*Digital Personalized Health and Medicine*
*L.B. Pape-Haugaard et al. (Eds.)*


*doi:10.3233/SHTI200196*

# Supervised Learning for the ICD-10 Coding of French Clinical Narratives

Clément DALLOUX [a], Vincent CLAVEAU [a], Marc CUGGIA [b], Guillaume BOUZILLÉ [b] and Natalia GRABAR [c]

[a] *Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France*
[b] *Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France*
[c] *UMR 8163 STL CNRS, Université de Lille, France*

**Abstract.** Automatic detection of ICD-10 codes in clinical documents has become a necessity. In this article, after a brief reminder of the existing work, we present a corpus of French clinical narratives annotated with the ICD-10 codes. Then, we propose automatic methods based on neural network approaches for the automatic detection of the ICD-10 codes. The results show that we need 1) more examples per class given the number of classes to assign, and 2) a better word/concept vector representation of documents in order to accurately assign codes.

**Keywords.** ICD-10 coding, clinical NLP, clinical narratives, French

## 1. Introduction

Automatic coding of clinical narratives in French with the International Classification of Diseases, 10th revision (ICD-10) has become a necessity, as ICD-10 is used for billing, epidemiology survey and many other tasks. In this work, we present 1) a corpus with French clinical documents annotated with the ICD-10 codes, and 2) methods relying on neural networks for document-level classification. Given the difficulty of this task, the aim of our work is to create a tool that suggests relevant codes to human coders.

Several shared tasks are related to the ICD coding. The first of the kind, Computational Medicine Challenge [1] in 2007, was dedicated to the assignment of the ICD-9-CM codes to clinical narratives (radiology reports). While [2] is the first work on ICD coding in French, since 2016, CLEF eHealth offers a clinical information extraction challenge which aim is to extract ICD-10 codes from health reports. Recently [3,4,5], the CépiDC task consisted in extracting ICD-10 codes from death reports in several languages (French in 2016, French and English in 2017, French, Hungarian and Italian in 2018). In 2019 [6], the task involved the automatic annotation of German non-technical summaries of animal experiments with ICD-10 codes. The 2020 task will focus on ICD-10 coding for clinical narratives in Spanish. As we are working with French clinical documents, we next describe the 2018 French data and the most efficient systems.

The French death certificates are stored in two files: one raw text file in which the physician writes the report (raw causes) and one file containing the corresponding ICD codes (computed causes). The computed causes file might contain the text supporting

| line | text | normalized text | ICD codes |
|---|---|---|---|
| 1 | choc septique | choc septique | A41.9 |
| 2 | peritonite stercorale sur perforation colique | peritonite stercorale | K65.9 |
| 2 | peritonite stercorale sur perforation colique | perforation colique | K63.1 |
| 3 | Syndrome de détresse respiratoire aiguë | syndrome détresse respiratoire aiguë | J80.0 |

**Table 1.** A sample document from the CépiDC French Death Certificates Corpus: aligned dataset [5]

| Aligned | | | | Raw | | | |
|---|---|---|---|---|---|---|---|
| Team | Precision | Recall | F-measure | Team | Precision | Recall | F-measure |
| IxaMed-2 | .841 | **.835** | **.838** | IxaMed-1 | .872 | **.597** | **.709** |
| IxaMed-1 | **.846** | .822 | .834 | IxaMed-2 | .877 | .588 | .704 |
| IAM-2 | .794 | .779 | .786 | LSI-UNED-1 | .842 | .556 | .670 |
| IAM-1 | .782 | .772 | .777 | LSI-UNED-2 | **.879** | .540 | .669 |

**Table 2.** Best CLEF eHealth 2018 results on the raw and aligned French dataset [5]

the coding decisions. Two datasets were provided for this task: one raw and one aligned. The raw dataset contains the original death certificates and the corresponding computed causes. The aligned dataset contains per line: raw text, normalized text and the corresponding ICD code, as exemplified in Table 1.

The *IxaMed* team [7] used a language-independent solution in which a neural sequence-to-sequence modeling system encodes a variable-length input sequence of tokens into a sequence of vector representations, and decodes those representations into a sequence of output tokens (the ICD10 codes). As shown in Table 2, *IxaMed* got the best results on the aligned and raw datasets. The *IAM-ISPED* team [8] used a dictionary-based approach to assign ICD-10 codes to each text line, and Levenshtein distance and synonym expansion system to detect typos. *IAM-ISPED* got the second best score on the aligned dataset and third on the raw dataset. The *LSI-UNED* team [9] submitted two runs for each dataset. Their best run used a multilayer perceptron and an One-vs-Rest (OVR) strategy supplemented with IR methods. The models were trained with the training data and dictionaries of CépiDC, estimating the frequency of terms weighted with Bi-Normal Separation (BNS). *LSI-UNED* was the second best system on the French raw dataset.

