ISSN 2007-9737

Generation of Feature Vectors for Identifying Medical Entities in Spanish

Gabriela A. García-Robledo[1], Alma Delia Cuevas-Rasgado[1,*], Maricela Bravo[2], José A. Reyes-Ortiz[2]

[1] Universidad Autónoma del Estado de México, Centro Universitario Texcoco, Mexico

[2] Universidad Autónoma Metropolitana, Unidad Azcapotzalco, Mexico

ggarciar015@alumno.uaemex.mx, adcuevasr@uaemex.mx, {mcbc, jaro}@azc.uam.mx

Abstract. Natural Language Processing (NLP) encompasses a range of high-impact techniques to enable computers to interact with humans more naturally. One such technique is the extraction of entities, which allows computers to identify relevant information within a text. This paper presents a methodology for recognizing medical entities within texts written in Spanish. The methodology combines syntactic, semantic and contextual features at the word level. The main aim of the feature-based approach is to identify drug, anatomy, and disease entities. A training evaluation was conducted on two machine learning algorithms, with an precision of 98% on an external set. In addition, an precision check was performed for each medical class.

Keywords. Information extraction, named entity recognition, natural language processing.

1 Introduction

Information extraction is a technique in Natural Language Processing (NLP) utilized to select relevant information within a text. Named entity recognition and classification is an application that facilitates the identification of knowledge within a set of texts by focusing on key concepts related to a specific domain.

Entities are objects in the real world that can be used to determine the subject matter of a text. The term ”named entity” was first used and defined in [6] to identify the names of organizations, people, and geographic locations in texts, as well as monetary expressions of time and percentages. This process is known as Named Entity Recognition (NER).

Initially, entity recognizers were employed in general contexts where the names of persons, dates, organizations, or locations were identified. However, recent advances have focused on searching for terms within a specific domain to understand a text's content. Entity recognizers have been used in the medicine, legal, and geology domains. In certain instances, these applications extract relationships between two entities, which facilitates a more efficient information search.

The recognition process of entities requires that each word be mapped to a form that can be understood by a computer, which, in this case, is a numerical representation. In order to assign a set of values to each word and make sense of the text, feature-based approaches are employed to differentiate between relevant and non-relevant terms.

A word vector may comprise various features, including part of speech (POS tag), word gender, the presence of prefixes or suffixes.

This paper presents a methodology for developing a feature generator that integrates syntactic, semantic, and contextual information

to recognize medical entities such as diseases, anatomy and drugs from the abstracts of scientific articles written in Spanish. The proposed approach leverages traditional machine learning algorithms, enabling high precision while avoiding the substantial computational cost typically associated with deep learning models.

Furthermore, the methodology efficiently uses expert-validated resources in the medical domain, such as, BioASQ and CoWeSe. The remainder of the article is organized as follows. Section 2 presents significant advances in the recognition of named entities. Section 3 presents the methodology used, while Section 4 displays the results and evaluations performed. Finally, Section 5 presents the conclusions and outlines future work.

2 Related Work

A traditional entity identifier classifies entities into general categories such as organizations, location, person, and time. In [5], they locate entities of person, organization, and location in English, very similar to what they do in [14] and [17] in Chinese with Large Language Model (LLM), which have recently become an innovative approach to named entity identification. Recognizing entities of a specific domain allow to go deeper into specialized information. For example, in [11], it is applied in geology to detect entities of geological time, geological structure, stratum, rock, mineral and physical space from an ontology of geological reports. In the field of medicine, similar advances can be observed.

For instance, in [7], different entities that belong to the same type are identified, this article presents an autocorrection algorithm and a dictionary of drugs with their commercial and product names. In [15], the authors identify specific entities such as proteins within Spanish texts, employing definitions derived from previous queries, biomedical texts, and pretrained models, which have emerged as a promising approach for recognizing specific medical entities, including proteins, chemicals, diseases, and genes.

Several recent works, such as [16], [1], and [4], have demonstrated the efficacy of this approach. In their implementation, they incorporate additional features, including POS-Tagging, using gazetteers of entities, affixes, and dependency trees. In this domain, there are also works employing the LLM approach. In [9], word embeddings, POS-tagging, and sequence features are utilized despite being in English. In contrast, development in Spanish uses BERT and a medical dataset but only to identify morphologies of tumor cell. In a more recent contribution, [10] examines the integration of recurrent neural networks with LLMs to extract clinical entities from Spanish electronic health records, constituting a substantial advancement in the development of medical NER systems for Spanish. The model proposed in this paper for recognizing medical entities in Spanish, including parts of anatomy, diseases, and drugs. The model combines syntactic, morphological, semantic, and contextual features.

