Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 41-52 recibido 27-03-2025 revisado 12-05-2025 aceptado 16-05-2025

Unifying Named Entity Recognition and Extreme Multi-Label Classification for Explainable Clinical Coding

Integración del Reconocimiento de Entidades Nombradas y la Clasificación Extrema Multi-Etiqueta para una Codificación Clínica Explicable

Alicia Ramirez-Arrabe,[1] **Andres Duque,**[1,2] **Juan Martinez-Romo**[1,2]

[1]Universidad Nacional de Educación a Distancia (UNED), 28040, Madrid
[2]Instituto Mixto UNED-ISCIII IMIENS
{aramirez, aduque, juaner}@lsi.uned.es

Abstract: Automatic clinical coding of medical reports sits at the intersection of healthcare and Natural Language Processing (NLP), facilitating the extraction of relevant information from unstructured clinical documents. This study introduces a three-stage explainable automatic coding system, developed within the experimental framework of the 2020 CodiEsp competition, a task devoted to automatic clinical coding in Spanish. The proposed system integrates two Named Entity Recognition (NER)-based models, a supervised text classification model, and an unsupervised similarity model enhanced with keyphrase extraction. This methodology allows for the detection of overlapped and discontinuous evidence texts, as well as for the inclusion of Out-Of-Distribution (OOD) codes. Our approach outperforms most state-of-the-art models, achieving an F1-score improvement of 4.2%, 0.2%, and 4.1% in the CodiEsp-D, CodiEsp-P and CodiEsp-X subtasks, respectively, and an increase of up to 2.4% in the MAP values.
Keywords: Automatic Clinical Text Coding, Keyphrase Extraction, Named Entity Recognition, Semantic Similarity.

Resumen: La codificación automática clínica de informes médicos sirve como intersección entre la atención sanitaria y el Procesamiento de Lenguaje Natural (PLN), facilitando la extracción de información relevante de documentos clínicos no estructurados. Este trabajo presenta un sistema de codificación automática explicable en tres etapas, desarrollado dentro del marco experimental de la competición CodiEsp 2020, una tarea orientada a la clasificación clínica automática en español. El sistema propuesto integra dos modelos basados en el Reconocimiento de Entidades Nombradas (NER), un modelo de clasificación de texto supervisado y un modelo de similitud no supervisado enriquecido con la extracción de frases clave. Esta metodología permite la detección de evidencias de texto superpuestas y/o discontinuas, así como la inclusión de códigos de fuera de la distribución. Nuestro enfoque supera a la mayoría de los modelos del estado del arte, logrando una mejora del 4,2%, 0,2% y 4,1% de la métrica F1 en las subtareas CodiEsp-D, CodiEsp-P y CodiEsp-X, respectivamente, además de un aumento de hasta el 2,4% en los valores de la métrica MAP.
Palabras clave: Codificación clínica automática de texto, Extracción de Frases Clave, Reconocimiento de Entidades Nombradas, Similitud Semántica.

1 Introduction

In the field of medical research and practice, the process of automatic coding and categorization of medical texts is of great importance, due to the vast amount of such texts that need to be processed every single day. This coding process significantly contributes to the structuring, storage, and retrieval of large volumes of medical data, directly influencing their subsequent use in various tasks in an effective and efficient manner. To this end, the International Classification of Dis-

ISSN 1135-5948 DOI 10.26342/2025-75-3

eases, in its tenth revision (ICD-10), aims to standardize the process of assigning unambiguous codes to diagnoses and procedures related to medical practice. However, manual processing of this amount of medical information is highly resource-consuming and prone to errors (O'malley et al., 2005). As a result, the development of automated systems based on Natural Language Processing (NLP) to tackle these tasks has become a particularly active area of research in this domain in recent years (Xie and Xing, 2018; Zhou et al., 2021; Yang et al., 2025). Additionally, one of the key concepts in the development of automated decision-support systems in the biomedical domain is explainability. In a field as sensitive as medicine, it is essential that the decisions made by an automated system are supported by evidence, ensuring that the stakeholders involved in the decision-making process (such as physicians, nurses, or patients, among others) can fully trust these decisions (Shaban-Nejad, Michalowski, and Buckeridge, 2021).

In this context, the CodiEsp competition (Miranda-Escalada et al., 2020), launched in 2020 as part of the eHealth shared task within the CLEF (Conference and Labs of the Evaluation Forum) conference, promotes the development of systems capable of automatically assigning ICD-10 codes to medical reports written in the Spanish language. In this case, the classification is referred to as CIE-10, following its Spanish acronym ("*Clasificación Internacional de Enfermedades*"). CodiEsp is divided into three main subtasks: the automatic assignment of diagnosis codes (CodiEsp-D), the automatic assignment of procedure codes (CodiEsp-P), and an explainability-related subtask (CodiEsp-X), which requires not only the assignment of codes to a medical report but also the identification of textual evidence (text spans) that justify the assigned codes.

In this study, we revisit the three subtasks proposed in CodiEsp and present a comprehensive system that addresses them from an integrated perspective. Our approach consists of a system divided into three different phases, which will be detailed later on. In the first phase, we perform textual evidence detection by training a Named Entity Recognition (NER) model to identify medical expressions that may lead to the assignment of an ICD-10 code, whether for diagnosis or procedure classification. The second phase leverages these textual evidences to train an extreme multi-label classification model capable of assigning an ICD-10 code to each text span extracted in the first phase. This approach ensures that the classification model is provided with more targeted information (the text spans containing medically relevant evidence), reducing the noise inherent to processing the full medical report. Finally, the third phase refines the system by handling those cases where the supervised model struggles to confidently assign a code. In these scenarios, also automatically detected in the second phase, an unsupervised model is applied, based on the semantic similarity between the textual evidences and keyphrases highly associated with the different assignable codes. In this final step, only those codes whose representative keyphrases exhibit high semantic similarity with the textual evidence extracted in the first phase are assigned.

