

Using Multimodal Clustering on Formal Contexts in Event Extraction with Neural Nets

Mikhail Bogatyrev¹, Dmitry Orlov¹ and Ilya Zapyantsev¹

¹Tula State University, 92 Lenin ave., Tula, Russia

Abstract

The most of the tasks of event extraction from textual data are solved using neural networks. One of the ways to train neural networks on texts is the use of n -grams. In this paper, we consider n -grams in the form of subgraphs of conceptual graphs that have a certain semantics. Each such n -gram can be treated as an event. The paper proposes to use an additional information resource in the form of a hierarchy of multimodal clusters based on a multidimensional formal context, the $n+1$ dimensional tensor elements of which consist of n -grams and elements selected as objects. The task of event extraction is solved by the joint use of a neural network and a hierarchy of multimodal clusters.

Keywords

Event extraction, conceptual n -grams, Formal Concept Analysis, multimodal clustering

1. Introduction

This paper represents the ongoing research on applying Formal Concept Analysis (FCA) [1] to neural-based solutions to certain text mining tasks. Among these tasks, there is Event extraction [2]. Event extraction constitutes an area of methods for acquiring additional information in unstructured text so that it represents an answers on so-called "5W1H" questions "Who did What to Whom, Where, When and How" [2]. The answers do not necessarily have to contain all the elements of the "5W1H" question, which creates a variety of event representations. Event extraction plays an important role in many areas of artificial intelligence applications. Among them there are social media, economics and finance, biomedical applications, etc. The event extraction task can be interpreted as a generalization of two text mining tasks: named entity recognition and relationship extraction. Accordingly, the arsenal of tools for solving these tasks is used in event extraction. Among these tools, neural networks currently play a key role. In this paper, we investigate the use of n -grams extracted from text to train a neural network instead of training it on full texts. This method of learning is known and is used in a number of tasks. But we use "meaningful" conceptual n -grams, which contain answers to the "5W1H" questions in whole or in part. Each such n -gram also can be treated as an event. However, when training a neural network on conceptual n -grams by standard way, the semantics of "meaningful" n -grams is not represented in the network. The paper proposes to use an additional

Published in Sergei O. Kuznetsov, Amedeo Napoli, Sebastian Rudolph (Eds.): The 11th International Workshop "What can FCA do for Artificial Intelligence?", FCA4AI 2023, co-located with IJCAI 2023, August 20 2023, Macao, S.A.R. China, Proceedings, pp. 25-34.

✉ okkambo@mail.ru (M. Bogatyrev)

🆔 0000-0001-8477-6006 (M. Bogatyrev); 0000-0003-0756-7491 (D. Orlov)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

information resource in the form of a hierarchy of multimodal clusters based on a multidimensional formal context, the $n+1$ -tensor elements of which are n -grams and the elements selected as objects. The task of event extraction is solved by the joint use of a neural network and a hierarchy of multimodal clusters. The neural network solves the classification problem and returns the class objects corresponding to the query text upon request. Then these objects are searched in multimodal clusters. These clusters contain combinations of events related to the classes of objects found by the network.

Using conceptual n -grams for learning allows one to apply neural network with a simple architecture. The use of an additional resource in the form of a hierarchy of clusters built on n -grams presents additional information to obtain a solution of event extraction task.

The paper is organized as follows. In the Section 2, there are brief descriptions of event extraction task and multimodal clustering problem in FCA with mentioning the main related works in these areas. In the Section 3, the proposed approach is outlined together with its functionality. The Section 4 contains conclusion. The paper ends with acknowledgements and references sections.

2. Preliminaries and Related Work

This work is related to two principles. The first principle of *query refining* has long been known in FCA community [3]. If a query is identified with some node of conceptual model (in FCA it is conceptual lattice), then its refinements or possible answers to it are located in neighboring nodes. In our method, class objects found by the neural network act as queries. Clusters as formal concepts corresponding to queries, form a lattice in which query refining principle operates. Having fixed the cluster that best matches the query, we find the clusters adjacent to it and use this principle to extract events.

The second principle of *concept-based explanations* [4] is applied in neural networks and aims to implement the notion of a concept in the layers of deep neural networks. In the work of [5] this principle is used to build an intrinsically interpretable document classifier using a combination of FCA and approaches from applied graph theory. At the current stage of research, we do not use the semantics of conceptual n -grams directly in the neural network, limiting ourselves to the use of standard configuration networks. However, the principle of concept-based explanations is applied when working with multimodal clusters.

