Deep Learning for ICD Coding: Looking for Medical Concepts in Clinical Documents in English and in French

Zulfat Miftahutdinov[1,2] and Elena Tutubalina[1,2(✉)]

[1] Kazan (Volga Region) Federal University, Kazan, Russia
zulfatmi@gmail.com, ElVTutubalina@kpfu.ru
[2] Neuromation OU, 10111 Tallinn, Estonia

Abstract. Medical Concept Coding (MCD) is a crucial task in biomedical information extraction. Recent advances in neural network modeling have demonstrated its usefulness in the task of natural language processing. Modern framework of sequence-to-sequence learning that was initially used for recurrent neural networks has been shown to provide powerful solution to tasks such as Named Entity Recognition or Medical Concept Coding. We have addressed the identification of clinical concepts within the International Classification of Diseases version 10 (ICD-10) in two benchmark data sets of death certificates provided for the task 1 in the CLEF eHealth shared task 2017. A proposed architecture combines ideas from recurrent neural networks and traditional text retrieval term weighting schemes. We found that our models reach accuracy of 75% and 86% as evaluated by the F-measure on the CépiDc corpus of French texts and on the CDC corpus of English texts, respectfully. The proposed models can be employed for coding electronic medical records with ICD codes including diagnosis and procedure codes.

Keywords: ICD coding · ICD codes · Medical concept coding
Medical record coding · Computer assisted coding
Recurrent neural network · Encoder-decoder model · Deep learning
Machine learning · Death certificates · CépiDc corpus · CDC corpus
Healthcare · CLEF eHealth

1 Introduction

Medical concept coding is an important task of biomedical information extraction (IE), which is also a central concern of the text mining research community in recent years. The goal of IE is to automatically detect a textual mention of a named entity in free-form texts and map the entity mention to a unique concept in an existing ontology after solving the homonymy problem [1]. There are several widely used ontologies of medical concepts such as the Unified Medical Language System (UMLS), SNOMED CT, and International Classification of Diseases (ICD, ICD-10).

P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 203–215, 2018.
https://doi.org/10.1007/978-3-319-98932-7_19

The problem of homonymy, i.e., of disambiguation of unrelated word meanings, is one of the well-known challenges in natural language processing (NLP), which could be found in each and every NLP sub-fields and related areas like information retrieval. The drug discovery application sub-field is no exception in that regard, but it also has its own unique features. Namely, it is typical for the field that semantic unit here is an entity consisting typically of two and more words or abbreviations. Thus, one needs to disambiguate the meaning of an entity rather than a single word. For example, "headache" could mean migraine, or dizziness, or a few additional discrepant medical terms. This task in the field is called medical concept mapping, and disambiguation is one of its main features.

In this paper, we focus on the problem of ICD-10 coding, the aim is to assign codes from the International Classification of Diseases to fragments of texts. Computer-assisted coding (CAC) can help reduce the coding burden. CAC systems are already in use in many healthcare facilities as a helpful tool for increasing medical coder productivity [2]. Thus, progress in automated methods for ICD coding is expected to directly impact real-world operations.

The problem of accurate identification of ICD codes based on verbal description of medical conditions naturally lends itself to using NLP approaches for the task at hand. Since manual coding is time-consuming and error-prone, automatic coding has been studied for many years. Two basic methods of identifying ICD codes are dictionary matching and pattern matching [3]. Recent advances in neural networks have deeply reshaped NLP research because of their capability to learn representations from data without feature engineering in an end-to-end manner. Recent studies treat the medical concept coding task as a supervised sequence labeling problem. For instance, Miftahutdinov and Tutubalina [4] proposed an encoder-decoder model based on bidirectional recurrent neural networks (RNNs) to translate a sequence of words into a sequence of medical codes; experiments were carried out on the English corpus of death certificates. Karimi et al. [5] leveraged a simple convolutional neural network with a fully-connected layer to assign a label (a diagnosis code) for entries in a dataset of radiology reports written in English. Duarte et al. [6] applied a deep neural network that processes the texts of clinical reports from the Portuguese Ministry of Health. These works demonstrate the first attempts to use deep learning methods for ICD coding.

