KCL-Health-NLP@CLEF eHealth 2018 Task 1: ICD-10 Coding of French and Italian Death Certificates with Character-Level Convolutional Neural Networks

Julia Ive[1][0000−0002−3931−3392], **Natalia Viani**[1][0000−0003−2205−2322], **David Chandran**[1][0000−0002−0123−666X], **André Bittar**[1][0000−0001−6587−0080], **Sumithra Velupillai**[1,2][0000−0002−4178−2980]

[1] King's College London, IoPPN, London, SE5 8AF, UK,
firstname.surname@kcl.ac.uk
[2] KTH, Sweden

Abstract. In this paper we describe the participation of the KCL-Health-NLP team in the CLEF eHealth 2018 lab, specifically Task 1: Multilingual Information Extraction - ICD10 coding. The task involves the automatic coding of causes of death in death certificates in French, Italian and Hungarian according to the ICD-10 taxonomy. Choosing to work on the two Romance languages, we treated the task as a sequence-to-sequence prediction problem. Our system has an encoder-decoder architecture, with convolutional neural networks based on character embeddings as encoders and recurrent neural network decoders. Our hypothesis was that a character-level representation would allow our model to generalise across two genealogically related languages. Results obtained by pre-training our Italian model on the French data set confirmed this intuition. We also explored the impact of character-level features extracted from dictionary-matched ICD codes. We obtained F-measures of 0.72/0.64 and 0.78 on the French aligned/raw and Italian raw internal test data, respectively. On the blind test set released by the task organisers, our top results were 0.65/0.52 and 0.69 F-measure, respectively.

Keywords: Encoder-decoder architecture · Convolutional neural networks · Recurrent neural networks

1 Introduction

The task of identifying medical entities mentioned in textual documents and linking them to external terminologies has been the subject of much research in the fields of natural language processing (NLP) and bioinformatics. Given the high volumes of unstructured text in electronic health records, such methods for normalising this information are potentially of great use in areas such as healthcare administration, quality of care and epidemiological research.

Task 1 of the 2018 CLEF eHealth Lab [14,12] is centred on the automatic identification of causes of death as mentioned in French, Italian and Hungarian

death certificates and their mapping onto the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [17].

Table 1. Some characteristics of the language used in death certificates for French and Italian.

i.	**Typographical/spelling errors**	**ICD10**
	FR *blesuure thoracique (blessure thoracique)*	S299
	bronchopneumonie d edéglutition (bronchopneumonie de déglutition)	J690
	IT *ARTEROPATIA (ARTERIOPATIA)*	I779
ii.	**Orthographical variants**	
	FR *adénocarcinome colique, adéno carcinome colique, adéno-carcinome colique, ADENOCARCINOME COLIQUE*	C189
	IT *SCOMPENSO CARDIOCIRCOLATORIO, SCOMPENSO CARDIO CIRCOLATORIO*	I516
iii.	**Omission of diacritics**	
	FR *ulcere duodenal – ulcère duodénal*	K269
iv.	**Lexical variants**	
	FR *arrêt cardio-respiratoire, arrêt ventilatoire, décompensation cardiorespiratoire*	R092
	IT *ENTERITE DA CLOSTRIDIUM, GASTROENTERITE DA CLOSTRIDIUM, ENTERITE DA CLOSTRIDIUM DIFFICILE, INFEZIONE DA CLOSTRIDIUM DIFFICILE*	A047
v.	**Acronyms**	
	FR *AVP – accident de la voie publique*	V892
	HTA – hypertension artérielle	I10
	IT *IRC – isufficienza renale cronica*	N189
	FA CRONICA – fibrillazione atriale cronica	I482
vi.	**Abbreviations**	
	FR *septicémie staph – septicémie à staphylocoque*	V892
	IT *INSUFF RENALE – insufficienza renale*	N19
vii.	**Morpho-syntactic variants**	
	FR *oedème pulmonaire aigu – oedème aigu du poumon*	I501
	IT *ADENOCARCINOMA DEL POLMONE – ADENOCARCINOMA POLMONARE*	C349

