SITIS-ISPED in CLEF eHealth 2018 Task 1 : ICD10 coding using Deep Learning

Kévin Réby[1,2], Sébastien Cossin[1,2], Georgeta Bordea[1], and Gayo Diallo[1]

[1] Univ. Bordeaux, INSERM, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France
[2] CHU de Bordeaux, Pôle de sante publique, Service d'information médicale, Informatique et Archivistique Médicales (IAM), F-33000 Bordeaux, France

Abstract. This paper presents SITIS-ISPED's participation in the Task 1 : Multilingual Information Extraction - ICD10 coding of the CLEF eHealth 2018 challenge which focuses on extraction of causes of death from a corpus of death reports. We used OpenNMT, a deep learning framework specialized in sequence-to-sequence models. An encoder-decoder model with attention transformed diagnostics into ICD10 codes. Our system obtained F-scores of 0.4368 for the raw corpus and 0.5385 for the aligned corpus.

Keywords: Information Extraction · Natural Language Processing · Deep Learning · Neural Machine Translation · Seq2seq · Causes of Death · Death certificates · CepiDC · CLEF eHealth · International Classification of Diseases · ICD10.

1 Introduction

The amount of medical data available is constantly increasing[4] with a large majority of existing medical data available in unstructured free text format[15]. For health professionals this is the most simple and natural way to record and transmit information. But valuable information is often hidden in such large amounts of unstructured data. However, important medical questions can be addressed by analyzing these already existing medical data[16]. In this context, the main ambition of Natural Language Processing (NLP) is to create systems that automatically understand human language[5]. In the healthcare domain, a key step to reach this goal is the identification of medical entities such as diseases. Medical entites recognition is an important task for search, classification and deep phenotyping[15,17,18]. However, many issues due to the unstructured nature of data and specificities of medical language remain. Recently, tools and algorithms based on Deep Learning[7,8] have shown promising results for tackling such issues.

2 CLEF eHealth 2018 Task 1

The main objective of the "Conference and Labs of the Evaluation Forum" (CLEF) is to "promote research and stimulate the development of innovation in multilingual and multimodal Information retrieval systems for European (and non-European) languages through the creation of an evaluation infrastructure and the organization of regular evaluation campaigns and the creation of a multi-disciplinary research community"[2]. CLEF organizes various challenges. One of the CLEF eHealth 2018 edition is the Task 1: Multilingual Information Extraction - ICD10 coding[22]. The purpose of this task is to automatically assign World Health Organization's (WHO) International Classification of Diseases 10th revision (ICD10) codes to the text content of death certificates. The languages addressed in this year's challenge are French, Hungarian and Italian, but our focus is on French data alone in this work.

2.1 Death certificates and ICD-10

The production of death statistics is a routine activity in many countries. As a result, there is an international standardization of procedures. The certification of deaths by physicians is supervised by the format of the death certificate and by the concept of root cause of death. Codification of causes of death by nosologists is based on ICD10. It is a classification of diseases, signs, symptoms, social circumstances and external causes of disease or injury, published by the WHO. This coding system ensures a certain level of quality and international comparability of mortality data[6]. As a diagnostic coding system, it is used to classify the causes of health problems and deaths, and provide information for clinical purposes. Each medical concept is defined by a unique identifier consisting of a unique alphabetical prefix and several digits. The unique alphabetic prefix represents a class of common diseases, and digits represent a specific type of disease(See examples below)1.

Table 1. Examples of ICD-10 codes

Medical concepts	ICD-10 codes
Malignant neoplasm of unspecified part of bronchus or lung	C349
Unspecified dementia	F03
Other and unspecified firearm discharge, undetermined intent	Y24

2.2 Dataset for Task 1

The data comes from the Centre for Epidemiology of Medical Causes of Death within the National Institute of Health and Medical Research in France, or CépiDC, and comprises free text descriptions of causes of death, as reported by physicians in death certificates. Each document was manually coded by experts with ICD10 according to WHO international standards.

The corpus includes text descriptions of causes of death as reported by physicians in the standard cause of death forms with selected metadata :

- DocID: death certificate ID
- YearCoded: year the death certificate was processed by CépiDC
- Gender: gender of the deceased
- Age: age at the time of death, rounded to the nearest five-year age group
- LocationOfDeath: Location of death
- LineID: line number within the death certificate
- RawText: raw text entered in the death certificate
- IntType: type of time interval the patient had been suffering from coded cause
- IntValue: length of time the patient had been suffering from coded cause.

The training data set contains 125,383 death certificates and ICD10 codes that will serve as the gold standard. These data are in the form of csv files and include raw text from the death certificates along with metadata. The objective is an information extraction task that uses the text provided to extract ICD10 codes from certificates, line by line.

3 Methods

Motivated by the success of Deep learning in NLP tasks[8,9,10,11,12,13] and by the work of the Kazan Federal University (KFU) team in the 2017 edition of CLEF eHealth[1] (ICD10 coding of English Death Certificates), we decided to use deep neural networks for this task. The system proposed by KFU gave promising results using a Long Short Term Memory (LSTM) encoder - decoder architecture.

