World Patent Information 54 (2018) 28–38

Contents lists available at ScienceDirect

World Patent Information

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

Automatic users extraction from patents

Filippo Chiarello[a,*], Andrea Cimino[b], Gualtiero Fantoni[c], Felice Dell'Orletta[b]

[a] Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126, Pisa, Italy
[b] Institute for Computational Linguistics of the Italian National Research Council (ILC-CNR), Via G. Moruzzi,1, Pisa, Italy
[c] Department of Mechanical, Nuclear and Production Engineering, University of Pisa, Largo Lucio Lazzarino, 2, 56126, Pisa, Italy

ARTICLE INFO

Keywords:
Patent analysis
Deep learning
Text mining
User of an invention

ABSTRACT

Patents contain a large quantity of information which is usually neglected. This information is hidden beneath technical and juridical jargon and therefore so many potential readers cannot take advantage of it. State of the art natural language processing tools and in particular named entity recognition tools, could be used to detect valuable concepts in patent documents.

The purpose of the present research is to design a method capable of automatically detecting and extracting one of the multiple entities hidden in patents: the users of the invention.

The method is based on a new approach tailored for users extraction by integrating state-of-the-art computational linguistics tools with a large knowledge base. Furthermore the paper shows a comparison among different machine learning algorithms with the twofold aim of achieving the highest recall and evaluating the performance in terms of precision and computational effort.

Finally, a case study on two patent sets has been conducted to evaluate the effectiveness and the output of the entire tool-chain.

1. Introduction

Nowadays patent data can be used for planning technological strategy [1]. The focus on the technological usefulness of patent data is certainly a great advantage, but this huge research area could hide other useful application for patents. For example, in Ref. [2] the authors consider on one side patents as a source to collect information about technologies and products, and on the other side manuals, handbooks and market reports to collect market information. Since patents are only technological documents many potential patent reader (e.g. designers, marketers) could be taken aside. Despite this problem, some researchers [3] affirm that there is an increasing variety of readers: not only technician and researchers but also marketers and designers who have grown an interest in patent analysis. Nevertheless, to our knowledge there are no researches that aim at facilitating information extraction for non-technological focused patent readers.

The bias that patent are only tech-oriented documents is due to two main reasons:

- Patents are produced to disclose and protect an invention, their content is mainly technical and legal.
- 80% of technical information is not available elsewhere [4,5], so patents are one of the most comprehensive resources for technical analysis.

Focusing on the second point, our hypothesis is that also a fraction of all the other kinds of information (e.g. marketing and sociological information) is not contained elsewhere and it will appear in public documents (e.g. manual handbooks and market reports) in 6–18 months [6].

Unfortunately there are four aspects reducing the non-tech readers' ability to analyze patents efficiently. First of all, an increasingly high number of patent filings generates a massive information overflow [7]; secondly, analyzing patents takes a long time and requires skilled personnel [8]; the quality of patent assessment process is decreasing [9,10] because of the reduced assessment time available for patent examiners; finally, activities like patent hiding, proliferation and bombing, contribute to the generation of confusion and to the loss of time in research and analysis phases [11]. These problems affect non-tech oriented patent readers as well as typical readers, even though the impact may be stronger on the firsts.

The main difference between typical and non-tech patent readers is the information they focus on. *Patent attorneys* and *Intellectual Property (IP) managers* are interested in reading patents for legal reasons to orient the IP direction. Analyzing patents is the core of their work, so they are experts in finding the information they need. Furthermore,

* Corresponding author.
E-mail address: filippo.chiarello@destec.unipi.it (F. Chiarello).

https://doi.org/10.1016/j.wpi.2018.07.006
Received 13 February 2018; Received in revised form 22 May 2018; Accepted 18 July 2018

they can spend most of their work-time on the activity. On the other hand, usually *marketers and designers* (taken as example of non-tech oriented readers) search users' behavioral changes and needs, market trends, designers' vision, R&D trends and competitors' strategies. In addition, they rarely work with patents, so they do not know what and how to search. Lastly, they have short time to spend on the activity, and they waste most of this time understanding the legal and technical jargon used in patents.

In this paper we present a *system for entity extraction from patents*, aimed at putting the focus on valuable non-technical information, making patent content more accessible and usable by non-technical readers like marketers and designers. Marketing and design business functions are always searching for new users and market niches where to experiment new products or to capture new trends, therefore an automatic system able to scan and explore a huge patent set could be useful. Moreover the extracted information could be used as keywords for a standard search in Facebook or Twitter in order to crosscheck such evidences coming from patents.

Entity extraction means to find textual elements which are unique identifiers of entities belonging to predefined classes [12]. To build a suitable system for this task we can not leave aside two aspects:

1. *effectiveness*: selection of useful entities contained in patents, in particular for non-technical readers.
2. *efficiency*: deep knowledge of state of the art techniques for data analysis required for the extraction of these entities from large corpora of documents.

In the present work we focus on the extraction of one of the possible entities hidden in patents: the users of the invention. The concept of user is of great interest to many disciplines, in particular to marketing and design.

Recent Natural Language Processing (NLP) tools and techniques have been proved to be adequate for the extraction of specific information from texts [13]. Unfortunately, NLP tools suffer from a dramatic drop of accuracy when tested on domain specific texts [14] such as patents. Consequently, tools must be adapted to the patent domain [15].

The paper is then organized as follows. In section 2 we present the concept of *user of the invention* and discuss why we want to extract this entity. In the same section we also provide some examples of the state-of-the-art patent text analysis systems. Section 3 presents the process we used to solve the problem. In Section 4 the process is applied to a case study and the outcomes are discussed. Finally, section 5 concludes with an overview of future developments.

2. Users: a key information hidden in patents

In this section we explains the concept of *user of the invention*. Section 2.1 gives a definition of users and presents the way that this concept is exploited in different knowledge fields. Section 2.2 briefly discusses the state of the art of automatic processing of patents, focusing on entity extraction systems.

2.1. Users: definition and usage

Patents are documents that must provide a detailed public disclosure of an invention [16]. An *invention* is a new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof.[1]

The notion of usefulness implies that the invention must have some value and not necessarily for a human entity. In fact, patents usually describe processes, machines or composition of matter which are useful for another process, machine or composition of matter.

Therefore, we distinguish between stakeholders and users, considering the definitions given by the authors in Ref. [17].

Definition 1. Stakeholder: *Stakeholders are entities on which the invention has or will have a positive or negative effect in order to show usefulness.*

This definition covers all possible entities that engage an active or passive relation with the invention. Given the logical condition of usefulness of patents, all patents must have stakeholder information. If a patent has not got any stakeholder information in it the patent application should be rejected.

Definition 2. User: *Users are animated or previously animated entities (human or animal, alive or dead), on which the invention has a positive or negative effect at an unspecified moment.*

Given definition 2, it is clear that every user is a stakeholder while non-users stakeholders include artifacts, machines, manufacturing or operational processes.

Corollary 1. *Multiple roles*: *Identities may have multiple roles as users.*

Our idea of users describes roles, not identities. Animated entities have an identity, as it happens for a specific person. A person has many roles as a user. For example, a working mother starts her day taking on the role of a mom, in which she is expected to feed her children and get them ready for school. At the office she shifts to the role of project manager, so she oversees projects in a timely and professional manner. Working mother, mom or project manager can be considered user roles attributed to the same person, or identity. From this definition it is clear that users are close to what social sciences define as social roles (a focus on the way that social sciences interpret social roles is shown in section 2.1.1).

