Technology Identification from Patent Texts: A Novel Named Entity Recognition Method

Giovanni Puccetti[a,e,1,*], Vito Giordano[b,e], Irene Spada[b,e], Filippo Chiarello[c,e,f], Gualtiero Fantoni[d,e,f]

[a]*Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy*
[b]*Department of Information Engineering, Via Girolamo Caruso 16, 56122 Pisa, Italy*
[c]*Department of Energy, Systems, Territory, and Construction Engineering, Largo Lucio Lazzarino 2, 56122 Pisa, Italy*
[d]*Department of civil and industrial engineering, Largo Lucio Lazzarino 2, 56122 Pisa, Italy*
[e]*Business Engineering for Data Science Lab—B4DS, University of Pisa, 56120 Pisa, Italy*
[f]*Erre Quadro S.R.L, Largo Padre Renzo Spadoni c/o Cittadella Galileiana, 56126 Pisa, Italy*

Abstract

Identifying technologies is a key element for mapping a domain and its evolution. It allows managers and decision makers to anticipate trends for an accurate forecast and effective foresight. Researchers and practitioners are taking advantage of the rapid growth of the publicly accessible sources to map technological domains. Among these sources, patents are the widest technical open access database used in the literature and in practice.

Nowadays, Natural Language Processing (NLP) techniques enable new methods for the analysis of patent texts. Among these techniques, in this paper we explore the use of Named Entity Recognition (NER) with the purpose to identify the technologies mentioned in patents' text. We compare three different NER methods, gazetteer-based, rule-based and deep learning-based (e.g. BERT), measuring their performances in terms of precision, recall and computational time.

We test the approaches on 1,600 patents from four assorted IPC classes as case studies. Our NER systems collected over 4,500 fine-grained technologies, achieving the best results thanks to the combination of the three methodologies. The proposed method overcomes the literature thanks to the ability to filter generic technological terms. Our study delineates a valid technology identification tool that can be integrated in any text analysis pipeline to support academics and companies in investigating a technological domain.

*Corresponding author
Email addresses: giovanni.puccetti@sns.it (Giovanni Puccetti), vito.giordano@phd.unipi.it (Vito Giordano), irene.spada@phd.unipi.it (Irene Spada), filippo.chiarello@unipi.it (Filippo Chiarello), gualtiero.fantoni@unipi.it (Gualtiero Fantoni)
[1]*Phone number*: +39 3519443040

Keywords: Information Retrieval, Named Entity Recognition, Natural Language Processing, Patents, Technology Analysis

---

1. Introduction

The analysis and observation of technologies is a fundamental part of technological management, especially for planning R&D policies by governments and companies (Yoon and Park, 2005). Given the present pace of innovation, market actors need a deeper understanding of a technological domain, to mitigate the uncertainty typical of the digital ages (Robinson et al., 2013).

In the literature there exists many methods for technological mapping and forecasting, that can help decision makers in predicting core and emerging technologies (Huang et al., 2020), diffusion of technologies (Daim et al., 2006), convergence (Karvonen and Kässi, 2013) and portfolio analysis (Ernst, 2003). Patents are among the best source of information to reach these goals (Joung and Kim, 2017) because they contain rich technical details. Anyway, they are also among the most complex textual sources to analyse automatically because of the mix of technical and juridical jargon.

A patent provides the technical description of an invention in a sufficiently clear and complete manner to enable a person skilled in the art (i.e. someone having the relevant technical information publicly disclosed at the time of the invention) to replicate the invention without any additional creative activity (Art. 83 EPC) (Lidén and Setréus, 2011). A patent is designed to disclose the minimum content that makes understandable and reproducible the invention. The use of juridical jargon makes the document legally binding, eligible to protect the invention.

Patent texts create a barrier to the access to one of the widest technical open access resources. It is believed that between 70% and 90% of the information about technologies can only be found in patents (Asche, 2017). Patent analysis is a valuable approach for deriving information about an industry or technology for forecasting (Daim et al., 2006), competitive analysis (Thorleuchter et al., 2010), technology trend analysis (Tseng et al., 2011), and for avoiding infringement (Yu and Zhang, 2019). This information may be obtained by the use of either bibliometric analysis and text mining. The former involves the analysis of meta-data, such as citations, assignee, inventors, and International Patent Classification classes (Cho and Kim, 2014). The latter aims to extract relevant information from unstructured text, that includes title, abstract, claims, state-of-the-art description, and other records (Tseng et al., 2007).

Given the complexity of patent texts, bibliometric analysis is more common in literature since it involves the analysis of well organized and structured data, more easily processable than unstructured data. However, the analysis of metadata is not able to capture the detailed technical content of patents (Kuhn et al., 2020; Righi and Simcoe, 2019). As affirmed by Vicente-Gomila et al. (2021), text mining enhances traditional measures based on bibliometric data in forecasting technological change.

Recently it has been shown that text mining is more effective than metadata analysis for measuring novelty and impact of a patent. In Arts et al. (2021), the authors use text mining techniques to identify new technologies and measure patent novelty, detecting uni-grams (single word, well-known as keywords), bi-

grams and tri-grams (two or three consecutive words, also called keyphrases) in title, abstract, or claims of a patent. They consider as emerging technologies a word or a words combination appearing for the first time in the text, achieving remarkable results with generic keywords and keyphrases. Other studies involve text mining from patent for identifying emerging technologies (Ranaei et al., 2020; Zhou et al., 2020; Porter et al., 2019; Jang et al., 2021; Sarica et al., 2020) or exploring the convergence phenomena among technological fields (Gustafsson et al., 2015; Song et al., 2017). At the state-of-the-art, text mining techniques for technological analysis focus on generic terms and not on specific ones for investigating the technological change. This is a limitation given the quantity and quality of technical information contained in patents, therefore novel approaches are needed to enable these contents.

In this paper, we use Natural Language Processing (NLP) to recognize the technologies mentioned in a patent, thus overcoming this gap in literature. We reach this goal by solving a well known task in NLP called Named Entity Recognition (NER). NER systems are widely used to extract general entities (such as persons' names, cities, dates and times), but there is still a lack of tools able to extract technically relevant information (Fantoni et al., 2013; Chiarello et al., 2018a, 2021, 2020). We measured the performances of three different NER methods in terms of precision, recall and computational time.

The three methods we tested are gazetteer-based, rule-based and distributional NER. Gazetteer-based and rule-based NER exploit pre-defined lists of entities (gazetteer) or rules-driven experts systems. Distributional NER exploits machine learning approach. In the present paper we use the state-of-the-art for entity extraction, the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018).

The tool for identifying technologies from text proposed in our work has two main contributions. It contributes to the field of scientometrics and technology analysis, overcoming the gap in literature related to the analysis of patents for understanding the technological change. It gives a contribution to practitioners, especially policy makers and companies, by supporting them in patent analysis to recognize emerging technologies, map technological convergence or investigate the content of a patent.

The results of this work show how the methods we propose extract $4,731$ technologies in a set of $1,600$ patents with a precision about 40%. We analyse in detail all methods proposed and show that some are more fit to search for small numbers of specific entities (gazettes), whereas others are more effective in gathering large numbers of technologies with less control over which they retrieve (rules, distributional methods).

The most novel result of our work is that it provides a first attempt to recognize technologies from texts, avoiding the collection of technological-related terms that are not technologies, such as function, user, advantage or drawback entities. Moreover, this paper extracts each type of technologies, overcoming the gap existing in literature, that is focused only on the collection of emerging technologies.

The paper is structured as follows, in Section 2 we describe the state-of-the-

art and propose a more precise definition of technology, in order to have a solid framework for the evaluation. In the following section, we present all the data and the methodologies we use. Afterwards, we present the results in Section 4 and finally draw conclusions and describe future steps.

2. State-of-the-art

2.1. Patent Analysis with Natural Language Processing Techniques

Patent analysis has been involved in both academic and industrial disciplines for different purposes. Scholars are generally interested in the analysis of patent data for identifying and studying new radical innovation and paradigm shifts. Whereas, industry actors are focusing also on incremental innovation, namely technical modifications on existing technologies (Sternitzke, 2010). Keeping this in mind, we attempt to provide the readers a comprehensive view of the state-of-the-art in patent analysis for both academic and industrial perspective.

Patent analysis techniques are capable of performing a wide range of tasks, vital from both legal and managerial perspectives (Liu et al., 2011; Yoon and Kim, 2011). Abbas et al. (2014) revise the literature on patent analysis identifying several purposes of the organization in analyzing intellectual property, among which there are novelty identification, technological forecast, technological road map and so on. Prior and current works have relied on two main approaches for the analysis of big amounts of patent data: bibliometric analysis and text mining, as affirmed in several works (Abbas et al., 2014; Li et al., 2019; Small et al., 2014).

The former comprehends a wide range of techniques focused on the analysis of structured data, such as patent assignee, inventor, citation, International Patent Classification (IPC) or Cooperative Patent Classification (CPC). Literature provides a plethora of indicators developed using patent meta-data for a specific purpose. For example, the number of patent applications, backward/forward citations and the number of non patent literature (NPL) citations are used for analysing technological diffusion (Chang and Fan, 2016; Magee et al., 2016). Co-inventors, co-citations, NPL citations and IPC co-classification are such metrics based on meta-data for identifying technological convergence (No and Park, 2010; Karvonen and Kässi, 2013; Cho and Kim, 2014) or emerging technologies (Small et al., 2014; Breitzman and Thomas, 2015; Kyebambe et al., 2017; De Rassenfosse et al., 2013).

Despite the large adoption by academics and practitioners, the bibliometric patent analysis has important limitations. Citation analysis is able to capture prior art but lacks the detail to reflect the technical content of each patent (Kuhn et al., 2020). On the other hand, patent classification is able to reflect the subject matter of the document, but patent classes are too broad to capture the detailed technical content of each document (Righi and Simcoe, 2019). Arts et al. (2021) illustrate and validate the improvement of text mining metrics over traditional measures (based on bibliometric data) for capturing the innovation phenomenon using patents.

Text mining aims to extract the information contained in unstructured data. A branch of text mining is the automatic processing of the human language (natural language in jargon) in written form, for a range of applications, such as machine translation, document classification, question answering, named entity recognition (Pennington et al., 2014). This stream of study is called Natural Language Processing (NLP).

One of the simplest algorithms in NLP is the Bag-of-Words (BoW), that is a representation of a set of documents (namely *corpus*) that describes the occurrence of words within each document. Hofmann et al. (2019) use BoW for measuring the text-based similarity between patents in order to generate technology-related network data that retraces elapsed patterns of technological change. An improvement of BoW model is weighting the word occurrences with the commonly used term frequency inverse document frequency (tf-idf) (Salton et al., 1983). In Niemann et al. (2017), a tf-idf BoW aims to structuring patent text to identify patterns of change over time.

This representation of the textual content of a patent is not useful for capturing the latent semantic of the document. Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) are two of the dimensionality reduction techniques used in the literature for overcoming this issue. In Magerman et al. (2010), the authors involve LSA for grasping similarities between patent and publication text documents. In Park et al. (2015), a patent similarity index based on LSA is developed for exploring potential R&D collaboration partners. SVD is involved by Han and Sohn (2015) to provide a concept of distance between a document and its backward or forward cited patents in terms of claims.

While the aforementioned NLP methods aim to compare the similarity between a set of patents, Latent Dirichlet Allocation (LDA) is employed in literature to extract topics and technological trends. LDA is a popular model used in the field of information retrieval from text corpora, developed in Blei et al. (2003). The primary benefit of LDA is predicting the future technological topics of specific firms (Suominen et al., 2017). In Kim et al. (2015), the authors apply LDA algorithm to identify a vacant technology clusters using patent documents. In Song and Suh (2019), the authors aim at identifying the convergence trajectory for safety technology development via topic modeling techniques.

