MAMTRA-MED at CLEF eHealth 2018: A Combination of Information Retrieval Techniques and Neural Networks for ICD-10 Coding of Death Certificates

Mario Almagro[1], Soto Montalvo[2], A. Díaz de Ilarraza[3], and A. Pérez[4]

[1] Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain
malmagro@lsi.uned.es
[2] Universidad Rey Juan Carlos (URJC), Madrid 28933, Spain
soto.montalvo@urjc.es
[3] University of the Basque Country UPV/EHU (IXA NLP), Donostia-San Sebastian 20018, Spain
a.diazdeilarraza@ehu.eus
[4] University of the Basque Country UPV/EHU (IXA NLP), Bilbao 48013, Spain
alicia.perez@ehu.eus

Abstract. This paper describes the systems proposed by LSI_UNED team in Task 1 of the CLEF eHealth 2018 challenge. The main objective is the automatic coding of death certificates in French, Italian and Hungarian languages according to the ICD-10. This task has been tackled through supervised learning methods such as neural networks, and techniques based on Information Retrieval (IR) systems. The first approach has been implemented by training one model for each of the most frequent ICD-10 codes in the corpus. For this purpose, a bag-of-words approach has been applied using the TF-BNS value for terms contained in death certificate statements. As for the IR approach, Lucene has been used as a search engine, indexing dictionaries and the content of the death certificates in the training corpus. Finally, a combination of both methods has been proposed to balance precision and recall, using the IR system for diseases not classified by any learning model. Similar F_1 scores are obtained on the test datasets of each language by supervised methods and the combined system giving the latter greater recall values.

Keywords: ICD-10 Coding · ICD-10 Codes · Information Retrieval · Neural Networks · Deep Learning · CepiDC · Biomedical Text Mining · CLEF eHealth.

1 Introduction

The amount of health text available in electronic format is immense —scientific papers, websites, forums, social networks or Electronic Health Records (EHRs)— so managing health information to support medical decisions is no easy task. An example of this complexity is the analysis of numerous clinical texts generated

by health care centres, which requires a large amount of resources that are often unavailable. The 2018 CLEF eHealth Evaluation Lab [9] is intended to address these challenges through different tasks aimed at facilitating access to health information.

Our proposals have focused on the resolution of the CLEF task [6] for the ICD-10 coding of death certificates in French, Italian and Hungarian. The classification of clinical texts according to the International Classification of Diseases (ICD) is one of the most pressing problems in hospital management due to its statistical purposes for morbidity and mortality. The 10th version of this coding assigns a unique identifier of between 3 and 7 alphanumeric symbols to disorders, grouping together nearly 16,000 possible diagnoses with a wealth of nuances.

As a possible resolution to the described task, different approaches supported by supervised learning and search engines have been proposed in this paper. Due to the nature of the ICD-10 codes, the data generally present a very biased distribution, with a small set of frequently occurring diagnoses [1]. The same distribution can be seen in the corpus used in this task, and therefore, it is considered the combination of approaches that could try to maximize precision —such as machine learning approximations— and methods that tend to maximize recall —such as the search engines on which Information Retrieval (IR) systems are built— could be of great interest. These two aspects will give rise to joint proposals.

2 Related Work

In general, the approaches used in the state of the art for the recommendation and assignment of ICD-10 codes can be divided into two groups: those based on medical language processing (MLP) and those based on classification techniques.

The first ones use unsupervised techniques to find correspondences between the concepts in standard descriptions and health concepts identified through medical knowledge bases and ontologies in health documents. For example, Ning et al. [7] apply an example-based model generated from a Chinese terminology containing correspondences with 4-digit ICD-10 codes, thus taking advantage of the hierarchical structure in the standard coding. Chen et al. [2] explore semantic similarity by applying the Longest Common Subsequence (LCS) method to the diagnoses and names given by ICD-10 codes. Other systems following this trend have participated in previous versions of the CLEF task [3, 10].

On the other hand, the second approaches generate classifiers by using supervised learning algorithms. Zweigenbaum and Lavergne [11] apply two classifiers: one trained with a set of EHRs, and the other trained with different medical dictionaries; Miftakhutdinov and Tutubalina [5] use word embeddings trained from a corpus of medical user opinions, along with recurrent neural networks to assign codes.

At the same time, mixed approaches combining both methods can be found. For example, Seva et al. [8] use an IR approach to search for possible candidate ICD-10 codes in different dictionaries, along with several classifiers to filter them.

Jatunarapit et al. [4] employ English corpus-based classifiers and a set of IR techniques to establish similarities with Thai terms.