This work is a significantly extended version of the previously reported study [4]; here, we extended experiments to employ novel RNN architectures. In addition to Long Short-Term Memory (LSTM), we utilize Gated Recurrent Units (GRU) used for sequence learning. We explore the impact of different word embeddings and the length of output sequences of ICD codes. We conduct extensive experiments on the French and English datasets from the CLEF eHealth shared task 2017 and demonstrate the efficiency of our approach.

2 Related Work

Different approaches have been developed for medical concept coding task, mainly falling into two categories: (i) knowledge-based methods [7–11]; and (ii) machine learning approaches [12–14].

The *ShARe/CLEF eHealth 2013* lab addressed the problem of identification and normalization of disorders from clinical reports in Task 1 [15]. Leaman et al. introduced a DNorm system for assigning disease mentions from PubMed abstracts [16]. The *CLEF Health 2016 and 2017* labs addressed the problem of mapping death certificates to ICD codes. Death certificates are standardized documents filled by physicians to report the death of a patient [17]. For the CLEF eHealth 2016 lab, five teams participated in the shared task 2 about the ICD-10 coding of death certificates in French [18]. Most submitted methods utilized dictionary-based semantic similarity and, to some extent, string matching. Mulligen et al. [9] obtained the best results in task 2 by combining a Solr tagger with ICD-10 terminologies. The terminologies were derived from the task training set and a manually curated ICD-10 dictionary. They achieved an F-measure of 84.8%. Zweigenbaum and Lavergne [19] utilized a hybrid method combining pre-processing steps (stop word removal, diacritic removal, correction of some spelling errors), simple dictionary projection, and mono-label supervised classification. They used Linear SVM trained on the full training corpus and the 2012 dictionary provided for CLEF participants. This hybrid method obtained an F-measure of 85.86%. The participants of the CLEF eHealth 2016 task 2 did not use word embeddings or deep neural networks.

The CLEF eHealth 2017 ICD-10 coding task provided datasets which consisted of death certificates in French and English [17]. Nine teams participated in the shared task 1. Cabot et al. [20] applied a combination of a dictionary-based approach and fuzzy match algorithms. Their system obtained an F-measures of 76.36% on French records and 80.38% on English records. Zweigenbaum and Lavergne extended their hybrid method [19] to multi-label classification. They obtained F-measures of 82.5% and 84.7% on French and English texts, respectively. Miftakhutdinov and Tutubalina [4] obtained the best results in the CLEF eHealth 2017 task 1, training an LSTM-based encoder-decoder architecture. As input, the network uses the certificates' text lines containing terms that could be directly linked to a single ICD-10 code or several codes. As output, the network predicts a sequence of codes. The model obtained an F-measure of 85% on English texts. In this paper, we extend experiments with neural networks on a corpus of French certificates.

Although deep neural network models and word embedding techniques are widely used in most natural language processing task, so far they have found limited use for the medical domain texts. Nevertheless, first studies towards using neural networks for medical concept coding could be noticed [4–6,21,22]. For instance, Karimi et al. [5] leveraged a simple convolutional neural network and fully-connected layer to assign a single label (an ICD code) on a dataset of radiology reports. Duarte et al. [6] applied bidirectional GRU-based neural networks for the assignment of ICD-10 codes to the death certificates, together with the associated autopsy reports and clinical bulletins, from the Portuguese Ministry of Health. We note that those works did not discuss experimental comparison of their methods for one-label and multi-label classification of clinical texts.

2.1 Materials and Methods

In this section, we discuss challenges in the task, our datasets, and proposed approaches. There are several challenges to concept coding as well as entity and word disambiguation:

- **Textual variations**. Clinical records have multiple mention forms, including lexical, morphological, and syntactic variations, synonyms (hypertension vs. high blood pressure disorder), abbreviations (attention deficit hyperactivity disorder vs. ADDH vs. ADHD), alternate spellings or grammatical errors (diarrheas vs. diarrhoea).
- **Multiple overlapping entities**. Boundaries of different entities in the text could be not well defined. For example, the sentence "metastatic adencarcinoma of lung to brain"is associated with two concepts:"Malignant neoplasm of unspecified part of bronchus or lung" (C349) and "Secondary malignant neoplasm of brain and cerebral meninges" (C793).
- **Ambiguity**. A single mention, like aspiration, can match multiple UMLS entries, e.g. Endotracheal aspiration, Pulmonary aspiration, Aspiration Pneumonia, Aspiration precautions. We note that a great number of ambiguous words in the biomedical domain are actually abbreviations [23].

