From Technology to Market. Bilingual Corpus on the Evaluation of Technology Opportunity Discovery

Amir Hazem, Chen Zhu, Kazuyuki Motohashi
The University of Tokyo
{amirhazem, motohashi}@tmi.t.u-tokyo.ac.jp, zhujohn0425@g.ecc.u-tokyo.ac.jp

Abstract

As companies aim to enhance and expand their product portfolios, Technology Opportunity Discovery (TOD) has gained increasing interest. To comprehend the role of emerging technologies in innovation, we introduce a novel technology-market corpus in English and Japanese languages, and conduct a comprehensive empirical evaluation of the linkage between technology and the market. Our dataset comprises English patents extracted from the USPTO database and Japanese patents extracted from the Japanese Patent Office (JPO), along with their associated products for each stock market company. We compare several static and contextualized word embedding methods to construct a technology-market space and propose an effective methodology based on a fine-tuned BERT model for linking technology to the market.

Keywords: Technology opportunity discovery, technology market corpus, patent product mapping

1. Introduction

Economic development and growth heavily depend on innovation and technological advances in the industrial sector. Therefore, more and more companies lean towards Technology Opportunity Discovery (TOD) to find the best and most effective opportunities for their business and research investments. A large variety of TOD approaches have been proposed so far (Yoon et al., 2015; Lee et al., 2020). They can be divided into two main directions: i) identifying and exploring emerging technologies (Kwon et al., 2018; Lee et al., 2022) and ii) exploiting existing technologies and products for diversification (Yoon et al., 2015; Kim et al., 2017; Motohashi and Zhu, 2023). Our work falls within the latter category which offers a more organic growth path for established companies (Cantwell and Piscitello, 2000), while the former carries a high range of uncertainty and is difficult to evaluate. TOD methods are usually based on patents to represent technologies and innovation (Lee et al., 2009, 2015, 2020), as they contain detailed and well-structured information about the developed technologies (patented inventions, domains, inventors, claims, etc.). Nonetheless, other data sources have also been explored such as Wikipedia (Kwon et al., 2018; Kim et al., 2019), online platforms and forums (Kwon et al., 2017; Kim and Lee, 2017), etc.

Patent data is undoubtedly the most appropriate source of information for storing technological knowledge across a wide spectrum of fields. Even though it has become the primary source for TOD methods, it exhibits some limitations in representing market-side information, specifically in terms of companies' products. (Motohashi and Zhu, 2023). Consequently, existing studies have turned to web content to supplement market-side innovation activities (Park and Geum, 2022). For instance, to monitor the marked-side innovation activities, Arora et al. (2013) and Motohashi and Zhu (2023) used web content by crawling companies websites, while Park and Geum (2022) collected technical news articles. This is mainly attributed to the widespread use of the Internet, and it has become commonplace for companies to disclose their commercial activities online by showcasing their products and services (Gök et al., 2014).

The main purpose of linking technology to market, and in other words, patents to products is to help inventors or patent owners to decide for which product a given patent can be used. If patent owners are usually companies that already have an idea about the product they will develop, it is not rare that many patents remain unused or can be applied in other fields to produce new products. To address this particular task, we investigate in this work several data representation and mapping techniques.

Within existing TOD methods based on crawled data to represent the market side, there are very limited publicly available technology-market datasets. Understanding the role of new technologies in innovation is the first step towards technology forecasting and technology opportunity discovery. For that purpose, we introduce a novel technology-market corpus in English and Japanese languages and conduct a comprehensive empirical evaluation of the linkage between technology and the market. Our dataset comprises English patents extracted from the USPTO database and Japanese patents extracted from the Japanese Patent Office (JPO), along with their associated products for each stock market company. The main contributions of this

LREC-COLING 2024, pages 7510–7520
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

work are as follows:

- We introduce two technology-market datasets in English and Japanese
- We conduct an extensive evaluation of existing methods for mapping technology to market in the US stock market
- We propose an effective way to address technology to market linkage by fine-tuning BERT on our proposed datasets

2. Related Work

Technology Opportunity Discovery (TOD), also known as technology opportunity analysis (Porter and Detampel, 1995), provides foresight analysis for developing technologies and insight into specific emerging technologies. TOD performs data analysis on collected bibliographic and/or patent information to provide the most adequate technology opportunities for firms and industries. Klevorick et al. (1995) distinguish three different sources of TOD: (i) advances in scientific understanding and technique; (ii) technological advances in other industries; (iii) feedback from technology. Olsson (2005) presents a model of technological opportunity and discusses the exploitation and regeneration of technological opportunities in terms of intentional incremental and radical innovations and unintentionally made discoveries.