Recently, the attention of NLP literature has focused on word embedding algorithms, these are able to represent words as dense vectors of real numbers. Unlike LSA and LDA, which also focus on estimating continuous representations of words, word embedding methods use an artificial neural network to represent words. In Mikolov et al. (2013b), the authors show that word embedding outperforms the LSA algorithm for preserving linear regularities among words and have greater computational efficiency than LDA with large text corpora (Mikolov et al., 2013a). Xu et al. (2019) use word embedding to automatically extract technical intelligence from news. Word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) are two recent and popular word embedding methods for analysing patent documents. In Lee et al. (2020a) the authors use word2vec developing a product landscape analysis in order to identify potential technology opportunities across multiple domains. In Trappey et al. (2019), a system based on word2vec is used for retrieving the best related patents of a target document aiming at analysing patent evolution in solar power technology. In Sarica et al. (2020), GloVe is applied to vectorize the patent terms and establish their relationships in a unified vector space for representing the technology semantic network. In Li et al. (2018), the authors propose a machine learning method for patent classification based on word embedding.

The recent advances in NLP research lead to develop a new class of techniques for representing words and phrases, called transformers (or contextual word embedding). In fact, word embeddings have several limitations, among others, these methods generate the same single vector for a given word, beyond the context in which the word appears. Transformers overcome this issue factorizing a word or a phrase based on the textual context. One of the most famous algorithm falling in this class is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which is trained on millions of Wikipedia pages. In Caragea et al. (2020), a BERT-based model is used on a dataset of patent abstracts and patent applications for building a taxonomy of Financial Technology (FinTech). Beltagy et al. (2019) fine-tune a BERT model based on scientific articles, called SciBERT, for processing the text of this huge amount of data makes available from scientific communities. The model is largely used in literature for many purposes, also related to the study of technological phenomena Zhang et al. (2021).

In Lee and Hsiang (2020), the authors fine-tune with about 2 millions of patents a pre-trained BERT model, called PatentBERT by the authors. Specifically,the authors use the model for classifying the patent documents based on the International Patent Classification. Recently, PatentBERT is adopted as a comparison in various patent analysis-based studies. For example, Hain et al. (2022) develop a new method to create vector representations of patent text for measuring the patent-to-patent similarity, and compare their performance with PatentBERT model. Similarly, Choi et al. (2022) propose an automated patent classification model for patent landscaping and compare the classification performance of their model with PatentBERT.

2.2. Technologies Identification with Named Entity Recogntion

In this paper, we apply Natural Language Processing techniques for technology identification and mapping. We use a branch of NLP, namely Named Entity Recogntion (NER). This is an Information Extraction technique that aims at recognizing unstructured text information units (e.g. foods, person names, companies, geographical entities) (Nadeau and Sekine, 2007). The NER methods used in the literature are:

1. *gazetteer-based NER* uses knowledge base resources (or gazette in jargon), consisting of lists of known instances (Pawar et al., 2012), to map mentions of entities within texts to knowledge base resources (e.g. Wikipedia). It is a fast and accurate approach for NER. The gazette-based NER is a semi-supervised method since list creation usually requires manual effort and domain-specific knowledge;
2. *rule-based NER* uses regular expressions and morphosyntactic information to express knowledge based systems able to extract a certain type of entity (Jiang et al., 2011). This system is unsupervised and top-down, and falls into the class of expert systems;
3. *distributional-based NER* uses manually annotated documents (training set) to train machine learning algorithms (Quinlan, 1986). These methods

are the-state-of-the-art in many fields of Natural Language Processing, but they require a large amount of effort to collect and annotate large data-set.

The evaluation of a NER system is one of the most important tasks in the literature for comparing different text mining methods. The key metrics used in the evaluation process are precision, recall and f-score. For further information about these fundamental metrics, we recommend to read the work of Tsai et al. (2006).

In the literature, there are various works that are able to achieve great results for the identification of named entities. In (Lee et al., 2020b), the authors aim to recognize biomedical entities for obtaining a f1 score of 89.61 for diseases, 94.26 for drugs and 78.58 for genes and proteins. In a similar manner, in (Fan et al., 2020), the authors achieve a level of f1 equals to 75.37 for the extraction of events and location names. Both techniques used contextual word embeddings (i.e. BERT).

On the contrary, to the best of our knowledge, the literature on NER systems for technology extraction is poor. In Jang et al. (2021), the authors aim at structuring textual information as a cornerstone of a lexical database for technology-related information, based on the patent data. Similarly, Sarica et al. (2020) develop a technology semantic network (TechNet) using natural language processing techniques to extract terms from massive patent texts. TechNet supports a wide range of applications, e.g., technical text summaries, search query predictions and relational knowledge discovery in the context of engineering and technology. Furthermore, these two articles construct a lexical database of technology-related words but they do not recognize a given technology in the patent text.

The most similar work is a pre-printed article (Hossari et al., 2019) focused on the extraction of Artificial Intelligence technologies. However, the authors give only qualitative criteria for evaluating their results. The state-of-the-art is mainly focused on the identification of emerging technologies and not established ones, exploiting patent classification (i.e. IPC code classes) (Kay et al., 2014; Gustafsson et al., 2015; Song et al., 2017) or using textual data with greedy methods, extracting generic terms and not only technologies.

For example, Ranaei et al. (2020) analyse the emergence of technologies using three text-based approaches: tf-idf metrics for capturing technological changes, LDA for evaluating the emerging topics, text-based score developed by Porter et al. (2019) and Carley et al. (2018). The authors found that the three different methods provide somewhat distinct perspectives improving the understanding of technological change. Despite the remarkable results of their work, it includes not only technologies but also technology-related terms, such as *defective*, *immune*, *delay fluorescence*.

While a training set for entities like chemical molecules and biomedical terms has been built to train corpus-based classifier, this is not true for technologies. This forces to manually evaluate the precision of the NER system, reviewing the list of extractions to filter out non-technologies. For what concerns recall, in absence of a training corpus, the grain of the extracted entities is the key element

to consider for evaluating it. Linguistically, we can refer to technologies using an abstract wording (e.g. *device* or *system*) or specific wording (e.g. *water proof smart watch* or *visual software modeling editor*). For a balanced identification of technological content in a patent corpus, it is important to avoid both the extraction of too generic and too specific technologies. Taking into account too generic technologies causes the risk that no technological change and evolution pattern would be identified by the patent analysis. For what concerns the extraction of too fine grained technologies, it is important to avoid collecting sub-systems instead of technologies.

In a recent paper, Giordano et al. (2021) involve gazetteer-based and rule-based system for identifying technologies from a set of 300,000 patents related to C4ISTAR, a defence related domain, for analysing the convergence phenom. They reached a precision of 35.39%, collecting about $1,000$ different technologies. However, the purpose of the authors is different from that of our article for three main reasons. First, the authors focused on the study of technological convergence phenomena and not on developing a NER methods able to extract technologies from the text. This different purpose of Giordano et al. (2021) leads to a lack of measurement testing their NER method. In fact, the authors only evaluated the precision of their method without calculating the recall or providing information about the time they spent to recognize technologies from the text. Second, Giordano et al. (2021) avoid to use state-of-the-art of NER system, namely distributional method, based on contextual word embeddings, but they only applied gazetteer-based and rule-based NER. As discussed before, these methods suffers some limitation, related to the recall measurement.

2.3. The Concept of Technology

As suggested by Nadeau and Sekine (2007), the use of NLP techniques for technology identification and mapping requires a formal definition of technology, precise enough to allow discerning these entities and comprehensive enough to include all of them. Therefore, we review several definitions from dictionaries and literature, reported in Table 1, to highlight the main characteristics of this entity and to identify what are the elements for which an entity can be classified as a technology.

The reported definitions in Table 1 convey different aspects, each relevant to this concept, the most insightful elements are:

(A) the technical means or in general the technical systems (Wikipedia, Waight 2014; Volti 2005; Ramanathan 1994);

(B) the anthropocentric view, related to the creation and the use of technology by human-kind (Wikipedia, Carroll 2017; Waight 2014; Volti 2005; Ramanathan 1994);

(C) the application of knowledge and science thanks to the skills of many individuals, who cooperate themselves to develop a technical system (Cambridge dictionary, Carroll 2017; Volti 2005; Ramanathan 1994);

(D) the purpose to solve practical problems and to perform a function (Carroll 2017; Waight 2014; Volti 2005).

We had the need for an operative description enabling us to define our scope and perform an efficient and strict selection of entities identifying a technology. Therefore, we attempt to distill all the multiple perspectives aforementioned into a comprehensive summary, as follows:

Definition 1. *A* technology *is a technical mean or in general a technical system (A) created by human-kind (B) through the application of knowledge and science (C) in order to solve a practical problem or perform a function (D).*

With this definition we do not have the ambition to define in a complete and exhaustive way such a broad and complex concept. Our aim is indeed to clarify the main characteristics of a technology in order to be able to distinguish the relevant information among the amount of data extracted with NLP techniques from patents.

Source	Definition
Cambridge dictionary	(The study and knowledge of) the practical, especially industrial, use of scientific discoveries.
Wikipedia	Technology ("science of craft", from Greek $\tau\epsilon\chi\nu\eta$, techne, "art, skill, cunning of hand"; and $-\lambda o\gamma\iota\alpha$, -logia)[1] is the sum of techniques, skills, methods, and processes used in the production of goods or services or in the accomplishment of objectives. [...] It can be the knowledge of techniques and processes, or it can be embedded in machines to allow for operation.
(Carroll, 2017)	Technology is "something that is always inherently intelligent enough either to function and to be used to function; anything devised, designed, or discovered that serves a particular purpose; [and] the knowledge that is used for a purpose, without itself necessarily being translated into something physical or material that does (e.g., instructional methodologies in education, processes, ideas)".
(Waight, 2014)	Technology is delineated as something that "improves and makes life easier, the artifacts which function to accomplish tasks, and the representations of advances in civilization".
(Volti, 2005)	Technology is "a system created by humans that uses knowledge and organization to produce object and techniques for attainment of specific goals".
(Ramanathan, 1994)	Technology is the manifestation of four elementary and interacting components: technoware, related to the tangible and palpable parts (i.e. tools and systems);humanware, related to the human resources who, with their knowledge and skills, produce, use and express the technoware; orgaware, referred to effective organizational practices, linkages, and related arrangements needed to make the best use of technoware and humanware; and inforware, that represents the accumulation of knowledge by human beings related to the other 3 components.

Table 1: Definition of technology

3. Methodology

We employed NLP systems to identify technologies in the title, abstract, claim and state-of-the-art description of the patents. We developed three different approaches based on gazetteer, rules and distributional methods. In this section, we describe each of them and outline their advantages and drawbacks. Throughout the rest of the paper, in order to ease the reading, we will occasionally refer to the items (technologies in our case) we extract as *entities*.

3.1. Data Collection

We set the experiment employing four case studies to measure the performance of the different methods proposed in this paper. The database chosen for retrieving the patent documents is the Erre Quadro S.r.l.[2] one. The database contains over 90 million patents and includes high quality bibliographical and legal status patent information from leading industrialized and developing countries. The Erre Quadro database was chosen due the fact that it indexes patents from different European and International databases, e.g., European Patent Office (EPO) and World Intellectual Property Organization (WIPO). The method we proposed may also be replicate using other patent database, such as United States Patent and Trademark Office (USPTO), EPO or WIPO.

In literature, the common strategies for retrieving patents are based on keywords (Maghrebi et al., 2011), patent classification systems (e.g. International Patent Classification system, IPC) (Ozcan and Islam, 2017) or their combination (Park et al., 2013). We defined three retrieving principles in order to simplify the measurement task:

- **Perform a IPC-based searching strategy**: The IPC is a hierarchical classification of patents based on the technological area. Each classification code is composed of 7 characters. The first character is a letter, it is called *section*, and indicates the technical field. Then, there are two numbers, called *class*, that indicate the subject matter of the patent. The next letter constitutes the *subclass*, i.e. the units in which the *classes* are divided. Afterward, there is the File Index classification (FI), a three-digit number (*IPC subdivision symbol*) and/or one letter (*file discrimination symbol*), that represents a *main group* or *subgroup* and indicates the theme of the patent (OECD, 2009). We decided to perform an IPC search because this classification well determines the boundary of each class, as established in the Strasbourg Agreement of 1971. So, since we aim to develop a system for automatically recognizing technologies in a given technological field, having well-defined technological domains eases measuring how much our extraction system is precise (precision) and complete (recall).

