Expert Systems With Applications 184 (2021) 115570

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A simple and fast method for Named Entity context extraction from patents

Giovanni Puccetti [a,*], Filippo Chiarello [b], Gualtiero Fantoni [c]

[a] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
[b] Department of Energy, Systems, Territory, and Construction Engineering, Largo Lucio Lazzarino 2, 56122 Pisa, Italy
[c] Department of Civil and Industrial Engineering Largo Lucio Lazzarino 2, 56122 Pisa, Italy

ARTICLE INFO

Keywords:
Natural Language Processing
Information retrieval
Patents

ABSTRACT

The process of extracting relevant technical information from patents or technical literature is as valuable as it is challenging. It deals with highly relevant information extraction from a corpus of documents with particular structure, and a mix of technical and legal jargon. Patents are the wider free source of technical information where homogeneous entities can be found. From a technical perspective the approaches refer to Named Entity Recognition (NER) and make use of Machine Learning techniques for Natural Language Processing (NLP). However, due to the large amount of data, to the complexity of the lexicon, the peculiarity of the structure and the scarcity of the examples to be used to feed the machine learning system, new approaches should be studied.

NER methods are increasing their performances in many contexts, but a gap still exists when dealing with technical documentation. The aim of this work is to create an automatic training sets for NER systems by exploiting the nature and structure of patents, an open and massive source of technical documentation. In particular, we focus on collecting the context where *users* of the invention appear within patents. We then measure to which extent we achieve our goal and discuss how much our method is generalizable to other entities and documents.

1. Introduction

Deep Learning based systems for Named Entity Recognition (NER) are achieving high performances on recognizing various entities (e.g. persons, organizations, geographical entities) from different domains (e.g. medicine, chemistry, history) and types of documents (e.g. tweets, Wikipedia, scientific literature) (Yadav & Bethard, 2018). A strong increase in the accuracy of NER systems has been achieved thanks to the recent development in contextual word embedding, that has proven to effectively recognize entities at the edges of distributions, like polysemous word (i.e. terms having different tags in different contexts) (Tenney, Das, & Pavlick, 2019). Indeed, the automatic analysis of text for information extraction, is shifting towards deep learning based methods (Bengio, Ducharme, Vincent, & Janvin, 2003).

While extremely successful in the analysis of many general purpose sources (e.g. reviews, tweets, Wikipedia) (Romero & Becker, 2019) and entities (e.g. personal names, dates, locations, events), these systems are weaker on domain specific tasks, preventing many potential users (both from academia and industry) to fully take advantage of the huge progress of Natural Language Processing (Chiarello, Cimino, Fantoni and Dell'Orletta, 2018).

This gap in the literature, is due to the fact that standard NER tasks (in terms of domain, documents and entities) rely on mature methodologies based on manually annotated data-sets (Konkol, Brychcín, & Konopík, 2015; Matin, Hansen, Hansen, & Mlgaard, 2019; Romero & Becker, 2019).

Unfortunately, it is not as easy to create annotated training sets for technical documents and entities, mainly for three reasons. First, domain specific entities are costly to tag, because they are rare with respect to more generic words (Chiarello, Fantoni, Bonaccorsi, et al., 2017); second, manually tagged data-set of technical documents have a big business value, thus they are not open-sourced by researchers or companies (Blanco-Fernández et al., 2020); third, manual tagging for entity specific tasks requires the time of experts in the chosen domain, and experts are known to have limited time (Chiarello, Trivelli, Bonaccorsi and Fantoni, 2018).

For these reasons, in the present paper we aim at using the state-of-the-art in Natural Language Processing in order to create a domain specific (in terms of entity and document type) NER system, without relying on manual annotated data. In particular, we propose (1) a novel method based on a knowledge based approach in order to create the training set for a domain specific entity and (2) a comparison

* Corresponding author.
E-mail addresses: giovanni.puccetti@sns.it (G. Puccetti), filippo.chiarello@unipi.it (F. Chiarello), gualtiero.fantoni@unipi.it (G. Fantoni).

https://doi.org/10.1016/j.eswa.2021.115570
Received 25 May 2020; Received in revised form 4 June 2021; Accepted 4 July 2021
Available online 13 July 2021
0957-4174/© 2021 Elsevier Ltd. All rights reserved.

of different supervised classification methods (i.e. classification trees, support vector machines and neural networks) in order to identify domain specific entities in domain specific documents.

The choice of the different classification methods to compare is driven by the fact that recent literature shows that neural networks represent the most effective tool in this kind of tasks (Devlin, Chang, Lee, & Toutanova, 2019; Mikolov, Chen, Corrado, & Dean, 2013), particularly since they are able to encode linguistic features (and this is also supported by our results). On the other hand, classification trees provide highly readable outputs, allowing us to better understand if the predictions made by the algorithm have reasonable explanations or are driven by a biased choice of the data (Guidotti et al., 2018).

The proposed methods are tested on *patents* documents for the extraction of the entity *user of the invention*.

For what concerns the domain specific documents, the decision to select *patents* is driven by preliminary works showing challenging tasks with interesting results. Patents are challenging because knowledge is highly encoded and hard to extract through text mixing due to domain specific, technical and juridical jargon. This makes the task of information retrieval from patents very difficult and time consuming (Fantoni, Apreda, Dell'Orletta, & Monge, 2013). Despite these challenges, recent works show that text-based metrics improve the traditional metrics based on patent metadata (Arts, Hou, & Gomez, 2021), untapping their great informative potential. Patents are in fact an exclusive source of high value technical information: even if it is hard to estimate the scope of their exclusivity (Asche, 2017), a large share of technical knowledge is only present there.

On the other side, for what concerns the domain specific entity to extract, we choose *users of the invention*, driven by preliminary works that have shown results and open challenges for this task. Chiarello, Cimino et al. (2018) demonstrate the possibility to extract users from patents. However, they underline the difficulty in developing a precise measure of the recall of the system, that is only possible through human evaluation. In this regard our approach may be helpful since we focus on contexts rather than single entities, allowing for a deeper evaluation of the extractions.

Our work contributes to the expert system literature dealing with Natural Language Processing. In fact, the novelty of the proposed method does not rely on the classification algorithm (for which we test different state of the art alternatives) but on the construction of the knowledge base to train the algorithm (Mirończuk & Protasiewicz, 2018), in line with recent literature that is moving towards the use of knowledge bases for Natural Language Processing tasks in technical domains (Binkhonain & Zhao, 2019; Burggräf, Wagner, & Weißer, 2020).

