Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 149-153 recibido 18-12-2021 revisado 22-01-2022 aceptado 29-01-2022

Biomedical entities recognition in Spanish combining word embeddings

Reconocimiento de entidades biomédicas en español combinando word embedding

Pilar López-Úbeda
SINAI, Department of Computer Science, CEATIC, Universidad de Jaén,
Campus Las Lagunillas s/n, 23071, Jaén (Spain)
plubeda@ujaen.es

Abstract: This is a summary of the Ph.D. thesis written by Pilar López Úbeda at Universidad de Jaén under the supervision of PhD. M. Teresa Martín Valdivia, Ph.D. L. Alfonso Ureña López and PhD. Manuel Carlos Díaz Galiano. The defense was held in Jaén on April 22, 2021. The doctoral committee was integrated by PhD. Rafael Muñoz Guillena from Universidad de Alicante, PhD. Paloma Martínez Fernández from Universidad Carlos III de Madrid, and Manuel Montes y Gómez from National Institute of Astrophysics, Optics and Electronics (Mexico). The thesis obtained the grade of *Summa Cum Laude* and the international mention.
Keywords: Natural language processing, Spanish corpora, biomedical entity recognition, word embeddings, deep learning.

Resumen: Este es un resumen de la tesis doctoral realizada por Pilar López Úbeda en la Universidad de Jaén bajo la dirección de los doctores Dña. M. Teresa Martín Valdivia, D. L. Alfonso Ureña López y D. Manuel Carlos Díaz Galiano. La defensa se realizó en Jaén el 22 de abril de 2021. La comisión de doctorado estuvo integrada por el PhD. Rafael Muñoz Guillena de la Universidad de Alicante, la PhD. Paloma Martínez Fernández de la Universidad Carlos III de Madrid, y Manuel Montes y Gómez del Instituto Nacional de Astrofísica, Óptica y Electrónica (México). La tesis obtuvo la calificación de Sobresaliente Cum Laude y mención de doctorado internacional.
Palabras clave: Procesamiento del lenguaje natural, corpus en español, reconocimiento de entidades biomédicas, representación de palabras, aprendizaje profundo.

1 Introduction

One of the main purposes of clinical text mining is the possibility to process and analyze the large volumes of textual information contained in medical records. Through this treatment of the information, we attempt to answer questions such as, which patients presented a certain condition? What kind of conditions were used to detect the disease? What were the results of the tests performed? What was the treatment given? These questions could seem quite simple for some medical professionals, but they become extremely complex when managed automatically by computational systems.

In the biomedical domain, we can find large collections of free textual information (medical reports, Electronic Health Records - EHR, scientific papers, among others) that contain very relevant data that need to be studied in depth. However, current health information systems are not prepared to analyze and extract this knowledge due to the time and cost involved in processing it manually. The field of artificial intelligence known as Natural Language Processing (NLP) is being applied to medical documents to build applications that can understand and analyze this huge amount of textual information automatically (Friedman y Johnson, 2006)

Many researchers in the NLP field focus on the area of Information Extraction (IE) in the biomedical domain to address these challenges. IE systems take natural language text as input and produce structured infor-

mation specified by certain criteria and that is relevant to a particular application. Depending on the different inputs of IE systems and expected outputs, many sub-tasks can be defined such as Named Entity Recognition (NER).

In this thesis, we focus on information extraction from Spanish biomedical texts, more specifically, on the NER task. Spanish has more than 480 million native speakers and nowadays there is a worldwide interest in processing medical texts in this language. With this study, we aim to advance the task of biomedical NER in this relevant language and thus answer the above-mentioned questions (López Úbeda, 2021).

To accomplish this study, we propose a methodology based on deep learning. Furthermore, different word embeddings are used in combination to obtain a better representation of each word. With this approach, we aim to achieve the desired final goal: to recognize biomedical entities accurately in different scenarios.