We report results for each of the three subtasks of the competition and compare them with those obtained by the original participants, as well as with other systems developed after the competition. The results clearly show that our system is highly competitive, outperforming the state of the art in most evaluated metrics and scenarios.

The rest of the paper is organized as follows: Section 2 reviews the main related works on ICD-10 code assignment, including some systems that addressed the CodiEsp task. Section 3 describes the developed system, providing a detailed explanation of each stage of the proposed methodology, as well as the corpora employed in the research. Section 4 presents the main results obtained by our system and compares them with the state of the art. Finally, Section 5 outlines the key conclusions of this work and discusses potential directions for future research.

2 Background

Most of the advances in the field of automatic ICD-10 coding have been made in the English language, and various methodologies have been explored to enhance coding accuracy. For instance, deep learning models, particularly those employing Transformer architectures like BERT, have been utilized to predict ICD-10 codes from clinical notes. One study (Chen et al., 2021) implemented

a deep neural network that automatically determined corresponding diagnosis and procedure codes based solely on free-text medical notes. Similarly, another approach (Zhou et al., 2020) employed deep learning techniques to create a semi-automatic system capable of assigning ICD-10 codes to clinical narratives. The system's performance was evaluated using a dataset of 12,000 medical records. In other study (Pereira et al., 2006) the authors developed a semi-automated system to assist in coding medical records with ICD-10 codes. The system utilized an automated MeSH-based indexing tool, mapping MeSH terms to ICD-10 codes via the UMLS Metathesaurus (Zweigenbaum, 1999). Additionally, the system incorporated drug prescription data, linking prescribed medications to relevant ICD-10 codes based on drug approvals.

Few studies have been published on the automatic coding of clinical texts in Spanish. This work (Almagro et al., 2020) evaluated different algorithms for assigning CIE-10-ES codes to Spanish clinical texts from hospital discharge reports. They compared binary classification methods, subset grouping approaches, and extreme classification (eXtreme Multi-label Text Classification). Their findings suggest that ensemble methods, which weight each code based on training frequency and performance, can enhance classification performance in highly imbalanced distributions like CIE-10-ES coding. Another research work (Blanco et al., 2019) investigated various deep learning models within a multi-label classification framework to tackle clinical coding using a dataset sourced from a public hospital in Spain. A previous study (Pérez et al., 2018) applied a latent Dirichlet allocation (LDA)-based approach for multi-label classification of electronic health records from the cardiology department at a Spanish public hospital, achieving positive results with the 124 most frequently occurring CIE-10-ES codes in the dataset. A later work (Duque et al., 2021) proposes an automatic ICD-10 code recommendation system based on keyphrase extraction. The dataset consists of Spanish EHRs, where ICD-10 codes are characterized by keyphrases, including both literal occurrences and statistically inferred expressions guiding human annotators. Results demonstrate its competitiveness with state-of-the-art machine learning approaches while offering interpretability.

One of the main handicaps of most of the works is that the datasets are made up of electronic reports from real patients and are not public due to privacy issues. In this sense, an initiative that allows comparing different systems against the same dataset is the CodiEsp shared task (Miranda-Escalada et al., 2020), which, as previously mentioned, is notable for being the first shared task focused on automatic coding of clinical cases in Spanish. To this end, the CodiEsp corpus was made available, a synthetic dataset comprising 1,000 clinical case samples manually curated by the task organizers. The evaluation metrics used were F1-score and MAP, and depending on the subtask and metric considered the best system differs. In general, the best systems employed NER techniques, classifiers and tuned dictionaries. A subsequent work on the task reported better results in the state of the art although only for the MAP metric. This study (López-García et al., 2021) fine-tuned different large language models (LLMs) for the clinical coding tasks using a multi-label sentence classification strategy. The domain-specific versions outperformed their general-domain counterparts, and an ensemble approach achieved the best results. Another later study (Barros et al., 2022) developed a novel architecture based on neural networks by analyzing the hierarchical nature of ontologies to create clusters based on semantic relationships. The results obtained were promising, although they did not employ the same evaluation methodology as the CodiEsp task by using a subset of the codes. A recent work (Barreiros et al., 2025) utilized the CodiEsp corpus to evaluate the performance of various LLMs, achieving strong results; however, it is not directly comparable to our study, as the models were given the gold standard codes as part of the input.

3 Materials and methods

3.1 Corpora

The corpora used throughout this work correspond to the so-called CodiEsp corpus, collected for the CodiEsp shared task of CLEF eHealth 2020 (Miranda-Escalada et al., 2020). As previously mentioned, the CodiEsp track encompasses three subtasks: CodiEsp-D, CodiEsp-P and CodiEsp-X, all of them requiring automatic ICD-10 code as-

	CodiEsp-D			CodiEsp-P		
	Train.	**Dev.**	**Test**	**Train.**	**Dev.**	**Test**
Documents	500	250	250	435	222	224
Total ICD codes	7209	3431	3665	1972	1046	1112
Unique ICD codes	1767	1158	1143	563	375	371
Unique ICD codes to each set	807	351	363	293	129	143
ICD codes per document (mean)	14.42	13.72	14.66	4.53	4.71	4.96

Table 1: Summary statistics of CodiEsp-D and CodiEsp-P corpora.

