Role-based model for Named Entity Recognition

Pablo Calleja, Raúl García-Castro, Guadalupe Aguado-de-Cea, Asunción Gómez-Pérez
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
{pcalleja,rgarcia,lupe,asun}@fi.upm.es

Abstract

Named Entity Recognition (NER) poses new challenges in real-world documents in which there are entities with different roles according to their purpose or meaning. Retrieving all the possible entities in scenarios in which only a subset of them based on their role is needed, produces noise on the overall precision. This work proposes a NER model that relies on role classification models that support recognizing entities with a specific role. The proposed model has been implemented in two use cases using Spanish drug Summary of Product Characteristics: identification of therapeutic indications and identification of adverse reactions. The results show how precision is increased using a NER model that is oriented towards a specific role and discards entities out of scope.

1 Introduction

Information extraction (IE) has become a popular research topic in the last three decades, specially in the biomedical field. Most of the works in this field are focused on corpora provided by conferences or challenges such as JNLPBA (Kim et al., 2004) and on the exploitation of paper abstracts (Hunter and Cohen, 2006). Other works, closer to health applications, exploit resources like Electronic Health Records (EHR) (Meystre et al., 2008) and drug Summary of Product Characteristics (SPC) (Boyce et al., 2012).

While EHRs are usually short simple phrases written by doctors, SPCs are official and detailed documents that collect the essential scientific information of a drug for healthcare professionals. In Spain, the authorization of SPCs depends on the *Agencia Española de Medicamentos y Productos Sanitarios* (AEMPS) and on the European Medicine Agency (EMA).

Named Entity Recognition (NER) is one of the most important tasks inside IE processes that consists in finding and classifying real-world entities denoted by a referent term or proper name (named entity). However, the state of the art is oriented to retrieve all the possible entities regardless if they are relevant or not to a concrete use scenario. Beyond 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 produce noise on the overall result of the NER task.

This is the case of SPCs, which are natural language documents that contain a lot of mixed valuable information for concrete use scenarios. This paper proposes a new method to create a NER model focused only on specific entities by taking into account the presented role of such entities in the corpus. Such role determines the general meaning and function of the entity in the corpus. The method has been implemented for two IE use cases over specific sections of SPCs in Spanish in collaboration with the AEMPS.

The paper is structured as follows. Section 2 describes related work and section 3 describes in detail the problem of documents with different entity roles. Section 4 proposes the method to create a NER model focused on entities which a specific role. Section 5 shows the application of the method over two use cases and section 6 discusses the obtained results. Finally, section 7 presents some conclusions and highlights future work.

2 Related Work

Normally, the main named entity types proposed in the literature are “person”, “organization”, “location” “dates”, “time expressions” and “mone-

Proceedings of Recent Advances in Natural Language Processing, pages 149–156,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_021

tary expressions" (Grishman and Sundheim, 1996; Ferro et al., 2005). The biomedical field defines its own named entity types such as such as "protein", "drug" and "disease" (Rindflesch et al., 2000; Zhou et al., 2005). In contrast, other researchers classify entities into taxonomic models where some types are considered subtypes (i.e., children) of a high level one (e.g., "geological region" or "address" are subtypes of "location") (Sekine et al., 2002).

Nowadays, the state of the art of NER models shows that the best results are provided by supervised machine learning techniques (Nadeau and Sekine, 2007; Campos et al., 2012). However, these techniques require big annotated corpora such as (Kim et al., 2003; Moreno et al., 2017) to be trained. Thus, the use of machine learning techniques is limited in some domains in which there are not such annotated training corpora or there are not in a specific language.

Roles are defined and attached to text segments or entities in an IE task called template filling (Schank and Abelson, 1975; Steimann, 2000). These roles are defined by its acts or meaning in a given context and normally are associated with lexical-syntactic patterns (Patwardhan and Riloff, 2007). Nevertheless, this task in the biomedical domain is focused on event relations like cause-effect or drug-drug interactions (Settles, 2004).