Afterward, we can outline knowledge fields using the concept of user. Sections 2.1.1, 2.1.2, 2.1.3 and 2.1.4 have a twofold aim: help the reader to understand how the concept of users is interpreted in different knowledge fields; explain the background of the methodology described in section 3.1).

2.1.1. Social sciences: social roles as users

In social sciences *social roles* are comparable with our definition of user. As defined in the psychological dictionary Alleydog,[2] ”*social roles refer to the expectations, responsibilities and behaviors we adopt in certain situations.*”. The example of the working mother shown in section 2.1 is the case of social roles.

The field of social sciences is the only one in which an attempt of automatic extraction of users has been done. In Ref. [18] the authors extracted social roles from Twitter using heuristic methods. The authors looked for all the words preceded by constructions like ”I'm a” and similar variations. This search resulted in 63.858 unique roles identified, 44.260 of which appeared only once. The result of the extraction process is noisy and only a low percentage of the extracted words are social roles. Despite of this noisy extraction, some entities are consistent with our definition of user, e.g. *doctor, teacher*, (*mother*) or *christian*.

Another work [19] tries to identify social roles on Twitter exploiting a set of assumptions. The authors take into account roles, each one with a set of related verbs: if someone uses verbs from a set, that person may cover that particular social role. To sanitize the collection of positively identified users, the authors crowd-sourced a manual verification procedure, using the Mechanical Turk platform.[3] Also here some interesting extractions are performed, obtaining users like *artist, athlete, blogger, cheerleader, christian, DJ*, or *filmmaker*. These two works differ from the present study for what concerns the analyzed texts and the

[1] http://www.uspto.gov.

[2] http://alleydog.com.

[3] https://www.mturk.com/mturk/.

methods to extract the entities. Nevertheless, the extracted set of entities is consistent with our definition of user.

2.1.2. Human Resources Management: workers as users

In organizations, Human Resources Management is the function designed to maximize employees performance [20]. Employees are key actors and they can be considered users according to our definition.

Human Resources Management has tried to classify employees, especially in sub-fields like insurance, social security or work psychology. Usually, we refer to those as lists of jobs. Classifications were made with the goal of grouping similar jobs for educational requirements, job outlooks, salary ranges or work environments to facilitate social analysis and the placement of new workers. Such lists are relevant because, even if they represent just one subset of all the possible users, they contain valid information. Many institutions developed lists of jobs [21].

2.1.3. Medicine: patients as users

Another field of interest is the medical one, since patients can be considered users. Also in this case there are many lists of patients, illnesses and diseases,[4] which are valuable in terms of information contained.

2.1.4. Design and marketing: between users and customers

In the field of *Design* the concept of user plays a central role and it overlaps with our definition of user. Many tools and theories like "User Centered Design" are based on the concept of user [22]. As stated by the authors in Ref. [23], the quality of the design process is proportional to the user needs' satisfaction. It implies that a designer has to understand the user needs; as a consequence he has to discover whom are potential users.

2.2. Computer aided systems for textual patent analysis

Due to the large amount of information contained in patents and the growing interest to exploit this information, huge efforts have been devoted to the development of systems source to automatically extract different kind of information from such an enormous and valuable data.

Many techniques introduced in order to extract textual information from patents come from extensive research advances in the Natural Language Processing field (NLP). NLP is an area of research and artificial intelligence which aims at teaching computers to understand and manipulate natural language text in order to performs different tasks such as information extraction, machine translation and sentiment analysis.

In literature two main approaches have been applied to extract structured data from patent texts:

1. *Keyword approaches*: methods able to produce vector representations of the analyzed documents. Computed vectors can be used for many applications such as patent retrieval by keyword, or patent similarity matching. Even though this approach can be used for several tasks, it is not suitable to catch semantic relationships between entities in sentences. Furthermore, these methods use a blacklist to remove noisy words [24] or use predefined lexicons [25]. The right design of such list dramatically impacts the final output of the analysis [26–28];
2. *Grammatical and syntactical approaches*: methods based on grammatical and syntactical structures extracted by natural language processing tools, such as Part-of-Speech taggers and syntactical parsers. Unlike the keyword based approach, these methods are able to capture the relationships between the entities mentioned in sentences [29,30,31].

The present work belongs to the second group and relies on a battery of natural language processing tools suited for *entity extraction*. Entity extraction focuses on automatically identifying words or phrases that correspond to predefined entity classes, depending on the extraction target, in our case, the user of the invention. Since our proposed computer aided system is based on entity extraction approaches, we analyze state of the art techniques for entity extraction and their application in patent analysis in section 2.2.1.

2.2.1. Entity extraction systems for patent analysis

Named Entity Recognition is the task of identifying entity names like people, organizations, places, temporal expressions or numerical expressions. An example[5] of an annotated sentence for a NER extraction system tailored for user entity extraction from patents, is the following:

Traditionally, < user > guitar players < user/ > or < user > players < user/ > of other stringed instruments may perform in any of a number of various positions, from seated, with the stringed instrument supported on the leg of the < user > performer < user/ > , to standing or walking, with the stringed instrument suspended from a strap.

Methods and algorithms to deal with the entity extraction task are different, but the most effective are the ones based on supervised methods. Supervised methods tackle this task by extracting relevant statistics from an annotated corpus. These statistics are collected from the computation of features values, which are strong indicators for the identification of entities in the analyzed text. Features used in NLP for NER purposes are divided in two main categories:

1. linguistically motivated features, such as n-gram of words (sequences of n words), lemma and part of speech;
2. external resources features as, for example, external lists of entities that are candidates to be classified in the extraction process.

The annotation methods of a training corpus can be of two different kinds: (a) human based, which is time expensive, but usually effective in the classification phase; (b) automatically based, which can lead to annotation errors due to language ambiguity. For instance *driver* can be classified both as a user (the operator of a motor vehicle), or not a user (a program that determines how a computer will communicate with a peripheral device).[6] Different training algorithms, such as Hidden Markov Models [32], Conditional Random Fields (CRF) [33] Support Vector Machines (SVM) [34], or more recently Bidirectional Long Short Term Mermory-CRF Neural Networks [35–38] are used to build a statistical model based on features that are extracted from the analyzed documents in the training phase.

For what concerns the extraction of specific entities in patents, improve the accuracy of domain specific patent retrieval systems [39] has been a matter of great interest both in academia and commercial organizations.

In Ref. [40] the authors propose a machine learning patent based NER system that identifies key terms in patent documents and recognizes products, services and technologies names in patent summaries and claims. In this work a study was conducted to identify the most relevant features for this classification task and by using lexical features like word unigrams, word bigrams and word trigrams, their NER system reached an F1 score of 65.40%. The authors compared their NER tagging system resulting from the optimal feature selection method, with the human tagged corpus, showing that the kappa coefficient was 0.67. This result was better than the kappa coefficient between two human taggers (0.60).

Other entity extraction systems for the patent domain were proposed for the CHEMDNER (chemical compounds and drug names

[4] http://www.cdc.gov/DiseasesConditions.

[5] Sentence belonging to patent US 20050022650 A1.