signment. The CodiEsp corpus comprises 1,000 clinical reports written in Spanish from a wide diversity of medical specialties. It is divided into three sets: training (500 documents), development (250 documents), and test (250 documents) (see Table 1). In this work, we focus on tackling the three CodiEsp subtasks, whose descriptions are as follows:

- CodiEsp-D: requires the automatic coding of the 2018 version of ICD-10-CM codes (CIE-10-ES *Diagnósticos*[1] in Spanish), used to diagnose medical conditions. Diagnosis codes follow a hierarchical structure, grouping codes from a minimum of three characters to a maximum of seven. The task organizers provided a list of 98,288 valid codes for this subtask, along with their descriptions in both Spanish and English.
- CodiEsp-P: entails the automatic assignment of the 2018 version of ICD-10-PCS codes (CIE-10-ES *Procedimientos*[2] in Spanish), used for coding medical procedures. These codes follow a multi-axial structure of seven characters, where each of them refers to a specific characteristic, such as anatomical location, organ system, or procedure type. A total of 87,170 valid procedure codes were provided by the organizers. Nonetheless, 229 of these valid codes are made up of only four characters and lack a description, as annotators considered the available contextual information insufficient to designate the complete seven-character code. Besides, it is worth mentioning that, while all CodiEsp documents contain at least one diagnosis code, not all include a procedure code, thereby reducing the number of available documents from each dataset for the CodiEsp-P subtask (see Table 1).
- CodiEsp-X: focuses on eXplainable Artificial Intelligence (XAI) and aims at providing supporting evidence texts alongside each assigned code. Such textual evidence must be returned by indicating their beginning and ending character positions, and can be either continuous or discontinuous (see Table 2). If an evidence text is said to be discontinuous, it contains non-adjacent spans. In this subtask, diagnosis and procedure codes are integrated together. By inspecting the subtask annotations, it can be noticed that a code may appear more than once in a document and that there exist overlapped evidence texts. In this study, two pieces of textual evidence have been considered as overlapping if they share at least one token. Among the 1,000 documents, there are a total of 4,476 overlapping cases. Figure 1 illustrates some examples of discontinuous and overlapped entities.

CodiEsp subtasks can be categorized as eXtreme Multi-label Text Classification (XMTC) problems, as their sets of labels contain hundreds of possible codes and more than one label can be assigned to each document. Furthermore, the three datasets (training, development and test) contain codes that are unique to them, that is to say, not present within the others. This implies that the test set includes Out-of-Distribution (OOD) codes, making it inappropriate to approach the subtasks in a fully supervised way. Additionally, the frequency of occurrence of the codes is notably low. Specifically, of the 2,557 unique codes comprised within the three sets of CodiEsp-D, 1,325 appear only once. Similarly, among the 870 unique codes from CodiEsp-P, 513 occur only once.

[1] https://www.sanidad.gob.es/estadEstudios/estadisticas/normalizacion/CIE10/CIE10ES_2018_norm_MANUAL_CODIF_DIAG_.pdf

[2] https://www.sanidad.gob.es/estadEstudios/estadisticas/normalizacion/CIE10/CIE10ES_2018_norm_MANUAL_CODIFICACION_PROCEDIMIENTOS_EDICION_2018.pdf

	Continuous entities		Discontinuous entities		Total
	Diagnoses	Procedures	Diagnoses	Procedures	
Train.	6162	1232	1047	740	9181
Dev.	2922	589	509	457	4477
Test	3141	682	524	430	4777
Total	12225	2503	2080	1627	18435

Table 2: Summary statistics of CodiEsp-X corpus.

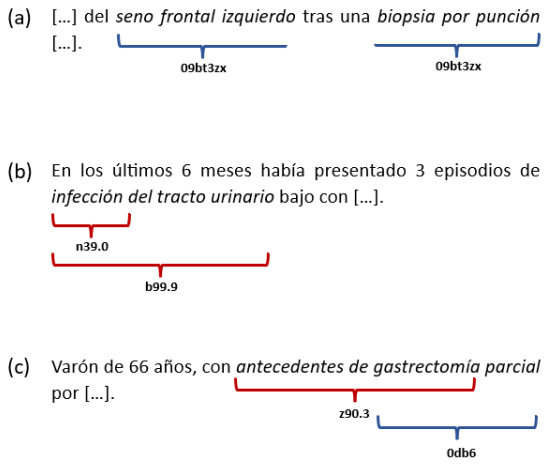

Figure 1: Three examples illustrating a discontinuous entity and two overlapped entities, where diagnoses are depicted in red and procedures in blue: (a) a discontinuous procedure, (b) two overlapping diagnoses, and (c) a diagnosis overlapping with a procedure.

3.2 Proposed system

Given the availability of CodiEsp-X annotations, and to unify the predictions across the three substasks, the problem has been approached by first identifying the evidence texts corresponding to a diagnosis or procedure mention and subsequently assigning them their possible codes. The workflow diagram illustrating the applied methodology in this work is shown in Figure 2. Since diagnosis and procedure codes differ in structure and conceptualization, their respective evidence texts have been identified and classified using two separate modules. However, in both cases, the same three-stage pipeline has been followed. In order to rely on a larger training dataset, we have merged the original training (500 docs) and development sets (250 docs) and then randomly split them into two updated training and development sets, following an 80-20% partition ratio. As a result, the updated training set consists of 600 documents, while the new development set comprises 150. A stratified partition was not feasible, as several codes appear only once throughout the entire set of documents. The three phases of the pipeline are described in the following subsections.