Existing TOD studies focused on evaluating the technology based on bibliography of technical data (Bengisu and Nekhili, 2006; Curran and Leker, 2011; Lee et al., 2014) and based on the company's portfolio to find technology opportunities (Cho et al., 2016; Choi et al., 2019; Motohashi and Zhu, 2023). Recently, Lee et al. (2022) considered the growth potential of technology in the new technology-based firms (NTBF) investment perspective using deep learning-based text mining and a Knowledge Graph.

Lee et al. (2020) proposed a product landscape analysis to identify product areas as potential technology opportunities across multiple domains. They built a patent-product database from the US Patent and Trademark Database (USPTO) and used word2vec to construct a product landscape as a vector space model. This work is similar to ours in terms of technology-to-market representation; however, it differs in the way they exploit products. Their word2vec representation is based on patents-products associations or co-occurrences and does not use any textual content or description of the used products.

3. Datasets

Our Technology/Market dataset comprises: (i) a list of stock market companies, (ii) a set of patents and, (iii) a set of products. We built datasets in two languages: English and Japanese. We refer to the English corpus as USPTO-Market and the Japanese corpus as the JPO-Market corpus.

USPTO and JPO are a set of patents that represent existing technologies while "Market" refers to a set of released products in the market and for which at least one patent exists in the USPTO or JPO database.

3.1. USPTO-Market Corpus

The USPTO-Market corpus construction can be divided into four steps: 1- USPTO patent extraction within a period of time; 2- assignee alignment with their corresponding patents; 3- USPTO-Market company name matching and 4- collecting companies webpages from the web.

3.1.1. Patent Extraction

From the USPTO website, we first download the *g_patent*[1] file which contains data on granted patents until 2023. Then, we select a period of time that we want to process (2015- 2023). We chose not to cover older periods to reduce the risk of having companies that do not exist anymore and for which we will not be able to collect their products from their website. Of course, our script, which is available on GitHub [2], allows one to choose any period of time to generate the USPTO-Market dataset.

The patents include several information present in different files, all linked together thanks to the patent id. The patent date, title and abstract are extracted from the *g_patent* file[3], while the patent description is extracted from the *g_detail_desc_text.csv*[4]. Finally, the claims and the figures description are respectively extracted from the *g_claim.csv*[5] and the *draw_desc_text.csv* files [6]. We store the selected database inputs from 2015 to 2023, into a csv file that we name patent_pool_2015_2023.csv. The file contains the patent Id ("Patent_id"), its date

[1] https://patentsview.org/download/data-download-tables

[2] https://github.com/hazemAmir/TOD

[3] https://patentsview.org/download/data-download-tables

[4] https://patentsview.org/download/detail_desc_text

[5] https://patentsview.org/download/claims

[6] https://patentsview.org/download/draw_desc_text

Technology				Market
Date	Inputs	Company	#Patent	Products
2015	326,969	43,944	302,454	975
2016	334,674	44,927	310,265	990
2017	352,586	48,057	326,986	1,070
2018	341,104	48,042	316,130	1,092
2019	392,618	53,592	364,532	1,175
2020	390,572	54,276	362,135	1,233
2021	363,829	53,203	336,962	1,279
2022	360,417	53,828	332,507	1,202
2023	84,772	20,538	78,045	845
≥2015	2,947,541	188,110	2,730,016	1,734

Table 1: Number of patents for all the listed companies in the USPTO database (Inputs), number of companies (Company), distinct patents (#Patent) and the stock market companies that have at least one patent in the USPTO database (Products).

("Patent_date"), its title ("patent_title"), and its abstract ("patent_abstract").

Table 1 shows for the USPTO-Market dataset, the number of its inputs, the number of organizations (Companies) that have been granted at least by one patent, as well as the total number of granted patents per year. It also shows for the listed companies of the US stock market, the number of matched companies with the USPTO organizations. The number of matched companies is given for all the US stock market companies which includes Nasdaq, NYSE and some other stock markets. It is important to note that the companies listed in the USPTO and the companies listed in the US stock market are not an exact match. Hence, same companies may have variants in their names. We applied a heuristic algorithm to match the companies. This resulted in the "Products" column of Table 1.

3.1.2. Assignee Extraction

The used *g_patent* file to extract patents does not contain information about the assignees or the companies that have been granted with patents. This information can be found in the *g_assignee_disambiguated.tsv*[1]file. Using the patent id, we can retrieve for each patent its corresponding company or organization name.

