IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 11, 2021, accepted November 27, 2021, date of publication December 13, 2021, date of current version January 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3135381

Cross-Domain Limitations of Neural Models on Biomedical Relation Classification

ILSEYAR ALIMOVA[1], ELENA TUTUBALINA[1,2,3], AND SERGEY I. NIKOLENKO[4,5]

[1]Alexander Butlerov Institute of Chemistry, Kazan Federal University, 420008 Kazan, Russia
[2]Sber AI, 117312 Moscow, Russia
[3]Faculty of Computer Science, National Research University Higher School of Economics, 101000 Moscow, Russia
[4]Steklov Institute of Mathematics, 117966 Saint Petersburg, Russia
[5]Department of Mathematics and Computer Science, Saint Petersburg State University, 199034 Saint Petersburg, Russia

Corresponding author: Elena Tutubalina (tutubalinaev@gmail.com)

The work of Ilseyar Alimova and Elena Tutubalina, shown in Sections II, III, and IV, was supported by the Russian Science Foundation under Grant 18-11-00284. The work of Sergey I. Nikolenko was supported by Saint Petersburg State University under Research Project 73555239 (Artificial Intelligence and Data Science: Theory, Technology, Industrial and Interdisciplinary Research and Applications).

ABSTRACT Relation extraction (RE) aims to extract relational facts from plain text, which is essential to the biomedical research field with the rapid growth of biomedical literature and generally large volumes of biomedicine-related text coming from various sources. Numerous annotated corpora and state-of-the-art models have been introduced in the past five years. However, there are no general guidelines about evaluating models on these corpora in single- and cross-domain settings with diverse entities and relation types. We aim to fill this gap for the task of detecting whether a relation holds between two biomedical entities given a text span. In this work, we present a fine-grained evaluation intended to perform a comparative evaluation of four biomedical benchmarks and understand the efficiency of state-of-the-art neural architectures based on Long Short-Term Memory (LSTM) with cross-attention and Bidirectional Encoder Representations from Transformers (BERT) for relation extraction across two main domains, namely scientific abstracts and electronic health records. We present a comparative evaluation of biomedical RE datasets, including the PHAEDRA, i2b2/VA, BC5CDR, and MADE corpora. Our evaluation of BioBERT and LSTM for binary classification shows significant divergence in in-domain and out-of-domain performance, finding an average drop in F1-measure of 34.2% for BioBERT. The cross-attention LSTM model developed in this work exhibits better cross-domain performance, with a drop of only 27.6% in F-measure.

INDEX TERMS Relation extraction, natural language processing, bioinformatics.

I. INTRODUCTION

Identification of semantic relations between entities found in text, known as *relation extraction* (RE), plays a central role in many areas of biomedical research and healthcare. For example, relation extraction aims to identify the relation between (*Lyrica*, *adverse reaction*, *autoimmune hemolytic anemia*) from the following sentence: ''She had just started on *Lyrica* and was thought perhaps that this had exacerbated her *autoimmune hemolytic anemia*''. Biomedical entity types include drugs/medications/chemicals, drug attributes, diseases, adverse drug reactions, proteins, and other biomedical objects, while relation types cover interactions among these types.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

With the rapid advances of deep learning in recent years, neural relation extraction models have defined state of the art performance for several years. In this work, we follow the currently most well-researched direction and focus on binary relations: given a pair of entity mentions in a text span, the goal is to detect if the text indicates a relation between this pair. Reported results of different neural networks vary substantially on different corpora, with, for example, the F1-measure (a common metric in this problem) ranging at least from 75% to 86% on scientific abstracts [2], [6], [11] and from 81% to 90% on electronic health records [1], [8]. The model performance is frequently evaluated under the implicit hypothesis that the training data (source) and the test data (target) come from the same underlying distribution, e.g., that both sets consist of PubMed abstracts from a specific narrow sub-domain (e.g., cardiology or oncology). Such an

assumption makes it hard to adopt supervised RE models in real-world applications: a new or underresearched domain may not have readily available large-scale labeled datasets. A recurring problem is the re-use of models trained on large-scale biomedical repositories (e.g., PubMed, clinical databases) on new data.

In this work, we perform an extensive evaluation of several neural RE models on four biomedical corpora. Each corpus contains manually annotated relations among entity mentions. Entity annotations can be clustered into those related to sign, symptom, or disease mentions, and those related to medication (drug name or chemical) mentions. We fine-tune:

(i) a classifier based on Bidirectional Encoder Representations from Transformers (BERT) [4] that currently achieves state of the art results on in-domain biomedical RE [2], [6], [11], and
(ii) a novel neural model where context and entities are processed by separate encoders based Long Short-Term Memory (LSTM) [7], which interact through cross-attention layers.

We aim to advance state of the art models in relation extraction with various entity types and context characteristics, concentrating on cross-domain evaluation. In particular, in this work, we seek to answer the following research questions:

RQ1: Do in-domain evaluation with training and testing on each benchmark separately lead to a significant overestimation of performance?

