Journal of Biomedical Informatics 103 (2020) 103382

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Multiple features for clinical relation extraction: A machine learning approach

Ilseyar Alimova[a,*], Elena Tutubalina[a,b,c]

[a] Kazan Federal University, 18 Kremlyovskaya Street, Kazan 420008, Russian Federation
[b] St. Petersburg Department of the Steklov Mathematical Institute, 27 Fontanka, St. Petersburg 191023, Russian Federation
[c] Insilico Medicine Hong Kong Ltd, Pak Shek Kok, New Territories, Hong Kong

ARTICLE INFO

Keywords:
Relation extraction
Electronic health records
Natural language processing
Machine learning
Clinical data
Features
MADE corpus
n2c2 corpus

ABSTRACT

Relation extraction aims to discover relational facts about entity mentions from plain texts. In this work, we focus on clinical relation extraction; namely, given a medical record with mentions of drugs and their attributes, we identify relations between these entities. We propose a machine learning model with a novel set of knowledge-based and BioSentVec embedding features. We systematically investigate the impact of these features with standard distance- and word-based features, conducting experiments on two benchmark datasets of clinical texts from MADE 2018 and n2c2 2018 shared tasks. For comparison with the feature-based model, we utilize state-of-the-art models and three BERT-based models, including BioBERT and Clinical BERT. Our results demonstrate that distance and word features provide significant benefits to the classifier. Knowledge-based features improve classification results only for particular types of relations. The sentence embedding feature provides the largest improvement in results, among other explored features on the MADE corpus. The classifier obtains state-of-the-art performance in clinical relation extraction with F-measure of 92.6%, improving F-measure by 3.5% on the MADE corpus.

1. Introduction

Drugs and diseases play a central role in many areas of biomedical research and healthcare. Aggregating knowledge about these entities across a broader range of domains is critical for information extraction (IE) applications. Relation extraction is a central task for various downstream applications such as knowledge base population and knowledge retrieval. At the same time, a large part of biomedical research has been focused on research abstracts; see a comprehensive overview of the field in [1]. In contrast to biomedical literature, research into the processing of electronic health records (EHRs) has not reached the same level of maturity.

There has been increasing interest from both industry and academia in automated computational models for IE from EHRs. There are still differences between relations expressed in research abstracts and EHRs that influence natural language processing (NLP) models: (i) there are clear differences in length and structure, (ii) the writing style and discourse in EHRs are different from that of scientific texts; for example, EHRs often consist of long sentences with multiple dependencies and clauses between entities from different sentences [2]; (iii) doctors and medical workers might use nonstandard vocabulary with shorter and less formal variations of medical concepts or abbreviations. Hence, increased availability of EHRs has represented an opportunity to tailor biomedical NLP algorithms to EHRs trained to extract knowledge contained in these texts.

As shown in Fig. 1, the goal of relation extraction is to detect relations between medical entities in raw texts. Following recent works [3–5], we view the relation extraction task as binary classification. The classifier takes as input preannotated pairs of entities and aims to identify the relation between them. For example, there is a relation between the drug *Ruxience plus CVP* and the duration of *4 cycles* in Fig. 1.

In recent years, there has been a surge of interest in relation extraction for biomedical texts, resulting in several competitive evaluations organized by the research community. Recent systems from the MADE shared task [6] are based on supervised methods and several features [7–10]. However, there is a common limitation in these works; namely, the contribution of features of different types has not been comprehensively investigated.

To fill this gap, we systematically evaluate four types of features on drug-related information extraction from EHRs: distance-based, word-based, knowledge, and embedding. In addition to popular features, we

* Corresponding author.
E-mail addresses: alimovailseyar@gmail.com (I. Alimova), ElVTutubalina@kpfu.ru (E. Tutubalina).

https://doi.org/10.1016/j.jbi.2020.103382
Received 21 August 2019; Received in revised form 27 January 2020; Accepted 28 January 2020
Available online 03 February 2020
1532-0464/

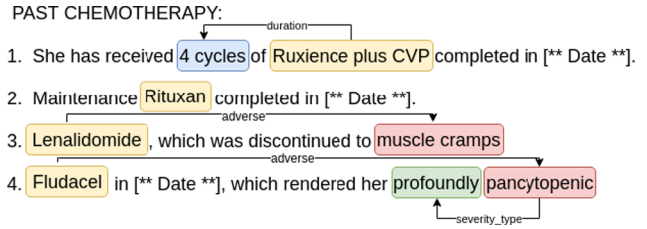

Fig. 1. Relation extraction. The example is chosen from the MADE corpus, where blue, yellow, red, and green circles denote Duration, Drug, ADE, and Severity entities, respectively; duration and adverse, severity_type denote different types of relations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

propose novel features: (i) number of sentences and punctuation characters between entities, (ii) previous co-occurrence of entities in biomedical documents from different sources, (iii) semantic types from Medical Subject Headings (MeSH), and (iv) sentence embedding feature obtained with a sent2vec model [11]. We apply a random forest model and perform experiments on the MADE and n2c2 corpora. For comparison, we evaluate classifiers based on Bidirectional Encoder Representations from Transformers (BERT), and approaches of teams that participated in the MADE and n2c2 shared tasks.

This work is a significantly extended journal version of the conference paper [12]. Compared to the conference version, we have (i) significantly extended the experimental part of this work to assess the performance of the feature-based model; specifically, we added new evaluation experiments on a second corpus from the 2018 n2c2 shared task 2; (ii) carried out an ablation study on two corpora; (iii) extended our description of related work, adding more summaries of the recent neural architectures for relation extraction from general domain (e.g., news); (iv) evaluated concept embeddings for the representation of entities; (v) compared the feature-based model with three versions of BERT-based models in terms of micro-averaged F-measures. Our results show that the length of context between entities has a drastic impact on performance, while other features show similar performance patterns on both corpora.

The paper is organized as follows. We begin with an overview of existing relation extraction methods in Section 2. In Section 3, we present description of two real-world datasets. We present a feature-based model in Section 4. Descriptions of our experimental setup, baselines, and the results are reported in Section 5. Finally, we summarize our contributions and discuss directions for further research in Section 7.

