An Encoder-Decoder Model for ICD-10 Coding of Death Certificates

Elena Tutubalina*
Kazan Federal University
Kazan, Russia 420008
elvtutubalina@kpfu.ru

Zulfat Miftahutdinov
Kazan Federal University
Kazan, Russia 420008
zulfatmi@gmail.com

Abstract

Information extraction from textual documents such as hospital records and health-related user discussions has become a topic of intense interest. The task of medical concept coding is to map a variable length text to medical concepts and corresponding classification codes in some external system or ontology. In this work, we utilize recurrent neural networks to automatically assign ICD-10 codes to fragments of death certificates written in English. We develop end-to-end neural architectures directly tailored to the task, including basic encoder-decoder architecture for statistical translation. In order to incorporate prior knowledge, we concatenate cosine similarities vector among the text and dictionary entry to the encoded state. Being applied to a standard benchmark from CLEF eHealth 2017 challenge, our model achieved F-measure of 85.01% on a full test set with significant improvement as compared to the average score of 62.2% for all official participants' approaches.

1 Introduction

Recent years have seen many new applications of Natural Language Processing (NLP) to biomedical information. Much of this work has been focused on a central task of information extraction, that is named entity recognition from the scientific literature or electronic health records (EHRs). The task of medical concept normalization is highly important for many clinical applications in the fields of health management and patient safety.

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-9, ICD-10). In particular, each medical concept in ICD 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 due to Streptococcus", "V25" covers "motorcycle rider injured in collision with railway train or railway vehicle").

In this work, we view ICD-10 coding as a sequence learning task. A sequence of codes is generated from a natural language text from medical notes by preserving the semantics of the text as much as possible. Motivated by the recent success of recurrent neural networks (RNNs), this work adopts RNN with an encoder-decoder architecture. For evaluation, we adopt a CDC corpus provided for the task of ICD-10 coding in CLEF eHealth 2017. This corpus contains free-text descriptions of causes of death in English reported by physicians. Table 1 contains examples of descriptions. There are several major challenges which information extraction methods face: (i) lexical, morphological, and syntactic variants; (ii) paraphrases, synonyms; (iii) abbreviations, ambiguity; (iv) misspellings and shortened forms of words.

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

NIPS 2017 Machine Learning for Health Workshop (NIPS ML4H 2017), Long Beach, CA, USA.

Table 1: Examples of raw texts from depth 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	A.FIB., D.M. TYPE II		
		Atrial fibrillation	I48
		Type 2 diabetes mellitus	E119
3	CAD / s/p CABG / Volume overload		
		Acute coronary artery disease	I251
		Fluid overload	E877
4	P.V.D.		
		Peripheral vascular disease	I739

We utilize 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. The 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. This work is an extended version of the conference paper [1].

2 Background

There exist many applications where a system needs to mediate between natural language expressions and elements of a vocabulary in an ontology. Huang and Lu [2] gave an overview of the work done in the organization of biomedical NLP (BioNLP) challenge evaluations up to 2014. We briefly give an overview of the major findings in previous research on terminology association. Many BioNLP evaluations have also focused on named entity recognition (NER) of disease names in clinical notes (e.g., ShARe/CLEF eHealth lab, SemEval 2014 lab). Automatic approaches to BioNLP tasks roughly fall into two categories: (i) linguistic approaches based on dictionaries, association measures, morphological and syntactic properties of texts [3, 4, 5, 6, 7]; (ii) machine learning approaches [8, 9, 10, 11, 1, 12]. 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 [13]. For the CLEF eHealth 2016 lab, 5 teams participated in the shared task 2 about the ICD-10 coding of death certificates in French [14]. For the CLEF eHealth 2017 lab, 9 teams participated in the shared task 1 about the ICD-10 coding of death certificates in French and English [15]. Mulligen et al. [3] 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 F-measure of 84.8%. Mottin et al. [4] applied pattern matching approach and achieved the F-measure of 55.4%. Dermouche et al. [10] 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-words features. This study did not focus on designing effective features to obtain better classification performance. Zweigenbaum and Lavergne [16] utilized a hybrid method combining simple dictionary projection and mono-label supervised classification. They trained Linear SVM on the full training corpus and the 2012 dictionary provided for CLEF participants. This hybrid method obtained an F-measure of 85.86%. The TUC-MI team [12] utilized fusion methods in conjunction with support vector machines with a large scale feature set. The SIBM team [7] developed a dictionary-based approach and fuzzy matching methods. The LIMSI team [17] explored the combination of a dictionary-based method and SVM. Overall, most methods utilized dictionary-based semantic similarity and, to some extent, string matching.