RQ2: Do models trained on texts from one source (e.g., scientific abstracts) perform better on texts from the same source rather than others, i.e., clinical texts (or vice versa)?

We answer the first question positively and show conflicting results on the second. Importantly, we show that the proposed model, while it loses to BioBERT on in-domain evaluation, shows significantly better results in cross-domain evaluation, with a much smaller drop in performance. This suggests that cross-domain RE may need to be decoupled from in-domain RE, and we suggest that the corresponding evaluation should become part of the standard evaluation of new RE models and part of the corresponding benchmarks and leaderboards.

The paper is organized as follows. Section II introduces four datasets that consist of scientific abstracts and/or clinical records. Section III describes neural architectures used in our experiments, including the newly proposed LSTM+CA model. Section IV presents the results of our cross-domain evaluation on the relation extraction task, and Section V concludes the paper.

II. CORPORA

We use the following publicly available benchmarks:

(i) Medication and Adverse Drug Events from Electronic Health Records (MADE) [8],
(ii) BioCreative V CDR (BC5CDR) [12],
(iii) PHArmacovigilence Entity DRug Annotation (PHAEDRA) [19],
(iv) 2010 i2b2/VA corpus [20].

Table 1 shows some basic descriptive statistics of these four datasets.

A. MADE

The Medication and Adverse Drug Events from Electronic Health Records (MADE) challenge [8] introduced a task for extraction of relations among medications, indications, and adverse drug events (ADEs) from electronic health record (EHR) notes taken from 21 randomly selected patients with cancer. These EHRs include discharge summaries, consultation reports, and other clinical notes. The total number of records is 1089, and the train set consists of 876 records. There are three types of relations (7 in total):

(i) drug–indication (reason to use),
(ii) drug–ADE, and
(iii) attribute relations (drug–route, drug–dosage, drug–duration, drug–frequency, other sign–severity).

This corpus contains the largest number of relations (27145) with the largest average and maximum context length between entities (29.9 and 981 respectively). Table 2 shows a summary of different relation types from the MADE corpus. Interestingly, two relation types, drug–indication and drug–ADE, have the maximum distance between entities exceeding 900 characters, which complicates the identification of relations between these entities.

B. i2b2

The second corpus of clinical records used in our experiments is the 2010 i2b2/VA corpus [20]. The 2010 i2b2/VA challenge proposed a task to identify relations between a treatment and a medical problem. The corpus includes 871 annotated documents containing statements of discharges and case histories. There are three types of relationships:

(i) medical problem–treatment,
(ii) medical problem–test,
(iii) medical problem–medical problem.

The first type includes cases in which treatment has improved, worsened, or caused a medical problem or has been prescribed due to a medical problem. The second type indicates that a test is conducted to diagnose a medical problem. The latter includes medical problems that reveal aspects of the same medical problem or cause other medical problems.

A summary of different relation types is shown in Table 2. Relations of the type ''Medical problem–treatment'' and ''Medical problem–test'' dominate in the corpus. The ''Medical problem–test'' type has the largest context in terms of the number of tokens (73), while the ''Medical problem–medical problem'' relation type has the shortest maximum context between entities. On average, the ''Medical problem–treatment'' type has the shortest context length (2.8 tokens), while the ''Medical problem–test'' has the longest context length (4.8 tokens). Compared with other

TABLE 1. Statistics of the datasets used in our experiments.

	MADE	i2b2	CDR	PHAEDRA
Domain	Clinical records	Clinical records	Abstracts	Abstracts
# of documents	1089	871	1500	597
Statistics on entity mentions				
# of entities	9	3	2	3
Entity types	indication, ADE, drug, dose, frequency, duration, route, severity, other signs	medical problem, treatment, test	disease, chemical	indication, drug, disease-related medical subject
Statistics on relations and context length between entities (in tokens)				
# of relation types	7	3	1	3
Avg. context length	29.9	3.7	14.8	15.0
Max. context length	981	73	394	262
Number of annotated relations				
Train set	23036	3120	3001	888
Test set	4109	6293	1512	248

TABLE 2. Statistics on different relation types from the MADE, i2b2, and PHAEDRA corpora. Context length is measured in tokens.

