Using Snomed to recognize and index chemical and drug mentions.

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano,
Maria-Teresa Martín-Valdivia, L. Alfonso Ureña-López

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda, mcdiaz, maite, laurena}@ujaen.es

Abstract

In this paper we describe a new named entity extraction system. Our work proposes a system for the identification and annotation of drug names in Spanish biomedical texts based on machine learning and deep learning models. Subsequently, a standardized code using Snomed is assigned to these drugs, for this purpose, Natural Language Processing tools and techniques have been used, and a dictionary of different sources of information has been built. The results are promising, we obtain 78% in F1 score on the first sub-track and in the second task we map with Snomed correctly 72% of the found entities.

1 Introduction

Research in biology in the past decade has generated a large volume of available biological data. These texts usually contain information related to drugs, medications, chemicals, reactions, interactions, etc.

Named Entity Recognition (NER) of chemical compounds is receiving increased attention from researchers, as it may facilitate the application of information extraction to the pharmaceutical treatment of diseases. The recognition of pharmaceutical drugs and chemical entities is a critical step required for the subsequent detection of relations of chemicals with other biomedically relevant entities. Biomedical named entity recognition aims to find entities in biomedical texts, an invaluable function that becomes very important for further processing such as information retrieval, information extraction and knowledge discovery. At present, it has referred to kinds of domains, such as protein (Liu et al., 2005; Mitsumori et al., 2005; Tsuruoka and Tsujii, 2003), gene (Liu et al., 2005; Leser and Hakenberg, 2005) or drug (Campillos et al., 2008).

This challenge arises to address the task of recognizing chemicals and drugs. This task has already been studied by several workshops in English, but it is important to continue researching in other languages that have a lot of clinical information. Thanks to this challenge, we can continue studying one of the most widely spoken languages in the world: Spanish. The main aim is to promote the development of named entity recognition tools of practical relevance, that is chemical and drug mentions in non-English content, determining the current-state-of-the art, identifying challenges and comparing the strategies and results to those published for English data.

In terms of English, there were several challenges presented recently such as *CHEMDNER Task: Chemical compound and drug name recognition task* (Krallinger et al., 2015) and *JNLPBA* (Kim et al., 2004) that served to determine the state of the art methodology and systems performance in addition of providing valuable datasets for developing new systems (Tanabe et al., 2005). Some of the most important corpus in this domain are GENIA (Kim et al., 2003), CRAFT (Bada et al., 2012), CALBC (Rebholz-Schuhmann et al., 2010) corpora or SCAI corpus (Kolárik et al., 2008).

In this paper we introduce the participation of the SINAI group in the challenge named PharmaCoNER (Pharmacological Substances, Compounds and proteins and Named Entity Recognition). PharmaCoNER (Gonzalez-Agirre et al., 2019) is one of the workshops presented at BioNLP 2019 and consists of two tracks:

1.1 Track 1: NER offset and entity classification

In this first sub-track the main objective is to find the chemicals and drugs within the text. For its later evaluation it is necessary to write down the beginning and end position of the concept, as well as the appropriate label. The types of entities may be the following:

- *NORMALIZABLES*: those mentions of chemical compounds and drugs that can be standardized with a unique identifier from a database.
- *NO NORMALIZABLES*: those mentions of chemical compounds and drugs that cannot be normalized
- *PROTEÍNAS*: includes peptides, proteins, genes, peptide hormones and antibodies.
- *UNCLEAR*: for plants, oils, essences, plant principles and general formulations/compositions of various compounds.

1.2 Track 2: Concept indexing

The objective of the second task was to assign a unique identifier to each concept detected in the previous task. The Snomed (Systematized Nomenclature of Medicine) (National Health Service, 2019) terminology was used for this. Snomed is an international standard distributed by the International Health Terminology Standards Development Organisation (IHTSDO)[1], an organisation to which Spain belongs as a member.

2 Data collection

The used corpus was Spanish Clinical Case Corpus (SPACCC). This corpus contains a manually classified collection of sections of clinical cases derived from open-access Spanish medical journals. The corpus contains a total of 1000 clinical cases and 396,988 words.

The organizers provided us 500 documents for training, 250 validation documents and finally, 3751 test documents. The final collection had a total of 16504 sentences, with an average of 16.5 sentences per clinical case. The SPACCC corpus contains a total of 396,988 words, with an average of 396.2 words per clinical case.