The rest of the paper is structured as follows, in Section 2 we discuss the state of the art in tasks related to the one we are tackling. We then continue with Section 3 where we describe the methodology we apply. Afterward we report the results in Section 4 and then conclude with Section 5 where we outline the principal limitations and ways to extend our work.

2. Related work

In the present paper we propose a classification method in order to identify domain specific entities in domain specific documents without relying on manually labelled data. This task is related to what in the literature is known as Named Entity Recognition (NER). There is a large literature about NER systems that focuses both on the development of the classification models and on their applications. In the present session, after showing related works on the context of generic NER, we focus on entity extraction from specific documents and on the extraction of specific entities.

2.1. Named Entity Recognition (NER)

Named Entity Recognition has been a hot topic in the community of Artificial Intelligence and Computer Science. It consists in detecting lexical units in a word sequence that refer to a predefined entity, thus determining what kind of entity it is referring to (e.g. persons, locations, organizations.). The most successful NER systems focuses on meaningful text representation through word embeddings (Cer et al., 2018; Gildea, 2001; Liu, Li, Xiong, & Cavallucci, 2020; Piskorski & Yangarber, 2013). The methods used for NER are various:

- terminology-driven NER: aims to map mentions of entities within texts to terminological resources (e.g. wikipedia) (Yadav & Bethard, 2018);
- rule-based NER: uses lexicons, regular expressions and lexical information to express knowledge based systems able to extract a certain type of entity (Sari, Hassan, & Zamin, 2010);
- corpus-based NER: uses manually tagged text corpora (training set) to train machine learning (ML) algorithms (Lafferty, McCallum, & Pereira, 2001; McCallum, Freitag, & Pereira, 2000; Yadav & Bethard, 2018).

Information about the entity to which a word belongs, can provide crucial, although shallow, semantic information for tasks such as question answering (Abujabal, Saha Roy, Yahya, & Weikum, 2018; Blanco-Fernández et al., 2020), topic disambiguation (Fernández, Arias Fisteus, Sánchez, & López, 2012) or detection (Lo, Chiong, & Cornforth, 2017) and elements relationships identification (Sarica, Luo, & Wood, 2019).

As stated in Section 1, NER is a classification task and thus NER systems need a set of annotated documents in order to use state of the art approaches in terms of accuracy (corpus-based NER uses deep neural networks) (Devlin et al., 2019). Since the annotation task is resource intensive and labelled training sets are rarely open, there are only few training sets of entities that are available. This is especially relevant for deep learning systems which need an high quantity of labelled data. Despite this, generic NER systems have proven to be successful in different languages (Konkol et al., 2015; Küçük & Yazıcı, 2012) and for several non-technical texts (Jung, 2012). Furthermore, a recent work shows that it is possible to overcome another important limitation of NER systems, that is the fact that these systems rely on external NLP tools and hand-crafted features. These information sources are not always accurate for various languages and contexts, limiting the effectiveness of NER systems. In Bekoulis, Deleu, Demeester, and Develder (2018b) the authors solved this problem developing a new joint neural model for entity recognition and relation extraction. Specifically, the entity recognition task has been modelled using a Conditional Random Field layer and the relation extraction task as a multi-head selection problem. The approach is inspiring for the present work since it shows that NER systems can be effective in various contexts (i.e., news, biomedical, real estate) and languages (i.e., English, Dutch).

2.2. NER systems for domain specific documents

As stated before, NER systems based on Deep Learning algorithms have lower performances when applied to texts belonging to specific domains, this problem has been widely faced in the biomedical field (Silvestri, Gargiulo, & Ciampi, 2019). The focus, in this domain, has emerged for the automation of the analysis of medical documents, such as Electronic Health Records (EHR). These systems can strongly increase the efficiency of the work of physicians and researchers to let them dedicate more time to focus on their core activities. Similar effort has been put on related domains, such as chemistry, where chemical entities are searched (Krallinger et al., 2015; Leaman, Wei, & lu, 2015).

Such a strong focus on a specific domain is rare with the exception of documents such as patents (Chiarello et al., 2019; Liu et al., 2020;

Park, Kim, Choi, & Yoon, 2013; Sarica et al., 2019), whereas for other domains, task specific solutions are found, for example for real estate ads (Bekoulis, Deleu, Demeester, & Develder, 2018a). Considering this we can affirm that a large amount of named entities are still hidden inside domain specific documents.

2.3. NER systems for domain specific entities

Applying NER systems trained to extract a specific entity in a specific domain (e.g. names of persons from scientific papers) or to extract another kind of entity from a novel domain (e.g. names of cities from Twitter) yields to a certain failure (Ciaramita & Altun, 2005). This problem is increased due to the partial inconsistencies in the manual tagging procedure, especially in domains where manual tagging requires very specific expertise (i.e. the users of the inventions in patents, the one chosen for the present work). Thus, at the state of the art, NER systems could be considered effective only if the system is designed by experts in NLP and domain experts (Chiarello, Trivelli et al., 2018). However, NER has proven to be effective in broader applications, such as user profiling (Nicoletti, Schiaffino, & Godoy, 2013) and biological entity recognition (Atkinson & Bull, 2012).

3. Methodology

The proposed NER method aims at identifying a specific entity, the user of the invention, briefly referred to as *user* from a specific domain (i.e. patents) without relying on manually annotated data. The method follows two main steps:

- automatically creating a training set using a knowledge based approach;
- training different supervised classification methods in order to solve the specific NER task.

In particular three classification methods were tested: classification trees, support vector machines and neural networks.

3.1. Knowledge driven training set

There exist different approaches for training set creation (Zesch & Gurevych, 2006). In the present paper, we propose a novel solution based on the conjecture that instances of the target entity are contained in sentences that are linguistically similar to the target entity itself, and it is thus possible to use these sentences as a training set. More specifically, entities contextually similar to those containing the entity to extract provide examples to use as a training set.

We first argue why we made this hypothesis and in the rest of the paper we show experiments aimed at understanding to which extent it is true. Let us remark that our hypothesis is rather strong for general purpose texts, where polysemy is a recurrent phenomenon, studied also in the NLP community (Wang, Wang, & Fujita, 2020), and similar words can assume different meanings. However, this applies to patents in a different way. Indeed, a patent contains the disclosure of the invention in a manner sufficiently clear and complete in order to allow a person skilled in the art to reproduce the invention without any creative activity (Art 83 EPC) (Lidén & Setréus, 2011). This requirement, driven by patents' legally binding nature, assures the presence of recurring patterns that can be exploited by rules to identify specific entities.

