ICD-10 coding based on semantic distance: LSI_UNED at CLEF eHealth 2020 Task 1 ⋆

Mario Almagro[1][0000−0003−4339−2959], Raquel Martínez[1][0000−0003−1838−632X], Víctor Fresno[1][0000−0003−4270−2628], Soto Montalvo[2][0000−0001−8158−7939], and Hegler Tissot[3][0000−0003−4635−451X]

[1] National University of Distance Education (UNED), 28040 Madrid, Spain
{malmagro, raquel, vfresno}@lsi.uned.es
[2] King Juan Carlos University (URJC), 28933 Madrid, Spain
soto.montalvo@urjc.es
[3] University College London (UCL), London NW1 2DA, UK
h.tissot@ucl.ac.uk

Abstract. This paper describes our contribution to the CLEF eHealth 2020 Task 1, consisting of the CIE-10-ES annotation of Spanish Electronic Health Records (EHRs). CIE-10-ES coding is the extended version of the ICD-10 in Spain. One of the sub-tasks is aimed at the interpretability of proposals, which is in line with the latest demands in Natural Language Processing (NLP). Moreover, ICD-10 entries generated by hospitals usually follow an extreme distribution, involving complex annotation challenges. For that reason, an unsupervised semantic similarity-based method has been explored using a representation based on SNOMED-CT clinical terminology. Since example-based learning is able to capture complex patterns, the proposal has been combined with Gradient Boosting methods to model the codes with more instances. mAP scores of 0.517 are achieved for CIE-10-ES codes associated with diagnoses and 0.398 for CIE-10-ES procedure codes. The mixed approach improves the strict supervised proposals by more than 38% and 13% respectively. Finally, the unsupervised component is used to provide code evidences in EHRs exploiting a greater interpretability.

Keywords: Semantic similarity · ICD-10 coding · CIE-10-ES coding · Ensemble method

⋆ Supported by the Spanish Ministry of Science and Innovation through the MAMTRA-MED Project, AEI/FEDER, UE, under Grant TIN2016-77820-C3-2-R, and the INDICA-MED Project, AEI/FEDER, UE, under Grant PID2019-106942RB-C32.

1 Introduction

Health services produce vast amounts of data every day. A significant proportion of this information is captured by clinicians in EHRs. Although EHRs may contain structured information such as medical test results, both the patient's history and the clinician's judgment are often described in natural language, requiring more flexibility and higher level of customization to capture all the details.

The clinical domain is characterized by a very diverse language, full of acronyms, misspellings, large specialized vocabularies, unstructured phrases, and a substantial richness of granularities. While there is a wide range of general NLP tools, the reduced accessibility of biomedical texts hinders the implementation of domain techniques to deal with all these complexities. In turn, the non-domain tools do not seem to fit properly into the requested tasks because clinical language involves a lack of lexical standardization. Therefore, syntactic rules are often relaxed with respect to texts in the general domain.

One of the tasks with the greatest impact on hospital funding is the ICD-10 coding, which aims to translate causes of morbidity and mortality written in natural language into structured data that can be quantified for statistical analysis. ICD-10 coding is a high-level task, which requires dealing with strong lexical variability, language understanding for document-level decision making, and extremely unbalanced data distributions. These restrictions are accompanied by the General Data Protection Regulation (GDPR) applied by Europe, which greatly reduces private access to patient data in order to preserve privacy. Limited access to clinical data and frequent associated biases often result in the scarcity of data and the absence of numerous codes within available data sets. These challenge raise questions about the viability of some data-driven models. Thus data augmentation, transfer learning, and unsupervised techniques become particularly relevant.

In this paper we present our contribution to the CLEF eHealth 2020 Task 1 [17], which aims at CIE-10-ES coding of EHRs. An unsupervised similarity-based method that can suggest codes not present in the training data is proposed. In addition, a boosting method is analysed to reveal the limitations of data-driven systems. Finally, both methods are combined to exploit the provided data but mitigating the scarcity of instances.

