ECSTRA-APHP @ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates

Rémi Flicoteaux[1]

INSERM, U1153 Epidemiology and Biostatistics Sorbonne Paris Cit Research Center (CRESS), ECSTRA team, Paris, F-75010 France
remi.flicoteaux@gmail.com

Abstract. This paper describes the participation of ECSTRA-APHP team at CLEF eHealth 2018, task 1. The task involved extracting ICD-10 codes from death certificates, mainly described with short plain texts. We casted the task as a machine learning problem the prediction of the ICD-10 codes (categorical variable) from the raw text transformed into word embeddings. We relied on probabilistic convolutional neural network for classification. Due to inbalanced representation of the ICD codes, we completed the prediction with dictionary-based lexical matching classifier for cases were there was less than 1,000 documents per code. Our best F1-score was 80.0% on a test set and 69.1% on the validate set (gold standard delivered by the organizers at the end of the challenge). This was the first time convolutional neural net were used for this multi-label classification task. The performance of our models were under the best neural predictor (recurrent network) described last year on the same task at CLEF eHealth (F1-score around 85%).

Keywords: ICD-10 coding, ICD-10 codes, cause of death extraction,convolutional neural network,lexical matching

1 Introduction

Completing death certificates is a routine task in hospitals and healthcare institutions. In France, the death certificates are produced by physicians and transmitted to the French Epidemiological Center for the Causes of Death (CépiDC) [1]. Beyond the administrative and personal information, the death certificates usually contain a free-text description of the cause(s) of death. Free texts are converted by the CépiDC into formal standardized codes to be processed for statistical purposes. Like many countries, the the World Health Organisation (WHO) International Classification of Diseases (ICD) taxonomy [2] is used for this normalized representation. The ICD taxonomy covers a wide range of diseases, symptoms, signs, and other content related to diseases. The WHO issues separate versions of ICD per language/country. In this paper, we use the French release of ICD, which is now at its 10 th revision (called ICD-10). It covers more

[1] http://www.cepidc.inserm.fr/
[2] http://www.who.int/classifications/icd/

than 38,000 codes of diagnoses, but only a subset of theses codes can be causes of death.

Requiring manual work and expertise, the task of ICD-10 code extraction from text is quite time-consuming because the ICD-10 taxonomy contains thousands of possible causes of death. Within the CLEF eHealth 2018, the task 1 focuses on the problem of automatic extraction of the causes of death from the textual description[1]. Classification for health-related text is considered a special case of multilabel text classification which may be approached either from a machine learning perspective (supervised classification) or a Natural Language Processing (NLP) perspective by using syntactic and/or semantic decision rules. For this purpose machine learning algorithms have been successfully applied, i.e. Support Vector Machines, Latent Dirichlet Allocation or neural network. Both approaches aim at automating the ICD-10 code extraction from death certificates. In this paper, we mainly focus on probabilistic Convolution Neural Network (CNN). Due to unbalanced ICD labels, we enriched prediction with dictionary-based lexical matching classifier.

2 Methods

CNNs utilize layers with convolving filters that are applied to local features [4]. Originally invented for computer vision, CNNs models have subsequently been shown to be effective for NLP and have achieved excellent results in text classification[2]. The filter can be seen as sliding over the columns of the features, performing an element-wise multiplication and summation on the current overlap, before moving one to the right. For NLP task, the filters dimension on the first axis is equal to the one, the resulting vector is only one-dimensional. One of the key differentiators between CNNs and traditional machine learning approaches is the ability for CNNs to learn complex feature representations.

State of the art for feature extraction is to use word vectors models, embedding, where each word is represented by a single real-valued vector. In this model sentences are then projected from a sparse vector space of the size of the vocabulary onto a lower dimensional vector space which encode semantic features of words in their dimensions. Then semantically close words are likewise close the new lower dimensional vector space as measured with vector distance operation like euclidean or cosine distance.

