KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks

Zulfat Miftahutdinov and Elena Tutubalina

Kazan (Volga Region) Federal University, Kazan, Russia
zulfatmi@gmail.com, ElVTutubalina@kpfu.ru

Abstract. This paper describes the participation of the KFU team in the CLEF eHealth 2017 challenge. Specifically, we participated in Task 1, namely "Multilingual Information Extraction - ICD-10 coding" for which we implemented recurrent neural networks to automatically assign ICD-10 codes to fragments of death certificates written in English. Our system uses Long Short-Term Memory (LSTM) to map the input sequence into a vector representation, and then another LSTM to decode the target sequence from the vector. We initialize the input representations with word embeddings trained on user posts in social media. The encoder-decoder model obtained F-measure of 85.01% on a full test set with significant improvement as compared to the average score of 62.2% for all participants' approaches. We also obtained significant improvement from 26.1% to 44.33% on an external test set as compared to the average score of the submitted runs.

Keywords: ICD-10 coding, ICD-10 codes, medical concept coding, recurrent neural network, sequence to sequence, sequence-to-sequence architecture, encoder-decoder model, deep learning, machine learning, death certificates, CepiDC, healthcare, CLEF eHealth

1 Introduction

Extracting and linking medical information from textual documents has attracted extensive interest from both academia and industry. Automatic matching of text phrases to medical concepts and corresponding classification codes is a highly important task for many clinical applications in the fields of health management and patient safety.

The International Classification of Diseases (ICD) is the diagnostic system that is used to monitor and classify causes of health problems and death and provide information for clinical purposes. Each medical concept is mapped onto a unique identifier which consists of a single alphabet prefix and several digits. Single alphabet prefix represents a class of common diseases (e.g. "J" covers diseases of the respiratory system, "V" covers external causes of morbidity) and digits represent specific type of disease (e.g. "J20.2" covers acute bronchitis

Table 1. Examples of causes of death from the international classification of diseases.

J189	Pneumonia, unspecified organism
I48	Atrial fibrillation and flutter
M726	Necrotizing fasciitis
G20	Parkinson's disease
F102	Alcohol dependence
A419	Sepsis, unspecified organism
D696	Thrombocytopenia, unspecified
E119	Type 2 diabetes mellitus without complications
V892	Person injured in unspecified motor-vehicle accident, traffic

due to streptococcus", "V25" covers "Motorcycle rider injured in collision with railway train or railway vehicle"). Table 1 contains examples of ICD-10 codes.

Machine learning methods have been widely successful in various NLP applications including named entity recognition and relation extraction [1–3], machine translation [4–6], opinion mining [7–9], detection of demographic information from health-related user posts [10, 11]. Recurrent Neural Networks (RNN), in particular, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are considered to be among the most powerful methods for sequence modeling [12–14, 4]. Motivated by the recent success of deep recurrent networks, herein we have explored an application of RNN-based encoder-decoder models to the task of automated ICD coding.

We describe participation of our team in the task 1 for English death certificates. The goal of this task is to assign one or more relevant ICD-10 codes to sentences in the death certificates. We employ an annotated corpus named the CepiDC Causes of Death Corpus, which contains free-text descriptions of causes of death reported by physicians. More specifically, we employ the part of the corpus with English texts. The CepiDC corpus of French texts was initially provided for the task of ICD-10 coding in CLEF eHealth 2016 (task 2) [15, 16]. The organizers recently extended this corpus with additional data for CLEF eHealth 2017 [17, 18]. Our neural network relies on two sources of information: word representations learned from unannotated corpora and a manually curated ICD-10 dictionary provided by the organizers of the task.

The rest of the paper is structured as follows. Section 2 contains our system description, Section 3 provides evaluation results. In Section 4, we discuss some related work from CLEF eHealth 2016. Finally, Section 5 provides concluding remarks.