The combination of these challenges makes concept coding especially challenging with simple string matching algorithms and dictionary-based approaches.

2.2 Corpora

We briefly describe two real-world datasets used in our study. **The CépiDc corpus** and **the CDC corpus** consist of free-form text death certificates in French and English, respectively. These corpora were provided for the task of ICD10 coding in CLEF eHealth 2017 (Task 1).

The CépiDc corpus was provided by the French institute for health and medical research (INSERM). It consists of free text death certificates collected from physicians and hospitals in France over the period of 2006–2014. The corpus consists of 65,844, 27,850, and 31,690 raw texts for training, developing and testing, respectively. The full set includes 131,426 codes (2,527 unique codes). Statistics of the corpus are presented in Table 1. We note that the CépiDc corpus contains 6 times more certificates than the CDC corpus. We utilize the 'raw' version of the CépiDc corpus for further experiments.

The CDC corpus was provided by the American Center for Disease Control (CDC). The corpus consists of free text death certificates collected electronically in the United States during the year 2015. The corpus consists of 13,330 and 14,833 raw texts for training and testing, respectively. Additionally, the CDC test set includes the "external" test set which is limited to textual fragments with ICD codes linked with a particular type of deaths, called"external causes" or violent deaths. The full set includes 18,928 codes (900 unique codes), while the "external" set includes only 126 codes (28 unique codes). Statistics of the corpus are presented in Table 2. Examples of raw texts from death certificates with medical concepts and ICD codes are presented in Table 3.

Table 1. Statistics of the CépiDc corpus from [24].

	Train	Development	Test
Certificates	65,844	27,850	31,690
Year	2006–2012	2013	2014
Lines	195,204	80,899	91,962
Tokens	1,176,994	496,649	599,127
Total ICD codes	266,808	110,869	131,426
Unique ICD codes	3,233	2,363	2,527
Unique unseen ICD codes	-	224	266

Table 2. Statistics of the CDC American death certificates corpus from [24].

	Train	Test
Certificates	13,330	6,665
Year	2015	2015
Lines	32,714	14,834
Tokens	90,442	42,819
Total ICD codes	39,334	18,928
Unique ICD codes	1,256	900
Unique unseen ICD codes	-	157

3 Our Approach

The basic idea of our approach is to map the input sequence to a fixed-sized vector, more precisely, some semantic representation of this input, and then unroll this representation in the target sequence using a neural network model. This intuition is formally captured in an encoder-decoder architecture. The output sequence is not a tagging sequence with one-to-one matching like in Part-of-Speech tagging task. It is the sequence of medical concepts corresponding to input sequence semantics. In fact, this architecture is aimed to solve multi-label classification problem, since output sequence could be interpreted as a set of labels for a sample input sequence.

3.1 Recurrent Neural Networks

RNNs are naturally used for sequence learning, where both input and output are word and label sequences, respectively. RNN has recurrent hidden states, which aim to simulate memory, i.e., the activation of a hidden state at every time step depends on the previous hidden state [25]. The recurrent unit computes a weighted sum of the input signal. There is the difficulty of training RNNs to capture long-term dependencies due to the effect of vanishing gradients [26], so

Table 3. Examples of raw texts from death certificates with medical concepts and ICD codes.

#	**Sample**	**Medical Concept Code**
1	CKD STAGE III, CHF, SEVERE OSTEOPOROSIS	
		Chronic kidney disease, stage 3 N183 Congestive ventricular heart failure I500 Osteoporosis M819
2	CAD / s/p CABG / Volume overload	
		Acute coronary artery disease I251 Fluid overload E877
3	F.T.T.	
		Failure to thrive syndrome R628
4	Neutropenic fever, pneumonia	
		Chronic Neutropenia D70 Fever R509 Pneumonia J189

the most widely used modifications of a RNN unit are the Long Short-Term Memory (LSTM) [27] and the Gated Recurrent Unit (GRU) [28].