In the present work the weights of words vectors are jointly learned as the first hidden layer of the classification itself, and we train a CNN that uses multiple filters (with varying window sizes) to obtain multiple features on top of word vectors. For this multi-label classification task, the prediction layer (last one) has the size of the number of distinct labels (entries in the ICD). Here we only focuses on codes that have already been used. Finally, the prediction is a vector of real numbers which are equivalent to a probability upon each ICD label (but the all vector do not sum to one). A grid search was perform to determine the threshold among which a code was chosen.

The dictionary-based lexical matching classifier rely on word recognition from a knowledge base build from several available dictionaries on the French ICD-10 classification : second volume of ICD, orphanet thesaurus, French SNOMED CT, and CépiDC dictionaries that were provided for the challenge.

From the detection in the text of entries of the index (i.e. words) ranking scores are usually computed individually for each concept mention. For this purpose we used a very simple score base on the probability of a code associated to a word and the number of words recognized in the text :

$$score = \Sigma p_{code|word} \times n_{match}$$

We used this approach to predict rare codes (represented by less than 1,000 lines), so we choose only one code per statement. Our main idea was to improve prediction of the CNN classifier, so we use this result to add a weight on the output vector of CNN predictor for the selected code. A grid search was performed to decide the size of this weight.

3 Dataset

The CépiDC corpus has been created by the French Center for Epidemiology and Medical Causes of Death (CépiDC) specifically for the CLEF eHealth task 1 [1, 9]. It is composed of separate train/test samples of death certificates. Only the textual description of the causes of death are available for analysis. CépiDC dataset is highly imbalanced: about 80% of documents are assigned to less than 20% of codes.

Table 1. Representativity of ICD diagnoses used in death certificates

	Number of distinct codes	Total number of codes
> 1,000 occurrences	72	240,301 (63.3%)
200-1,000 occurrences	180	80,316 (21.1%)
< 200 occurrences	3007	59,185 (15.6%)
Total	3,259	379,802

The task was defined at the level of each statement (line) in a death certificate: one statement could be associated with 0, 1 or more ICD-10 codes which represent causes of death at various levels in the causal chain which led to the death. Each line was tagged with 0 code (n=7,598 - 2.5%), 1 code (n=238,929 - 78.5%), 2 codes (n=40,572 - 13.38%) up to 14 codes codes (n=1). The dataset included 70656 lines, it was divided into train and test set (25 000 lines). A validation set was also provided at the end of the challenge, with 70,656 lines. On the validation set, there was 1,431 (2.0%) lines without code, 51,383 (72.7%) with one code, 11,981 with 2 codes (16.9%) and the maximum number of code by line was 16 (n=1).

We remove stopwords and numerics. After an homemade spelling correction algorithm based on Levenstein distance, only words from the knoweledge base were sustain for classification. The preprocessed text vectorize in tokens are used as input for CNN and dictionary-based lexical matching.

4 Results

We first build the knowledge base from the available thesaurus (cf methods) and the 2015 (most recent) CépiDC dictionaries that was provided with for the task. At the end we had a 216,110 entries ICD thesaurus, were entries goes from 1 to 314 words. An word/code index was then build with 19,606 distinct entries.

For statistical approach with convolusional network, we use words embeddings features that were fitted on the data together with classification task, and no external weights were used. Models were fitted with and without adding data from knowledge base. Our best prediction gives a F1-score to 79.6% on the test set. We performed two non official runs at the challenge, and on the validation set our best F1-score was of 69.1%.

We studied also performance at the code level given the level of document for each class. Result are presented table 2.

Table 2. Global results

Model	Precision	Recall	F1-Score
Results on test set			
CNN alone without knowledge base	83.4%	76.8%	80.0%
CNN with knowledge base	82.0%	77.4%	79.6%
CNN with knowledge base plus lexical matching	81.6%	77.7%	79.6%
Results on validation set			
CNN alone without knowledge base	63.3%	76.1%	69.1%

The results on the badly represented classes were expected, and indeed a significant reduction in efficiency was recorded below 1000 representatives per class.

Table 3. Results on test set by categories of codes frequencies - CNN alone

Freq codes	Prop. true positives	Prop. false negatives
$[1000; inf[$	84.4%	15.6%
$[500; 1000[$	75.0%	25.0%
$[200; 5000[$	67.7%	32.3%
$[0; 200[$	28.5%	71.5%

We use the second classifier based on word recognition, which did not improve performance on the global criteria on the test set, but which allowed to gain on the F1-score on the very low represented categories.