Sequence-to-sequence (Seq2seq) and Neural Machine Translation (NMT) methods are effective for many NLP and language-related applications such as dialogue, image captioning, and summarization[14]. In Seq2seq models, a sequence input is used to generate another sequence as an output.

A popular approach to transform one sequence into another is based on an encoder-decoder architecture consisting of two recurrent neural networks (RNNs) combined together with an attention mechanism that aligns target with source tokens[14] : an encoder network condenses an input sequence into a vector and a decoder network unfolds that vector into a new sequence. The attention mechanism lets the decoder learn to focus over a specific range of the input sequence[20].

Among all the frameworks available[13], we have chosen OpenNMT, an open source (MIT) initiative for Neural Machine Translation and neural sequence modeling. It is a general-purpose attention-based seq2seq system, originally developed by Yoon Kim, and then improved thanks to collaboration between the Harvard NLP group and Systran Paris[3]. OpenNMT has currently three main implementations : in Lua, Python and a TensorFlow version. We used OpenNMT-py, an OpenNMT-lua clone using PyTorch. OpenNMT requires a corpus of sentences: in our case, diagnostics and their corresponding ICD10 codes. First, the diagnostics and their ICD10 codes are extracted from the csv files, and then respectively splitted into a source text file and a target text file. This extraction is made by a simple bash program. In this way the data consists of parallel source (diagnosis) and target (ICD10 codes) data containing one sentence per line with words separated by a space. Then those data are splitted into two groups: one for training and one for validation. Validation files are required and used to evaluate the convergence of the training process. For source files, a first preprocessing step converts upper cases into lower cases. A tokenization process is applied on sources files and on target files which are used as input for the neural network (see Figure 1).

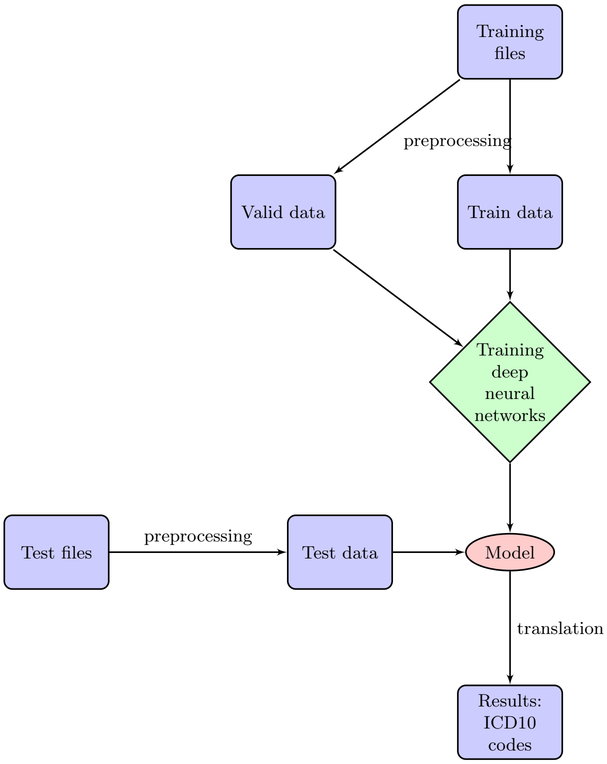

Fig. 1. Workflow of the training process.

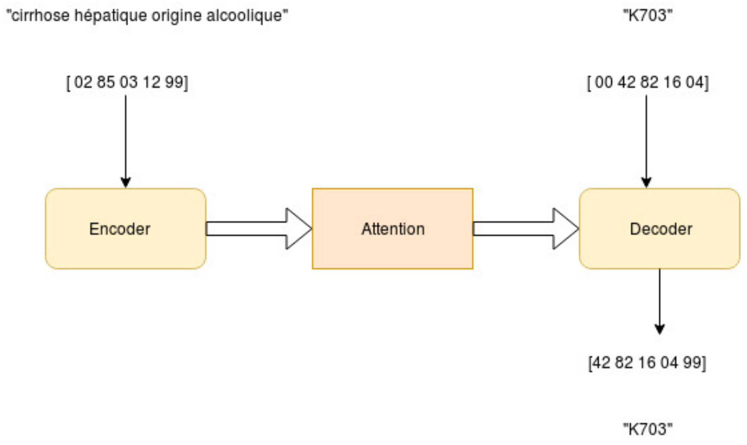

Fig. 2. Encoder/Decoder with attention scheme (adapted from Pytorch seq2seq tutorial).

The used encoder/decoder model consists of a 2 layers LSTM with 500 hidden units on both the encoder and decoder. The encoder encodes the input sequence into a context vector which is used by the decoder to generate the output sequence. More precisely, after the processing steps of a source sentence, the encoder takes as input the sequence of source tokens and produces a context vector, and it is given to the decoder. The context vector, also called the attention vector, is calculated as a weighted average of the source states. The decoder is an RNN that predicts the probability of a target sequence. The probability of each target token is predicted based on the recurrent state in the decoder RNN, the previous words, and the context vector (See figure2). The training process goes on for 13 epochs and provide a model.