2 Related Work

This is not the first time that the Conference and Labs of the Evaluation Forum (CLEF) encourages the proposal of methods to solve ICD coding task. Between 2016 and 2018, tasks have been organized for the classification of death notes. The documents used consist of a couple of short lines of text describing the diagnoses. The correspondence per sentence is 'zero-to-many' codes per sentence. Although some sentences do not contain diagnoses, the tendency is for the relevant information to be highly condensed.

Unsupervised approaches have been explored in these conditions. For example, Van Mulligen et al. [25] and Ho-Dac et al. [13] focus on applying Information Retrieval (IR) techniques to search for similar instances, Cabot et al. deal with coding as a Name Entity Recognition (NER) task [5] and Cossin et al. use lexical similarity based on Levenshtein distance to assign codes [7]. Multiple feature-based approaches have also been proposed, such as linear regression [16], Naive Bayes (NB) [4], and random forests [15]. In addition, sequence-to-sequence approaches have been explored [23, 10, 3], in which an encoder-decoder structure is used to transform a sequence of words into a sequence of codes.

In 2019, the codification of summaries of animal experiments into ICD-10 subchapters was proposed [19]. The summaries provided include paragraphs where the relevant information is more sparse. In contrast, the number of classes is severely reduced to a few categories, significantly simplifying the problem. Two models based on BERT [8] are proposed exploiting fewer and more frequent labels. Amin et al. use the multilingual BERT-Base model directly on the German reports [2], and Sänger et al. fine-tune the BioBERT model [14] on the reports translated into English [22]. The latter reaches the best results. In line with the unsupervised proposals, Ahmed et al. explore the application of k-Nearest Neighbor (kNN) [1].

The current proposed task is more in line with the need of hospitals. It consists of coding complete EHRs, which are longer than the documents of previous years, using all the official codes. In this paper we have followed proposals for semantic similarity between terms and codes such as those proposed by Ning et al. [20] and Chen et al. [6], but at the document level, with all the difficulties that this entails.

3 Proposed Approach

We explore a semantic similarity-based method for CIE-10-ES coding to deal with the abundance of biases in the data.

We start by using an NER method by associating SNOMED-CT concepts to official CIE-10-ES descriptions and EHRs text. Subsequently, the method estimates the similarity between concepts using the hierarchical structure of SNOMED-CT. Once the similarities between concepts are known, code affinity to text is computed as the similarity between sets: the concepts defining a code and those concepts within a piece of EHRs text. Finally, the code ranking associated to a document is established by ordering code relatedness to all pieces of EHRs text. In addition, the combination of the resulting ranking with a supervised learning method has been explored. An overview of our approach and the overall pipeline is shown in Figure 1. More details of each process are given below.

3.1 SNOMED-CT association

As early mentioned, access to Spanish clinical data is very restricted, which complicates modeling Spanish representations based on distributional semantics