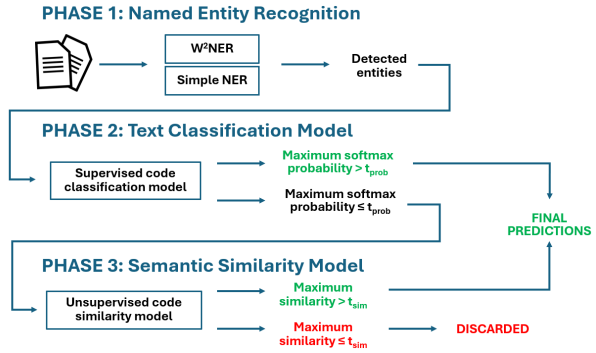

Figure 2: Workflow diagram of the three-stage approach implemented in this study. Diagnoses and procedures are identified and classified separately. t_{prob} represents the probability threshold defined in phase 2 to filter out the predictions, while t_{sim} stands for the similarity threshold from phase 3.

3.2.1 Phase 1: Named Entity Recognition

As already mentioned, we have decided to address the task by first identifying the textual evidence rather than applying text classification straight away. In particular, two different Named Entity Recognition systems have been implemented:

- SimpleNER: The texts have been annotated at the token level using the BIO (Beginning, Inside, Outside), or IOB, annotation scheme (Ramshaw and Marcus, 1999). In this annotation process, each token can only be labeled once; hence, an issue regarding overlapping entities in the corpus has arisen. To address this challenge, when annotating a new entity, if a token within its span has already been labeled as part of another entity (either as B or I), the new entity is discarded and omitted to avoid labeling conflicts.

After the annotation process, we have trained a NER model using a RoBERTa (Liu et al., 2019) model pre-trained with different clinical and biomedical resources written in the Spanish language (Carrino et al., 2022). Other evaluated pre-trained models that were discarded because their lower results include a different RoBERTa-based model (de la Iglesia et al., 2023) and a Longformer-based model (Carrino et al., 2022), both pre-trained on biomedical and clinical Spanish language. The main limitation of the selected pre-trained model for this task is its maximum input size of 512 tokens, because some documents exceed this length. Therefore, documents surpassing such limitation have been divided into chunks of up to 512 tokens. The models have been trained for 33 epochs for diagnoses and 21 epochs for procedures, employing the macro F1 metric for validation and applying an early stopping criteria based on the development set performance.

- W^{2}NER: Due to the presence of overlapped entities in the corpus, many of which have been discarded in the SimpleNER approach, we have applied the W^{2}NER scheme (Li et al., 2022). This innovative system models documents as a 2D grid of token pairs, capturing all possible entity relationships each token may have, and thus allowing us to include the entire set of entities within the corpus without disregarding any.

  The W^{2}NER method has been trained using the same RoBERTa pre-trained model (Carrino et al., 2022) as in the SimpleNER approach. Consequently, some documents have been chunked to comply with the 512-token upper constraint. During this process, a small percentage of entities has been lost as they spanned across two different chunks. Specifically, in both the training and development sets, 94 diagnosis entities out of 10,640 and 98 procedure entities out of 3,018 were affected. After annotating the corpus in accordance with the requirements of the W^{2}NER scheme, both the diagnosis and procedure models have been trained for a broader range of epochs than the SimpleNER models as the learning pace of the W^{2}NER is slower.

3.2.2 Phase 2: Text classification model

Once the entities have been detected by the NER models, an extreme multi-label classification model is implemented to assign a code to each piece of textual evidence. This model is trained using the same RoBERTa pre-trained model (Carrino et al., 2022) as in the previous phase, considering a set of possible labels made up of only the codes present in the training set and performing a hyperparameter optimization using the Optuna framework (Akiba et al., 2019). In this phase, text chunking has not been required, as all identified entities fulfilled the pre-trained model's input length requirement. By applying an early stopping criteria based on the macro F1 performance on the development set, the models have been trained for 33 epochs for diagnoses and 28 epochs in the case of procedures. However, the development set, which serves to validate the models, contains labels that are not seen throughout the supervised model training and can be considered OOD instances. To identify such instances, we have implemented a minimum confidence threshold that filters out predictions where the text classification model does not showcase sufficient confidence, based on the maximum softmax probability assigned to each instance. By defining this threshold and discarding low-confidence predictions, we ensure that codes likely absent from the set of labels used in training are excluded.

The probability thresholds were evaluated at intervals of 0.05, taking values from 0 to 1. The CodiEsp task assigns equal importance to the F1-score and MAP metrics. Therefore, two different thresholds have been computed, one to maximize each metric. As shown in Figure 3, if the objective is to maximize the F1-score on the development set, the probability threshold is set to 0.50 for the evidence texts corresponding to diagnosis codes and 0.55 for procedures. On the other hand, if we intend to maximize the MAP metric on the development set, a less restrictive threshold is applied, taking values of 0.40 and 0.35, respectively. Lastly, predictions that do not meet the selection criteria proceed to the third and last stage of the system.

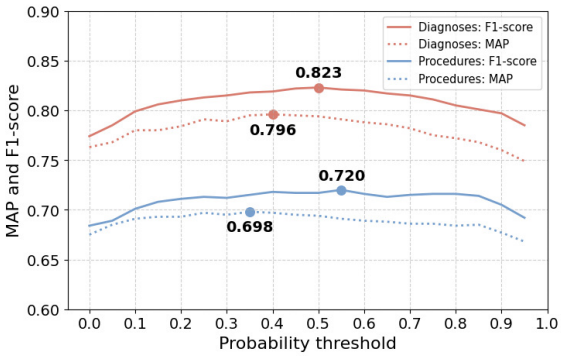

Figure 3: MAP and F1-score values with respect to the predefined probability thresholds for diagnoses (red) and procedures (blue) predictions, generated by the phase 2 text classification model on the development set.