3 Methodology

Generating a feature vector for medical entity recognition requires applying NLP techniques for data preparation, a medical term labeling for validation, and a medical resource validated by domain experts to serve as the specialist. Figure 1 illustrates the methodology implemented in this paper.

Processing is performed on a set of medical texts in Spanish to eliminate special characters in each one. The clean text set is entered into a medical term labeling system in IOB(Inside-Outside-Beginning) format. This labeling uses a lexicon developed by domain experts to determine which word belongs to a disease, anatomy, or drug. The output of this term labeling process is utilized to train machine learning algorithms and generate syntactic, semantic, and contextual feature vectors.

Training with this set of features yields a specialized model capable of determining whether an entity is ANAT (anatomy), MEDI (drug), or ENFE (disease). The efficacy of this methodology is assessed using a distinct set of medical data that differs from those employed during the training phase.

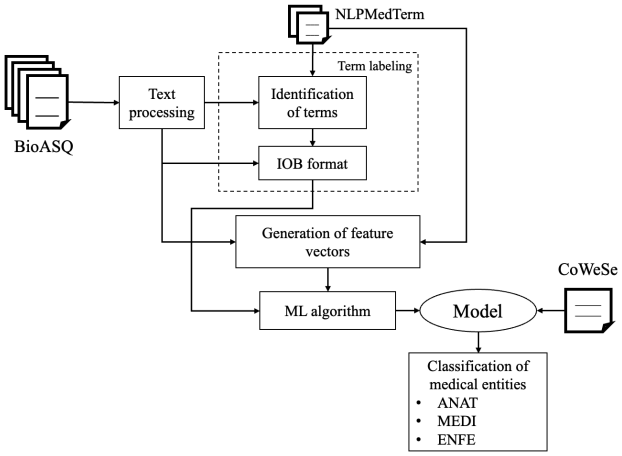

Fig. 1. Methodology for the identification of medical entities

3.1 Dataset

The datasets used for the extraction of medical features and the term labeling are explained in this section.

— **NLPMedTerm.** The deliverable 1[3] is a medical lexicon for the Spanish language created by experts in the domain. This resource is used for term labeling and the generation of vectors as it contains medical features that contribute to domain-specific knowledge. The resource was obtained from the NLPMedTerm (Natural Language Processing for Medical Terminology) project of the Autonomous University of Madrid. This deliverable comprises 100,886 medical terms that experts have evaluated.

— **BioASQ Challenge.** The abstracts of scientific articles in JSON format from the BioASQ Challenge[2] are used dedicated to biomedical semantic indexing and large-scale query answering. This competition includes, among its tasks, the recognition of the medical entity named MedProcNER, which is co-organized with the Barcelona Supercomputing Center (BSC). The methodology involved the use of 13,067 texts from version 2021 for the purposes of training and testing the algorithms. The values of the keys ”id” and ”abstractText,” which serve as identifiers and abstracts of medical scientific articles, were extracted from the dataset.

— **Spanish Biomedical Crawled Corpus (CoWeSe).** The medical texts used to evaluate the trained models are a sizeable biomedical corpus in Spanish comprising 1.5 million documents in plain text and preprocessed. The corpus was generated by a text mining crawl conducted by the Text Mining Unit at the Barcelona Supercomputing Center[8] on 3,000 Spanish domains in 2020. The text extraction process only considers HTML tags of paragraph and heading types. The domains involved in the text mining process include medical and scientific communities, medical journals, research centers, pharmaceutical companies, informative websites about health issues, patient associations, personal blogs from healthcare professionals, hospital websites, and public health organizations.

3.2 Text Processing

The text processing phase allows for identifying relevant information, which can then be used to inform more effective training outcomes. The punctuation marks (dot, semicolon, comma, underscore, forward slash, brackets, parentheses, colon, quotation marks, apostrophe) have been removed. Regular expressions are employed to eliminate digits and digits combined with special characters (dot, forward slash, plus, minus, hashtag) and characters combined with signs (greater than, less than, per cent, and sign). The signs were selected based on an analysis of the scientific articles, with less relevance to the possible terms identified.