2. Related work

The first attempts to relation extraction from EHRs were made in 2008. Roberts et al. proposed a machine learning approach for relation extraction from oncology narratives [13]. The model is based on SVM with several features, including lexical and syntactic features assigned to tokens and entity pairs. The system achieved an F-measure of 70%. This model has implemented in a Clinical E-Science Framework (CLEF), which aims to extract clinically significant information from EHRs. The full CLEF IE system, including automatic entity recognition, was used to extract 6 million relations from over half a million patient documents.

One of the challenges of an i2b2 2010 competition was devoted to assigning relation types between medical problems, tests, and treatments in clinical health records [14]. This challenge aimed to classify relations between pairs of given reference standard concepts from a sentence. The model based on semantic features from Medline abstracts and parsing trees feature achieved the best performance among other participants [15]. The system obtained F-measure of 73.7%. The model developed by the team from NRC Canada achieved an F-measure of 73.1% [16]. This model is based on the maximum entropy classification algorithm with the following set of features: parsing trees; word surface; concept mapping; context, section, sentence, document-level features; Pointwise Mutual Information between two entities calculated on Medline abstracts. Besides, the authors applied category balancing and semi-supervised training. The system in third place adopted a hybrid approach that combines machine learning techniques and matching of constructed linguistic patterns [17]. The authors trained an SVM with three types of features: surface, lexical, and syntactic. The system obtained an F-measure of 70.9%. The rest of the participants applied supervised approaches and obtained results varying from 70.2% to 65.6% in terms of F-measure [18–22]. One of the main problems faced by participants was an unbalanced number of examples for each relation type. The developed classifiers could capture larger classes accurately by using basic textual features. However, handcrafted rules have to be developed in order to recognize less common relation types.

Further studies on the i2b2 competition corpus were devoted to hybrid systems based on rule-based and machine learning approaches to improve classification performance. D'Souza and Ng employed a combination of rule-based and machine learning models with a rich set of knowledge-based features [23]. This approach yields a 17–24% relative reduction in error over a state of the art learning-based baseline system. Sahu et al. investigated the potential of a convolutional neural network (CNN) for relation extraction [24]. The model takes as input the whole sentence and generates a feature vector for every word in the sentence. The resulting vectors go through convolutional, dense, and softmax layers. The results indicate that CNNs can learn global features that can capture contextual features quite well and thus help to improve the performance. Lv et al. adopted conditional random fields and applied a deep learning model for features optimization by the employment of autoencoder and sparsity limitation [25].

Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE) was organized in 2018 [6]. The competition aimed to extract ADRs and detect relations between drugs, their attributes, and diseases. In contrast to the i2b2 competition, only entities are defined in the corpus. Thus, it is necessary to make candidate pairs and then determine if there is a relation between them. The system in first place utilized a random forest model with the following features: (i) candidate entity types and forms, (ii) the number of entities between and their types, (iii) tokens and part-of-speech tags between candidate entities and adjacent to them [7]. According to the performance metrics table of the competition, the described system obtained micro-averaged F_1 of 86.8%. Dandala et al. applied a combination of bidirectional long short-term memory (biLSTM) and attention network and achieved second place results with micro-averaged F-measure of 84% [8]. The system in third place adopted a support vector machine (SVM) model [10]. The classifier uses four types of features: position, distance, a bag of words, and a bag of entities and obtained an micro-averaged F_1 measure of 83.1%. Magge et al. employed a random forest model with entity types, number of the word in entities, number of words between entities, averaged word embeddings of each entity, and indicator of presence in the same sentence as a feature [9]. This approach obtained micro-averaged F_1 of 81.6%. As one can see from the above, most participating teams applied machine learning models, and the only one utilized neural networks while the results were on par.

Munkhdalai et al. conducted additional experiments on MADE corpus and explored three supervised machine learning systems for relation identification: (1) an SVM model, (2) an end-to-end deep neural network system, and (3) a supervised descriptive rule induction baseline system [26]. The authors used the following features for the SVM model: entity types, a number of clinical entities, tokens between entities, n-grams between two entities and of surrounding tokens, character n-grams of named entities. The combination of biLSTM and attention was utilized as a neural network model. The maximum averaged F-measure of 89.1% was obtained by the SVM classifier, while the neural network achieved only an F-measure of 65.72%.

The modern approaches for relation extraction are based on neural networks. The most widespread model is the convolutional neural network (CNN) [27–29]. Zeng et al. proposed CNN with multi-instance learning, which can extract relation extraction based on distant supervision data [27]. In this model, CNN aims to create a semantic representation of a sentence. The model was evaluated on dataset generated by aligning Freebase relations with the New York Times corpus and achieved 78.3% of precision. However, the method assumes that at least one sentence that mentions these two entities will express their relations and drop a large amount of rich information containing in neglected sentences. To address this problem, Yankai et al. modified the described model adding selective attention over sentence instances [28]. The model achieved 72.2% of precision. Several models have been tested on a benchmark dataset from the SemEval-2010 shared task. Zeng et al. developed the convolutional deep neural network, which takes word embeddings as input and produces lexical and sentence features automatically [29]. This architecture performs relation extraction without complicated text preprocessing. The model obtained 82.7% of F-measure. Zhang et al. applied biLSTM with additional features derived from the lexical resources and NLP systems, including WordNet, dependency parser, and named entity recognizer [30]. This model performed 84.3% of F-measure. Zhou et al. improved biLSTM with an attention layer [31] and achieved 84% of F-measures. This model does not rely on features obtained with additional resources or with NLP tools and takes a row text as an input, while the results stay on par with the model from [30]. Zhi-Xiu et al. studied the problem of a few-shot relation classification [32]. The authors proposed a multi-level matching and aggregation network, which encodes query instances and class prototypes in an interactive fashion. The model achieves a state-of-the-art performance of 92.66% accuracy on a FewRel dataset. This dataset consists of 70,000 sentences on 100 relations derived from Wikipedia and annotated by crowd workers. To sum up, recent approaches for relation extraction from the general domain utilize advanced neural architectures. However, there is a substantial number of differences between EHRs and general texts (e.g., from news or Wikipedia) that make relation extraction from EHRs a specific challenge.