3.1.3. USPTO-Market Match

The USPTO-Market corpus draws a linkage between patents (Technology) and products (Market) for companies listed in the US stock market. To list the companies of the US stock market, two main sources can be used : (i) Crunchbase[7] and (ii) Yahoo finance [8] combined with Nasdaq website[9]. Crunchbase provides business information about private and public companies. While it covers a large amount of entries (over 2M companies can be listed), the download process is limited (1000 inputs per download) which makes it tedious for large queries. On the other hand, Yahoo Finance via the Nasdaq website does not limit the download size, however, the listed companies are fewer (around 7k). In order to collect as many companies as possible, we used Crunchbase and ran several requests to download a list of US stock market companies. Our list contains about 7,246 listed companies that are located in the United States. Crunchbase also provides useful information about each company that is: its stock symbol, name, URL, and description.

Once the listed companies of the US stock market have been downloaded, we need to match them with the companies extracted from the USPTO database. It is noted that the same company can have name inconsistencies as illustrated in Table 2. To match company names, we apply a string match heuristic algorithm and obtain 1,734 matched companies.

3.1.4. Webpage Collection

It has become common for companies to exhibit their commercial activities (products and services) on the Internet. Hence, companies' websites are a valuable source of information about companies' products and so, about the market.

Crunchbase provides the homepage link for each listed company. Based on that, we crawl companies' web pages to construct their market representation. However, we do not use the entire content of the company's website, on the contrary, we limit our selection to the main page of the company and to the pages that describe each product. At the end of the crawling process, we extract two types of information: i) the full content of each selected page and ii) a subpart that corresponds to its keywords. The keywords are extracted using a dual attention model as described in (Motohashi and Zhu, 2023). We assume that the extracted keywords and key phrases are representative of each product.

Table 4 illustrates two examples extracted from the USPTO-Market dataset. The first example concerns a patent-product pair of the TXG stock symbol and the second example concerns a patent-

[7] https://www.crunchbase.com/

[8] https://finance.yahoo.com/screener/predefined/ms_basic_materials/

[9] https://www.nasdaq.com/market-activity/stocks/screener

Company/Organization Name	
Crunchbase	USPTO
airbnb	airbnb,inc.
netflix	netflix,inc
acer therapeutics	acer therapeutics inc.
axcella	axcella health inc.
monotype	monotype imaging inc
medidata	medidata solutions, inc
biote	biote medical, llc
acacia research	acacia research group llc
aegion	aegion coating services,llc
biospecifics technologies	biospecifics technologies corp.
blueprint medicines	blueprint medicines corporation
adobe	adobe systems incorporated
bentley systems	bentley systems, incorporated
orthofix	orthofix s.r.l.
xoma	xoma (us) llc
anavex	anavex life sciences corp.
bel fuse	bel fuse (macao commercial offshore)
braze	tokyo braze co., ltd.
bumble	bumble be holdings , llc
clene nanomedicine	clene nanomedicine, a nevada corp
ipower	ipower technology limited
kalvista pharmaceuticals	kalvista pharmaceuticals limited
lanzatech	lanzatech new zealand limited
keyw corporation	the keyw corporation
zagg	zagg intellectual property holding
web.com	web.com group, inc.
urban outfitters	urban outfitters wholesale, inc.
tuesday morning	tuesday morning partners, ltd.
the goodyear tire	rubber, the goodyear tire & rubber company
tenable	tenable network security, inc.

Table 2: Examples of company name inconsistencies between Crunchbase (Market) and USPTO database (Technology)

product pair of the RGEN stock symbol. Each example shows a patent abstract (Patent), a product description (Product) and a product web page content (Product(Web)).

3.2. JPO-Market Corpus

For the Japanese corpus that we refer to as the JPO-Market corpus, we followed the procedure described in (Motohashi and Zhu, 2023) to construct the dataset.

Patents were extracted from the database of the Ministry of Economy, Trade and Industry (METI)[10] for a period ranging from 2012 to 2021. METI only discloses the patent application number for each company. In order to obtain the patent content, we further linked the extracted inputs to the Japan Patent Office (JPO)[11].

Technology		Market
Date	#Patent	Products
2012	99,937	1,125
2013	104,931	1,122
2014	113,997	1,135
2015	118,312	1,185
2016	119,003	1,197
2017	112,142	1,228
2018	54,112	1,033
2019	18,923	788
2020	8,360	599
2021	2,545	321

Table 3: Number of patents for all the listed companies in the JPO database (#Patent), and the stock market companies that have at least one patent in the JPO database (Products).

For the market side, we collected financial statements of listed companies released by the Japanese Financial Service Agency (FSA) [12]. We obtained 3,189 listed companies for which web information was available based on the publicly available website [13] to search for the company's homepage using its unique corporate identifier (houjinbango).