Relation Type	Number of relations			Average context length			Maximum context length		
	Train	Test	All	Train	Test	All	Train	Test	All
MADE									
do	5176	866	6042	8.4	7.7	8.3	215	143	215
reason	4523	870	5393	89.3	63.8	85.2	981	868	981
fr	4417	729	5146	17.7	18.6	17.8	201	178	201
severity	3475	557	4032	2.6	1.8	2.5	259	188	259
adverse	1989	481	2470	59.4	45.6	56.7	937	718	937
route	2550	455	3005	13.5	12.9	13.4	191	137	191
du	906	147	1053	18.5	15.0	18.0	272	121	272
All	23 036	4109	27 145	30.6	26.0	29.9	981	868	981
i2b2									
Medical problem — treatment	1206	2447	3653	2.7	2.8	2.8	58	61	61
Medical problem — test	1159	2398	3557	4.5	4.8	4.7	73	46	73
Medical problem — medical problem	755	1448	2203	3.9	3.4	3.7	33	48	48
All	3120	6293	9413	3.7	3.7	3.7	73	61	73
PHAEDRA									
Subject_Disorder	493	130	623	3.7	3.6	3.7	42	22	42
is_equivalent	229	67	296	1.1	2.1	1.6	15	23	23
Coreference	166	51	217	35.7	43.6	39.7	137	262	262
All	888	248	1136	13.5	16.4	15	137	262	262

corpora, the number of relations in the i2b2 test set exceeds the training set's number of relations. This corpus has the smallest context window between entities (3.7 on average and a maximum of 73 tokens).

C. BioCreative V CDR

BioCreative V CDR (BC5CDR) [12] introduces a task for the extraction of chemical–disease relations from abstracts. The corpus annotations contain entities denoting diseases (Disease) and chemical preparations (Chemical), and relations between these entities. The corpus is divided into three subsets: training, test, and development. The corpus consists of 1500 documents, with the test set taking up one third of the texts.

D. PHAEDRA

The PHArmacovigilence Entity DRug Annotation (PHAEDRA) corpus consists of 597 PubMed abstracts [19]. The corpus contains three relation types:

(i) subject–disorder that includes disorders corresponded to a complaint suffered by the subject(s),
(ii) is_equivalent that shows links between different names of the same concept,
(iii) co-reference relation.

The first relation type is especially important since subjects are frequently characterized by their existing medical conditions while discussing drug effects.

As shown in Table 1, both CDR and PHAEDRA corpora of abstracts contain a significantly lower number of context tokens between entities than the MADE corpus (14.8 and 15.0 respectively vs. 29.9). The PHAEDRA corpus includes the lowest number of annotated relations (1136 in total). Summary statistics of different relation types are shown in Table 2; in particular, it shows that 55% of relations in the corpus are subject–disorder.

E. GENERATION OF NEGATIVE EXAMPLES

Since each corpus contains only positive examples of entity pairs that appear in some relation, we have generated negative samples for the i2b2, CDR, and PHAEDRA sets according to the intra- and inter-sentence level rules proposed by [5]. In particular, for intra-sentence relations, we apply heuristic rules as follows:

(i) the token distance between the two mentions should be less than 10,
(ii) if there are multiple mentions in a sentence that refer to the same entity, keep the nearest pair.

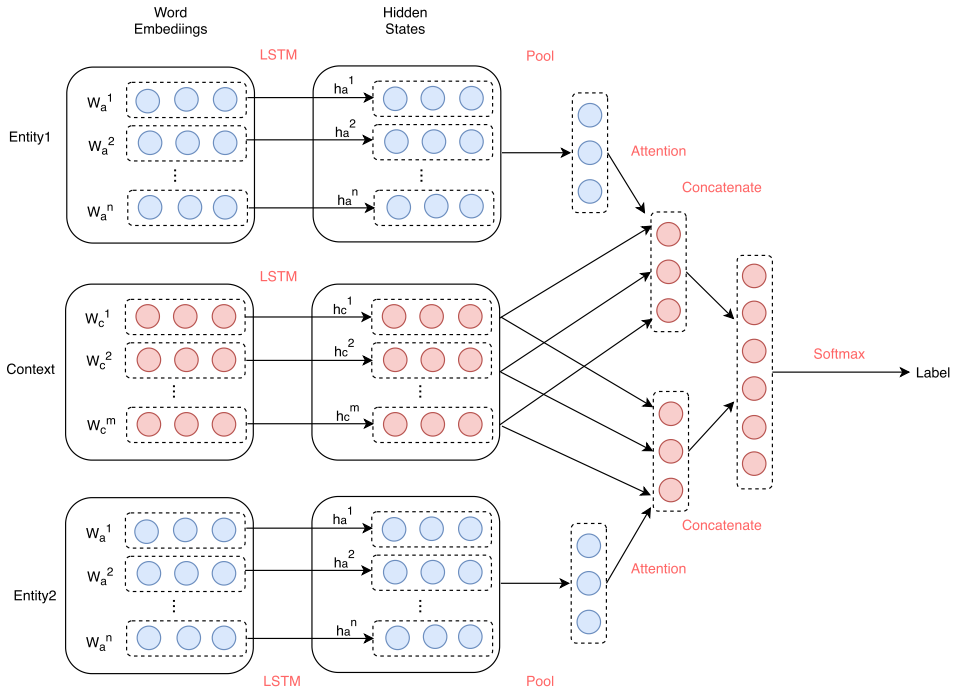

FIGURE 1. Architecture of the proposed LSTM+CA for relation extraction.

