Res. Pol. 34, 321–342.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

Dey, D., Sarkar, S., De, P., 2002. A distance-based approach to entity reconciliation in heterogeneous databases. IEEE Trans. Knowl. Data Eng. 14, 567–582.

Dong, A., Pourmohamadi, M., 2014. Knowledge matching in the technology outsourcing context of online innovation intermediaries. Technol. Anal. Strateg. Manag. 26, 655–668.

Ganesh, M., Srivastava, J., Richardson, T., 1996. Mining entity-identification rules for database integration. In: Proceedings of the Second International Conference on Knowlegde Discovery and Data Mining, pp. 291–294.

Grimpe, C., Fier, H., 2010. Informal university technology transfer: a comparison between the United States and Germany. J. Technol. Tran. 35, 637–650.

Hong, S., Kim, J., Woo, H.G., Kim, Y.C., Lee, C., 2021. Screening ideas in the early stages of technology development: a word2vec and convolutional neural network approach. Technovation, 102407.

Hsu, D.W.L., Shen, Y.C., Yuan, B.J.C., Chou, C.J., 2015. Toward successful commercialization of university technology: performance drivers of university technology transfer in Taiwan. Technol. Forecast. Soc. Change 92, 25–39.

Jeon, D., Ahn, J.M., Kim, J., Lee, C., 2022. A doc2vec and local outlier factor approach to measuring the novelty of patents. Technol. Forecast. Soc. Change 174, 121294.

Kim, J., Lee, C., 2017. Novelty-focused weak signal detection in futuristic data: assessing the rarity and paradigm unrelatedness of signals. Technol. Forecast. Soc. Change 120, 59–76.

Kim, Y., Ahn, J., Kwon, O., Lee, C., 2019. Valuation of university-originated technologies: a predictive analytics approach. IEEE Trans. Eng. Manag. 68, 1813–1825.

Kim, J., Hong, S., Kang, Y., Lee, C., 2022a. Domain-specific valuation of university technologies using bibliometrics, Jonckheere–Terpstra tests, and data envelopment analysis. Technovation, 102664.

Kim, J., Lee, G., Lee, S., Lee, C., 2022b. Towards expert–machine collaborations for technology valuation: an interpretable machine learning approach. Technol. Forecast. Soc. Change 183, 121940.

Knoll, A., Röhrbein, F., Kuhn, A., Akl, M., Sharma, K., 2017. Neurorobotics. Informatik-Spektrum 40, 161–164.

Ko, N., Jeong, B., Yoon, J., Son, C., 2020. Patent-trademark linking framework for business competition analysis. Comput. Ind. 122, 103242.

Kotha, R., Crama, P., Kim, P.H., 2018. Experience and signaling value in technology licensing contract payment structures. Acad. Manag. J. 61, 1307–1342.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196.

Lee, C., Jeon, D., Ahn, J.M., Kwon, O., 2020. Navigating a product landscape for technology opportunity analysis: a word2vec approach using an integrated patent-product database. Technovation 96–97, 102140.

Lee, G., Kim, D., Lee, C., 2019. A sequential pattern mining approach to identifying potential areas for business diversification. Asian J. Technol. Innovat. 28, 21–41.

Lee, C., 2021. A review of data analytics in technological forecasting. Technol. Forecast. Soc. Change 166, 120646.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, pp. 1–12.

Mom, T.J.M., Oshri, I., Volberda, H.W., 2012. The skills base of technology transfer professionals. Technol. Anal. Strat. Manag. 24, 871–891.

Motohashi, K., Koshiba, H., Ikeuchi, K., 2021. New Indicator of Science and Technology Inter-relationship by Using Text Information of Research Articles and Patents in Japan. Research Institute of Economy, Trade and Industry (RIETI).

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2227–2237.

Qiang, B.H., Zhang, L., Xi, J.Q., 2008. Research on entities matching across heterogeneous databases. In: 2008 International Conference on Wireless Communications, Networking and Mobile Computing. WiCOM 2008, pp. 9–12.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP), pp. 3980–3990.

Sengupta, A., Ray, A.S., 2017. Choice of structure, business model and portfolio: organizational models of knowledge transfer offices in British universities. Br. J. Manag. 28, 687–710.

Shane, S., 2002. Selling university technology: patterns from MIT. Manag. Sci. 48, 122–137.

Siegel, D.S., Waldman, D.A., Atwater, L.E., Link, A.N., 2004. Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: qualitative evidence from the commercialization of university technologies. J. Eng. Technol. Manag. 21, 115–142.

Soares, T.J., Torkomian, A.L.V., 2021. TTO's staff and technology transfer: Examining the effect of employees' individual capabilities. Technovation 102, 102213.

Thayer, P., Martinez, H., Gatenholm, E., 2020. History and trends of 3D bioprinting. In: Methods in Molecular Biology, pp. 3–18.

Thursby, J.G., Jensen, R., Thursby, M.C., 2001. Objectives, characteristics and outcomes of university licensing: a survey of major U.S. universities. J. Technol. Tran. 26, 59–72.

Thursby, J.G., Thursby, M.C., 2004. Are faculty critical? Their role in university-industry licensing. Contemp. Econ. Pol. 22, 162–178.

Tseng, F.-C., Huang, M.-H., Chen, D.-Z., 2020. Factors of university–industry collaboration affecting university innovation performance. J. Technol. Tran. 45, 560–577.

Wu, A., Li, H., Dong, M., 2020. A novel two-stage method for matching the technology suppliers and demanders based on prospect theory and evidence theory under intuitionistic fuzzy environment. Appl. Soft Comput. 95, 106553.

Wu, Y., Welch, E.W., Huang, W.-L., 2015. Commercialization of university inventions: individual and institutional factors affecting licensing of university patents. Technovation 36–37, 12–25.