Comparative analysis of context representation models in the relation extraction task from biomedical texts*

Ilseyar Alimova
Kazan Federal University
Kazan, Russia
alimovaIlseyar@gmail.com

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

Abstract

This paper focuses on the task of extracting relations between entities in biomedical texts. This study aims to identify the most effective method for representing context between entities. We compare several context representation methods such as a bag of words representation, average word embeddings, sentence embedding, representations obtained by convolutional, recurrent neural networks, and bidirectional encoder representations from Transformers (BERT). We conduct a set of experiments on two benchmark corpora of patient electronic health records and scientific articles in English. As expected, thehighestclassificationresultswereobtainedwiththestate-of-theart neural architecture BERT.

1 Introduction

Relation extraction is one of the crucial problems in the field of natural language processing and information extraction. Relation extraction aims to extract from unstructured text entities, which are semantically connected. Relation extraction is a main step for developing different systems in the fields of natural language processing, including, question-answering systems [34], ontology [41], information retriever [4]. In this paper, we focus on extracting relations from biomedical texts [32]. In the field of biomedical text processing, relation extraction is applied to extract adverse drug reaction and drug-related information [30], detecting protein-protein interactions [6], identifying the influence of chemical on disease [42].

The context between two entities is essential for relation extraction. Two entities can be related that depends on the context between two entities. For example, in the passage of receipt given to a patient “Lorazepam 1 mg every 6 hours in case of nausea, Omeprazole 20 mg in a day” nausea is the indication of Lorazepam; therefore entities Lorazepam and nausea are related to each other. However, in the sentence “Prochlorperazine 10 mg every 6 hours in case of nausea, Valacyclovir 500 mg 2 times a day, Lorazepam 1 mg in case of insomnia” the entities Lorazepam and nausea are not related.

Inthispaper, we perfor man extensive comparison of context representation methods in order to identify the most effective method for relation extraction in the biomedical domain. We consider several methods of context representations: (i) a bag of words representation; (ii) averaged word embeddings from a word2vec model [27]; (iii) sentence embeddings from a sent2vec [7], (iv) representations obtained by convolutional neural networks (CNN) [22], long short-term memory (LSTM) [14], and bidirectional encoder representations from Transformers (BERT) [11]. We conduct a set of experiments on MADE and CDR corpora of texts from various sources. Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE) corpus consists of annotated electronic health records [16]. BioCreative V chemical

disease relation (CDR) corpus [42] includes annotations of scientific articles on a biomedical domain. This study examines the relationship between drugs and their attributes and between chemicals and diseases.

2 Related Work

There are various approaches to the problem of identifying related entities in biomedical texts [3,5,15,18,24,28, 38]. Earlier works on relation extraction adopted frequency-based methods. This method calculates the frequency of the entities occurrences within the given context length. If the resulting number is greater than the specified threshold, then it is considered that the entities are related. The advantage of this approach is the simplicity of its implementation, no need for linguistic analysis and labeled sampling. However, the significant drawback of this approach is that it does not take into account the semantic interpretation of context between entities presented in a text.

The template-based approach is grounded in finding a match for linguistic patterns, represented as regular expressions. Templates are generated automatically or manually based on context. The advantage of this approach is that there is no need for an annotated corpus. However, a wide variety of contests generates a large number of templates, which significantly reduces the quality of the system [8,12,35].

The increase of annotated corpora of biomedical text number leads to experiments with machine learning methods to the problem of the relation extraction [2,19,20,31,36,39]. According to this approach, the context is encoded as the feature vectors. The most common features are:

- bag of words: a feature vector that consists of the words before, after, and between entities;
- part of speech tags: a feature vector consisting of parts of speech words before, after and between entities;
- distance between the entities: the number of the words between the entities, the number of indicator wordsbetween the entities, for example, specific verbs that indicate the existence of a connection;
- shortest syntactic tree path: the encoded shortest path from one entity to another in the syntactic parse tree.