3 Problem Setting

In this context, the current classifications covered in NER systems just deal with taxonomic types and are not meant to represent the entities' role. Nevertheless, as the conceptual model proposed by Steimann shows (Steimann, 2000), entity roles can also be represented as a classification model. For example in the biomedical domain, the taxonomic hierarchy of diseases is normally represented by the affection type such as "mental disorder" and "gastric disease". But, "adverse reaction" or "contraindication" are roles that an entity may have in a given context, and which can also be represented in a taxonomic form. This work proposes to introduce roles into the NER task in order to identify entities according such specific roles.

The following use cases are driven by a NER need in the AEMPS. In them, the agency needs to identify entities in drug SPC documents. However, it usually happens in those documents that there are entities with the same type (disease) but

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

with different role, and the agency is only interested in those entities with a specific role.

The **therapeutic indication section** provides information about the diseases to be treated with the drug. However, it is sometimes verbose and includes information from other sections such as 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 (e.g., *who does not have a recent history of*).

The **adverse reactions section** provides information about unwanted effects caused by the administration of a drug (diseases or disorders) and their frequency. However, sometimes it also contains therapeutic indication information to specify the adverse reactions. Figure 1 shows an example of the adverse reaction section; all the disease entities are in italics, but only the entities surrounded with a continuous black box are adverse reactions. The entities surrounded with a segmented black box represent other roles.

4 NER model for specific entity roles

The proposed NER model for specific entity roles requires the classification of the reflected roles of the named entities in taxonomic models. These models have to be created through manual tasks using a representative gold standard corpus as reference. The annotated entities must be classified by its named entity type and their position in the text (initial and final offset) must be identified.

Besides, the next assumptions are declared. First, that the entities annotated in the gold standard corpus have the role that must be recognized by the NER model (target role). Second, that the annotated entities always belong to a general named entity type (e.g., person, disease, etc.).

In this work, we define the terms pattern, entity type, entity role and role classification model as: A) A *pattern* is a particular contextual sequence

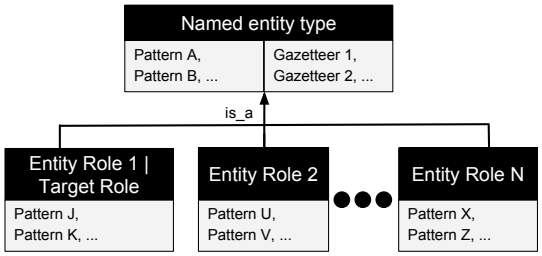

Figure 2: Role classification model representation

of hints that allows to identify an entity by its type or its role; patterns can be lexical, syntactic and layout-based. B) An *entity type* is defined as the classification group of an entity and it composed by sets patterns and gazetteers. C) An *entity role* is defined as a subtype of a named entity type that is characterized by a set of patterns in a specific corpus and describes an entity by its act or meaning in a given context. D) A *role classification model* describes the hierarchy of the different roles for a named entity type in a corpus; the role classification model is composed of a named entity type, one role defined as the target role and a set of zero or more non-target roles. The different roles are disjoint between them and inherit patterns and gazetteers from the entity type.

Figure 2 shows a representation of the role classification model. Once a role classification model is defined, the proposed NER algorithm uses it to identify named entities with a specific role.

4.1 Creation of the role classification model

The method to create a role classification model is composed of the following six tasks that are performed over a gold standard corpus:

1- Extraction of annotated entities. The first task is to collect all the annotated entities of one named entity type in the gold standard corpus.

2- Pattern detection of the annotated entities. The collected annotated entities are studied to detect lexical patterns (character combinations or affixes) that could represent the named entity type. Since lexical patterns do not represent role information, they are attached to the named entity type. Then, the context of the annotated entities is studied to find syntactic patterns (word combinations) and layout patterns (the format in which the information is presented); both types of patterns can represent the named entity type or the target role. Lexical and syntactical patterns are used to identify new possible entities in documents; however, layout patterns are meant to attach roles or named entity types to entities previously identified.