Moreover, for the case of *users* we can further sustain our hypothesis. Let us report the definition of *user* provided by Chiarello, Cimino et al. (2018):

Definition 1. A *User* is an animated or previously animated entity (human or animal, alive or dead), on which the invention has a positive or negative effect at an unspecified moment.

The legal requirements imply that patents are technology centred documents. This fact and the definition of *user*, imply that, within a patent, all references to humans or to human characters identify users with high probability. In particular, those pronouns that can only refer to humans can be used as pointers to users.

Based on this theoretical analysis of patents' text, we proceed to test two knowledge based approaches for sentences identification, selecting generic taggers (1) personal pronouns and (2) the entity name itself. In our case the first class of *taggers* is composed by **anybody, anyone, everybody, everyone, he, she, somebody, someone**; for the second approach the *taggers* are the words *user* and *users*. We argue that pronouns are a better choice.

3.2. Training set validation

The first step towards the experiments consists of choosing metrics in order to test the quality of the proposed approaches for automatic training set definition. From the results we can also understand how to best refine the training data, before moving to the classification phase.

In order to validate the two approaches we make a statistical analysis of the sentences containing the *taggers* and the words *user* and *users*. At the same time we also attempt to identify possible cues implying less likely presence of a user. We assume that the word *device* has this property. We then make a statistical analysis of the verbs content in the two sets of sentences (the user-related and the device related) in order to identify which verbs most likely represent the actions of a user.

We proceed to a second step where statistical classifiers are employed. Two classifiers, classification trees and linear support vector machines are used to attempt and classify the sentences where our *taggers* are present, thinking that this will generalize to new users. We believe that understanding which, between the two different proxies (i.e. the *taggers* or the words *user* and *users*) perform better, and inspecting the most relevant features used by these classifiers, can allow us to choose for the best method.

3.3. Classification methods

In the recent NLP literature several attempts are being made to understand the encoding performed by the most effective language models (Devlin et al., 2019). In particular, the focus is on finding out if the vectors, fitted by the model to encode each sentence, carry a deeper meaning than plain vocabulary aggregation, being able to express linguistic content. The way this is most often done is via diagnostic classifiers (Conneau, Kruszewski, Lample, Barrault, & Baroni, 2018), used to inspect the vectors generated by the neural networks. Moreover, the fine tuning of neural networks seems to be able, in relatively short time compared to the long times needed to pre-train the network, to achieve good results in several tasks.

We fine tune the models developed in Cer et al. (2018) and in Devlin et al. (2019) to address our specific task, since they are able to encode the meaning of the whole sentence at once. This is crucial for the identification of a context-dependent entity such as users. In doing this we focus on two main aspects. The first is checking if the sentence encoding is able to extract the information of the user presence, keeping in mind that this task can be hard even for a person non expert in the field. The second is the fast applicability of this method, in particular we add a feed forward neural network on top of the encoding and then let the training propagate back to the whole network. We are able to perform deep analysis in short enough time for this to be applicable to real problems. This method also provides us with a work around for the problem of the different categories behaviour, since, retraining the classifier on a new set can be relatively fast.

3.4. Metrics and evaluation

Due to the limited number of attempts so far into the user extraction task, we will limit ourselves to the list of users collected by Chiarello, Cimino et al. (2018) to compute accuracy and related scores for our classifiers. In particular, we will use the following metrics when we compare with past work: true positives *tp*, true negatives *tn*, false positives *fp* and false negatives *fn* as follows:

$$tp = |\{\text{Sentences with a user}\} \cap \{\text{Positive predictions}\}|, \quad (1)$$

$$tn = |\{\text{Sentences without a user}\} \cap \{\text{Negative predictions}\}|, \quad (2)$$

$$fp = |\{\text{Sentences without a user}\} \cap \{\text{Positive predictions}\}|, \quad (3)$$

$$fn = |\{\text{Sentences with a user}\} \cap \{\text{Negative predictions}\}|. \quad (4)$$

Doing this allows us to keep the usual definitions for the scores:

$$Precision = \frac{tp}{tp + fp}, \quad (5)$$

$$Recall = \frac{tp}{tp + fn}, \quad (6)$$

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}, \quad (7)$$

$$F1 = 2 * \left(\frac{1}{Precision} + \frac{1}{Recall}\right)^{-1}. \quad (8)$$

With these definitions in mind we can compare the results we achieve with a baseline and validate our methodology, since the metrics we choose reflect the core hypothesis that the *taggers* can redirect us to the *users*.

4. Results

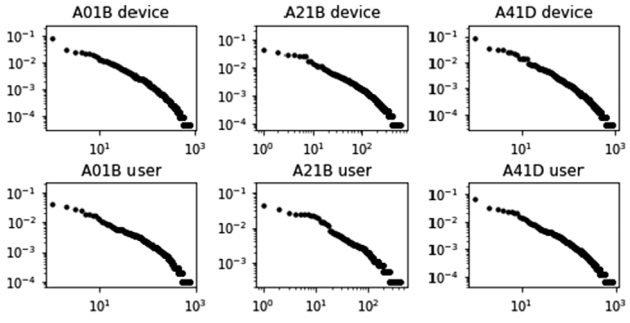
The data-set consists of patents belonging to 6 International Patent Classification categories (IPC). Given that different IPC have different technological focus, each of them has a different vocabulary. The application of our method on such a corpus, makes it possible to achieve higher generalization of the classifier. Moreover the IPC class, A, was chosen since it involves *Human Necessities* where the probability to find users is higher (e.g. with respect to Class C "Chemistry and Metallurgy" or H "Electricity") (Chiarello, Cimino et al., 2018). In Table 1 each searched IPC is described together with the number of patents from which the sentences were taken. These patents are identified according to whether they contained one of the taggers.

As discussed in Section 3, the analysis focuses on sentences. To argue that the data-set is consistent, the ratio of sentences including one of the pronouns has been measured, inspecting a set of 150,000 patents, considering each patent contains between 250 and 300 sentences. The most occurring pronoun is *he*, which appears in approximately 13 000 sentences, far above any of the other pronouns, while on the other side we have *everybody* which appears 91 times. Although quite different in magnitude, both these occurrence rates are low (*he* appears only in less than 1% of the sentences). This is an evidence that pronouns are a good proxy for classifying users, since they are as rare as the entity we want to classify (Chiarello, Cimino et al., 2018). To have enough data to build the classifier, we need to collect enough patents to cope with the low occurrence of pronouns. Particularly, for the statistical classifications a set of 192,000 sentences of at most 50 words was used, from the mentioned *IPC* categories. The ratio of positive and negative cases is set to 50%: in other terms, in the training set half sentences hold one of the pronouns. The ratio is decided following a common choice in the literature (Mitchell, 1997).