An important modification of the basic RNN architecture is bidirectional RNNs, where the past and the future context is available in every time step [29]. Bidirectional LSTMs, developed by Graves and Schmidhuber [30,31], contain two chains of LSTM cells flowing in both forward and backward direction, and the final representation is either a linear combination or simply concatenation of their states.

3.2 Encoder-Decoder Model

As shown in Fig. 1, the model consists of two components based on RNNs: an encoder and a decoder. The encoder processes the input sequence, while the decoder generates the output sequence.

We adopted the architecture as described in [4,28]. The input layer of our model is vector representations of individual words. Word embedding models represent each word using a single real-valued vector. Such representation groups together words that are semantically and syntactically similar [32].

In order to incorporate prior knowledge, we additionally concatenated cosine similarity vector to the encoded state using on tf-idf representation. CLEF participants were provided with a manually created dictionary. The tf-idf score of a word, as defined by Salton and Buckley [33], is a reasonable measure of word importance. This score privileges the words that not only mention frequently in a given document, but also appear rarely in other documents of a corpus.

Cosine similarity vector was calculated as follows. First, for each ICD-10 code present in the dictionary, we construct a document by simply concatenating diagnosis texts belonging to that code. For the resulting document set, the

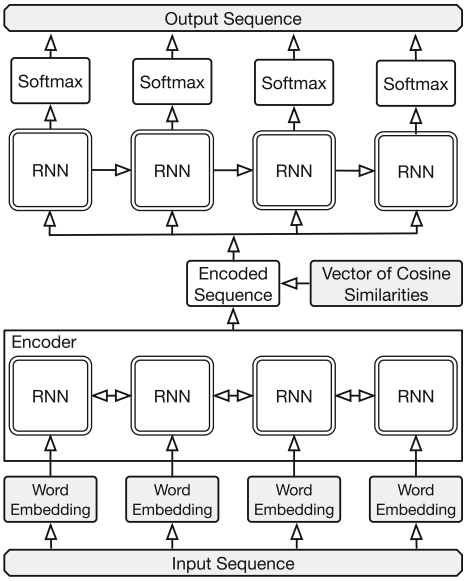

Fig. 1. An illustration of the encoder-decoder architecture.

tf-idf transformation was computed; thus, every ICD-10 code was provided with a vector representation. For a given input sequence, the tf-idf vector representation was calculated. Using the vector representation of the input sequence and each ICD-10 code, the vector of cosine similarities was constructed such as follows: the i-th position of vector is the cosine distance between input sequence representation and i-th ICD code representation. We have made the implementation of our model available at the github repository[1]. We consider pairs (diagnosis text, ICD1) from the dictionary for our system since most entries in the dictionary are associated with these codes.

Neural networks require word representations as inputs. We investigate the use of several different pre-trained word embeddings. We utilize word embeddings named *HealthVec*: publicly available 200-dimensional embeddings that were trained on 2,607,505 unlabeled user comments from health information websites using the Continuous Bag-of-Words model in [34]. We adopt 300-dimensional embeddings trained on the French version of Wikipedia using fasttext [35]. We also experiment with another published 200-dimensional embeddings named *PubMedVec*, which were trained on biomedical literature indexed in PubMed [36].

4 Experiments

In this section, we discuss the performance of neural networks.

[1] https://github.com/dartrevan/clef_2017.

Table 4. ICD-10 coding performance of the encoder-decoder model on the CDC test set of English texts (left) and the CépiDc test set of French texts (right).