According to the reviewed studies, machine learning approaches have a high potential for the clinical relation extraction task. However, for real-world biomedical applications, the results need to be improved [6]. The error analysis of systems shows three common errors:

(i) related entities more than two sentences away from each other;
(ii) a model wrongly marks entities as related if these entities occur together in a small distance;
(iii) there is more than one entity related to the same entity and only the closest relation is detected.

Also, most of the previously proposed studies devoted to relation extraction from EHRs largely ignore valuable supportive information, such as the context and knowledge sources. Therefore, the machine learning approach proposed in this paper can be viewed as an extension of the previous work on extracting relations from clinical notes.

3. Corpora

We evaluate our model on two corpora of de-identified EHRs: (i) a Medication and Adverse Drug Events from Electronic Health Records 2018 (MADE) corpus [6] and (ii) a National NLP Clinical Challenge 2018 (n2c2) corpus [33]. Each corpus contains manually annotated relations between drugs, diseases, and drugs' attributes. The summary statistics of the MADE and n2c2 corpora of annotated relations are presented in Tables 1 and 2, respectively. The summary of each dataset includes the number of relations, the average and maximum length of context between entities (in characters).

3.1. MADE

MADE corpus consists of EHRs from 21 cancer patients [6]. These EHRs include discharge summaries, consultation reports, and other clinic notes. The overall number of records is 1089, where 876 records were selected for training and 213 notes for testing. Several annotators participated in the annotation process, including physicians, biologists, linguists, and biomedical database curators. Each document was annotated with two annotators, one of which carried out the initial annotation, the second reviewed the annotations, and modified them to produce the final version. The agreement between five annotators computed in a set of three documents was 0.424, which is falls in the fair-to-significant agreement range.

Each record annotated with the following types of entities: drug, adverse drug reaction (ADR), indication, dose, frequency, duration, route, severity, and SSLIF (other signs/symptoms/illnesses). There are 7 types of relations: drug–ade (adverse), sslif–severity (severity), drug–route (route), drug–dosage (do), drug– duration (du), drug–-frequency (fr), drug–indication (reason). As shown in Table 1, the most common relation types are drug-dose, drug-indication, and frequency. Two types of relationships (reason and adverse) have the maximum distance between entities more than 900 characters, which complicates the identification of relations between them.

3.2. n2c2

The n2c2 corpus consists of 505 discharge summaries obtained from the MIMIC-III (Medical Information Mart for Intensive Care-III) clinical resource [33]. A total of 303 annotated files were used as a training set, and 202 files utilized for testing. Each record was screened to contain at least one ADR. The records were annotated by two independent annotators while a third annotator resolved conflicts. The agreement rates were not provided.

The corpus contains the following entity annotations: drug, strength, form, dosage, frequency, route, duration, reason, ADR. There are 8 types of relations: strength–drug (severity), form–drug (form), dosage–drug (do), frequency–drug (fr), route–drug (route), duration–-drug (du), reason–drug (reason), ADR–Drug (adverse). As shown in Table 2, the most common relation types are form–drug, frequency–-drug, and strength–drug. The reason and adverse relation types have the longest context between entities, the same as in the MADE corpus.

Hence, both corpora contain 7 common types of relations, while the n2c2 corpus contains an additional form-drug (form) type.

4. Features

We divide features into four categories: (i) distance-based, (ii) word-based, (iii) embedding, and (iv) knowledge-based. Distance features are based on counting different metrics between entities. Word features are derived using various properties of context and entity words. Embedding features are received from word embedding models pre-trained on a large number of biomedical texts. Knowledge features are obtained from biomedical resources. We describe each type of features below.

1. Distance features:
   - *word distance* (word_dist): the number of words between entities.
   - *char distance* (char_dist): the number of characters between entities.
   - *sentence distance* (sent_dist): the number of sentences between entities.
   - *punctuation* (punc_dist): the number of punctuation characters between entities.
   - *position* (position): the position of the entity candidate (drug or

Table 1
The summary statistics for the MADE corpus.

	# relations			Avg. distance			Max. distance		
Relation type	**train**	**test**	**all**	**train**	**test**	**all**	**train**	**test**	**all**
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

Table 2
The summary statistics for the n2c2 corpus.

	# relations			Avg. distance			Max. distance		
Relation type	**train**	**test**	**all**	**train**	**test**	**all**	**train**	**test**	**all**
do	4225	2695	6920	22.3	23.9	22.9	389	505	505
reason	5169	3410	8579	62.9	63.9	63.3	792	908	908
fr	6310	4034	10 344	30.4	32.4	31.2	413	348	413
severity	6702	4244	10 946	3.6	4.4	3.9	398	313	398
adverse	1107	733	1840	49.9	44.8	47.9	823	500	823
route	5538	3546	9084	26.4	28.6	27.2	402	514	514
du	643	426	1069	40.2	39.3	39.8	350	361	361
form	6654	4374	11 010	20.3	20.7	20.5	404	344	404
all	36 384	23 462	59 810	27.3	28.7	27.9	823	908	908

SSLIF type entity) with respect to the attribute among the entire entity candidates of the attribute, where the position of medical attribute is set to 0.