2 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 a encoder-decoder architecture. In the following subsections, we provide a brief description of recurrent neural networks (RNNs) and the encoder-decoder model.

2.1 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 [12]. 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 [19], so the most widely used modification of a RNN unit is the Long Short-Term Memory (LSTM) [20]. LSTM provides the "constant error carousel" and does not preclude free gradient flow. The basic LSTM architecture contains three gates: input gate, forget gate, and output gate, together with a recurrent cell. LSTM cells are usually organized in a chain, with outputs of previous LSTMs connected to the inputs of subsequent LSTMs.

An important modification of the basic RNN architecture is bidirectional RNNs, where the past and the future context is available in every time step [13]. Bidirectional LSTMs, developed by Graves and Schmidhuber [14, 21], 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.

2.2 Encoder-Decoder Model

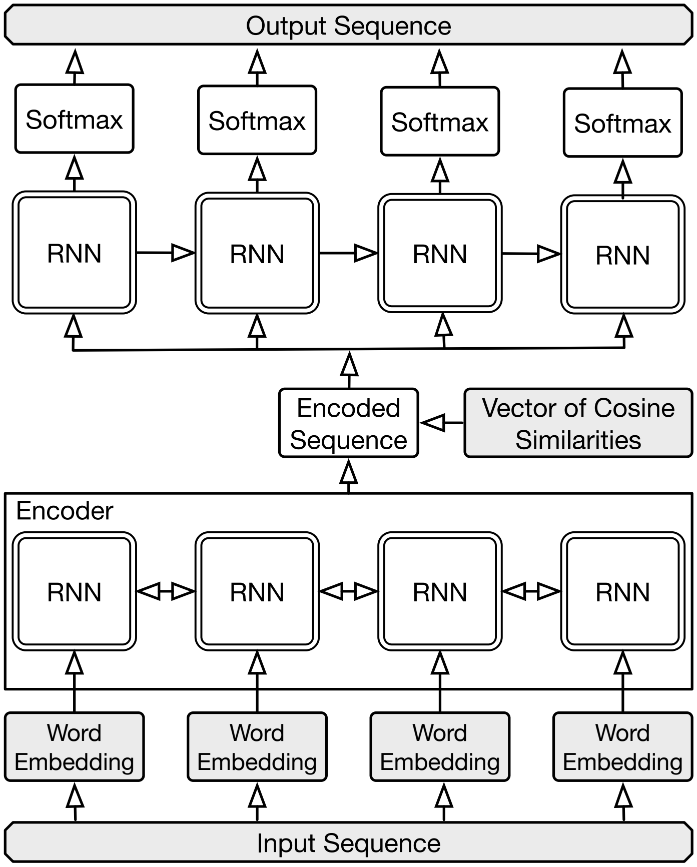
We introduce the sequence-to-sequence architecture, more precisely, an encoder-decoder model proposed earlier [4, 6] for the ICD-10 coding task. As shown in Figure 1, the model consists of two components based on RNNs: an encoder and a decoder. The encoder processes the input sequence, while the decoder generates the output sequence.

We adopted the architecture as described in [4]. As encoder RNN we used bidirectional LSTM, as decoder RNN we used left-to-right LSTM. 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 [22]. The word embeddings are trained using an unlabelled corpus of user reviews.

In order to incorporate prior knowledge, we additionally concatenated cosine similarities vector to the encoded state. CLEF participants were provided with a manually created dictionary. This dictionary named AmericanDictionary contains quadruplets (diagnosis text, codes Icd1, IcdC, Icd2). We only consider pairs (diagnosis text, Icd1) for our system since most entries in the dictionary are associated with these codes.

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

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, vector of cosine similarities was constructed such as to have in the i-th position the cosine similarity measure between input sequence representation and i-th ICD-10 code representation.

We have made the implementation of our model available at the github repository[1].

3 Experiments

In this section, we discuss the performance of our LSTM-based encoder-decoder model for ICD coding.