Table 3 shows for the JPO-Market dataset, the number of organizations and companies that have been granted at least one patent (Products), as well as the total number of granted patents per year (#Patents). Unlike the USPTO database, there are no inconsistencies in companies names for the JPO-Market corpus.

4. Task Definition

Given a set of stock market listed companies[14], $C = \{C_1, C_2, ..., C_n\}$. Each company ($C_i$) is represented by a technology/market pair (C_i^t, C_i^m) of patents and products, where C_i^t is the technology representation of the company and C_i^m its market representation. The technology representation of a company i is $C_i^t = \{t_1, t_2, ..., t_m\}$ where t_1 for instance is a given technology of the company C_i. The market representation of the same company i is $C_i^m = \{m_1, m_2, ..., m_k\}$ where m_1 is its corresponding market.

To represent technology, we use patent abstracts. Hence, the technology representation of a company

[10] https:////info.gbiz.go.jp/hojin/DownloadTop

[11] https://www.gazette.jpo.go.jp/scciidl010

[12] https://www.fsa.go.jp/

[13] https://houjin.jp

[14] Nasdaq companies of the Us Market for instance

	Patent-Product pair (English)
	Example 1 (TXG)
Patent	This disclosure provides methods and compositions for sample processing, particularly for sequencing applications. Included within this disclosure are bead compositions, such as diverse libraries of beads attached to large numbers of oligonucleotides containing barcodes. Often, the beads provides herein are degradable. For example, they may contain disulfide bonds that are susceptible to reducing agents...
Product	10x Genomics is creating revolutionary DNA sequencing technology to help researchers better identify subtle variations that are overlooked by technologies that shred biological samples into tiny fragments before sequencing the short stretches and using computers to assembling them into a genome.
Product (Web)	Single cell gene expression of the transcriptome and epigenome in every profile open chromatin from same with chromium, multiply your power discovery to characterize types states, uncover regulatory programs view product writing buyer two detection accessibility, define new interactions rare populations precision flexible hundreds tens thousands cells per sample, streamlined data simultaneously software efficient lab library days hidden profiling at resolution...
	Example 2 (RGEN)
Patent	Methods and systems of harvesting a cell product from a cell culture by culturing cells in a fluid medium until the cells have produced a cell product at a harvest concentration are disclosed. the cells are cultured in a cell culture system including a bioreactor connected to an atf device. The methods include draining fluid medium from the bioreactor through the outlet and the atf device until the bioreactor volume reaches a predetermined volume, and the atf column yields at an atf outlet...
Product	Repligen corporation is a life sciences company focused on the development and commercialization of high-value consumable products used in the process of manufacturing biological drugs. our bioprocessing products are sold to major life sciences and biopharmaceutical companies worldwide. we are the leading manufacturer of protein a affinity ligands, a critical component of protein a resins that are used to separate and purify monoclonal antibody therapeutics...
Product (Web)	Contact repligen login shop menu solutions modalities unit operations products support company overview process intensification atmp ultrafiltration/diafiltration continuous manufacturing fluid management analytics covid antibodies | proteins viral vectors mrna pdna emerging cell culture chromatography uf/df upstream filtration downstream supplements ligands product configurators dialysis knowledge base about leadership careers...

Table 4: Examples of Patents (Technology) and their corresponding products (Market) extracted from the English USPTO-Market dataset. Product stands for a product using its description provided by Crunchbase while Product(Web) stands for a product using its web content.

i can be rewritten as follows: $C_i^t = \{p_1, p_2, ..., p_m\}$ where p_1 for instance is a granted patent of the company C_i. For the market side, we use the products of companies which can be represented either by company descriptions (provided by Crunchbase) or by companies web content crawled from the web. Hence, the market representation can be rewritten as $C_i^m = \{d_1, d_2, ..., d_k\}$ where d_1 for instance is the product' description of company C_i or as $C_i^m = \{w_1, w_2, ..., w_k\}$ where w_1 for instance is the product' web content of company C_i.

The task consists in linking, each company's technology (C_i^t) to its corresponding market (C_i^m). In other words, we aim at building a concordance matrix or a mapping space between technology (patents) and the market (products) which leads to a technology-market matrix. This is done based on textual data which consists in patent abstracts for the technology side and company description or web pages content for the market side.

A company may have several patents and several products, in that case, we need a single representation for all the patents and a single representation for all the products of the company. Each representation is computed by aggregating all the patents in one embedding vector (the same process is done for the products). Finally, each technology representation for a company C_i^t is given by:

$$C_i^t = \sum_{j=1}^{n}(Emb(p_j)) \quad (1)$$

where n represent the number of patents p of the company C_i^t.

$$Emb(p_j) = \sum_{l=1}^{k} Emb(w_l) \quad (2)$$

where w are the words contained in the patent's abstract.