For the MADE corpus, where the average distance between entities is larger than in the other three corpora, we use the generation of negative samples proposed by [22]. We apply the following rules:

(i) the number of characters between entities should be less than 1000,
(ii) the number of other entities that participate in relations and occur between candidate entities should not be exceeding 3.

We refer to [5], [22] for more details on negative example generation.

III. MODELS FOR RELATION EXTRACTION

In this section, we describe our proposed relation extraction framework in detail. Formally, each collection consists of a set of documents $D = \{d_1, d_2, \ldots, d_n\}$. Each document includes a set of annotated entities $e^1 = \{w_1^1, w_1^2, \ldots, w_1^{|e^1|}\}$, $e^2 = \{w_2^1, w_2^2, \ldots, w_2^{|e^2|}\}$, $\ldots$, $e^n = \{w_n^1, w_n^2, \ldots, w_n^{|e^n|}\}$. Entities e_i and e_j are in a relationship if the text indicates their semantic interaction or influence on each other. The task definition is as follows. For each pair of entities from one document $e_i^k, e_j^k \in d_k, i, j \in [1, |e^k|], i \neq j$, we determine whether a relation r holds between them, $r(e_i^k, e_j^k) = 1$ or not, $r(e_i^k, e_j^k) = 0$. We view this task as a binary classification problem.

In recent years, state of the art neural methods adopt different variations of attention mechanisms in order to de-emphasize noisy or less important context (in terms of tokens or sentences) for relation extraction [3], [9], [13], [18], [23], [24], [26]. For comparative evaluation, we utilize three models: (i) a BERT-based classifier, (ii) attention-based convolutional neural network (CNN) (further named CNN+A) [18], (iii) a novel cross-attention LSTM-based model (further named LSTM+CA). In particular, we follow the ideas originating in sentiment and multimodal classification [14], [28] and propose to utilize the attention mechanism associated with two entities to obtain important information from the context to compute the final text representation for classification. We note that our goal is not to achieve state of the art performance with large pretrained fine-tuned language models such as BERT on each dataset separately but to ask which architecture can better transfer knowledge from one domain to another.

A. CROSS-ATTENTION LSTM

Our proposed cross-attention LSTM network architecture is presented in Figure 1. The model has three input layers. As input, two layers take the entities encoded by the word embeddings, while the third layer is fed with context between entities, encoded by word embeddings and position embeddings [25]. Position embeddings are based on the assumption that words close to the target entities are usually more informative for determining the relations between entities. To determine position embeddings, we construct two vectors that contain the relative distance from each token in the context to each target entity. If a context token appears in the text after the entity, then the relative position is a positive number, otherwise the position is defined by a negative number. Initially, each position is encoded with randomly initialized vectors of length 5, and then optimal values of the vectors are learned in the process of training the network.

Outputs of all embedding layers are fed to LSTM layers separately. Further, the attention mechanism is applied to the outputs of the LSTM layers.

Let $[h_c^1, h_c^2, \ldots, h_c^n]$ be the hidden context representation obtained from the LSTM layer and let $h_{e_1} = [h_{e_1}^1, h_{e_1}^2, \ldots, h_{e_1}^n]$ and $h_{e_2} = [h_{e_2}^1, h_{e_2}^2, \ldots, h_{e_2}^n]$ be the hidden representations of entities. The average value is calculated for each entity:

$$e_{1,\text{avg}} = \frac{1}{m}\sum_{i=1}^{n} h_{e_1}^i, \quad e_{2,\text{avg}} = \frac{1}{m}\sum_{i=1}^{n} h_{e_2}^i.$$

For the context vector $[h_c^1, h_c^2, \ldots, h_c^n]$, an attention vector α is generated with respect to each entity using the average value of the entity vector $e_{1,\text{avg}}$ and $e_{2,\text{avg}}$. The attention vector for the first entity is defined as

$$\alpha_{1,i} = \frac{\exp(\gamma(h_c^i, e_{1,\text{avg}}))}{\sum_{j=1}^{n}\exp(\gamma(h_c^j, e_{1,\text{avg}}))},$$

where γ is a function that shows the degree of importance of h_c^i in the context. This parameter is defined as

$$\gamma(h_c^i, e_{1,\text{avg}}) = \tanh\left(h_c^i \cdot W_a \cdot e_{1,\text{avg}}^\top + b_a\right),$$

where W_a and b_a are the weight matrix and the offset matrix respectively, and $e_{1,\text{avg}}^\top$ denotes the transposition of $e_{1,\text{avg}}$. Similarly, the vector of the context attention relative to the second entity is calculated as

$$\alpha_{2,i} = \frac{\exp(\gamma(h_c^i, e_{2,\text{avg}}))}{\sum_{j=1}^{n}\exp(\gamma(h_c^j, e_{2,\text{avg}}))},$$

where

$$\gamma(h_c^i, e_{2,\text{avg}}) = \tanh\left(h_c^i \cdot W_a \cdot e_{2,\text{avg}}^\top + b_a\right).$$

The final context vector representation is calculated based on the resulting attention vectors:

$$c_{e_1} = \sum_{i=1}^{n}\alpha_{1,i}h_c^i, \quad c_{e_2} = \sum_{i=1}^{m}\alpha_{2,i}h_c^i.$$

The context presentation vectors c_{e_1} and c_{e_2} are concatenated into one vector for further classification in the linear layer with the softmax activation function.