[6] http://wordnetweb.princeton.edu.

recognition) community challenge [39]. Organizers wanted to promote the development of novel, competitive and accessible chemical text mining systems. The best results were obtained by the *tmChem* system [41], achieving a 0.8739 f-measure score. The authors proposed an ensemble system composed of two Conditional Random Fields based classifiers, each one using hard feature engineering such as lemmatization, stemming, lexical and morphological features. In addition, external lists of entities were exploited to recognize whether a token matched the name of a chemical symbol or element, each one used to compute features to be added in the final statistical model.

In chemical field, NER systems have good performances mainly for two reasons:

- chemical entity names (such as molecular formulas) have very frequent orthographic patterns
- contexts that surround such entities are very similar to each other

In our case, these two features can not be exploited for users extraction, since users are named entities that have more confuse surrounding contexts.

In addition, given our definition of user as a subset of the stakeholders of an invention, it is reasonable to state that these two kinds of entities share very similar lexical contexts. For this reason, an automatic user extraction system may tag as user other type of entities like components, systems, products, processes or services which are effectively stakeholders of an invention described in patents. A machine that benefits from a innovative process that extends its life or improve its performance is described in the text as an user that benefits from a new medical device that extends his/her life or improve his/her performance. This source of ambiguity influences the output of our classification system.

Another important key factor concerning the high performances of the systems in the chemical field is that many external resources, such as lists of chemicals or product names, are available: similar external knowledge can not be exploited in the case of users extraction, since lists of users exist, but they are spread among different knowledge bases, as shown in section 2.1, and need human review in order to be useful for our purposes.

Finally, all the above described systems use manually annotated training sets, while we resorted to a method that automatically creates a large training set. Since analyzed patents in our experiments belong to different domains and building a manually annotated corpus is expensive in terms of human resources, the procedure described in section 3.3.3 allows to automatically collect informative sentences to be used in the training phase.

3. Approach for automatic users extraction from patents

In this section we will show the approach used to extract the users of the invention described in a patent. The proposed process is shown in Fig. 1 and its phases are:

1. *Generation of an input list of users*: search all possible sources with the aim of creating an input list of users with the largest possible coverage (section 3.1);
2. *Patent set selection*: select the set of documents from which extract the users (section 3.2);
3. *Patent text pre-processing*: application of natural language processing tools on the documents with the aim of preparing them for the automatic user extraction;
4. *Automatic patent set annotation 1*: projection of the input list of users on the text to generate the Automatically Annotated Patent Set 1;
5. *Relevant sentences extraction*: selection of sentences containing at least one user to generate an informative training set;
6. *Automatic patent set annotation 2*: generation of a statistical model by a machine learning algorithm based on the training set sentences and automatically tagging the patent set to generate the Automatically Annotated Patent Set 2;
7. *Difference computation*: generation of the new list of users by computing the difference between the lists of users found in the automatically annotated patent set 1 and 2;
8. *Manual review*: manual selection of the entities that, in the new list of users, are effectively users. This new list will enrich the original list of users. This phase is described in section 3.4.

3.1. List of users generation

To generate the input list of users, we used two different approaches: a bottom-up approach and a top-down approach. The bottom-up approach is based on the merge of lists from heterogeneous sources. In the present work we used the following lists of entities:

- *Lists of jobs*: obtained by using the document [21]. Such list was merged with more recent lists[7] collecting a total of 11.142 users;
- *Lists of sports and hobbies*: obtained by the union of lists[8] [9] for a total of 9.660 users;
- *List of animals*: obtained by parsing a web-page[10] for a total of 600 users;
- *Lists of patients*: obtained by merging two web pages[11] [12] for a total of 14.609 users;
- *List of generic words*: manually generated. It contains users with a higher level of abstraction (such as *person* or *human being*), 56 users.

Bottom-up approach produced a list of 35.767 entries.

Afterwards, a top-down approach was applied. Starting from the list generated with the bottom-up approach, we looked for alternative methods to indicate a user, finding defined word patterns. The most relevant are:

- Patterns like ”hobby_term + practitioner” for the hobbies;
- Patterns like ”person who has + disease_term” or ”suffering from + disease_term” for the diseases;
- Patterns like ”practitioner of + sport_term” for sports.

Top-down approach generated a total of 41.090 entries.

The whole process generated a total of 76.857 users and gave us a reasonable number of terms to be used in the next step of the process.

Obviously our lists have a limited coverage and, therefore, they do not contain all variations of a certain user. For instance, the lists miss some users belonging to the classes mentioned above (e.g. new jobs emerged in the last years) and all the alternative ways for referring to a user we do not spotted in the top-down approach. For example our lists miss jobs like *data analyst*, *lap dancer*, *undertaker*, *mortician* and *thief* or patients with emerging diseases like *work-alcoholic* and *web-addicted*. In addition, our lists miss a class of users related to religious groups, containing users like *christians* or *jewish*. Such terms have intentionally **not** been introduced in the input list because we considered these terms as candidates to be extracted by the process in our case study (section 4).

3.2. Patent set selection

Our choice of patent sets aimed at challenging our system to find new users missing in the input list. To reproduce a patent set selection,

[7] http://www.careerplanner.com/DOTindex.cfm.
[8] http://www.notsoboringlife.com/list-of-hobbies/.
[9] http://discoverahobby.com/listofhobbies.
[10] http://a-z-animals.com/animals/.
[11] http://www.medicinenet.com/diseases_and_conditions/alpha_a.htm.
[12] http://www.cdc.gov/DiseasesConditions/az/a.html.