3 Encoder-Decoder Model

The basic idea of our approach can be intuitively explained as follows: when we try to link a sentence to medical concepts, we do not really go word by word but rather first construct some semantic representation of this sentence and then unroll this representation in the target sequence using a neural

Table 2: Statistics of the CDC American Death Certificates Corpus from [15].

	Train	**Test**
Certificates	13,330	6,665
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

network model. For instance, the sequence "Neutropenic fever, pneumonia" is mapped to "D70 R509 J189". This intuition is formally captured in the encoder-decoder architecture. We adopted the architecture as described in [18].

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 [19]. An important modification of the basic RNN architecture is bidirectional RNNs. One of the most widely used such modifications of RNNs is called the *Long Short-Term Memory* (LSTM) [20]. Our system utilizes LSTM to map the input sequence into a vector representation, and then another LSTM to decode the target sequence from the vector. The primary goal of applying bidirectional encoder to ICD-10 coding is to capture "semantic representation" based on not only the past but also the future context on every time step. We utilize left-to-right LSTM as the decoder.

In order to incorporate prior knowledge, we additionally concatenated cosine similarities vector between the text and dictionary's entries 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 presented the ICD-10 code as a single document by concatenating diagnosis texts associated with this code. In order to provide a ICD-10 code and an input sequence with vector representations, we computed the TF-IDF transformation and calculated the cosine similarity between these vectors. We only consider pairs (diagnosis text, Icd1) for our system since most entries in the dictionary are associated with these codes.

4 Evaluation

The CLEF e-Health 2017 Task 1 participants were provided with data from 13,330 and 14,833 raw texts from death certificates for training and testing, respectively. The full 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.

We applied the word embeddings trained on 2,5 millions of health-related reviews from [21]. 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. Additionally, we applied word embeddings trained on biomedical literature indexed in PubMed[22] and a part of Google News dataset[2]. Statistics of the word embeddings are presented in Table 3. For out-of-vocabulary words with the pre-trained word model, we used representations randomly sampled. In order to find optimal neural network configuration and word embeddings, the five-fold cross-validation procedure was applied to the training set. Embedding layers are trainable for all networks. Table 4 shows the five-fold cross-validation results on the training dataset. It shows that all models with prior knowledge obtained better results. Models with different word embeddings obtained similar results.

We have implemented networks with the Keras library [23]. 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 applied dropout of 0.5

[2]https://code.google.com/archive/p/word2vec/

Table 3: Statistics of *word2vec* embeddings.

Embeddings	Dim.	#tokens	% of tokens (train data)	% tokens (test data)
HealthVec	200	73,644	68%	70%
PubmedVec	200	2,351,706	87%	88%
GoogleNewsVec	300	3,000,000	73%	76%

Table 4: Five-fold cross-validation results on the training dataset.

Encoder-decoder LSTM	Embeddings	P	R	F
with prior knowledge	HealthVec	.876	.811	.842
	PubmedVec	.881	.816	.847
	GoogleNewsVec	.879	.811	.843
without prior knowledge	HealthVec	.857	.802	.828
	PubmedVec	.842	.796	.819
	GoogleNewsVec	.844	.790	.816

[24]. We used categorical cross entropy as the objective function, HealthVec as input, and the Adam optimizer [25] with the batch size of 20. We trained our model for 10 epochs.

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). For comparison, we present our results and several official results of participants' methods (TUC-MI, SIBM teams, etc.) which did not resort to RNNs [12, 7, 15] in Table 5. Our encoder-decoder model obtained F-measure of 85.0% on a full test set with significant improvement as compared to the average score of 62.2% for all official CLEF participants' approaches that were based on machine learning or knowledge-based algorithms. Our model 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. The difference of results on two sets is explained by a small number of codes in the latter case.