[2]Available: https://www.errequadrosrl.com/, Accessed: July 28, 2021.

- **Select the first 5 digits for each IPC**: The number of digits selected for each IPC class is 5 for a balanced identification of technological content, avoiding to consider too generic or too fine grained technologies as explained in Section 2;
- **Collect different IPC section**: We choose to consider different IPC sections (namely first letter of the IPC code) for performing the most extensive and exhaustive comparisons among the various methods. In fact, the section is the broadest patents classification level.

We selected the following IPC codes:

- **A63F 3/00** Board games, Raffle games;
- **C23C 2/00** Hot-dipping or immersion processes for applying the coating material in the molten state without affecting the shape;
- **E04B 5/00** Floors; Floor construction with regard to insulation;
- **H04J 1/00** Frequency-division multiplex systems.

Whereas the chosen *subclass* codes are the following:

- **A63F** CARD, BOARD OR ROULETTE GAMES;
- **C23C** COATING METALLIC MATERIAL;
- **E04B** GENERAL BUILDING CONSTRUCTIONS;
- **H04J** MULTIPLEX COMMUNICATION.

We searched for the higher diversity in terms of languages and domains. We selected cases where the inventions is user-centered (A63F, E04B) or not (C23C, H04J); where technical subject matter may represent a product (A63F), an apparatus (H04J) or a process (C23C, E04B); where technical subject matter may be a physical device (A63F, C23C, E04B) or an electronic/software system (H04J). The selection process is fundamental for identifying difference among the methods developed in this study and understanding which biases are embedded in each one. For each selected IPC code, we retrieved also the patents set of the relative *subclass* (e.g. the *subclass* of A63F 3/00 class is A63F). These data will be fundamental for training the distributional NER system as we explain later in this Section. The *subclass* is more general than the complete IPC code and includes diversified patents, so it can help in the identification of different technologies. This leads to improve the models with reference to the recall, but it may produce minor losses regarding the precision. We will show this is not a dramatic drawback, indeed for one of the methods it leads to a small improvement for both measures.

In the rest of the paper, we refer to the complete IPC code as *IPC group* and the related IPC *subclass* as *IPC category* to avoid misinterpretations.

For each IPC category and IPC group we selected 400 patents. Regarding the IPC category, we chose this number to have a data-set that is a good fit for the training of the models. On average we found patents to contain around 300 sentences. A classical benchmark data-set for NER (Tjong Kim Sang and De Meulder, 2003) contains around 14,000 sentences for training, therefore we selected this amount of patents in order to reach a comparable amount. We gather a larger number of sentences closer to 100,000 from our patent set. Indeed, we need to compensate the lower precision of rules as opposed to human tagging: a larger amount of sentences helps us average out part of the imperfections in the data.

We chose 400 patents for the IPC group as well for the sake of comparison. Although we analyse as well the number of entities extracted, our focus is on a fair comparison among the methods. We need to make sure that we do not give an advantage to the trained methods by providing them broader sets of technologies, thus we need IPC categories and IPC groups to be equal in number. Moreover, this does not pose any limitation on the patent set validity.

3.2. Extraction methodologies

3.2.1. Gazetteer

The gazetteer based method is the simplest available, and it consists in using gazettes of technologies. For the selection of the sources we relied on the work by Giordano et al. (2021) that uses freely available online sources:

- ***Wikipedia***: the Wikipedia contributors list a set of emerging and potentially disruptive technologies[3]. It contains 397 different technologies. In Bonaccorsi et al. (2020), it is demonstrated that Wikipedia's list of emerging technologies has a degree of coverage in the range 90%–95% with respect to academic journals, consultancy reports and specialized blogs;

- ***O*NET***[4]: it is an occupational framework developed by the U.S. Department of Labor which is made of 974 occupations classified on the basis of the Standard Occupational Classification (SOC) system and their corresponding skills, knowledge, abilities and technologies. The framework includes 30,173 different technologies. O*NET divides those technologies in (i) 21,267 machines and equipment (70.48 %) and (ii) 8,906 information, communication and software technologies (29.52 %).

The use of gazettes allows us to work in a controlled environment, where one can know in advance the detectable type of entities. Therefore, given the gazette source and specifications, we focus only on a particular kinds of technologies.

This solution is extremely fast to deploy and can be used to inspect several thousands of documents in very short time. Looking for the flaws of this

[3]Available: https://en.wikipedia.org/wiki/List_of_emerging_technologies, Accessed: July 28, 2021.

[4]Available: https://www.onetonline.org/, Accessed: July 28, 2021.

approach, the biggest one is the difficulty in gathering high quality gazetteer containing large numbers of technologies. The cleansing of a gazetteer is performed by highly trained people, so such a process is hard to scale to several thousands of entities.

Furthermore, the gazettes used in our paper aim to identify only a part of technologies mentioned in patents. The Wikipedia list contains only emerging technologies. These last are well defined in literature and have five main characteristics: (i) radical novelty, (ii) relatively fast growth, (iii) coherence, (iv) prominent impact, and (v) uncertainty and ambiguity (Rotolo et al., 2015). The O*NET gazette derives from an occupational framework. So, it includes the technologies related to job occupations with a coarse-grained level of detail, mentioning only the commercially available technologies. Moreover, there is an unbalanced focus on information and software technologies (more than 20 % of the total), reflecting the current situation of job landscape, where the digital transformation had large effects (Frey and Osborne, 2017).

Beyond this, gazetteer are hard to adapt to different contexts, as demonstrated also in biomedical domain by Odat et al. (2015). Patents are an extremely broad data-set with all possible kinds of technologies and a gazette able to cover this variegated type of knowledge would need to contain a vast amount of entities, hard to develop and to maintain as well. Furthermore, the gazetteer NER suffers from ambiguity (Carlson et al., 2009), where words like *python* can have different meanings (e.g. an animal or a programming language) depending on the context.

Given the above considerations, pre-built gazettes aims to minimize (i) the human effort in listing all possibly technologies, and (ii) the machine time to recognize technologies in text. However, we believe this methodology will extract a smaller number of entities with some blurred technologies.

3.2.2. Rules

The second family of NER systems we employed is based on rules. In a general sense, this means we define morphosyntactic patterns that identify the presence of a technology in a patent text. This type of extraction can be made in several ways and there is literature about it (Hearst, 1992; Roller et al., 2018; Chiarello et al., 2018b). This methodology, if properly adapted to each case, turns out to be quite effective especially in terms of coverage.

Let us first make a short premise to introduce these methodologies. We will use the hypernym/hyponym relation indicating more and less abstract concept. An example of such relation is the one between phone and smartphone. Indeed whenever something is identified as a smartphone it can be identified also as a phone and at the same time there are phones that are not smartphones.

We use two different rule-based methods:

- ***Extractor 4.0***: regular expressions for extracting technologies 4.0 developed in Chiarello et al. (2018b);
- ***Hearst*** regular expressions for searching hypernym/hyponym relation (Roller et al., 2018).

The first rule-based approach consists of a list of automatically generated regular expressions aimed at extracting Industry 4.0 terms. Regular expressions are pattern matching queries that can be built to match virtually any pattern. More complex patterns require for more complex queries and this poses a limit to the generality they can achieve. Nevertheless, the list built in (Chiarello et al., 2018b) contains about 1,600 expressions and covers several different cases, making it a valuable extraction tool.

We expect an high application field dependence due to the fact this approch was originally developed for Industry 4.0. On the other hand, the flexibility of regular expressions and the data driven way used to develop them give this approach good generality, therefore we expect it to show higher recall than gazetteer.

The rules based on hypernym/hyponym relation instead are derived from Hearst (1992), it builds upon the methodology that assigns to each word its role in the sentence (Petrov et al., 2012), e.g. nouns, indicated as *NOUN* or adjectives as *ADJ*.

We search for *NOUN* tokens and then we use fixed patterns as described in Hearst (1992) to select the surrounding specifiers. These patterns are developed to match hypernym/hyponym pairs. However, they do not automatically identify technologies. we add a step to find the technologies, indeed, we keep only pairs including one of the following words as their hypernym: *technology, machine, device, apparatus, mechanism, sensor, network, system, unit.*

The words we select identify general terms that are hypernyms to large families of technologies. To choose them, we used the work from Jang et al. (2021) where the authors define words such as *system*, *unit*, *device*, *apparatus* and so on as high level concepts in terms of automotive technology.

The choice of words introduces a bias into this methodology making it lean towards more machinery and tools based fields. This will affect its ability to perform in certain IPC groups.

We expect a decrease in precision in the use of rules compared to the gazetteer, since we use not too strict patterns to extract a relevant amount of entities. On the other hand, the recall of our method, namely the fraction of technologies extracted, is supposed to increase.

3.2.3. Distributional methods

We train several distributional NER models, using BERT (Devlin et al., 2018) to obtain contextual word embeddings, a 2 layer bidirectional Long Short Term Memory (LSTM) on top of the embeddings and conditional random fields.

This architecture is similar to the one from Peters et al. (2018) and it leads to near state-of-the-art performance on NER benchmarking datasets such as CoNLL-03 data-set (Tjong Kim Sang and De Meulder, 2003). Moreover, these models need to be trained on tagged data. This is done using the set of patents extracted from the categories for each IPC.

For each IPC group, namely A63F 3/00, C23C 2/00, E04B 5/00 and H04J 1/00, and for both rules extractor, namely *Extractor 4.0* and *Hearst*, we build a different extractor based on distributional methods. Each of these methods

will be called ***Dist Hearst*** or ***Dist 4.0*** respectively. Notice how we omit the IPC group name to ease the notation and in the rest of the paper we will always explicitly state which IPC group is under study.

Let us describe the pipeline used to build each extractor. We proceed as follows:

1. We select a IPC group and a NER method;
2. We take the associated IPC category, for example A63F for A63F 3/00, as described at the beginning of this Section;
3. We annotate the patents in this IPC category using the method we selected;
4. We save the output which is now a set of sentences with tagged technologies;
5. We train the distributional model on this set of sentences;
6. Finally, we use this trained model to extract from the IPC group we selected in the first step.

Beyond this approach, for both *Hearst* and *Extractor 4.0*, we train one model following the same steps outlined above with the difference that each of the category associated with an IPC group is substituted with a patent set built randomly selecting 100 patents from each of the IPC category. The models obtained will be called ***Dist Hearst All*** and ***Dist 4.0 All***. Summing up:

- ***Dist 4.0*** is the distributional method using the *Extractor 4.0* NER trained on IPC cateogry;
- ***Dist Hearst*** is the distributional method using the *Hearst* NER trained on IPC cateogry;
- ***Dist 4.0 All*** is the distributional method using the *Extractor 4.0* NER trained on a set of 100 randomly selected patents from each of the IPC category;
- ***Dist Hearst All*** is the distributional method using the *Hearst* NER trained on a set of 100 randomly selected patents from each of the IPC category.

Our hypothesis is that the wider range of technologies available in the higher level IPC, together with the adaptability of rule based methodologies, will lead to a high quality training set. In turn the generalization skills of contextual word embedding should be able to make use of a data-set encompassing such knowledge in order to learn from it in a proficient way. We believe that, once trained, these models can be used in the IPC groups to extract an higher number of technologies compared to rules.

We expect an improvement in recall, whereas it is hard to make predictions about precision, since the more general IPC group could be both an improvement or a drawback for this point of view.