3.2.3 Phase 3: Semantic similarity model

To classify the instances excluded in the previous stage, and assuming that the set of possible codes is unknown, we have implemented an unsupervised approach consisting of a semantic similarity model. As task organizers provided a list of 98,288 valid diagnosis codes and 87,170 valid procedure codes along with their descriptions, we can use these descriptions to find the most similar one for each piece of textual evidence. However, in some cases, the phrasing of a code's description may not be suitable for assigning the correct code. Therefore, in order to build a more enriched list of descriptions, we have leveraged the keyphrase-based methodology presented in (Duque et al., 2021). This methodology has been applied by combining the CodiEsp training and development sets with the dataset used in the original study, which consists of 12,966 medical discharge reports from 2016. We have extracted the two most significant keyphrases for each code appearing in those datasets. In total, 6,545 diagnosis codes and 3,337 procedure codes include a description and their top two keyphrases, except for the 229 procedure codes that, as mentioned previously, were provided without a description. Then, using the Arctic-Embed 2.0 multilingual embedding model (Yu et al., 2024), we have computed the embeddings for each piece of textual evidence, as well as for each code description and keyphrases. Subsequently, applying the Sentence Transformers framework (Reimers and Gurevych, 2019), the entities' embeddings have been compared to the codes' embeddings, and then the code with the highest cosine similarity is assigned to such entity. Once the entities have been classified according to their most similar code, a similarity threshold has been defined to filter out mappings where the similarity value is not high enough. Figure 4 shows three cases where the code's keyphrases are more helpful than its description, with two of them even matching the evidence text exactly.

Textual evidence	*tibia vara*	*ambos ojos fondo de ojo*	*oxigenoterapia*
Code	m92.50	08j1xzz	3e0f7sf
Description	*Osteocondrosis juvenil de tibia y peroné, pierna no especificada*	*Inspección de ojo, izquierdo, abordaje externo*	*Introducción en tracto respiratorio de gas, otro gas, abordaje orificio natural o artificial*
	Similarity: 0.4669	Similarity: 0.5737	Similarity: 0.3705
Keyphrase 1	*tibia vara*	*ojo*	*oxigenoterapia*
	Similarity: 1	Similarity: 0.5918	Similarity: 1
Keyphrase 2	*tosca fascies*	*de ojo fondo*	*fibrosis bilateral pulmonar*
	Similarity: 0.3997	Similarity: 0.7760	Similarity: 0.3176

Figure 4: Three examples illustrating the code assignment effectivity of keyphrases in cases where the code's description may not be as suitable. Underlined keyphrases indicate the most similar matches to each piece of textual evidence.

The similarity thresholds were evaluated at intervals of 0.05, taking values from 0.50 to 1. As shown in Figure 5, the similarity threshold that maximizes the F1-score value on the development set is 0.85 for diagnoses and 0.90 for procedures. Furthermore, when it comes to maximizing the MAP value on the development set, the similarity threshold values are lower, specifically 0.70 for diagnoses and 0.80 for procedures. It is also worth mentioning than an increase in the maximum F1-score and MAP values with respect to the maximum ones obtained in phase 2 (see Figure 3) is observed. Despite this increase being modest due to the fact that the number of instances processed in phase 3 is much smaller than those processed in phase 2, it demonstrates the positive contribution of applying an unsupervised semantic-based approach to include the OOD codes.

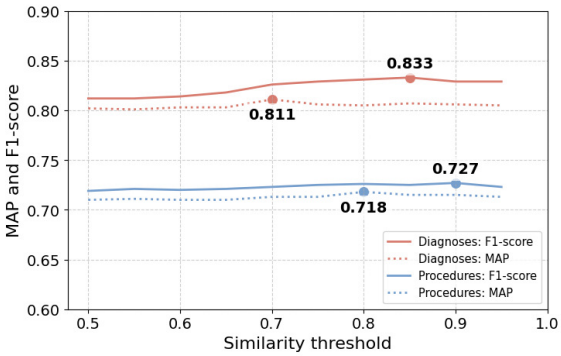

Figure 5: MAP and F1-score values with respect to the predefined similarity thresholds for diagnoses (red) and procedures (blue) predictions, generated by applying the phase 3 semantic similarity model on the development set to the predictions obtained throughout phase 2.

4 Results

Once we have the textual evidence for each document along with its corresponding code, we evaluate the CodiEsp-D and CodiEsp-P subtasks by extracting the predicted codes per document and sorting them. First, we place the codes predicted in phase 2 (ordered by the model's assigned probability), followed by the codes predicted during phase 3 (sorted by maximum similarity). For the assessment of CodiEsp-X subtask, we merge the diagnosis and procedure predictions, extracting the code, position and type (diagnosis or procedure) of each entity per document. The results of the three subtasks can be evaluated based on the three ranking configurations considered in the CodiEsp shared task:

- **CodiEsp-D/P/X**: it evaluates the system considering the complete set of codes present in the test set. This ranking provides a general view of its performance in a more realistic scenario, where there may be unseen codes.
- **CodiEsp-D/P/X train+dev**: the system is evaluated using only the subset of codes present in the training and development sets, that is, it focuses on the system's performance with previously seen data.
- **CodiEsp-D/P/X categories**: this evaluation takes into account only the first three digits of diagnosis codes and the first four digits of procedure codes. Thus, this ranking evaluates the system's ability to assign the broader categories of codes instead of their exact matches.

Table 3 presents the results obtained by the different systems across the three evaluations of the three subtasks, CodiEsp-D, CodiEsp-P and CodiEsp-X. In total, three systems are evaluated. The first two rows show results from the SimpleNER and W2NER systems, respectively, using classification models from phases 2 and 3. The third row (Ensemble) presents results from combining the predictions of both systems. Additionally, a final row includes the upper-bound results based on perfect NER performance, that is, applying the classification models directly to the true entities of the test set. This row provides insight into the performance of both the NER and classification models.