Table 4. Results on test set by categories of codes frequencies - CNN plus plus lexical matching

Freq codes	Prop. true positives	Prop. false negatives
$[1000; inf[$	85.3%	14.7%
$[500; 1000[$	76.0%	24.0%
$[200; 5000[$	68.5%	31.5%
$[0; 200[$	34.2%	65.8%

Finally, we also looked at the performances of the models on lines which were labeled with 0 ICD code. On the validation set, the performance of our final predictor for theses lines was very poor : precision = 19.7%, recall=5.2% and F1-score = 8.2%.

5 Discussion

The problem of ICD code extraction has been investigated from a larger perspective involving code assignment to various types of medical documents. The CLEF eHealth is one of the only international conferences to propose a specific task on this subject each year, which makes it possible to have a particularly interesting follow-up of the methods and their performances. Mainly two parallel approaches are often developed statistic and entity recognition, and in case the compilation of both. The convolutive networks have shown their effectiveness for automatic classification of documents, particularly medical documents [2, 3]. To our knowledge this is the first time they are used in the CLEF eHealth Challenge. We observe a 10% difference on the model performance from test to validation data. This could be due rather to overfitting on training data and to the fact that our test set might not represent the true distribution of data and labels.

Other neural net architectures are used for text classification. Recurrent networks have becoming more popular in the NLP domain and seems to outperform performance of CNN. Miftahutdinov et al. report the use of RNN in the CLEF eHealth 2017 with success and obtained F-measure of 85% [5]. New character based approach which allowed a huge reduction of time for features engineering seems to very promising also for their performances[6], and will be to investigate for this specific task. But it was expected that classifier performance would be lower for the less well represented classes, an issue that will be challenging for every machine learning algorithm. Especially for rare diseases or documents reporting unusual situations, knowledge-based methods will be powerful complements.

Various methods have been explored in this area. In 2016 Mulligen et al. obtained the best results by combining a Solr tagger with ICD-10 terminologies at the CLEF eHealth. The terminologies were derived from the task training set and a manually curated ICD-10 dictionary. They achieved F-measure of 84.8%. Moreover the contribution of mixed methods makes perfect sense. Our dictionary-based lexical matching was too simplistic and only marginally improved the performance of CNN classifier even if the gain in the less well represented categories is interesting. Zwegembaum et al reported a similar combined approach and very promising results [7, 8].

6 Conclusion

We have studied CNN performance for multi-label classification in ICD-10 of death certificates. In order to take into account the low performance of machine learning methods for situations where data are reliably represented, we combine a dictionary based method with small improvement of performance on the most rare situations.

References

1. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Nvol, A., Ramadier, L., Robert, A., and Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018. In: 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2018.
2. Kim, Y.:Convolutional neural networks for sentence classification. In:arXiv preprint 2014. arXiv:1408.5882.
3. Hughes, M., Li, I., Kotoulas, S., Suzumura, T.: Medical text classification using convolutional neural networks. In: Stud Health Technol Inform. 2017;235:246250.
4. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 86(11):22782324, November 1998.
5. Miftakhutdinov Z., Tutubalina E.: KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks. In : CLEF 2017 Online Working Notes. CEUR-WS
6. Howard, J.,Ruder,S.: Universal Language Model Fine-tuning for Text Classification. In:arXiv preprint 2018.arXiv:1801.06146 [cs.CL]
7. Zweigenbaum, P., Lavergne, T.: Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In : CLEF 2017 Evaluation Labs and Workshop: Online Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, September 2017.
8. Zweigenbaum, P., Lavergne, T.: Hybrid methods for ICD-10 coding of death certificates. In: Seventh International Workshop on Health Text Mining and Information Analysis, pages 96-105, Austin, Texas, USA, November 2016. EMNLP 2016.
9. Nvol, A., Robert, A., Grippo, F., Lavergne, T., Morgand, C., Orsi, C., Pelikn, 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.