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

3.1 Evaluation Dataset

The CLEF e-Health 2017 Task 1 participants were provided with data from 13,330 death certificates for training. Each certificate contains information about the demographic attributes of each person (gender, age), other metadata (e.g., a location of death) and one or more codes of the primary cause of death. Diagnostic statements with multiple codes were repeated for each code assigned by physicians. The test set contained 14,833 certificates.

The experiments were also carried out on the following sets:

1. The full version of the CepiDC test set named the “ALL” set.
2. The part of the full test set named the “EXTERNAL” set.

The “ALL” test set consists of texts associated with all ICD codes. The “EXTERNAL” test set is limited to textual fragments with ICD codes linked with a particular type of deaths, called “external causes” or violent deaths. The ”EXTERNAL” set was selected due to two reasons: (i) there is a special interest for the public health policies that can target ICD codes specifically, e.g. suicide prevention; (ii) the semantic analysis of the context associated with these deaths is more complex in terms of comorbidity, affected people and language models used to describe the event. External causes are characterized by codes V01 to Y98. Please refer to the task overview paper [18] for more details.

3.2 Experimental Setting

Word embeddings We used the word embeddings trained on 2,5 millions of health-related reviews from [1]. Statistics of there reviews is presented in Table 2. The embeddings were trained with the Continuous Bag of Words model with the following parameters: vector size of 200, the length of local context of 10, negative sampling of 5, vocabulary cutoff of 10.

Table 2. Summary of statistics of data sources.

Data source	**# reviews**	**# tokens**	**# uniq. tokens**	**avg. len**
webmd.com	284 055	20 794 273	103 935	73.21
askapatient.com	113 836	13 649 150	79 036	119.90
patient.info	1 472 273	160 750 980	720 380	109.19
dailystrength.org	214 489	13 880 025	76 384	64.72
drugs.com	93 845	9 191 434	51 530	97.42
amazon health reviews	428 777	36 499 681	135 523	85.13

Model tuning To find optimal neural network configuration and word embeddings, the 5-fold cross-validation procedure was applied to the training set. We compared architectures with different numbers of neurons in hidden layersof encoder and decoder LSTM. The best cross-validation F-score is obtained for the architecture with 600 neurons in the hidden layer of encoder LSTM and 1000 neurons in the hidden layer of the decoder LSTM. We tested bidirectional LSTM as decoder but did not achieve an improvement over the left-to-right LSTM. We also established that 10 epochs are enough for stable performance on the validation sets.

We have implemented networks with the Keras library [23]. LSTM is trained on top of the embedding layer. We use the 600-dimensional hidden layer for the encoder RNN chain. Finally, the last hidden state of LSTM chain output concatenated with cosine similarities vector is 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 [24]. We used categorical cross entropy as the objective function and the Adam optimizer [25] with the batch size of 20.

In addition, we have evaluated word embeddings trained on biomedical literature indexed in PubMed from [26] as well as on health-related reviews from [1]. Embeddings on health-related reviews showed better results during cross-validation. We also tried to exploit meta-information along with cosine similarities vectors but we did not observe any significant improvement.

3.3 Results

Our neural models were evaluated on texts in English using common evaluation metrics such as precision, recall and balanced F-measure. We trained our model for 10 epochs (Run1) and 15 epochs (Run2). The reported results are presented in Tables 3 and 4.

Table 3. ICD-10 coding performance on the "ALL" test set.

	Precision	**Recall**	**F-measure**
Run1	0.893	0.811	0.850
Run2	0.891	0.812	0.850
Average score	0.670	0.582	0.622
Median score	0.646	0.606	0.611

Table 4. ICD-10 coding performance on the "EXTERNAL" test set.

