Expanding SNOMED-CT through Spanish Drug Summaries of Product Characteristics

Pablo Calleja
Universidad Politécnica de Madrid
Madrid, Spain
pcalleja@fi.upm.es

Raúl García-Castro
Universidad Politécnica de Madrid
Madrid, Spain
rgarcia@fi.upm.es

Guadalupe Aguado-de-Cea
Universidad Politécnica de Madrid
Madrid, Spain
lupe@fi.upm.es

Asunción Gómez-Pérez
Universidad Politécnica de Madrid
Madrid, Spain
asun@fi.upm.es

ABSTRACT

Terminologies in the biomedical field are one of the main resources used in the clinical practice. Keeping them up-to-date to meet real-world use cases is a critical operation that even in the case of well maintained terminologies such as SNOMED-CT involves much effort from domain experts. Pharmacological products or drugs are constantly being approved and made available in the market and their clinical information should be also updated in terminologies. Each new drug is provided with its Summary of Product Characteristics (SPC), a document in natural language that contains its essential information.

This paper proposes a method for populating the Spanish extension of SNOMED-CT with drug names using SPCs and representing their clinical data sections in the terminology. More precisely, the method has been applied to the therapeutic indication and the adverse reaction sections, in which disease names are recognized as named entities in the document and mapped to the terminology. The relations between the drug name and the mapped entities are also represented in the terminology based on the specific roles that they have in the document.

1 INTRODUCTION

Terminologies and lexicons from the early moment they are created, face the problem of not being up-to-date to meet real-world use cases. Maintaining them can be quite expensive and the need to constantly be updated is well known [5, 8, 16]. For example, the biomedical field deals with new drug products that are constantly being approved and made available on the market. Medical surveil-lance and medical records need to keep in track the prescribed drugs. In this area, one of the main terminologies used for many medical purposes is SNOMED-CT.1 This multilingual terminology represents the whole knowledge of the medical domain by means of concepts and descriptions associated with codes, related between them and nested in different levels of detail and categories. How-ever, its drug catalog is not complete and is composed mainly of generic products.

The commercialization of new drugs is granted by regulatory agencies in a country or region. In Spain, this authorization depends on the *Agencia Española de Medicamentos y Productos Sanitarios* (AEMPS) and on the European Medicine Agency (EMA). Such agencies verify the new product clinical research through its Summary of Product Characteristics (SPCs). SPCs are natural language documents in which the essential information of the drug is reflected (e.g., drug name, indications, dosage, contraindications, adverse reactions, etc.). This valuable information can be captured by information extraction (IE) systems [2].

Most of the IE works in the biomedical field are focused on corpora provided by conferences or challenges such as JNLPBA [13] and on the exploitation of paper abstracts [11]. Other works, closer to health applications, exploit resources like Electronic Health Records (EHR) [18]. While EHRs are usually short simple phrases written by doctors, the SPCs are complex and detailed documents.

Named Entity Recognition (NER) is one of the most important tasks inside IE processes; it involves natural language processing that consists in finding and classifying real-world entities denoted with a referent term or proper name (named entity). However, the state of the art is oriented towards retrieving all the possible entities regardless of whether they are relevant or not to a concrete use scenario. Apart from paper abstracts and conference corpora, natural language documents display mixed information in which there are entities that are not relevant to a use scenario and that

[1] http://www.snomed.org/

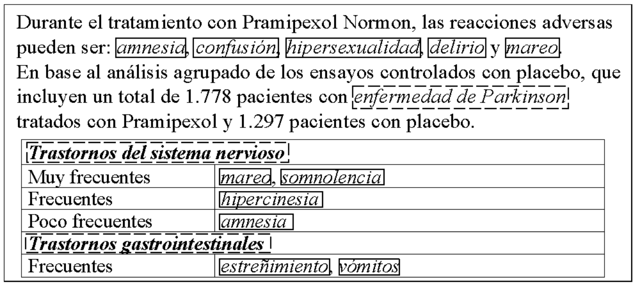
produce noise on the overall result of the NER task. This is the case of the SPC in which, although they are divided in sections referring to those specific use scenarios, there is a lot of mixed information in them.

The current classifications covered in NER systems just deal with taxonomic types and are not meant to represent the entities' role. Such role determines the general meaning and function of the entity in the corpus. However, as the conceptual model proposed by Steimann shows [27], entity roles can also be represented as a classification model. The taxonomic hierarchy of diseases is normally represented by the affection type such as "mental disorder" and "gastric disease". However, "adverse reaction" and "contraindication" are roles that an entity may have in a given context, and which can be also represented in a taxonomic form.

To prove the heterogeneity of the information presented in the sections of the SPCs, some research founded by the AEMPS has been carried out. This research was focused on two specific sections and it reflected that, in each of them, there are entities with the same named entity type (disease) but with different roles, apart from the proper role of the section.

Therapeutic indication section. This section of the SPC provides information about the disease or diseases to be treated with the drug. However, this section is sometimes verbose and includes information from other sections. Commonly, this extra information deals with contraindications, diseases for which the drug should not be prescribed (e.g., *should not be used as a treatment for*), or diseases that refer to the medical record of the patient (*who does not have a recent history of*).