3.3 Term Labeling

The labeling process employs a medical term identification format (IOB). This phase requires the processed articles from which the information is extracted and the NLPMedTerm resource. The

Table 1. Semantic groups

Entity	Semantic group
Anatomy	Body area
	Body part or organ
	Body system
Drug	Pharmacological substance
	Antibiotic
Disease	Disease or syndrome
	Injury or poisoning
	Anatomical abnormality
	Virus
	Bacterium

```
<data>
  <documento id="ibc-194909">
    <termino id="T1">
      <nombre>covid-19</nombre>
      <inicio>208</inicio>
      <final>216</final>
      <clase>ENFE</clase>
      <multipalabra>False</multipalabra>
    </termino>
    <termino id="T2">
      <nombre>covid-19</nombre>
      <inicio>991</inicio>
      <final>999</final>
      <clase>ENFE</clase>
      <multipalabra>False</multipalabra>
    </termino>
  </documento>
  <documento id="ibc-ET6-1764">
    <termino id="T1">
      <nombre>antiinflamatorio</nombre>
      <inicio>1128</inicio>
      <final>1144</final>
      <clase>MEDI</clase>
      <multipalabra>False</multipalabra>
    </termino>
```

Fig. 2. The XML file segment was generated with the identified entities

entities identified in the texts are diseases, drugs, and anatomy.

3.3.1 Identification of Terms

NLPMedTerm deliverable [3] was employed to identify the terms in the text. This resource comprises approximately 127 semantic groups, of which 11 were utilized to meet the characteristics of the entities identified, which are of the disease, drug, and anatomy. The semantic groups utilized are presented in Table 1.

In this phase, an XML file is generated to record the terms identified for each text. The file contains a list of terms for each scientific article, together with the following information:

— Document. The current identifier of each document (ID) was derived from the set of input scientific articles.

— Name. Word with which the identified term appears.

— Start of the term. The initial position of the term in the text.

— Start of the term. The initial position of the term in the text.

— End of term. The final position of the term given in the text.

— Class. The entity is defined according to its semantic group, which can be ANAT (anatomy), ENFE (disease), or MEDI (drug).

— Multiword. Displays a boolean value *True* if the current term contains more than one word. The value is *False* otherwise.

This section is designed to train dedicated machine learning models to identify medical entities. Additionally, it serves to validate the results of each model. Figure 2 illustrates a segment of the generated XML containing the identified terms from all input documents.

3.3.2 IOB Format

The annotation format is necessary to determine the number of classes identified and to determine whether the entity consists of more than one word. It indicates the parts of the entity. The format utilized in this methodology is IOB[12], where B-*EntityType* signifies the beginning of the entity, I-*EntityType* denotes the center or end of the entity, and O is used when the entity does not belong to any class. This format is applied at the word level and is used by learning algorithms

Table 2. Example of IOB format

Token	IOB format
potente	O
inhibidor	B-MEDI
de	I-MEDI
la	I-MEDI
proteasa	I-MEDI
principal	O
del	O
sarscov-2	B-ENFE

to identify the start and end of entities. For example, the text fragment *"potente inhibidor de la proteasa principal del sarscov-2"* (potent sarscov-2 major protease inhibitor) the term *"inhibidor de la proteasa"* (protease inhibitor) is a drug, and its class is MEDI. The "sars-cov2" disease has a class of ENFE. The format IOB of the text fragment is shown in Table 2.

3.4 Generation of Feature Vectors

For each token in every text, a feature vector generator is automatically created. Each vector comprises eight numerical values representing semantic, morphological, and syntactic features.

This combination is intended to incorporate contextual information about words to improve the identification of terms. For example, the dependency of a term on its parent node may vary depending on the context. The semantic groups corresponding to features one or two, extracted from NLPMedTerm, are listed in Table 1. The subsequent section will describe the features corresponding to each position in the generated vectors.