Recent relation extraction approaches are based on neural networks, where context and entities are encoded with word embeddings as an input [10, 25, 37]. Sahu et al. applied CNN for extracting relations from patients' electronic health records [37]. The model utilized as input the whole sentence encoded with word embeddings. The obtained vectors sequentially passed through convolutional and dense layers. The results show that CNN can extract global features, which can give good context representation and improve the quality of the system. Lv X. et al. adopted autoencoder for context representation [25]. The experiments indicate that the proposed model is effective, and the method of optimizing functions by the deep learning model has great potential. Dandala et al. employed bidirectional long short term memory network with attention for extracting relations from electronic health records [10]. The proposed approach achieved 84% of F-measure.

A review of the literature shows that machine learning models are the most widely used method, and the most common method for representing context is a bag of words. There are no studies that utilize sentence embeddings to solve the problem of context representations.

3 Context Representation Methods

Let **context** be text between two entities. For the evaluation, we select several approaches for context representation, ranging from the simplest methods, such as a bag of words and an average vector representation of words, to more complex, such as a vector representation of sentences, convolutional, and recurrent neural networks. A classifier takes the context representation between two entities as input and predicts whether it express a relation. **Bag of words** (bow) is one of the first models of text presentation, proposed by Zelling Harris in 1954 [13]. Currently, a bag of words is actively used for text classification and information retrieval. According to this model,

the number of occurrences of each word from the dictionary is calculated for the text, where the dictionary is a set of unique words of all the texts of the training dataset. The model does not take into account the word order in the text, which is one of its main disadvantages. Besides, the final text representation vector has a large dimension.

The averaged word embeddings (word2vec) is calculated by summing the embedding of each word in the context divided by the total number of words in the context. Tomas Mikolov proposed the word embedding model in 2013 [27]. It is based on a neural network trained to predict a word by context on a large corpus of text, the hidden states of which are later used as vectors for words. The advantage of this model is the ability to consider the semantic meaning of the words. Thus, the vectors of words that are close in meaning will be close to each other in the vector space. However, this property can be lost on the text level due to averaging vectors. Also, this representation has a fixed dimension for all texts, equal to the length of the word embedding vector.

Sentence embeddings (sent2vec) are one of the variations of word embedding representation model [33]. However, the neural network trains not only on separate words but also on word n-grams and the averaged embeddings for the words in a sentence. Thus, the model can better represent the semantic meaning of the sentence than a simple averaging of word embeddings.

Convolutional neural network (CNN) is widely used for context modeling [17,21,22]. The network takes as an input a matrix E consisting of context words encoded with word embeddings. We apply a standard convolutional layer over the matrix E. It is followed by a global max-pooling layer to produce the text embedding:

$$b_{ij} = \sum\sum k \odot E_{[i-s,i+s]}$$

$$z_s = max_i B, \quad ,$$

where $k \in R^{v \times d}$ is a kernel matrix, v is the width of a kernel; $B \in R^{(n-v) \times d}$ is a matrix composed of elements b_{ij}. The j axis is computed using different parallel kernels. The max operation is applied alongside the i axis.

Thus, each neuron on the next layer is connected not with all neurons, but only with a small localized subset of neurons in the previous layer. This fact allows for identifying the most significant features for each of the input matrix fragments. The pooling layer is used to reduce the size of the feature map. Most often, the function of maxpooling or weighted average pooling is used.

Recurrent neural network (RNN) is used to process sequential data such as time series or word sequences [26]. The network utilizes information from the previous network states, which is one of the critical advantages of this model. The model takes context words encoded with word embeddings as an input. At each step, the network calculates the weights using the word embedding vector and the output obtained at the previous step. We use the last cell state as the context representation. $y_s = c_n$,

$$h_i = \text{RNN}(w_i, c_{i-1}),$$

$$a_i = \frac{\exp\left(h_i^\top A y_s\right)}{\sum_{j=1}^{n} \exp\left(h_i^\top A y_s\right)}$$

where c_n is the RNN memory state after reading the entire input sequence; h_i is the RNN output produced using w_i (a word embedding) and c_{i-1} (memory state from the previous time step) as inputs.

BERT (Bidirectional Encoder Representations from Transformers) is a recent neural network model for NLP presented by Google [11]. The model obtained state-of-the-art results in various NLP tasks, including question answering, dialog systems, text classification, and sentiment analysis. BERT neural network based on bidirectional attention-based transformer architecture [40]. One of the main model advantages is the ability to give it a row text as the input. In our experiments, we calculated the averaged vector of each word in the context. We utilize a biomedical version of BERT called BioBERT [23].