We trained LSTM+CA with 300 hidden units for 10 epochs using 200-dimensional BioWordVec embeddings [27], learning rate of 0.001, batch size of 32, and Adam optimizer [10]. The BioWordVec embeddings were trained on texts from PubMed and the MIMIC-III Clinical Database.

Note that our in-domain experiments have demonstrated that LSTM+CA outperforms LSTM with one input layer for a text span of entities and context between them.

B. ATTENTION-BASED CNN

We utilize an attention-based convolutional neural network (CNN+A) from [18], for which we introduce and consider a simplified version. This network uses word-level attention to select relevant words with respect to the target entities.

The network consists of two parts: the first part generates a vector representation of the context using the CNN network, while the second part extracts features based on the attention layer. A concatenation of the obtained vectors is then fed as input to the classification layer. Word representations, positional features, and parts of speech are fed to the CNN layer. Positional features are constructed in the same way as in the cross-attention LSTM (Section III-A). Our experiments showed that part-of-speech vectors did not yield any gains in performance and thus were excluded.

The vector of attention weights is calculated as follows. Let each sentence contain T words, and denote their vectors by w_{it}, where $t \in [1, T]$ is the index of a word in the ith sentence, and entity vectors by e_{ij}, where $j \in [1, 2]$ represents the jth entity in the ith sentence. Further, word vectors w_{it} and entity vectors e_{ij} were concatenated to obtain a new representation of words in the sentence:

$$h_{it}^j = \left[w_{it}, e_{ij}\right].$$

After that, u_{it}^j is calculated as the degree of relevance of each word in the sentence in relation to the jth entity in the ith sentence:

$$h_i^{tj} = [w_{it}, e_{ij}],$$
$$u_j^{it} = W_a[\tanh(W_w e)h_i^{tj} + b_{we})] + b_a.$$

The weight vector α_j^{it} is then normalized by softmax:

$$\alpha_j^{it} = \frac{\exp(u_j^{it})}{\sum_t \exp(u_j^{it})}.$$

Finally, the representation of a sentence is computed based on the obtained values of attention weights as follows:

$$s_{ij} = \sum_t \alpha_j^{it} w_{it}.$$

In-domain evaluation has shown that this architecture yield results inferior to LSTM with cross-attention (see Table 3).

C. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a language model based on the bidirectional multilayer Transformer architecture [21]. We utilize BioBERT$_{base}$ v1.0 (+PubMed 200K +PMC 270K) [11] and the authors' implementation of the relation extraction classifier[1] BioBERT was trained on the texts of research paper abstracts from PubMed and PMC. We trained BioBERT for 10 epochs with batch size 32.

IV. CROSS-DOMAIN EVALUATION

We train models on a source train set and evaluate on a target test set. Let us use a *A-B* pair notation to indicate the

[1]Made available at https://github.com/dmis-lab/biobert/blob/master/run_re.py

TABLE 3. In-domain performance of LSTM with cross-attention, attention-based CNN, and BioBERT of macro F1.

Corpus	Model	No-Related (class 0)			Related (class 1)			Average		
		P	R	F	P	R	F	P	R	F
CDR	LSTM+CA	**.850**	.628	.722	.503	**.772**	.609	.676	.700	.666
	CNN+A	.798	.718	.756	.521	.628	.570	.660	.674	.663
	BioBERT	.839	**.812**	**.825**	**.637**	.679	**.657**	**.738**	**.745**	**.741**
Phaedra	LSTM+CA	.971	.962	.966	.727	.782	.753	.849	.872	.860
	CNN+A	.948	.973	.961	.742	.593	.659	.845	.783	.810
	BioBERT	**.978**	**.985**	**.982**	**.884**	**.827**	**.854**	**.931**	**.906**	**.918**
i2b2	LSTM+CA	.912	.852	.881	.664	.779	.716	.788	.816	.799
	CNN+A	.888	.873	.880	.675	.704	.688	.782	.788	.784
	BioBERT	**.919**	**.889**	**.904**	**.726**	**.787**	**.755**	**.822**	**.838**	**.829**
MADE	LSTM+CA	.981	.978	.979	.847	.862	.854	.914	.920	.917
	CNN+A	.973	.982	.978	.863	.806	.834	.918	.894	.906
	BioBERT	**.988**	**.989**	**.989**	**.923**	**.916**	**.920**	**.956**	**.953**	**.954**

TABLE 4. Cross-domain performance of LSTM with cross-attention (LSTM+CA) and BioBERT of macro F1. We report in-domain results on the diagonals (grey cells), the average out-of-domain F1 and drop scores.