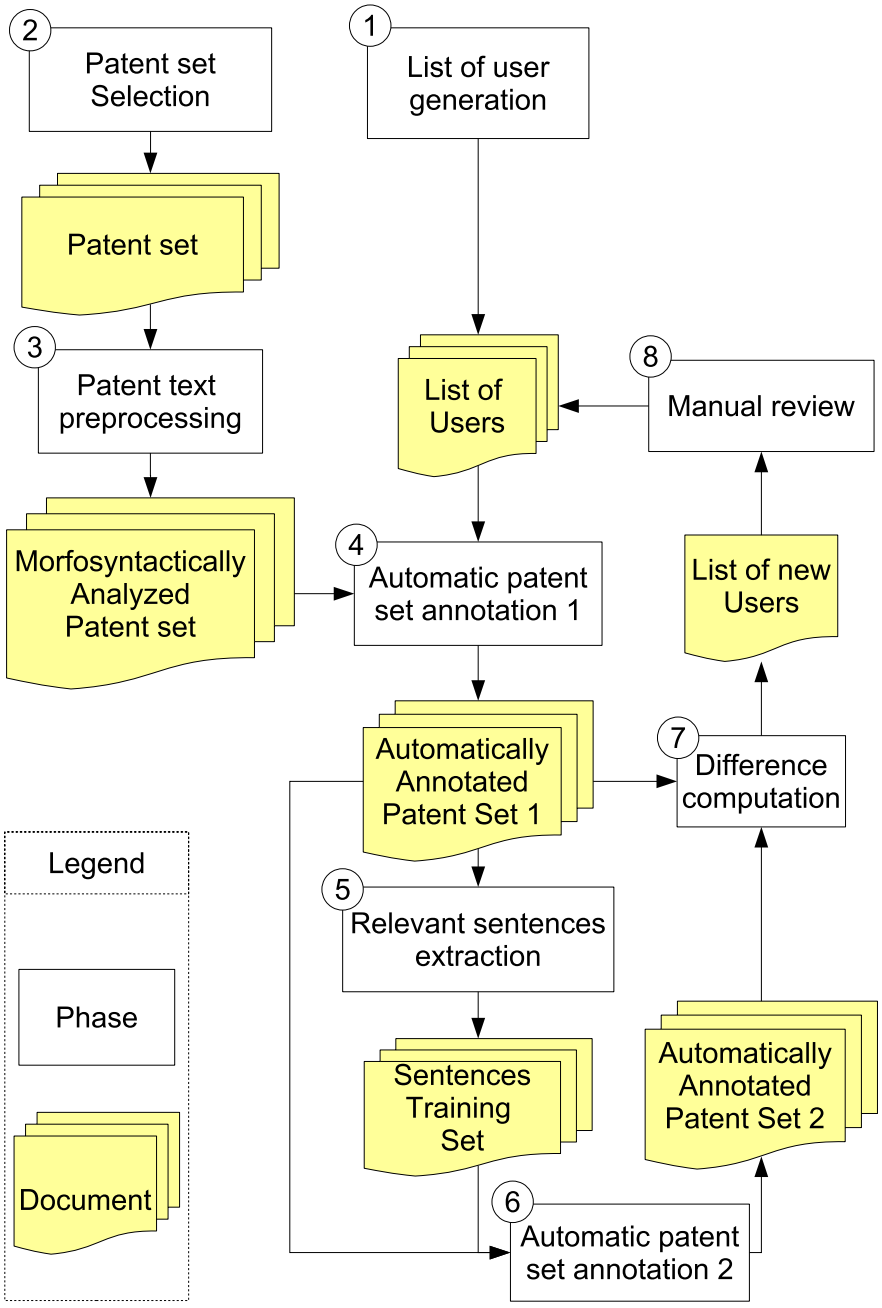

Fig. 1. Process flow diagram of the proposed automatic user extraction system from patents. The diagram contains the representation of the documents and the operations performed on them. The process takes in input a patent set and a list of users and produces a list of new users as output.

we took into consideration the International Patent Classification (IPC).[13] IPC is a hierarchical system of patent classes representing different areas of technology. Then, we wondered which classes could contain new users according to our seed list. Furthermore, IPC class A, which is the first level in IPC differentiation, is based on human necessities. For this reason, we assumed that in this class we would have found likely users from patents texts. Following these guidelines we chose two patent sets to be analyzed as described in section 4.1.

[13] http://www.wipo.int/classifications/ipc/en/.

3.3. Patent text analysis

Our Entity Extraction system is composed by a set of sequential phases. The first three phases are related to the linguistic annotation: sentence splitting and tokenization, part of speech tagging and lemmatization. Then, the patent set is analyzed by the entity extractor, specialized for users extraction. A more detailed description of each phase is given in the following sections.

3.3.1. Sentence splitting and tokenization

These processes split the text into sentences and then segment each sentence in orthographic units called tokens. In our system, sentence splitting plays a key role since thanks to a given word, it is possible to find sentences where the word is used. Finding correct boundaries for a

specific word allows to dramatically reduce the space to retrieve its surrounding contexts.

3.3.2. POS tagging and lemmatization

The Part-Of-Speech tagging (or POS tagging) is the process of assigning unambiguous grammatical categories to words in context. It plays a key role in NLP and in many language technology systems. For the present application we used the most recent version of the Felice-POS-tagger described in Ref. [42]. Once the computation of the POS-tagged text is completed, the text is lemmatized according to the result of this analysis.

3.3.3. Semi-automatic users annotation

The Users Extraction tool is based on supervised methods. Such methods require an entity annotated corpus in order to extract new entities from unseen documents. A semi-automatic method has been used to generate an annotated corpus of users to avoid manual annotation of a patent set. The method is a projection of the list of users on the patent set defined in section 3.2. The list of users described in section 3.1 is cleaned to avoid linguistic ambiguities when projecting these entities on the corpus. For example, the term "*guide*" has two different meanings when used as a verb or as a noun. Furthermore, as a noun it could indicate a component of a system (guide for mechanical parts) or a person (someone employed to conduct others) and therefore a user. Avoiding ambiguities is a crucial aspect to produce an informative training set, so ambiguous words were pruned.

The entity annotation schema for a single token is defined using a widely accepted BIO annotation scheme [43]:

- **B-USE**: the token is the beginning of an entity representing an User;
- **I-USE**: the token is the continuation of a sequence of tokens representing an User;
- **O**: for all the other cases.

3.3.4. User entity extraction

The Users extraction problem is tackled by the implementation of a supervised classifier that is trained on an annotated patent set. Thus, the patent set is linguistically-annotated, using the steps described above and entity-annotated, exploiting the semiautomatic annotation process executed in the previous steps.

Given a set of features the classifier trains a statistical model using the feature statistics extracted from the corpus. For each new document the trained model assigns to each word the probability of belonging to one of the classes previously defined (B-USE, I-USE, O).

In our experiments the classifier has been trained using two different learning algorithms: Support Vector Machines (SVM) using the LIBSVM library [44] configured to use a linear kernel and Multi Layer Perceptron (MLP) implemented using the Keras library [45]. It has been proven that LSTM methods are well suited for similar NER task. Anyway, we chose SVM and MLP method to study how two wheel established state of the art classifiers perform on the specific task of user extraction from patents and to evaluate their performance in terms of precision and computational effort. We also think that the popularity of these methods increment the reproducibility of the work.

The classifier uses different kind of features extracted from the text:

- *linguistic features*, i.e. lemma, Part-Of-Speech, prefix and suffix of the analyzed token;
- *contextual features*, the linguistic characteristics of the context words of the analyzed token; in addition the entity category of the previous token is considered;
- *compositional features*, combinations of contextual features and linguistic features. i.e. Part-Of-Speech of the previous word and the lemma of the current word. These extra features allow to infer statistics on the interaction of the combined features that can not be captured by a linear SVM model.
- **word2vec features**: vector representations of words computed by the *word2vec* [46] tool.

Word2vec is a NLP tool able to produce word representations exploiting big corpora. The main property of the vectors produced by *word2vec* is that words sharing similar contexts have similar vector representations. By using word vectors instead of the corresponding words we were able overcome the problem of the limited lexical knowledge in the training phase. Using these features and excluding all the others (delexicalized model) we expected that the resulting user extraction system had a lower precision and an higher recall in the classification phase. We presumed to find new users not contained in the input seed list.

3.4. Manual review of the new list of users

It is still possible that the classification process creates false positive results (words labeled as users that do not match the definition in section 2.1). Thus, it is necessary to make a manual review of the extracted entities with the aim of evaluating the output.

4. Results: case study

The following section describes the performances of the automatic users extraction process on two different patent sets, which are described in section 4.1. To test the system four experiments were conducted, as described in section 4.2. Finally, in 4.3 the performances and the outcomes of the system are shown and discussed.

4.1. Patent set

Following the guidelines for the patent set selection described in section 3.2, we examined two patent sets belonging to the IPC class A:

- **A47G33.** The IPC definition of the subclass is "*religious or ritual equipment in dwelling or for general*".
- **A61G1-A61G13.** The IPC definition of the subclass A61G1 is "*Stretchers*" while the definition of the subclass A61G13 is "*Operating tables; Auxiliary appliances therefor*".