The ICD-10 terminology contains codes for diseases, symptoms, and causes of injury or death, along with other medical concepts. Each concept is represented by a normalised form that is associated with a unique code. Each code is made up of a letter, which indicates the type of disease or other concept, followed by a

sequence of numbers that further specify it. For example, the letter G concerns *Diseases of the nervous system*, and G00.1 is the unique code for *Pneumococcal meningitis*. The ICD-10 terminology is developed for numerous languages, and the base classification contains thousands of unique codes. For the shared task, participating teams were provided with death certificates in which each line was annotated with gold standard ICD codes corresponding to the cited causes of death. For the challenge, systems were required to automatically assign correct codes to each line of unseen death certificates.

There are numerous challenges inherent to this task, many of which are due to the nature of the data itself. In particular, the correspondence between the actual texts written by doctors and the normalised entries in ICD dictionaries and terminologies is not guaranteed. Table 1 lists some examples of the difficulties that were to be confronted in the 2018 CLEF shared task data sets for French and Italian[3].

Firstly, although concise, the unstructured text data of death certificates sometimes contains certain "imperfections", such as typographical and spelling errors (i). Furthermore, clinicians often record terms using different orthographical variants (ii) which must be accounted for by an automated system. In languages such as French, arbitrary use of diacritics (iii) adds a further complexity to the task. A single medical concept may correspond to a range of lexical variants (iv), including acronyms (v) and abbreviations (vi). Finally, morpho-syntactic variants may also exist for a single concept (vii). These characteristics of the source texts, among others, limit the efficacy of simple dictionary lookup techniques and complicate the selection of the appropriate cause of death code from the standardised ICD terminology.

Our main contributions to this lab are two-fold. Firstly, we explored the application of character-level convolutional neural networks (CNNs) to the automatic ICD-10 coding of death certificates written in French and Italian. We paid particular attention to the ability of our approach to generalise across languages. To our knowledge, this is the first attempt to apply such techniques to this task. Secondly, we also studied the extent to which character-level features extracted from dictionary-matched ICD codes contributed to performance.

2 Related Work

In previous years of the CLEF ICD coding shared task, participants have used a variety of methods to tackle the problem, including rule- or dictionary-based approaches, and systems based on machine learning. In the 2017 challenge [10], the target languages were English and French.

That year, rule-based approaches to the task included a system using dictionary look-up and a set of priority rules [7], a multilingual, fuzzy matching dictionary-based system [2], and an acronym translation system coupled with a binary weighting and tf-idf similarity measure to match dictionary entries to the death certificate texts [11].

[3] We do not provide examples for Hungarian as we did not deal with this language.

Among approaches implementing traditional machine learning techniques, the system of [21] used dictionary projection coupled with a linear SVM classifier trained on bag-of-words features.

The hybrid system of [16] combined rule-based Named Entity Recognition and two-strategy dictionary matching for candidate code selection with machine learning classifiers for candidate ranking.

Finally, state-of-the-art neural network approaches also featured in the 2017 shared task. [9] used an encoder-decoder RNN-based approach initialised with word embeddings pre-trained on social media user posts. This approach fed death certificates into the network as sequences of words, ignoring the division into lines. The authors also used ICD code similarity vectors as additional information. This team achieved the highest performance on the English dataset.

Outside of the CLEF shared tasks, systems based on neural networks have also been developed for other languages and other types of clinical documents. [6] implemented a hierarchical recurrent neural network (RNN) to code death certificates in Portuguese, so that document-level representations used for classification decisions were composed of sentence representations, which are in turn composed of word representations. [3] propose an algorithm which uses a similarity measure based on sub-string matching to map disease names found in Chinese clinical notes to a dictionary of ICD-10 codes. For the coding of diagnoses in English, [13] also deploy a hierarchical RNN neural architecture with an attention mechanism. Their system builds sentence representations from word representations, which are, in turn, built from character representations.