Similarly, each market representation for a company C_i^m using products descriptions is given by:

$$C_i^m = Emb(d_i) \quad (3)$$

where d_i represents the description of the company C_i^m.

$$Emb(d_i) = \sum_{l=1}^{k} Emb(w_l) \quad (4)$$

where w are the words contained in the product's description or in its web content.

5. Technology to Market Methods

In order to build a Technology-Market linkage we consider three categories of methods based on the following assumptions:

1. no mapping is needed to link technology and market. We assume that technology and market can be represented in the same static word embedding space and linked together using the cosine similarity. We refer to this category of methods as: Static Space.

2. a linear mapping is needed to map technology to market. We use a linear regression model as baseline and VecMap (Artetxe et al., 2016) method originally used for cross-lingual embedding mapping. We refer to this category of methods as: Linear Mapping Space.

3. a non linear mapping is needed to build a joint space between technology and market. We use BERT (Devlin et al., 2018) as a binary classifier to link both representations. We refer to this category of methods as: Joint Space

We describe each category and its corresponding methods in the following Sections.

5.1. Static Space

In the static space approach, we represent both technology and the market in the same embedding space. Technology representation for each company is computed by aggregating all its patents in one embedding vector. Also, the market representation for each company is computed by aggregating all its products (product description or web content) in one embedding vector. To link technology to the market, we simply compute the Cosine similarity between the aggregated technology and market embedding vectors. We consider several embedding representations [15] but we only report the two embedding models that obtained the best results in our experiments: i) Word2Vec (W2V) (Mikolov et al., 2013) and ii) sentence Bert (SBERT) (Reimers and Gurevych, 2019). When using W2V for patent representation, for instance, we sum up the embeddings of each word of the patent's abstract. While using SBERT, we obtain a sentence embedding representation of the patent's abstract. This can be translated in the following equations:

Each technology representation for a company C_i^t can be given by:

$$C_i^t = \sum_{j=1}^{p}\sum_{l=1}^{k} W2V(w_{jl}) \quad (5)$$

where p is the number of granted patents of the company and k is the number of words contained in the patent's abstract.

Similarly, each market representation for a company C_i^m using products descriptions is given by:

$$C_i^m = \sum_{l=1}^{k} W2V(w_l) \quad (6)$$

where k is the number of words contained in the description or in the web content of the company.

Using SBERT, the aggregation is already embedded in its representation which gives the following equations:

$$C_i^t = \sum_{j=1}^{n} SBERT(p_j) \quad (7)$$

where n is the number of granted patents (p) of the company. Its market representation is given by:

$$C_i^m = SBERT(d_i) \quad (8)$$

where d_i is the description of the company' products.

5.2. Linear Mapping

In this scenario, we assume that there is a linear mapping between the technology and the market space. We experiment with two linear mapping

[15] Other embedding representations such as fastText, Glove and Gpt2 were tested but the results were lower than word2vec and SBERT

techniques. The first one is the well-known linear regression that we refer to as: LReg in our experiments, and the second approach is VecMap (Artetxe et al., 2016).

VecMap is a mapping technique that uses the orthogonality constraint and a global preprocessing with length normalization and dimension-wise mean centering to map data from a source space into a target space. If it was initially proposed for bilingual induction tasks using cross-lingual word embeddings, it can be applied in any scenario where mapping data is required as long as a mapping dictionary exists. We adapt the VecMap method in our case and use it to map technology to the market.

In the bilingual scenario, VecMap uses a bilingual dictionary to learn a linear mapping that minimizes the distances between equivalences listed in a bilingual dictionary. In our case, instead of a bilingual dictionary, we use a patent-product dictionary to train VecMap.

Both LReg and VecMap use word embeddings to initialize the training model. In the same way as for the static space scenario, we use W2V and SBERT for initialization.

5.2.1. Joint Space

In our third scenario, we assume the existence of a nonlinear relation between technology and the market. To test our hypothesis we use BERT (Devlin et al., 2018) as a binary classifier to build a technology-market joint space.

BERT is a supervised learning model that has been trained on the Masked Language Model (MLM) objective and Next Sentence Prediction (NSP). For NSP, the model takes as input pairs of sentences and learns to predict if the second sentence is the subsequent sentence of the first one. We similarly fine-tune our model, providing positive and negative technology/market sentence pairs. In a similar methodology to NSP, we select 50% of the training pairs where the second sequence is a product of the company, while the other 50% pairs consist of randomly chosen products. Therefore, for each positive training pair, we randomly select negative patent-product pairs. The intuition here is that BERT will learn shared features between patents and products the same way it does for subsequent pairs of sentences. For the USPTO-Market corpus, we experiment Bert-base and for the JPO-Market corpus, we experiment with BERT for Japanese. Also, for both datasets, we experiment with multilingual BERT.