We decided to not manually check the training set for constructing the distributional methods for two reasons. First of all, the manual check is time consuming and expensive: it requires a domain expert to read each technology extracted by the rule-based NER carefully. Second, we aim at observing the achievable levels of precision and recall of distributional methods in condition of low effort in terms of time and cost.

In parallel, we did not perform any information based cleaning (e.g. tf-idf filtering) for two reasons. First, custom pre-processing would influence the results. Second, this process would require an assessment on how it is performed (per IPC, per corpora, per method, etc.). In addition, just like manual filtering, it constitutes a large effort on its own. However, we believe that this might be useful for improving a model performance in future works.

The implementation of the distributional methods was done using the Python package *Flair*, developed in (Akbik et al., 2018). The training is performed on a Quadro RTX 6000 with 24 gigabytes of memory. We kept the same parameters over all experiments, the batch size is fixed at 64, for the recurrent neural network we use a hidden size of 256 which we found to be good for all cases, as embeddings we use BERT-base and a learning rate of $5 * 10^{-3}$ for the training that lasted 20 epochs[5]. The high computational cost of the experiments limited our possibility to make an exhaustive hyper-parameters search. We tried some preliminary experiments changing learning rate that didn't lead to significant variations in performance and we thus chose a single set of parameters for all models. In the rest of this work we identify which approach among the several tested is best and in future works, focusing on a single model we will optimize it further through a large set of possible parameters.

3.3. Performance Evaluation

The list of technologies recognized by our NER systems was manually checked to compute precision and recall. In literature, the results of several NLP tasks have traditionally been evaluated using human subjects (Belz and Reiter, 2006) or using previously annotated data (Lee et al., 2020b). As discussed in Section 2, in the technological domain there is a lack of annotated documents. For these reasons, we rely on human evaluation in a manner similar to Giordano et al. (2021). The validation was performed in a straightforward way by two PhD students in Smart Industry at the University of Pisa. We provided both students with a table containing the list of extracted entities for each IPC group (the same for both students). The assignment was "*Read each extracted entity and decide whether the entity is a technology or not in the context where it appears, comparing each entity with the Definition 1 of technology, provided in Section 2.3.*". The output of the extractions with manual check will be made available upon publication.

[5]While we can not make the code available we are very open to discuss the implementation upon contact.

We remark how this task is itself complex, since all the entities, even those that are not technologies in a strict sense, belong to close semantic fields. Several ambiguous cases are found, for example *high frequency signal*, which to a non expert might appear as a technology, is physical behaviour. According to the Definition 1, one can see that this example is too generic. On the contrary the similar entity *high frequency signal interface* is a example of technology and the noun *"interface"* helps us to disambiguate it with respect to the physical behaviour.

This case demonstrates how the human validation is itself very challenging. Indeed, from an algorithmic perspective, the two aforementioned entities are extremely similar, since they differ in only one word, while being mostly composed of rare words, making it hard to infer that the fourth one is what makes the difference (Bernier-Colborne and Langlais, 2020).

For understanding the degree of agreement among the independent observers who assess the completed list of the technologies we calculated inter-rate agreement using the Fleiss' Kappa (Fleiss, 1971). We selected a sample of 300 technologies and provided the 5 authors of this study the same instructions described above for evaluating whether the entity extracted by our algorithm is a technology or not. We stress the fact that the 5 authors have a different background (1 Mechanical Engineer, 2 Management Engineers, 1 Mathematician and 1 Aerospace Engineer) and experience (1 Professor, 1 Researcher, 3 PhD students), adding value to the evaluation of the technologies because it is carried out by experts from different perspectives. The inter-rate agreement is 0.516 and the authors give the same rating for the 50.30% of the technologies in the sample. Moreover, we found that 4 authors agree on an additional 29.00% of the samples and the remaining 20.70% meet the agreement of only 3 out of 5 authors, highlighting how complex is evaluating what could be defined as technology.

Several studies in literature offer rules of thumb for interpreting the inter-rate agreement. Many authors agree that an inter-rate agreement value greater than 0.50 is the target condition to be confident on the results (Landis and Koch, 1977; Fleiss et al., 1981; Regier et al., 2013). It is possible to reach higher values with more complex evaluations. For example, Zhang et al. (2020) attempt to build a fine-grained entity annotation corpus of clinical items, manually annotating more than $10,000$ clinical records in five rounds. In each round, 100 records are randomly selected, and the inter-rate agreement is calculated: the initial value of 0.40 increased step by step reaching 0.94 at the fifth round (Zhang et al., 2020). In our case, we perform one review round, reaching 0.516 of inter-rate agreement. This result, higher than the initial value presented in Zhang et al. (2020), shows a moderate but confident agreement.

We are aware an iteratively process of annotation can help in improving the inter-rate agreement, especially for engineering a system able to recognize technologies in textual data. However, our goal is first to demonstrate that is possible to use NER methods for extracting technologies from textual data and to identify which NER system perform better than others.

4. Results

In Sections 4.1 to 4.3, we provide insights about the performances of the different approaches used to extract technologies from patents, we analyse qualities and flaws of each method and we study how they work together. In Section 4.4, using the four IPC groups outlined in Section 3.3, we describe how our methods allow researchers and practitioners to map a given knowledge domain. The Table 2 summarise the different methodologies, to which we refer from now on as reported in the column *Method*.

Type	Method	Explanation
Gazetteer	Wikipedia	List of emerging technologies from Wikipedia.
	O*NET	List of technologies from the occupational framework O*NET.
Rules	Extractor 4.0	Regular expressions for extracting technologies 4.0.
	Hearst	Regular expressions for searching hypernym/hyponym relation.
Distributional	Dist 4.0	Distributional method using the Extractor 4.0 NER trained on each IPC cateogry using BERT.
	Dist Hearst	Distributional method using the Hearst NER trained on each IPC cateogry using BERT.
	Dist 4.0 All	Distributional method using the Extractor 4.0 NER trained on all the IPC cateogries using BERT.
	Dist Hearst All	Distributional method using the Hearst NER trained on each IPC cateogries using BERT.

Table 2: Summary of methods per types with explaination

4.1. Performance of Technological Named Entity Recognition

The total number of unique entities extracted by our NER systems is $12,572$ in a set of $1,600$ patents. After the human validation process described in Section 3.3, the portion of entities that are considered technologies by the revision is 38 % ($4,731$ different technologies).

This score is in line with the similar work of Giordano et al. (2021), which attempts to identify technologies from 300,000 patents on defence sector, reaching an overall precision of 35.39 % and collecting $1,090$ different technologies. However, the score is low if compared with other NER systems; but, as mentioned in Section 2.2, the absence in literature of a training set for technologies negatively affects the performance of the extraction systems.

Let us report insights about the number of technologies per patent. The median of extractions per patent is 81 technologies (counted with repetition) and 10 distinct ones. So, on average per sentence there are 0.35 technologies and 0.04 distinct ones. This means that, one needs to read about 3 sentences to encounter a technology and about 25 sentences for a new unseen one in a patent.

These numbers, although only averages, highlight that patents are generally densely populated with technologies while only containing few unique ones. The difficulty of technologies extraction emerges: this tasks requires to cherry-pick the few interesting ones within each patent.

Based on the work of Chiarello et al. (2018a), we use the following metrics for the evaluation of the whole technologies extraction process:

- *Training time*: time needed to create the statistical model using the training set in the distributional-based NER;

- *Extraction time*: time needed to extract the entities on the patent set;
- *Precision*: number of technologies that a system correctly detected in textual data divided by the total number of entities identified by the system;
- *Relative recall*: proportion that the NER system retrieves of the total technologies retrieved by all systems together for a given IPC group.

In particular, we measure the precision and recall based on the human validation. As described in 3.3, evaluating the recall is a challenging task for a domain where an annotated corpus in not available, as in our case. For this reason, we relied on relative recall, instead of traditional recall. Sampson et al. (2006) define relative recall as the proportion that any NER system retrieves of the total technologies retrieved by all systems considered to be working as a composite.

For the scope of this paper, these metrics allow us to consider both the effectiveness (Precision and Relative recall) and efficiency (Training time and Extraction time). The latters are evaluated for a good reproducibility of our work. The training time is suitable for researchers and practitioners who want to develop their own distributional NER methods for recognizing technologies in textual data. The extraction time mainly helps practitioners that want to apply our NER methods for their own purposes and textual information. In general, the measures of training and extraction time allow to better estimate the duration of the analysis and plan the related tasks. Big data technologies, such as Hadoop and Spark, can be easily integrated with the proposed methodology, resulting in a time reduction in the training of distributional methods or in the extraction of technologies from the text. These frameworks are widely adopted in current academic and business landscape, however they requires expertise and knowledge to set up the architectures to process the data.

In Table 3 we compare all the metrics across all methodologies and IPC groups. We report also the number of unique entities automatically extracted in a patent set (Found) and the number of distinct entities considered as technologies after the manual review (Correct). These values were used in the computation of Precision and Relative Recall. In the next paragraph, we describe and discuss each class of methods presented in section 3.

4.1.1. Performance of Gazetteer-based NER

The gazettes achieve the highest precision over all tasks, as shown in Table 3. They are hand crafted and thus this result is not unexpected. However, they provide a smaller number of entities when compared to the other methodologies, as the recall points out.

The *O*Net* gazette is the most consistent method in terms of found technologies (Correct) across different IPC groups. As pointed out in Section 3, this result is due to its structure, where almost all descriptions (70.48%) contain names of machines and equipment. It behaves more coherently than rules-based approaches or distributional methods, though at the cost of a lower amount of retrieved technologies (Correct).

IPC	Training Time	Extraction Time	Method	Found	Correct	Precision	Relative Recall	F1
A63F 3/00 Board Games	-	3s	O*Net	286	148	0.517	0.152	0.235
		2s	Wikipedia	27	13	0.481	0.013	0.024
		732s	Extractor 4.0	547	284	0.519	0.292	0.373
		348s	Hearst	398	216	**0.543**	0.222	0.315
	900m	1,720s	Dist 4.0	771	378	0.490	0.388	0.433
		1,714s	Dist Hearst	2,362	743	0.315	**0.764**	0.446
	500m	1,729s	Dist 4.0 All	732	375	0.512	0.385	0.439
		1,721s	Dist Hearst All	1,063	414	0.389	0.425	0.406
C23C 2/00 Coating solutions	-	3s	O*Net	302	178	**0.589**	0.318	0.413
		2s	Wikipedia	31	10	0.323	0.018	0.034
		745s	Extractor 4.0	865	123	0.142	0.220	0.172
		353s	Hearst	274	96	0.350	0.171	0.229
	600m	1,702s	Dist 4.0	301	118	0.392	0.211	0.274
		1,685s	Dist Hearst	791	183	0.231	**0.327**	0.270
	500m	1,706s	Dist 4.0 All	305	127	0.416	0.227	0.294
		1,702s	Dist Hearst All	605	171	0.283	0.305	0.293
E04B 5/00 Floors construction	-	3s	O*Net	283	184	**0.650**	**0.399**	0.494
		2s	Wikipedia	17	5	0.294	0.015	0.028
		771s	Extractor 4.0	494	71	0.144	0.211	0.171
		376s	Hearst	313	62	0.198	0.185	0.191
	500m	1,858s	Dist 4.0	267	76	0.285	0.226	0.252
		1,837s	Dist Hearst	548	86	0.157	0.256	0.194
	500m	1,811s	Dist 4.0 All	254	81	0.319	0.241	0.275
		1,849s	Dist Hearst All	442	79	0.179	0.235	0.235
H04J 1/00 Multiplex systems	-	5s	O*Net	248	174	**0.702**	0.071	0.128
		3s	Wikipedia	22	13	0.591	0.005	0.010
		1,010s	Extractor 4.0	1,332	528	0.396	0.215	0.278
		527s	Hearst	1,019	640	0.628	0.261	0.368
	800m	2,674s	Dist 4.0	1,634	930	0.569	0.379	0.455
		2,671s	Dist Hearst	3,132	1,651	0.527	**0.673**	0.591
	500m	2,583s	Dist 4.0 All	1,597	906	0.567	0.369	0.447
		2,563s	Dist Hearst All	1,373	680	0.495	0.277	0.355

Table 3: Performance of proposed NER systems

The *Wikipedia* gazette has lower performances in terms of Relative Recall than the other methods. The amount of found technologies (Correct) is lower than 15 per IPC group. However, we may consider that *Wikipedia* gazette includes only 397 different technologies, while *O*Net* has more than 30,000 technologies. Although the sizes of two gazettes differ by more than 2 orders of magnitude, the number of Found and Correct technologies differs only by an order of magnitude. This may be due to the fact that the technologies in *O*NET* are indicated with commercial names, rare to find in patents.