2. Word features:
   - *bag of words* (bow): all words within a 10-word window before and after the entities plus the entities text. We utilized as features only words that appeared in such context windows with frequencies ⩾500 across the dataset.
   - *bag of entities* (boe): the counts of all annotation types between the entities.
   - *entity types* (type): binary vector indicating the types of entities.
3. Embedding features:
   - *entities embeddings* (ent_emb): the vectors obtained from pre-trained word embedding models for each entity. We explored two word embedding models, including trained on the concatenation of Wikipedia and PubMed, PMC abstracts [34], and BioWordVec created using PubMed and the clinical notes from MIMIC-III Clinical Database [35]. For entities represented by several words the averaged vector value was applied.
   - *concept embeddings* (concept_emb): the vectors obtained from pre-trained medical concept embeddings for each entity. We utilized the set of 500-dimensional embeddings for UMLS concepts trained on insurance claims for 60 million Americans, 1.7 million full-text PubMed articles, and clinical notes from 20 million patients at Stanford [36].
   - *sentence embedding* (sent_emb): the vectors obtained from pre-trained BioSentVec model for words between two entities [11]. BioSentVec was obtained using sent2vec library and consists of 700-dimensional sentence embeddings.
   - *similarity* (sim): similarity measure between entities embedding vectors. Four types of similarity measures were employed: taxicab, Euclidean, cosine, coordinate. The vectors were obtained from BioWordVec model [11].
4. Knowledge features:
   - *UMLS concept types* (umls): UMLS[1] (Unified Medical Language System) semantic types of entities represented with binary vector. We used a publicly available system QuickUMLS [37] and UMLS 2018AA version for extracting UMLS concepts.
   - *MeSH concept types* (mesh): MeSH[2] (Medical Subject Headings) categories of entities represented with a binary vector.
   - *Occurrence in FDA clinical trials* (fda): the number of co-occurrence of both entities in approval document received from FDA[3] for each drug of dataset.
   - *Occurrence in biomedical literature* (bio_texts): the number of entities co-occurrence in biomedical texts. The detailed description of this feature is provided below.

Prior knowledge retrieved from available sources is essential for today's health specialists to keep up with and incorporate new health information into their practices [38]. This process of retrieving relevant information is usually carried out by querying and checking medical articles. We propose a set of features based on primary sources of information to analyze the influence of this process on clinical decision making. In particular, we utilize statistics from various resources using *Pharmacognitive*[4]. This system provides access to databases of grants, publications, patents, clinical trials, and others.

For our experiments, we focus on three sources: (i) scientific abstracts from MEDLINE, (ii) USPTO patents, and (iii) projects from the grant-making Agencies of USA, Canada, EU, and Australia. The *Pharmacognitive* system allows retrieving statistics such as the number of documents or overall funding per year matching a query. The queries

[1] https://www.nlm.nih.gov/research/umls/.
[2] https://www.nlm.nih.gov/mesh/meshhome.html.
[3] https://www.fda.gov/.
[4] https://pharmacognitive.com.

are generated using terms from entities of three types: Drug, Indication, and ADR. We extend all queries with terms' synonyms provided by the Pharmacognitive tools. We consider the following features for a individual query:

- the number of publications/patents/projects published in the particular year (3 features for each year from 1952 to 2018).
- the number of publications/patents/projects published before the particular year (3 features for each year from 1953 to 2018).
- the total number of publications/patents/projects (3 features).
- the average and sum of projects' funding published in the particular year (2 features for each year from 1974 to 2018).
- the average and sum of projects' funding published before the particular year (3 features for each year from 1975 to 2018).
- the average and sum of projects' funding (2 features).

We also generate features based on statistics of publications and projects for joint queries of two terms: *Drug* and a disease-related entity (*ADR* or *Indication*).

5. Experiments

In this section, we describe our classifier model, entity pair generation, experiments, and results. We make the source code available at the github repository https://github.com/Ilseyar/relation-extraction-ehr.

5.1. Classifier

We build a system to resolve the task as a set of independent Random Forest classifiers, one for each relation type. The Random Forest model was implemented with a Scikit-learn library [39]. We tuned the parameters on 5-fold cross-validation and set the number of estimators equal to 100 and the weight balance: 0.7 for positive and 0.3 for negative classes to mitigate the imbalanced class issues.

5.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is a recent neural network model for NLP presented by Google [40]. BERT is based on bidirectional attention-based Transformer architecture [41]. We investigated three models: 1) general-domain BERT[5] [40] 2) BioBERT[6] [42] 3) Clinical BERT[7] [43]. General-domain BERT was pre-trained on BooksCorpus and English Wikipedia. In particular, we utilize BERT_{base}, Uncased, which has 2-heads, 12-layers, 768-hidden units per layer, and a total of 110 M parameters. BioBERT was initialized with General-domain BERT and in addition pre-trained on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) (version: BioBERT v1.0 (+ PubMed 200 K + PMC 270 K)). Clinical BERT was initialized with a general-domain BERT model and pre-trained on clinical texts from the approximately 2 million notes in the MIMIC-III v1.4 database. For general-domain BERT and ClinicalBERT, we ran classification tasks and for the BioBERT relation extraction task. We utilized the entity texts combined with a context between them as an input. All models were trained without a fine-tuning or explicit selection of parameters. We observe that loss cost becomes stable (without significant decrease) after 30–35 epochs.

5.3. Entities pair generation

For each entity, we obtained a set of candidate entities following the rules from [10]: the number of characters between the entities is smaller than 1000, and the number of other entities that may participate in relations and locate between the candidate entities is not more than 3. These restrictions allow to reduce infrequent negative pairs and mitigate the imbalanced class issues, while more than 97% of the positive pairs remain in the dataset.

5.4. Experiments and results

We utilize the model with distance and word features as a baseline. Besides, we compare our results with two state-of-the-art approaches for MADE corpus: proposed by Munkhdalai et al. [26] and by Li et al. [44]. Munkhdalai et al. applied SVM with following features: (i) token distance between the 2 entities, (ii) number of clinical entities between the 2 entities, (iii) n-grams between the 2 entities, (iv) n-grams of surrounding tokens of the 2 entities, (v) one-hot encoding of the left and right entities types, (vi) character n-grams of the named entities. Li et al. utilized modern capsule networks. For n2c2 corpus, we utilize Xu et al. [45] approach as a baseline. This approach based on a combination of biLSTM and CRF models and obtained the best performance on this dataset [33].

For distance and word features evaluation, we removed each of the features individually and in combination. To determine the most significant features from embedding and knowledge features sets, we add each of the features separately to the baseline model. The F-measure for each relation type and micro-averaged over all classes F_1 were used as evaluation metrics. The evaluation scripts provided by competitions' organizers were applied to compute these values. The results for each relation type and micro-averaged F-measure are shown in Tables 3 and 4.