Adverse reactions section. This section of the SPC provides information about unwanted effects caused by the administration of a drug (diseases or disorders) and their frequency. However, sometimes this section also contains therapeutic indication information to specify the adverse reactions studied for a particular patient with a disease. Figure 1 shows an example of the adverse reaction section. All the disease entities in the figure are in italics. However, only the entities surrounded with a continuous black box are adverse reactions. The entities surrounded with a segmented black box represent other roles.



Durante el tratamiento con Pramipexol Normon, las reacciones adversas pueden ser: *amnesia*, *confusión*, *hipersexualidad*, *delirio* y *mareo*.
En base al análisis agrupado de los ensayos controlados con placebo, que incluyen un total de 1.778 pacientes con *enfermedad de Parkinson* tratados con Pramipexol y 1.297 pacientes con placebo.

Trastornos del sistema nervioso	
Muy frecuentes	*mareo*, *somnolencia*
Frecuentes	*hipercinesia*
Poco frecuentes	*amnesia*
Trastornos gastrointestinales	
Frecuentes	*estreñimiento*, *vómitos*

Figure 1: Excerpt of the adverse reactions section

This paper proposes a method for populating the Spanish extension of SNOMED-CT with drug names using SPCs and representing their clinical data sections in the terminology. More precisely, the method has been applied to the therapeutic indication section and the adverse reaction section in which disease names are recognized as entities in the document and mapped to the terminology. The relations between the drug name and the mapped disease entities are based on the specific roles they have in the document.

The paper is structured as follows. Section 2 describes related work and section 3 addresses the proposed method for populating SNOMED-CT from SPCs. Section 4 evaluates and discusses the obtained results. Finally, Section 5 presents the conclusions of the work and highlights future research lines.

2 RELATED WORK

There are many works in the literature focused on the exploitation of SNOMED-CT for clinical purposes [7]. However, there is a reduced number of works interested in the population of the terminology. Researchers such as [10] have aimed to expand the Swedish extension with synonyms extracted from Electronic Health Record. For Spanish, there are few projects that use the Spanish extension of SNOMED-CT and most of them have focused on mapping processes without a specific work area, such as [6]. In general mapping projects use the SNOMED-CT terminology in English [20, 24] and in other languages such as Swedish [26].

Nowadays, the state of the art of NER shows that the best results are provided by supervised machine learning techniques [4, 21]. However, these techniques require big annotated corpora to be trained such as [14, 19]. Thus, the use of machine learning techniques is limited in some domains in which such annotated corpora do not exist. Furthermore, annotated corpora do not reflect the entities' roles.

Roles are defined and attached to text segments or entities in an IE task called template filling [23, 27]. These roles are defined by its acts or meaning in a given context and normally are associated with lexical-syntactic patterns [22]. Nevertheless, the template filling task in the biomedical domain is focused on event relations like cause-effect or drug-drug interactions [25].

As for the mapping techniques used in the biomedical area, a thorough analysis has been carried out. Some works are general-purpose studies [12, 17], whereas others do not focus on recognizing entities that may include numerous descriptors such as diseases. Likewise, it has been observed that current techniques are not oriented to mapping terms extracted from natural language texts [1, 9]. These current techniques are very sensitive to any linguistic variation and do not exploit the syntactic categories of words that compose the terms.

3 METHOD

The proposed method for populating the Spanish extension of SNOMED-CT with drug names using SPCs and representing their clinical data sections in the terminology is divided into three main processes: named entity recognition of diseases and drug names, entity mapping with SNOMED-CT, and creation of the SNOMED-CT extension.

The first process is focused on the recognition of named entities in different sections. This process is divided into two tasks. The first one aims to find drug names that represent the product. The second task is focused on the recognition of diseases in clinical data sections with the same role. Although SPCs have specific separated sections for each piece of information, they present the problem defined in the introduction, mixed information. The named entity recognition process is based on gazetteers and rules oriented for Spanish language.

The second process aims to map the identified disease entities with SNOMED-CT to identify their equivalent concepts. The mapping process tries to find the best candidate, by creating all the possible variants of the entity in which the description of the concept may appear in the terminology, using word normalization and synonyms. Finally, the SNOMED-CT extension is created: populating the product names, creating their adverse reaction and indication concepts and relating them to the appropriate-role mapped diseases.

The method was developed in a collaboration project with the AEMPS focused only on two concrete sections of the clinical data: the therapeutic indication and the adverse reaction sections. The implementation of the method relies on the GATE text processing framework. The AEMPS provided one set of more than 1,000 SPCs in Spanish divided into different files per section. From this set, 120 randomly selected SPCs had been annotated by domain experts from the agency. The annotation process was made separately in the two required sections of the SPC. In each section, the annotated diseases represent the target role of their section ("therapeutic indication" or "adverse reaction"). These annotated SPCs have been used to create two gold standard corpora to evaluate the NER process; a gold standard corpus for the therapeutic indication section and a gold standard corpus for the adverse reaction section. Moreover, the AEMPS provided a list of 500 disease entities with their equivalent SNOMED-CT description and concept code as a gold standard for the mapping process.

3.1 External resources

As mentioned before, the problem of named entity recognition is to identify named entities according to a role. However, gazetteers of named entities are needed to support the process. For drug names, a gazetteer has been created using the SNOMED-CT descriptions that contain the label "(product)" in the preferred term. As this work aims to populate SNOMED-CT with product concepts, if the concept already exists it is not necessary to create an instance and it is possible to refer to it directly.