Figure 1 shows an example in which the annotated entities are diseases inside a continuous black box. A lexical pattern is represented by the prefix "hiper_" that is commonly used in disease names. Also, a syntactic pattern represented by the word combination *reacciones adversas pueden ser:* (the adverse reactions could be:). This syntactic pattern is reflecting the target role of the gold standard corpus (adverse reactions) and all the entities under the scope of the pattern (until the full stop) have the target role. Finally, 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 target role (adverse reactions).

3- Identification of gazetteers for the named entity type. Gazetteers are a common NER resource and there are many of them available for the most common named entity types. This task aims to identify gazetteers aligned with the annotated named entities to support the model on the identification of named entities with the same type.

4- Discovery of entities with different role. Using the selected gazetteers and the patterns identified in the previous tasks, a review of the corpus is made to detect named entities of the same type that have not been reflected in the gold standard corpus. Figure 1 shows three disease named entities inside a segmented black box that are not reflected in the gold standard corpus.

5- Pattern detection of the entities with different role. This task aims to detect syntactic and layout patterns of the entities that are not reflected in the gold standard corpus studying their context. These identified patterns represent other roles present in the corpus. The most common role interpretation is the complementary of the target role (¬Target role). However, it is possible to define and classify different roles studying the gold standard corpus.

Figure 1 shows two examples with a syntactic pattern and a layout pattern. 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. The second one is represented by rows in the table that are in bold font that is reflecting the classification role of the mentioned diseases.

6- Creation of the role classification model. This task aims to create the role classification

model based on the named entity type. The named entity type is represented in the model as the upper class and it is composed by the set of patterns discovered in task 2 and the gazetteers identified in task 3; the patterns associated to the named entity type lack of role information. Then, the different roles are specified. First, the target role is represented by the set of patterns associated to the role discovered in task 2. Second, the rest of the roles are specified by the sets of patterns discovered in task 5. Finally, the role classification model of the named entity type could be represented as in Figure 2.

4.2 Role-based NER algorithm

The NER model executes the proposed Algorithm 1 for each role classification model to identify named entities with the target role, along with its span text and position.

The input of the algorithm is the document to be processed, the role classification model and whether a closed world assumption holds. The output is the set of identified named entities. In the algorithm, two sets of entities are defined: the set of entities without role (*entities*) and the set of entities with the target role (*targetEntities*). The algorithm is divided into five main steps. The first one (lines 5 to 11) is oriented to identify named entities by using the gazetteers of the named entity type. The second one (lines 13 to 16) uses the patterns of the entity type to identify named entities. Patterns are detected with the function *detectPattern*. These two tasks store their results in the *entities* set. The next task (lines 17 to 19) executes the patterns of the target role and stores the results in the *targetEntities* set. Then, the patterns of the non-target roles of the model are executed to identify and delete entities of the two sets (lines 20 to 27). Finally, if the NER model is oriented to a close world assumption (lines 28 to 32), the results are composed only by the entities of the *targetEntities* set. In other case, the results are the union between the *entities* set and *targetEntities* set.

Algorithm 1 Role-based NER algorithm

Input: Document d, RoleClassificationModel rcm, boolean cwa
Output: : Set of entities in document d

```
1:  target= rcm.TargetRole
2:  type= rcm.EntityType
3:  entities ← { }
4:  targetEntities ← { }
5:  for all gaz ∈ Gazetteers do
6:      for all entry ∈ gaz do
7:          if (termMatches(entry,d)) then
8:              add(entities,entry)
9:          end if
10:     end for
11: end for
12: aux ← { }
13: for all p ∈ type.Patterns do
14:     add(aux, detectPattern(p,d))
15: end for
16: entities ← entities ∪ aux
17: for all p ∈ target.Patterns do
18:     add(targetEntities, detectPattern(p,d))
19: end for
20: for all role ∈ rcm.NonTargetRoles do
21:     aux ← { }
22:     for all p ∈ role.Patterns do
23:         add(aux, detectPattern(p,d))
24:     end for
25:     entities ← entities - aux
26:     targetEntities ← targetEntities - aux
27: end for
28: if (cwa) then
29:     return targetEntities
30: else
31:     return entities ∪ targetEntities
32: end if
```

5 Method Implementation

As mentioned in section 3, two projects in collaboration with the AEMPS have implemented the proposed NER model for specific roles. Both projects were oriented to the exploitation of different sections of the SPCs. The first aimed at automatically identifying therapeutic indications in the section with the same name. The second aimed at improving pharmacological surveillance processes through the identification of adverse reactions in the section with the same name.