The combination of baseline selected features achieved 86.6% and 85% of micro F-measure on MADE and n2c2 corpora, respectively. The result for MADE corpus stays on par with the best an F-measure of 86.84% achieved in the MADE competition, while for the n2c2 corpus, the baseline performed micro F-measure lower by 11.3% in comparison to results obtained at the competition. The combination of baseline and sentence embedding features achieved the best results of 92.6% of micro-averaged F-measure on MADE corpus. Thus our model outperformed the Munkhdalai et al. results on 3.5%, Li et al. approach on 5.4%, and baseline approach on 6% on MADE corpus. All reported improvements of the baseline model with sentence embedding feature over baseline and both state-of-the-art methods are statistically significant with p-value < 0.01 based on the paired sample t-test.

For the n2c2 corpus, the models with a combination of baseline and UMLS features obtained the best results of 85.2% F-measure. This result did not outperform the first-place results of the n2c2 competition (96.3%). Further, we provided a more detailed analysis of the presented results.

According to the results, the classifier with only distance features achieves the micro-averaged F-measure of 76.6% and 69.1% on MADE and n2c2 corpora, respectively. Word, char, and punctuation features seemed to be complementary to each other due to the absence of one of them lead to approximately the same loss in results on both corpora. The sentence feature is significant for the n2c2 corpus (-1%), while for the MADE corpus, this feature did not give the improvement in results. The position feature is significant for the MADE corpus (-0.8%), while on the n2c2 corpus, the F-measure of the baseline method increased on 0.01% without position feature. The baseline model without distance set of feature (see rows 'word' in Tables 3 and 4) decrease results of micro F-measure on 19% and 37.2% on MADE and n2c2 corpora respectively, which evidences the importance of these parameters for relation classification.

The word-based features also improved the performance of the relation extraction system. The most significant improvement of micro F-measure obtained with a bag of words feature (+3.8 % on MADE and +12.7% on n2c2), which can be explained by a larger vector size

[5] This model is available at https://github.com/google-research/bert.
[6] This model is available at https://github.com/naver/biobert-pretrained.
[7] This model is available at https://github.com/EmilyAlsentzer/clinicalBERT.

Table 3
The results of F-measure for each relation type and averaged micro F-measure of all relation types for MADE corpus. The distance and word features are applied as a baseline.

Features	severity	route	reason	do	du	fr	adverse	all
baseline: distance & word feat-s	.933	.918	.806	.906	.905	.896	.729	.866
Munkhdalai et al. [26]	.950	.960	.750	.880	.910	**.950**	.850	.891
Li et al. [44]	–	–	–	–	–	–	–	.872
baseline-word_dist	.923	.922	.812	.900	.860	.909	.716	.864
baseline-char_dist	.929	.916	.810	.908	.869	.890	.731	.864
baseline-sent_dist	.933	.919	.807	.910	.880	.906	.719	.866
baseline-punc_dist	.926	.912	.798	.907	.836	.906	.735	.863
baseline-position	.931	.917	.803	.897	.865	.883	.723	.858
distance	.918	.843	.683	.859	.713	.780	.525	.766
baseline-boe	.932	.897	.775	.888	.861	.868	.715	.845
baseline-bow	.918	.906	.726	.895	.810	.843	.712	.828
baseline-type	.934	.906	.779	.899	.891	.891	.562	.839
word	.542	.777	.645	.662	.718	.846	.511	.672
baseline + emb_pubmed_pmc_wiki	.927	.898	.730	.887	.684	.900	.605	.827
baseline + emb_bio	.920	.903	.772	.893	.602	.908	.613	.833
baseline + concept_emb	.920	.897	.764	.902	.910	.889	.610	.841
baseline + sent_emb	.936	.954	**.937**	.929	.854	.938	**.869**	**.926**
sent_emb	.932	.935	.909	.915	.854	.835	.782	.884
baseline + sim	.920	.908	.796	.905	.880	.902	.737	.862
baseline + umls	.936	.915	.815	.922	.883	.891	.734	.870
baseline + mesh	.938	.918	.812	.910	.856	.904	.730	.868
baseline + fda	.936	.912	.808	.906	.895	.909	.730	.868
baseline + bio_text	.934	.918	.805	.906	.905	.896	.749	.866
baseline + knowledge	.936	.914	.806	.916	.889	.896	.736	.848
BERT	.951	.976	.845	**.934**	**.946**	.950	.767	.905
BioBERT	**.953**	**.978**	.851	. 930	.940	**.951**	.770	.910
Clinical BERT	.952	.972	.856	.926	.930	.930	.765	.900

Table 4
The results of F-measure for each relation type and micro-averaged F-measure of all relation types for the n2c2 corpus. The distance and word features are applied as a baseline.

Features	severity	route	reason	do	du	fr	adverse	form	all
baseline: distance & word feat-s	.874	.896	.715	.872	**.781**	.839	.706	.912	.850
Xu et al. [45]	–	–	–	–	–	–	–	–	.963
baseline-word_dist	.872	.893	.712	.872	.769	.836	.702	.913	.848
baseline-char_dist	.869	.889	.705	.871	.762	.833	.697	.904	.843
baseline-sent_dist	.875	.892	.690	.868	.763	.825	.666	.902	.840
baseline-punc_dist	.870	.891	.707	.869	.768	.832	.689	.908	.844
baseline-position	.875	.896	.711	.871	.768	.843	.706	.917	.851
distance	.636	.646	.646	.619	.626	.803	.601	.847	.691
baseline-boe	.862	.873	.701	.864	.758	.828	.706	.899	.837
baseline-bow	.662	.696	.672	.652	.639	.816	.648	.886	.723
baseline-type	**.877**	.897	.712	.875	.757	.842	.694	.910	.850
word	.586	.410	.344	.531	.378	.351	.352	.590	.478
baseline + emb_pubmed_pmc_wiki	.740	.781	.569	.803	.535	.632	.453	.807	.740
baseline + emb_bio	.738	.836	.587	.830	.554	.754	.491	.823	.755
baseline + concept_emb	.801	.843	.605	.846	.698	.802	.467	.858	.789
baseline + sent_emb	.870	.874	.593	.846	. 704	. 817	.586	.909	.822
sent_emb	.607	.637	.528	.589	.592	.820	.524	.876	.670
baseline + sim	.817	.893	.701	**.882**	.764	**.850**	.694	.903	.838
baseline + umls	.875	.896	.716	.874	.769	.843	.696	**.921**	**.852**
baseline + mesh	.874	.895	.705	.872	.768	.838	**.708**	.907	.847
baseline + fda	.874	**.907**	.710	.871	.768	.840	.698	.909	.850
baseine + bio_text	.874	.896	.647	.875	.770	.842	.485	.915	.840
baseline + knowledge	.873	.897	.706	.875	.771	.840	.698	.909	.848
BERT	.576	.634	.216	.670	.409	.531	.103	.624	.556
BioBERT	.676	.738	**.726**	.815	.656	.786	.623	.810	.752
Clinical BERT	.678	.735	.725	.817	.654	.783	.619	.808	.746