1.1 Motivation

Over the years, the recognition of biomedical entities has motivated the scientific community to continue developing automatic systems to facilitate the extraction of medical knowledge. NER is a difficult task to solve that can help in many other medical-related systems such as those presented below:

- **Clinical decision support**. Automated NER systems can provide real-time results, which means that entities such as diseases can be detected immediately. This evidence can be used to help professionals identify emerging health problems, for instance, to alert them to the presence of certain unexpected findings (López-Úbeda et al., 2020b).
- **Entity representation**. In the NER task, different words can have similar meanings. This problem is caused by the multiple ways in which a particular entity can be represented and written. For instance, "*adriamicina*"(adriamycin) and "*doxorrubicina*"(doxorubicin) refer to the same drug widely used in cancer chemotherapy.

  On the other hand, an acronym does not always have a unique description, it can be interpreted as two different entities depending on the context. For instance, in Spanish, PCR can be referred to "*parada cardiorrespiratoria*"(cardiorespiratory arrest) or "*Reacción en Cadena de la Polimerasa*"(Polymerase Chain Reaction). Finally, as we see in the examples, biological entities may also have multi-word names, so the problem is additionally complicated by the need to determine name boundaries and resolve overlap of candidate names.
- **Basis for other NLP tasks**. Biomedical entity recognition serves as the basis for many other crucial areas of information management, such as classification tasks, question answering, information retrieval, and text summarization (López-Úbeda et al., 2020a). For instance, the use of NER becomes important for analyzing the clinical text and obtaining the most relevant tags in each report, allowing the classification of documents.
- **Extracting structured information**. Biomedical NER is a task that facilitates medical professionals in structuring reports contributing to solutions such as providing a summary of patient conditions or serving as a tool to organize the documentation of the physician's decision-making process, plan development, and patient outcomes.

1.2 Objectives

The main objective of this thesis focuses on the study, analysis, and development of NLP techniques and tools for the NER task in the biomedical domain in Spanish. Specifically, it focuses on the study and applicability of different combinations of word embeddings as word representations.

This general objective has been defined through the following specific objectives:

- Collect resources available in Spanish annotated with biomedical entities used in different challenges.
- Study and select the existing word embeddings in Spanish serving as input to the network.
- Propose a deep learning-based method for NER in the biomedical domain that

can take a combination of different word embeddings as input.

- Generate a new word embedding for Spanish focused on the biomedical domain to see how effective it is in comparison to existing ones.
- Evaluate the performance of the proposed method on the NER problem using three application scenarios: pharmacological domain, oncological domain, and knowledge discovery in biomedical texts.
- Conduct a results analysis comparing our system with the state-of-the-art.
- Perform an error analysis to understand the capabilities and drawbacks of our system.
- Identify open issues from the conclusions in order to propose future research.

1.3 Hypotheses

In this thesis, we address the problem of biomedical entity extraction in Spanish through deep learning and combinations of word embeddings using NLP methods. Based on the objectives set out above, our general hypothesis can be summarized as follows:

NLP techniques applied to the NER task can improve biomedical systems.

However, since this hypothesis is very ambitious, we have decided to subdivide it into three specific hypotheses:

Hypothesis 1 *(H1). Deep neural networks in NLP leverage the advantage of existing relevant information from the Spanish biomedical textual data and the NER task, outperforming models that do not integrate this information properly.*

Hypothesis 2 *(H2). Combining different types of word embeddings by concatenating each embedding vector to form the final word vectors is an important part of the biomedical entity recognition task. The probability of recognizing a specific entity in a text should increase as optimal representations of that word are combined because they are more comprehensively represented and integrate relevant knowledge.*

Hypothesis 3 *(H3). Integrating domain-specific knowledge into the training corpus can be beneficial for improving the quality of word embeddings. Thus, this resource provides a more accurate representation of words in a particular context and domain.*

2 Thesis outline

This thesis is organized into six chapters and an appendix as described below:

Chapter 1 contains an introduction explaining the motivation and objectives that led us to carry out the study. Furthermore, we have presented the hypotheses with the research questions we intend to solve and the methodology we will carry out.