2.1. Neural-based event extraction from texts

The main definitions in the event extraction are the following [2]. An *event mention* usually is a phrase or sentence that describes an event in which a trigger and corresponding arguments are included. Not every sentence contains event mention, so that sentences without events must be recognized and omitted. *Event trigger* is usually a verb or a noun that most clearly expresses the core meaning of an event. We use verb-oriented algorithm for acquiring conceptual graphs from text and construct n -grams. *Event type* refers to the category to which the event corresponds. In most cases, event types are predefined manually, categorized by event triggers. For instance, we may be interested on existing some verbs in the text, i.e. attack, shoot, etc. *Event arguments* are the main attributes of events. They are usually entity mentions describing the event state change, involving who, what, when, where, and how. An *argument role* is a function or position that an event argument performs in the relationship between the event argument and the trigger.

There are two directions in event extraction. They are *Closed-domain event extraction* and *Open-domain event extraction*. *Closed-domain event extraction* uses predefined event schema to discover and extract desired events of particular type from text. An event schema contains several event types and their corresponding event structures. *Open-domain event extraction* aims at detecting events from texts and in most cases, also clustering similar events via extracted event keywords. Event keywords refer to those words/phrases mostly describing an event, and sometimes keywords are further divided into triggers and arguments.

There are a large number of works devoted to event extraction from the text. The review [2] contains 245 references. Among the works closest to our topic and using neural networks, we highlight the following two papers.

In the work of [6] the Abstract Meaning Representation (AMR) graphs acquired from text are used. AMR graph is a rooted directed acyclic graph where the nodes represent concepts and the edges represent relations between these concepts. Events are considered as subgraphs of AMR graph. Neural network is used to identify an event subgraphs. The neural network is trained on a tagged textual corpus. Tagging includes the definition of events and their arguments. Methodology of event extraction presented in the work of [6] needs external information resources. Since this work is about biomedical event extraction, these resources are knowledge base containing relations between proteins and large corpus of unannotated text that contain protein mentions.

The work of [7] represents recent results of nested event extraction from biomedical domain data. In this work, event extraction model named as DeepEventMine is proposed. It extracts multiple overlapping directed acyclic graph structures from a sentence. An event is called nested event when it has other events in its arguments, while an event is called flat event when there are only entities in its arguments. According to DeepEventMine model, an event consists of a trigger and zero or more arguments. A trigger is a textual mention that denotes the presence of an event in text. Triggers, events, arguments and roles are all united in graph structures. The model is strongly tied to pre-trained networks, primarily to the BERT network. The implementation of the model requires large computing resources.

Based on the mentioned works and other works in this research area, the following conclusions can be drawn.

1. Graph models of event representation are used in the works.
2. Neural networks used in event extraction have a complex architecture.
3. To extract events, external resources are needed in the form of text corpora, thesauri, and databases.

2.2. Multimodal clustering in FCA

In FCA, multimodal clustering is formulated as follows.

If $R \subseteq D_1 \times D_2 \times \dots \times D_n$ is a relation on data domains D_1, D_2, \dots, D_n then *formal context* is an $n + 1$ set:

$$\mathbb{K} = \langle K_1, K_2, \dots, K_n, R \rangle \quad (1)$$

where $K_i \subseteq D_i$. *Multimodal clusters* on the context (1) are $n -$ sets

$$\mathbb{C} = \langle X_1, X_2, \dots, X_n \rangle \quad (2)$$

which have the closure property [8]:

$$\forall u = (x_1, x_2, \dots, x_n) \in X_1 \times X_2 \times \dots \times X_n, u \in R \quad (3)$$

and $\forall j = 1, 2, \dots, n, \forall x_j \in D_j \setminus X_j \langle X_1, \dots, X_j \cup \{x_j\}, \dots, X_n \rangle$ does not satisfy (3).

A multimodal cluster is a subset in the form of combinations of elements from different sets K_i . It is also defined as a closed n -set [9] since the closure property (3) provides its “self-sufficiency”: it cannot be enlarged without violating (2).