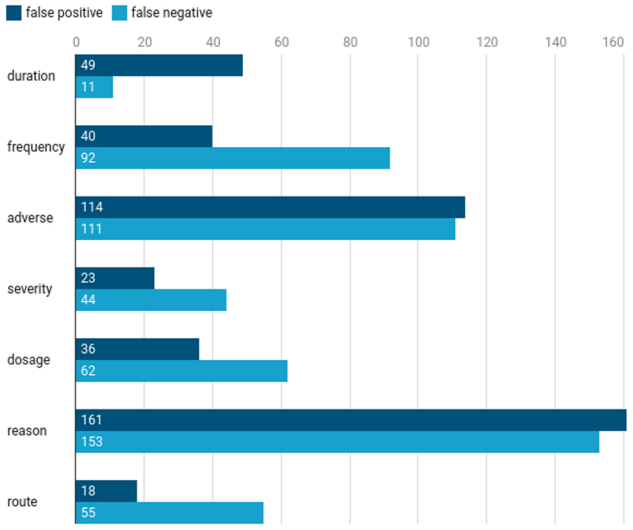

Fig. 2. Error statistics of different relation types for the MADE corpus.

compared to the rest of the word-based features. The bag of entities feature also increased the results of the baselines on 2.1% and 1.3% on MADE and n2c2 corpora, respectively (see rows 'baseline-boe'). The entity type feature improved micro F-measure only on MADE corpus (+2.7%).

The results for embedding features show that entity embeddings and similarity features decrease the results regardless of the word embedding model used. The sentence embedding feature achieved the most considerable improvement of baseline results and obtained 92.6% of micro F-measure on MADE corpus. Moreover, the model trained only with the sentence embedding feature, outperformed the baseline by 1.8%. However, the same results are not observed on the n2c2 corpus. The sentence embedding feature decreased the results of the baseline F-measure on 2.8%.

It is better to consider the results for different relation types separately to evaluate knowledge features. The supplement of UMLS based feature to baseline model increased the results of baseline for severity (0.3%), reason (0.9%), dose (1.6%), and adverse (0.5%) relation types on the MADE corpus. For the n2c2 corpus, this feature increased results on severity (0.1%), reason (0.1%), dose (0.2%), frequency (0.4%) and form (0.9%) types. The combination of baseline and UMLS features performed the best results among knowledge features on both corpora; moreover, this model achieved the best results on the n2c2 corpus (see rows 'baseline + umls'). On the MADE corpus, the supplement of MeSH semantic types increased the results of baseline on the largest number of relation types, including, severity (+0.3%), reason (+0.6%), dose (+0.4%), frequency (0.8%) and adverse (0.1%) types. However, on the n2c2 corpus, the "baseline + mesh" model increased results only for adverse (+0.2%) relation type. The FDA co-occurrence model achieved the most significant increase of F-measure on frequency type (1.3%) for MADE corpus and route type (1.1%) for n2c2 corpus in comparison to the baseline model. The number of co-occurrence in the biomedical texts feature improved the classifier performance for adverse relation type on 2% of F-measure on MADE corpus, while for n2c2 corpus, the same model increased results on 0.3% of F-measure for dose, frequency and form models (see rows 'baseline + bio_text'). Thus, all knowledge features both individually, and in combination, increased results of severity and adverse relation types on the MADE corpus. The knowledge features did not increase the results on route and duration types for the MADE corpus. The UMLS feature increased results on more relation types among knowledge features on the n2c2 corpus. The MeSH feature is more effective for MADE corpus.

The BioBERT model performed the best results among BERT-based models on both corpora (91% on MADE and 75.2% on n2c2). The BioBERT model also achieved the best results for severity (95.3%), route (97.8%), and frequency (95.1%) relation types on MADE corpus and reason (72.6%) type on n2c2 corpus. However, for a reason and adverse types, this model obtained F-measure approximately lower on 10% than the best-achieved results on both corpora. We suppose that the results reducing for adverse and reason types can be caused for two reasons: (i) the same disease in different cases could be an adverse drug reaction and a reason, (ii) the average length of the context for these relation types is too long to catch the relation between entities. It should be noted that on the MADE corpus, the average micro F-measure of BERT-based models stays on par, while on n2c2 corpus, the general-domain BERT model performed significantly lower results in comparison to the rest BERT-based models.

A comparison of results for different types of relation shows that the best result was achieved for route (97.6%) on MADE corpus and form (92.1%) on n2c2. This result roughly stays on par with the best results for severity, reason, dose, duration, and frequency types, while the best results for adverse type lower on 10.7% and 21.3% on MADE and n2c2 corpora, respectively. This difference in results could be due to the greater lexicon variety of adverse drug reaction entity type.

To sum up this section, three important conclusions can be drawn. First, the distance and word-based features are beneficial for the classifier. Second, the sentence embedding has more impact on entities' relations than entities' embeddings. Finally, the prior knowledge improves the results on particular relation types and the most improvement on MADE corpus achieved for adverse relation type with biomedical text co-occurrence feature (+2%) and on n2c2 corpus for route relation type with FDA co-occurrence feature (+1.1%).