We extracted from the private Errequadro s. r.l.[14] database a random sample of 2.000 patents from each IPC class. For each patent set we applied the semiautomatic set annotation process by projecting the input list of users on the morphosyntactically analyzed patent set. After this process, each semi-automatically annotated patent set was split in two parts: the first was used as training set for the user extractor, and the second one was used as test set.

To build an informative training set, from the semi-automatically patent set we selected a subset of sentences containing at least one user. The size of the training set in both cases is approximately composed by 600.000 tokens. For each patent set Table 1 shows the number of sentences of the training set, the number of sentences of the test set, and the number of distinct users in the training set (re-projected by the semi-automatic annotation process).

4.2. Experimental setup

We chose two orders of magnitude for the sentences test-set to test the efficiency of multiple configurations of the system.

To test the performances of the implemented user extractor, we devised four different configurations. Each configuration uses a specific learning algorithm and a set of features to build the statistical model. The main purpose of this procedure is to find the configurations that

[14] http://www.errequadrosrl.com/.

Table 1
Statistics related to the patent set groups analyzed in the case study.

Patent set group	#Sentences - training	#Sentences - test	#Distinct users projected on training
A47G33	13.364	214.029	126
A61G1-A61G13	15.108	2.520.350	121

better perform in the user extraction task. In addition, the different behaviour of the system in the classification phase is studied. In Tables 2 and 3 are reported the detailed configurations used in our experiments.

By using the first and the second configuration we expected to have a higher precision in the classification phase, since explicit lexical information is used in the training phase. For the same reason we expected to have low recall in classification phase. On the other hand, the third and fourth configurations are delexicalized: lexical information is provided by word vectors computed by *word2vec*. In these two configurations we expected to have an higher recall and a lower precision, due to the characteristics of the computed vectors explained in section 3.3.4. To limit errors when using the *word2vec* features, some linguistically motivated filtering rules were introduced. Specifically, sequences of tokens classified as users were constrained from the following categories: verbs, adjectives not preceded by articles, articles and adverbs.

4.3. Output of the experiments and measurements

To evaluate the whole user extraction process in each experiment, we defined some evaluation measures. Each measure was introduced to evaluate the characteristics of the extraction system concerning the configuration applied.

These measures are:

- Training time: time needed to create the statistical model using the training set;
- Test time: time needed to re-annotate the semi-automatically annotated patent set;
- Number of extracted users: number of unique entities classified as user in the automatically annotated patent set;
- Number of known users: number of distinct extracted users in the automatically annotated patent set and belonging to the list of user in input;
- Number of new users: number of distinct entities classified as user in the automatically annotated patent set and not belonging to the input list of users;
- Number of new correct users: number of distinct entities considered as user and as correct after a manual review;
- Precision: ratio between the number of new distinct correct users and the total number of new distinct users;

Table 3
Context windows of the extracted features considering 0 as the current analyzed token.

Feature group	Context Window
Lemma unigrams	[−2, −1,0,1]
Lemma bigrams	[(−1,0), (0,1)]
Word bigrams	[(−1,0), (−2, −1), (0,1), (1,2)]
Word trigrams	[(1,0,1)(−2,1,0)]
Pos unigrams	[−2, −1,0,1]
Pos bigrams	([(−2, −1)(−1,0), (0,1)])
Compositional feature#1	(POS_{-1}, $Lemma_0$)
Compositional feature#2	($Lemma_{-1}$, $Lemma_0$)
Compositional feature#3	($Lemma_0$, $Lemma_1$)
Compositional feature#4	(POS_0, $Lemma_1$)
Compositional feature#5	(NER_{-1}, $Lemma_0$)
Word2vec	[-2, −1, 0, 1, 2]

- Gain: ratio between the number of new distinct correct users and the number of re-projected distinct users on the training set.

Table 4 reports the values of the defined metrics across all the experiments run on the two patent sets.

For what concerns training and test time of the automatic patent set annotation, it's clear that the configuration based on the SVM learning algorithm without the *word2vec* features performs better in both the experiments (1, 5). When the features based on *word2vec* are introduced, the configuration based on the MLP learning algorithm is the fastest both in training and test time (3, 6): it is due to the fact that keras [45] implementation of this algorithm exploits all the available CPU cores of the system. On the other side, the MLP algorithm does not scale properly with a higher number of features, as seen in training and annotation time in the experiment (2). In addition, we could not perform the patent set annotation in the experiment (6), since it would have required more than 60 machine days to complete the process. When *word2vec* features are introduced, the patent set annotation based on the SVM algorithm is 10 times slower than the MLP algorithm.

For what concerns the precision in the automatic patent set annotation, the SVM configuration without *word2vec* features is clearly the more reliable: the precision values are from 1.5 to 2 times higher in the experiments (1, 5) in contrast to the other experiments. The higher precision is justified by the fact that the configurations based on *word2vec* features lack explicit lexical information: words with very similar contexts are represented by similar *word2vec* vectors, probably leading to errors in the classification phase. On the other hand, the use of *word2vec* vectors aims at extracting entities that would not be extracted by considering explicit lexical information only.

Finally, for what concerns information gain, the same amount of new information (21–37%) is extracted in the experiments on the A47G33 patent set. The gain values drastically change in the experiments on the A61G1-A61G13 patent set: in the experiments (5, 8) a

Table 2
Configurations used in our experiments.

Feature group	Configuration 1	Configuration 2	Configuration 3	Configuration 4
Lemma unigrams	✓	✓	✗	✗
Lemma bigrams	✓	✓	✗	✗
Word bigrams	✓	✓	✗	✗
Word trigrams	✓	✓	✗	✗
POS unigrams	✓	✓	✗	✗
POS bigrams	✓	✓	✗	✗
Compositional features	✓	✓	✗	✗
NER Previous Token	✓	✓	✓	✓
Word2vec	✗	✗	✓	✓
Learning algorithm	SVM	MLP	SVM	MLP
Linguistic filtering rules	✗	✗	✓	✓

Table 4
Comparison of the values of the defined metrics across all the experiments. The patent set annotation in the experiment (6) was not performed due to the computational costs. All the experiments were run on a machine provided with 10 AMD Opteron (tm) 6376 processors.

Experiment	Training time	Test Time	Extracted	Known	New	New correct	New wrong	Prec. (%)	Gain (%)
A47G33									
1 (SVM)	83 m	321 m	161	93	68	47	21	69.11	37.30
2 (MLP)	1911 m	9091 m	196	55	141	27	114	19.15	21.42
3 (MLP-W2V)	165 m	246 m	162	35	127	45	82	35.43	35.71
4 (SVM-W2V)	1265 m	4310 m	121	29	92	45	47	48.91	35.71
A61G1-A61G13									
5 (SVM)	148 m	3443 m	302	120	182	88	108	48.35	72.72
6 (MLP)	1818 m	–	–	–	–	–	–	–	–
7 (MLP-W2V)	333 m	3530 m	305	38	267	44	230	16.48	36.36
8 (SVM-W2V)	1268 m	47020 m	313	49	264	74	197	28.03	61.15

gain between 61% and 72% is obtained: it is due to the size of this patent set in comparison to the A47G33 one. In the experiment (7), despite the introduction of *word2vec* features, a gain of 36% is obtained. This fact, in conjunction with the non-feasibility of the experimental configuration 6, shows how MLP systems lack in efficacy and efficiency (in entity extraction in patent domain) when the test-set has an order of magnitude of millions of sentences. We think that this result is relevant, based on our experience with practical applications.