The results are consistent across the three evaluations for the three subtasks. In each case, the SimpleNER approach achieves the best Precision, while the highest values for the other metrics (Recall, F1-score, and MAP) are reached by the Ensemble system. However, despite leading in Precision, the SimpleNER system falls behind in Recall and MAP. The W^{2}NER system, although it has lower Precision, achieves higher Recall, which leads to a better F1-score. Finally, the Ensemble system serves as a balance between the two approaches, achieving the highest values for Recall, F1-score, and MAP.

It is also worth mentioning that classifying diagnoses appears to be an easier task than classifying procedures. This may be due to three factors: (1) A higher proportion of annotations correspond to diagnoses, specifically, 77.8% (Miranda-Escalada et al., 2020), (2) As previously stated, 229 out of the 870 unique procedure codes are incomplete, and (3) There are more abbreviations in procedures than in diagnoses (Miranda-Escalada et al., 2020).

Furthermore, it can be stated that the first type of evaluation of all subtasks (CodiEsp-D/P/X) is the most challenging, offering a more realistic assessment of how the systems would behave in real-world scenarios. The train+dev evaluation shows better performance across all systems, suggesting that the models are more familiar with this subset of codes, and evidencing the need for improvements in the phase 3 similarity model. Similarly, in the type of evaluation regarding

CodiEsp-D System	CodiEsp-D				CodiEsp-D train+dev				CodiEsp-D categories			
	P	R	F1	MAP	P	R	F1	MAP	P	R	F1	MAP
SimpleNER MF1	**0.818**	0.507	0.626	0.484	**0.831**	0.584	0.686	0.567	**0.875**	0.587	0.703	0.565
W^{2}NER MF1	0.802	0.641	0.713	0.586	0.820	0.738	0.777	0.682	0.865	0.725	0.788	0.674
Ensemble MF1	0.778	0.663	**0.716**	0.608	0.795	0.764	**0.779**	0.710	0.844	0.749	**0.794**	0.700
PerfectNER MF1	*0.902*	*0.751*	*0.819*	*0.724*	*0.918*	*0.862*	*0.889*	*0.843*	*0.951*	*0.817*	*0.879*	*0.798*
SimpleNER MMAP	0.756	0.520	0.616	0.492	0.793	0.595	0.679	0.573	0.834	0.616	0.708	0.589
W^{2}NER MMAP	0.730	0.659	0.693	0.597	0.778	0.752	0.765	0.690	0.810	0.760	0.784	0.702
Ensemble MMAP	0.701	**0.681**	0.691	**0.617**	0.750	**0.778**	0.764	**0.717**	0.784	**0.786**	0.785	**0.728**
PerfectNER MMAP	*0.823*	*0.774*	*0.798*	*0.743*	*0.872*	*0.879*	*0.875*	*0.856*	*0.895*	*0.864*	*0.879*	*0.844*

CodiEsp-P System	CodiEsp-P				CodiEsp-P train+dev				CodiEsp-P categories			
	P	R	F1	MAP	P	R	F1	MAP	P	R	F1	MAP
SimpleNER MF1	**0.772**	0.358	0.490	0.456	**0.793**	0.434	0.561	0.534	**0.816**	0.383	0.522	0.495
W^{2}NER MF1	0.743	0.461	0.569	0.522	0.722	0.568	0.636	0.598	0.772	0.508	0.613	0.573
Ensemble MF1	0.683	0.504	**0.580**	0.553	0.699	0.611	**0.652**	0.631	0.793	0.545	**0.627**	0.608
PerfectNER MF1	*0.876*	*0.629*	*0.732*	*0.667*	*0.880*	*0.760*	*0.816*	*0.760*	*0.920*	*0.666*	*0.773*	*0.710*
SimpleNER MMAP	0.747	0.365	0.490	0.463	0.762	0.443	0.560	0.542	0.786	0.389	0.520	0.502
W^{2}NER MMAP	0.715	0.469	0.566	0.522	0.722	0.568	0.636	0.598	0.772	0.508	0.613	0.573
Ensemble MMAP	0.675	**0.509**	**0.580**	**0.557**	0.689	**0.617**	0.651	**0.636**	0.727	**0.546**	0.624	**0.611**
PerfectNER MMAP	*0.819*	*0.661*	*0.732*	*0.684*	*0.837*	*0.800*	*0.818*	*0.783*	*0.869*	*0.707*	*0.780*	*0.742*

CodiEsp-X System	CodiEsp-X				CodiEsp-X train+dev				CodiEsp-X categories			
	P	R	F1	MAP	P	R	F1	MAP	P	R	F1	MAP
SimpleNER MF1	**0.772**	0.450	0.568	-	**0.786**	0.525	0.629	-	**0.815**	0.504	0.623	-
W2NER MF1	0.755	0.578	0.654	-	0.770	0.674	0.719	-	0.798	0.631	0.705	-
Ensemble MF1	0.729	**0.602**	**0.660**	-	0.746	**0.702**	**0.723**	-	0.775	**0.657**	**0.711**	-
PerfectNER MF1	*0.895*	*0.720*	*0.798*	-	*0.908*	*0.837*	*0.871*	-	*0.942*	*0.776*	*0.851*	-

Table 3: Results for CodiEsp-D, CodiEsp-P and CodiEsp-X, evaluated using CodiEsp's official metrics. For each system, Precision (P), Recall (R) and F1-score (F1) values are reported. For subtasks CodiEsp-D and CodiEsp-P, Mean Average Precision (MAP) values are also included. MF1 systems use the probability and similarity thresholds that maximize the F1-score, whereas MMAP systems use the thresholds that maximize MAP values. The best performance of each metric and evaluation is highlighted in bold, while the second-best performance is underlined.

the classification of categories, which involves a simpler classification task, the systems also obtain better results, indicating that they are able to effectively comprehend the broader structure of the codes.