6. Error analysis

In this section, we present an analysis of classification errors. We applied the baseline model for both MADE and n2c2 corpora. Figs. 2 and 3 present error statistics of different relation types for MADE and n2c2 test sets. According to the statistic, the number of false-negative errors exceeds the number of false-positive errors for both corpora. The least number of errors the classifier makes for *duration* relation type, while the higher rate of errors noted for *reason* relation type on both corpora. Further, we provide a more detailed analysis of *reason* reason relation type.

Table 5 outlines the main categories of errors found when

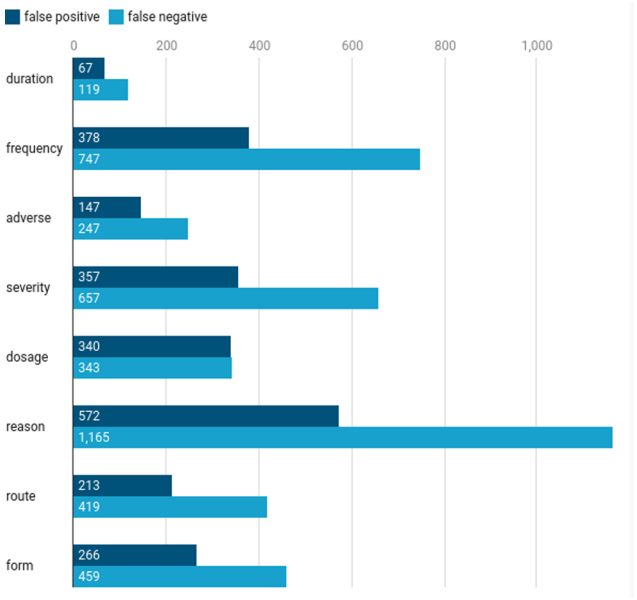

Fig. 3. Error statistics of different relation types for the n2c2 corpus.

Table 5
Error statistics for reason relation type on test sets. Percentages in brackets indicate the proportion of the total number of errors.

Error type	MADE	n2c2
false-negative errors		
more than 10 words in the context	115 (75%)	876 (76%)
entity between first and second entity	98 (64%)	687 (59%)
entities in different sentences	98 (64%)	751 (64%)
more than 5 punctuation in context	33 (22%)	223 (19%)
false-positive errors		
less than 10 words in the context	91 (57%)	497 (87%)
no entities between first and second entity	144 (89.4%)	385 (67.3%)
entities was annotated in a train set as positive	117 (73%)	484 (85%)

evaluating F-measure of *reason* relation type classification on test sets. According to the table, the main reason for false-negative errors is a broad context between entities (75% on MADE and 76% on n2c2). The presence of entities in different sentences equally complicates the identification of relations on the same level for both corpora (64%). The presence of other entities in the context leads to the wrong classification result. This type of error is 5% more common for the MADE corpus. The presence of more than five punctuation marks in the context also complicates the identification of relations (22% on MADE and 19% on n2c2). Such contexts typically describe enumerates or lists. For a false-positive type, the absence of entities and short context are the most common reasons for errors. The short context produces more errors in the n2c2 corpus (87%), while the absence of entities is the most common for the MADE corpus (89%). The presence of entities in the train set as positive examples also leads to the false-positive error types (73% on MADE and 85% on n2c2).

The best results achieved by the developed model for the MADE corpus outperformed the best results obtained on the n2c2 corpus on 7.4% in terms of F-measure (92.6% on MADE and 85.2% on n2c2). We suppose that this is due to a higher disbalance of positive and negative examples in n2c2 corpus (7% of positive samples) in comparison to the MADE corpus (11% of positive samples). We also find out that contexts in the n2c2 corpus contain more abbreviations and terms, including 'WBC', 'HIT', 'PRN'. It makes context representation of n2c2 more complicated and leads to the wrong classification of relation. We assume that more careful pre-processing of texts, including removing punctuation marks, abbreviations, low frequently terms, and reducing the class imbalance, can improve the quality of relation extraction on the n2c2 corpus.

7. Conclusion

In this study, we have investigated different types of features for drug-related information extraction tasks from EHRs. Our evaluation on MADE and n2c2 corpora shows that distance-based and word-based features prove to be the most beneficial for the relation extraction task. The resulting classifier, using a combination of these sets of features with sentence embedding, outperformed state of the art results. These results lead to the conclusion that the context between entities plays a crucial role in relation detection. A detailed analysis of our results has shown that prior knowledge about the entities' co-occurrence improves the results for adverse and form relation types. Besides, we evaluated the general-domain and two biomedical BERT models. The results indicate that these models need fine-tuning for their parameters, including learning rate and batch size. In future research, we plan to focus on the investigation of modern neural networks for relation extraction from EHRs. We also plan to analyze various context representation methods and extend our experiments to other biomedical relation types.

CRediT authorship contribution statement

Ilseyar Alimova: Methodology, Software, Validation, Investigation, Writing - Original Draft. **Elena Tutubalina:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Work on problem definition and development of feature-based models was supported by the Russian Foundation for Basic Research Grant No. 19-07-01115. Work on experiments with BERT-based models was carried out by E.T. and supported by the Russian Science Foundation Grant No. 18-11-00284. The authors would like to thank the anonymous reviewers for their instructive and thoughtful comments and Sergey Nikolenko for helpful discussions.