Chapter 2 presents an overview of the methodologies based on ML commonly used in the NER task and which are necessary to understand the later parts of this thesis.

Chapter 3 summarizes previous work on NLP tasks based on ML in the biomedical domain and shows an extensive literature review of the NER task with regard to present state-of-the-art studies. Since the interest of this thesis lies in word representation, this chapter details the review of existing methods for word representations up to the moment.

Chapter 4 describes the proposed model to solve the biomedical entity extraction problem. After an extensive review of previously applied methodologies, we propose an approach based on a Bidirectional Long Short-Term Memory (BiLSTM) neural network with a final CRF layer.

Chapter 5 presents the experimentation carried out using the approach proposed. The experimental framework was developed in three scenarios belonging to different biomedical sub-domains including pharmacology, oncology, and knowledge discovery. For each scenario, this chapter contains a description of the problem, the dataset, the results obtained, error analysis, and a discussion.

Chapter 6 contains our conclusion where we summarize our findings and main contributions. Moreover, this chapter provides an outlook into the future, the publications derived from the study, and the research results transferred.

Finally, **Appendix A** contains additional results of the NER task performance in the different scenarios proposed.

3 Main contributions

This research has carried out a series of studies, analyses, and development of NLP tech-

niques designed to address the task of NER in Spanish biomedical texts. This has resulted in several contributions to the research that we have considered on the basis of the hypotheses.

To support hypothesis H1, we can summarize the following contributions:

Contribution 1 We have investigated and implemented different machine learning approaches. First, we have reviewed unsupervised models and then advanced to supervised models using traditional models such as CRF and deep neural networks.

Contribution 2 In our review of the state-of-the-art in deep learning, we have exposed what kind of architectures are used by the scientific community interested in NER.

Contribution 3 We have proposed a model based on neural networks. Specifically, the architecture is composed of a BiLSTM network and a CRF layer (López-Úbedaa et al., 2020).

To support hypothesis H2, we provide the following contributions:

Contribution 4 In our review of related literature, we have found that word representations and, more specifically, word embeddings are the most commonly used methods.

Contribution 5 We have selected different word embeddings to include in the neural network to address the NER problem in biomedicine.

Contribution 6 We have presented a model based on a combination of word embeddings for a more exhaustive representation of the words, thus improving entity identification systems.

The contributions that support hypothesis H3 can be summarized as follows:

Contribution 7 We have collected an unannotated corpus by extracting documents from different corpora and websites related to the biomedical domain, obtaining a vocabulary of 1,704,151 words.

Contribution 8 We have generated new word embeddings specifically for the biomedical domain in Spanish (López-Úbeda et al., 2020c).

Acknowledgements

This work has been partially supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project [RTI2018-094653-B-C21], and the Government of Andalusia [PY20_00956].

Bibliografía

Friedman, C. y S. B. Johnson. 2006. Natural language and text processing in biomedicine. En *Biomedical Informatics*. Springer, páginas 312–343.

López Úbeda, P. 2021. Biomedical entities recognition in spanish combining word embeddings.

López-Úbeda, P., M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, y M. T. Martín-Valdivia. 2020a. Covid-19 detection in radiological text reports integrating entity recognition. *Computers in Biology and Medicine*, 127:104066.

López-Úbeda, P., M. C. Díaz-Galiano, T. Martín-Noguerol, A. Ureña-López, M.-T. Martín-Valdivia, y A. Luna. 2020b. Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches. *Expert Systems with Applications*, 160:113647.

López-Úbeda, P., M. Dıaz-Galiano, M. T. Martın-Valdivia, y L. A. Ureña-López. 2020c. Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. *Proceedings of IberLEF*.

López-Úbedaa, P., J. M. Perea-Ortegab, M. C. Díaz-Galianoa, M. T. Martín-Valdiviaa, y L. A. Ureña-Lópeza. 2020. Sinai at ehealth-kd challenge 2020: Combining word embeddings for named entity recognition in spanish medical records.