For disease and disorder names, the selected gazetteers were extracted from the Spanish version of the Medical Dictionary for Regulatory Activities (MedDRA) [3] and from the *Diccionario de siglas médicas* [29]. MedDRA is a rich and highly specific standardized medical terminology to facilitate sharing of regulatory information internationally for medical products. The terminology is a disease term list classified in hierarchies by their degree of depth and it is distributed in various languages, including Spanish. The two lowest hierarchy levels are called Preferred Terms (PT) and Lowest Level Terms (LLTs). Both levels contain more than 70,000 terms that reflect simple observations reported in the clinical practice. Due to its exhaustive list of diseases with non-specific terms, both levels have been used to create the gazetteer of entities with named entity type "disease".

Using MedDRA rather than a subset of the disorders and clinical findings of SNOMED-CT is due to 3 reasons: (1) A preliminary study has found that the coverage of MedDRA for common used terms in the SPCs is better; MedDRA was based on a terminology belonging to the Medicines and Healthcare products Regulatory Agency (MHRA) of UK and developed for pharmaceutical purposes. (2) MedDRA is easier to use due to its classification in fewer hierarchies. (3) SNOMED-CT has concepts with high detailed descriptions that are not relevant for constructing a gazetteer (e.g., *anemia por deficiencia de hierro en la madre que complica el nacimiento* - iron deficiency anemia in the mother complicating birth).

Another gazetteer used for disease and disorder names has been extracted from the *Diccionario de siglas médicas*. Although MedDRA contains a lot of international terms, the SPCs contain a lot of abbreviations used colloquially in Spain. This gazetteer was created by extracting only the abbreviations.

Moreover, in collaboration with the AEMPS, a synonym and acronym catalog has been created to enrich the mapping process with SNOMED-CT. Using their expert knowledge and resources such as the *Diccionario de siglas médicas*, sets of terms have been grouped according to their semantic relation. This catalog is not focused on detecting all the possible cases, but only the most relevant ones. (e.g., *cáncer* and *tumor maligno* (malign tumour), HBP and *hiperplásia benigna de próstata* (benign prostate hyperplasia) or the set of adjectives *incrementada* increased, *elevada* elevated and *aumentada* augmented).

3.2 Named entity recognition process

For each SPC, the named entity recognition process is performed in two different tasks. The first one identifies drug names over the name section. In this section the product is defined with the different dosages and pharmaceutical forms in which it is commercialized (e.g., Celecoxib Lesvi 100 mg cápsulas duras EFG / Celecoxib Lesvi 200 mg cápsulas duras EFG). So, each of the terms are considered different named entities (as SNOMED-CT represent products). In this task, the gazetteer created from SNOMED-CT is used to identify already instantiated drugs. If the drug name does not exist in the terminology, the different terms that are in the section are considered new entities.

The second task is focused on the identification of diseases in the clinical data sections; in this case, the therapeutic indication and the adverse reaction sections. The main challenge in each section is to identify not only diseases but the correct ones according to their role in their corresponding section. As for the disease recognition, the gazetteers extracted from MedDRA and from the *Diccionario de siglas médicas* are used to discover entities. To support the disease recognition, some coded patterns are used to find entities that are not reflected on the gazetteers. Table 1 shows the lexical patterns (character combinations or affixes) and syntactic patterns (word combinations) used to identify diseases.

Figure 1 shows an example in a document in which there are some examples of patterns commonly used in medicine. There is a lexical pattern represented by the suffix "_itis" that means inflammation (e.g., sinusitis). Lexical patterns are used to identify nominal phrases (NP) as a disease in which the noun contains at least one of the affixes. The nominal phrases include the adjectival phrases (AdjP) that are joined to the noun. Also, there is a syntactic pattern represented by the word combination *enfermedad de* (disease of) and an eponim (Parkinson). The syntactic patterns represented in

Table 1 show between brackets which part of the span text is considered as an entity. The syntactic patterns 19 to 21 are diseases that are represented by fluctuation disorders of biological substances of the organism and are presented in sets having the same meaning with different words. For example, pattern 21 represents decrease of biological substances and there are four words to compose the pattern: *disminución* (decrease), *reducción* (reduction), *pérdida* (loss) and *descenso* (decline). The optionality of the words is presented between parenthesis and separated by the symbol '|'.

Once the disease named entities are identified, a new set of coded patterns are executed in order to classify diseases according to their role. These role patterns could be syntactic patterns and layout patterns (the format in which the information is presented). Figure 1 shows three examples with a syntactic pattern and two layout patterns. The syntactic pattern *las reacciones adversas pueden ser* (the adverse reactions could be) is reflecting that all the entities under the scope of the pattern (until the full stop) have the adverse reaction role. On the other hand, a layout pattern is also represented in the example. All the entities in the table that are not headers in bold are also diseases with the adverse reactions role. Other roles are represented with different patterns. The first one is represented with the word combination *pacientes con* (patients with) that is reflecting diseases with the role for which the drug is prescribed (therapeutic indication role). The second one is represented by rows in bold font and it reflects the classification role of the mentioned diseases (classification role).