References

[1] C.-C. Huang, Z. Lu, Community challenges in biomedical text mining over 10 years: success, failure and the future, Briefings Bioinformat. 17 (1) (2015) 132–144.
[2] J. Zheng, H. Yu, Methods for linking ehr notes to education materials, Informat. Retrieval J. 19 (1–2) (2016) 174–188.
[3] J.M. Cejuela, S. Vinchurkar, T. Goldberg, M.S.P. Shankar, A. Baghudana, A. Bojchevski, C. Uhlig, A. Ofner, P. Raharja-Liu, L.J. Jensen, et al., Loctext: relation extraction of protein localizations to assist database curation, BMC Bioinformat. 19 (1) (2018) 15.
[4] Y. Zhang, Z. Lu, Exploring semi-supervised variational autoencoders for biomedical relation extraction, Methods (2019).
[5] D. Ningthoujam, S. Yadav, P. Bhattacharyya, A. Ekbal, Relation extraction between the clinical entities based on the shortest dependency path based lstm, arXiv preprint arXiv:1903.09941.
[6] A. Jagannatha, F. Liu, W. Liu, 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. (2018) 1–13.
[7] A.B. Chapman, K.S. Peterson, P.R. Alba, S.L. DuVall, O.V. Patterson, Detecting adverse drug events with rapidly trained classification models, Drug Saf. (2019) 1–10.
[8] B. Dandala, V. Joopudi, M. Devarakonda, Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks, Drug Saf. (2019) 1–12.
[9] A. Magge, M. Scotch, G. Gonzalez-Hernandez, Clinical ner and relation extraction using bi-char-lstms and random forest classifiers, in: International Workshop on Medication and Adverse Drug Event Detection, 2018, pp. 25–30.
[10] D. Xu, V. Yadav, S. Bethard, Uarizona at the made1. 0 nlp challenge, Proc. Machine Learn. Res. 90 (2018) 57.
[11] Q. Chen, Y. Peng, Z. Lu, Biosentvec: creating sentence embeddings for biomedical texts, 2019 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2019, pp. 1–5.
[12] I. Alimova, E. Tutubalina, A comparative study on feature selection in relation extraction from electronic health records, in: A. Elizarov (Ed.), Data Analytics and Management in Data Intensive Domains, 2523 CEUR Workshop Proceedings, 2019, pp. 34–45.
[13] A. Roberts, R. Gaizauskas, M. Hepple, Y. Guo, Mining clinical relationships from patient narratives, BMC Bioinformat. 9 (11) (2008) S3.
[14] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.
[15] K. Roberts, B. Rink, S. Harabagiu, Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/va shared task, Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.
[16] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, X. Zhu, Nrc at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features, Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.
[17] C. Grouin, A.B. Abacha, D. Bernhard, B. Cartoni, L. Deleger, B. Grau, A.-L. Ligozat, A.-L. Minard, S. Rosset, P. Zweigenbaum, Caramba: concept, assertion, and relation annotation using machine-learning based approaches, in: i2b2 Medication Extraction Challenge Workshop, 2010.
[18] J. Patrick, D. Nguyen, Y. Wang, M. Li, I2b2 challenges in clinical natural language processing 2010, Proceedings of the 2010 i2b2/VA Workshop on Challenges in

Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.
[19] S. Jonnalagadda, G. Gonzalez, Can distributional statistics aid clinical concept extraction, in: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data. Boston, MA, USA: i2b2, 2010.
[20] G. Divita, O. Treitler, Y. Kim, et al., Salt lake city vas challenge submissions, Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.
[21] I. Solt, F.P. Szidarovszky, D. Tikk, Concept, assertion and relation extraction at the 2010 i2b2 relation extraction challenge using parsing information and dictionaries, in: Proc. of i2b2/VA Shared-Task. Washington, DC, 2010.
[22] D. Demner-Fushman, E. Apostolova, R. Islamaj Dogan, et al., Nlms system description for the fourth i2b2/va challenge, Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.
[23] J. DSouza, V. Ng, Knowledge-rich temporal relation identification and classification in clinical notes, Database 2014.
[24] S. Sahu, A. Anand, K. Oruganty, M. Gattu, Relation extraction from clinical texts using domain invariant convolutional neural network, Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 206–215.
[25] X. Lv, Y. Guan, J. Yang, J. Wu, Clinical relation extraction with deep learning, IJHIT 9 (7) (2016) 237–248.
[26] T. Munkhdalai, F. Liu, H. Yu, Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning, JMIR Public Health Surveillance 4 (2) (2018).
[27] D. Zeng, K. Liu, Y. Chen, J. Zhao, Distant supervision for relation extraction via piecewise convolutional neural networks, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1753–1762.
[28] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2124–2133.
[29] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 2335–2344.
[30] S. Zhang, D. Zheng, X. Hu, M. Yang, Bidirectional long short-term memory networks for relation classification, Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 2015, pp. 73–78.
[31] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 207–212.
[32] Z. Ye, Z. Ling, Multi-level matching and aggregation network for few-shot relation classification, CoRR abs/1906.06678. arXiv:1906.06678. URL: http://arxiv.org/abs/1906.06678.
[33] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, n2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc. (2018).
[34] S. Moen, T.S.S. Ananiadou, Distributional semantics resources for biomedical text processing, Proc. LBM (2013) 39–44.
[35] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, Biowordvec, improving biomedical word embeddings with subword information and mesh, Scientific Data 6 (1) (2019) 52.
[36] A.L. Beam, B. Kompa, I. Fried, N.P. Palmer, X. Shi, T. Cai, I.S. Kohane, Clinical concept embeddings learned from massive sources of medical data, CoRR abs/1804.01486. arXiv:1804.01486. URL: http://arxiv.org/abs/1804.01486.
[37] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction, in: MedIR workshop, sigir, 2016.
[38] M.L. Pao, S.F. Grefsheim, M.L. Barclay, J.O. Woolliscroft, M. McQuillan, B.L. Shipman, Factors affecting students' use of medline, Comput. Biomed. Res. 26 (6) (1993) 541–555.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Machine Learn. Res. 12 (2011) 2825–2830.
[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[42] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, BioinformaticsBtz682. arXiv:http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz682/30132027/btz682.pdf, doi:10.1093/bioinformatics/btz682. URL: https://doi.org/10.1093/bioinformatics/btz682.
[43] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.
[44] F. Li, H. Yu, An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models, J. Am. Med. Inform. Assoc. 26 (7) (2019) 646–654.
[45] J. Xu, H.-J. Lee, Z. Ji, J. Wang, Q. Wei, H. Xu, Uth_ccb system for adverse drug reaction extraction from drug labels at tac-adr 2017, in: TAC, 2017.