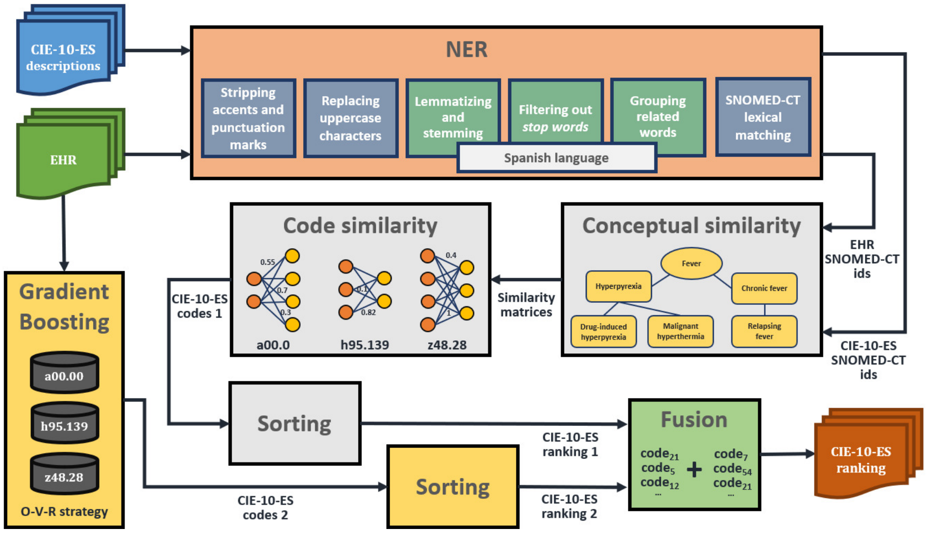


Fig. 1: Diagram of the pipeline. The red block *NER* assigns SNOMED-CT concepts to both ICD descriptions and EHRs; the grey blocks *Conceptual similarity*, *Code similarity* and *Sorting* compute the similarities between the SNOMED-CT concept sets, and sort the codes according to their affinity; the yellow blocks *Gradient Boosting* and *Sorting* predict and rank codes based on Bags of Words (BoW) features; and finally the green block *Fusion* combines both rankings in the final output.

with a sufficiently diverse vocabulary. Accordingly, domain knowledge bases are used to provide an overall representation of all the different concepts.

One of the most widely used standardized clinical terminology is SNOMED-CT [9], which hierarchically encompasses more than 300,000 unique clinical concepts and 70,000 unique clinical terms. SNOMED-CT is designed as a low-level terminology to deal with lexical diversity rather than abstract concepts, covering a broad scope. The assignment of SNOMED concepts to descriptions and EHRs has been done through a partial lexical matching. For this purpose, a pre-processing step has been firstly carried out by stripping accents and punctuation marks, converting text to lowercase, lemmatizing and stemming, removing stop words, grouping pertainym words (terms "pertaining to" others, such as pulmonary and lung), and handling term exclusions.

A Spanish lemmatizer built from WordNet [11] and ConceptNet [24] has been used to standardize the text with a greater coverage. Lexical disambiguation does not seem to be particularly relevant to the task, so the use of this knowledge-based lemmatizer has been preferred to other tool based on supervised models, such as spaCy[4] for general domain and IxaMed [12] for clinical domain. All the words not found by the lemmatizer are subsequently stemmed. In addition, a

[4] https://spacy.io

list of pertainyms has also been generated using WordNet in conjunction with machine translation techniques and human supervision.

3.2 Similarity computation

SNOMED-CT organizes concepts using hierarchical 'is-a' relationships, so it is relatively easy to estimate similarities between nodes according to some of the well-known semantic similarity measures in graphs. Multiple similarity measures based on path or Information Content (IC) are described in [21]. Although path-based measures are simpler, linearity in the hierarchy is assumed, i.e. 'is-a' relationships are equally relevant in general and specific concepts. For this reason, IC-based measures seem to better fit this task, suggesting greater similarity for specific nearby concepts than for general close concepts. In particular, the Lin measure (Equation 1) has been applied in this proposal.

$$S_{lin}(c_1, c_2) = \frac{2 * IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{1}$$

where c_1 and c_2 are the couple of concepts, lcs is the lowest common subsumer, and IC is the Information Content, which is usually computed with the frequency of concepts in large corpus. A fixed IC value is assumed based on the depth of the node in the branch in order to avoid corpus dependency and accessibility constraints.

The similarity between code sets (S_G in Equation 2) can be seen as a problem of maximizing pair arguments. Pair assignments can be defined as a bipartite graph $G = (V, E)$, with the vertices V being the two sets of codes and the edges E being the similarities between codes. We use the Kuhn–Munkres algorithm [18] to solve the optimization problem.

$$S_G = \frac{max \sum_{i=1}^{N_{code}} \sum_{j=1}^{N_{ehr}} M_{lin}(i,j)B(i,j)}{N_{code}} \tag{2}$$

where N_{ehr} is the number of SNOMED-CT concepts in the piece of document, N_{code} is the number of SNOMED-CT concepts in the code description, $M_{lin}(i,j)$ is a matrix with the Lin similarity values, and $B(i,j)$ is a binary value which is only active if concept i has been paired with concept j. There is only one positive value of B for each i.