3 Methodology

Our group has participated in both sub-tasks proposed by PharmaCoNER. In each sub-task we have sent 4 runs. For the first sub-track we have created machine learning and deep learning approaches providing extra information with features. In the second sub-track we have used the outputs of the first task using a dictionary-based approach.

3.1 Track 1: NER offset and entity classification

3.1.1 Machine learning with CRF.

Conditional Random Fields (CRF) (Lafferty et al., 2001) are a probabilistic framework for the labeling or segmentation of sequential data. We used CRFsuite, the implementation provided by Okazaki (Okazaki, 2007), as it is fast and provides a simple interface for training and modifying the input features.

Similar to most machine learning-based systems, the token-level CRF requires a tokenization module at first. The tokenizer used is WordPunctTokenizer of the NLTK[2] library in Python.

Run 1. CRF + basic features. For the first experiment, we incorporate to CRF some basic features of each word such as isLower, isUpper, isTitle, isDigit, isAlpha, isBeginOfSentence and isEndIfSentece.

Run 2. CRF + basic features + features based on medical terminology. For this experiment, we decided to add a new feature to CRF using medical terminology to provide extra information for each word. This feature indicated if the word was contained in The Spanish Medical Abbreviation DataBase (AbreMES-DB), dictionary of chemicals, compounds, and drugs in Spanish (Nomenclator for Prescription) or Snomed in Spanish. Nomenclator for Prescription and AbreMES-DB are resources provided by the organizers and available on the workshop website[3]. On the other hand, Snomed was reduced using only the concepts of products and substances.

3.1.2 Deep learning with BiLSTM and CNN

For this sub-track, we present a hybrid model of bi-directional LSTMs and CNNs that learns both character and word-level features, based on the model of Chiu (Chiu and Nichols, 2016).

The first neural network, use the Convolutional Neural Network (CNN) to extract character features. For each word we employ a convolution and

[1] https://www.ihtsdo.org/

[2] https://www.nltk.org/

[3] http://temu.bsc.es/pharmaconer/index.php/resources/

a max layer to extract a new feature vector from the per character feature vectors such as character embedding and character type. This model also uses Bi-directional recurrent neural network with Long Short-Term Memory (BiLSTM) to transform word features into named entity.

The word embedding used for this task is Spanish Billion Word Corpus. The corpus for creating this embedding contain 1,000,653 words and the vector dimension is 300 (Cardellino, 2016).

Finally, the hyper-parameters used in these Neural Networks are those proposed by Chiu.

Run 3. BiLSTM + CNN + basic features. For this experiment, we used the model of Chiu (Chiu and Nichols, 2016). The Bi-LSTM take the concatenation of the output of CNN with the word embedding of each word.

Run 4. BiLSTM + CNN + basic features + features based on medical terminology. For the last experiment sent, we used the dictionary explained in the Run 2. In this case, the Bi-LSTM neural network used the CNN output, the word embedding of each word and if the word was contained in the new dictionary created.

3.2 Track 2: Concept indexing

Our approach is to create a large medical terminology dictionary to help map the named entity recognized in sub-track 1 with Snomed identifiers.The process for developing this task can be seen in Figure 1 and each step of the process followed is detailed below:

1. **Construction of the drug name dictionary.**

   At first, a drug name dictionary was build with different sources of knowledge related to chemicals, drugs and medicines. Our goal in creating this dictionary was to generate the maximum number of synonymous concepts. The sources of information used are explained below:

   (a) **Wikidata**: is a document-oriented database, focused on items, which represent topics, concepts, or objects. We downloaded from Wikidata all the elements that were *instances of*: *disease*, *gene*, *syndrome*, *protein*, *group or class of chemical substances*, *structural class of chemical compounds*, *medication*, *drug*, *chemical substance* and *chemical compound*. To make this query easy we used SPARQL and obtained the English and Spanish *alias* for each found object. We use both languages because most synonyms come with information in English.

   (b) **AbreMES-DB**: the Spanish Medical Abbreviation DataBase are extracted from the metadata of different biomedical publications written in Spanish, which contain the titles and abstracts.

   (c) **Nomenclator for prescription**[4]: is a medicine database designed to provide basic prescription information to healthcare information systems.