Finally, Table 4 compares the systems implemented in this study with the best state-of-the-art (SOTA) models. The results show that, except for the MAP value in CodiEsp-D, our systems outperform all of them. In the CodiEsp-D subtask, the highest MAP is obtained using a Transfer Learning (TL) approach, where three Transformer architectures were pre-trained on Spanish oncology clinical cases and fine-tuned for a multi-label sentence classification task (López-García et al., 2021). This methodology also achieves the second-best F1-score in the CodiEsp-P subtask. Other notable approaches to the task include a NER model combined with dictionary lookup (Cossin and Jouhet, 2020), fine-tuning of pre-trained models, such as a multilingual BERT model (Costa et al., 2020) and a BETO model (Blanco, Pérez, and Casillas, 2020), and other non-deep learning approaches, including a semantic and hierarchical clustering (Barros et al., 2022) and XGBoost for text classification (Blanco, Pérez, and Casillas, 2020).

5 Conclusions and Future Work

In this study, we implement a three-stage system for explainable ICD-10 coding within the experimental framework of the CodiEsp 2020 shared task. We compare the performance obtained using two different NER-based schemas, followed by an extreme multi-label supervised classification model and an unsupervised similarity model enriched with keyphrase extraction. The system incorporates an innovative NER-based schema, W^{2}NER, which enables the detection and inclusion of discontinuous and overlapped medical entities, thereby enhancing the explainability of the system. Moreover, the phase 3 unsupervised model allows for the inclusion of OOD codes, which were unseen during model training, providing a methodology aligned with real-world scenarios.

The results are consistent across all evaluations and metrics, proving the reliability of the proposed system. Furthermore, they outperform the SOTA in most scenarios, demonstrating the competitive potential of the system. Interestingly, the best F1-score does not always lead to the best MAP, suggesting that different metrics highlight different aspects of the performance. There is also a clear relationship between higher Recall and

CodiEsp-D System	CodiEsp-D		CodiEsp-D train+dev		CodiEsp-D categories	
	F1	MAP	F1	MAP	F1	MAP
IAM (Cossin and Jouhet, 2020)	0.687	0.521	0.748	0.605	0.773	-
IXA-AAA (Blanco, Pérez, and Casillas, 2020)	0.009	0.593	0.009	0.698	0.021	-
Clinical Coding (López-García et al., 2021)	0.677	**0.662**	-	-	-	-
Divide and Conquer (Barros et al., 2022)	-	-	-	-	0.746	0.682
SimpleNER MF1	0.626	0.484	0.686	0.567	0.703	0.565
W^{2}NER MF1	0.713	0.586	0.777	0.682	0.788	0.674
Ensemble MF1	**0.716**	0.608	**0.779**	0.710	**0.794**	0.700
SimpleNER MMAP	0.616	0.492	0.679	0.573	0.708	0.589
W^{2}NER MMAP	0.693	0.597	0.765	0.690	0.784	0.702
Ensemble MMAP	0.691	0.617	0.764	**0.717**	0.785	**0.728**

CodiEsp-P System	CodiEsp-P		CodiEsp-P train+dev		CodiEsp-P categories	
	F1	MAP	F1	MAP	F1	MAP
IAM (Cossin and Jouhet, 2020)	0.522	0.493	0.586	0.569	0.579	-
The Mental Strokers (Costa et al., 2020)	0.488	0.445	0.531	0.509	0.541	-
Clinical Coding (López-García et al., 2021)	0.579	0.544	-	-	-	-
Divide and Conquer (Barros et al., 2022)	-	-	-	-	0.560	0.562
SimpleNER MF1	0.490	0.456	0.561	0.534	0.522	0.495
W^{2}NER MF1	0.569	0.522	0.636	0.598	0.613	0.573
Ensemble MF1	**0.580**	0.553	**0.652**	0.631	**0.627**	0.608
SimpleNER MMAP	0.490	0.463	0.560	0.542	0.520	0.502
W^{2}NER MMAP	0.566	0.522	0.636	0.598	0.613	0.573
Ensemble MMAP	**0.580**	**0.557**	0.651	**0.636**	0.624	**0.611**

CodiEsp-X System	CodiEsp-X		CodiEsp-X train+dev		CodiEsp-X categories	
	F1	MAP	F1	MAP	F1	MAP
IAM (Cossin and Jouhet, 2020)	0.611	-	0.667	-	-	-
FLE (García-Santa et al., 2020)	0.611	-	0.670	-	-	-
Explainable Clinical Coding (López-García et al., 2023)	0.633	-	-	-	-	-
SimpleNER MF1	0.568	-	0.629	-	0.623	-
W^{2}NER MF1	0.654	-	0.719	-	0.705	-
Ensemble MF1	**0.660**	-	**0.723**	-	**0.711**	-

Table 4: Comparison of SOTA models on the three configurations of the three subtasks, using CodiEsp's official metrics. For each system and ranking, Precision (P), Recall (R) and F1-score (F1) values are reported. For subtasks CodiEsp-D and CodiEsp-P, Mean Average Precision (MAP) values are also included. MF1 systems use the probability and similarity thresholds that maximize the F1-score, whereas MMAP systems use the thresholds that maximize MAP values. The best performance of each metric and evaluation is highlighted in bold, while the second-best performance is underlined. The blank cells indicate that the respective research did not report a result for the corresponding metric or evaluation.

better MAP values, with systems focused on improving MAP also achieving higher Recall, although including more instances may lead to a decrease in Precision. Additionally, the upper-bound results demonstrate that there is still room for improvement in both the classification and NER-based models.