All the results are obtained using the python packages spacy (Honnibal, Montani, Van Landeghem, & Boyd, 2020), scikit-learn (Pedregosa et al., 2011) and tensorflow (Abadi et al., 2015).

Table 1
Number of patents in each IPC class.

IPC category	Description	#Patents
A01B	Soil working in agriculture	1109
A21B	Bakers' ovens	216
A41D	Outerwear	3854
A42B	Hats; head coverings	1344
A61B	Diagnosis; surgery	8221
A62C	Fire fighting	829
A63B	Apparatus for physical training	7703



	A01B		A21B		A41B	
	α	$x_{\min}$	α	$x_{\min}$	α	$x_{\min}$
device	2.082	0.002	2.149	0.002	1.941	0.001
user	2.440	0.003	2.277	0.002	2.240	0.003

Fig. 1. Log–log plots of word frequency for the words *user* and *device* for each patents' category, below estimate parameter of the power laws that would best fit the frequencies data.

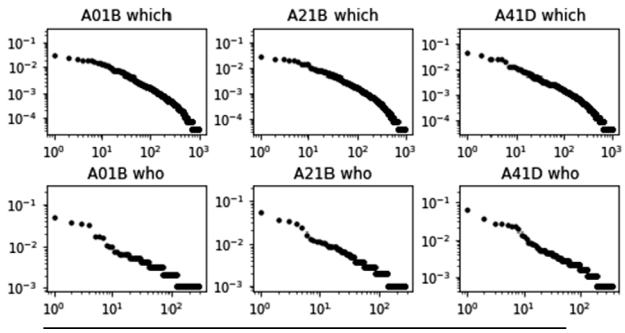
4.1. Data

Targets such as Chemical entities, Product names and dates (Tsai et al., 2006), have already been studied also in the context of patents, particularly with NER methodologies. Users are different from this kind of entities, since they are not identified in any way other than their context (while for other entities features such as capital letters and symbols can be used). For this reason, as stated in Section 3, in order to perform the classification task on users, we will use lexical features only. This makes the classification task harder.

Furthermore, users are rare and their lexical context (words that indicates their presence) are sparse. To measure that, we compare the word co-occurrence distribution of the nouns *user* and *device* and the pronouns *who* and *which*. We choose *device* and *which*, because they represent artefacts in patents (i.e. technologies, products, components). These are particularly challenging to distinguish from users, because they have similar context (Chiarello, Trivelli et al., 2018). Furthermore, *user* and *device* are chosen to validate the difficulty of the task in terms of features, whereas *who* and *which* to assess if the pronouns replicate this behaviour and are thus a good proxy for users identification. Looking at Definition 1 we focus on verbs since they are the most indicative features for this entity.

In Fig. 1 we show the log–log plot of the co-occurrence with other words of the words *device* and *user*. In order to show that no relevant difference exists between different IPC classes, we make the measures for each of them separately.

The graph is equally spaced on the x-axis, while on the y-axis there are the sorted numbers of co-occurrences, then logarithm is taken on both scales so that a power law becomes a (approximately) straight line. For example, in the top-left plot measures of co-occurrence with the pronoun *device* in the IPC A01B is shown. For each distribution we also compute α that is the power law parameter. It is evident that for the noun *user* α is higher, thus the distribution has a fatter tail. A highly



	A01B		A21B		A41D	
	α	$x_{\min}$	α	$x_{\min}$	α	$x_{\min}$
which	2.055	0.001	2.130	0.002	2.165	0.004
who	2.950	0.003	2.483	0.006	2.643	0.002

Fig. 2. Log–log plots of word co-occurrence with the words *who* and *which* for each patents' category, below estimate parameter of the power laws that would best fit the frequencies data.

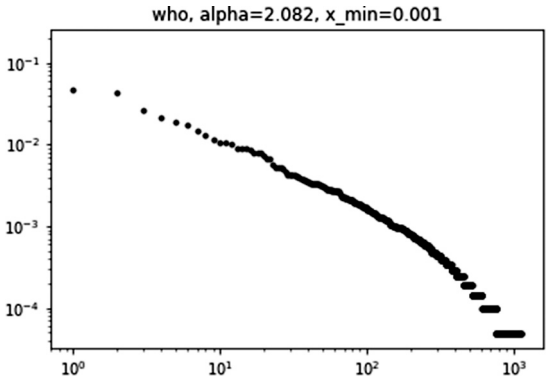


Fig. 3. Log–log plots of verbs' frequencies contextual with *who*, below estimate parameter of the power laws that would best fit the frequencies data.

similar behaviour can be seen in Fig. 2, where a similar log–log plot is shown for the words *which* and *who*. From this we can draw three conclusions:

1. The higher value of α of the co-occurrence distribution on users makes the relevant features for classification (verbs that co-occur with the word users) more rare. This makes the classification harder as more relevant features are needed for a correct analysis.
2. The behaviours of nouns and pronouns are closely related, supporting the use of the second as a proxy for the first.
3. The task at hand does not need the IPC class separation (see Fig. 3).

4.2. Training set validation

As stated in Section 3, the first hypothesis under which we work is that pronouns occur in contexts similar to those where users do, so that learning to classify these pronouns' contexts can generalize to finding users. This follows from the fact that pronouns identify persons, which, in a patent, almost certainly indicate users (Chiarello, Trivelli et al., 2018).

To measure the validity of this assumption, we use standard scores: accuracy, precision, recall and F1 score. We define these measures with respect to a predefined lexicon of users (Chiarello, Cimino et al., 2018) as defined in Section 3.

Table 2
Baseline for each pronoun.