3 Methodology

To address the problem of automatic ICD-10 coding of death certificates we implemented a hierarchical encoder-decoder architecture [15,1]. This is an architecture for sequence-to-sequence prediction problems. It has become state-of-the-art in addressing various NLP problems, where inputs and outputs to models are often sequential.

The architecture works as follows: an input sequence is first encoded into an internal representation (roughly speaking, a set of automatically learned features). An output sequence is then generated from this representation. Current best practice is to implement both encoders and decoders using RNNs. RNNs often combined with attention mechanisms applied to this task have been shown to achieve good performance [9,6,13]. The performance of CNNs remains less well studied. Our intuition was that CNNs would be particularly effective for this task. This is because, when applied at the character level, they are known to perform well for noisy data, such as that found in death certificates. CNNs are designed to extract only crucial input features [8,20]. Furthermore, given that French and Italian, both Romance languages, are genealogically related, and similar on a character level, we hypothesised that using a character-level representation would make for models that would generalise across these languages.

We used the French data set to pre-train the architecture and tested whether this pre-training could be beneficial for our Italian system. The only external data we used are the ICD-10 dictionaries (for both French and Italian) provided by the organisers of the task.

With this in mind, and drawing inspiration from the work of [18], who use a character-level representation for document classification, we combine CNN encoders with RNN decoders in our system.[4] The system's general architecture is presented in Figure 1.

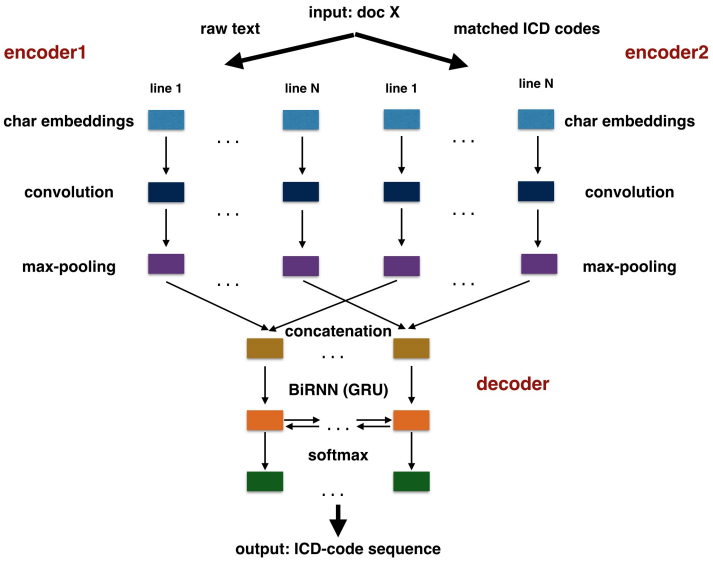

Fig. 1. Our hierarchical encoder-decoder architecture.

The input to our model is a batch of death certificates, each one made up of lines of raw text. Each line is encoded at the character level by a standard series of 3 CNN layers – a convolutional layer, a max-pooling layer and a flattening layer. Each sequence of line representations is fed into a bidirectional RNN (biRNN, decoder). The resulting sequence of representations is provided to the `softmax` output layer that produces a probability distribution over the set of possible labels.

In our model we also experiment with a second CNN encoder of exactly the same structure. The input to this encoder is a batch of death certificates each represented by their lines, which are, in turn, represented by their matching ICD-10 codes (see Section 3.2). We also extract features from these code sequences at the character level. The intuition behind this is that matched codes can be

[4] The code is publicly available online: https://github.com/KCL-Health-NLP/clef2018-char-cnn-death-certificates.

partially correct and give clues as to the correct codes with a difference of several characters. The outputs of this second encoder are concatenated to the outputs of the first encoder and provided directly to the decoder.

Our architecture fits both aligned and raw French tasks: the only difference is that for the aligned data the length of the input line sequence is the length of the output code sequence for a certificate, whereas these length values are different for the raw data. These regularities are captured by the network during training.