The task addressed during this challenge is interesting although the death certificates cannot be considered as free-text but rather as controlled text documents. This is especially true for the aligned dataset, which makes it very convenient for the methods proposed during the challenge. In this work, we use clinical narratives containing mostly free-text, therefore, the methods used during the challenge are not appropriate.

## 2. Method

*Dataset.* Our dataset contains 28,000 clinical documents from the Rennes university hospital. Each document contains one or several notes for a single patient issued from a single stay at the hospital. Table 3 presents statistics on our corpus. As it shows, although the tokenization methods might differ, our documents are much longer than death certificates. Indeed, the latter only contain on average 3 lines, 10 tokens and 4 ICD codes per document while our clinical narratives have on average 1,345 tokens per document and twice as many codes. It also shows that a majority of ICD codes have very few examples in the corpus which will be an issue for the deep learning methods. A typical document describes the patient's medical history, reasons for hospitalization, laboratory measurements, findings, treatment and conclusion. Our set of data also contains some noise (e.g.

| | Training | Development | Test | Total |
|---|---|---|---|---|
| Documents | 22,400 | 2,800 | 2,800 | 28,000 |
| Tokens | 30,307,091 | 3,718,715 | 3,654,599 | 37,680,405 |
| Total ICD codes | 182,112 | 23,010 | 22,561 | 227,683 |
| Unique ICD codes | 5,735 | 2,824 | 2,887 | 6,116 |
| Unique unseen ICD codes | - | 191 | 212 | - |
| Codes with less than 10 examples | 3,885 | 2,320 | 2391 | - |
| Codes with 100 examples or more | 382 | 32 | 31 | - |

**Table 3.** Descriptive statistics of the Rennes University Hospital dataset

phone numbers, names, email addresses, titles, etc.) which was not preprocessed. As our data contain confidential medical information, we cannot show any examples.

*Tasks.* The first task addressed is the multi-label classification done on a document level for all codes. As Table 3 shows, there are 6,116 different classes (ICD codes) to identify: 5,735 codes are present in the training set among which 3,885 examples have less than 10 examples. For the second task, we reduced the hierarchical level of the ICD-10 codes. All codes with 3 digits and more were reduced to 2 digits: A01.3 *Paratyphoid fever C* becomes A01 *Typhoid and paratyphoid fevers*. Therefore, only codes A00 to Z99 are exploited, which decreases the number of classes to 1,549 overall, among which 1,501 are present in the training set.

*Methods.* The purpose of our work is to design an approach for the automatic and efficient detection of the ICD codes. The approaches proposed and tested rely on specifically trained word embeddings and supervised learning techniques. In the existing work, various methods have been used to represent words as vectors. Recently, several machine learning-based approaches have been introduced for a better representation of semantic relations between words. We experiment with fastText vectors [10], which use subword information to represent words missing in the training vocabulary. Each word is represented as a bag of all possible n-grams of characters it contains. Our model is trained on our dataset with the *Skip-Gram* algorithm, 300 dimensions, a window of 5 words before and after each word, a minimum count of five occurrences for each word and negative sampling. Besides, we experiment with two neural network architectures: recurrent and convolutional. A recurrent neural network is a class of network which is capable of adapting its decision by taking into account the previously seen data, in addition to the currently seen data. This operation is implemented thanks to the loops in the architecture of the network, which allows the information previously seen (previous words) to persist in memory. In our experiments, we used a bidirectional recurrent neural network with long short-term memory cells (BiLSTM). Convolutional neural networks (CNN) are widely used in Computer Vision and Natural Language Processing. The convolution layer consists of a set of filters, where each filter is convolved with the input forming feature maps. These filters are randomly initialized and become parameters that will be learned by the network.

*Experiments.* We trained several CNN models based on the architecture proposed in [11] for 30 epochs. We used four filter windows (2, 3, 4 or 5 width) with either 100, 500, 1,000 or 2,000 feature maps each (ReLU activation), 0.5 dropout, global max pooling and a fully connected layer with ReLU activation of either 100, 1,000 or 2,000 dimensions for a more compact representation. Regarding our BiLSTM, the dimensionality of the

output space was set to 600 units per layer (backward/forward) with 0.5 dropout. Both architectures use our pre-trained fastText vectors with 0.5 dropout and a fully connected sigmoid layer for prediction. Performance was measured with the standard metrics: precision, recall and F-measure.