Table 1: Disease patterns identified

<table>
<tr><td colspan="4"><b>Lexical Pattens</b></td></tr>
<tr><td>1) _oma</td><td>2) _itis</td><td>3) _osis</td><td>4) _algia</td></tr>
<tr><td>5) _ema</td><td>6) _asis</td><td>7) _emia</td><td>8) _orrea</td></tr>
<tr><td>9) _penia</td><td>10) _plasia</td><td>11) hiper_</td><td>12) hipo_</td></tr>
<tr><td colspan="4"><b>Syntatic patterns</b></td></tr>
<tr><td colspan="2">13) {infección de + NP}</td><td colspan="2">14) {enfermedad de + NP}</td></tr>
<tr><td colspan="2">15) {enfermedad + AdjP}</td><td colspan="2">16) {afección de + NP}</td></tr>
<tr><td colspan="2">17) {virus de + NP}</td><td colspan="2">18) {empeoramiento de + NP}</td></tr>
<tr><td colspan="4">19) {(prolongación | incremento | elevación | aumento) de + NP}</td></tr>
<tr><td colspan="4">20) {(alteración | anormalidad | cambios | descompensación) de + NP}</td></tr>
<tr><td colspan="4">21) {(disminución | reducción | pérdida | descenso) de + NP}</td></tr>
</table>

All of the identified role patterns had been found in a prior study of each corpus in collaboration with experts from AEMPS. Table 2 shows the patterns identified and their role. In the therapeutic indication section, the diseases with the section's role are represented by the syntactic patterns 22 to 28. Also, two more roles have been observed in this section: "medical record" and "contraindication". "Medical record" advises the patient that the therapeutic indication of the drug is dependent of diseases of the patient's clinic history. "Contraindication" represents the diseases for which the use of the drug is not recommended.

Table 2: Role patterns identified

<table>
<tr><td><b>Syntatic patterns</b></td><td><b>Role</b></td></tr>
<tr><td>22) tratamiento de + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>23) pacientes con + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>24) alivio de los síntomas de + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>25) asociado a + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>26) prevención de + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>27) administración en combinación en + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>28) en ensayos clínicos de + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>29) en estudios clínicos de + {NP}</td><td>Therapeutic indication</td></tr>
<tr><td>30) sin + {NP}</td><td>Medical record</td></tr>
<tr><td>31) que se hayan excluido + {NP}</td><td>Medical record</td></tr>
<tr><td>32) excluyendo + {NP}</td><td>Medical record</td></tr>
<tr><td>33) que no tiene + {NP}</td><td>Medical record</td></tr>
<tr><td>34) siempre que no exista + {NP}</td><td>Medical record</td></tr>
<tr><td>35) pero no + {NP}</td><td>Contraindication</td></tr>
<tr><td>36) no protege + {NP}</td><td>Contraindication</td></tr>
<tr><td>37) no debe ser utilizado + {NP}</td><td>Contraindication</td></tr>
<tr><td>38) no se recomienda + {NP}</td><td>Contraindication</td></tr>
<tr><td>39) no debe utilizarse + {NP}</td><td>Contraindication</td></tr>
<tr><td>40) no se ha demostrado | estudiado + {NP}</td><td>Contraindication</td></tr>
<tr><td>41) no se han realizado estudios + {NP}</td><td>Contraindication</td></tr>
<tr><td>42) potenciado por + {NP}</td><td>Medical interaction</td></tr>
<tr><td>43) con el fin de evitar + {NP}</td><td>Medical interaction</td></tr>
<tr><td>44) no se asocia + {NP}</td><td>Non adverse reaction</td></tr>
<tr><td>45) no se observó + {NP}</td><td>Non adverse reaction</td></tr>
<tr><td>46) sin indicios de</td><td>Non adverse reaction</td></tr>
<tr><td colspan="2"><b>Layout Patterns</b></td></tr>
<tr><td>47) Headers in bold with tags <b>, <i> or <u></td><td>Classification headers</td></tr>
<tr><td>48) Entities between labels <table> and </table></td><td>Adverse reaction</td></tr>
<tr><td>49) Entities in below classification headers</td><td>Adverse reaction</td></tr>
</table>

In the adverse reaction section, it has been observed that most of the adverse reactions presented in SPCs are represented in tables, indicating the affection classification type and their frequency. At the same time, documents that do not contain tables also use the affection classification type as headers to enounce adverse reactions. Both layout patterns had been used to identify entities with the adverse reaction role. Pattern 48 associates entities that are inside the HTML table tag as an entity and pattern 49 identifies entities that are in the text under a header and associates them to the adverse reaction role. The role "therapeutic indication" was also present in the section. In specific cases SPCs enounce the disease for which the drug is prescribed. Other roles that have appeared are "medical interaction", "non adverse reaction" and "classification headers". "Medical interaction" represents diseases that only appear in a specific case or diseases that could produce more adverse reactions. The "non adverse reaction" role represents diseases that have not been discovered as adverse reactions during the drug clinical research. The last role refers to "classification headers", i.e., general diseases that classify and enounce the adverse reactions.

The named entity recognition process based on role patterns verifies that the named entities belong to the corresponding section and discards the ones that belong to the other ones. Considering as an assumption that the presented information is representative of its section, if an entity does not belong to any role, it takes the same one as the section. Roles can be represented in a taxonomic form as shown in Figure 2. However, this work is only focused on the roles "therapeutic indication" and "adverse reaction".