	P	R	F
Encoder-decoder LSTM	.907	**.817**	**.860**
Official results from [24]			
KFU-run1 (ours)	.893	.811	.850
TUC-MI-run1	**.940**	.725	.819
SIBM-run1	.839	.783	.810
WBI-run1	.616	.606	.611
LIRMM-run1	.691	.514	.589
Average score	.670	.582	.622
Median score	.646	.606	.611
Non-off. from [24]			
LIMSI-run2	.899	.801	.847

	P	R	F
Encoder-decoder LSTM	.848	.673	.750
Official results from [24]			
SIBM-run1	.857	.689	.764
LITL-run2	.666	.414	.510
LIRMM-run1	.541	.480	.509
LIRMM-run2	.540	.480	.508
Average score	.475	.358	.406
Median score	.541	.414	.508
Non-off. from [24]			
LIMSI-run2	.872	**.784**	**.825**
TUC-MI-run1	**.883**	.539	.669

4.1 Settings

To find optimal neural network configuration and word embeddings, the 5-fold cross-validation procedure was applied to the training CDC set. We compared architectures with different numbers of neurons in hidden layers of the encoder and the decoder. The best cross-validation F-score was obtained for the architecture with 600 neurons in the hidden layer of the encoder and 1000 neurons in the hidden layer of the decoder. We tested bidirectional LSTM as decoder but did not achieve an improvement over the left-to-right LSTM. Additionally, we utilized the encoder with attention mechanism but did not achieve an improvement on the validation set. We also established that 10 were enough for stable performance on the validation sets. In contrast with our previous model [4], we set the decoder to predict ICD codes from the training set, not all codes from the dictionary. We adopted the train and validation sets of the CépiDc corpus for training.

We have implemented networks with the Keras library [37]. LSTM is trained on top of the embedding layer. We used the 600-dimensional hidden layer for the encoder RNN chain. Finally, the last hidden state of LSTM chain output concatenated with cosine similarities vector was fed into a decoding LSTM layer with 1000-dimensional hidden layer and softmax activation. In order to prevent neural networks from overfitting, we used dropout of 0.5 [38]. We used categorical cross entropy as the objective function and the Adam optimizer [39] with the batch size of 20.

4.2 Results

Our neural models were evaluated on texts in English using evaluation metrics of task 1 such as precision (P), recall (R) and balanced F-measure (F).

Table 4 presents results of the LSTM-based encoder-decoder model trained with PubMedVec and several official results of participants' methods (TUC-

Table 5. Performance of the encoder-decoder model on the CDC test sets.

Networks' settings			= 1 code			≥2 codes			Full set		
encoder	decoder	emb	P	R	F	P	R	F	P	R	F
biLSTM	LSTM	random, 100 d.	.935	.899	.916	**.837**	.605	.702	.908	.813	.858
biLSTM	LSTM	random, 200 d.	.934	.900	.917	**.837**	.603	.701	.903	.816	.857
biLSTM	LSTM	random, 300 d.	.932	.899	.915	.827	.606	.699	.904	.814	.857
biLSTM	LSTM	HealthVec	.932	.899	.915	.813	.601	.691	.902	.814	.856
biLSTM	LSTM	PubMedVec	**.937**	**.904**	**.920**	.803	.623	.702	**.907**	.817	.860
biGRU	LSTM	PubMedVec	.931	.901	.916	.829	**.631**	**.717**	.904	**.823**	**.861**
biGRU	GRU	PubMedVec	.927	.896	.912	.800	.627	.703	.892	.819	.854

Table 6. Performance of the encoder-decoder model on the CépiDc full sets.

Networks' settings				= 1 code			≥2 codes			Full set		
encoder	decoder	emb	sim.	P	R	F	P	R	F	P	R	F
biLSTM	LSTM	random, 100 d.	no	.868	.721	.787	.799	.340	.477	.832	.658	.735
biLSTM	LSTM	HealthVec	no	.874	.725	.793	.799	.340	.477	.836	.660	.737
biLSTM	LSTM	PubMedVec	no	.876	.728	.795	.806	.350	.488	.838	.669	.744
biLSTM	LSTM	PubMedVec	yes	.877	.728	.796	.815	.350	.490	.847	.673	.750
biLSTM	LSTM	French Wiki	no	**.879**	**.730**	**.798**	.815	**.355**	**.495**	.845	**.677**	**.752**
biLSTM	LSTM	French Wiki	yes	.874	.723	.792	**.821**	.350	.491	.848	.673	.750

MI, SIBM, LIMSI teams, etc.) which did not resort to RNNs [19,20,24]. On the CDC test set, LSTM-based encoder-decoder model obtained F-measure of 86.0% with significant improvement as compared to other methods. The neural network obtained comparable results with the LIMSI team that combined SVM with the dictionary for multi-label classification and submitted unofficial runs due to conflict of interest. On the CépiDc test set, our neural network obtained F-measure of 75.2% (without additional knowledge) and 75.0% (with similarity vector) which is comparable results with SIBM team (F-measure of 76.4%).