   (d) **Snomed**: we used the dictionary explained in Section 3.1.1, Snomed was reduced using only the concepts of products and substances in Spanish.

   (e) **Chemical symbols in Spanish**: abbreviated signs used to identify chemical elements of the periodic table and compounds.

   All these sources of information have something in common, they all contain synonyms, acronyms or other ways of referring to the same entity.

2. **Text pre-processing.**

   The second step of this architecture was to normalize the texts in order to make the matching of concepts. To do this we use the spaCy library because it is a free open source library for Natural Language Processing in Python. with *'es_core_news_sm'* module in Spanish. This pre-processing consists of:

   - Change the text to lower case.

   - Remove accents.

   - Use the lemma of each word with spaCy.

   - Remove punctuation marks.

   - Remove stop-word.

3. **Match with drug name dictionary.** At this point, we try to match the input text (recognized entity) with texts in the previously generated dictionary. If we can match them, then we will increase the list of possible synonyms

[4] https://cima.aemps.es/cima/publico/nomenclator.html

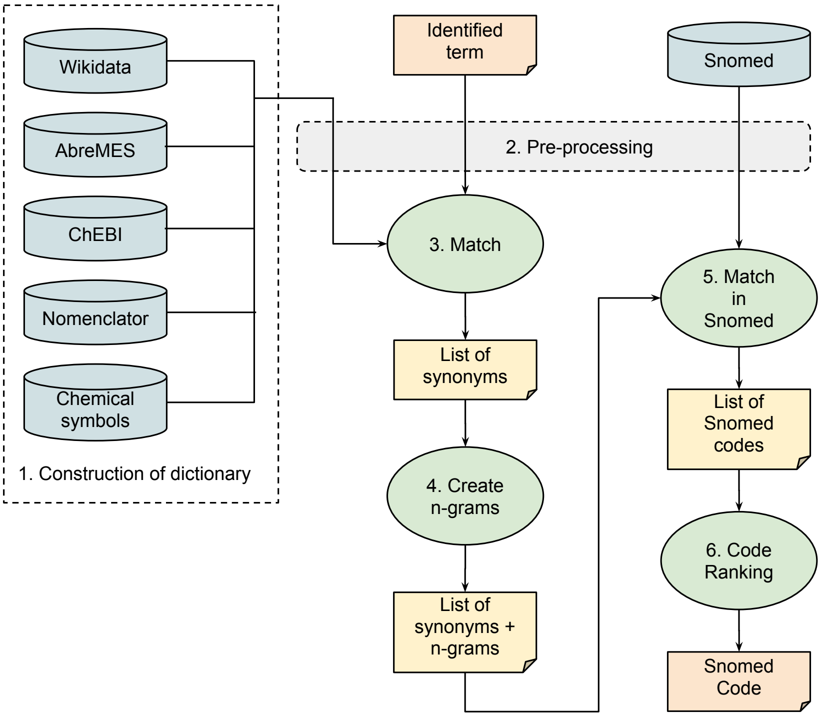

Figure 1: Snomed code indexing process.

to have more options to find the concept in Snomed.

4. **N-grams of terms.** After several tests with the development collection, we found that in some cases, the order of the multi-word concepts did not match well. For this reason we decided to create n-gram, where *n* is the size of the multi-word concept with all possible word combinations. This list of new n-grams was added to the list of possible synonyms of the concept.

5. **Match Snomed concepts.** Using the list of synonyms extracted from the previous steps, we try to match any of the synonyms with Snomed concepts. To get as many Snomed concepts as possible, we use a library called Hunspell[5].

   Hunspell uses a special dictionary to which we have added Snomed concepts. With this library, we use the *hunspell_suggest* function where we can get similar words to the given word. This function will return many concepts of Snomed so later we must choose one of them.

6. **Ranking of concepts by Levenshtein distance.** Finally, we must choose a single Snomed concept. For this we use the Levenshtein distance (LD). LD is a measure of the similarity between two strings, which we will refer to as the source string and the target string. The distance is the number of deletions, insertions, or substitutions required to transform each synonym included with a Snomed concept. Lastly, we chose the Snomed concept that has the least distance with the input text.

   In Table 1 we can see some examples of how the resources and tools applied in the architecture can contribute to the achievement of Snomed concept mapping.