	Accuracy	Precision	Recall	F1
she	0.72	0.61	0.79	0.69
he	0.69	0.53	0.78	0.63
anyone	0.66	0.45	0.79	0.57
someone	0.64	0.44	0.74	0.55
anybody	0.63	0.40	0.73	0.52
somebody	0.63	0.44	0.71	0.54
everybody	0.62	0.40	0.72	0.51
everyone	0.57	0.31	0.64	0.42
which	0.52	0.11	0.64	0.19
user	0.51	0.07	0.58	0.13
device	0.51	0.09	0.60	0.15

To measure a baseline for the classification task, for each of this pronouns we compose a set of sentences where half of them hold one of the pronoun and then we measure the scores as mentioned above in Table 2, for example as compared to formula (1), the true positives for the baseline becomes

$$tp = |\{\text{Sentences with a user}\} \cap \{\text{Sentences with a pronoun}\}|.$$

This can be used as a baseline, since half of the sentences in the data set we use for testing contain one of the pronouns. Notice that by construction each accuracy in Table 2 is bounded from below by 50%, since half of the sentences contain a pronoun. One more aspect that can be inferred from Table 2, as we expected, is that some pronouns co-occur with users sensibly more often than others, in particular, the last three rows of the table induce us not to consider the words *user* and its plural in the list of taggers.

The second hypothesis is that the lexical context of a noun or a pronoun, is a strong indicator of whether it is a user or not. The fact that the lexical context of a word is a strong indicator of the meaning of the word itself is demonstrated in state of the art systems, as explained in Devlin et al. (2019). For what concerns users, in order to understand which classes of words are best indicators of the presence of this entity, we have to consider its definition (Chiarello, Cimino et al., 2018) reported as Definition 1. From this we can deduce that users interact with the invention (actively or passively). Thus, we can better specify our second hypothesis, considering that certain verbs (indicating an interaction) are strong indicators of the presence of a user in a sentence.

To find out which are these verbs, we made a first investigation, based on the observations of the words *user* and *device*. The assumption behind this choice is that *user* will most often indicate a user, while *device* can provide good examples of sentences where the focus is on the invention being patented (or a technology, product, component), so that the verbs in that context most likely do not identify users.

In Fig. 4 we demonstrate the validity of our second hypothesis. The plot represents the most co-occurrent verbs with the words *"user"* and *"device"*. On the x-axis the plot measures the frequency of co-occurrence with the word *user*; on the y-axes the frequency of co-occurrence with the word *device*. It is evident that the words are separated in two clusters. We notice In particular that among those closer to user there are verbs such as *desires* and *feel* which are typical to humans (users) only.

Notice that, to make all the figures of this chapter, a few highest frequency verbs are ignored, since they are not informative, being mainly variations of the verbs *be* and *have*, which show a number of occurrences out of scale when compared to all others.

4.3. Classification

4.3.1. Baseline

As outlined in Section 3, a bias in the data we have to cope with, is the co-occurrence between users and *taggers*. We aggregate the results shown in Section 4.2 and compare them with a random sentence

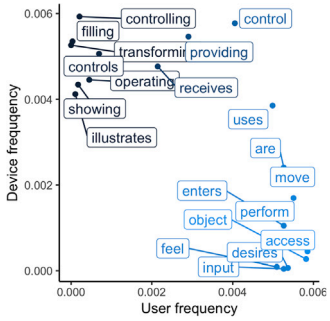


Fig. 4. The co-occurrence frequency with the word *user* is plotted against the co-occurrence frequency with *device*, so that in the top left words occurring with *device* more than with *user* are found, while the opposite in the bottom right.

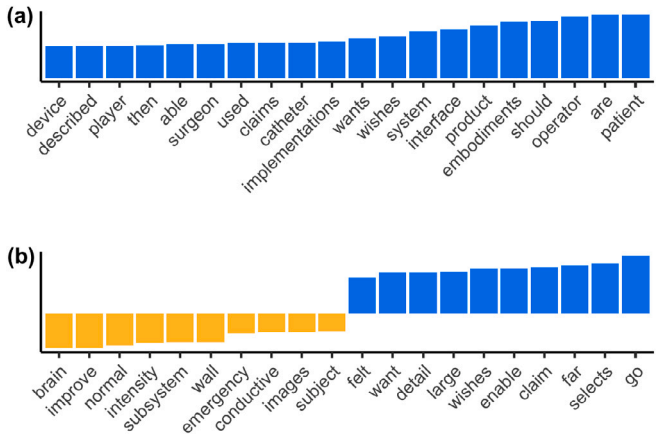


Fig. 5. In (a) the most relevant features for tree classifier and in (b) the most relevant positive (yellow) and negative (blue) features for SVM.

sampling. We report the ratio of sentences with a user in two different cases, randomly selected sentences or sentences containing one of the *taggers*. For the first the ratio of sentences containing a user is 0.22 whereas for the second it is more than double, 0.55. Using the user list in Chiarello, Cimino et al. (2018), for both cases we considered sets of 10^4 sentences and averaged over 10 repetitions. This result implies, as expected, that there is a trade-off between learning how to classify the sentences with a pronoun and the effective generalization to users. Indeed if a trivial method was used, selecting all the sentences where one of the *taggers* appears, in this case, the result would be limited by a precision score of 0.55. Another constraint is that the users co-occurring with the words we choose, are likely to have specific characteristics, again limiting the generalization power of out system. This poses a challenge that will be reflected in the results we obtain: a classifier performing better on *taggers* will eventually start performing worse than a weaker one on users, once a certain accuracy threshold is reached.

4.3.2. Machine learning methods

As a first experiment we employ classification trees and linear support vector machines. The reason behind this choice, is that using these algorithms we are able to inspect the decisions made by the system. In this way we can inspect the possible *taggers* choices, allowing us to select the best data-set to tackle the user extraction challenge. One more way to refine the analysis is inspecting the most relevant features used by these classifiers.

The experiments are set of 192 thousand sentences for the *taggers* task is studied and $5 * 10^5$ sentences for the *users*. In each set 50% of the sentences are a positive matches. We use 80% of the sentence for training, and the rest for testing. Of each sentence we only kept nouns, verbs, adjectives and adverbs, give that the rests of the Part of Speeches are not good indicators for users (Chiarello, Trivelli et al., 2018). The results of the experiments are shown in Table 3. As already known in the literature, SVM outperform decision trees in this kind of tasks (Isozaki & Kazawa, 2002).

For all our experiments there are two measures to consider. The first consists of measuring the performance regarding pronouns classification and how well that is performed. The second regards the users, since indeed our intuition stands in the idea that pronoun classification can generalize to the users classification. To measure the performance on the second task, we use the list of users by Chiarello, Cimino et al. (2018).