4.1.2. Performance of Rule-based NER

The effectiveness of rule-based NER is domain dependent. Both *Hearst* and *Extractor 4.0* suffer a performance loss on the IPC groups C23C 2/00 and E04B 5/00 in terms of both precision and recall. In general, all rule-based methods perform better on H04J 1/00 than on any other IPC group. These results may be explained by the design of those methods:

(i) *Extractor 4.0* developed by Chiarello et al. (2018b) aims at identifying Industry 4.0 technologies, by definition referred to "cyber-physical-system"

(Lu, 2017), that are more related with electronic and software technologies, as in the case of H04J 1/00 (Frequency-division multiplex systems) and also for A63F 3/00 (Board games, Raffle games);

(ii) *Hearst* method derived from Hearst (1992) was designed to recognize modular technologies, namely technologies structured in parts, sub-parts, systems, devices and so on.

Therefore, IPC groups like C23C 2/00 and E04B 5/00, rooted in processes and constructions, are farther apart from the focus of these tools. Whereas H04J 1/00 and A63F 3/00 better fit the features of *Hearst* and *Extractor 4.0*, as confirmed by the performance over these IPC groups. These results may reflect the biases embedded in our methods that uphold the choice to select different case studies (IPCs), as explained in Section 3.

4.1.3. Performance of Distributional NER

Different behaviours demonstrated by the distributional methods root on different reasons we explain hereinafter.

The performance of the rules-based methods reflects on the performance of the distributional models trained on them. That is, if the rules perform worse on a certain IPC group the distributional methods will behave the same way. However, we notice an improvement of the precision in IPC group C23C 2/00. This is likely due to higher precision and recall of the extractions in the IPC category. Indeed, we are not able to understand which patterns lead to a decision about a technology selection since distributional methods are inherently hard to explain (Pedreschi et al., 2019).

Let us now underline the most unexpected behaviours in Table 3. The high number of technologies found by *Dist Hearst* in H04J 1/00 is noteworthy. In general this group has higher number than all others in terms of retrieved technologies (Correct), the highest number of extracted entities (Found) among IPC groups, and the highest precision (Precision) among the others as well. One more characteristic of this IPC is that, even though *Extractor 4.0* and *Dist 4.0* detect a comparable amount of entities ($1,332$ and $1,634$, respectively), *Extractor 4.0* achieves this with a lower precision. Such a behaviour is not easy to interpret given the black box nature of distributional models (Pedreschi et al., 2019).

One more general trend is that rules, particularly *Hearst*, are only partially outperformed by distributional methods. The recall is higher in some IPC groups (H04J 1/00 and A63F 3/00) than in the others. However, the time consumption is about three times longer than Hearst (without considering the training time).

The recall of distributional methods is what makes it worth training them. Despite they do not provide improvements in terms of precision, they increase the recall in all IPC groups. This means that they are able to recover a larger amount of technologies in the same amount of text. This difference depends on the strictness of rules, which only identify specific patterns, opposed to the

flexibility of distributional methods, that, though highly data dependent, can learn to identify any pattern.

One more information conveyed by Table 3 is that the methodology developed by training on a merged set of patents, *Dist Hearst All* and *Dist 4.0 All*, are also a valid approach. Indeed, they achieve similar precision and recall, they are often within 0.05 variation in precision and 0.1 variation in recall over all IPC groups (except for H04J 1/00) compared to the respective *Dist* method despite being trained only once.

Regarding the time costs, *Dist Hearst All* and *Dist 4.0 All* provide a faster method since they need to be trained only once with a training time equal to 500 minutes, instead of one time in each IPC groups, where the overall time of the training is 2,800 minutes (900 for A63F 3/00, 600 for C23C 2/00, 500 for E04B 5/00 and 800 for H04J 1/00). Furthermore, the performance of these methods is good on all IPC groups, though partially worse than those trained in the respective categories, *Dist Hearst* and *Dist 4.0*.

4.1.4. Performance of Global vs Local Training for the Distributional Methods

Looking at Table 3 as a whole we see that the differences in precision and recall of distributional methods among IPC groups support the choice to select a diversified data-set, as we will further investigate in this section.

One limitation of our work is that we trained distributional models on a tagged data-set created using rules. For example, *Dist Hearst* was trained on a data-set tagged by *Hearst*, thus one can expect a large overlap between the technologies found by *Hearst* on the IPC category A63F and by *Dist Hearst* on the IPC group A63F 3/00.

IPC	Dist Hearst	Category Hearst (%)
A63F 3/00	743	307 (41)
H04J 1/00	**1,651**	330 (20)
C23C 2/00	183	48 (26)
E04B 5/00	86	26 (30)
Total	*2,264*	*589 (26)*
	Dist 4.0	Category Extractor 4.0 (%)
A63F 3/00	378	177 (47)
H04J 1/00	**930**	233 (25)
C23C 2/00	118	43 (36)
E04B 5/00	76	23 (30)
Total	*1,198*	*298 (25)*

Table 4: Performance of Global vs Local Training

We provide a measure of the fact that though this dependence exists, there is a limited overlap between the technologies we found in the categories and those we found in the IPC groups. In Table 4, we show for both *Dist Hearst*

and *Dist 4.0* how many (in absolute and percentage values) of the technologies found in the groups by distributional methods (Dist Hearst, Dist 4.0) are also found by the rules in the IPC categories (Category Hearst, Category Extractor 4.0).

As we can see for all IPCs and for all methods, a large portion of the extractions in the IPC group was not previously found in the IPC category, indeed, in all cases less than 50% and in several cases below 30% of the entities were found by the rules. This supports the validity of our methodology, in particular, it makes the training effort worthy. In this table we can find as well better performances in IPC group H04J 1/00, the reasons are the same as outlined above and they depend on biases of the algorithms related to the patent IPC groups.

4.2. Quantitative Comparison of the proposed NER

Table 5 shows the number of shared technologies per pair of NER systems, the diagonal of the table reports the number of technologies found by each extractor.

We expect that the maximum number of technologies contained in the patents, though this amount is unknown, influences the level of intersection among different methodologies. Therefore, as a general remark, Table 5 can indicate the complementarity among the methodologies.

The distributional methodologies show a good level of intersection among them because of the high number of extracted entities and the training processes executed on similar data. Each of them has a stronger intersection with the rules used to build their respective training data. For example, the number of shared entities is 377 for the pair *Dist Hearst* and *Hearst* and 203 for *Dist Hearst* and *Extractor 4.0*. Therefore, the rules used to tag the training data of the distributional approaches leads to higher intersection between the respective couple of methods. In addition, the distributional methods appear consistent because they are trained on the same data-set, albeit tagged differently. Therefore, they register high intersection values.

O*Net	454							
Wikipedia	3	22						
Extractor 4.0	38	9	660					
Hearst	73	6	86	922				
Dist 4.0	51	10	311	206	1,211			
Dist Hearst	130	11	203	377	565	2,303		
Dist 4.0 All	50	9	253	192	658	532	1,158	
Dist Hearst All	68	4	100	209	318	573	339	1,051
	O*Net	Wikipedia	Extractor 4.0	Hearst	Dist 4.0	Dist Hearst	Dist 4.0 All	Dist Hearst All

Table 5: Pairwise comparison of proposed NER methods

On the contrary, methods based on gazetteer show little overlap with other methodologies. As a meaning of example, the *O*Net* gazetteer retrieves a significant amount of entities, though lower than the rules, but only shares a little part of them with the other methods. *O*Net* and *Dist Hearst* register an apparently high overlap, i.e., 130 over 454 technologies found by *O*Net*. However this

number is not as meaningful as it may appear at first given the large quantity of technologies found by *Dist Hearst*.

One of the leads that can be inferred from Table 5 is that a combination of different methodologies results in a higher number of retrieved technologies. Indeed, a large amount of technologies are only found by specific methodologies, though there is a non negligible part of shared entities. This supports our initial claim: a blended extractor, resourcing to a suite of approaches, is a viable solution and appears to be the best one.

Cases	Gazetteer	Rules	Distributional	Amount	Examples
1			✓	2,870	aircraft, air motor, antenna feeder, card game machine, card shuffler, digital assistant, digital cable, optical transport network, computer hard drive, radio access node, ram memory, reactor pump, relay system, remote control bar, rf circuit
2		✓		749	adsl transceiver, air vent, air vent hole, antenna port, integrate circuit lead frame, information communication technology, pentium ii, pc network station, proximity detector, record playback apparatus, repeater, rfid tag, robot gun, satellite broadcast center, server machine
3	✓			285	air lift pump, alarm system, conveyor system, allen wrench, ball valve, bridge crane, external hard drive, fire alarm system, flowmeter, ground table, pressure transducer, refrigeration unit, roulette wheels, smoke detector, tape recorder
4		✓	✓	639	airplane, antenna, central processing unit, access terminal, adsl, barcode, bluetooth, cellular network device, coaxial cable, lte network, magnetic tape, smart tv, touchscreen, transistor, tv camera
5	✓		✓	80	boiler, centrifuge, computer workstation, control valve, disk, earphone, freezer, mass storage device, motorcycle, optical filter, plasma screen, truck, vacuum chamber, web server, wireless communication system
6	✓	✓		17	air conditioner, air knife, decoiler, deflector, hemodynamic monitor, inclinometer, microsoft windows, pacemaker, php, pointing device, radiant heater, signal generator, track ball, water jet cutter, wearable computer
7	✓	✓	✓	91	defibrillator, desktop computer, card reader, forklift, javascript, linux, mainframe computer, microprocessor, optical sensor, pager, robotic arm, router, slotmachine, touch screen, workstation

Table 6: Quantitative Comparison of the different NER systems **Note:** The table shows how many entities are found by each subset of the technologies, the first three columns indicate which families are considered in each row and the fourth the number of entities found by them, finally the fifth shows some examples belonging to the respective group of extractors.

Table 6 reports the number of entities found by each subset of methods. The technologies identifies by more than one approach are counted only in the cases that contains all those methods. For example, *microsoft windows* is found by rules and gazetteer, so it is counted in the sixth case and not in the second nor

in the third.

The highest intersection is between rules and distributional methods. It may depend on the fact that these two methods provide the most numerous among all extractions.

A large share of the extractions performed by the gazettes are only found by them (285), underlining again the ability of this solution to provide few high quality results, as expected by a humanly curated algorithm.

The least intersection is among *Rules* and *Gazetteer* (17), while the amount of technologies shared by all methods is higher (91). Therefore, we can infer that when a technology is found by both *Gazetteer* and *Rules* than it is also found by *Distributional* methods. A consequence of this analysis is that one could consider giving up rules for the sake of simplicity, whereas gazetteer appear to provide good improvements compared to how much information they recover.

In Table 6 we also report a qualitative comparison among the technologies found by different families of methodologies. We provide a number of examples of the the extractions to underline how different methodologies are highly complementary.

Some very specific technologies (e.g., *air lift pump*) are only found by gazetteer, whereas other simpler ones either only by rules (e.g., *air vent*), or only by distributional methods (e.g., *air motor*). There is also an opposite pattern, that is a general entity found by the rules and by distributional methods (e.g., *antenna*) is then generalized through the latter, identifying new entities (e.g., *antenna feeder*), though not all of them (e.g., *antenna port*).