The experiments with neural networks are presented in Tables 5 and 6. Each dataset was divided into two parts: the one part contains records with only one corresponding label, so we may consider this task to be single-label classification; the other part contains records with two or more corresponding labels which makes it multi-label classification task. The full dataset is also considered as multi-labeled.

Table 5 presents results for the English dataset. The best achieved F-measure on single-label classification task is 92% for biLSTM with PubMedVec. For two multi-label classification tasks (on the second part of the dataset and on the full dataset) the best model was biGRU with PubMedVec achieving 72% and 86% of F-measure, respectively. The second result is the best among all the participants of this challenge. Interestingly the best precision on the experiment with second

class only is achieved by systems using random vectors. Overall the quality of underlying vectors has limited influence on system performance.

Table 6 presents results for the French dataset. These results are comparable with approaches presented by challenge participants, but our solution does not use large vocabulary as additional input. The lowered system performance in comparison with English dataset could be explained by two main reasons: (I) the large number of Out-of-Vocabulary (OOV) words (app. 64% words of the vocabulary) for French language which were not associated with embeddings, (II) we did not perform language-dependent pre-processing steps including diacritic removal and correction of some spelling errors (as in the LIMSI's system), and (III) unlike the CDC dataset, the CépiDc train and test sets have records from different years, so the results could be influenced by changes in ICD-10 itself. Interestingly, the vectors for the English language actually improve system's performance, which can be explained by the significant percentage of French loan words in English language and consequently vocabulary sharing between these two datasets.

5 Conclusion

In this paper, we introduce a neural network architecture with a specific application to medical concept coding, i.e. linking the free-form language of clinical records to particular entries in the International Classification of Diseases. We find that by combining the encoder-decoder framework with cosine similarity metrics and a traditional tf-idf weighting scheme, we achieve the state-of-the-art results on the CDC corpus of English texts. Although we focus on ICD-10 coding of death certificates, our model is extensible without any task-specific manual feature engineering effort to other multi-label document tagging tasks, including prediction of diagnoses and procedures.

We foresee three directions for future work. First, we plan to carry out experiments on other datasets for medical code prediction including both MIMIC-II and MIMIC-III datasets. Second, we believe attention should be given to infrequent codes since ICD-10-CM has more than 70,000 codes. From the system perspective, future research might focus on embedding code descriptions and ICD hierarchy to a latent space. If we can better incorporate prior knowledge about codes, we may be able to infer rare medical events. From the medical side, future work might focus on applying our automatic coding model to find misclassification in clinical records coded manually. The third promising direction for research is to investigate multilingual models on datasets provided by CLEF eHealth 2017 and 2018 challenges.

Acknowledgements. This work was supported by the Russian Science Foundation grant no. 18-11-00284. The authors are grateful to Prof. Alexander Tropsha and Valentin Malykh for useful discussions about this study.