One more step in this direction is inspecting which features are most relevant. This can be easily done for classification trees, known to create highly interpretive models. However, looking at Fig. 5 we notice how an intuitive meaning can be found in the relevant features also for SVM. In particular, we notice the presence of words regarding human sensations. We also notice how some features indicate strong biases present in text under consideration. What happens is that some standard sentences typical to the writing of a patent, introduce biases. The strongest one is the word *claim*, which is often used in a patent to refer to the claims of the patent itself or of other patents. This fact creates false positive for our system, but shows that patents texts rises interesting linguistic challenges worth of investigation.

Indeed, one more consideration following from the experiments is that all our analyses are extremely dependent on the IPC. In some categories other than *Human Necessities*, the presences of the users in the final list obtained in Chiarello, Cimino et al. (2018) is below 1%. The highest categories distinction, the first letter, turns out to be crucial to the effectiveness of our work. In other terms, a system trained on a IPC cannot be used to tag sentences coming from others IPCs. We expected this from the beginning as it is already pointed out in Chiarello, Cimino et al. (2018).

Indeed, also patents belonging to the *Human Necessities* categories but to different sub classes show variations in the kind of words and frequencies, as shown in Section 4. At this stage of the work we join the data from all the sub-categories, both due to the small amount of sentences in some of them and in the attempt to increase the robustness of the classifiers.

We also consider that the classification experiments attempted so far can only account for the vocabulary used in the text. Indeed, since a one-hot encoding of the sentences was used, these classifiers only choose one vector from the other based on the words that compose a sentence. Word order and possible dependencies between words are completely ignored by these methods. At the same time, the state of the art in computational linguistics, is currently obtained by focusing on sentence encodings able to embed linguistic meaning, one of the goal of this work is to find out if this solution can help in this specific task.

Our next argument is that comparing the results obtained with the *taggers* and with the word *user* is meaningful even though the actual data we trained the classifiers on are not the same. The sentences composing the two sets however different are obtained in the same way. The procedure used is as follows: from each patent category considered, all the sentences containing one of the words we look for were extracted. This way not only the effectiveness of the task is compared, but, at the same time, the best data set is chosen, since for any word the entire set of available patents was exhausted.

4.3.3. Neural networks methods

In the current NLP literature there exists a strong focus on interpreting the encoding performed by state of the art language models (Abnar, Beinborn, Choenni, & Zuidema, 2019; Chowdhury & Zamparelli, 2019; Jawahar, Sagot, & Seddah, 2019). In particular the large models developed by Devlin et al. (2019), but also smaller ones such as the one

Table 3
Scores achieved by classification tree and support vector machine on both the *user* sentences and the *taggers*.

	Accuracy	Precision	Recall	F1 score
Taggers *SVM*	0.82	0.78	0.86	0.81
Taggers *Tree*	0.77	0.78	0.76	0.77
User *SVM*	0.86	0.86	0.86	0.86
User *Tree*	0.77	0.78	0.78	0.78

Table 4
Scores achieved by deep models on *taggers*, BERT indicates the model in Devlin et al. (2019) while UE indicates the universal sentence encoder from Cer et al. (2018).

	Accuracy	Precision	Recall	F1 score
BERT fixed weights	0.80	0.76	0.76	0.78
BERT fine tuned	0.94	0.93	0.95	0.94
UE fixed weights	0.77	0.83	0.67	0.74
UE fine tuned	0.91	0.91	0.89	0.90

from Cer et al. (2018), seem to have the ability to express meaning beyond pure vocabulary. For the task tackled here, this can be of fundamental importance, given the complexity of the words that need to be discovered, which can change their nature from users to not, depending on the context where they appear.

We performed the classification task in the same setting as described in Section 4.3.2, with a deep learning approach. In particular, different attempts are made, where the main difference consists in whether the encoding provided by pre-trained networks can be relied upon, or if instead there is need to retrain the model for our specific task. What we show is that in general, since the frequency of pronouns is extremely low in patents, the classifier can exploit information present in the rest of the text.

Both the models trained in Cer et al. (2018) and the bigger model present in Devlin et al. (2019) are fine-tuned to address the specific task at hand. This is done focusing on two main aspects. The first is checking if the sentence-encoding is able to extract the information of the user presence, keeping in mind that this task can be hard for a person non expert in the field. The second is the fast applicability of this method, in particular we add a feed forward neural network on top of the encoding and then let the training propagate back to the whole network. Fine tuning the bigger model (Devlin et al., 2019) is more time consuming as the number of parameters to tune is higher.

In Table 4 we show the scores achieved by the classifiers. They show that the largest fine-tuned model (BERT fine tuned) fits better than all others on the data-set. While using it as a non trainable sentence encoder (BERT fixed weights) or using the model from Cer et al. (2018) (UE) performs worse on this data-set though they are faster to train. We notice that despite these scores, when used in general text for the user classification task their performance can be different and we investigate this in the conclusions.

In doing these experiments, for both architectures, classical soft max cross entropy is used, with a fixed learning rate of $2 * 10^{-5}$ and 3 epochs in the training, employing an Adagrad optimizer. The reason behind these particular choices are purely empirical, that is the best performances out of several attempts were achieved with these settings. While keeping the training time always below the 8 h.

5. Discussion and conclusion

In the present paper we employed different classification methods (i.e. decision trees, support vector machines and neural networks) in order to solve an entity specific Named Entity Recognition task in patents documents. To do that, we did not rely on a manually tagged training set, but on a knowledge base that exploit the characteristics of the entity to be extracted (i.e. users of the invention). The obtained results confirm that the proposed method can be employed to achieve improvements in the user identification task with respect to previous works (Chiarello, Cimino et al., 2018).

Table 5
Precision achieved by the models on repeated tests over sets of 1000 patents from the same 6 IPC categories on all sentences in the patents description.

Precision table	User	Device
Baseline	0.22	0.14
SVM	0.40	0.25
BERT fixed weights	0.44	0.36
BERT fine tuned	0.14	0.13
UE fixed weights	0.34	0.26

5.1. Discussion

In the present section we discuss the performances on the users identification task on which we tested the system. We selected two test sets of patents from the same IPC classes as before. These sets are completely unseen by the models to provide a fair testing ground. We create two distinct sets of patents, dividing those containing the word *user* and the word *device*. These two sets are used to compare the performance of the systems. In fact, we speculate that patents containing the word *user* are more likely to contain other user related words with respect to the ones containing the word "device". As shown in Table 5, the hypothesis is confirmed by the differences in the results as will be further discussed.