Source	Model	Target set				Out-of-domain scores	
		CDR	PHAEDRA	i2b2	MADE	Average F	Average drop
CDR	LSTM+CA	66.6	52.0	51.4	59.0	54.1	-12.5
	BioBERT	74.1	51.9	57.4	51.6	53.6	-20.5
PHAEDRA	LSTM+CA	46.7	86.0	50.9	47.1	48.2	-37.8
	BioBERT	39.9	91.8	47.7	46.5	44.7	-47.1
i2b2	LSTM+CA	57.9	62.8	79.9	54.1	58.3	-21.6
	BioBERT	58.3	59.3	82.9	52.3	56.6	-26.3
MADE	LSTM+CA	52.5	46.8	60.6	91.7	53.3	-37.4
	BioBERT	57.9	44.6	55.6	95.4	52.7	-42.7

training set (A) and the test set (B). In this work, the models were evaluated according to the recommendations presented in [16].

Table 3 presents the results of in-domain models evaluation for RE task in terms of macro F1. The BioBERT model showed the highest results in the average F-measure on all corpora. The highest results were obtained on the MADE case (95.4%). LSTM+CA outperformed the CNNA model on all sets. The largest gain of the LSTM+CA model in comparison with CNNA was achieved on the Phaedra corpus (+5%).

Table 4 presents the results of cross-domain models evaluation for RE task in terms of macro F1. Several observations can be drawn to answer **RQ1**. First, LSTM+CA and BioBERT achieve 81.1%, and 86.1% averaged across four sets on in-domain RE. With the in-domain setup, BioBERT achieves 83.0% F1 on abstracts on 89.2% on clinical texts. From the latter, it follows RE could be considered as a largely solved task on EHRs. Yet, out-of-domain results are significantly worse: LSTM+CA and BioBERT achieve F1 of 53.5% and 51.9% averaged across 12 scores for each model. To answer **RQ2**, we compare two averaged results achieved (i) by models on four pairs (i2b2/MADE)-(CDR/PHAEDRA) and (ii) by models on two pairs of PubMed abstracts CDR-PHAEDRA and PHAEDRA-CDR. To our surprise, results of LSTM+CA on CDR-PHAEDRA are similar to averaged results on (i2b2/MADE)-PHAEDRA (52.0% vs. 54.8%), yet F1 on i2b2-PHAEDRA is higher (62.8%). Although CDR and PHAEDRA have equal context length (app. 15 tokens), context vectors from PHAEDRA are closer to context vectors from i2b2 than to CDR (see below). Altogether, the average difference between the results' pairs is 5-6% F1. We summarize that the loss of quality is not closely related to a change of a source (PubMed vs. EHRs). We also refer to studies on protein-protein interaction corpora of abstracts [16], [17], where the performance of a CNN model on interaction detection varies 15-30% macro F1 on test sets with the unified set of relations [17].

To analyze context similarity between corpora, we compute the Euclidean distance between context representations. For LSTM, we obtained context representations from the LSTM's output layer, which takes the context as input. For BioBERT, we computed context representations as an average of the last four layers' hidden states. These representations were normalized by dividing by maximum.

First, we apply the t-SNE algorithm [15] on each pair of target and source sets to see that context representations do overlap with each other. Figure 2 shows the visualization of context representations. We observe that LSTM's context representations are denser than BioBERT's vectors. Second, we clustered representations and computed pairwise Euclidean distance between cluster centers. According to the calculated metrics of the distance between clusters, LSTM generates closer vectors for different corpora compared to BioBERT. In particular, the distance scores for LSTM's vectors are as follows: 4.25 for PHAEDRA-i2b2, 5.96 for PHAEDRA-CDR, and 5.02 for CDR-i2b2. For BioBERT, the distance scores as follows: 7.19 for PHAEDRA-i2b2, 6.29 for PHAEDRA-CDR, and 6.59 for CDR-i2b2. These observations also confirm that there is room for improving the transferability of BERT-based models, that is, the ability to maintain large-scale performance.