4 Datasets

We conduct experiments on two annotated corpora of biomedical texts: MADE [16] and CDR [42]. The overall corpora statistic is presented in Table 1.

Table 1: The overall statistic of corpora.

Corpus	# Relations	Avg. context len.	Max. context len. (in characters)	# Unique context words
MADE	27 145	29.9	981	17 443
CDR	3 013	167.1	1 021	16 197

4.1 MADE corpus

Indication and Adverse Drug Events from Electronic Health Record Notes (MADE) corpus consist of 1089 anonymized electronic health records of patients with cancer [16]. Electronic records include an extract statement, inspectionresults, and othernotes. The corpus containsnine typesof entities andseven typesofrelations. Annotated entities can be divided into two groups: related to the disease or the drug. Entities of the first group: adverse drug reaction (ADE), a reason to use the drag (Indication), the severity of the disease (Severity), and other symptoms and diseases not included in previous groups (SSD). Entities related to drugs: name (Drug name), dose (Dose), duration of taking a drug (Duration), frequency of taking a drug (Frequency), route of taking a drug (Route).

The corpus includes seven types of relationships, 4 of which are between the name of the drug and its attributes:

- Drug name – Dose
- Drug name – Route
- Drug name – Frequency
- Drug name – Duration
- Drug name - Indication
- Drug name - ADE
- SSD - Severity, includes the relationship between the severity of the disease and all types of entities includedin the group of diseases: ADE, Indication, SSD.

Entities in relations can be found both in one sentence and indifferent ones. The corpus is divided into training and test subsets.

4.2 CDR corpus

The CDR corpus was developed for the BioCreative V competition [42]. The corpus consists of abstracts of scientific articles collected from the PubMed resource. The corpus annotations contain the entities denoting diseases (Disease) and chemical preparations (Chemical), and the relations between these entities. The corpus is divided into three subsets: training, test, and development. In this work, the training and development subsets are combined into one common train subset; the model is evaluated on a test subset.

4.3 Generation of negative examples for training

Manual annotations in both corpora contain only positive examples, denoting related entities. It is necessary to generate negative examples to train models for binary classification. For each entity, we obtained a set of candidate entities following the rules from [16]: 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 MADE dataset, and 100% remain in CDR corpus.

5 Experiments and Results

We applied word vectors trained on the texts of PubMed and PMC resource articles and Wikipedia texts [29] for the average vector of context representations. The length of the vectors is 200. The vocabulary coverage is 93%

for CDR and 89% for the MADE corpus. For sentence, embeddings were obtained from the BioSentVec model, pre-trained on the text corpus consisting of articles from the PubMed resource and electronic patient cards of the MIMIC-III base [7]. The model is trained on bigrams, with a window size of 20 words, the length of the resulting vectors is 700.

We utilized freezed weights from the last layer of BioBERT model [23]. 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 200K + PMC 270K)).

Following [21], we trained convolutional neural network with the following parameters: the number of layers is 3, the size of the layer filters are 5, 4, 3, the number of epochs is 10, the batch size is 32, the weights for classes is 0.7 for related entities, and 0.3 for unrelated entities. A recurrent neural network was trained with the following parameters: the number of hidden states is 200, the dropout is 0.2, the number of epochs is 20, the size of the input data block is 64, the weight for the classes is 0.75 for entities that have a connection and 0.25 for unrelated entities. All implementation is based on Keras and TensorFlow libraries [1,9].

We employed a support vector machine (SVM) as a classifier. The classifier takes as an input various context representationssequentially. Theclassifierwasevaluatedwithstandardmetrics: precision(P),recall(R),F-measure (F). The results are presented in Table 2.

Table 2: Results of SVM with different context presentations.