4.3.1. Aggregate results of automatic patent set annotation processes

In the previous section we have shown that each configuration of the patent set annotator contributes to the extraction of users that do not belong to the input user list: the result of this process is an informative gain. A way to maximize the overall informative gain is to merge the results of all manually reviewed user extractions obtained by executing the patent set annotation process with all possible configurations.

The overall informative gain of the merging process is related to intersections that occur among the results obtained by the patent set annotation process in each configuration: the less the intersections, the more the overall informative gain obtained. In Table 5 is shown the overall gain obtained by merging results of the manually reviewed extractions in each patent set.

The table shows that the merging process of manually reviewed entities extracted from each patent set annotation run effectively contributes to increase the overall informative gain. For instance in the A47G33 patent set an overall gain of 103.17% is obtained, tripling the best result achieved by the extraction performed using the best single configuration. Good results are also achieved in the A47G33 patent set user extraction. In this case an overall gain of 140.49% is obtained, doubling the best result achieved by the extraction performed using the best single configuration.

Table 5
Gain obtained by merging correct entities extracted from each patent set annotation.

Configuration	A47G33 - Gain (%)	A61G1 + A61G11 - Gain (%)
SVM	37.30	72.72
MLP	21.42	–
MLP-W2V	35.71	36.36
SVM-W2V	35.71	61.15
SVM ∪ MLP	52.38	–
SVM ∪ MLP-W2V	69.84	126.44
SVM ∪ SVM-W2V	73.01	103.30
MLP ∪ MLP-W2V	55.55	–
MLP ∪ SVM-W2V	57.14	–
MLP-W2V ∪ SVM-W2V	59.52	76.30
SVM ∪ SVM-W2V ∪ MLP-W2V	90.47	140.49
SVM ∪ MLP ∪ MLP-W2V	82.53	–
SVM ∪ MLP ∪ SVM-W2V	85.71	–
MLP ∪ SVM-W2V ∪ MLP-W2V	77.77	–
SVM ∪ MLP ∪ SVM-W2V ∪ MLP-W2V	103.17	–

The results shown in section 5 prove that if the goal of the extraction is to reach the maximal recall, an ensemble method (combining the output of multiple classifier) over-performs every single classifier method. Anyway, the ensemble approach has clear efficiency issues, because the time of analysis will be the sum of every single approach time (in hypotheses of non-parallelization). This leads to a trade off between the speed of the system and the quality of the results, and whoever would use the presented system can decide to gain benefit in one or in another direction.

4.3.2. The extracted users

Tables 6 and 7 show an overview of extracted users randomly chosen from the A47G33 patent set (the only one in which were able to perform all experiments). Each table is divided in two blocks, representing the results of the extraction performed using a specific

Table 6
Extracted users from the A47G33 patent set using the SVM and DL configurations. New users extracted by the system are reported in bold.

SVM			MLP		
Lemma	Frequency	# Patents	Lemma	Frequency	# Patents
female	801	109	child	402	102
child	426	108	cleregy member	128	5
guy	156	17	patient	113	11
patient	115	11	man	50	26
parent	70	31	young	48	32
man	51	26	**angel**	29	23
merchant	50	6	dog	20	7
soon	46	29	artisan	12	12
engineer	45	45	**male/female**	12	4
adult	39	23	hockey player	7	1
young	35	24	**professional**	7	7
society	32	21	tennis player	7	4
angel	29	23	football player	6	3
fund raiser	27	4	**ghost**	5	3
priest	22	4	children	5	5
cheerleader	15	4	manager	5	5
fund-raiser	11	4	**spider**	5	5
athlete	10	9	**vandal**	5	1
ghost	5	5	**athlete**	4	3
adulterant	3	3	mother	4	2
jew	3	3	soccer player	4	3
maid	3	1	squirrel	3	2
tourist	3	3	**maid**	3	1
indian	2	2	**god**	3	2
beginner	1	1	**mariner**	3	3
christians	1	1	**male-female**	2	2
datum entry operator	1	1	**manufacturer**	2	2
expert	1	1	**jew**	1	1
jewish	1	1	**merchandizers**	1	1
marinaro	1	1	**parishioner**	1	1

Table 7
Extracted users from the A47G33 patent set using the SVM-W2V and MLP-W2V configurations. New users extracted by the system are reported in bold.

SVM-W2V			MLP-W2V		
Lemma	Frequency	# Patents	Lemma	Frequency	# Patents
child	152	68	clergy member	124	5
clergy member	124	5	**crowd**	36	3
man	50	26	basketball player	20	5
engineer	45	45	**him**	17	8
young	29	24	woman	16	8
choir	17	1	**saint**	14	2
infirm	13	8	**youth**	14	2
bride	9	4	**angel**	8	4
volunteer	8	6	**choir**	8	1
musician	6	6	musician	6	6
boy	3	1	**god**	5	1
children	3	3	children	3	3
girl	3	2	guy	3	3
creature	2	1	infant	3	3
deceased	2	1	priest	3	3
jewish	2	2	**bride**	2	2
person	2	2	**consumer**	2	2
mother	2	2	**everyone**	2	2
audience	1	1	**him/her**	2	2
boyfriend	1	1	**spectator**	2	2
derby member	1	1	farmer	2	1
gift giver	1	1	youngster	2	2
handicapped	1	1	**boyfriend**	1	1
jesus	1	1	**grandparent**	1	1
saint	1	1	**subject**	1	1
husband	1	1	clown	1	1
lady	1	1	husband	1	1
runner	1	1	runner	1	1
society	1	1	society	1	1
teenager	1	1	tennis player	1	1

configuration. For each extracted user is shown the corresponding lemma (the root form), the frequency (how many times that user appears in the whole corpus) and the total number of patents containing the user. Users not contained in the starting user list, are highlighted in bold.

The table shows that the system was able to extract characteristic users of the patent set. The results are in fact not unexpected for the IPC class under analysis: this is an evidence of the correct performances of the proposed system. In other words, the results presented in the table show that it is possible to train a NER systems able to extract sparse and valuable information. Such users are the ones that an expert would manually extract but the NER system does it with an enormous saving in terms of time and efforts. In the case of the patent set extracted from the IPC class A47G33 (whose definition in section 4.1 is "*religious or ritual equipment in dwelling or for general*") the system was able to extract users such as:

The table shows that the system was able to extract characteristic users of the patent set. The results are in fact not unexpected for the IPC class under analysis: this is an evidence of the correct performances of our system. In other words, the table shows that it is possible to train a NER systems able to extract rare and valuable information that are in line whit the ones that an expert would manually extract but with enormous saving in terms of time and efforts. In the case of the patent set extracted from the IPC class A47G33 (whose definition in section 4.1 is "*religious or ritual equipment in dwelling or for general*") the system was able to extract users such as:

- *angel*, *jew* or *christian*, belonging to the religious groups;
- *datum entry operator* and *fund-raiser* missing from the list of jobs;
- *cartridge-player* and *choir* missing from the list of sports and hobbies;
- *infirm* missing from the list of patients.
- *him/her* and *male/female* that are combination of users

Other remarkable results are:

- many newly extracted entities have very low frequency in the patent set: it shows that the developed system is able to extract rare entities.
- Table 7 shows that configurations using *word2vec* features are able to find new users with a higher frequency in the patent set: it was an expected result, since the *word2vec* configurations are not explicitly lexicalized and more able to generalize during extraction phase.
- The system is able to extract single words and multi-words.
- Taking into consideration the definition we have given in section 2.1, the system extracts unusual and sometimes borderline users. Examples like *saint*, *angel*, *god* and *ghost* need discussion that is far beyond the purposes of the present paper. These results are a remarkable evidence of the human-like generalization ability of the described method.