This example highlights again how each methodology is helpful to the final goal of technologies identification and that a blended solution using all the strategies we attempted would be the best.

We can notice also that the technologies detected by all methods (last row of Table 6) belong in large part to electronics, as expected by the strongest performance all solutions have on IPC group H04J 1/00.

In addition, the extracted entities present all types of morphological patterns. Among the examples reported in Table 6 we find:

- *pentium ii, linux, javascript:* proper nouns;
- *truck:* common noun;
- *vacuum chamber, air conditioner:* noun + noun;
- *fire alarm system, card game machine:* noun + noun + noun;
- *digital assistant, smart tv:* adjective + noun;
- *remote control bar:* adjective + noun + noun;
- *computer hard drive:* noun + adjective + noun;
- *integrate circuit lead frame:* adjective + noun + adjective + noun.

We remark that beyond the few examples we reported in Table 6, inspecting all the extractions, we have seen that for entities shorter than three words we find all patterns in large numbers and high quality, for entities that are 4 words long or more, they are rarer but keep a consistent quality.

4.3. Efficiency of the NER Systems

We conclude the quantitative analysis of the results discussing on the efficiency of the NER systems. Figure 1 reports the number of correctly found technologies in each IPC group by each method versus the time it took to perform the extraction.

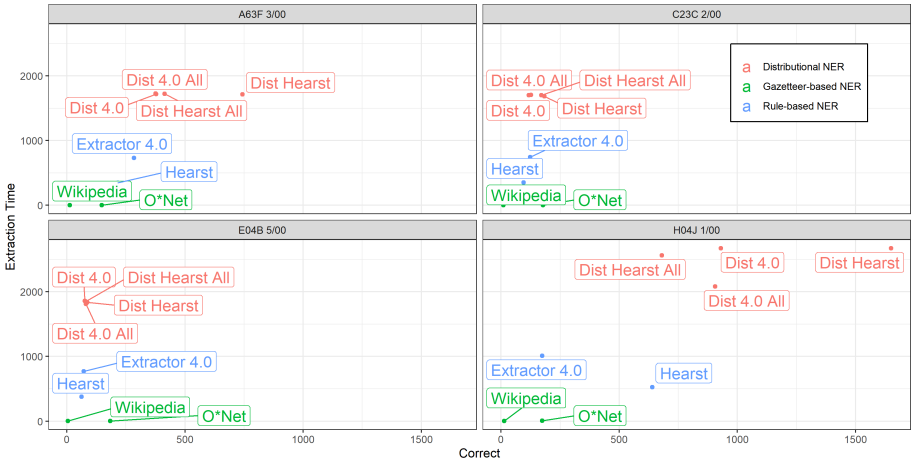

Figure 1: Number of Correct Technologies vs Extraction Time for IPC group

In general, all the methodologies are coherent with reference to extraction time. The gazettes are the fastest approach, *Extractor 4.0* and *Hearst* are placed as second best, and all the Dist methodologies appear to be the slowest. These results are consistent with the number of technologies each solution finds, though not directly dependent.

A question that comes to mind looking at these results is whether the value provided by distributional methods is worth their extra time cost. Different approaches can be used for different tasks. In the case of emerging technologies mapping, distributional methods can lead to better performances, since on average they are able to find more technologies. For mapping more stable sector, gazetteer approaches can be more efficient, since they require less manual revision. In cases of high level studies involving large corpora of patents (Arts et al., 2021; Jang et al., 2021; Giordano et al., 2021), the distributional approach can make the time costs prohibitive for tagging the technologies in the text, while adding a limited amount of significant information. Indeed, a broader limitation of *Distributional* methods is the higher effort in development phase than *Gazetteer* or *Rules*.

Beyond time consumption, Figure 1 shows another interesting fact. The performance of the distributional methodologies depends on the performance of the method they are based on, as observable by the horizontal components of the figure. We discussed above on the reasons for this dependence, here we want to highlight the magnitude. For instance, *Dist Hearst* depends on *Hearst*. Indeed, whenever *Hearst* retrieves a larger number of technologies (Correct) in a class with respect to another, a larger improvement in the number of technologies found by *Dist Hearst* is registered. To see this, we notice how between IPC groups A63F 3/00 and H04J 1/00 there is an increase of 423 in the number of Correct for *Hearst*, while for *Dist Hearst* the increase is 908, and a similar behaviour happens whenever *Hearst* increases. This is relevant since it hints that larger experiments could lead to a strong gain in terms of collected technologies. In addition, as previously mentioned, the better performance of *Hearst* in A63F 3/00 and even more in H04J 1/00 are due mainly to the set of words we use as general technologies' hypernyms, which are a good match for these two IPC groups related to tools and machinery. On the contrary, for IPC groups C23C 2/00 and E04B 5/00, dealing with processes, the chosen words perform more poorly.

4.4. Benchmarking trough Comparative Assessment

In this section, we attempt to compare our methodologies to previous studies in literature. As discussed in Section 2, Jang et al. (2021) develop a method to extract technological information from textual data of patents, aiming at investigating a given technological domain. Keeping in mind the different goals between this work and our analysis, we replicate their methodology. So we show a qualitative comparison on the extractions obtained with the two approaches (i.e., our methodologies and the one in Jang et al. (2021)) . Note that we only replicate the part of the work from Jang et al. (2021) that can be compared to ours, namely the vocabulary construction, and that we did not perform human evaluation on this set of technologies.

We analysed the most recurring technologies and the number of patents where they were found for each IPC group to compare the type of the extracted technologies and their frequency. Figure 2 reports the results of our approach and Figure 3 depicts the ones obtained through the method developed in Jang et al. (2021).

Our approaches are able to identify field specific technologies. For example, in Figure 2 we can find *valve* in C23C 2/00 (in the top right chart), defined as "Hot-dipping or immersion processes for applying the coating material in the molten state without affecting the shape" which involves working with liquids; or *clip* in E04B 5/00 (in the bottom left chart), described as "Floors; Floor construction with regard to insulation" and thus involving constructions. Anyway, some entities were found in other less suited IPC groups. For instance, *level*, which is a good match for E04B 5/00 (in the bottom left chart), was found also in 71 patents of A63F 3/00 (in the top left chart). However, since the revision performed manually takes into account field specificity in this IPC group it was removed. These results can prove the validity of choosing diversified IPC

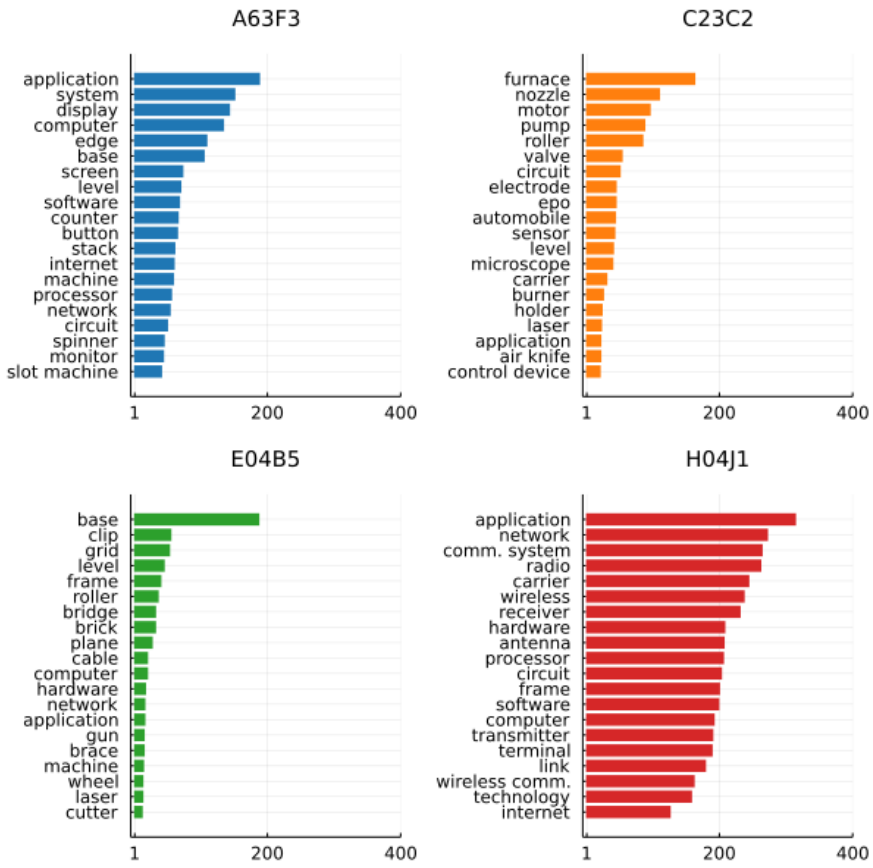

Figure 2: Top 20 frequently technologies for each IPC group using our approach

groups and the generality of the methods we propose, which are able to detect field specific items though with different performances on different IPC group. Some general entities are found as well in many IPC groups and they may be not informative. As a meaning of example, *computer* is listed in three charts in Figure 2 since it is found in all IPC groups except C23C 2/00. However, those general terms can not be removed a priori because they do represent technologies although in a very general sense. The loss of term specificity in some cases and the extraction of too general technologies are both limitations of our work which we intend to further investigate in the future.

The IPC group H04J 1/00 registers the highest value of occurrences, the second is A63F 3/00 and last are C23C 2/00 and E04B 5/00. The large margin of H04J 1/00 can be explained with the biases in the methodologies we employ, as already discussed.

The terms reported in the charts of Figure 3, such as *steel* in C23C 2/00

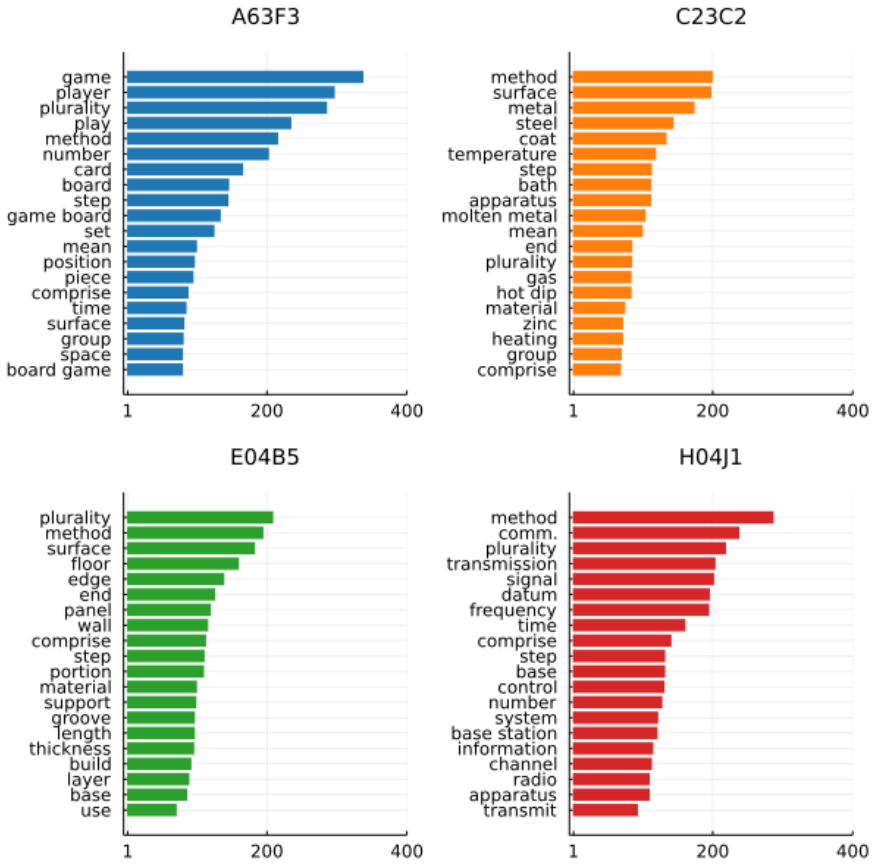

Figure 3: Top 20 frequently found entities for each IPC group using the approach of Jang et al. (2021)

(in top right chart) and *base station* in H04J 1/00 (in bottom right chart), appear very general. Comparing Figure 3 and Figure 2, we can highlight the specificity of the technologies detected in most IPC groups by our solution (e.g. *brick* in E04B 5/00 and *pump* in H04J 1/0). This behaviour is due to the higher specificity we impose on our methods, using the words shown in Section 3, *system*, *unit*, *device*, *apparatus* and so on. A direct implication of this aspect is the difference in the number of occurrences of words, since more specific terms appear more rarely than more general ones. Indeed, the number of occurrences is always lower for our extractions except for IPC groups H04J 1/00, where the top 20 extractions are more general for our methods as well and thus the occurrences are closer. The solutions therefore differ in terms of generality and specificity of lexical information extracted. Let us remark how this difference is not negative for neither of the works, indeed it pushes each one towards its goal and does not make an alternative one of the other.

