Data-Driven Model for Identifying Related Pharmaceutically-Significant Entities in Clinical Texts

Alexander Sboev,[1, a)] Gleb Rylkov,[1, b)] Roman Rybka,[1, c)] Artem Gryaznov,[1, d)] and Sanna Sboeva[2, e)]

[1)]*NRC "Kurchatov Institute", Moscow, Russia*
[2)]*I.M. Sechenov First Moscow State Medical University (Sechenov University)*

[a)]*Corresponding author: sag111@mail.ru*
[b)]*Electronic mail: gvrylkov@mail.ru*
[c)]*Electronic mail: rybkarb@gmail.com*
[d)]*Electronic mail: artem.official@mail.ru*
[e)]*Corresponding author: sboevasanna@mail.ru*

INTRODUCTION

To the date, large amount of useful medical data on undesirable effect of pharmaceuticals have been accumulated in electronic health records and Internet users' feedback. To analyze such data so that to find correlations among medicines, their administering, adverse effects they cause, and other entities of significance for pharmaceutics, is a task undoubtedly relevant, but at the same time laborious and requiring automation.

The goal of this work is to create a method for automatically establishing relations among entities. The method is based on the token encoding component made in the architecture of an encoder from the Transformer topology (Bio+Discharge Summary BERT) [1], a Bidirectional Long Short-Term Memory (BiLSTM) layer, and the attention mechanism. The model proposed in this work, unlike many existing in the literature [2, 3], allows analyzing not only text alone, but also quantitative characteristics computed for each pair of entities, in order to deeper examine the entity pair for the presence of a relation, i.e. the input features of the model include additional characteristics important for relation check: distance between entities and the type of the attribute entity. These features were previously [2] used only for balancing data on which to train the model proper.

Validation of the method proposed is performed on the task of determining all attributes associated with a given medicine in a text fragment of a health record extract with entities preliminarily labeled.

DATA

We use electronic health records from the data of the n2c2 2009 [4] competition from Massachusetts Medical School, where *attribute entities* are defined as entities of the following: dosage, duration of admission, frequency of admission, modes of admission, and reasons for admission. *Medication entities* include prescription substances, over-the-counter medications, active ingredients and active substances, established names for groups of pharmaceuticals. A relation between entities is unidirectional, directed from an attribute to a medication. Table I shows the corpus statistics.

Table I. Statistics of the electronic health record corpus

Texts	Relations	Entity types	Entities labeled
260	14672	6	24360

NEURAL NETWORK ARCHITECTURE

The developed method is an improved version of an existing model [2]. It employs a neural network model which predicts a relation between a medicine entity and a probable attribute entity (or vice versa) by the text between them.

The scheme of the model is presented in Fig. 1. The text inputted into Discharge BERT is a medicine entity, its

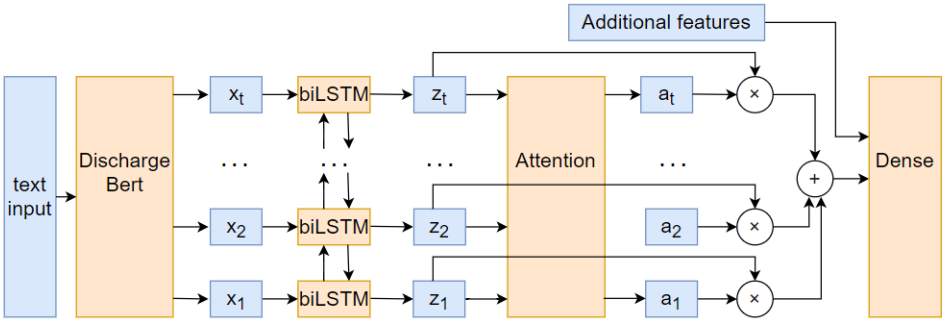

Figure 1. Topology of proposed neural network model

supposed attribute entity, and text between them. As a separate input to the fully-connected layer (denoted as Dense in the figure), additional features are inputted vectorized. These additional features are: the number of word tokens between entities, the number of entity instances between the considered entity pair, type of the supposed drug attribute entity, and the pair order flag (what comes first in the text, the medication or its supposed attribute.)

The components the model is comprised of are explained below using the following notation:

$$\text{softmax}(\overrightarrow{x})_i = \frac{e^{x_i}}{\sum_k^K e^{x_k}}, K = \dim \overrightarrow{x} \tag{1}$$

$$\text{Linear}(\overrightarrow{x}) = W\overrightarrow{x} + b \tag{2}$$

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{3}$$

Denoted in our scheme as *Discharge BERT* is Bio+Discharge Summary BERT [1], the neural-network-based language model that had been preliminarily trained on discharge summary texts from the MIMIC-III [5] dataset of electronic health records. The BERT model is based on the encoder component of the transformer neural network topology [6]. An encoder component, in its turn, is based on l self-attention mechanisms working in parallel.