4.3.3. Users metrics

In this section we present a preliminary analysis of the user entities. The total number of users, extracted by the different systems from the process described in section 3, is 109. 28,2% (564 on 2.000) of patents in analysis contains at least one user. This result is an evidence of the fact that patents actually contain users information, and, considering the approach we followed, this percentage is an accurate lower approximation of the actual percentage of patents containing at least one user.

In Fig. 2 for each user on the x axes is shown the number of patents in which the user is contained. The distribution is skewed, with some occurrences showing large numbers and many others with just one or few occurrences. It is clear that there is a Pareto like distribution, with the first 20% of users covering 70% of total users in terms of occurrence. It means that some users are more likely to be cited in patents and many more users that rarely appear. Following this observations,

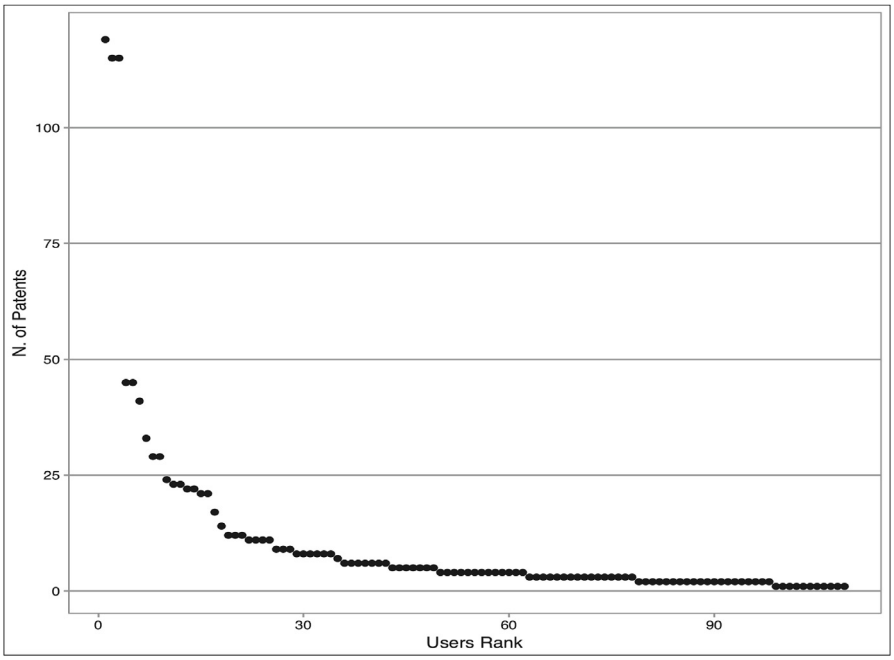

Fig. 2. Distribution of the number of patents per user.

we can divide users in three groups:

- *Group A*: users that appear in more than 100 patents (5% of the patent set). In our case these are *male*, *child* and *female*.
- *Group B*: users that appear in more than 20 patents (1% of the patent set). This group is composed by 13 different users. Some of these are *engineer*, *person*, *player*, *adult*, *angel* and *guy*.
- *Group C*: users that appear in less then 20 patents. This group is composed by 93 different users. Some of these are *mother*, *athlete*, *priest*, *adulterant*, *golfer* and *hockey player*.

Further research means to study how these users differ from patent set to patent set. We expect to see similar distribution but with different content of users. Frequent and non-specific users comprise Group A: in other patent set we could see differences in terms of entities contained in this class but its content will stay non-specific. These results seem to be generic social roles indicating the gender or the age of a person. Group B is composed of mainly non-specific users and some specific users that change from patent set to patent set. This class helps to identify the core users of the patent set. Lastly, Group C contains non-frequent users that are both specific and non-specific, making it the most interesting of the three for the purposes of our work. In this group we find users that are market niches, so the patent that contains these users is of great interest for marketers and designers. These are both samples of the more generic users (for example a *mother* is a *female* and a *hockey player* is a *player*) or specific users of the patent-set (like *priest*, fund-*raiser*, *doll*, *spouse* and *clergy member*.

5. Conclusions

This paper has proposed a method to automatically extract users from patents. Starting from an input seed list of users, the proposed system is able to extract new users from a patent set of interest. The extraction process is composed of several steps which do not require human effort. Human intervention is required only in the final review of extracted new users. The approach we used is based on an extraction pipeline of Natural Language Processing tools that provides linguistic features for the classification process. Support Vector Machines and Multi Layer Perceptron have been exploited as learning algorithms for the system. Furthermore, lexical and non-lexical features have been adopted as input for learning algorithms. We obtained four different configurations that we compared. To reduce the limited lexical knowledge problem, we used as features word vectors generated by *word2vec* from a big corpora of patents. The proposed approach was tested against two different patent sets and the obtained results have shown the effectiveness of the extraction process. Experiments demonstrated that delexicalized models perform better in terms of recall at the expense of precision. On the other side, lexicalized models were more stable for what concerns the precision of the system, while the information gain was reduced. Based on these results, we tested an ensemble method which outperforms all the other single systems. In addition, the conducted experiments showed that the system is able to extract low frequency single and multi-words related to users. Finally, some preliminary analysis on the extracted users were conducted showing that there is a difference between generic and specific users in terms of patents citing these entities.

As further development, we want to extract new entities of interest for designers and marketers that could be connected with the users of the invention, giving a wider and more valuable representation of the inventions described in patents. Once this goal is achieved, the method we proposed in the present work could help to answer to further research questions and hypothesis, such as:

1. *Patents can be exploited to assist user driven innovation*: a deepened analysis of patents that contain specific users, could help designers to make new products for specific market niches.
2. *Large companies grow by adding new categories of users over time*: successful companies expand their product range over time by integrating new technologies and/or by covering new markets. By following the development over time of the number of users it will be possible to observe these dynamics. Furthermore we want to apply forecasting techniques on the information we will extract, since patent trends are a great source for forecasting [47].
3. *New entrants enter the industry by targeting highly specific categories of users (niches)*: New entrants may develop innovations that target niche categories, instead of large categories. The relevant literature from population ecology, particularly the contributions on niche creation, might be used to examine this issue.

4. *Novel patents include novel categories of users*: It will be possible to examine the issue of the emergence of novelty in patents by defining a new dimension of novelty. i.e. user novelty. Novel categories of users may constitute a separate dimension of novelty, in addition to the technological one (new combination of IPC classes).

Needless to say, the construction of ensemble methods that puts together engineering, economics and computational linguistics will inevitably iterative process. However, the prospect of building up a platform for re-assessing many of the most interesting (and often unsettled) issues in economics and management of innovation, suggests that the effort is worthwhile.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.wpi.2018.07.006.