3 Proposed Approach

In this paper we have explored two alternative approaches for the assignment of ICD-10 codes to death certificates. On the one hand, a supervised approach is proposed through Vector Support Machines (SVMs) and neural networks that aims to take advantage of the training corpus by generating One-Vs-Rest (OVR) models for the most frequent ICD-10 codes. The dependence on examples makes this approach less robust in the face of the possibility of coding new diseases, given the immense number of ICD-10 codes with little or no representation in the corpus. For this reason, it is also proposed to complement learning models with an unsupervised approach based on IR techniques to achieve greater recall.

The machine learning approach is based on the training of a binary model for each of the target ICD-10 codes, indicating the presence or absence of the code. As this is a multiclass and multilabel problem, the coding of diseases is carried out by grouping the results of all the binary classifiers. The particularization of a model for each ICD code allows the processing of the data adapted to each class, as will be seen later when applying the weighting with Bi-Normal Separation (BNS). To implement these models, different configurations have been developed with linear SVMs and Multi-Layer Perceptrons (MLPs).

The proposed IR approach consists of a search engine in which information relating to codes has been indexed, both terminology from provided dictionaries and associated sentences from training data. In this way, the coding of death certificates is reduced to the generation of queries based on their terms, choosing the result with the highest score. As a drawback, retrieving a fixed number of results (in this case only one) implies the loss of the ability to adapt the number of codes assigned to a line in death certificates, which may contain several disorders.

These two approaches yield different results. As expected in the experimentation, while the supervised approach achieves higher precision values, a hybrid method involving IR techniques ensures a better balance between the correct coding rate and the number of different codes capable of coding.

4 Experiments

4.1 Datasets

The training data is organized in three separate corpus, one for each language: French, Italian and Hungarian. Although each corpus is structured in two modes —*aligned*, for line-level annotation, and *raw*, for document-level annotation—, line-level annotation is only available for the French corpus. For this reason, this proposal has focused only on coding at document level for all three languages.

Each corpus has different metadata on diagnoses, dictionaries with equivalences between ICD-10 codes and terms, the text of death certificates and line-level equivalences between its processed content and ICD-10 codes.

Table 1. Overall corpus statistics. The overlap of codes is intended to measure the percentage of the different types of ICD-10 disorders that are present in other corpora.

	Italian	Hungarian	French	Total
Number of certificates in training	14,502	84,702	65,843	165,047
Number of certificates in test	3,617	21,175	24,375	49,167
Number of ICD-10 codes	60,954	392,019	527,940	980,913
Number of unique ICD-10 codes	1,442	3,123	3,829	5,011
Overlapping with Italian codes	100%	34%	30%	-
Overlapping with Hungarian codes	73%	100%	57%	-
Overlapping with French codes	79%	70%	100%	-

A general summary of the amount of data grouped by corpus is given in Table 1. Although the Hungarian corpus contains more death certificates, the number of ICD-10 codes present in the French corpus far exceeds that of the rest. As can be seen, most of the ICD-10 codes in the Italian corpus are also present in some of the other corpora, with an average overlap of 76%. Given this overlapping, a large part of the results achieved on the Italian corpus could be considered extrapolable since the model is expected to behave similarly in at least the same codes. For this reason, the experimentation shown in this paper is only carried out on it, taking advantage of its lower volume.

In terms of distribution, the frequency of codes follows a power law, with most of the entries corresponding to a small group of codes. This implies that a supervised approach alone has a more restrictive limit to improvement than other techniques.

4.2 Experimental Setting

Regardless of the approach used, a common pre-processing has been developed. A lowercase conversion and accent removal has been applied, as well as a stop word filter and a stemming process for each language.

Supervised approach Here the problem has been addressed through the combination of different binary classifiers, each one determining the presence or absence of a specific code. Due to the scarcity of data in training collections on some ICD-10 disorders, model generation has been limited to only those ICD codes that appear more than a certain number of times. With this it is understood that the rest of the ICD codes (those absent in the corpora or with little presence) cannot be represented by supervised models since data lack sufficient examples with which to abstract the corresponding patterns.

In order to find the configuration that best suits this task, multilayer perceptrons with different numbers of neurons and hidden layers have been implemented, as well as variations of the rest of the hyperparameters. In addition, linear SVMs have been trained to compare the efficiency of both models in ICD-10 coding.