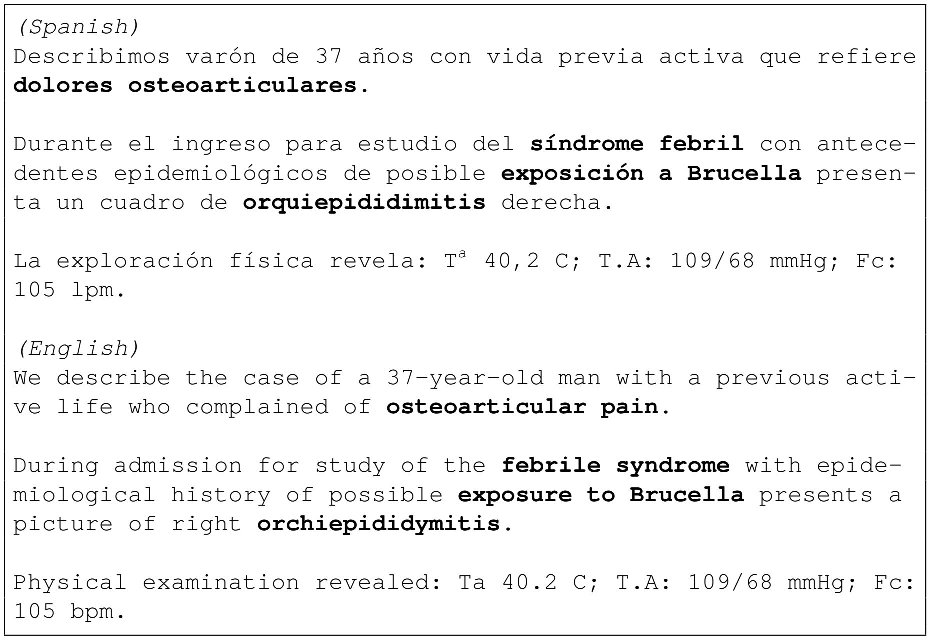
Once the similarity values between all codes and each of the pieces of a given document (S_G) are calculated, a first ranking is created by sorting the codes by S_G at the document level. The final ranking is subsequently produced by recalculating all code similarity values through an iterative exclusion of the SNOMED-CT concepts already used by the codes at the top of the first ranking. With this second computation, the number of codes associated with a single SNOMED-CT concept subset is limited to only one.

3.3 Supervised learning

A priori, one would expect worse accuracy values for sub-tasks 1 and 2 due to lack of learning for complex patterns and better recall values by accessing all possible codes and a wide variety of vocabulary. We have explored such complex decisions by implementing a Gradient Boosting multi-label algorithm, based on binary classifiers using a ‘one-vs-the-rest’ (OvR) strategy. These classifiers chain a series of consecutive learning models, iteratively emphasizing the mistakes made by the previous model. Boosting techniques seem to produce better results for these types of problems where significant imbalance is the main factor. Finally, we rank the codes by prioritizing the predictions of the Gradient Boosting classifiers, which have been ordered according to the confidence values. The predictions are followed by the codes suggested by the similarity-based approach.

4 Experiments

The following subsections describe the used data, the proposal setup and the achieved results.



(Spanish)
Describimos varón de 37 años con vida previa activa que refiere **dolores osteoarticulares.**

Durante el ingreso para estudio del **síndrome febril** con antecedentes epidemiológicos de posible **exposición a Brucella** presenta un cuadro de **orquiepididimitis** derecha.

La exploración física revela: T^{a} 40,2 C; T.A: 109/68 mmHg; Fc: 105 lpm.