1. **Multiword expression belonging to a medical semantic group.** A value of one (1) is assigned if the token is part of a medical semantic group and is the initial element of the term. If the token is part of a medical semantic group and is part of a multiword expression without being the start of the term, a value of two (2) is assigned. In the absence of these conditions, a value of zero is assigned.
2. **Medical semantic group.** Value that indicate the semantic group of the current token; otherwise, a zero is assigned. The possible values for this feature are presented in Table 3.
3. **Dependency type.** A value assigned the dependency relationship of the current token from the Hash Table is determined by Spacy[13].
4. **Dependency type of the syntactic parent node.** The value of the dependency relationship between the current token and the syntactic parent node.
5. **POS tag.** The value represents the general grammatical component of the Part of Speech (POS) tag set utilized by Spacy[13].
6. **POS tag of the syntactic parent node.** A value representing the grammatical component of the syntactic parent node, as defined by the Spacy POS tag set [13].
7. **Child nodes to the left.** The value represents the existence of child nodes to the left of the current token. This value is one (1) if the condition is met and zero otherwise. The window under consideration is two tokens to the left.
8. **Child nodes to the right.** The Value represents the existence of child nodes to the right of the current token. A value of one (1) is assigned in the event of existence, while zero is assigned in the absence. The window of consideration is two tokens to the right.

3.5 Machine Learning Algorithm

The proposed feature combination was initially evaluated through a series of experiments employing classification algorithms based on decision trees and random forests to recognize medical entities in Spanish-language texts. Classical entity classification techniques were applied to assess the feature combination's

Table 3. Medical semantic group

Value of the semantic group	Semantic group
1	Body location or region
2	Body part, organ or organ component
3	Body system
4	Pharmacologic substance
5	Antibiotic
6	Clinical drug
7	Disease or sindrome
8	Injury or
9	Anatomical abnormality
10	Virus
11	Bacterium

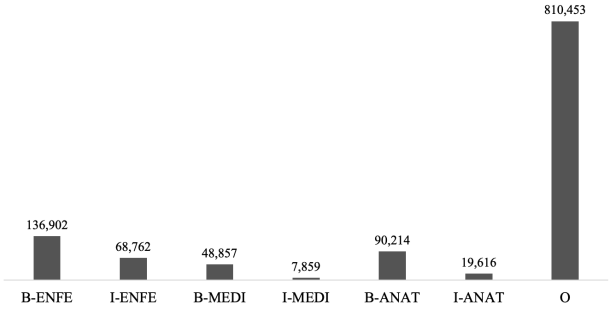

Fig. 3. Class distribution of the BioASQ Challenge partial

performance. One million one hundred eighty-two thousand six hundred sixty-three feature vectors were generated from 13,067 abstracts of BioASQ Challenge.

The articles were annotated using terms provided by the NLPMedTerm resource, which has been validated by experts in the medical domain. It ensures the quality and consistency of the applied labels.

One of the main challenges during the training and validation of the algorithms was the class imbalance in the dataset, which is a common but significant issue in entity recognition tasks. Although medical abstracts contain specialized terminology, the overall proportion of medical terms is significantly lower than the total words. It is due to many general-purpose vocabularies, such as connectors, verbs, auxiliary expressions, and methodological descriptions, that are essential for structuring the content and conveying complete information. Consequently, general words outnumber domain-specific terms in most abstracts.

Figure 3 illustrates the class imbalance in the dataset. In this figure, the class 'O' represents general words not part of medical entities classified under ANAT, ENFE, or MEDI. An imbalance is also observed between the B-*term* and I-*term* labels, which can be attributed to the fact that most medical entities in the dataset consist of single-word terms.

Despite the class imbalance, the dataset was divided into an 80% training set and a 20% validation set. The results were promising: The decision tree algorithm achieved an precision of 97.84%, while the random forest algorithm slightly outperformed it with an precision of 97.86%.

4 Results

As a preliminary evaluation, the performance of the proposed feature combination was analyzed by training classification models using the feature vectors generated with the Decision Tree and Random Forest algorithms. A five-fold cross-validation process was conducted to ensure the results' robustness and generalizability. Despite the inherent randomness in the data partitioning, the precision metric remained stable across the different experiments, demonstrating the proposed approach's consistency.

The exclusive use of precision as the evaluation metric is motivated by the dataset's characteristics, which exhibit a significant class imbalance. Precision provides a more appropriate measure of the model's ability to correctly identify positive instances, which is particularly important in the case of medical entities, as they are often underrepresented.

Table 4 presents the class-wise precision distribution for the Decision Tree model across the five experiments conducted using the BioASQ dataset. Similarly, Table 5 shows the class-wise precision results obtained by the Random Forest model under the same experimental conditions.