	MADE			**CDR**		
Method	**P**	**R**	**F**	**P**	**R**	**F**
bow	.878	.573	.693	.395	.341	.367
word2vec	.760	.800	.779	**.557**	.312	.400
sent2vec	.894	.873	.883	.437	.376	.405
CNN	.725	.825	.772	.446	.334	.382
RNN	.482	.404	.440	.297	**.516**	.377
BERT	**.929**	**.882**	**.905**	.473	.385	**.424**

According to the results, all models outperformed the baseline results of the bag of words model, which obtained 69.3% and 36.7% F-measures on MADE and CDR corpora, respectively. The best method of context representation is BERT for both corpora. This model achieved 90.5% and 42.4% of F-measures on MADE and CDR corpora, respectively. The averaged sent2vec method performed the second results. For the CDR corpus, the difference between sent2vec and BERT models is 1.9, while on the MADE corpus, the difference is 2.2%, which is more significant. CNN outperformed the RNN on MADE and CDR corpora on 33.2%, while the result for CDR corpus state on par. The highest results in terms of precision and recall for CDR corpus was achieved by averaged word embeddings method (55.7% of precision) and recurrent neural network (51.6% of recall). On the MADE corpus, the highest results of precision and recall were achieved with the BERT model (92.9% and 88.2%, respectively).

The results show that the F-measure on the MADE corpus is higher than on the CDR corpus in common. Such a difference in results could be due to the MADE corpus has significantly more examples of relations, which allows the classifier to learn the parameters better and make a better classification.

[*] This model is available at https://github.com/naver/biobert-pretrained.

6 Conclusion

In this paper, we have investigated several methods for representing the context in the task of extracting relations between biomedical entities. The study aims to identify the most effective methods of context representation. The experiment results showed that the BERT model performed the highest results. In the future, we plan to evaluate models considered in the article for the protein-protein relation extraction task.

Acknowledgments This research was supported by the Russian Foundation for Basic Research grant no. 190701115.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, GeoffreyIrving, MichaelIsard, etal., *Tensorflow: Asystemforlarge-scalemachinelearning*, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

[2] Syed Toufeeq Ahmed, Radhika Nair, Chintan Patel, and Hasan Davulcu, *Bioeve: bio-molecular event extraction from text using semantic classification and dependency parsing*, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, Association for Computational Linguistics, 2009, pp. 99–102.

[3] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski, *All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning*, BMC bioinformatics **9** (2008), no. 11, S2.

[4] Omar Alonso, Jannik Strötgen, Ricardo A Baeza-Yates, and Michael Gertz, *Temporal information retrieval: Challenges and opportunities.*, Twaw **11** (2011), 1–8.

[5] William A Baumgartner, K Bretonnel Cohen, and Lawrence Hunter, *An open-source framework for largescale, flexible evaluation of biomedical text mining systems*, Journal of biomedical discovery and collaboration **3** (2008), no. 1, 1.

[6] Christian Blaschke, Miguel A Andrade, Christos A Ouzounis, and Alfonso Valencia, *Automatic extraction of biological information from scientific text: protein-protein interactions.*, Ismb, vol. 7, 1999, pp. 60–67.

[7] Qingyu Chen, Yifan Peng, and Zhiyong Lu, *Biosentvec: creating sentence embeddings for biomedical texts*, The 7th IEEE International Conference on Healthcare Informatics (2019).

[8] Yong Suk Choi, *Tree pattern expression for extracting information from syntactically parsed text corpora*, Data Mining and Knowledge Discovery **22** (2011), no. 1-2, 211–231.

[9] François Chollet et al., *Keras: The python deep learning library*, Astrophysics Source Code Library (2018).

[10] Bharath Dandala, Venkata Joopudi, and Murthy Devarakonda, *Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks*, Drug safety **42** (2019), no. 1, 135–146.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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.

[12] Andrew D Fox, William A Baumgartner, Helen L Johnson, Lawrence E Hunter, and Donna K Slonim, *Mining protein-protein interactions from generifs with opendmap*, Linking Literature, Information, and Knowledge for Biology, Springer, 2010, pp. 43–52.

[13] Zellig S Harris, *Distributional structure*, Word **10** (1954), no. 2-3, 146–162.

[14] S.HochreiterandJ.Schmidhuber, *LongShort-TermMemory*, NeuralComputation**9**(1997), no.8, 1735–1780, Based on TR FKI-207-95, TUM (1995).

[15] Minlie Huang, Xiaoyan Zhu, and Ming Li, *A hybrid method for relation extraction from biomedical literature*, International journal of medical informatics **75** (2006), no. 6, 443–455.

[16] Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong 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 safety (2018), 1–13.

[17] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom, *A convolutional neural network for modelling sentences*, arXiv preprint arXiv:1404.2188 (2014).