	Precision	**Recall**	**F-measure**
Run1	0.584	0.357	0.443
Run2	0.631	0.325	0.429
Average score	0.405	0.267	0.261
Median score	0.279	0.262	0.274

As shown in Tables 3 and 4 our performance results are significantly better than the average and median score of all submitted runs. The system obtained

F-scores of 85.01% and 44.33% on the full test set and the “EXTERNAL” set, respectively. The difference of results on these sets is explained by a small number of codes in the latter case. The “ALL” set includes 18,928 codes (900 unique codes), while the “EXTERNAL” set includes only 126 codes (28 unique codes). We note that RNNs and word embeddings can be successfully applied to medical concept coding tasks without any task-specific feature engineering effort.

4 Related Work

Different approaches have been developed for ICD coding task, mainly falling into two categories: (i) knowledge-based methods [27–29]; and (ii) machine learning approaches [30, 31].

In the CLEF eHealth 2016, five teams participated in the shared task 2 about the ICD-10 coding of death certificates in French [15]. Most methods utilized dictionary-based semantic similarity and, to some extent, string matching. Mulligen et al. [27] obtained the best results 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 F-measure of 84.8%. Cabot et al. [28] applied an approximate string matching method and obtained F-measure of 68.0%. Mottin et al. [29] used a pattern matching approach and obtained F-measure of 55.4%. Dermouche et al. [30] applied two machine learning methods: (i) a supervised extension of Latent Dirichlet Allocation (LDA), i.e., Labeled-LDA and (ii) Support Vector Machine (SVM) based on bag-of-word features. For Labeled-LDA, they used ICD-10 codes from the training set as documents classes. The Labeled-LDA and SVM classifier archived F-measures of 73.53% and 75.19%, respectively. This study did not focus on designing effective features to obtain better classification performance. Zweigenbaum and Lavergne [31] proposed a classifier with TF-IDF transformer for tokens and used cosine similarity for ranking of classification codes. They studied the problem of learning to accurately rank a set of candidate codes obtained as a result of classification. The authors explored the effectiveness of several groups of features including meta-information and n-grams of normalized tokens. They focused only on statements which are associated with a singular code. The proposed approach obtained F-measure of 65.2% due to low recall of 56.8%. In recent work [32], Zweigenbaum and Lavergne utilized a hybrid method combining 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%. Overall, the participants of task 2 did not use word embeddings or deep neural networks, which are proved useful in many natural language processing tasks.

Besides experiments on CLEF eHealth data sets, the medical concept coding task has also been studied by several researchers. Ontologies of medical concepts such as the Unified Medical Language System (UMLS) [33], SNOMED CT [34], and ICD-9 or ICD-10 are widely used for this task. In order to map texts to medical concepts in the UMLS, the National Library of Medicine (NLM) developed

MetaMap [35]. This system is based on a linguistic approach using variants of terms and rules. Recent studies applied machine learning methods such as learning-to-rank methods [36] and convolutional neural networks [37]. Leaman et al. introduced a DNorm system based on pairwise learning-to-rank technique with a predefined set of features [36]. Features were based on a dictionary of diseases derived from the UMLS Metathesaurus. Recently, Limsopatham and Collier [37] experimented with convolutional and recurrent neural networks with pre-trained word embeddings for mapping social media texts to medical concepts. The authors observed that training can be effectively achieved at 40-70 epochs for corpora of tweets and user reviews. Experiments showed that both neural networks outperformed the DNorm system and a multi-class logistic regression. Word embeddings trained on a Google News corpus improved significantly over embeddings on medical articles downloaded from BioMed Central. In [1], using word embeddings trained on social media produces better scores than using embeddings trained on PubMed articles for disease named entity recognition. We also mark word embeddings trained on electronic health records [38–40] for future work.

5 Conclusion

In this paper, we have developed RNN-based encoder-decoder models for ICD-10 coding on Task 1 of the 2017 CLEF eHealth evaluation lab. Our results show that the neural network performs significantly better than the official median and average computed using the participants' runs, reaching F-measure of 85.01% on the full test set. In further studies, we plan to implement other encoder-decoder architectures and convolutional neural networks. We also plan to carry out a qualitative analysis on the extracted codes. Additionally, we would like to explore alternative distributed word representations trained on medical notes from electronic health records.