References

[1] H. Ernst, Patent information for strategic technology management, World Patent Information 25 (3) (2003) 233–242.
[2] G. Jin, Y. Jeong, B. Yoon, Technology-driven roadmaps for identifying new product/market opportunities: use of text mining and quality function deployment, Adv. Eng. Inf. 29 (1) (2015) 126–138.
[3] D. Bonino, A. Ciaramella, F. Corno, Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics, World Patent Inf. 32 (1) (2010) 30–38.
[4] I. P. Rights, Quick Scan, a novelty Search Service in the Framework of Eu-&d Programmes.
[5] P. Terragno, Patent as technical literature, IEEE Trans Prof Commun PC 22 (2) (1979) 101–104 cited By 2.
[6] D. Golzio, Wwwwwhow Read a Patent!, (2012).
[7] I. Bergmann, D. Butzke, L. Walter, J. Fuerste, M. Moehrle, V. Erdmann, Evaluating the risk of patent infringement by means of semantic patent analysis: the case of dna chips, R D Manag. 38 (5) (2008) 550–562.
[8] Y. Liang, R. Tan, Trends in computer aided innovation, Second IFIP Working Conference on Computer Aided Innovation, October 8–9 2007, Michigan, USA, Springer US, Boston, MA, 2007, pp. 89–96 Ch. A Text-Mining-based Patent Analysis in Product Innovative Process.
[9] P. Burke, M. Reitzig, Measuring patent assessment quality-analyzing the degree and kind of (in)consistency in patent offices' decision making, Res. Pol. 36 (9) (2007) 1404–1430, https://doi.org/10.1016/j.respol.2007.06.003.
[10] M. Philipp, Patent filing and searching: is deflation in quality the inevitable consequence of hyperinflation in quantity? World Patent Inf. 28 (2) (2006) 117–121.
[11] G. Fantoni, R. Apreda, F. Dell'Orletta, M. Monge, Automatic extraction of function-behaviour-state information from patents, Adv. Eng. Inf. 27 (3) (2013) 317–334.
[12] N. Chinchor, P. Robinson, Muc-7 named entity task definition, Proceedings of the 7th Conference on Message Understanding, 1997, p. 29.
[13] J. Piskorski, R. Yangarber, Information extraction: past, present and future, Multi-source, Multilingual Information Extraction and Summarization, Springer, 2013, pp. 23–49.
[14] D. Gildea, Corpus variation and parser performance, Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 2001, pp. 167–202.
[15] J. Wang, W. Lu, H. Loh, A two-level parser for patent claim parsing, Adv. Eng. Inf. 29 (3) (2015) 431–439.
[16] K. Idris, Wipo Intellectual Property Handbook: Policy, Law and Use, Geneva: WIPO publication (489).
[17] F. Bonaccorsi, Chiarello D'amico, Mapping users in patents. towards a new methodology and the definition of a research agenda, EPIP 2017 Conference BORDEAUX, 2017.
[18] C. Beller, R. Knowles, C. Harman, S. Bergsma, M. Mitchell, B. Van Durme, I'm a belieber: social roles via self-identification and conceptual attributes, ACL (2), 2014, pp. 181–186.
[19] C. Beller, C. Harman, B. Van Durme, Predicting fine-grained social roles with selectional preferences, ACLPPinforma (2014) 50.
[20] P. Johnson, 2 hrm in Changing Organizational Contexts, Strategic HRM, (2009), p. 19.
[21] Check Suffix Codes for Jobs Defined in the Dictionary of Occupational Titles, third ed., United States Employment Service, U.S. Dept. of Labor.
[22] I. ISO, 13407: Human-centred Design Processes for Interactive Systems, Geneva: ISO.
[23] U. K, Users, Experts, and Institutions in Design, Handbook of New Product Development Management, (2008), p. 421438.
[24] A. Blanchard, Understanding and customizing stopword lists for enhanced patent mapping, World Patent Inf. 29 (4) (2007) 308–316.
[25] B. Chiarello, Fantoni, Product description in terms of advantages and drawbacks: exploiting patent information in novel ways, Proceedings of the 21st International Conference on Engineering Design, 2017.
[26] S. Lee, B. Yoon, Y. Park, An approach to discovering new technology opportunities: keyword-based patent map approach, Technovation 29 (67) (2009) 481–497.
[27] C. Lee, B. Kang, J. Shin, Novelty-focused patent mapping for technology opportunity analysis, Technol. Forecast. Soc. Change 90 (2015) 355–365.
[28] T. Montecchi, D. Russo, Y. Liu, Searching in cooperative patent classification: comparison between keyword and concept-based search, Adv. Eng. Inf. 27 (3) (2013) 335–345.
[29] J. Yoon, H. Park, K. Kim, Identifying technological competition trends for r&d planning using dynamic patent maps: sao-based content analysis, Scientometrics 94 (1) (2013) 313–331.
[30] H. Park, J. Yoon, K. Kim, Identifying patent infringement using sao based semantic technological similarities, Scientometrics 90 (2) (2011) 515–529.
[31] H. Park, K. Kim, S. Choi, J. Yoon, A patent intelligence system for strategic technology planning, Expert Syst. Appl. 40 (7) (2013) 2373–2390.
[32] S.R. Eddy, Hidden markov models, Curr. Opin. Struct. Biol. 6 (3) (1996) 361–365.
[33] J. Lafferty, A. McCallum, F. C. Pereira, Conditional Random fields: Probabilistic Models for Segmenting and Labeling Sequence Data.
[34] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkopf, Support vector machines, intelligent systems and their applications, IEEE ASME J. Microelectromech. Syst. 13 (4) (1998) 18–28.
[35] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, arXiv preprint arXiv:1603.01360.
[36] X. Ma, E.H. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, vol. 1, 2016 Long Papers http://aclweb.org/anthology/P/P16/P16-1101.pdf.
[37] C. Lee, LSTM-CRF models for named entity recognition, IEICE Trans. 100-D (4) (2017) 882–887 http://search.ieice.org/bin/summary.php?id=e100-d_4_882.
[38] S. Misawa, M. Taniguchi, Y. Miura, T. Ohkuma, Character-based bidirectional LSTM-CRF with words and characters for Japanese named entity recognition, Proceedings of the First Workshop on Subword and Character Level Models in NLP, Copenhagen, Denmark, September 7, 2017, 2017, pp. 97–102 URL https://aclanthology.info/papers/W17-4114/w17-4114.
[39] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, Chemdner: the drugs and chemical names extraction challenge, J. Cheminf. 7 (S-1) (2015) S1.
[40] T.-S. LEE, S.-S. KANG, Optimizing the features of crf-based named entity recognition for patent documents, Technology 18 2–56.
[41] R. Leaman, C.-H. Wei, Z. Lu, tmchem: a high performance approach for chemical named entity recognition and normalization, J. Cheminf. 7 (S-1) (2015) S3.
[42] F. DellOrletta, Ensemble system for part-of-speech tagging, Proceedings of EVALITA.
[43] L.A. Ramshaw, M.P. Marcus, Text chunking using transformation-based learning, Natural Language Processing Using Very Large Corpora, Springer, 1999, pp. 157–176.
[44] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[45] F. Chollet, Keras, https://github.com/fchollet/keras (2015).
[46] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv preprint arXiv:1301.3781.
[47] R.S. Campbell, Patent trends as a technological forecasting tool, World Patent Inf. 5 (3) (1983) 137–143.