6. Experiments and Results

6.1. Experiments

To evaluate the technology-market mapping, we conducted two sets of experiments. The first experiment was conducted on the USPTO-Market dataset which comprises around 1200 training companies having granted patents from 2015 to 2016 and 460 test companies having granted patents from 2017 to 2023. The second experiment was conducted on the the Japanese JPO-Market dataset which comprises around 1300 training companies and 300 test companies. From the market side, for the JPO-Market experiment, the information about products is extracted from the company's website. Hence, we consider two scenarios, the first one consists of using the entire content of the website, while in the second scenario, to get rid of noise and extra information present in companies' webpages, we only extract keywords from the website. For the USPTO-Market corpus, Crunchbase, used for extracting the stock market companies, provides a description for each company. We use these descriptions as a clean source of information about the products that we contrast with their corresponding web page information.

6.2. Results

Table 5 shows the obtained results on the technology to market alignment for the USPTO-market dataset. The results are divided into three types of methods: 1) Static space approaches using word2vec (W2V) (Mikolov et al., 2013) and sentence BERT (SBERT) (Reimers and Gurevych, 2019); 2) Linear mapping space approaches using the two mapping techniques: VecMap (Artetxe et al., 2016) and the linear regression model (LREG); and finally 3) Joint space approaches based on a fine-tuned BERT classifier using Bert-base-cased (Bert-base) and multilingual BERT (Bert-multi) (Devlin et al., 2018). Table 5 also shows two data source categories: 1) Using company descriptions (Desc) provided by Crunchbase; and 2) using companies web content (Web(Full)) previously collected from the web.

For static space methods, SBERT obtained the best results for both company sources (Desc and Web) with an overall best accuracy of 88.70% and a Map score of 37.14%. We notice a drop in performance when using web content. This is not surprising due to the noise carried by the web content. However, the drop in performance is more important for W2V than the drop observed for SBERT.

For the linear mapping methods, LReg(SBERT) showed the best performance with an overall Map score of 34.75% while VecMap(SBERT) obtained the best accuracy (90.65%). Here also we note that

	Desc		Web(Full)	
	AC	MAP	AC	MAP
Static Space				
W2V	71.74	14.73	53.48	8.63
SBERT	88.70	37.14	78.91	33.00
Linear Mapping				
LReg(W2V)	72.39	13.59	57.61	10.34
LReg(SBERT)	88.04	34.75	78.91	31.55
VecMap(W2V)	77.39	14.17	58.70	8.49
VecMap(SBERT)	**90.65**	28.22	77.83	22.28
Joint Space				
Bert-base	87.66	**42.02**	**81.15**	**35.55**
Bert-multi	85.25	37.72	77.11	32.68

Table 5: Technology to Market Alignment Accuracy (Ac@100) and Mean Average Precision (Map) for the USPTO-Market test set. Best results of each block are underlined and overall best results are given in bold.

	Web(Full)		Web(KW)	
	AC	MAP	AC	MAP
Static Space				
W2V	37.66	2.34	34.66	1.62
SBERT	34.00	3.50	27.33	2.44
Linear Mapping				
LReg(W2V)	66.33	6.52	66.00	6.08
LReg(SBERT)	61.66	8.05	59.33	6.68
VecMap(W2V)	44.66	5.88	49.00	6.82
VecMap(SBERT)	37.66	3.15	31.66	1.82
Joint Space				
Bert-Japanese	**79.33**	**11.65**	**76.00**	9.73
Bert-multi	76.33	10.10	64.33	**11.12**

Table 6: Technology to Market Alignment Accuracy (Ac@100) and Mean Average Precision (Map) for the JPO-Market test set. Best results of each block are underlined and overall best results are given in bold.

using company description allows to obtain better performance than using web content. Surprisingly, VecMap which is very effective in bilingual mapping (Artetxe et al., 2016) showed lower performance in MAP score. We also remark that using SBERT embeddings for initialization is more effective than using W2V for both LReg and VecMap.

Finally, for the joint pace methods, Bert-base obtained better results than multilingual BERT (Bert-multi) for both data types. Here also we denote a drop in performance when using the web content of companies.

Overall, if we compare static space, linear mapping, and joint space methods, we see that Bert-base used to construct a fine-tuned joint space model obtained the best performance with a Map score of 42.02% (Desc) and a Map score of 35.55% (Web(Full)). We note also that VecMap(SBERT) obtained the highest accuracy of 90.65% (Desc). If we look at the results using the web content, we see that Bert-base obtained the highest accuracy score of 81.15%.