4 Results

4.1 NER offset and entity type classification.

The first evaluation consist in the classical entity-based or instanced-based evaluation that requires that system outputs match exactly the beginning and end locations of each entity tag, as well as match the entity annotation type of the gold standard annotations.

The results obtained by our team for this subtrack are shown in Table 2.

[5] http://hunspell.github.io/

Resource	**Input text**	**Snomed Term**	**Snomed Code**
Wikidata	adriamicina	doxorrubicina	372817009
Chemical symbols	Na	sodio	39972003
AbreMES-DB	Hb	hemoglobina	38082009
Hunspell Library	6-Metil-Prednisolona	metilprednisolona	116593003

Table 1: Examples of Snomed concept indexing.

Run	Precision	Recall	F1
1	**0.92602**	0.61835	0.74154
2	0.88507	**0.69815**	**0.78058**
3	0.84404	0.64929	0.73397
4	0.85992	0.69653	0.76965

Table 2: Results of Track 1. NER offset and entity classification.

Run	Precision	Recall	F1
1	**0.87879**	0.55849	0.68295
2	0.85207	**0.63267**	**0.72616**
3	0.8335	0.57846	0.68295
4	0.82887	0.6184	0.70833

Table 3: Results of Track 2. Concept indexing.

In these results we can see that applying features in both methods (*Run 2* and *Run 3*) improves the base model (*Run 1* and *Run 3*). In the case of CRF the precision decreases but the recall increases and finally the F1 measure improves from 74% to 78%. For the RNN, in all the measures the use of new features improves, obtaining 76% of F1. For future occasions we will continue to exploit the use of new features in the different strategies.

4.2 Concept indexing.

For this sub-track the main objective was to index each document of the previous task and each concept detected with a unique Snomed code. Table 3 shows the evaluation of the systems for this sub-track.

The results are concordant to the previous task, we use the output of each Run of sub-track 1 to index the concepts detected with Snomed codes. For this reason we get 72% F1 score in *Run 2*, and in *Run 1* we obtain 87% precision.

We are still analyzing the results obtained, in this way, in the future we will know how we can improve the task of indexing with terminologies, which dictionaries to use or which Natural Language Processing (NLP) tools we can apply.

5 Conclusion and future work

The SINAI group presents its participation in the PharmaCoNER challenge. The first task was to find chemical and drug mentions in the text and assign a specific label. In the next task, the main objective was to index each found concept with a Snomed code.

For sub-track 1 we have developed four systems with different machine learning and deep learning approaches adding some relevant features obtained from the Snomed terminology in Spanish. The goal of sub-track 2 is to assign a unique identifier to each detected concept of sub-track 1 and for this we have developed a system based on a large dictionary of medical terminology according to the task, this dictionary provided us with a long list of synonyms for each entity to match with a Snomed code.

The results obtained have been as expected, adding extra information from Snomed terminology helps classifiers to detect relevant entities within medical texts. On the other hand, apply NLP techniques and tools and the creation of a medical dictionary has contributed to find synonyms for later assigning a single Snomed code. Using our methodology, we found the correct code for example to the input text *IgG* although in Snomed this concept is described as *immunoglobulin G*.

In future works we will continue working on machine learning approaches and different features of improvement. Specifically, we will create more sophisticated Neural Networks and explore different embeddings in Spanish. Normalization plays an important role in this track so we will use NLP to continue improving.

Acknowledgments

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):161.

Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. 2008. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266.

Cristian Cardellino. 2016. Spanish billion words corpus and embeddings.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST)*, pages 1–X, Hong Kong, China. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel 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*, pages 70–75. Citeseer.

Corinna Kolárik, Roman Klinger, Christoph M Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. 2008. Chemical names: terminological resources and corpora annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, 7(1):S1.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Ulf Leser and Jörg Hakenberg. 2005. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4):357–369.

Hongfang Liu, Zhang-Zhi Hu, Jian Zhang, and Cathy Wu. 2005. Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105.

Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi. 2005. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC bioinformatics*, 6(1):S8.

National Health Service. 2019. SNOMED International. `http://www.snomed.org/`.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. Calbc silver standard corpus. *Journal of bioinformatics and computational biology*, 8(01):163–179.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):S3.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2003. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 41–48. Association for Computational Linguistics.