References

1. Pradhan, S., Elhadad, N., Chapman, W.W., Manandhar, S., Savova, G.: SemEval-2014 task 7: analysis of clinical text. In: SemEval@ COLING, pp. 54–62 (2014)
2. Dougherty, M., Seabold, S., White, S.E.: Study reveals hard facts on CAC. J. AHIMA **84**(7), 54–56 (2013)
3. Stanfill, M.H., Williams, M., Fenton, S.H., Jenders, R.A., Hersh, W.R.: A systematic literature review of automated clinical coding and classification systems. J. Am. Med. Inform. Assoc. **17**(6), 646–651 (2010)
4. Miftahutdinov, Z., Tutubalina, E.: KFU at CLEF ehealth 2017 task 1: ICD-10 coding of English death certificates with recurrent neural networks. In: CEUR Workshop Proceedings, vol. 1866 (2017)
5. Karimi, S., Dai, X., Hassanzadeh, H., Nguyen, A.: Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In: BioNLP 2017, pp. 328–332 (2017)
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. J. Biomed. Inform. **80**, 64–77 (2018)
7. Zhang, Y., et al.: Uth_CCB: a report for SemEval 2014-task 7 analysis of clinical text. In: SemEval 2014, p. 802 (2014)
8. Ghiasvand, O., Kate, R.J.: UWM: disorder mention extraction from clinical text using CRFS and normalization using learned edit distance patterns. In: SemEval@ COLING, pp. 828–832 (2014)
9. Van Mulligen, E., Afzal, Z., Akhondi, S.A., Vo, D., Kors, J.A.: Erasmus MC at CLEF eHealth 2016: concept recognition and coding in French texts. In: CLEF (2016)
10. Cabot, C., Soualmia, L.F., Dahamna, B., Darmoni, S.J.: SIBM at CLEF eHealth evaluation lab 2016: extracting concepts in French medical yexts with ECMT and CIMIND. In: CLEF (2016)
11. Mottin, L., Gobeill, J., Mottaz, A., Pasche, E., Gaudinat, A., Ruch, P.: BiTeM at CLEF eHealth evaluation lab 2016 task 2: multilingual information extraction. In: CEUR Workshop Proceedings, vol. 1609, pp. 94–102 (2016)
12. Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., Taright, N.: ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. In: CLEF (2016)
13. Zweigenbaum, P., Lavergne, T.: LIMSI ICD10 coding experiments on CépiDC death certificate statements. In: CLEF (2016)
14. Leaman, R., Khare, R., Lu, Z.: NCBI at 2013 shARe/CLEF ehealth shared task: disorder normalization in clinical notes with DNorm. Radiology 42(21.1), 1–941 (2011)
15. Suominen, H., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 212–231. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_24
16. Leaman, R., Islamaj Doğan, R., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics **29**(22), 2909–2917 (2013)
17. Névéol, A., et al.: CLEF ehealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2017)

18. Névéol, A., et al.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2016 (2016)
19. Zweigenbaum, P., Lavergne, T.: Hybrid methods for ICD-10 coding of death certificates. In: EMNLP 2016, p. 96 (2016)
20. Cabot, C., Soualmia, L.F., Darmoni, S.J.: SIBM at CLEF ehealth evaluation lab 2017: multilingual information extraction with CIM-IND. In: CLEF (2017)
21. Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., Malykh, V.: Medical concept normalization in social media posts with recurrent neural networks. J. Biomed. Inform. **84**, 93–102 (2018)
22. Rios, A., Kavuluru, R.: EMR coding with semi-parametric multi-head matching networks. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), vol. 1, pp. 2081–2091 (2018)
23. Schuemie, M.J., Kors, J.A., Mons, B.: Word sense disambiguation in the biomedical domain: an overview. J. Comput. Biol. **12**(5), 554–565 (2005)
24. Névéol, A., et al.: CLEF eHealth 2017 Multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings (2017)
25. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
26. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
27. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2016)
28. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
29. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Proc. **45**(11), 2673–2681 (1997)
30. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 799–804. Springer, Heidelberg (2005). https://doi.org/10.1007/11550907_126
31. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings of IEEE International Joint Conference on Neural Networks, IJCNN 2005, vol. 4, pp. 2047–2052. IEEE (2005)
32. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
33. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Proc. Manag. **24**(5), 513–523 (1988)
34. Miftahutdinov, Z., Tutubalina, E., Tropsha, A.: Identifying disease-related expressions in reviews using conditional random fields. In: Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog, vol. 1, pp. 155–167 (2017)
35. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
36. Moen, S., Ananiadou, T.S.S.: Distributional semantics resources for biomedical text processing (2013)

37. Chollet, F., et al.: Keras (2015). https://github.com/fchollet/keras
38. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
39. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)