In Table 5 we report the precision achieved by the different models for the different patents' sets. As stated in Section 1, in the present paper we propose a classification method in order to identify domain specific entities in domain specific documents. For this reason we measure the precision achieved by concurring systems considering the task of classifying sentences containing a user. The classification is performed among all the sentences of groups of 1000 patents, only considering their description section. We repeated the task 5 times (on 5 different sets of sentences). In Table 5 we report the mean of the precision scores. We only report the precision, since the recall, due to the overall low appearance rates of users (about 20% of the sentences) is not informative.

The baseline model for the performances of the system, is obtained by selecting random sentences among the one in the chosen patents. In other words the baseline is the probability of randomly finding a sentence containing a user in the patents we selected.

The most interesting fact in the table is that the system is able to generalize. This is shown by the fact that results are higher than the baseline for the model with fixed BERT weights. At the same time, we see that the fine tuned model has surpassed the threshold where the precision on the taggers becomes a negative factor and indeed on the users classification task it performs worse than both the universal sentence encoder and the fixed weights model.

We underline how all this is achieved without any human effort in the data-set creation and how the generalization achieved by the models is a clear sign confirming the validity of the methodology of selecting the pronouns as a proxy for users.

We can interpret the results in Table 5 in a more detailed way as follows. First of all we notice how all the models, except BERT fine tuned, surpass the baseline, meaning that our assumption about the sentence selection does work, since we are able to partially recover the users in the paper (Chiarello, Cimino et al., 2018). We remark that the way we defined the measures only accounts for the users as found in Chiarello, Cimino et al. (2018).

A special discussion is needed by the BERT fine tuned model. The reason why this model fails in our task is because it is over-fitting the pronouns. In other words, the approach of using generic entities of users (pronouns such as he, she, etc.) as the training set leads to poor results with a model that somehow over-fits these entities. Anyway, it has to be stressed that this result cannot be generalized to other domain. In fact, pronouns are rare in patents (with respect to other documents type). Their very rare appearance rates, as noted

in Section 4, makes it so that the model perfectly fitting the training set performs poorly (due to generalization problems) on the new test sentences.

On the other hand we notice how the support vector machine based system is able to perform comparably well to the best performing model (BERT fixed weights) on the set of patents including the word *user*. This is not the case on those including *device*. Notice how this result is coherent with the fact that the support vector machine highly relies on vocabulary, while the BERT language model can encode syntactic features as well (Tenney et al., 2019). In fact as we noted in Section 4, the vocabulary is more closely related between the "user" and the *taggers* than it is for "device", and this is well reflected by this results.

Finally the universal encoder model, of which we only performed the fixed weights version (knowing that the BERT system was performing better using this paradigm) is found in between the behaviours of support vector machines and BERT model, and only partially exploits both features, resulting in lower precision.

We can therefore summarize the results of this paper as twofold: we found an extremely fast way to retrieve a set of sentences largely containing the entities we wish to extract (*users*) and we managed to exploit the information contained in such data-set to find other users in new unseen text. These poses a fast way to start tackling the more general issues we presented in the introduction and that we proposed to address.

5.2. Limitations and future work

The proposed approach has several limitations worth discussing to help directing future studies in the task of domain specific Named Entity Recognition.

As we mention, the weakest step in the proposed process is the difference between the *taggers* and actual users. Particularly, we remark that there is an unknown threshold after which performing better on the *taggers* diminishes the result quality on users. One possible refinement would be to try to clean the sentences where the pronouns occur, removing the remaining ones that do not contain users, by looking more closely at the linguistic features of these sentences, (e.g. number and types of nouns, subjects, etc.).

The second limitation of our work is that, though we find evidence of the quality of our methodology for data-set creation, a larger suite of tests with human validation would be needed. As stated in Section 1, this was out of scope for the present paper, but future works can employ more human resources to test the method in greater detail.

Finally, applying the proposed methodology together with the one developed in Chiarello, Cimino et al. (2018), could lead to a more precise and exhaustive approach to tackle the user extraction task.

These are research lines we will certainly investigate closely in the future.

CRediT authorship contribution statement

Giovanni Puccetti: Software, Formal analysis, Methodology, Data curation, Writing – original draft. **Filippo Chiarello:** Data curation, Writing – review & editing, Conceptualization. **Gualtiero Fantoni:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL workshop blackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 191–203). Florence, Italy: Association for Computational Linguistics.

Abujabal, A., Saha Roy, R., Yahya, M., & Weikum, G. (2018). Never-ending learning for open-domain question answering over knowledge bases. In *WWW '18, Proceedings of the 2018 world wide web conference* (pp. 1053–1062). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, http://dx.doi.org/10.1145/3178876.3186004.

Arts, S., Hou, J., & 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), Article 104144. http://dx.doi.org/10.1016/j.respol.2020.104144.

Asche, G. (2017). "80% of technical information found only in patents" – Is there proof of this [1]? *World Patent Information*, *48*, 16–28. http://dx.doi.org/10.1016/j.wpi.2016.11.004.

Atkinson, J., & Bull, V. (2012). A multi-strategy approach to biological named entity recognition. *Expert Systems with Applications*, *39*(17), 12968–12974. http://dx.doi.org/10.1016/j.eswa.2012.05.033.

Bekoulis, G., Deleu, J., Demeester, T., & Develder, C. (2018a). An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, *102*(C), 100–112. http://dx.doi.org/10.1016/j.eswa.2018.02.031.

Bekoulis, G., Deleu, J., Demeester, T., & Develder, C. (2018b). Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, *114*, 34–45. http://dx.doi.org/10.1016/j.eswa.2018.07.032.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Binkhonain, M., & Zhao, L. (2019). A review of machine learning algorithms for identification and classification of non-functional requirements. *Expert Systems with Applications: X*, *1*, Article 100001. http://dx.doi.org/10.1016/j.eswax.2019.100001.

Blanco-Fernández, Y., Gil-Solla, A., Pazos-Arias, J. J., Ramos-Cabrer, M., Daif, A., & López-Nores, M. (2020). Distracting users as per their knowledge: Combining linked open data and word embeddings to enhance history learning. *Expert Systems with Applications*, *143*, Article 113051. http://dx.doi.org/10.1016/j.eswa.2019.113051.

Burggräf, P., Wagner, J., & Weißer, T. (2020). Knowledge-based problem solving in physical product development—-A methodological review. *Expert Systems with Applications: X*, *5*, Article 100025. http://dx.doi.org/10.1016/j.eswax.2020.100025.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., et al. (2018). Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-2029.