For each vector of an encoded word $x_i \in [x_1 \dots x_t]$ from the sequence of words with length t, the l-th self-attention mechanism forms three types of vectors: queries q_i^l, values v_i^l, and keys k_i^l. Further, in order to form the output representation of the original word x_i, all the generated value vectors $[v_1^l \dots v_t^l]$ are added, taking into account their contextual significance with respect to the word x_i. The contextual significance a_j^l of all words v_j relative to x_i is calculated using the corresponding query word vector q_i^l and key vectors of the entire sequence $[k_1^l \dots k_t^l]$:

$$b_j^l = \frac{\overrightarrow{q}_i^l \cdot \overrightarrow{k}_j^l}{\sqrt{\dim \overrightarrow{k}_j^l}}, j \in [1...t] \tag{4}$$

$$a_j^l = \text{softmax}(\overrightarrow{b}^l)_j \tag{5}$$

$$\text{SelfAttention}(\overrightarrow{x_i})^l = \sum_{j=1}^{t} a_j^l \cdot \overrightarrow{v_j}^l, \tag{6}$$

Further, all the l vectors obtained using equation (6) are concatenated into one vector, which after the transformation (2) produces the output vector y_i, which describes the word x_i while incorporating information about the entire sequence.

BiLSTM is a component that combines two LSTM [7] layers, where one LSTM layer processes the sequence of encoded words $[x_1 \dots x_t]$ in the order they appear in the text, and the other LSTM layer processes the word sequence in reverse order. For each word x_i, output vectors from the two LSTM layers are concatenated into the resulting output vector y_i, which allows to incorporate information about words both before and after x_i.

Attention is the component that sums up all vectors of encoded words $[x_1 ... x_t]$ of the input sequence,packed into the

matrix X, taking into account their significance a_i:

$$a_i = softmax(\overrightarrow{v_a} X^T)_i, \tag{7}$$

$$\overrightarrow{y} = \sum_{i=1}^{t} a_i \overrightarrow{x_i}. \tag{8}$$

Here v_a is the weight vector of the attention layer, and y is the output vector of this component.

Denoted as *Dense* is a fully connected layer with sigmoid (3) activation function. The output of this layer is what determines the model's decision about the existence of a relation: relation exists if the output exceeds 0.5.

The neural network model described above has been trained to predict the presence or absence of a relation between entity pairs extracted from the original dataset. Weights of Discharge BERT were initialized and frozen during the training phase. The training has been performed on 13,202 entity pairs with relations labeled as present and 132,752 pairs with relations absent.

RESULTS

Applying the proposed model to the corpus has been performed under the following algorithm:

1. for each fragment of input text, a set of entity pairs is formed, containing all such pairs that one entity is a medication and the other is a probable attribute;

2. additional attributes listed in previous section are extracted for each pair;

3. all the formed pairs of entities, the text between them, additional features of the formed pair, are processed with the neural network model to detect relations: every attribute's word is marked by the fact of the existence of a connection with a given medication.

The efficiency of proposed method was estimated using f1_score for each attribute entity type separately, and with the f1_micro score for the overall accuracy:

$$\text{f1\_micro} = \frac{TTP}{TTP+E}, \tag{9}$$

$$\text{f1\_score} = \frac{2 \cdot (P \cdot R)}{P+R}, \quad P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}. \tag{10}$$

Here Total True Positive (TTP) is the number of correctly labeled words in the analyzed fragment, Errors (E) is the number of incorrectly marked words, True Positive (TP) is the number of correctly labeled words related to the one entity type. False Positive (FP) is the number of words related to the one entity type mistakenly marked as related to a given medication, False Negative (FN) is the number of words related to the one entity type mistakenly marked as unrelated to a given medication.

Accuracies of identifying relations between medications and attribute entities of particular types are presented in table II, where the overall accuracy is measured by the f1_micro score, and accuracies for particular entity types of attributes are measured by f1_score. For comparison with the existing method [3], we used the same set of test fragments.

Table II. F1-scores of identifying relations of medication entities with attributes of different types

Method	Dosage	Reception mode	Frequency	Duration	Reason	Overall f1_micro
[3]	0.80	0.86	0.81	0.70	0.67	0.82
Our method	0.90	0.90	0.89	0.86	0.81	0.89

CONCLUSION

An algorithm for identifying relations between entities has been created that is based on a neural network model involving a component of the vector representation of tokens pre-configured on a large body of medical texts (Bio+Discharge Summary BERT), bidirectional LSTM layers and attention mechanisms.

The developed model achieved the accuracy of 0.89 (by the f1-micro score) on the n2c2-2009 corpus, which is higher than results of the existing methods described in the literature, and thus confirms the efficiency of the solution proposed in this work.

ACKNOWLEDGMENTS

The reported study was funded by RSCF project No. 20-11-20246. This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru/.

REFERENCES

1. A. E. et al., "Publicly available clinical bert embeddings." (2019).
2. D. M. Dandala B., Joopudi V., "Clinical notes by jointly modeling entities and relations using neural networks," Drug Saf. Springer International Publishing. Math. **42**, 135–146 (2019).
3. G. L. et al., "Named entity recognition in electronic health records using transfer learning bootstrapped neural networks," Neural Networks. Elsevier Ltd **121**, 132–139 (2020).
4. C. E. Uzuner Ö., Solti I., "Extracting medication information from clinical text," J. Am. Med. Informatics Assoc **17**, 514–518 (2010).
5. J. et. al, "Mimic-iii, a freely accessible critical care database," Scientific Data **3**, 1–9 (2016).
6. V. A. et al, "Attention is all you need," Adv. Neural Inf. Process. Syst. **2017-Decem**, 5999–6009 (2017).
7. S. J. Hochreiter S., "Long short-term memory," Neural Comput **9**, 1735–1780 (1997).