Future work should focus on further refining the NER and classification stages, particularly focusing on improving Recall and MAP values. Additionally, addressing the challenges with procedures, such as the existence of incomplete codes, could increase the system's performance. A promising direction for future research includes improving the phase 3 similarity model, as suggested by the better results in the train+dev evaluation.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the OBSER-MENH Project (MCIN/AEI/10.13039 and NextGenerationEU"/PRTR) under Grant TED2021-130398B-C21, and EDHER-MED under grant PID2022-136522OB-C21 as well as by the Universidad Nacional de Educación a Distancia (UNED), Spain within project SICAMESP (2023-VICE-0029).

References

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Almagro, M., R. M. Unanue, V. Fresno, and S. Montalvo. 2020. ICD-10 coding of Spanish electronic discharge summaries: An extreme classification problem. *IEEE Access*, 8:100073–100083.

Barreiros, L., I. Coutinho, G. M. Correia, and B. Martins. 2025. Explainable ICD

Coding via Entity Linking. *arXiv preprint arXiv:2503.20508*.

Barros, J., M. Rojas, J. Dunstan, and A. Abeliuk. 2022. Divide and conquer: An extreme multi-label classification approach for coding diseases and procedures in Spanish. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 138–147.

Blanco, A., A. Casillas, A. Pérez, and A. D. de Ilarraza. 2019. Multi-label clinical document classification: Impact of label-density. *Expert Systems with Applications*, 138:112835.

Blanco, A., A. Pérez, and A. Casillas. 2020. IXA-AAA at CLEF eHealth 2020 CodiEsp. Automatic Classification of Medical Records with Multi-label Classifiers and Similarity Match Coders. In *CLEF (Working Notes)*.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199.

Chen, P.-F., S.-M. Wang, W.-C. Liao, L.-C. Kuo, K.-C. Chen, Y.-C. Lin, C.-Y. Yang, C.-H. Chiu, S.-C. Chang, F. Lai, et al. 2021. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Medical Informatics*, 9(8):e23230.

Cossin, S. and V. Jouhet. 2020. IAM at CLEF eHealth 2020: Concept Annotation in Spanish Electronic Health Records. In *CLEF (Working Notes)*.

Costa, J., I. Lopes, A. V. Carreiro, D. Ribeiro, and C. Soares. 2020. Fraunhofer AICOS at CLEF eHealth 2020 Task 1: Clinical Code Extraction From Textual Data Using Fine-Tuned BERT Models. In *CLEF (Working Notes)*.

de la Iglesia, I., A. Atutxa, K. Gojenola, and A. Barrena. 2023. EriBERTa: A bilingual pre-trained language model for clinical natural language processing. *arXiv preprint arXiv:2306.07373*.

Duque, A., H. Fabregat, L. Araujo, and J. Martinez-Romo. 2021. A keyphrase-based approach for interpretable ICD-10 code classification of Spanish medical reports. *Artificial Intelligence in Medicine*, 121:102177.

García-Santa, N., K. Cetina, L. Cappellato, C. Eickhoff, N. Ferro, and A. Nevéol. 2020. FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding. In *CLEF (Working Notes)*.

Li, J., H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, and F. Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

López-García, G., J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas. 2021. Transformers for clinical coding in Spanish. *IEEE Access*, 9:72387–72397.

López-García, G., J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas. 2023. Explainable clinical coding with in-domain adapted transformers. *Journal of Biomedical Informatics*, 139:104323.

Miranda-Escalada, A., A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger. 2020. Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. *CLEF (Working Notes)*, 2020.

O'malley, K. J., K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton. 2005. Measuring diagnoses: ICD code accuracy. *Health services research*, 40(5p2):1620–1639.

Pereira, S., A. Névéol, P. Massari, M. Joubert, and S. Darmoni. 2006. Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. In *MIE*, pages 845–850. Citeseer.

Pérez, J., A. Pérez, A. Casillas, and K. Gojenola. 2018. Cardiology record multi-label classification using latent Dirichlet allocation. *Computer methods and programs in biomedicine*, 164:111–119.

Ramshaw, L. A. and M. P. Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, pages 157–176.

Reimers, N. and I. Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Shaban-Nejad, A., M. Michalowski, and D. L. Buckeridge. 2021. Explainability and interpretability: keys to deep medicine. *Explainable AI in healthcare and medicine: Building a culture of transparency and accountability*, pages 1–10.

Xie, P. and E. Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076.

Yang, Y., H. Lin, Z. Yang, Y. Zhang, D. Zhao, and L. Luo. 2025. LCDL: Classification of ICD codes based on disease label co-occurrence dependency and Long-Former with medical knowledge. *Artificial Intelligence in Medicine*, 160:103041.

Yu, P., L. Merrick, G. Nuti, and D. Campos. 2024. Arctic-Embed 2.0: Multilingual Retrieval Without Compromise. *arXiv preprint arXiv:2412.04506*.

Zhou, L., C. Cheng, D. Ou, and H. Huang. 2020. Construction of a semi-automatic ICD-10 coding system. *BMC medical informatics and decision making*, 20:1–12.

Zhou, T., P. Cao, Y. Chen, K. Liu, J. Zhao, K. Niu, W. Chong, and S. Liu. 2021. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957.

Zweigenbaum, P. 1999. Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, 2(3):27–47.