The models generated using both algorithms were evaluated on a medical text dataset distinct

Table 4. Precision by class using Decision Tree model in the BioASQ dataset

Class	# iteration				
	1	2	3	4	5
B-ENFE	100%	100%	100%	100%	100%
I-ENFE	96.5%	97%	96.8%	96.7%	96.8%
B-MEDI	100%	100%	100%	100%	100%
I-MEDI	18.1%	17.2%	18.9%	16.8%	18.7%
B-ANAT	100%	100%	100%	100%	100%
I-ANAT	7.7%	7.8%	7.3%	8%	8%
O	100%	100%	100%	100%	100%
Total	**97.69%**	**97.76%**	**97.6%**	**97.73%**	**97.73%**

Table 5. Precision by class using Random Forest model in the BioASQ dataset

Class	# iteration				
	1	2	3	4	5
B-ENFE	100%	100%	100%	100%	100%
I-ENFE	96.8%	97.3%	97.2%	97%	97.4%
B-MEDI	100%	100%	100%	100%	100%
I-MEDI	19%	17.5%	19.7%	17.6%	19.5%
B-ANAT	100%	100%	100%	100%	100%
I-ANAT	7.2%	7%	7%	7.1%	7.4%
O	100%	100%	100%	100%	100%
Total	**97.71%**	**97.77%**	**97.69%**	**97.76%**	**97.76%**

from the one used for training. For this phase, 5,000 texts from the CoWeSe corpus were selected and annotated using the terms provided by NLPMedTerm, a resource validated by experts in the medical domain, thus ensuring the consistency and reliability of the annotations.

A total of 20,369 feature vectors were generated, achieving a precision of 98.8% for both the Decision Tree and Random Forest models.

Table 6 presents the class-wise precision distribution for the Decision Tree model evaluated on the CoWeSe dataset. Similarly, Table 7 shows the class-wise precision results for the Random Forest model under the same conditions.

The performance difference between the two models was minimal in both evaluations. However, the Random Forest model demonstrated slightly superior performance, as indicated by marginally higher precision scores.

Table 6. Distribution of precision by class of Decision Tree model in the CoWeSe dataset

Class	Vectors classified correctly	Total vectors for evaluation	Class precision
B-ENFE	2,285	2,285	100%
I-ENFE	988	1,010	97.8%
B-MEDI	855	855	100%
I-MEDI	15	86	17.4%
B-ANAT	1,593	1,593	100%
I-ANAT	11	157	7%
O	14,383	14,383	100%
Total	**20,130**	**20,369**	**98.83%**

Table 7. Distribution of precision by class of Random Forest model in the CoWeSe dataset

Class	Vectors classified correctly	Total vectors for evaluation	Class precision
B-ENFE	2,285	2,285	100%
I-ENFE	993	1,010	98.3%
B-MEDI	855	855	100%
I-MEDI	14	86	16.2%
B-ANAT	1,593	1,593	100%
I-ANAT	8	157	5.1%
O	14,383	14,383	100%
Total	**20,131**	**20,369**	**98.83%**

Table 8 presents a comparative analysis between the proposed approach and existing methods in the medical domain. The performance differences observed among the studies are mainly due to variations in the entities considered and the datasets used for evaluation.

It is crucial to note that the proposed model does not simplify entity classification, which is inherently complex. The improvement in classification performance is due to syntactic, semantic, and contextual features, which enhance the model to categorize entities accurately.

Table 8. Comparison of related works

Method	Precision
Bin, J. et al.[7]	91.2%
Yoon, W. et al.[16]	85.1%
Xiong, Y. et al.[15]	92.2%
Armengol-Estapé, J. et al.[1]	92.3%
Moreno-Barea, F. et al.[10]	96.3%
Our work	**97.7%**

5 Conclusion and Future Work

This paper presents a methodology for identifying medical entities in Spanish texts using a word-level feature combination approach. In addition, a training and validation process was carried out using different datasets to assess the effectiveness of the proposed method.

The results obtained using traditional classification algorithms, such as Decision Trees and Random Forests, were highly promising regarding model evaluation. These algorithms were used as a preliminary assessment starting point, demonstrating the methodology's potential for medical entity recognition tasks. The proposed approach offers the following main contributions: (a) a medical entity tagger that follows the IOB format and generates XML-formatted output; (b) a feature vector generator that integrates syntactic, semantic, and contextual information; (c) functional validation of the approach on texts different from those used in training, employing two traditional machine learning algorithms; (d) a preliminary evaluation on imbalanced datasets, with results that support the feasibility of the proposed method.