3.1 Dataset

Both French datasets (raw and aligned) included a training set (2006-2012), a development set (2013), and a test set (2014). The Italian dataset, however, was provided as a whole, without any further division. For this reason, we split it into three subsets of equal size, which were used as internal training, development, and test sets (see Table 2). While creating our split we took into account the statistics of unique and unseen ICD-codes from the official French raw data split.[5]

Table 2. Statistics over the internal raw Italian data split.

internal set	certificates	number of lines	unique codes
training	4835	16590	966
development	4833	16651	972
test	4833	16583	1013

3.2 Dictionary string matching

We developed a baseline string matching method in order to provide additional cues to the neural model for each certificate. We used the dictionaries for each language that were provided by the challenge organisers.

We preprocessed the text of each death certificate using the Natural Language Toolkit (NLTK)[6] in the following ways: removing diacritics and stopwords (using default stopwords as provided by NLTK), performing stemming (using the NLTK Snowball stemmer for each language) and converting all text to lower case. We also preprocessed all strings associated with an ICD code in the dictionaries in the same way. These steps were intended to help overcome some of the irregularities in the texts (as mentioned in the introduction) and improve the correspondence with the dictionaries.

[5] We took into account the average number of codes per certificate, the total number of unique codes per split and the total number of unique unseen codes per split for the raw French dataset.

[6] https://www.nltk.org/, version 3.2.5

We further preprocessed the texts of the certificate by generating all possible word n-grams (1-5), to be matched against the dictionary entries. Then, to match a potential ICD code for each preprocessed certificate text, we applied set intersection to identify string overlap between the certificate text and all strings associated with each ICD code. This resulted in a list of potential matched ICD codes for each certificate text. In addition, we sorted the matched codes by their frequency in the training data, i.e. if one of the matched codes was frequent in the training data it was placed earlier in the list.

For the aligned French data, we also added a step to split the list of found ICD codes based on their position in the certificate text. For each text where more than one code was found, the index position of where the strings matched was saved in ascending order, and recorded separately for each line. A paraphrased example is given in Table 3.

Table 3. Paraphrased example from the French aligned data, illustrating the dictionary string matching approach split by the position in which an ICD code was found.

raw text	standard text	gold	matched
Tumeur broncho-pulmonaire maligne, avec métastases cutanées	tumeur broncho-pulmonaire maligne	C349	['C349', 'C340']
Tumeur broncho-pulmonaire maligne, avec métastases cutanées	métastases cutanées	C792	['C799', 'C792']

3.3 Implementation Details

We implemented our hierarchical architecture using the `Keras` [5] toolkit. The character embedding dimensionality was set to 300. For encoders, we used convolution layers with 256 hidden units with a window of 3 and the ReLU activation function. The size of the max-pooling layer was set to 2. We initialised the weights using the Glorot uniform initialiser.[7] For the decoder, we used Gated Recurrent Units (GRUs) [4] as RNNs. The size of the hidden units of the RNN decoder was 300.

For both French and Italian configurations,[8] we limited the raw text length of input lines to 49 characters (empirically chosen value, which corresponds to the third quartile of the overall distribution of French line length values). We limited the number of considered matched codes to 5 (≈20 characters, empirically chosen value). We limited the certificate length to 6 for both (empirically chosen value, which corresponds to the third quartile of the overall distribution

[7] This configuration was inspired by [20].

[8] The French configuration was re-used for Italian to ensure model compatibility for shared training experiments. Statistics for both languages are similar.

of French certificate length values, to which we add two lines to match the resulting ICD-code sequence length[9]). The labels present in the French training and development data were considered as possible output labels (the dimensionality of the `softmax` layer).[10] All the characters from French or Italian data were used in the vocabulary.

For training, we used a mini-batch size of 50. The model was trained to minimize the categorical cross-entropy error loss using the Adadelta optimiser [19]. We employed early stopping with the patience of 5 epochs based on validation loss.

