IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence approach

A. Atutxa[1], A. Casillas[2], N. Ezeiza[3], V. Fresno[4], I. Goenaga[3], K. Gojenola[1], R. Martínez[4], M. Oronoz[3], O. Perez-de-Viñaspre[1]

[1] Dpt. Languages and Computer Systems, School of Engineering, UPV/EHU Bilbao
{aitziber.atucha,koldo.gojenola,olatz.perezdevinaspre}@ehu.eus
[2] Dpt. Electricity and Electronics, Fac. of Science and Technology, UPV/EHU Leioa
arantza.casillas@ehu.eus
[3] Dpt. Languages and Computer Systems, Faculty of Informatics, UPV/EHU Donostia-San Sebastian
{iakes.goenaga,n.ezeiza,maite.oronoz}@ehu.eus
[4] Dpt. Languages and Computer Systems. School of Computer Engineering, UNED Madrid
{vfresno,raquel}@lsi.uned.es

Abstract. Hospital systems routinely assign disease codes (ICD10 codes) to medical records. The challenge stands on treating natural and non-standard language in which doctors express their diagnoses and, additionally, to solve a large-scale classification problem, as there are thousands of possible codes. In this working notes paper, we present our system and the results of the CLEF 2018 eHealth Evaluation Task 1 on Multilingual Information Extraction - ICD10 coding. This benchmark addresses information extraction in written text with focus on several languages, specifically Hungarian, Italian and French. The goal is to automatically assign ICD10 codes to diagnostic terms of death certificates. The problem can be cast in different ways, for example as a multilabel classification task or as sequence-to-sequence prediction. Our proposal follows this last approach, with promising results, well above the average results for the task. It only relies on the material provided by the task organizers, allowing the application of the same system to all datasets.

Keywords: Natural language processing · Clinical texts · ICD10 coding · Death certificates · Machine learning.

1 Introduction

The aim of this paper is to explore computer aided approaches to classify Medical Records following the World Health Organization's *International Classification of Diseases* (ICD). These records are written in different languages and the CLEF 2018 eHealth Evaluation Task 1 on Multilingual Information Extraction consists of assigning the right ICD10 coding [1, 2] according to the diagnostic terms provided for each Medical Record. Medical Records belong to several services (pharmacy, documentation, etc.) and achieving their right coding is crucial

to exchange and consult medical information on a daily basis as the ICD codes serve as a reference to exchange information (e.g., billing, epidemiologies or mortality) between hospitals in a country and even between countries. So far, it is common practice in the hospitals to classify the records manually, but there is an increasing interest in the evolution of the automatic or semi-automatic classification, amongst others, due to economic factors. According to [3], the approximate cost of ICD-9-CM coding clinical records and correcting related errors is estimated to be about $25 billion per year in the US. The ICD-10-CM coding is more complex than the previous ICD-9-CM and the costs will be presumably higher. For the Clinical Documentation Services, automatically classifying 1% of the electronic health records would have an outstanding impact in terms of person-months work.

However, the encoding of diagnoses with ICD codes is a difficult, time consuming and expensive task for health services. These records are written in a non-standard medical language causing problems for retrieving and exchanging information due to elements such as misspellings or colloquial and specific language. This lack of standardization also poses a challenge for the automatic classification process due to:

- Acronyms: the adoption of non standard contractions for the word-forms.
- Abbreviations.
- Omissions: often prepositions, articles or verbs are omitted in an attempt to write the word-form quickly.
- Synonyms: some technical words are typically replaced by others.
- Misspells: sometimes words are incorrectly written.

The IxaMed group has approached the automatic ICD10 coding for French, Italian and Hungarian with a neural model that tries to map the input text snippets with the output ICD10 codes. Our solution does not make assumptions about the content of the input and output data, treating them by means of a machine learning approach that assigns a set of labels to any input line. The solution is language-independent, in the sense that treating a new language only needs a set of (input, output) examples, making no use of language-specific information apart from terminological resources such as ICD10 dictionaries, when available.

2 Related Work

Computer aided classification of medical records can be seen as a pattern recognition task, as the aim is to recognize unknown instances of expressions and assign them one or more elements from a set of possible labels. This problem has been approached in several tasks and challenges using different techniques.

The 2007 Computational Medicine Challenge [7], the first shared task related to ICD coding, was designed: (i) to facilitate advances in mining clinical free text and (ii) to create a publicly available gold standard that could serve as the seed for a larger, open source clinical corpus. This Challenge involved the

classification of English clinical free texts by automatically assigning ICD-9-CM codes in a limited domain devoted to radiology reports. [3] addressed this shared task employing machine learning approaches. Their results showed that handcrafted systems could be reproduced by replacing several laborious steps in their construction with machine learning models, reporting an F1-measure of 0.8893. By contrast to [3] we focus on the entire scope of the ICD10 catalog. That is, while they were dealing with 45 classes, we have to cope with thousands of classes.