The AEMPS provided one set of more than 1,000 SPCs in Spanish. From this set, 120 randomly selected SPCs were annotated by domain experts from the agency. The annotation process was made separately in two sections of the SPC: the therapeutic indication one and the adverse reaction one. In each section, the annotated diseases represent the target role of their section ("therapeutic indication" or "adverse reaction"). From these annotated SPCs, two gold standard corpora

have been created with the two different sections.

In both projects, 80 SPCs of each gold standard corpus had been used to train their respective model and 40 to test them. The SPCs were selected randomly once for the two use cases. As the named entity type in both use cases is 'disease', the proposed gazetteers for task 3 of the role classification model method were extracted from the Spanish version of the medical dictionary MedDRA (Brown et al., 1999) and from the *Diccionario de siglas médicas* (Yetano Laguna, J., Alberola Cuñat, 2003).

5.1 Therapeutic indication section use case

The first use case aimed at automatically identifying those diseases that have the therapeutic indication role. The gold standard corpus used in this project was created from the therapeutic indication sections of 80 annotated SPCs.

The first two tasks of the method are to extract the annotated entities in the gold standard corpus and to identify patterns. Table 1 shows the patterns discovered through these tasks, which were reviewed by experts from the agency. Lexical patterns are represented by affixes in nouns that are commonly used in medicine like "_itis" (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. The identified patterns had to be classified into patterns that belong to the named entity type and those that belong to the target role therapeutic indication. Lexical patterns are attached directly to the named entity type due to the lack of role information. Similarly, syntactic patterns 13 to 17 are language structures that represent only a disease. Syntactic patterns 18 to 22 represent the common structure to present therapeutic indications in SPCs, describing the target role of the entities.

This use case involved the identification of the named entity type disease, so the gazetteers proposed in the third task are MedDRA and *Diccionario de siglas médicas*. The next task involves the identification of disease entities that are not reflected in the gold standard by using the gazetteers and the patterns identified in the previous task. The context of the discovered entities was then studied to discover patterns of entities with different role. Table 2 shows the patterns identified in the task. Experts from the AEMPS determined that patterns 23 to 27 attach the role "medical record" of the patient, while patterns 28 to 33 represent the role "contraindication", i.e., diseases for which the use of the drug is not recommended.

Lexical Pattens			
1) _oma	2) _itis	3) _osis	4) _algia
5) _ema	6) _asis	7) _emia	8) _orrea
9) _penia	10) _plasia	11) hiper_	12) hipo_
Syntatic patterns			
13) {infección de + NP}		14) {enfermedad de + NP}	
15) {enfermedad + AdjP}		16) {afección de + NP}	
17) {virus de + NP}		18) tratamiento de + {NP}	
19) asociado a + {NP}		20) pacientes con + {NP}	
21) prevención de + {NP}		22) alivio de los síntomas de + {NP}	

Table 1: Patterns identified from the annotated entities in the therapeutic indication use case

Syntatic patterns	
23) sin + {NP}	24) que se hayan excluido + {NP}
25) excluyendo + {NP}	26) que no tiene + {NP}
27) siempre que no exista + {NP}	28) pero no + {NP}
29) no protege + {NP}	30) no se recomienda + {NP}
31) no debe ser utilizado + {NP}	32) no debe utilizarse + {NP}
33) no se ha demostrado/ documentado/ estudiado + {NP}	34) no se han realizado estudios + {NP}

Table 2: Patterns identified for entities with other role in the therapeutic indication use case

Finally, the role classification model was created representing the named entity type as the parent class with its patterns and gazetteers and the different discovered roles and their patterns as children classes.