(English)
We describe the case of a 37-year-old man with a previous active life who complained of **osteoarticular pain.**

During admission for study of the **febrile syndrome** with epidemiological history of possible **exposure to Brucella** presents a picture of right **orchiepididymitis.**

Physical examination revealed: Ta 40.2 C; T.A: 109/68 mmHg; Fc: 105 bpm.

Fig. 2: Examples of Spanish and English sentences in the CodiEsp-SPACCC corpus – expressions associated with CIE-ES-10 codes are shown in bold.

4.1 Data sets

CIE-10-ES coding challenge has been evaluated in the CodiEsp-SPACCC corpus[5]. It consists of 1000 EHRs, with an average of 16.5 long sentences per document, and split into training, development, and test data sets. Each document usually contains a wide range of information, including medical history, medical examinations, test results, and clinical judgments, all without a predefined structure. A piece of document is shown in Figure 2.

The examples illustrate a series of evidences distributed along the text. The CIE-ES-10 associated with the expressions found in the example are collected in Table 1. The terms used by clinicians in EHRs differ in granularity from those defining code descriptions. ICD descriptions tend to be more abstract in order to capture multiple cases, which gives the coding task a high degree of complexity. Fortunately, SNOMED-CT handles these granularities, extending the more general concepts in the hierarchical structure.

Table 1: English evidences from the example in Figure 2.

CIE-ES-10 code	Description	Evidence
m25.50	Pain in unspecified joint	osteoarticular pain
r50.9	Fever, unspecified	febrile syndrome
z20.818	Contact with and (suspected) exposure to other bacterial communicable diseases	exposure to Brucella
n45.3	Epididymo-orchitis	orchiepididymitis

In terms of data distribution, ICD-10 codes tend to follow extreme distributions, characterized by a large imbalance, a scarcity of instances for most codes,

[5] https://zenodo.org/record/3837305#.XwMHiBHtZH5

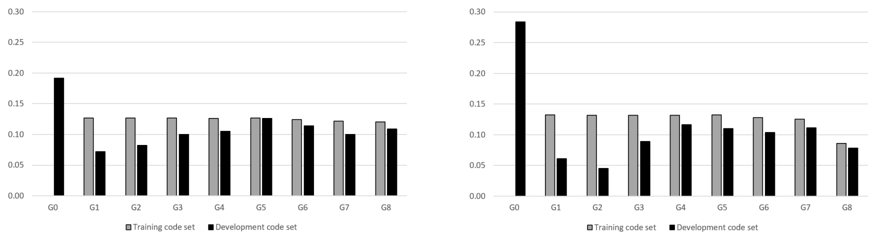
Table 2: Code group statistics.

	Diagnoses			Procedures		
Groups	Freq. range	# labels	# instances	Freq. range	# labels	# instances
G1	1	704	704	1	193	193
G2	2	499	704	2	186	192
G3	3-4	254	704	3	84	192
G4	5-7	144	701	4-6	45	192
G5	8-11	82	704	7-9	26	193
G6	12-20	46	688	11-19	13	186
G7	21-45	22	675	20-33	7	182
G8	>45	9	666	>33	2	125

and the presence of hospital biases. In this case, such a power-law distribution can be seen in Table 2. The training codes are ordered according to frequency and clustered in 8 groups, trying to gather a similar number of instances.

Table 2 shows that the nine most frequent diagnoses in group G8 appear about the same number of times as the 704 least frequent diagnoses in group G1. In the case of procedures, those 13 codes that appear between 11 and 19 times (group G6) are equivalent in volume of entries to the 84 codes of group G3, which appear 3 times each.

Figure 3 plots a normalized histogram of these groups for the training (in grey) and development (in black) data sets. A ninth group (G0) has been included to represent all unseen codes in the training data set. The training and development distributions for both diagnoses and procedures are different, with the unseen codes reaching almost 20% and 30% of instance volume, respectively. The lack of information on labels implies a huge problem for data-driven approaches as a considerable percentage of the codes are left out of the model. For this reason, we have proposed a method that directly uses CIE-10-ES descriptions and does not require training data.