3.3 Disease mapping process

The disease mapping process provides the equivalent concept of the SNOMED-CT terminology for a disease entity. This process has two tasks. The first one preprocesses the entity creating all

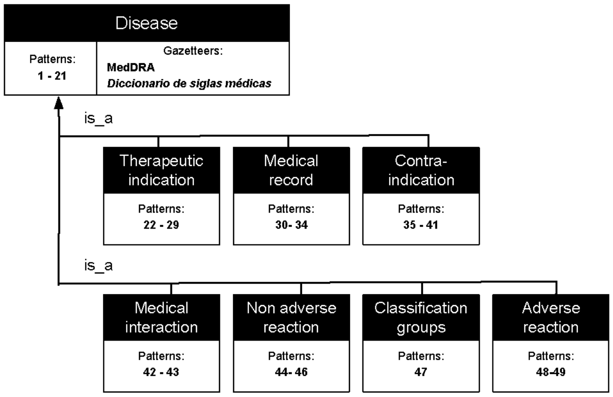


Figure 2: Disease role classification

the possible variants in which the disease could be written in the descriptions of the terminology, using word normalization and synonyms. The second task identifies the candidate SNOMED-CT concept using a mapping algorithm based on the entity variants and in the syntactical function of the words.

Entity preprocessing

The entity preprocessing aims to prepare the disease entities obtained in the last process to identify their equivalent concepts in SNOMED-CT in the next task. The task cleans the input entity and creates all the disease entity variants that the entity can adopt. All the words are lowercase converted, the Spanish accent is removed and the stopwords are deleted. Once the text is clean, the words are classified into two main groups: keywords and descriptors. The keywords are the nouns that will be the main reference in the mapping process. The rest of the words such as adverbs, adjectives or numbers are classified as descriptors. All the classified words are used to create disease entity variants, being the initial one composed by the original tokens of the entity. In the case that there are words in plural, a new disease entity variant is created with these tokens in singular (normalization). Finally, disease entity variants are checked with the synonym and acronym catalog created with the AEMPS. If one word matches, a new disease entity variant is created for each synonym as showed in Fig. 3. The figure shows the different disease entity variants for the disease term *cáncer de testículos avanzado* (advanced cancer of testicles). Keywords are represented in bold and underlined.

Concept identification

The concept identification task aims to provide a SNOMED-CT concept for each disease entity using a mapping algorithm based on the classification of the words (keywords and descriptors). The algorithm takes all the disease entity variants and searches for each one through the descriptions of the terminology concepts. As mapping is focused on diseases, only concepts that belong to the category "disease" and "clinical finding" are taken into account. The algorithm uses a threshold that counts the number of equal words, by exact comparison, between a disease entity variant and a SNOMED-CT description. The pseudocode of the algorithm is presented in Algorithm 1.

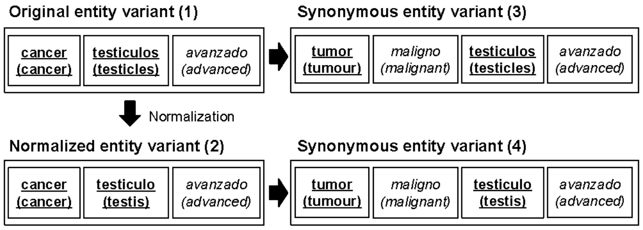


Figure 3: Entity variants of *Cáncer de testículos avanzado*

The algorithm stores SNOMED-CT descriptions that have the same threshold value as the candidates, discarding those that are below that value. If this threshold is exceeded, all previous candidates are deleted, the threshold is updated, and the algorithm starts saving new candidates beginning with the one that produced the change. The threshold distinguishes between keywords and descriptors. Names or keywords are prioritized, so when the algorithm is processing a SNOMED-CT description, it firstly compares the names that have been recognized in the disease entity variant and in the SNOMED-CT description. If the number of keywords is bigger than the threshold, it is updated by erasing the previous candidates. If the number of keywords is equal to the threshold, the algorithm does the same comparison with descriptors. With this approach, a term that has two keywords and no descriptors would be a better choice than another term that has only one keyword and three descriptors in common.

For example, for the disease entity variant *cáncer testículos* (testicles cancer), the algorithm goes through SNOMED-CT storing candidate terms of the terminology that exceed the threshold. In this case, SNOMED-CT has no terms containing the two keywords *testículos* (testicles) and *cáncer* (cancer), but only one (e.g., *cáncer de piel* (skin cancer), *herida expuesta de los testículos* (open wound of testicles), etc.). With another disease entity variant of the same disease, *tumor maligno testículo* (testis malignant tumour), the algorithm finds new candidates with two keywords, discarding previous candidates (e.g., *tumor simple de testículo* (simple tumour of testis)). Finally, the algorithm discards the previous candidates when it finds the term *tumor maligno de testículo* (malignant tumour of testis) containing a descriptor above the previous candidates. The result for the *cáncer de testículos* disease entity is the SNOMED-CT term *tumor maligno de testículo* (Malignant tumour of testis - 363449006).