Table 5: ICD-10 coding performance from [15] on the full test set (left) and the "external" test set (right).

	P	R	F
Official runs submitted			
Encoder-decoder LSTM	.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			
LIMSI	.899	.801	.847

	P	R	F
Official runs submitted			
Encoder-decoder LSTM	.584	.357	.443
TUC-MI-run1	.880	.175	.291
SIBM-run1	.426	.389	.407
WBI-run1	.246	.119	.160
LIRMM-run1	.232	.524	.322
Average score	.405	.267	.261
Median score	.279	.262	.274
Non-off			
LIMSI	.723	.373	.492

5 Conclusion

In this work, we have applied deep neural networks, in particular, LSTM-based encoder-decoder architecture, to the problem of ICD-10 coding. We have obtained very promising results, both quantitatively and qualitatively. We outline three directions for future work. First, the use of novel architectures and multilingual neural networks remains to be explored. We would like to explore alternative distributed word representations trained on medical notes from electronic health records. Second, a promising research direction is the integration of linguistic knowledge into the models. Third, future research might focus on developing extrinsic test sets for medical concept normalization.

References

[1] Zulfat Miftakhutdinov and Elena Tutubalina. KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks. *CEUR Workshop Proceedings*, 1866, 2017.

[2] Chung-Chi Huang and Zhiyong Lu. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144, 2015.

[3] E Van Mulligen, Zubair Afzal, Saber A Akhondi, Dang Vo, and Jan A Kors. Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts. CLEF, 2016.

[4] Luc Mottin, Julien Gobeill, Anaïs Mottaz, Emilie Pasche, Arnaud Gaudinat, and Patrick Ruch. BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction. In *CLEF (Working Notes)*, pages 94–102, 2016.

[5] Omid Ghiasvand and Rohit J Kate. UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. In *SemEval@ COLING*, pages 828–832, 2014.

[6] Yaoyun Zhang1 Jingqi Wang1 Buzhou Tang, Yonghui Wu1 Min Jiang, and Yukun Chen3 Hua Xu. UTH_CCB: a report for semeval 2014–task 7 analysis of clinical text. *SemEval 2014*, page 802, 2014.

[7] Chloé Cabot, Lina F Soualmia, and Stéfan J Darmoni. SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with CIM-IND. CLEF, 2017.

[8] Robert Leaman, Ritu Khare, and Zhiyong Lu. NCBI at 2013 ShARe/CLEF eHealth Shared Task: disorder normalization in clinical notes with DNorm. *Radiology*, 42(21.1):1–941, 2011.

[9] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[10] Mohammed Dermouche, Vincent Looten, Rémy Flicoteaux, Sylvie Chevret, Julien Velcin, and Namik Taright. ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. CLEF, 2016.

[11] Pierre Zweigenbaum and Thomas Lavergne. LIMSI ICD10 coding experiments on CépiDC death certificate statements. CLEF, 2016.

[12] Mike Ebersbach, Robert Herms, and Maximilian Eibl. Fusion Methods for ICD10 Code Classification of Death Certificates in Multilingual Corpora. CLEF, 2017.

[13] Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. CLEF 2017 eHealth Evaluation Lab Overview. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[14] Aurélie Névéol, Lorraine Goeuriot, Liadh Kelly, Kevin Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Grégoire Rey, Aude Robert, Xavier Tannier, 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.

[15] Aurélie Névéol, Robert N. Anderson, K. Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire 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. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2017.

[16] Pierre Zweigenbaum and Thomas Lavergne. Hybrid methods for icd-10 coding of death certificates. *EMNLP 2016*, page 96, 2016.

[17] Pierre Zweigenbaum and Thomas Lavergne. Multiple methods for multi-class, multi-label icd-10 coding of multi-granularity, multilingual death certificates. CLEF, 2017.

[18] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[19] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[20] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *CoRR*, abs/1503.04069, 2015.

[21] Z.Sh. Miftahutdinov, E.V. Tutubalina, and A.E. Tropsha. Identifying Disease-related Expressions in Reviews using Conditional Random Fields. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog*, volume 1, pages 155–167, 2017.

[22] Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of Languages in Biology and Medicine*, 2013.

[23] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[24] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[25] D Kinga and J Ba Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.