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

Amir Hazem, Chen Zhu and Kazuyuki Motohashi

RCAST, The University of Tokyo

LREC-COLING 2024

Outline

- Context and Motivation
- Task and Applications
- Datasets
- Methods and Evaluation
- Conclusion and perspectives

Context and Motivation

- Economic development and growth heavily depends on innovation and technological advances in the industrial sector
- Companies aim to enhance and expand their product portfolios
- Technology Opportunity Discovery (TOD) has gained increasing interest
- More and more companies lean towards TOD to find the best and most effective opportunities for their business and research investments

Contributions

- 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

Related Work

- Variety of TOD approaches have been proposed so far Yoon et al. (2015); Lee et al. (2020)
- Can be divided into two main directions:
  - identifying and exploring emerging technologies Kwon et al. (2018); Lee et al. (2022)
    - carries a high range of uncertainty and is difficult to evaluate
  - exploiting existing technologies and products for diversification Yoon et al. (2015); Kim et al. (2017); Motohashi and Zhu (2023)
    - offers a more sustainable growth path for companies Cantwell and Piscitello (2000)

Technologies and Innovation Data

TOD methods

- usually based on patents to represent technologies Lee et al. (2009, 2015, 2020)
- patents contain detailed and well structured information about technologies (patented inventions, domains, inventors, claims, etc.)
- other data sources have 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)

Patent Limitations

Technology side

- patent data has become the primary source for TOD methods
- patents show some limitations to represent the market-side information Motohashi and Zhu (2023)

Market side

- to reflect the marked-side, companies exhibit their commercial activities on the Internet by introducing their products and services Gök et al. (2014)
- web content by crawling companies websites Arora et al. (2013) and Motohashi and Zhu (2023) or using technical news articles Park and Geum (2022)

Task and Applications

Task linking technology to market (patents to products)

- help inventors or patent owners to decide for which product a given patent can be used
- many patents remain unused or can be applied in other fields to produce new products

Technology/Market Datasets

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

- USPTO and JPO are a set of patents that represents existing technologies
- "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

Datasets: USPTO-Market Corpus

The USPTO-Market corpus construction can be divided into four steps:

1. USPTO patent extraction within a period of time (2015-2023)
2. assignee alignment with their corresponding patents
   - *g_patent* does not contain information about the assignees or the companies that have been granted with patents
3. USPTO-Market company name matching
   - We used Crunchbase to list the companies of the US stock market (stock symbol, name, URL, and description)
   - match listed companies from Crunchbase with the companies extracted from the USPTO database
4. collecting companies webpages from the web

Datasets: USPTO-Market Corpus (Name matching)

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
adobe	adobe systems incorporated
bel fuse	bel fuse (macao commercial offshore)
the goodyear tire	rubber, the goodyear tire & rubber company
tenable	tenable network security, inc.

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

Datasets: USPTO-Market Corpus

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
$\geq$2015	2,947,541	188,110	2,730,016	1,734

Table: 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)

Datasets: USPTO-Market Corpus (Example)

Patent-Product pair (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...
Prod	10xGenomics 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.
Prod (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...

Table: Example of a Patent (Technology) and its corresponding product (Market) extracted from the English USPTO-Market dataset

Datasets: JPO-Market corpus

- Patents were extracted from the database of the Ministry of Economy, Trade and Industry (METI)[1] for a period ranging from 2012 to 2021
- METI only disclose 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)
- For the market side, we collected financial statements of listed companies released by the Japanese Financial Service Agency (FSA)
- We obtained 3,189 listed companies for which web information was available based on the publicly available website to search for the companies homepage using its unique corporate identifier (houjinbango)

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

Datasets: JPO-Market Corpus

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

Technology to Market Alignment Methods

Technology-Market linkage $--> $ three assumptions

1. no mapping is needed to link technology and market
   - can be represented in the same static word embedding space
   - linked together using the cosine similarity
2. a linear mapping is needed to map technology to market
   - linear regression model
   - VecMap Artetxe et al. (2016)
3. a non linear mapping is needed to build a joint space
   - use BERT Devlin et al. (2018) as a binary classifier
   - fine-tuned on a small technology-to-market training dataset

Experiments

- USPTO-Market
  - train: 1200 patent/product pairs
  - test: 460 patent/product pairs
- JPO-Market
  - train: 1300 patent/product pairs
  - test: 300 patent/product pairs

Results (USPTO-Market)

	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: 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

Results (JPO-Market)

	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: 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

Conclusion

- introduced two technology to market datasets in English and Japanese
- 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
- proposed an effective BERT-based approach to map technology to market by fine-tuning BERT

Future Work

- In-domain analysis and evaluation
- Inject complementary information given by patent claims and description
- Propose novel architectures specifically dedicated to technology-market linkage

Thank you very much for your attention!

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.

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

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.

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.

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.

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.

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.