Acknowledgements

This work was supported by the Russian Science Foundation grant no. 15-11-10019.

References

1. 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. Volume 1. (2017) 155–167
2. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In: EMNLP. (2015) 1753–1762

3. Solovyev, V., Ivanov, V.: Knowledge-driven event extraction in Russian: corpus-based linguistic resources. Computational intelligence and neuroscience **2016** (2016) 16
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
5. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
6. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112
7. Dos Santos, C.N., Gatti, M.: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In: COLING. (2014) 69–78
8. Liu, P., Joty, S.R., Meng, H.M.: Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In: EMNLP. (2015) 1433–1443
9. Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., Jaggi, M.: Swiss-Cheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. Proceedings of SemEval (2016) 1124–1128
10. Tutubalina, E., Nikolenko, S.: Automated Prediction of Demographic Information from Medical User Reviews. In: International Conference on Mining Intelligence and Knowledge Exploration, Springer, Cham (2016) 174–184
11. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data, EACL (2017)
12. Elman, J.L.: Finding structure in time. Cognitive science **14**(2) (1990) 179–211
13. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11) (1997) 2673–2681
14. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005 (2005) 753–753
15. Névéol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., 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)
16. Lavergne, T., Névéol, A., Robert, A., Grouin, C., Rey, G., Zweigenbaum, P.: A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage. BioTxtM 2016 (2016) 60
17. Goeuriot, L., Kelly, L., Suominen, H., Nvol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth Evaluation Lab Overview. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2017)
18. Nvol, 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. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings. (2017)
19. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks **5**(2) (1994) 157–166

20. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems (2016)
21. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. Volume 4., IEEE (2005) 2047–2052
22. 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. (2013) 3111–3119
23. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
24. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1) (2014) 1929–1958
25. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). (2015)
26. Moen, S., Ananiadou, T.S.S.: Distributional semantics resources for biomedical text processing (2013)
27. 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, CLEF (2016)
28. Cabot, C., Soualmia, L.F., Dahamna, B., Darmoni, S.J.: SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND, CLEF (2016)
29. Mottin, L., Gobeill, J., Mottaz, A., Pasche, E., Gaudinat, A., Ruch, P.: BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction
30. Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., Taright, N.: ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates, CLEF (2016)
31. Zweigenbaum, P., Lavergne, T.: LIMSI ICD10 coding experiments on CépiDC death certificate statements, CLEF (2016)
32. Zweigenbaum, P., Lavergne, T.: Hybrid methods for icd-10 coding of death certificates. EMNLP 2016 (2016) 96
33. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(suppl_1) (2004) D267–D270
34. Spackman, K.A., Campbell, K.E., Côté, R.A.: SNOMED RT: a reference terminology for health care. In: Proceedings of the AMIA annual fall symposium, American Medical Informatics Association (1997) 640
35. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (2001) 17
36. Leaman, R., Islamaj Doğan, R., Lu, Z.: DNorm: disease name normalization with pairwise learning to rank. Bioinformatics **29**(22) (2013) 2909–2917
37. Limsopatham, N., Collier, N.: Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In: ACL. (2016)
38. Grnarova, P., Schmidt, F., Hyland, S.L., Eickhoff, C.: Neural Document Embeddings for Intensive Care Patient Mortality Prediction. arXiv preprint arXiv:1612.00467 (2016)
39. Fries, J.A., Center, M.: Brundlefly at SemEval-2016 Task 12: Recurrent Neural Networks vs. Joint Inference for Clinical Temporal Information Extraction. Proceedings of SemEval (2016) 1274–1279

40. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association **24**(3) (2017) 596–606