To assess the relationship between the obtained classification results and the context representations, we calculated the Spearman correlation coefficient between the F-measures

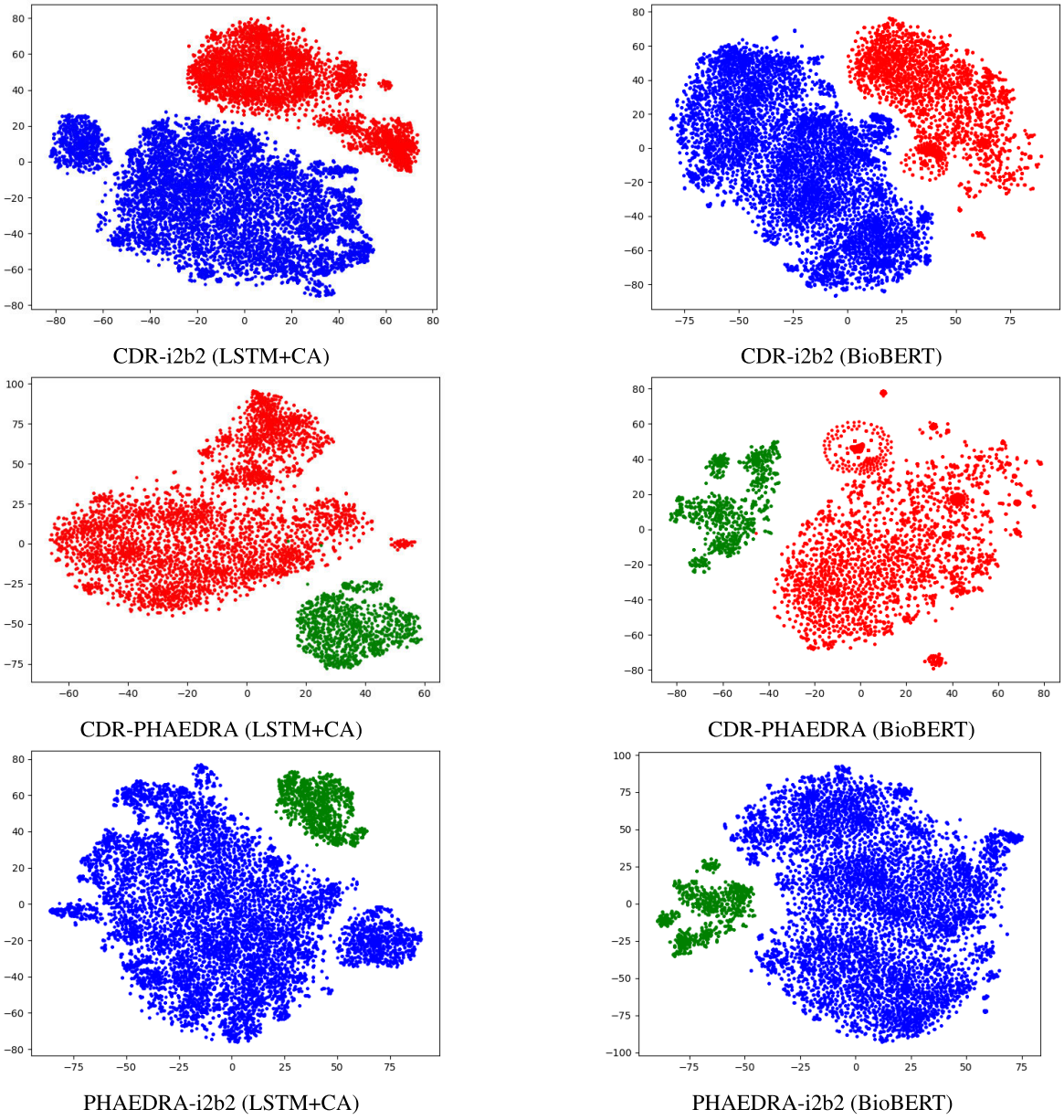

FIGURE 2. Visualization of context representations between entities for CDR (red), Phaedra (green) and i2b2 (blue). Representations are derived from LSTM+CA and BioBERT.

and the distance between the contexts. For the cross-attention LSTM, Spearman's correlation coefficient is -0.717, and for the BioBERT model, the coefficient is -0.517, which proves the inverse relationship between the F-measure and the distance: the larger the F-measure, the closer the distance between sets and vice versa. Moreover, for the cross-attention LSTM, this dependence is stronger (coefficient is closer to -1). This determines higher results for the cross-attention LSTM compared to BioBERT.

V. CONCLUSION

In this work, we have presented a comparative evaluation of BERT-, CNN-, and LSTM-based neural models for relation extraction on four biomedical datasets of scientific abstracts and clinical records. While BioBERT has outperformed CNN and LSTM when measured in terms of in-domain performance, our cross-domain experiments have demonstrated that LSTM with the proposed cross-attention layers outperformed BioBERT by 1.6% F1-measure on out-of-domain relation extraction. This indicates that a fine-tuned language model, even a large one, has limited capacity to decide if a relation holds between two entities given a text span, and this capacity is hard to carry over across datasets and domains, even closely related ones. We have observed that the average drop in performance does not differ greatly depending on the text domain. We believe that this evaluation can serve as a step toward reliable evaluation of relation extraction models and improving restricted leaderboards of

the corresponding competitions. Moreover, this opens up a natural question of developing models that have better cross-domain performance; while in this work we have already presented a model that improves over BioBERT in cross-domain performance, this is merely a first step, and further work is definitely required.