(a) Diagnosis Code Frequency Histogram (b) Procedure Code Frequency Histogram

Fig. 3: Comparison of code relative frequency histograms for the training and development data. The codes have been grouped by frequency ranges so that similar numbers of training instances are held (except for the first group where unseen codes are shown).

4.2 Experimental Setting

The CLEF eHealth 2020 Task 1 [17] is structured into 3 sub-tasks related but with different objectives: the suggestion of CIE-10-ES codes corresponding to diagnoses (CodiEsp-D), the recommendation of CIE-10-ES codes associated with procedures (CodiEsp-P), and the prediction of both codes facilitating evidences within the report (CodiEsp-X). The first two sub-tasks consist of generating a ranking of codes ordered by their affinity to a given EHR, while the objective of the last task is to retrieve only those codes that are most probable, together with the evidence. Based on these objectives, we propose different settings for each sub-task.

We explore a similarity-based method including the specifications described in Section 3. The evidences of the codes in the training data set has been used as part of the descriptions to feed the method with more specific information about the codes. Furthermore, we exploit a digitized version of the CIE-10-ES tabular list, including additional code descriptions in combination with the official entries provided by the organizers. Finally, we propose two approaches in the first two sub-tasks, one method without grouping pertainym words (**SIM-BASIC**) and another using these related words (**SIM-EXT**). As for the third sub-task, we implement the approach that does not group related words, using different similarity thresholds to choose the retrieved codes per document. The thresholds 0.7 (**SIM-BASIC-7**), 0.8 (**SIM-BASIC-8**), and 0.9 (**SIM-BASIC-9**) have been applied.

As for the supervised learning method, we explore a Gradient Boosting algorithm for sub-tasks 1 and 2. The model is also trained on Spanish abstracts from Lilacs and Ibecs annotated with CIE-10-ES codes in addition to the training data set. These data sets are also provided by the organizers and reach the amount of 355,840 abstracts, with an average of 2.5 codes per abstract. The distribution of these codes is also extreme unbalanced, so a subsampling is performed during the training to avoid excessively increasing the negative instances per code. Regarding the representation, classic BoW features are applied due to the presence of a large volume of codes with less than 20 instances. In particular, label-specific features (**GB-BNS**) are used in order to focus learning and prediction on code-relevant patterns. For this purpose, we extract the term frequency weighed by Bi-normal separation (TF-BNS). Global features such as TF-IDF (**GB-IDF**) are also used for procedures as this data set has few labels and less statistical information. None of these approaches are proposed in the last sub-task because those lack sufficient interpretability.

We implement a final approach (**GB-SIM**) for the first two sub-tasks by combining the Gradient Boosting method that uses TF-BNS features (**GB-BNS**) and the similarity-based method that groups related words together (**SIM-EXT**). As a result, the codes predicted by the classifiers are ranked according to the confidence value and merged with the ranking generated by the similarity method. Codes suggested by classifiers are placed at the top of the new ranking assuming that supervised learning methods generally lead to higher precision values.

4.3 Evaluation

Two different evaluation metrics have been defined for the CLEF eHealth 2020 Task 1 according to the objectives of each sub-task: mAP and F1.

The first two sub-tasks aim to generate a list of codes per EHR ranked by relevance and are evaluated with mAP in order to quantify how many significant codes are in the top positions. mAP is specified below.

$$AP = \frac{\sum_{k}^{n} Precision@k \times rel@k}{TP + FN} \tag{3}$$

$$mAP = \frac{\sum_i^N AP_i}{N} \tag{4}$$

where n and N are the total number of retrieved codes, $Precision@k$ is the precision considering the top k codes, $rel@k$ is a relevance function, and AP_i is the Average Precision for the i top codes.

In contrast, the last sub-task is designed to associate CIE-10-ES codes to EHRs in an explainable way, providing the text fragment containing the keywords. In this case, the number of retrieved and successful codes is assessed computing F1.