5.2 Adverse reaction use case

The second use case aimed at identifying adverse reactions. As in the first project, the target entities for the NER model are diseases, but their role is "adverse reaction". The gold standard corpus used for this project was created from the adverse reaction section of the 80 annotated SPCs.

The first two tasks of the method are to extract the annotated entities and to identify patterns; table 3 shows the patterns identified for the entities. Lexical patterns are the same as presented in Table 1 (1 to 12). Also, other syntactic patterns were repeated (13 to 17). The new syntactic patterns (35 to 38) are diseases that are represented by fluctuation disorders of biological substances of the organism. The syntactic patterns in Table 3 are presented in sets having the same meaning with different words. For example, pattern 35 represents decrease of biological substances and there are four words to compose the pattern: *dismin-*

Lexical Pattens	
Patterns 1-12	
Syntatic patterns	
Patterns 13- 17	35) { (disminución | pérdida | reducción | descenso) de + NP}
36) { (prolongación | incremento | elevación | aumento) de + NP}	37) { (alteración | anormalidad | cambios | descompensación) de + NP}
38) { empeoramiento de + NP}	
Layout Patterns	
39) Entities inside tag <table>	40) Entities inside tag <p> below headers in tags <b>, <i> or <u>

Table 3: Patterns identified from the annotated entities in the adverse reaction use case

Syntatic patterns	
Patterns 18-21	41) potenciado por + {NP}
42) con el fin de evitar + {NP}	43) no se asocia + {NP}
44) no se observó + {NP}	45) sin indicios de + {NP}
46) en ensayos clínicos de + {NP}	47) en estudios clínicos de + {NP}
48) administración en combinación en + {NP}	
Layout Patterns	
49) Entities in tags <b>, <i> or <u>	

Table 4: Patterns identified for entities with other role in the adverse reaction use case

ución (decrease), *reducción* (reduction), *pérdida* (loss) and *descenso* (decline). None of these patterns contains information about the role. However, 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 39 associates entities that are inside the HTML table tag as an entity and pattern 40 identifies entities that are in the text under a header and associates them to the adverse reaction role.

As in the other use case, the next step was to identify other entities with the NE type disease that had not been reflected in the gold standard using the gazetteers and the identified patterns. The syntactic patterns found for these entities are presented in Table 4. In specific cases SPCs repeat the diseases with the "therapeutic indication" role (18, 21, 46-48). Other roles that have appeared are "medical interaction", "non adverse reaction" and "classification headers". "Medical interaction" (41-42) represents diseases that only appear in a specific case or diseases that could produce more adverse reactions. The "non adverse reaction" (43-45) role represents diseases that have not been discovered as adverse reactions during the drug clinical research. The last role is "classification headers" (49), general diseases that classify and enounce the adverse reactions. Finally, it was possible to represent the role classification model.

6 Evaluation

The model evaluation has been performed by measuring the results obtained over the 40 test documents of the gold standard corpus. Each use case has been evaluated separately with their own role classification model. The evaluation consisted of four experiments related with the main steps in which the role-based NER algorithm is divided. The first experiment only takes the results obtained by gazetteers. The second one adds the results obtained by the patterns of the named entity type disease. The third one uses the target role patterns to add or associate entities to the target role. Finally, the fourth one uses the patterns of the other roles to discard entities that are not associated to the target role. Both use cases work under the open world assumption; i.e., entities with no role are considered to be part of the target role.

The evaluation metrics used 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 (Tsai et al., 2006). 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 adjectives that describes the intensity or degree (e.g., *grave* (severe)). Thus, the partial criteria allows that one of the span texts offsets can be different.