As for the input data, once the pre-processing has been applied, different textual representations have been used. On the one hand, the models have been generated with the frequency of terms weighted with Inverse Document Frequency (IDF) and Bi-Normal Separation (BNS) values. IDF is calculated at document level, determining the relevant terms based on the number of documents in which they appear. This may penalize those terms relevant to a class but too frequent. For this reason, the BNS feature is introduced in the experimentation, since it estimates the representation of terms at class level, avoiding this type of error. This weight is defined as $BNS = |Icdf(\ P(W \mid class+)\ ) - Icdf(\ P(W \mid class-)\ )|$, where *Icdf* is the inverse cumulative distribution function, $P(W \mid class+)$ is the probability of finding a word in the positive classes and $P(W \mid class-)$ is the probability of finding a word in the negative classes. Based on both measures, n-grams of two and three words have been considered. On the other hand, a feature filtering has been performed using Chi-Square ($\tilde{\chi}^2$). In Table 2 different configurations implemented in the experimentation are presented, which include the different options mentioned above. The structure of the MLPs shown consists of 4 hidden layers and 80 neurons each.

Table 2. Configurations for supervised learning models. The Bias parameter is the minimum frequency in the collection for considering an ICD code in the training process.

System	Models	Bias	Bag Of Words	Measure	Filtering
S_1	SVMs	10	Unigrams	Tf-Idf	50
S_2	SVMs	10	Unigrams	Tf-Bns	50
S_3	SVMs	10	Unigrams	Tf-Bns	5000
S_4	SVMs	10	Unigrams	Tf-Bns	10
S_5	MLPs	40	Unigrams	Tf-Bns	1000
S_6	MLPs	100	Unigrams	Tf-Bns	1000
S_7	MLPs	200	Unigrams	Tf-Bns	1000
S_8	MLPs	100	Bigrams	Tf-Bns	1000
S_9	MLPs	100	Trigrams	Tf-Bns	1000
S_{10}	MLPs	100	Trigrams	Tf-Idf	1000

Unsupervised IR approach This approach uses Lucene as the search engine. To enrich the indexes, the diseases present in the CépiDC dictionaries have been added as well as the content of the death certificates in the training corpus for each language. The aim is to make each of the possible ICD-10 codes accessible through a set of descriptions and associated terms. There are fewer descriptions for less common codes, so it has been decided to remove duplicate descriptions during indexing to avoid penalties. In addition, it has been considered to include the official description of codes as it appears in the ICD standard, taking advantage of the electronic versions provided by some governments.

Each query has been generated from the terms contained in each line of the death certificates. As it is a multilabel problem, the number of classes assigned to a document line varies. Since Lucene's output consists of a ranking of results, the evaluation has been carried out according to the number of results chosen for each query (1 or 2). The different configurations are presented in Table 3.

Table 3. Configurations of the IR techniques based on Lucene.

System	Method	Number of results	Using Official Descriptions
S_{11}	IR	1	Yes
S_{12}	IR	2	Yes
S_{13}	IR	1	No
S_{14}	IR	2	No

Combined approach The method based on the combination of supervised models together with search engine aims to take advantage of the effectiveness of the first ones with an increase in robustness for codes that are absent or hardly present in the training corpus. Since less common diseases —no learning model— should be left without ICD-10 code assigned after applying multiple trained models, it seems reasonable to use search engines only with those death certificate statements not classified by the supervised approach. Table 4 shows the combinations of learning model and IR system configurations that give the best results.

Table 4. Configurations for the combination of methods.

System	Combination
S_{15}	S_5 + S_{13}
S_{16}	S_6 + S_{13}
S_{17}	S_7 + S_{13}
S_{18}	S_8 + S_{13}
S_{19}	S_{10} + S_{13}

4.3 Results

The results of the configurations are only shown in the Italian corpus, as this represents to a large extent the type of disorders present in the other corpora. This choice is based on the lower number of certificates and the higher percentage of common diseases in other corpora. Different configurations have been evaluated using a k-fold cross-validation of 5 folds and a 94/6 split. The results are shown in Table 5.

Table 5. Results of the different configurations on the Italian training dataset.

Method	System	Precision	Recall	F_1 score
Supervised approach	S_1	0.91	0.59	0.71
	S_2	0.90	0.78	0.83
	S_3	0.87	0.48	0.63
	S_4	0.89	0.79	0.84
	S_5	0.94	0.88	0.90
	S_6	**0.95**	0.84	0.89
	S_7	**0.95**	0.78	0.86
	S_8	**0.95**	0.79	0.86
	S_9	0.94	0.74	0.83
	S_{10}	0.97	0.83	0.89
IR approach	S_{11}	0.71	0.58	0.64
	S_{12}	0.43	0.71	0.54
	S_{13}	0.75	0.61	0.67
	S_{14}	0.45	0.74	0.56
Mixed approach	S_{15}	0.92	0.87	0.89
	S_{16}	0.93	**0.89**	**0.91**
	S_{17}	0.92	**0.89**	0.90
	S_{18}	0.92	0.82	0.87
	S_{19}	0.94	0.87	0.90