We pre-trained model weights on the raw French data. These weights were used to initialise the Italian model. To ensure the compatibility of models the character vocabulary contained all the characters found in both French and Italian training data. The labels present in both French and Italian training and development were considered as possible output labels.[11]

4 Results

Table 4 shows the results obtained on internal test sets, for both French (aligned and raw) and Italian (raw) data. The performance is reported in terms of precision (P), recall (R), and F-measure (F), and was computed with the evaluation script provided by the task organisers. For each dataset, the first row represents the dictionary string matching baseline (see Section 3.2), while the second and the third rows report results for the different encoder-decoder (enc.-dec.) models. For French, we ran experiments using only raw text inputs ("enc.-dec. NN") and including additional features extracted from lookup codes ("enc.-dec. NN w/ lookup codes"). For Italian, we first used this same architecture ("enc.-dec. NN w/ lookup codes"), and we then investigated the effects of pre-training of weights on the French data ("enc.-dec. NN w/ lookup codes, pre-trained").

As mentioned, each ICD-10 code is a sequence of characters, starting with a letter (a general Chapter) followed by 2-4 numbers (identifying the specific concept). From a manual analysis of results, we found that some of the labels predicted by the models overlap only partially with the provided gold standard labels. To evaluate the extent of this, we re-computed performance by requiring only the first n characters of ICD-10 codes to match. Results are shown in Table 5, for matches on the first three ($F_{n=3}$), two ($F_{n=2}$), or one ($F_{n=1}$) characters.

[9] 90% of ICD-code sequences are maximum 2-lines longer than corresponding certificates

[10] Labels from the development set were added to correctly estimate the criteria for early stopping and at the same time avoid adding an `UNK` (unknown) label. We noticed that adding this label to possible output labels decreased the final performance as the model tended to resort to this label in case of uncertainty.

[11] Note that a model trained on the concatenation of the French and Italian data did not outperform this configuration.

Table 4. Results on internal test sets: official evaluation (precision, recall, F-measure).

Dataset	System	P	R	F
FR aligned	dictionary string matching baseline	0.2962	0.7248	0.4205
	enc.-dec. NN	0.7677	0.6301	0.6921
	enc.-dec. NN w/ lookup codes	0.8012	0.6493	0.7173
FR raw	dictionary string matching baseline	0.3101	0.6790	0.4258
	enc.-dec. NN	0.7547	0.5046	0.6048
	enc.-dec. NN w/ lookup codes	0.7998	0.5327	0.6394
IT raw	dictionary string matching baseline	0.3528	0.5994	0.4442
	enc.-dec. NN w/ lookup codes	0.8131	0.6973	0.7507
	enc.-dec. NN w/ lookup codes, pre-trained	0.8447	0.7246	0.7801

Table 5. Results on internal test sets: evaluation on first n characters (F-measure).

Dataset	System	F	$F_{n=3}$	$F_{n=2}$	$F_{n=1}$
FR aligned	dictionary string matching baseline	0.4205	0.5016	0.5458	0.6237
	enc.-dec. NN	0.6921	0.7177	0.7416	0.8054
	enc.-dec. NN w/ lookup codes	0.7173	0.7487	0.7725	0.8255
FR raw	dictionary string matching baseline	0.4258	0.5114	0.5614	0.6472
	enc.-dec. NN	0.6048	0.6349	0.6658	0.7445
	enc.-dec. NN w/ lookup codes	0.6394	0.6731	0.7046	0.7744
IT raw	dictionary string matching baseline	0.4442	0.4601	0.5275	0.6649
	enc.-dec; NN w/ lookup codes	0.7507	0.7682	0.7988	0.8515
	enc.-dec. NN w/ lookup codes, pre-trained	0.7801	0.8012	0.8339	0.8782

The proposed models were developed and evaluated internally. Table 6 shows the results of the evaluation conducted on the official test sets (following the same structure as for Table 4).

Table 6. Results on official test sets (precision, recall, F-measure).