4 Results

From the test data provided by the CLEF organization, we extracted the diagnostics, preprocessed them and used the model we created to "translate" them into their respective ICD10 codes. Table 1 shows the results obtained on the raw dataset with average and median preformance scores of the runs of all rask participants, Table 2 shows the results obtained on the aligned dataset. We achieved a precision of 0.6764 and a recall of 0.225 for the raw dataset, and a precision of 0.6649 and a recall of 0.4525 for the aligned dataset. Overall in both evaluations our results are lower than the average and median score of others participants.

Table 2. Results on the raw dataset

	Precision	Recall	F-measure
SITIS-ISPED run	0.676	0.323	0.4378
average	0.712	0.581	0.634
median	0.771	0.544	0.641

Table 3. Results on the aligned dataset

	Precision	Recall	F-measure
SITIS - ISPED run	0.665	0.453	0.539
average	0.723	0.410	0.507
median	0.798	0.475	0.579

5 Conclusion and discussion

This paper describes ICD10 coding by deep neural networks for the Task 1 of CLEF eHealth 2018. SITIS - ISPED participated for the first time to an evaluation challenge. This task allowed us to try a deep learing approach. We have used OpenNMT, an open-source toolkit for neural machine translation. We obtained promising results that we will try to improve. A possible explanation for the better results of KFU team is that it is due to the use of word embeddings and cosine similarities in their implementation[1]. In 2016 and 2017 for a similar task better performances are obtained from machine learning methods relying on knowledge based-sources (team LIMSI) and with a combination of knowledge based and NLP methods (team SIBM)[21]. In consequence for future works we will try other model architectures, arrange for the model to predict several ICD10 codes and use word embeddings. We envision a hybrid approach that combines a dictionary based approach with the deep learning approach described here.

Acknowledgments

This work is supported by the University of Bordeaux and the Bordeaux University Hospital. We want to thank the CLEF eHealth 2018 organization.

References

1. Z. Miftahutdinov, E. Tutubalina, KFU at CLEF eHealth 2017 Task1: ICD-10 coding of english death certificates with recurrent neural networks.
2. N. Ferro, CLEF 15th birthday: past, present, and future, ACM SIGIR Forum, Vol. 48 2 December 2014.
3. G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush, OpenNMT: Open-Source Toolkit for Neural Machine Translation, ArXiv e-prints, 2017.
4. W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health Information Science and Systems, 2014,2:3.
5. R. Collobert et al., Natural Language Processing (almost) from Scratch, Journal of Machine Learning Research, Vol.12, 2011.
6. G. Pavillon, F. Laurent, Certification et codification des causes médicales de décès, BEH №30-31, 2003.
7. Li Deng, Deep Learning: Methods and Applications, Foundations and Trends® in Signal Processing: Vol. 7: No. 3–4, pp 197-387.
8. Y. Kim, Convolutional Neural Networks for Sentence Classification, 2014
9. B. Shin, F. H. Chokshi, T. Lee, J. D. Choi, Classification of Radiology Reports Using Neural Attention Models, 2017.
10. Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, H. Xu, Entity recognition from clinical texts via recurrent neural network, The International Conference on Intelligent Biology and Medicine (ICIBM) 2016.
11. Y. Wu, M. Jiang, J. Lei, H. Xu, Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network, Stud Health Technol Inform, 2015, 216: 624–628.
12. Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, Anita Burgun, Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings, BMC Bioinformatics, 2014, 15:266.
13. B. J. Erickson, P. Korfiatis, Z. Akkus, T. Kline, K. Philbrick, Toolkits and Libraries for Deep Learning, J Digit Imaging (2017) 30:400–405.
14. D. Britz, A. Goldie, M. Luong, Q. Le, Massive Exploration of Neural Machine Translation Architectures, Google Brain, 2017.
15. B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis, 2018.
16. Richard Jackson, Rashmi Patel, Sumithra Velupillai, George Gkotsis, David Hoyle, Robert Stewart, Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records, F1000Research 2018, 7:210.
17. R. Miotto, F. Wang, S. Wang, X. Jiang and J. T. Dudley, Deep learning for healthcare: review, opportunities and challenges, Briefings in Bioinformatics , 2017, 1–11.
18. Riccardo Miotto, Li Li, Brian A. Kidd, Joel T. Dudley, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, Nature Scientic Reports, 2016, 6:26094.

19. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Palotti, J., Jimmy, 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).

20. Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, Neural Machine Translation by jointly learning to align and translate, published as a conference paper at ICLR 2015.

21. Aurelie Neveol, Robert N. Anderson, K. Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Gregoire Rey, Aude Robert, Claire Rondet, and Pierre Zweigenbaum, CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French, 2017.

22. 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. CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September, 2018.