Pérez et al. [4, 5] proposed the use of *Finite-State Transducers* (FSTs) that constrain the allowed input diagnostic string, synchronously producing the output ICD class. FSTs are versatile and efficient to implement soft-matching operations between terms expressed in natural language to standard terms and, hence, to the final ICD code. The FSTs were built up from a corpora and standard resources such as the ICD-9-CM and SNOMED CT amongst others. An F1-measure of 0.9120 was achieved on a test-set of 2,850 randomly selected diagnostic terms. A difference with the present work is that in their system the input diagnostic terms were correctly aligned by physicians one by one, while at the 2018 shared task most ICD10 codes are aligned only at the document level, which makes the task harder.

Pérez et al. [11] tackle diagnostic term normalization employing Weighted Finite-State Transducers (WFSTs) that learn how to translate sequences into standard representations given a set of samples. They are highly flexible and easily adaptable to terminological singularities of different hospitals and practitioners. They also implemented a similarity metric to enhance spontaneous-standard term matching. Looking at their results, they found that only 7.71% of the diagnostics were written in their standard form matching the ICD. This WFST-based system enabled matching spontaneous ICD codes with a Mean Reciprocal Rank of 0.68, which means that, on average, the right ICD code for each diagnosis is found between the first and second position among the normalized set of candidates. Similarly, Almagro et al. [6] experiment a combination of techniques for ICD-10 coding in Spanish.

CLEF eHealth 2017 Task 1 is a similar challenge but multilingual since the texts were both English and French, more extensive because it was not limited to an specific service and employed ICD-10-CM for coding instead of ICD-9-CM. [8] implemented recurrent neural networks to automatically assign ICD10 codes to fragments of death certificates written in English. Their system used *Long Short-Term Memory* (LSTM) to map the input sequence into a vector representation, and then another LSTM to decode the target sequence from the vector. They initialized the input representations with word embeddings trained on user posts in social media. Their encoder-decoder model obtained an F-measure of 0.8501 on a test set, with significant improvement as compared to the average score of 0.6220 for all participants approaches.

Other systems presented at the CLEF 2017 shared task made use of varied approaches. In [10], they composed a large scale feature set comprising more than 40k features based on bag of words, bag of 2-grams, bag of 3-grams, latent

Dirichlet allocation, and the ontologies of WordNet and UMLS. [9] used concept detection and normalization experiments, starting upon dictionary projection and supervised multi-class, mono-label text classification using simple features, and extending the system in several dimensions with multi-label classification and new features, including a combination of dictionary and classifier.

To summarize, we can say that the problem presents a complex characterization due to multiple factors, like non-standard language variation, spontaneous writing, or large-scale multilabel classification. Accordingly, there are plenty and varied approaches to tackle it, ranging from knowledge-based solutions to statistical and deep learning ones.

3 Resources and Methods

3.1 Corpus

In the present challenge [2], French, Italian and Hungarian are the languages under study. There are two sources of information:

- ICD-10 dictionaries.
- Different sets of documents and their corresponding (text lines, ICD10 code) pairs.

The sets of document-ICD10 codes come in two different formats: raw and aligned, though the aligned version is only available for French. For the raw version, the diagnostic terms as expressed in the original death certificate are stored in one file (*CausesBrutes*, see Table 1) separately from the coding which is stored in another one (*CausesCalculees*, see Table 2). The link between them can be carried out through indexing information common to both: document identifier, year of the death certificate, and line number within the death certificate representing the exact location in the text. As previously stated, diagnostic terms in the *CausesBrutes* files appear as originally expressed in the death certificates and therefore they show orthographic misspellings (*infacrtus* vs. *infarctus*) and abbreviations (*HTA* vs. *hypertension artrielle*, see Table 1).

It is important to notice that a one-to-one correspondence between the raw diagnostic term and the ICD is not assured. Missmatches occur like the ones shown in document 100569, where line 5 in Table 1 has no correspondence in Table 2. The correspondence appears in line 6. It might happen to find more than one diagnostic term in one line separated by commas, and coordination by means of complementizers or prepositions.

In the aligned version, the original text is accompanied with the standard text and the ICD associated (see Table 3).