Finally, the identified candidates are evaluated to select the best one. SNOMED-CT is very large and represents very detailed concepts so it is necessary to select the best in some way. Each candidate has an incorporated measure based on the approximation distance to the SNOMED-CT description. The measure is calculated by subtracting the number of words in the SNOMED-CT description (stopwords do not count) and the sum of all words (keywords and descriptors) that the algorithm has found for the disease entity variant. This measure gives a quantification of the similarity between them. The candidate SNOMED-CT descriptions obtained by the algorithm are organized from the lowest to the highest distance value. The description that has the lowest value is selected. Following the example of the *cáncer testículos* (cancer of testicles), candidates retrieved from the algorithm are those that contain the words *tumor* (tumour), *maligno* (malignant) and *testículo* (testi). The optimal candidate is *tumor maligno de testículo* (Malignant tumor of testis - 363449006) but the algorithm also returns *tumor maligno de testículo de células de Leydig* (Malignant Leydig cell tumour of testis - 278055006). In the first case, the distance is zero because the SNOMED-CT description is composed by three words (not counting stopwords) and three of them matched. In the second case, the distance is two, the SNOMED-CT description has five words (not counting stopwords) and only three of them matched. The best candidate is the one with the lowest distance: *tumor maligno de testículo* (Malignant tumour of testis - 363449006). If the algorithm retrieves more than one description, it is considered as a bad mapping.

Algorithm 1 Mapping Algorithm

```
Require: Entity_Variants
Candidates ← { }
Keywords_threshold ← { 0 }
Descriptors_threshold ← { 0 }
for all Var ∈ Entity_Variants do
    for all Desc ∈ SNOMED_Description do
        Keywords ← {countCommonKeyWords(Var,Desc)}
        Descriptors ← {countCommonDescriptors(Var,Desc)}
        if ( KeyWords > Keywords_threshold ) then
            Candidates← {Desc}
            Keywords threshold← { KeyWords }
        end if
        if ( Keywords == Keywords_threshold ) then
            if ( Descriptors > Descriptors_threshold ) then
                Candidates ← { Desc }
                Words threshold ← { Words }
            end if
            if ( Descriptors == Descriptors_threshold ) then
                Candidates ← {Candidates, Desc}
            end if
        end if
    end for
end for
Candidates ← $selectBestCandidates(Candidates)$
```

3.4 SNOMED-CT extension creation process

The third process consists in creating the SNOMED-CT extension. The identifiers for new concepts in the extension, according to the IHTSDO manual,[2] have to be created using UUIDs. For each SPCs various concepts must be created. First, the identified drug names are used to find the general drug name without dosage or pharmaceutical form (e.g., Celecoxib Lesvi for the drug names Celecoxib Lesvi 100 mg and Celecoxib Lesvi 200 mg) to create a new concept subtype of the SNOMED-CT concept "Pharmaceutical/Biological product (product) - 373873005". Then, new subtypes of the general drug concept are created using the drug names. If the general drug name already exists in the terminology, the subtypes are referenced using the SNOMED-CT identification number (SCTID) of the identified product.

For the therapeutic indication section, a new concept subtype of "Drug indicated (situation) - 135794007" is created showing the general drug name in the description. All the mapped disease entities with the therapeutic indication role are referenced as subtypes. The same process is made for the adverse reaction section. A new concept subtype of "Adverse reaction section (disorder) - 281647001" is created with the general drug name in the description. Also, the mapped disease entities with the adverse reaction role are referenced as subtypes.

4 EVALUATION

The evaluation has been made separately for the two main processes: the NER process and the mapping process. The former has been performed by measuring the results obtained over the gold standard corpus of the therapeutic indication and the adverse reaction sections. The latter has been performed through the gold standard of mapped disease entities.

4.1 NER evaluation

Each section has been evaluated separately with four experiments related with the resources used in the NER process. The first experiment only takes the results obtained by gazetteers. The second experiment adds the results obtained by the patterns coded for disease named entities. The third experiment uses the role patterns of the section to add or associate entities to the section's role. Finally, the fourth experiment uses the patterns of the other roles to discard entities that are not associated to the section's role. The experiment for both sections works under the open world assumption; i.e., entities with no role are considered to be part of the target role.

The evaluation metrics used for the NER process are precision (P), recall (R) and F-measure (F). The evaluation measures the detected entities under two matching criteria as proposed in other biomedical evaluations [28]. Normally, NER systems use the strict or exact matching criteria; the entity detected by the system and the entity annotated in the gold standard corpus must have the same span text and named entity type. However, the annotated entities of the provided gold standard corpus have problems in the consensus between annotators, i.e., the same entity with the same span text is annotated with different length (different offset in one side). Normally, the difference between annotations are adjectives that experts have taken or not into account in the annotation process, such as adjectives that describe a particular case of the patient's disease (e.g., *recurrente* (recurrent)) and intensity or degree (e.g., *grave* (severe)). Thus, the partial criteria allows that one of the span text offsets can be different.

Table 3 presents the obtained results of the 4 experiments in the therapeutic indication section. Firstly, experiment 1 denotes that gazetteers cover most of the entities that are presented in the corpus, but not representative enough to cover all of them. Experiment 2 shows how lexical and syntactic patterns detect more disease named entities thus improving the results of the gazetteers.

[2] http://doc.ihtsdo.org/download/doc_TechnicalImplementationGuide_Current-en-US_INT_20150131.pdf

Experiment 3 shows how the patterns of the section's role increase the recall. However, precision decreases because these syntactic patterns are overlapped with other role patterns. For example, sometimes pattern 18 *tratamiento de* (treatment of) overlaps with pattern 31 *no debe ser utilizado* (must not be used) in the sentence *no debe ser utilizado para el tratamiento de la rinitis* (must not be used in the treatment of the rhinitis). Experiment 4 demonstrates that applying the patterns of other roles to discard entities increases the final precision, with a minimal decrease in the recall. All the experiments reflect how the corpus is influenced by the annotation problems mentioned previously. Only the partial matching criteria obtained the best F-measure in experiment 4.