As future work, the following objectives are considered: (a) apply the generated feature vectors to neural networks, deep learning models, and other classical algorithms in order to conduct a comparative performance analysis; (b) use the recognized medical entities for semantic relationship extraction; (c) incorporate new features into the vectors to address ambiguity in medical terminology, since some words may correspond to multiple categories; (d) enhance medical terminological resources through expert validation, to classify new entity within the biomedical domain; (e) integrate additional evaluation metrics such as recall, F1-score, and class-specific performance indicators to provide a more comprehensive assessment of model effectiveness, particularly in the context of imbalanced datasets.

References

1. **Armengol-Estapé, J., Soares, F., Marimon, M., Krallinger, M. (2019).** PharmacoNER tagger: a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts. Genomics & informatics, Vol. 17, No. 2.
2. **BioASQ (2022).** `http://bioasq.org/`.
3. **Campillos-Llanos, L. (2019).** First steps towards building a medical lexicon for Spanish with linguistic and semantic information. **Demner-Fushman, D., Cohen, K. B., Ananiadou, S., Tsujii, J.**, editors, Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, pp. 152–164. DOI: `10.18653/v1/W19-5017`.
4. **Chen, L., Gu, Y., Ji, X., Lou, C., Sun, Z., Li, H., Gao, Y., Huang, Y. (2019).** Clinical trial cohort selection based on multi-level rule-based natural language processing system. Journal of the American Medical Informatics Association, Vol. 26, No. 11, pp. 1218–1226.
5. **Dernoncourt, F., Lee, J. Y., Szolovits, P. (2017).** NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. arXiv preprint arXiv:1705.05487.
6. **Grishman, R., Sundheim, B. (1996).** Message Understanding Conference- 6: A brief history. COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.
7. **Ji, B., Liu, R., Li, S., Yu, J., Wu, Q., Tan, Y., Wu, J. (2019).** A hybrid approach for named entity recognition in Chinese electronic

medical record. BMC medical informatics and decision making, Vol. 19, pp. 149–158.

8. **Krallinger, M., Armengol-Estapé, J., De Gibert, O., Carrino, C. P., Gonzalez-Agirre, A., Gutiérrez-Fandiño, A., Villegas, M. (2021).** Spanish biomedical crawled corpus. DOI: 10.5281/zenodo.4561971.
9. **Landolsi, M. Y., Ben Romdhane, L., Hlaoua, L. (2024).** Hybrid medical named entity recognition using document structure and surrounding context. The Journal of Supercomputing, Vol. 80, No. 4, pp. 5011–5041.
10. **Moreno-Barea, F. J., López-García, G., Mesa, H., Ribelles, N., Alba, E., Jerez, J. M., Veredas, F. J. (2025).** Named entity recognition for de-identifying Spanish electronic health records. Computers in Biology and Medicine, Vol. 185, pp. 109576. DOI: https://doi.org/10.1016/j.compbiomed.2024.109576.
11. **Qiu, Q., Tian, M., Xie, Z., Tan, Y., Ma, K., Wang, Q., Pan, S., Tao, L. (2023).** Extracting named entity using entity labeling in geological text using deep learning approach. Journal of Earth Science, Vol. 34, No. 5, pp. 1406–1417.
12. **Ramshaw, L., Marcus, M. (1995).** Text chunking using transformation-based learning. Third Workshop on Very Large Corpora.
13. **Spacy (2022).** https://spacy.io/.
14. **Tang, X., Huang, Y., Xia, M., Long, C. (2023).** A multi-task BERT-BiLSTM-AM-CRF strategy for Chinese named entity recognition. Neural Processing Letters, Vol. 55, No. 2, pp. 1209–1229.
15. **Xiong, Y., Chen, S., Tang, B., Chen, Q., Wang, X., Yan, J., Zhou, Y. (2021).** Improving deep learning method for biomedical named entity recognition by using entity definition information. BMC bioinformatics, Vol. 22, pp. 1–13.
16. **Yoon, W., So, C. H., Lee, J., Kang, J. (2019).** Collabonet: collaboration of deep neural networks for biomedical named entity recognition. BMC bioinformatics, Vol. 20, pp. 55–65.
17. **Zhou, Y., Zheng, X.-Q., Huang, X.-J. (2023).** Chinese named entity recognition augmented with lexicon memory. Journal of Computer Science and Technology, Vol. 38, No. 5, pp. 1021–1035.

Article received on 03/06/2024; accepted on 03/07/2025.
**Corresponding author is Alma Delia Cuevas-Rasgado.*