The use of official descriptions in the IR system worsens both Precision and Recall, which could be an indication of how different the diagnoses in practice and descriptions in the standard are. Thus, although the use of official descriptions does not seem advisable in itself due to noise, it would be interesting to use synonyms to enrich them. The configuration with the highest F_1 score is the combination of MLPs models to assign ICD-10 codes with frequencies greater than 100 occurrences on the training corpus, and search engines selecting only the result with the highest affinity. The models chosen have been trained with the Tf-BNS of the 1,000 most relevant features.

Finally, the proposals S_5 and S_{16} —called LSI_UNED-run2 and LSI_UNED-run1 respectively—- have been used on the official test dataset provided by each language. S_6 has been chosen as the system offering the best results in the supervised approach. In [6] you can see the ranking of the task published. Our results are shown in Table 6.

In principle, it appears from this data that combining IR techniques with supervised approaches decreases the Precision value to the same extent as it increases the Recall value, so the F_1 score does not change. Nevertheless, in our opinion the combination of both approaches results in a more robust system compared to other possible distributions with a greater number of infrequent codes, so it would be preferable to a single approach based on Machine Learning.

Table 6. Results of the S_5 and S_{16} configurations on the test datasets.

Dataset	System	Precision	Recall	F_1 score
Italian	S_5	0.93	0.86	0.89
	S_{16}	0.91	0.87	0.89
Hungarian	S_5	0.94	0.91	0.92
	S_{16}	0.93	0.92	0.92
French	S_5	0.88	0.54	0.66
	S_{16}	0.84	0.55	0.67

5 Conclusions and Future Work

Different methods have been proposed for the automatic coding of diseases according to the ICD-10 standard. Although a supervised learning approach seems an appropriate solution at first glance, we understand that in a distribution as complex as the one presented by the data, it is necessary to extend these models with other techniques that offer greater coverage, such as the IR approach.

The development of automatic systems for coding death certificates can provide a major boost to health administrations in managing their resources. And to this end, the results published in the CLEF task seem promising.

One of the main problems of natural language processing in health scope is multilingualism, since it is a very broad and specialized domain, and at the same time it requires a large amount of textual resources that do not yet exist for certain languages. Therefore, in the near future we hope to limit the dependence on these textual resources by improving IR techniques and advance in different ways of combining the methods described.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation MAMTRA-MED Project (TIN2016-77820-C3-2-R).

References

1. Almagro, M., Martínez, R., Fresno, V., Montalvo, S. (2018). Estudio preliminar de la anotación automática de códigos CIE-10 en informes de alta hospitalarios. Procesamiento del Lenguaje Natural, Revista nº 60, pp. 45-52. DOI 10.26342/2018-60-5.
2. Chen, Y., Lu, H., Li, L. (2017). Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. In PloS one, vol. 12(3).
3. Ho-Dac, L. M., Fabre, C., Birski, A., Boudraa, I., Bourriot, A., Cassier, M., Delvenne, L., Garcia-Gonzalez, C., Kang, E., Piccinini, E., Rohrbacher, C., Séguier, A. (2017). LITL at CLEF eHealth2017: automatic classification of death reports. CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.

4. Jatunarapit, P., Piromsopa, K., Charoeanlap, C. (2016). Development of thai text-mining model for classifying ICD-10 TM. In Proceedings of ECAI 2016, pages 1–6.
5. Miftakhutdinov, Z., Tutubalina, E. (2017). Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.
6. Névéol, A., Robert, A., Grippo, F., Lavergne, T., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., Zweigenbaum, P. (2018). CLEF eHealth 2018 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.
7. Ning, W., Yu, M., Zhang, R. (2016). A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation. In BMC Medical Informatics and Decision Making, vol. 1:16–30.
8. Ševa, J., Kittner, M., Roller, R., Leser, U. (2017). Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017. CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.
9. Suominen, Hanna, Kelly, Liadh, Goeuriot, Lorraine, Kanoulas, Evangelos, Azzopardi, Leif, Spijker, Rene, Li, Dan, Névéol, Aurélie, Ramadier, Lionel, Robert, Aude, Palotti, Joao, Jimmy, Zuccon, Guido. (2018). Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer.
10. van Mulligen, E. M., Afzal, Z., Akhondi, S. A., Vo, D., Kors, J. A. (2017). Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts. CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.
11. Zweigenbaum, P., Lavergne, T. (2017). Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.