Table 3: Therapeutic indication section evaluation

	Strict			Partial		
	P	R	F	P	R	F
Exp. 1	0.8513	0.8716	0.8613	0.9122	0.9339	0.9229
Exp. 2	0.9266	0.9266	0.9266	0.9633	0.9633	0.9633
Exp. 3	0.9135	0.9303	0.9218	0.9622	0.9798	0.9709
Exp. 4	0.9249	0.9266	0.9258	0.9744	0.9761	**0.9753**

The evaluation of the adverse reaction section is presented in Table 4. Experiment 1 shows again that gazetteers cover most of the entities. However, the main problem gazetteers have is the representation of disorders: MedDRA represents disorders with adjectives (e.g., *glucosa aumentada* (increased glucose)) and SPCs represent them in a nominal form (e.g., *aumento de glucosa* (increase of glucose)). This problem is solved by using syntactic patterns that represent the nominal form of the disorders in experiment 2. Patterns of the named entity type increase recall significantly in the partial matching criteria. The results in the exact matching criteria show again the annotation problems of the corpus. Experiment 3 shows no modification over the results. Layout patterns are associating discovered entities (in tables or under headers) to the target role and not discovering new ones. In spite of not improving the results, associating entities to the target role is critical if the experiment is not under the open world assumption. Experiment 4 finally shows that patterns from other roles are used to discard entities and the precision on the overall result increases. As in the previous evaluation, the partial matching criteria obtained the best F-measure.

Table 4: Adverse reaction section evaluation

	Strict			Partial		
	P	R	F	P	R	F
Exp. 1	0.8861	0.8117	0.8472	0.9528	0.8728	0.911
Exp. 2	0.8510	0.8237	0.8371	0.9411	0.9108	0.9257
Exp. 3	0.8510	0.8237	0.8371	0.9411	0.9108	0.9257
Exp. 4	0.8633	0.8215	0.8419	0.9538	0.9076	**0.9302**

4.2 Disease mapping evaluation

The evaluation of the mapping algorithm has been performed with four experiments based on the different forms in which it is possible to create disease entity variants. The measure used is the accuracy of the obtained results over the gold standard. The first experiment addresses the creation of the disease entity variant using the tokens of the entity, without synonyms or normalization. The second one adds the possibility of normalizing the disease entity in case some words are in plural. The third experiment is focused only on the creation entity variants using synonyms. The last one evaluates the results obtained with the complete creation of entity variants task, with normalization and using synonyms.

Table 5: Disease entity mapping evaluation

	Accuracy	
	Therapeutic indication section	Adverse reaction section
Exp. 1	0.8323	0.8990
Exp. 2	0.9221	**0.9357**
Exp. 3	0.8502	0.8990
Exp. 4	**0.9401**	**0.9357**

Table 5 shows the mapping results of the experiments performed over the disease entities recognized in the two sections. The results of the therapeutic indication section prove that experiment 2 (normalization) improved significantly the results over the simple mapping process using only the original tokens, as normally SPCs present diseases in plural. Experiment 3 also improved the results using synonyms over experiment 1. The combination of both techniques to create disease entity variants obtained the best accuracy in experiment 4. In contrast, the results of the adverse reaction section differ. The results are also improved by normalization in experiment 2 but they are not affected by the use of synonyms in experiment 3. Experiments 2 and 4 have the same accuracy. Despite the good results, the mapping evaluation is not complete and representative because the list of 500 gold disease mapped entities only covers 169 diseases from 307 entities obtained in the therapeutic indication section and 109 from 2042 entities obtained in the adverse reaction section.

The evaluations have been performed only with the entities reflected and all the unknown cases have not been taken into account. In the unknown results, a subset of 18 in the therapeutic indication section and a subset of 540 in the adverse reaction section contain the same SNOMED-CT description text as the entity. Furthermore, most of the other unknown results are entities that differ in irrelevant adjectives like frequency *alopecia frecuente* with *alopecia - 56317004*. Thus, an extension of the gold standard should be developed by experts to cover all the possible cases to validate, for example, the real utility of disease entity variants with synonyms in the adverse reaction section.

5 CONCLUSIONS AND FUTURE WORK

The evaluations of the NER process driven by roles show that precision is increased in those natural language documents in which only some of the entities are required. This process has been a key factor to represent the future relations of the diseases with the drug. Roles and their patterns allow to represent a classification model of the entities and to identify named entities with a specific role.

Normally, the proposed role classification model could be represented with a target role and all the other roles joined as the complementary role (¬Target role). However, these two specific use cases have also demonstrated that to precisely classify and define roles benefits the overall work because different sections have repeated roles and patterns. Thus, it is possible to create a complete role classification model for all the roles that could be reused in different use cases for homogeneous domain-specific documents without searching and reimplementing patterns.

The main disadvantage of the method is that it requires very time-consuming tasks involving domain experts. Although the method is proposed for real use cases in which it is better to annotate only the required entities instead of annotating and classifying them all, the patterns and the role they represent have been discovered manually. This method performs the first approach to introduce entity roles inside the NER task in natural language documents in which detecting specific entities according to a role is required.