Table 6 shows the obtained results on the JPO-Market dataset. Here, the company descriptions are not provided. Instead, we only use web based content in two manners: 1) by using the full content of the web pages (Web(Full)); 2) by selecting web page keywords that describe companies products (Web(KW)).

For static space methods, and on the contrary to previous experiments where SBERT was the most effective method, in the Japanese experiments however, W2V obtained the best accuracy with an overall score of 37.66%. In terms of MAP however, SBERT obtained the best overall Map score of 3.50%. We Also notice that using keywords only (Web(KW)), produces a drop in performance. If further investigations are certainly needed, this suggests that important joint information is lost when selecting keywords.

For linear mapping techniques, LReg(W2V) obtained the best results in terms of accuracy while LReg(SBERT) obtained the best Map score (8.05%) for Web(Full) and VecMap(W2V) obtained the best Map score (6.82%) for Web(KW). We also note that linear mapping methods outperformed the static space methods which is different from what we observed in the English USPTO-Market experiments. However, it is important to note that we do note have clean descriptions in the Japanese dataset.

Finally, for the joint space models, the fine-tuned Bert-Japanese obtained the best results except for Web(KW) where Bert-multi obtained the highest Map score of 11.12%.

Overall, by comparing the three types of representations we see that building a joint space model with Bert obtained, here also, the best results with 79.33% of accuracy and a Map score of 11.65% for Bert-Japanese (Web(Full)) and 76.00% of accuracy (Web(KW)). Multilingual BERT (Bert-multi) obtained the best Map score of 11.12% (Web(KW)).

It is to note that BERT is less sensitive to using full web page content or only keywords. If BERT has been trained on full sentences, its masked language model technique as well as an appropriate fine-tuning make BERT efficient for both: full sentences and for sequences of keywords.

7. Discussion

Using a static space is a simple and straightforward procedure as it does not require any training or fine-tuning beforehand and also it can be easily deployed in a real-world scenario. If this approach did not obtain the best results it remains a good alternative as it showed some encouraging results under some conditions using SBERT for English and W2V for Japanese. It is worth noticing that SBERT showed competitive results when compared to Bert-base for the English dataset. SBERT is a BERT-based model that has been trained for efficient sentence representation. It can also be considered as a joint space model for semantic sentence pair similarity.

VecMap, which has shown high performance on bilingual lexicon induction task (Artetxe et al., 2016, 2018b,a), obtained mitigating results on the alignment of patents to products and so, failed to build an effective technology to market space that maximizes the Map score. This may be due to the fact that patent-product relations are somehow different from translation pair characteristics.

Another interesting remark is that static space methods showed better performance than linear mapping methods. This suggests that linear mapping may not be the best choice to link technology to market (This remark can be tempered by the observations on the Japanese dataset). In fact, this suggests to use of non-linear mapping methods which we proposed in this work by fine-tuning BERT.

Overall, building a joint space by fine-tuning BERT on a patent-product pair of sentences is the most effective way to map technology to market. The obtained results on both English and Japanese datasets confirm the capacity of BERT to capture similarities between pairs of sentences in our task and hence, to build an effective joint space that maps together patents and products. These remarkable results pave the way to the next step for future work on technology opportunity discovery. Indeed, at this stage, we are able to map technology to market but we did not experiment with the prediction of new technologies for the same company. If we let this work for the near future, our corpus provides timestamp information that allows us to train our model on a given period of time and to predict the company's products for another period.

8. Conclusion

In order to address the technology opportunity discovery task, we introduced two technology-to-market datasets in English and Japanese. We conducted a comprehensive evaluation of existing methods and showed under which conditions static-based and contextual-based word embeddings can be used for the alignment of patents to products. We also proposed an effective BERT-based approach to map technology to market by fine-tuning BERT using a set of patent-product training pairs. Our findings suggest that it is possible to build a recommendation system for companies to help them choose new strategies based on their existing technologies and to open up diversification in other fields. If more investigations and real-world case studies are certainly needed, we hope that these promising results along with the proposed data sets will encourage future investigations and work in the technology opportunity discovery field.

9. Limitations

Our proposed methodology allows us to empirically evaluate the alignment between patents and their corresponding products which shows that mapping technology to market is possible. However, as of now, it does not evaluate the technology opportunity discovery, to do so, we need to portray for each company its evolution in terms of technology and products given a period of time. Then, try to predict their trajectory using our proposed methods. If this evaluation has not been conducted yet, the proposed datasets contain the timeline information and can be used for such evaluation. We are working on building such an annotated evaluation methodology for our future works.