Dataset	System	P	R	F
FR aligned	enc.-dec. NN	0.7868	0.5525	0.6492
	enc.-dec. NN w/ lookup codes	0.7694	0.5365	0.6322
FR raw	enc.-dec. NN	0.738	0.405	0.5229
	enc.-dec. NN w/ lookup codes	0.7235	0.3939	0.5101
IT raw	enc.-dec. NN w/ lookup codes	0.7459	0.6358	0.6865
	enc.-dec. NN w/ lookup codes, pre-trained	0.7252	0.6162	0.6662

5 Analysis and Discussion

On the internal data split, utilising the results of the string matching algorithm as additional features to the neural network model showed a marginal but

promising improvement in terms of F-measure (Table 4, "FR aligned" and "FR raw" rows). Starting from this observation, applying a string matching algorithm could effectively support the training of neural network models, whenever external dictionaries are available. As regards the Italian dataset, it can be seen that pre-training on the French datasets shows an improvement over running the model without pre-training (Table 4, "IT raw" row). This shows that the model can successfully operate on a multi-lingual level. This represents an interesting result, that should be further investigated in other multi-lingual tasks.

An interesting observation from Table 4 concerns the different results obtained on the two French datasets. In terms of the baseline string matching algorithm, the F-measures were very similar for raw and aligned data, illustrating no difference between either dataset for that approach. Our encoder-decoder model instead obtained a substantially higher F-measure for the aligned data than the raw data, irrespective of whether or not lookup codes were used. This can be explained by the intrinsic comparative simplicity of the aligned task in which the length of the input sequence is equal to the length of the output sequence, a characteristic which is trivially captured by the neural model.

Another notable observation is that the precision was very similar (for both the neural network approaches) for both the raw and aligned datasets, with the greatest improvement being due to an increase in recall. This creates a potential case for the utility of aligned data of tasks that are dependent on maximising recall.

Comparing the results obtained on French and Italian raw data, it can be observed that our architecture performed substantially better for the latter, both with and without pre-training. There are a number of possible reasons for this. One key explanation is that there is a greater level of consistency in the Italian certificates, naturally favouring higher performance.

As another interesting observation, comparing the results of string matching and neural network models over French and Italian data, there is a difference in terms of recall. For both the French datasets, the string matching algorithm produces improved recall over the neural network models (with dramatically lower levels of precision). This characteristic of recall is reversed for the Italian data, where both the proposed models demonstrated substantially higher recall than the string matching algorithm (while the string matching algorithm's precision remained low). We conclude that, for Italian, the use of external dictionaries could have a lower impact on recall than for the French data.

A key observation from Table 5 is the general performance improvement going from looking at the full ICD-10 code ("F" column) to looking at the code on a chapter level (the first character). A graphical representation of this is provided in Figure 2.

For the baseline string matching algorithm, a substantial increase in F-measure can be observed across all datasets. However, the linearity of these improvements varies across French and Italian datasets. In the case of French aligned and raw datasets, the string matching algorithm showed substantial increases between F and $F_{n=3}$, a smaller increase between $F_{n=3}$ and $F_{n=2}$, and a

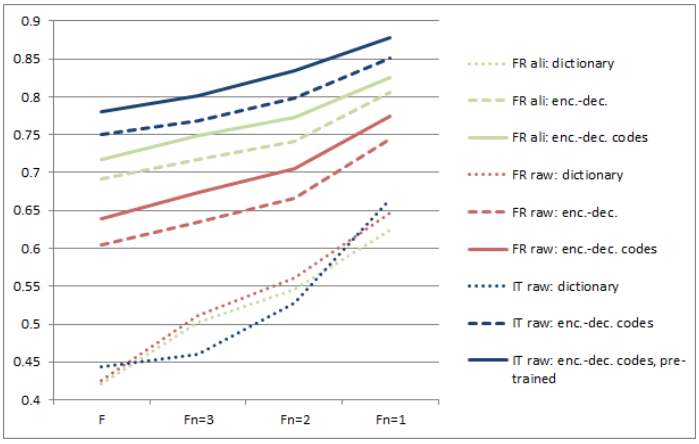

Fig. 2. Evaluation on different numbers of code characters.

large increase again between $F_{n=2}$ and $F_{n=1}$. The Italian dataset, on the other hand, showed an almost steady increase between F and $F_{n=2}$, before showing a large increase between $F_{n=2}$ and $F_{n=1}$. For all neural network models, an improvement can also be seen when considering only the first characters of ICD-10 codes. However, the level of improvement is much smaller if compared to the string matching baseline.[12] Nevertheless, some of the codes that are incorrectly extracted by the system could actually be used to identify the correct ICD-10 code's chapter or higher-level specification.