4.4 Results

Metrics are calculated using all the codes described in the CIE-10-ES coding and only those that appear in the training data set. The results of sub-tasks 1 and 2 are shown in Table 3. Approaches based on semantic similarity obtain better results than the supervised method, considering both all codes and only those seen during training. SIM-BASIC achieves an mAP score of 0.51, significantly higher than the 0.37 mAP score for GB-BNS, with a difference of 0.14. This seems to support the ability of SIM-BASIC to predict unseen or rare codes as opposed to the need for higher volumes of instances in GB-BNS. Indeed, the difference between scores when evaluating only seen codes is slightly reduced, reaching the mAP scores of 0.596 and 0.475 respectively. Apparently, Gradient Boosting method fails to model rare codes. Moreover, better results are generally obtained with the GB-SIM approach, which combines the characterization of codes with little information from SIM-BASIC and the learning of more complex patterns from SIM-BNS.

Table 3: Results of the approaches for CodiEsp-D and CodiEsp-P sub-tasks.

Approach	CodiEsp-D		CodiEsp-P	
	(All codes) mAP	(Only seen codes) mAP	(All codes) mAP	(Only seen codes) mAP
SIM-BASIC	**0.517**	0.596	0.366	0.421
SIM-EXT	0.493	0.571	0.376	0.440
GB-IDF	-	-	0.351	0.403
GB-BNS	0.372	0.475	0.310	0.345
GB-SIM	0.511	**0.612**	**0.398**	**0.457**

Precision, Recall and F1 are shown in Table 4 for third task. There are different methods to improve the interpretability of supervised models such as distillation into decision trees. However, exploring these lines has not been the purpose of this paper. Thus, only similarity-based approaches have been used to predict codes by retrieving textual evidence. In particular, the SIM-BASIC approach is applied with different similarity thresholds. Using a threshold of 0.9, SIM-BASIC-9 achieves an F1 score of 0.451 with all codes and 0.494 with only those seen in the training data set. Unrelated codes are apparently associated when choosing the less restrictive thresholds, resulting in a decrease in F1.

Table 4: Results of the approaches for CodiEsp-X sub-task. Precision (P) and Recall (R) are shown for all and only seen codes.

Approach	All codes			Only seen codes		
	P	R	F1	P	R	F1
SIM-BASIC-7	0.268	0.414	0.326	0.351	0.465	0.400
SIM-BASIC-8	0.397	0.413	0.405	0.443	0.464	0.453
SIM-BASIC-9	0.508	0.406	**0.451**	0.537	0.457	**0.494**

5 Conclusion and Future Work

ICD-10 coding, and in particular the Spanish extension of this task (CIE-10-ES coding), cannot be easily automated with the existing techniques. One of the main problems is the biased distribution of data that prevents the availability of sufficient data for all codes, hindering the development of supervised learning methods.

In this work, we propose an unsupervised method that improves recall by suggesting also rare or unseen training codes. Our method achieves the characterization of less frequent codes through a representation based on the clinical terminology SNOMED-CT. Finally, we found that an ensemble that introduces supervised learning methods is able to provide a better characterization of frequent codes.

We plan to extend our work by adding other intrinsic relationships of the SNOMED-CT concepts that provide alternative information. We also plan to explore the generation of similarity-based features to address few-shot and zero-shot learning.

References

1. Ahmed, N., Arıbaş, A., Alpkocak, A.: DEMIR at CLEF eHealth 2019: Information Retrieval based Classification of Animal Experiment Summaries. In: CLEF (Working Notes) (2019)
2. Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K., Wixted, M.: MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In: CLEF (Working Notes) (2019)