Chiarello, F., Cimino, A., Fantoni, G., & Dell'Orletta, F. (2018). Automatic users extraction from patents. *World Patent Information*, *54*, 28–38. http://dx.doi.org/10.1016/j.wpi.2018.07.006.

Chiarello, F., Cirri, I., Melluso, N., Fantoni, G., Bonaccorsi, A., & Pavanello, T. (2019). Approaches to automatically extract affordances from patents. In *Proceedings of the design society: International conference on engineering design, Vol. 1* (pp. 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* (pp. 101–110).

Chiarello, F., Trivelli, L., Bonaccorsi, A., & Fantoni, G. (2018). Extracting and mapping industry 4.0 technologies using wikipedia. *Computers in Industry*, *100*, 244–257. http://dx.doi.org/10.1016/j.compind.2018.04.006.

Chowdhury, S. A., & Zamparelli, R. (2019). An LSTM adaptation study of (un)grammaticality. In *Proceedings of the 2019 ACL workshop blackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 204–212). Florence, Italy: Association for Computational Linguistics.

Ciaramita, M., & Altun, Y. (2005). Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS workshop on advances in structured learning for text and speech processing, Vol. 2005*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 2126–2136). Melbourne, Australia: Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Fantoni, G., Apreda, R., Dell'Orletta, F., & Monge, M. (2013). Automatic extraction of function–behaviour–state information from patents. *Advanced Engineering Informatics*, *27*(3), 317–334. http://dx.doi.org/10.1016/j.aei.2013.04.004.

Fernández, N., Arias Fisteus, J., Sánchez, L., & López, G. (2012). IdentityRank: Named entity disambiguation in the news domain. *Expert Systems with Applications*, *39*(10), 9207–9221. http://dx.doi.org/10.1016/j.eswa.2012.02.084.

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of the 2001 conference on empirical methods in natural language processing*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93:1–93:42. http://dx.doi.org/10.1145/3236009.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). Spacy: industrial-strength natural language processing in python. Zenodo, http://dx.doi.org/10.5281/zenodo.1212303.

Isozaki, H., & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *COLING 2002: The 19th international conference on computational linguistics*. URL: https://www.aclweb.org/anthology/C02-1054.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language?. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3651–3657). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1356.

Jung, J. J. (2012). Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Systems with Applications*, *39*(9), 8066–8070. http://dx.doi.org/10.1016/j.eswa.2012.01.136.

Konkol, M., Brychcín, T., & Konopík, M. (2015). Latent semantics in named entity recognition. *Expert Systems with Applications*, *42*(7), 3470–3479. http://dx.doi.org/10.1016/j.eswa.2014.12.015.

Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics*, *7*(1), S1. http://dx.doi.org/10.1186/1758-2946-7-S1-S1.

Küçük, D., & Yazıcı, A. (2012). A hybrid named entity recognizer for Turkish. *Expert Systems with Applications*, *39*(3), 2733–2742. http://dx.doi.org/10.1016/j.eswa.2011.08.131.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01, Proceedings of the eighteenth international conference on machine learning* (pp. 282–289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Leaman, R., Wei, C.-H., & lu, Z. (2015). TmChem: A high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, *7*, S3. http://dx.doi.org/10.1186/1758-2946-7-S1-S3.

Lidén, C., & 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. http://dx.doi.org/10.1517/13543776.2011.573786.

Liu, L., Li, Y., Xiong, Y., & Cavallucci, D. (2020). A new function-based patent knowledge retrieval tool for conceptual design of innovative products. *Computers in Industry*, *115*, Article 103154. http://dx.doi.org/10.1016/j.compind.2019.103154.

Lo, S. L., Chiong, R., & Cornforth, D. (2017). An unsupervised multilingual approach for online social media topic identification. *Expert Systems with Applications*, *81*, 282–298. http://dx.doi.org/10.1016/j.eswa.2017.03.029.

Matin, R., Hansen, C., Hansen, C., & Mlgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, *132*, 199–208. http://dx.doi.org/10.1016/j.eswa.2019.04.071.

McCallum, A., Freitag, D., & Pereira, F. C. N. (2000). Maximum entropy Markov models for information extraction and segmentation. In *ICML '00, Proceedings of the seventeenth international conference on machine learning* (pp. 591–598). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio, & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, workshop track proceedings*.

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36–54. http://dx.doi.org/10.1016/j.eswa.2018.03.058.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Nicoletti, M., Schiaffino, S., & Godoy, D. (2013). Mining interests for user profiling in electronic conversations. *Expert Systems with Applications*, *40*(2), 638–645. http://dx.doi.org/10.1016/j.eswa.2012.07.075.

Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, *40*(7), 2373–2390. http://dx.doi.org/10.1016/j.eswa.2012.10.073.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization* (pp. 23–49). Springer, http://dx.doi.org/10.1007/978-3-642-28569-1_2.

Romero, S., & Becker, K. (2019). A framework for event classification in tweets based on hybrid semantic enrichment. *Expert Systems with Applications*, *118*, 522–538. http://dx.doi.org/10.1016/j.eswa.2018.10.028.

Sari, Y., Hassan, M. F., & Zamin, N. (2010). Rule-based pattern extractor and named entity recognition: A hybrid approach. In *2010 international symposium on information technology, Vol. 2* (pp. 563–568). http://dx.doi.org/10.1109/ITSIM.2010.5561392.

Sarica, S., Luo, J., & Wood, K. L. (2019). Technet: Technology semantic network based on patent data. *Expert Systems with Applications*, Article 112995. http://dx.doi.org/10.1016/j.eswa.2019.112995.

Silvestri, S., Gargiulo, F., & Ciampi, M. (2019). Improving biomedical information extraction with word embeddings trained on closed-domain corpora. In *2019 IEEE symposium on computers and communications (ISCC)* (pp. 1129–1134). http://dx.doi.org/10.1109/ISCC47284.2019.8969769.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4593–4601). Florence, Italy: Association for Computational Linguistics.

Tsai, T., Chou, W.-C., Wu, S.-H., Sung, T.-Y., Hsiang, J., & Hsu, W.-L. (2006). Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. *Expert Systems with Applications*, *30*(1), 117–128. http://dx.doi.org/10.1016/j.eswa.2005.09.072.

Wang, Y., Wang, M., & Fujita, H. (2020). Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems*, *190*, Article 105030. http://dx.doi.org/10.1016/j.knosys.2019.105030.

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2145–2158). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Zesch, T., & Gurevych, I. (2006). Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the workshop on linguistic distances* (pp. 16–24). Sydney, Australia: Association for Computational Linguistics.