3.2 Description of the System

Preprocessing. With the aim of boosting the ICD assignation, we preprocessed the raw corpora to organize the information at three levels: document level, line

Table 1. Example of diagnostic terms in a Causes Brutes file for French (raw).

indexing info.	**Diagnostic Term (DT)**
100644 2014 1	insuffisance cardiaque
100644 2014 2	infacrtus du myocarde
100644 2014 5	HTA. AIT trouble mnésiques hypercholestrolmie
100569 2014 1	Défaillance Cardiaque
100569 2014 2	Infarctus du Myocarde étendu
100569 2014 5	BAV appaareillé avec décharge du PM
100569 2014 6	AVC et Sd démensiel

Table 2. Example of standard terms in a Causes Calculees file for French (raw).

indexing info.	**preferred DT**	**ICD Code**
100644 2014 1 1	insuffisance cardiaque	I509
100644 2014 2 1	infarctus myocarde	I219
100644 2014 6 1	hta	I10
100644 2014 6 2	ait	G459
100644 2014 6 3	troubles mnésiques	R413
100644 2014 6 4	hypercholestémie	E780
100569 2014 1 1	défaillance cardiaque	I509
100569 2014 2 1	infarctus myocarde étendu	I219
100569 2014 6 1	bav appareillé	I443
100569 2014 6 2	dysfonction sonde pm	Z950
100569 2014 6 3	avc	T821
100569 2014 6 4	syndrome démentiel vasculaire	I640
100569 2014 6 5	NULL	F019

Table 3. Example of standard diagnostic terms in a for French (aligned).

indexing info.	**Diagnostic Term (DT)**	**preferred DT**	**ICD Code**
100569 2014 1 80 2 1	Défaillance Cardiaque	défaillance cardiaque	I509
100569 2014 1 80 2 2	Infarctus du Myocarde étendu	infarctus myocarde tendu	I219
100569 2014 1 80 2 5	BAV appaareillé avec décharge du PM 4 0	—	—
100569 2014 1 80 2 6	AVC et Sd démensiel	bav appareillé	I443
100569 2014 1 80 2 6	AVC et Sd démensiel	dysfonction sonde pm	Z950
100569 2014 1 80 2 6	AVC et Sd démensiel	NULL	F019
100569 2014 1 80 2 6	AVC et Sd démensiel	avc	T821
100569 2014 1 80 2 6	AVC et Sd démensiel	syndrome démentiel vasculaire	I640

level and finally ICD level. At the document level and line level, we grouped all diagnostic terms and ICD codes by document and by line respectively, hoping that the system could capture dependencies among the different ICD codes. It seems logic to think that ICD codes within a document or within a line are related to each other and, if so, ensemble recognition might be helpful. The preprocess to obtain the line level information consisted mostly in trying to overcome the alignment mistakes in the original corpus as shown in Tables 1 and 2. At the ICD level, we treated separately each (diagnostic term ICD) pair aiming to simplify the assignment process but at the cost of missing any inter-relation that could exist. This level required a more refined preprocessing since the original information was set at the line level. Remember that certain lines showed several diagnostic terms.

As a first step in normalization, the input texts were preprocessed in the following order: tokenization, lowercasing and substitution of accents. These are standard operations in sequence-to-sequence learning, that help to improve the results.

ICD10 coding. In neural sequence-to-sequence modeling, the encoder-decoder model has been used to encode a variable-length input sequence of tokens into a sequence of vector representations, and to then decode those representations into a sequence of output tokens, in this case ICD10 codes.

This decoding is conditioned on information from both the latent input vector encodings as well as its own continually updated internal state, motivating the idea that the model should be able to capture meanings and interactions beyond those at the word level [12, 13].

The supplied data was divided in three subsets. A training set was iteratively evaluated on a second hold-out evaluation set and, finally, the best performing system was evaluated on an independent third set. For the final submission, the training and hold-out sets were merged, using the third subset for iterative evaluation, and applying the best system on the unseen test set.

4 Results and Discussion

Table 4 presents the results obtained by our system for all the languages. We can see that our results are significantly above the average in all languages. We obtained our best results in Hungarian with a improvement of 16 points over the average. For Italian we obtained similar results, while French shows the worst average in both aligned and raw versions with respect to Italian and Hungarian. For French we obtained an improvement of 20 points with respect to the average for both the aligned and raw versions of the data.

Table 4 shows the results obtained when applying the system at line level, that is to say, one input sequence and the corresponding ICD codes per line as training instances. We obtained our worst results when training at the document level or when we trained the system using 1:1 pairs of diagnostic terms and ICD codes. This might support the idea that there is an interrelation among

the diagnostic codes which is kept at the line level but cannot be assumed at document level.