REFERENCES

[1] I. Alimova and E. Tutubalina, "Multiple features for clinical relation extraction: A machine learning approach," *J. Biomed. Informat.*, vol. 103, Mar. 2020, Art. no. 103382.

[2] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 3606–3611.

[3] Q. Dai, N. Inoue, P. Reisert, R. Takahashi, and K. Inui, "Distantly supervised biomedical knowledge acquisition via knowledge graph based attention," in *Proc. Workshop Extracting Struct. Knowl. Sci.*, 2019, pp. 1–10.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[5] J. Gu, L. Qian, and G. Zhou, "Chemical-induced disease relation extraction with various linguistic features," *Database*, vol. 2016, pp. 1–11, Mar. 2016.

[6] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020, *arXiv:2007.15779*.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] A. Jagannatha, F. Liu, W. Liu, and H. Yu, "Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0)," *Drug Saf.*, vol. 42, no. 1, pp. 99–111, 2019.

[9] W. Jia, D. Dai, X. Xiao, and H. Wu, "ARNOR: Attention regularization based noise reduction for distant supervision relation classification," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1399–1408.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[12] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: A resource for chemical disease relation extraction," *Database*, vol. 2016, pp. 1–10, Jan. 2016.

[13] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 2124–2133.

[14] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4068–4074.

[15] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[16] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski, "Comparative analysis of five protein-protein interaction corpora," in *BMC Bioinformatics*, vol. 9. Springer, 2008, p. 6.

[17] A. Ramponi, B. Plank, and R. Lombardo, "Cross-domain evaluation of edge detection for biomedical event extraction," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1982–1989.

[18] Y. Shen and X.-J. Huang, "Attention-based convolutional neural network for semantic relation extraction," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 2526–2536.

[19] P. Thompson, S. Daikou, K. Ueno, R. Batista-Navarro, J. Tsujii, and S. Ananiadou, "Annotation and detection of drug effects in text for pharmacovigilance," *J. Cheminformatics*, vol. 10, no. 1, p. 37, Dec. 2018.

[20] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 552–556, 2011.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[22] D. Xu, V. Yadav, and S. Bethard, "UArizona at the MADE1. 0 NLP challenge," in *Proc. Conf. Mach. Learn. Res.*, vol. 90, 2018, pp. 57–65.

[23] Z.-X. Ye and Z.-H. Ling, "Distant supervision relation extraction with intra-bag and inter-bag attentions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 2810–2819.

[24] Z. Yi, S. Li, J. Yu, Y. Tan, Q. Wu, H. Yuan, and T. Wang, "Drug-drug interaction extraction via recurrent neural network with multiple attention layers," in *Proc. Int. Conf. Adv. Data Mining Appl.* Springer, 2017, pp. 554–566.

[25] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. 25th Int. Conf. Comput. Linguistics, Tech. Papers*, 2014, pp. 2335–2344.

[26] X. Zhang, P. Li, W. Jia, and H. Zhao, "Multi-labeled relation extraction with attentive capsule network," in *Proc. AAAI Conf. Artif. Intelligence*, vol. 33, 2019, pp. 7484–7491.

[27] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," *Sci. Data*, vol. 6, no. 1, 2019, Art. no. 52.

[28] Y. Zhou, S. Mishra, M. Verma, N. Bhamidipati, and W. Wang, "Recommending themes for ad creative design via visual-linguistic representations," in *Proc. Web Conf.*, 2020, pp. 2521–2527.

ILSEYAR ALIMOVA is a Research Scientist at Kazan Federal University. She wrote her Ph.D. thesis at Kazan Federal University and defended it at the Institute for System Programming, Russian Academy of Sciences, in 2021. Her research interests include deep learning and natural language processing.

ELENA TUTUBALINA is a Lead Research Scientist at Sber AI and Kazan Federal University. She wrote her Ph.D. thesis at Kazan Federal University and defended it at the Institute for System Programming, Russian Academy of Sciences, in 2016. Her research interests include natural language processing, especially of biomedical data and graph construction.

SERGEY I. NIKOLENKO received the M.Sc. degree from Saint Petersburg State University, Saint Petersburg, Russia, in 2005, and the Ph.D. degree from the Steklov Institute of Mathematics, Saint Petersburg, in 2009. He is the Head of the Artificial Intelligence Laboratory, Steklov Institute of Mathematics, an Assistant Professor at Saint Petersburg State University, the Chief Research Officer at Neuromation, Tallinn, Estonia, and the Head of AI at Synthesis AI, San Francisco, CA, USA. He is doing research in machine learning (deep learning, Bayesian methods, and natural language processing), analysis of algorithms (algorithms for networking, competitive analysis, and theoretical computer science), and mathematics.