In Figure 3, we find as well terms that don't represent technologies. Words like *player* in A63F 3/00, *support* in E04B 5/00, *comprise* in C23C 2/00 and *number* in H04J 1/00 are not technologies on their own. Indeed, the goal of Jang et al. (2021) is to develop a lexicon of technology related terms. So that the identification of *player* in the patents of the IPC group A63F 3/00 concerning board and raffle games is a good result from their point of view. Anyway, this results does not fit our goal of identifying technologies. These considerations motivates the large difference in the type of terms.

The compared analysis underlines the choices made by the two works and so the respective purposes. Let us use an example to show how these two methodologies can be used in combination for different aspects of the same problem. The studies on the evolution of technological domain require a wide perspective enabling a high level analysis of the changes it undergoes over time. In these cases, Jang et al. (2021) is able to create a broad map of the concepts used in the description of the relevant entities. The analysis on targeted technologies over selected sets of patents demands for a fine grained approach. Thus, our proposed methodologies are more suited for an in-depth study on a certain technology. Therefore, we believe these two solutions can work in parallel towards the challenging goal of technology-related information extraction from patents.

5. Conclusions

Our work offers a viable Named Entity Recognition system based on rules, gazetteer and machine learning techniques. The system attempts to extract technologies from patent documents, providing quantified metrics to compare the approaches developed in this paper. These metrics are based on traditional measures, namely precision and recall. We estimate the time to perform each extraction to support researchers and practitioners in understanding advantages and limitations of our system. We test our method using four case studies collected with patent classification system-based (IPC) search strategy. For each field of analysis, we show the ability of the NER system in unveiling the most cited technologies. The proposed method may be a valid technology identification tool that can be integrated or used in conjunction to other text analysis pipelines to support academics and industrial actors in investigating a technological domain.

Our system is able to collect $4,731$ technologies from $1,600$ patents. The paper outperforms previous literature in terms of precision and recall (Giordano et al., 2021). Moreover, we compare the results of the technology extraction process with the work of Jang et al. (2021). Our paper allows to analyse a technological domain with a fine grained level, avoiding the noise brought by generic terms. As pointed out in section 4, the NER system enables us to identify an average of 0.35 technologies per sentence (approximately 1 technology every 3 sentences) and a median of 10 distinct technologies per patent. Therefore, technologies are not rare words in patents if compared with other recognizable entities, such as users (Chiarello et al., 2018a), advantages/drawbacks (Chiarello et al., 2017), affordances (Chiarello et al., 2019) and biases (Melluso et al., 2021).

From an academic point of view, the methodologies proposed in this paper may benefit other streams of literature regarding technological forecasting. The technology extractor may be used in conjunction with the current methods for the fostering of emerging technologies. In their seminal work, Porter et al. (2019) use text mining and a R&D emergence indicator based on four criteria for mapping the pathways of emerging technologies: novelty, persistence, growth and community. However, the method proposed by the authors recognizes as technologies also terms that may not be defined as such. Our paper may be used in the work of Porter et al. (2019) to extract technologies and calculate its emergent indicator (called EScores by the authors). The technology extractor can be of interest to the identification of emerging technologies in synergy with other text analysis methods. For example, in (Ranaei et al., 2020), the authors may apply our method to perform LDA and tf-idf to evaluate technological emergence in two case studies. Our study could improve the results of the NER system proposed by Giordano et al. (2021) and Melluso et al. (2020) for extracting technologies in the study of convergence. Indeed, the authors demonstrate how the identification of technologies before analysing technological convergence is a valuable method to improve the precision of the results.

From an industrial point of view, the system developed in this paper may improve the analysis of technological domains to map the landscape and fore-

cast the diffusion of technologies. Traditional methods for the identification of competitors and partners using paper documents may benefit from our system to include additional information in the analysis. For example, a technology extraction process may be involved in the text mining pipeline of Vicente-Gomila et al. (2017) to consider the competitors also from a technological point of view. Fareri et al. (2020) use text mining to identify industry 4.0 technologies and estimating the impact of the fourth industrial revolution on the work force of a multinational company. Similarly, practitioners may employ the method developed here for the analysis of technological competences to assess the need of re-skilling or up-skilling at company level.

After mentioning the possible applications, let us outline the limitations of our work. The strongest limitation lies in the precision our models achieve. Although we manage to extract a large number of technologies, only about 40% of the entities we extract are identified as correct by the manual evaluation, while in other fields (Fan et al., 2020) higher performances are achieved. The low value can be understood considering two main aspects. First, we did not use any manually tagged gold training set because, to the best of our knowledge, there are no examples of such a set for patents. We resource to gazettes and rules to create a data-set. This solution for developing a training set poses difficulties to improve beyond a certain threshold. The development of a gold set is a challenge for future work that we intend to tackle hopefully learning from the findings of this one. In parallel, we also mention the absence of a proper evaluation of the recall. Though the large number of technologies we extracted and doubled checked makes us confident that we do manage to extract a non-trivial share of the knowledge available in the patents we analyse, its time cost prevents us from a manual evaluation of the recall.

The second limitation is posed by the absence of extensive research on the topic. Indeed, the NER solutions in other fields are the product of several works relying on one another. For our case, besides few exceptions (Sarica et al., 2020; Jang et al., 2021), which are still not entirely in line with what we do, and very few works that are closely related (Giordano et al., 2021), the methodology still needs refinement. However, our job poses a strong and structured starting point for such a development which we intend to try and further investigate.

We also remark that relying on search and retrieval of patent records from patent database may influence the results because of the access conditions of the database. Indeed, there are not many examples of freely available databases. Excepting very few (notably, Freepatentsonline[6]), most patent databases require licensing or give a limited access. And this is probably one of the sources of boundaries in this line of research.

The third limitation is the use of a specific and not too extensive set of patents, generally information extraction from these documents are more extensive (Arts et al., 2021). The choice of a smaller number of patents was made on purpose to be able to manually evaluate the entire set of extractions and

[6]Available: `https://www.freepatentsonline.com/`, Accessed: July 28, 2021

to have a contained range of possibilities in order to achieve a good assessment of all positive and negative sides of our work. Note also that, though not too large, the set of patents is carefully constructed, as described in Section 3 in order to balance between a controlled setting and a variegated one. Different IPC groups, with different types of technologies and applications, allow us to assess which aspects of the text are properly caught by our extractors and which instead are not.

The fourth limitation is the constrained numbers of pre-trained BERT models used for constructing the NER distributional methods and type of documents used as textual corpus. In the article we observe that the distributional methods reach a higher level of performance if compared to other ones. However, we only use the BERT model pre-trained with Wikipedia pages that is not specific for the technological phenomena analysis. In future, we will try to use PatentBERT (Lee and Hsiang, 2020) and compare its performance with those reached by our methods for the task of technological entity recognition. Moreover, we tested our approach only on patents, but there are other types of text where our methodology could be used, such as the scientific papers. Patents and scientific publications share several aspects: they are both technical documents, they describe new ideas with the purpose to spread the acquired knowledge, they use rare words and specific vocabulary as the evolution of every field of knowledge imposes a continuously renewing lexicon. This creates a level of similarity among these two text corpora which we speculate would make our solution extendable to the one we did not attempt yet. In future investigations, it might be possible to use this different corpus of text (i.e. scientific articles) with a distributional NER method fine-tuned with the SciBERT model (Beltagy et al., 2019), that is demonstrated to be more suitable for this particular type of documents.

The future steps we intend to tackle are those that would allow overcoming the limitations mentioned above. First of all, the attempt to create a golden set from the work we did so far to develop more effective and precise models. One possible method for saving time and cost in the creation of training set using rule-based NER is a human evaluation of the training set in the form of Active Learning. The expert checks the quality of a sample of the initial training dataset obtained with the rule-based NER (i.e., 30% of the training dataset can be used) and remove terms that are not technologies. Then, the distributional NER is trained on the sample of the training dataset for a quality check. The quality of the distributional method is then measured, if it reaches an acceptable quality we stop the training, otherwise the same methodology can be re-applied on more data until a certain degree of training data quality is reached. Afterwards, the development of a test set will enable a more precise assessment of the recall of our systems. We are also interested in an expert developed baseline, indeed, to the best of our knowledge, an accurate assessment of how well humans can identify technologies in patents is still missing in literature. Finally, new experiments on larger patent sets or on different data sources (e.g., scientific papers) can be carried out. For the design of future experiments this work suggests that a multi-domain model working on different IPC classes is the road to follow since we see limited advantages at a high cost from training

IPC focused models.

More in general, we believe there is a large amount of valuable knowledge in patents that is hardly extracted with the right level of detail and that is awaiting to be found.

References

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

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Arts, S., Hou, J., and Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Asche, G. (2017). "80% of technical information found only in patents"–is there proof of this? *World Patent Information*, 48:16–28.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Belz, A. and Reiter, E. (2006). Comparing automatic and human evaluation of nlg systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Bernier-Colborne, G. and Langlais, P. (2020). HardEval: Focusing on challenging tokens to assess robustness of NER. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1704–1711, Marseille, France. European Language Resources Association.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bonaccorsi, A., Chiarello, F., Fantoni, G., and Kammering, H. (2020). Emerging technologies and industrial leadership. a wikipedia-based strategic analysis of industry 4.0. *Expert Systems with Applications*, 160:113645.

Breitzman, A. and Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research policy*, 44(1):195–205.

Caragea, D., Chen, M., Cojoianu, T., Dobri, M., Glandt, K., and Mihaila, G. (2020). Identifying fintech innovations using bert. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1117–1126. IEEE.

Carley, S. F., Newman, N. C., Porter, A. L., and Garner, J. G. (2018). An indicator of technical emergence. *Scientometrics*, 115(1):35–49.

Carlson, A., Gaffney, S., and Vasile, F. (2009). Learning a named entity tagger from gazetteers with the partial perceptron. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 7–13.

Carroll, L. S. L. (2017). A comprehensive definition of technology from an ethological perspective. *Social Sciences*, 6(4).

Chang, S.-H. and Fan, C.-Y. (2016). Identification of the technology life cycle of telematics: A patent-based analytical perspective. *Technological Forecasting and Social Change*, 105:1–10.

Chiarello, F., Belingheri, P., Bonaccorsi, A., Fantoni, G., and Martini, A. (2021). Value creation in emerging technologies through text mining: the case of blockchain. *Technology Analysis & Strategic Management*, pages 1–17.

Chiarello, F., Bonaccorsi, A., and Fantoni, G. (2020). Technical sentiment analysis. measuring advantages and drawbacks of new products using social media. *Computers in Industry*, 123:103299.

Chiarello, F., Cimino, A., Fantoni, G., and Dell'Orletta, F. (2018a). Automatic users extraction from patents. *World Patent Information*, 54:28–38.