As mentioned above, one of the most time-consuming tasks was pattern discovery and the creation of the role classification model. Among the future lines of work regarding the proposed method is to explore automatic pattern discovery using algorithms based on distributional semantics such as Latent Semantic Analysis (LSA) [15]. In the same way, the next steps should explore other sections of SPCs such as contraindication to extract disease entities and reflect them into SNOMED-CT.

The proposed mapping process based on the creation of all the possible disease entity variants shows how the results are improved over using the original tokens. Also, the algorithm used based on the syntactic functions of the words allows to obtain closer results in case there is no clear mapping and reduces the retrieval of false positives. Some of these cases are reflected in the gold standard, e.g., experts from the AEMPS agreed that the entity *artritis reumatoide activa* (Active rheumatoid arthritis) is represented by the SNOMED-CT concept *artritis reumatoide*- 69896004 and could be mapped through the algorithm that does not take into account the descriptor word *activa*. Further evaluations should be done with experts to increase the gold standard with the unknown mappings in which many of these cases are represented. Finally, the next step is to request a SCTID to the IHTSDO organization to formalize the created extension with the clinical data and publish it.

ACKNOWLEDGMENTS

This work was funded by the Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), by the Programa Propio of the Universidad Politécnica de Madrid and by project Datos 4.0 (TIN2016-78011-C4-4-R), of the Agencia Estatal de Investigación MINECO and Fondos FEDER.

REFERENCES

[1] S. A. Akhondi, K. M. Hettne, E. van der Horst, E. M. van Mulligen, and J. A. Kors. 2015. Recognition of chemical entities: combining dictionary-based and grammar-based approaches. *Journal of Cheminformatics* 7, 1 (2015).

[2] R. Boyce, G. Gardner, and H. Harkema. 2012. Using Natural Language Processing to Identify Pharmacokinetic Drug-drug Interactions Described in Drug Package Inserts. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP '12)*.

[3] E. G. Brown, L. Wood, and S. Wood. 1999. The Medical Dictionary for Regulatory Activities (MedDRA). 20, 2 (1999).

[4] D. Campos, S. Matos, and J. L. Oliveira. 2012. Biomedical named entity recognition: a survey of machine-learning tools. In *Theory and Applications for Advanced Text Mining*. InTech.

[5] A. Coden, D. Gruhl, N. Lewis, M. Tanenblatt, and J. Terdiman. 2012. Spot the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*. IEEE.

[6] J. Cruanes, M. T. Romá-Ferri, and E. Lloret. 2012. Measuring lexical similarity methods for textual mapping in nursing diagnoses in Spanish and SNOMED-CT. In *MIE*.

[7] K. Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. 121 (2006).

[8] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web* 8, 6 (2017).

[9] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. 2005. Resolving abbreviations to their senses in Medline. *Bioinformatics* 21, 18 (2005).

[10] A. Henriksson, M. Skeppstedt, M. Kvist, M. Duneld, and M. Conway. 2013. Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics.

[11] L. Hunter and K. B. Cohen. 2006. Biomedical language processing: What's beyond PubMed? *Molecular Cell* 21, 5 (2006).

[12] A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. In *BMC bioinformatics*, Vol. 9. Issue Suppl. 3.

[13] J. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*.

[14] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - A semantically annotated corpus for bio-textmining. In *Bioinformatics*, Vol. 19. Issue Suppl. 1.

[15] M. Konkol, T. Brychcín, and M. Konopík. 2015. Latent semantics in named entity recognition. *Expert Systems with Applications* 42, 7 (2015).

[16] K. Lee, A. Qadir, S. A Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri. 2017. Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

[17] M. Meizoso García, J. L. Iglesias Allones, D. Martínez Hernández, and M. J. Taboada Iglesias. 2012. Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations. *International journal of medical informatics* 81, 8 (2012).

[18] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook of Medical Informatics* 2008, 1 (2008).

[19] I. Moreno, E. Boldrini, P. Moreda, and M. T. Romá-Ferri. 2017. DrugSemantics: a corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics. *Journal of Biomedical Informatics* (2017).

[20] A. Mottaz, Y. L. Yip, P. Ruch, and A. Veuthey. 2008. Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC bioinformatics* 9, 5 (2008).

[21] D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30, 30 (2007).

[22] S. Patwardhan and E. Riloff. 2007. Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Vol. 7.

[23] R. C. Schank and R. P. Abelson. 1975. Scripts, Plans, and Knowledge. *Proceedings of the 4th International Joint Conference on Artificial Intelligence* (1975).

[24] L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. 2011. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* 40, 1 (2011).

[25] B. Settles. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. (2004).

[26] M. Skeppstedt, M. Kvist, and H. Dalianis. 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text.. In *Proceedings of the eight international conference on language resources and evaluation*.

[27] F. Steimann. 2000. On the representation of roles in object-oriented and conceptual modelling. *Data and Knowledge Engineering* 35, 1 (2000).

[28] R. T. Tsai, S. Wu, W. Chou, Y. Lin, D. He, J. Hsiang, T. Sung, and W. Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC bioinformatics* 7, 1 (2006).

[29] V. Yetano Laguna, J., Alberola Cuñat. 2003. *Diccionario de siglas médicas*. Ministerio de sanidad y consumo.