10. Acknowledgements

This research is supported by the Research Center for Advanced Science and Technology (RCAST) of The University of Tokyo.

11. Bibliographical References

Sanjay K. Arora, Jan Youtie, Philip Shapira, Lidan Gao, and TingTing Ma. 2013. Entry strategies in an emerging technology: a pilot web-based study of graphene firms. *Scientometrics*, 95(3):1189–1207.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Murat Bengisu and Ramzi Nekhili. 2006. Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting and Social Change*, 73(7):835–844.

John Cantwell and Lucia Piscitello. 2000. Accumulating technological competence: Its changing impact on corporate diversification and internationalization. *Industrial and Corporate Change*, 9:21–51.

Chanwoo Cho, Byungun Yoon, Byoung-Youl Coh, and Sungjoo Lee. 2016. An empirical analysis on purposes, drivers and activities of technology opportunity discovery: the case of korean smes in the manufacturing sector. *R&D Management*, 46(1):13–35.

Jaewoong Choi, Byeongki Jeong, and Janghyeok Yoon. 2019. Technology opportunity discovery under the dynamic change of focus technology fields: Application of sequential pattern mining to patent classifications. *Technological Forecasting and Social Change*, 148:119737.

Clive-Steven Curran and Jens Leker. 2011. Patent indicators for monitoring convergence – examples from nff and ict. *Technological Forecasting and Social Change*, 78(2):256–273. Using Technological Intelligence for Strategic Decision Making in High Technology Environments.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abdullah Gök, Alec Waterworth, and Philip Shapira. 2014. Use of web mining in studying innovation. *Scientometrics*, 102:653 – 671.

Hyunwoo Kim, Suckwon Hong, Ohjin Kwon, and Changyong Lee. 2017. Concentric diversification based on technological capabilities: Link analysis of products and technologies. *Technological Forecasting and Social Change*, 118:246–257.

Jieun Kim and Changyong Lee. 2017. Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. *Technological Forecasting and Social Change*, 120:59–76.

Juram Kim, Seungho Kim, and Changyong Lee. 2019. Anticipating technological convergence: Link prediction using wikipedia hyperlinks. *Technovation*, 79:25–34.

Alvin K. Klevorick, Richard C. Levin, Richard R. Nelson, and Sidney G. Winter. 1995. On the sources and significance of interindustry differences in technological opportunities. *Research Policy*, 24(2):185–205.

Heeyeul Kwon, Jieun Kim, and Yongtae Park. 2017. Applying lsa text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation*, 60:15–28.

Heeyeul Kwon, Yongtae Park, and Youngjung Geum. 2018. Toward data-driven idea generation: Application of wikipedia to morphological analysis. *Technological Forecasting and Social Change*, 132:56–80.

Changyong Lee, Daeseong Jeon, Joon Mo Ahn, and Ohjin Kwon. 2020. Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database. *Technovation*, 96-97:102140.

Changyong Lee, Bokyoung Kang, and Juneseuk Shin. 2015. Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change*, 90:355–365.

MyoungHoon Lee, Suhyeon Kim, Hangyeol Kim, and Junghye Lee. 2022. Technology opportunity discovery using deep learning-based text mining and a knowledge graph. *Technological Forecasting and Social Change*, 180:121718.

Sungjoo Lee, Byungun Yoon, and Yongtae Park. 2009. An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, 29(6):481–497.

Yongho Lee, So Young Kim, Inseok Song, Yongtae Park, and Juneseuk Shin. 2014. Technology opportunity identification customized to the technological capability of smes through two-stage patent analysis. *Scientometrics*, 100(1):227–244.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Kazuyuki Motohashi and Chen Zhu. 2023. Identifying technology opportunity using dual-attention model and technology-market concordance matrix. *Technological Forecasting and Social Change*, 197:122916.

Ola Olsson. 2005. Technological opportunity and growth. *Journal of Economic Growth*, 10(1):35–57.

Mingyu Park and Youngjung Geum. 2022. Two-stage technology opportunity discovery for firm-level decision making: Gcn-based link-prediction approach. *Technological Forecasting and Social Change*, 183:121934.

Alan L. Porter and Michael J. Detampel. 1995. Technology opportunities analysis. *Technological Forecasting and Social Change*, 49(3):237–255.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Janghyeok Yoon, Hyunseok Park, Wonchul Seo, Jae-Min Lee, Byoung youl Coh, and Jonghwa Kim. 2015. Technology opportunity discovery (tod) from existing technologies and products: A function-based tod framework. *Technological Forecasting and Social Change*, 100:153–167.