Chiarello, F., Cirri, I., Melluso, N., Fantoni, G., Bonaccorsi, A., and Pavanello, T. (2019). Approaches to automatically extract affordances from patents. In *Proceedings of the Design Society: International Conference on Engineering Design*, volume 1, pages 2487–2496. Cambridge University Press.

Chiarello, F., Fantoni, G., Bonaccorsi, A., et al. (2017). Product description in terms of advantages and drawbacks: Exploiting patent information in novel ways. In *DS 87-6 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 6: Design Information and Knowledge, Vancouver, Canada, 21-25.08. 2017*, pages 101–110.

Chiarello, F., Trivelli, L., Bonaccorsi, A., and Fantoni, G. (2018b). Extracting and mapping industry 4.0 technologies using wikipedia. *Computers in Industry*, 100:244 – 257.

Cho, Y. and Kim, M. (2014). Entropy and gravity concepts as new methodological indexes to investigate technological convergence: Patent network-based approach. *PloS one*, 9(6):e98009.

Choi, S., Lee, H., Park, E., and Choi, S. (2022). Deep learning for patent landscaping using transformer and graph embedding. *Technological Forecasting and Social Change*, 175:121413.

Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological forecasting and social change*, 73(8):981–1012.

De Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., and de la Potterie, B. v. P. (2013). The worldwide count of priority patents: A new indicator of inventive activity. *Research Policy*, 42(3):720–737.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, 25(3):233–242.

Fan, C., Wu, F., and Mostafavi, A. (2020). A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access*, 8:10478–10490.

Fantoni, G., Apreda, R., Dell'Orletta, F., and Monge, M. (2013). Automatic extraction of function–behaviour–state information from patents. *Advanced Engineering Informatics*, 27(3):317–334.

Fareri, S., Fantoni, G., Chiarello, F., Coli, E., and Binda, A. (2020). Estimating industry 4.0 impact on job profiles and skills using text mining. *Computers in industry*, 118:103222.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fleiss, J. L., Levin, B., Paik, M. C., et al. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Frey, C. B. and Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280.

Giordano, V. F., Chiarello, F., Melluso, N., Fantoni, G., and Bonaccorsi, A. (2021). Text and dynamic-network analysis for measuring technological convergence: A case study on defence patent data. *IEEE Transactions on Engineering Management*.

Gustafsson, R., Kuusi, O., and Meyer, M. (2015). Examining open-endedness of expectations in emerging technological fields: The case of cellulosic ethanol. *Technological Forecasting and Social Change*, 91:179–193.

Hain, D. S., Jurowetzki, R., Buchmann, T., and Wolf, P. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.

Han, E. J. and Sohn, S. Y. (2015). Patent valuation based on text mining and survival analysis. *The Journal of Technology Transfer*, 40(5):821–839.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Hofmann, P., Keller, R., and Urbach, N. (2019). Inter-technology relationship networks: Arranging technologies through text mining. *Technological Forecasting and Social Change*, 143:202–213.

Hossari, M., Dev, S., and Kelleher, J. D. (2019). Test: A terminology extraction system for technology related terms. In *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*, pages 78–81.

Huang, Y., Zhu, F., Porter, A. L., Zhang, Y., Zhu, D., and Guo, Y. (2020). Exploring technology evolution pathways to facilitate technology management: From a technology life cycle perspective. *IEEE Transactions on Engineering Management*.

Jang, H., Jeong, Y., and Yoon, B. (2021). Techword: Development of a technology lexical database for structuring textual technology information based on natural language processing. *Expert Systems with Applications*, 164:114042.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., and Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606.

Joung, J. and Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological Forecasting and Social Change*, 114:281–292.

Karvonen, M. and Kässi, T. (2013). Patent citations as a tool for analysing the early stages of convergence. *Technological Forecasting and Social Change*, 80(6):1094–1107.

Kay, L., Newman, N., Youtie, J., Porter, A. L., and Rafols, I. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology*, 65(12):2432–2443.

Kim, G. J., Park, S. S., and Jang, D. S. (2015). Technology forecasting using topic-based patent analysis. *JSIR Vol.74(05)*.

Kuhn, J., Younge, K., and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1):109–132.

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., and Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change*, 125:236–244.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

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

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020b). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lee, J.-S. and Hsiang, J. (2020). Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965.

Li, S., Hu, J., Cui, Y., and Hu, J. (2018). Deeppatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2):721–744.

Li, X., Xie, Q., Jiang, J., Zhou, Y., and Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, 146:687–705.

Lidén, C. and Setréus, E. (2011). Patent prosecution at the european patent office: what is new for life sciences applicants? *Expert opinion on therapeutic patents*, 21(6):813–817.

Liu, S.-H., Liao, H.-L., Pi, S.-M., and Hu, J.-W. (2011). Development of a patent retrieval and analysis platform–a hybrid approach. *Expert systems with applications*, 38(6):7864–7868.

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of industrial information integration*, 6:1–10.

Magee, C. L., Basnet, S., Funk, J. L., and Benson, C. L. (2016). Quantitative empirical trends in technical performance. *Technological Forecasting and Social Change*, 104:237–246.

Magerman, T., Van Looy, B., and Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2):289–306.

Maghrebi, M., Abbasi, A., Amiri, S., Monsefi, R., and Harati, A. (2011). A collective and abridged lexical query for delineation of nanotechnology publications. *Scientometrics*, 86(1):15–25.

Melluso, N., Bonaccorsi, A., Chiarello, F., and Fantoni, G. (2020). Rapid detection of fast innovation under the pressure of covid-19. *PloS one*, 15(12):e0244175.

Melluso, N., Pardelli, S., Fantoni, G., Chiarello, F., and Bonaccorsi, A. (2021). Detecting bad design and bias from patents. In *Proceedings of the Design Society: International Conference on Engineering Design*, volume 1. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Niemann, H., Moehrle, M. G., and Frischkorn, J. (2017). Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. *Technological Forecasting and Social Change*, 115:210–220.

No, H. J. and Park, Y. (2010). Trajectory patterns of technology fusion: Trend analysis and taxonomical grouping in nanobiotechnology. *Technological Forecasting and Social Change*, 77(1):63–75.

Odat, S., Groza, T., and Hunter, J. (2015). Extracting structured data from publications in the art conservation domain. *Digital Scholarship in the Humanities*, 30(2):225–245.

OECD (2009). *OECD Patent Statistics Manual*.

Ozcan, S. and Islam, N. (2017). Patent information retrieval: approaching a method and analysing nanotechnology patent collaborations. *Scientometrics*, 111(2):941–970.

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

Park, I., Jeong, Y., Yoon, B., and Mortara, L. (2015). Exploring potential r&d collaboration partners through patent analysis based on bibliographic coupling and latent semantic analysis. *Technology Analysis & Strategic Management*, 27(7):759–781.

Pawar, S., Srivastava, R., and Palshikar, G. K. (2012). Automatic gazette creation for named entity recognition and application to resume processing. In *Proceedings of the 5th ACM COMPUTE Conference: Intelligent & scalable system technologies*, pages 1–7.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. (2019). Meaningful explanations of black box ai decision systems. In *Proc. of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and 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, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Porter, A. L., Garner, J., Carley, S. F., and Newman, N. C. (2019). Emergence scoring to identify frontier r&d topics and key players. *Technological Forecasting and Social Change*, 146:628–643.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Ramanathan, K. (1994). The polytrophic components of manufacturing technology. *Technological Forecasting and Social Change*, 46(3):221–258.

Ranaei, S., Suominen, A., Porter, A., and Carley, S. (2020). Evaluating technological emergence using text analytics: two case technologies and three approaches. *Scientometrics*, 122(1):215–247.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., and Kupfer, D. J. (2013). Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, 170(1):59–70.

Righi, C. and Simcoe, T. (2019). Patent examiner specialization. *Research Policy*, 48(1):137–148.

Robinson, D. K., Huang, L., Guo, Y., and Porter, A. L. (2013). Forecasting innovation pathways (fip) for new and emerging science and technologies. *Technological Forecasting and Social Change*, 80(2):267–285.

Roller, S., Kiela, D., and Nickel, M. (2018). Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 358–363, Melbourne, Australia. Association for Computational Linguistics.

Rotolo, D., Hicks, D., and Martin, B. R. (2015). What is an emerging technology? *Research Policy*, 44(10):1827–1843.

Salton, G., Fox, E. A., and Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Sampson, M., Zhang, L., Morrison, A., Barrowman, N. J., Clifford, T. J., Platt, R. W., Klassen, T. P., and Moher, D. (2006). An alternative to the hand searching gold standard: validating methodological search filters using relative recall. *BMC medical research methodology*, 6(1):1–9.

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

Small, H., Boyack, K. W., and Klavans, R. (2014). Identifying emerging topics in science and technology. *Research policy*, 43(8):1450–1467.

Song, B. and Suh, Y. (2019). Identifying convergence fields and technologies for industrial safety: Lda-based network analysis. *Technological forecasting and social change*, 138:115–126.

Song, C. H., Elvers, D., and Leker, J. (2017). Anticipation of converging technology areas—a refined approach for the identification of attractive fields of innovation. *Technological Forecasting and Social Change*, 116:98–115.

Sternitzke, C. (2010). Knowledge sources, patent protection, and commercialization of pharmaceutical innovations. *Research Policy*, 39(6):810–821.

Suominen, A., Toivanen, H., and Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, 115:131–142.

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010). A compared r&d-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological forecasting and social change*, 77(7):1037–1050.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Trappey, A. J., Chen, P. P., Trappey, C. V., and Ma, L. (2019). A machine learning approach for solar power technology review and patent evolution analysis. *Applied Sciences*, 9(7):1478.

Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., Sung, T.-Y., and Hsu, W.-L. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC bioinformatics*, 7(1):1–8.

Tseng, F.-M., Hsieh, C.-H., Peng, Y.-N., and Chu, Y.-W. (2011). Using patent data to analyze trends and the technological strategies of the amorphous silicon thin-film solar cell industry. *Technological forecasting and social change*, 78(2):332–345.

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information processing & management*, 43(5):1216–1247.

Vicente-Gomila, J., Artacho-Ramírez, M., Ting, M., and Porter, A. (2021). Combining tech mining and semantic triz for technology assessment: Dye-sensitized solar cell as a case. *Technological Forecasting and Social Change*, 169:120826.

Vicente-Gomila, J. M., Palli, A., de la Calle, B., Artacho, M. A., and Jimenez, S. (2017). Discovering shifts in competitive strategies in probiotics, accelerated with techmining. *Scientometrics*, 111(3):1907–1923.

Volti, R. (2005). *Society and technological change*. Macmillan.

Waight, N. (2014). Technology knowledge: High school science teachers'conceptions of the nature of technology. *International Journal of Science and Mathematics Education*, 12(5):1143–1168.

Xu, J., Guo, L., Jiang, J., Ge, B., and Li, M. (2019). A deep learning methodology for automatic extraction and discovery of technical intelligence. *Technological Forecasting and Social Change*, 146:339–351.

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

Yoon, J. and Kim, K. (2011). Identifying rapidly evolving technological trends for r&d planning using sao-based semantic patent networks. *Scientometrics*, 88(1):213–228.

Yu, X. and Zhang, B. (2019). Obtaining advantages from technology revolution: A patent roadmap for competition analysis and strategy planning. *Technological Forecasting and Social Change*, 145:273–283.

Zhang, M., Fan, B., Zhang, N., Wang, W., and Fan, W. (2021). Mining product innovation ideas from online reviews. *Information Processing & Management*, 58(1):102389.

Zhang, T., Wang, Y., Wang, X., Yang, Y., and Ye, Y. (2020). Constructing fine-grained entity recognition corpora based on clinical records of traditional chinese medicine. *BMC medical informatics and decision making*, 20(1):1–17.

Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., and Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics*, 123(1):1–29.