[18] Halil Kilicoglu and Sabine Bergler, *Adapting a general semantic interpretation approach to biological event extraction*, Proceedings of the BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics, 2011, pp. 173–182.

[19] Mi-Young Kim, *Detection of gene interactions based on syntactic relations*, BioMed Research International **2008** (2008).

[20] Sun Kim, Soo-Yong Shin, In-Hee Lee, Soo-Jin Kim, Ram Sriram, and Byoung-Tak Zhang, *Pie: an online prediction system for protein–protein interactions from text*, Nucleic acids research **36** (2008), no. suppl_2, W411–W415.

[21] Yoon Kim, *Convolutional neural networks for sentence classification*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[22] Yann LeCun, Yoshua Bengio, et al., *Convolutional networks for images, speech, and time series*, The handbook of brain theory and neural networks **3361** (1995), no. 10, 1995.

[23] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, *Biobert: pre-trained biomedical language representation model for biomedical text mining*, arXiv preprint arXiv:1901.08746 (2019).

[24] Florian Leitner, Scott A Mardis, Martin Krallinger, Gianni Cesareni, Lynette A Hirschman, and Alfonso Valencia, *An overview of biocreative ii. 5*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **7** (2010), no. 3, 385–399.

[25] Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu, *Clinical relation extraction with deep learning*, International Journal of Hybrid Information Technology **9** (2016), no. 7, 237–248.

[26] Larry Medsker and Lakhmi C Jain, *Recurrent neural networks: design and applications*, CRC press, 1999.

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, *Distributed representations of words and phrases and their compositionality*, Advances in neural information processing systems, 2013, pp. 3111–3119.

[28] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii, *A rich feature vector for protein-protein interaction extraction from multiple corpora*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 121–130.

[29] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou, *Distributional semantics resources for biomedical text processing*, Proceedings of LBM (2013), 39–44.

[30] Tsendsuren Munkhdalai, Feifan Liu, and Hong Yu, *Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning*, JMIR public health and surveillance **4** (2018), no. 2, e29.

[31] Yun Niu, David Otasek, and Igor Jurisica, *Evaluation of linguistic features useful in extraction of interactions from pubmed; application to annotating known, high-throughput and predicted interactions in i2d*, Bioinformatics **26** (2009), no. 1, 111–119.

[32] Stanley Chika ONYE, Arif AKKELEŞ, and Nazife DIMILILER, *Review of biomedical relation extraction*.

[33] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi, *Unsupervised learning of sentence embeddings using compositional n-gram features*, arXiv preprint arXiv:1703.02507 (2017).

[34] Deepak Ravichandran and Eduard Hovy, *Learning surface text patterns for a question answering system*, Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 41–47.

[35] Dietrich Rebholz-Schuhmann, Antonio Jimeno-Yepes, Miguel Arregui, and Harald Kirsch, *Measuring predictioncapacityofindividualverbsfortheidentificationofproteininteractions*, Journalofbiomedicalinformatics **43** (2010), no. 2, 200–207.

[36] Rune Sætre, Kenji Sagae, and Jun'ichi Tsujii, *Syntactic features for protein-protein interaction extraction.*, LBM (Short Papers) **319** (2007).

[37] Sunil Kumar Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu, *Relation extraction from clinical texts using domain invariant convolutional neural network*, arXiv preprint arXiv:1606.09370 (2016).

[38] Isabel Segura-Bedmar, Paloma Martínez, and César de Pablo-Sánchez, *A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents*, BMC bioinformatics, vol. 12, BioMed Central, 2011, p. S1.

[39] Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, and Yves Van de Peer, *Extracting protein-protein interactions from text using rich feature vectors and feature selection*, 3rd International symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku Centre for Computer Sciences (TUCS), 2008, pp. 77–84.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[41] Dmitriy Yur'yevich Vlasov, Dmitriy Yevgen'yevich Pal'chunov, and Pavel Andreyevich Stepanov, *Avtomatizatsiya izvlecheniya otnosheniy mezhdu ponyatiyami iz tekstov yestestvennogo yazyka*, Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnyye tekhnologii **8** (2010), no. 3.

[42] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu, *Overview of the biocreative v chemical disease relation (cdr) task*, Proceedings of the fifth BioCreative challenge evaluation workshop, vol. 14, 2015.