As shown in Table 6, running the developed models on the blind test data, we did not see the same types of improvement as for the internal data split. For both French datasets, utilising string matching codes as inputs to the neural model actually led to a slight decrease in the F-measure. Similarly, pre-training the Italian model with weights derived from the French models using string matching codes did not support the neural network training. This represents an interesting result, showing that there might be differences between the two datasets which have an impact on the performance of the developed models. Further work will be needed to test the generalisability of the proposed architecture, as well as to investigate how to leverage dictionary-based approaches.

6 Conclusions and Future Work

To address the automatic extraction and normalisation of ICD-10 codes in French and Italian death certificates, we proposed an encoder-decoder approach relying on character-level embeddings. This allowed us to reuse the same architecture to process texts written in two different languages, without the need for specific

[12] This could be explained by the limitations of the architecture taking only the observed labels into account (see Section 3.3).

preprocessing steps. In an effort to improve the system performance, we also ran experiments by searching for string matches in external dictionaries, thus obtaining additional inputs to guide the training of the network. This lookup step led to improvements in both precision and recall in our internal data split.

In future, it would be interesting to test the generalisability of our character-level approach to other languages. We believe that our approach could be effective even for distantly related languages since medical terminology shares a certain amount of similarity across languages.

We also plan to extend our architecture with attention mechanisms and study their performance. As future development, we plan to explore the contribution of other features, for instance other ways of informing the model with matched ICD-10 codes e.g. by chapter or other hierarchical attributes, exploiting knowledge about changes in coding standards over time, and incorporating information about gender and age.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of International Conference on Learning Representations (ICLR) (2015)
2. Cabot, C., Soualmia, L.F., Darmoni, S.J.: SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with p. 11
3. Chen, Y., Lu, H., Li, L.: Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. PLOS ONE **12**(3) (Mar 2017)
4. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR **abs/1406.1078** (2014)
5. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
6. Duarte, F., Martins, B., Pinto, C.S., Silva, M.J.: Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. Journal of Biomedical Informatics **80**, 64–77 (Apr 2018)
7. Jonnagaddala, J., Hu, F.: Automatic coding of death certificates to ICD-10 terminology p. 8
8. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. CoRR **abs/1508.06615** (2015)
9. Miftahutdinov, Z., Tutubalina, E.: KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English death certificates with recurrent neural networks p. 11
10. Neveol, A., Anderson, R.N., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Rondet, C., Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French p. 17
11. Nunzio, G.M.D., Beghini, F., Vezzani, F., Henrot, G.: A lexicon based approach to classification of ICD10 Codes. IMS Unipd at CLEF eHealth Task p. 9 (2017)
12. Névéol, A., Robert, A., Grippo, F., Lavergne, T., 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)

13. Shi, H., Xie, P., Hu, Z., Zhang, M., Xing, E.P.: Towards automated ICD coding using deep learning. arXiv:1711.04075 [cs] (Nov 2017)
14. 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)
15. Sutskever, I., Vinyals, O., Le, Q.V.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27, pp. 3104–3112 (2014)
16. Ševa, J., Kittner, M., Roller, R., Leser, U.: Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017 p. 8
17. World Health Organization: International statistical classification of diseases and related health problems. (2016), oCLC: 910334285
18. Xiao, Y., Cho, K.: Efficient character-level document classification by combining convolution and recurrent layers. CoRR **abs/1602.00367** (2016)
19. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method. CoRR **abs/1212.5701** (2012)
20. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15). pp. 649–657. MIT Press, Cambridge, MA, USA (2015)
21. Zweigenbaum, P., Lavergne, T.: Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates p. 11