## 3. Results

As Table 4 shows, on both tasks, our CNN, even at the lowest settings, performs better than our BiLSTM. Our experiments also show that increasing the number of feature maps per convolution filter and hidden units in the fully connected layer both improves the results. Indeed, we get an 8.21 point increase at the most expensive settings when compared to the lowest settings. We might be able to slightly improve the results if we raise parameters further. However, it would be a low increase with a great computing cost. Indeed, the results on Task 2 show a very low improvement (0.28 point) when doubling the value of the parameters.

| | System | Feature maps | Hidden units | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| Task 1 | BiLSTM | - | 600 | **0.6513** | 0.1387 | 0.2287 |
| | CNN | 100 | 100 | 0.6510 | 0.2091 | 0.3165 |
| | CNN | 100 | 1,000 | 0.4916 | 0.2491 | 0.3306 |
| | CNN | 500 | 100 | 0.5624 | 0.2652 | 0.3605 |
| | CNN | 500 | 1,000 | 0.4935 | 0.3042 | 0.3763 |
| | CNN | 1,000 | 100 | 0.5353 | 0.2597 | 0.3498 |
| | CNN | 1,000 | 1,000 | 0.5052 | 0.3200 | 0.3918 |
| | CNN | 2,000 | 2,000 | 0.5029 | **0.3301** | **0.3986** |
| Task 2 | BiLSTM | - | 600 | **0.6309** | 0.2837 | 0.3914 |
| | CNN | 1,000 | 1,000 | 0.6306 | 0.4427 | 0.5202 |
| | CNN | 2,000 | 2,000 | 0.6018 | **0.4625** | **0.5230** |

**Table 4.** Results for the ICD coding tasks. The results are given in terms of Precision, Recall and F-measure. The best scores are in bold.

## 4. Discussion

As expected, we get better results on our second task which shows that the reduction of the number of classes is efficient. However, as we can only assign each class once to a document, we might lose some information as a document can have multiple codes belonging to the same hierarchically higher class. Besides, we get much lower results than the systems presented at CLEF eHealth 2018. This can be explained by the nature of the documents: short, structured death certificates cannot be compared to clinical narratives with free text. The tasks are therefore different. Another explanation is that some codes might not be explained by the textual data but rather by the structured data in the hospital database.

## 5. Conclusion

The interest for the automatic detection of ICD-10 codes has increased in recent years. Yet, the data used in previous shared tasks do not seem representative of real clinical

narratives. In our work, after a brief reminder of the existing work, we presented a new body of French clinical narratives annotated with the ICD-10 codes. Another contribution of our work is the exploitation of fastText vectors and neural networks for the automatic ICD-10 coding. The experiments are conducted and evaluated on French clinical narratives (free-text), which have not been experienced much in the previous work of this type. From a more technical point of view, our work also indicates that the convolutional neural architectures are more efficient than the recurrent neural architectures for the multi-label classification. We plan to improve our neural network performance by providing a richer feature set. Indeed, adding available structured data and using recent embedding techniques, such as BERT [12], may provide more accurate representation of the documents.

**Acknowledgments.** This work was partly funded by the French government support granted to the CominLabs LabEx managed by the ANR in Investing for the Future pro-gram under reference ANR-10-LABX-07-01.

## References

[1] Pestian, John P., Brew, Christopher, Matykiewicz, Pawe l, Hovermale, D. J., John-son, Neil, Cohen, K. Bretonnel, Duch, Wlodzislaw: A Shared Task Involving Multi-label Classification of Clinical Free Text. In: Proceedings of the Workshop onBioNLP 2007: Biological, Translational, and Clinical Language Processing, BioNLP'07, pp. 97–104. Association for Computational Linguistics, Stroudsburg, PA, USA2007.

[2] Ruch, P., Gobeill, J., Tbahriti, I., & Geissbühler, A. (2008). From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. In AMIA Annual Symposium Proceedings (Vol. 2008, p. 636). American Medical Informatics Association.

[3] Névéol, A., Cohen, K. B., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., ... & Zweigenbaum, P. (2016, September). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In CEUR workshop proceedings (Vol. 1609, p. 28). NIH Public Access.

[4] Névéol, A., Robert, A., Anderson, R., Cohen, K. B., Grouin, C., Lavergne, T., ... & Zweigenbaum, P. (2017, September). CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. In CLEF (Working Notes).

[5] Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., ... & Zweigenbaum, P. (2018, September). CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. In CLEF (Working Notes).

[6] Kelly, L., Suominen, H., Goeuriot, L., Neves, M., Kanoulas, E., Li, D., ... & Palotti, J. (2019, September). Overview of the CLEF eHealth evaluation lab 2019. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 322-339). Springer, Cham.

[7] Atutxa A, Casillas A, Ezeiza N, Goenaga I, Fresno V, Gojenola K, Martinez R, Oronoz M and Perez-de-Viñaspre O (2018). IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence approach. CLEF 2018 Online Working Notes. CEUR-WS

[8] Cossin S, Jouhet V, Mougin F, Diallo G, and Thiessard F (2018). IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates. CLEF 2018 Online Working Notes. CEUR-WS

[9] Almagro M, Montalvo S, Diaz de Ilarraza A, and Pérez A (2018). LSI UNED at CLEF eHealth 2018: A Combination of Information Retrieval Techniques and Neural Networks for ICD-10 Coding of Death Certificates. CLEF 2018 Online Working Notes. CEUR-WS

[10] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

[11] Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751. Association for Computational Linguistics, Doha, Qatar.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.