Table 4. Performance.

Language	Run	Precision	Recall	F-measure
French (aligned)	run1	0.8412	0.8347	0.8338
	run2	0.8412	0.8347	0.8380
	frequencyBaseline	0.4517	0.4504	0.4511
	average	0.7123	0.5808	0.6342
	median	0.7712	0.5445	0.6407
French (raw)	run1	0.8724	0.5966	0.7086
	run2	0.8773	0.5874	0.7037
	frequency Baseline	0.3410	0.2005	0.2525
	Average	0.7227	0.4101	0.5066
	Median	0.7981	0.4750	0.5790
Italian (raw)	run1	0.9599	0.9450	0.9524
	run2	0.9453	0.9223	0.9337
	frequencyBaseline	0.1648	0.1723	0.1685
	average	0.8441	0.7606	0.7992
	median	0.8995	0.8239	0.8630
Hungarian (raw)	run1	0.9678	0.9543	0.9610
	run2	0.9700	0.9554	0.9627
	frequencyBaseline	0.2425	0.1735	0.2023
	average	0.8266	0.7830	0.8025
	median	0.9221	0.8972	0.9095

5 Conclusions

This work tackles the classification of medical records following the ICD10 standard. The classification problem is tough for several reasons: 1) the gap between spontaneous written language and standard one; and 2) it is a large-scale classification system, being the number of possible classes the number of different diseases within the ICD10 catalogue.

Our best system showed high-quality results, and this fact opens a promising avenue for the task of automatically assigning ICD10 codes to medical documents. Moreover, the method is language independent and it allows efficient training, given only a set of annotated documents, not requiring complex feature engineering.

6 Acknowledgements

This work has been partially funded by:

– The Spanish ministry (projects TADEEP: TIN2015-70214-P, PROSA-MED: TIN2016-77820-C3-1-R).
– The Basque Government (projects DETEAMI: 2014111003, ELKAROLA:KK-2015/00098).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Palotti, J., Zuccon, G. Overview of the CLEF eHealth Evaluation Lab 2018. In: CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2018.
2. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., Zweigenbaum, P.: CLEF eHealth 2018 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2018.
3. Farkas, R., Szarvas, G.: Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics, **9**(Suppl. 3), 1–9 2008.
4. Pérez, A., Casillas, A., Gojenola, K., Oronoz, M., Aguirre, N., Amillano, E.: The aid of machine learning to overcome the classification of real health discharge reports written in Spanish. Revista de Procesamiento de Lenguaje Natural (ISSN: 1135-5948) 2014.
5. Pérez, A., Gojenola, K., Casillas, A., Oronoz, M., Díaz de Ilarraza, A.: Computer aided classification of diagnostic terms in spanish. Expert Systems with Applications, **42**(6), 2949–2958 .2015.
6. Almagro, M., Martínez, R., Fresno, V., Montalvo, S.: Estudio preliminar de la anotacin automtica de cdigos CIE-10 en informes de alta hospitalarios. Revista de Procesamiento de Lenguaje Natural (ISSN: 1135-5948) (60) 2018.
7. Pestian, John P., Brew, Christopher, Matykiewicz, Paweł, Hovermale, D. J., Johnson, Neil, Cohen, K. Bretonnel, Duch, Wlodzislaw: A Shared Task Involving Multi-label Classification of Clinical Free Text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, BioNLP '07, pp. 97–104. Association for Computational Linguistics, Stroudsburg, PA, USA 2007.
8. Miftahutdinov, Z., Tutubalina, E.: KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks. In: CLEF 2017 Conference and Labs of the Evaluation Forum, Online Working Notes, CEUR-WS, September .2017.
9. Zweigenbaum, P., Lavergne, T.: Multiple Methods for Multi-class, Multi-label ICD-10 Coding of Multi-granularity, Multilingual Death Certificates. In: CLEF 2017 Conference and Labs of the Evaluation Forum, Online Working Notes, CEUR-WS, September 2017.
10. Ebersbach, M., Herms, R., Eibl, M.: Fusion Methods for ICD10 Code Classification of Death Certificates in Multilingual Corpora. In: CLEF 2017 Conference and Labs of the Evaluation Forum, Online Working Notes, CEUR-WS, September 2017.

11. Pérez, A., Atutxa, A., Casillas, A., Gojenola, K., Sellart, A.: Inferred joint multigram models for medical term normalization according to ICD. International Journal of Medical Informatics, **110**, pp. 111–117 2018.
12. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems (NIPS), pp. 3104–3112 2014.
13. Cho, K., van Merrienboer, B., Gulcehre, G., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) 2014.