Table 5 presents the obtained results of the 4 experiments in use case 1. Firstly, experiment 1 denotes that gazetteers cover most of the entities in the corpus, but they are not representative enough to cover all of them. Experiment 2 shows how lexical and syntactic patterns detect more named entities thus improving the results of

	Strict			Partial		
	P	**R**	**F**	**P**	**R**	**F**
Exp. 1	0.8162	0.8531	0.8343	0.9243	0.9661	0.9448
Exp. 2	0.905	0.9153	0.9101	0.9665	0.9774	0.9719
Exp. 3	0.9176	0.9435	0.9304	0.9615	0.9887	0.9749
Exp. 4	0.9382	0.9435	**0.9408**	0.9831	0.9887	**0.9859**

Table 5: Therapeutic indication evaluation

	Strict			Partial		
	P	**R**	**F**	**P**	**R**	**F**
Exp. 1	0.8901	0.809	0.8476	0.9502	0.8636	0.9049
Exp. 2	0.8598	0.8274	0.8433	0.9429	0.9074	0.9248
Exp. 3	0.8598	0.8274	0.8433	0.9429	0.9074	0.9248
Exp. 4	0.872	0.8256	**0.8482**	0.9557	0.9048	**0.9296**

Table 6: Adverse reaction evaluation

the gazetteers. Experiment 3 shows how the patterns of the target role increase the recall. However, precision decreases because these syntactic patterns are overlapped with patterns 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.

The evaluation of the adverse reaction use case is presented in Table 6. Experiment 1 shows again that gazetteers cover most of the entities. The main problem that gazetteers have in this domain 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 in experiment 2 by using syntactic patterns that represent the nominal form of the disorders. Patterns of the named entity type increase recall significantly in the partial matching criteria. The results in the exact matching criteria reflect how the gold standard corpus is highly affected by the annotation problems. 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.

7 Conclusions and Future Work

The evaluations of both use cases show how a NER model oriented by roles increases the precision of the obtained results in those natural language documents in which only some of the entities are required. Roles and their patterns allow to represent a classification model of the entities in a corpus and the NER algorithm uses the role classification model to identify named entities with a specific role. The proposed method to create a role classification model requires a gold standard corpus in which only the required entities are annotated, saving time in the annotation process.

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 and extended in different use cases for homogeneous domain-specific documents.

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 all of them, the patterns and the role that they represent have been discovered manually. This method performs the first approach to introduce entities roles inside the NER task in natural language documents in which the detection of specific entities according to a role is required.

As mentioned above, one of the most time-consuming tasks was the pattern detection. One of the first future lines of work regarding the proposed method is to explore automatic pattern detection using algorithms based on distributional semantics such as Latent Semantic Analysis (Konkol et al., 2015).

Acknowledgments

This work has been funded by the Agencia Española de Medicamentos y Productos Sanitarios and by project Datos 4.0 (TIN2016-78011-C4-4-R), of the Agencia Estatal de Investigación MINECO and Fondos FEDER.

References

R. Boyce, G. Gardner, and Henk H. 2012. Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts. *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, (BioNLP).

E. G. Brown, L. Wood, and S. Wood. 1999. The Medical Dictionary for Regulatory Activities (MedDRA).

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.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions.

R. Grishman and B. Sundheim. 1996. Message Understanding Conference-6: A Brief History. *Proceedings of the 16th conference on Computational linguistics*, 1.

L. Hunter and K. B. Cohen. 2006. Biomedical language processing: What's beyond PubMed?

J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - A semantically annotated corpus for bio-textmining. In *Bioinformatics*, volume 19.

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

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

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).

Isabel Moreno, Ester Boldrini, Paloma Moreda, and M Teresa Romá-Ferri. 2017. Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of Biomedical Informatics*.

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

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

T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.

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

S. Sekine, K. Sudo, and C. Nobata. 2002. Extended named entity hierarchy. In *Third International Conference on Language Resources and Evaluation (LREC 2002)*.

B. Settles. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets.

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

R. T. Tsai, S. Wu, .i 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).

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

G. Zhou, D. Shen, J. Zhang, J. Su, and S. Tan. 2005. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC bioinformatics*, 6 Suppl 1.