3. Atutxa, A., Casillas, A., Ezeiza, N., Goenaga, I., Fresno, V., Gojenola, K., Martinez, R., Oronoz, M., Perez-de Vinaspre, O.: Ixamed at clef ehealth 2018 task 1: Icd10 coding with a sequence-to-sequence approach. In: CLEF (Working Notes) (2018)
4. Bounaama, R., El Amine Abderrahim, M.: Tlemcen university at celf ehealth 2018 team techno: Multilingual information extraction-icd10 coding. In: CLEF (Working Notes) (2018)
5. Cabot, C., Soualmia, L. F., Darmoni, S. J.: Sibm at clef ehealth evaluation lab 2017: Multilingual information extraction with cim-ind. In: CLEF (Working Notes) (2017)
6. Chen, Y., Lu, H., Li, L.: Automatic icd-10 coding algorithm using an improved longest common subsequence based on semantic similarity. PloS one **12**(3), e0173410 (2017)
7. Cossin, S., Jouhet, V., Mougin, F., Diallo, G., Thiessard, F.: Iam at clef ehealth 2018: Concept annotation and coding in french death certificates. In: CLEF (Working Notes) (2018)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), http://arxiv.org/abs/1810.04805
9. Donnelly, K.: SNOMED-CT: The advanced terminology and coding system for eHealth. In: Studies in health technology and informatics, vol. 121, pp. 279 (2006).
10. eblee, S., Budhkar, A., Milic, S., Pinto, J., Pou-Prom, C., Vishnubhotla, K., Hirst, G., Rudzicz, F.: Toronto cl at clef 2018 ehealth task 1: Multi-lingual icd-10 coding using an ensemble of recurrent and convolutional neural networks. In: CLEF (Working Notes) (2018)
11. Fellbaum, C., Miller, G.: WordNet: An Electronic Lexical Database. 2nd edn. MIT press (1998)
12. Gojenola, K., Oronoz, M., Pérez, A., Casillas, A.: IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts. In: The 8th International Workshop on Semantic Evaluation, SemEval@ COLING 2014, vol. 1, pp. 361–365. Dublin, Ireland (2014).
13. Ho-Dac, L.-M., Fabre, C., Birski, A., Boudraa, I., Bourriot, A., Cassier, M., Delvenne, L., Garcia-Gonzalez, C., Kang, E.-B., Piccinini, E., et al.: Litl at clef ehealth2017: automatic classication of death reports. In: CLEF (Working Notes) (2017)
14. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **4**(34), 1234–1240 (2020)
15. Li, M., Xu, C., Wei, T., Bao, D., Lu, N., Yang, J.: Ecnu at 2018 ehealth task1 multilingual information extraction. In: CLEF (Working Notes) (2018)
16. López-Úbeda, P., Diaz-Galiano, M., Martin-Valdivia, M., Urena-López, L. A.: Machine learning to detect icd10 codes in causes of death. In: CLEF (Working Notes) (2018)
17. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of eHealth CLEF 2020. In: CLEF (Working Notes) (2020)
18. Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics **1**(5), 32–38 (1957)
19. Neves, M., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the CLEF eHealth 2019 Multilingual Information Extraction. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., et al. (eds.) Experimental

IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Lecture Notes in Computer Science. Springer, Berlin Heidelberg, Germany (2019)

20. Ning, W., Yu, M., Zhang, R.: A hierarchical method to automatically encode chinese diagnoses through semantic similarity estimation. BMC medical informatics and decision making **16**(1), 1–12 (2016)
21. Pedersen, T., Pakhomov, S., Patwardhan, S., Chute, C.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics **3**(40), 288–299 (2007)
22. Sänger, M., Weber, L., Kittner, M., Leser, U.: Classifying German Animal Experiment Summaries with Multi-lingual BERT at CLEF eHealth 2019 Task 1. In: CLEF (Working Notes) (2019)
23. Seva, J., Sänger, M., Leser, U.: Wbi at clef ehealth 2018 task 1: Language independent icd-10 coding using multi-lingual embeddings and recurrent neural networks. In: CLEF (Working Notes) (2018)
24. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4444–4451. San Francisco, California USA (2017).
25. Van Mulligen, E. M., Afzal, Z., Akhondi, S., Dang, D., Kors, J.: Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. In: CLEF (Working Notes) (2016)