Formal concepts on multimodal formal context are those multimodal clusters where *for all* $u = (x_1, x_2, \dots, x_k) \in X_1, X_2, \dots, X_k, u \in R$ and k is maximally possible. In other words, they are the largest possible k -dimensional hypercubes completely filled with units. The concept of the density of a multimodal cluster is introduced in FCA and formal concepts are interpreted as absolutely dense clusters [9].

3. Methodology

The considered approach is implemented by performing the following steps.

1. Conceptual n -grams are built on the text under processing. We consider every such n -gram as a graph in the form of a tree. For a given sentence, it may be dependency parse tree, AMR-graph, conceptual graph or any other n -gram model. Depending on the type of n -grams, their construction can be performed by an appropriate software tool. All the currently most well known variants of n -grams are supported by parser programs. It is necessary to ensure the storage of n -grams, for example, in a database. Some text corpora store n -grams as tagging.

2. A multidimensional formal context is formed as an $n+1$ -dimensional tensor. Every point of this tensor contains an object and a n -gram it corresponding. Objects can be documents containing texts, terms, entity names, text topics, etc. Multimodal clusters are built on this formal context.

3. The task of event extraction is solved by the joint use of a neural network and a hierarchy of multimodal clusters. The neural network solves the classification problem and, upon request, returns objects-classes corresponding to the query text. Then the clusters corresponding to these objects are determined. They contain combinations of events related to the query text.

3.1. Experiments and Early Results

Consider implementation of the outlined methodology and its experimental study.

Data sets. We explored our methodology on two datasets. The first dataset is The Stanford Question Answering Dataset (SQuAD) [10]. In this dataset there are topics, corresponding texts, questions and answers. This dataset is often used to train neural networks applied in question-answering systems. Extracting events in SQuAD corresponding to the semantic template “who, what, with whom and where and when” will allow one to form answers to questions related to the template elements.

The second dataset (KaggleTDC) is a collection of approximately 1000 newsgroup documents from 10 different newsgroups [11]. This dataset differs from SQuAD in that its newsgroups more overlap by words. Therefore, we expected greater uncertainty in determining the newsgroup from the text by the neural network.

Conceptual n -grams. The sequence of n tokens in the text is called an n -gram. The token can be a phoneme, a syllable, a letter, a word or base pairs according to the application. We construct n -grams by acquiring conceptual graphs [12] from text and selecting their subgraphs having three, four or five concepts connected by *agent*, *patient* or *attribute* relations. Accordingly, we have trigrams, 4-grams and 5-grams.

Figure 1 shows an example of conceptual graphs corresponding to the sentence "Dioxins are highly toxic, and health effects on humans include reproductive, developmental, immune and hormonal problems" related to the topic of the Immune system.

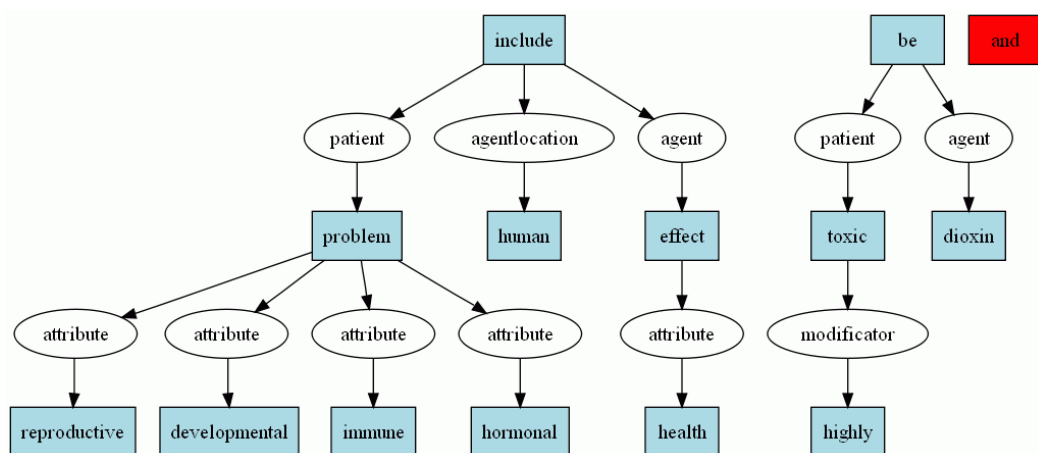


Figure 1: An example of conceptual graphs corresponding to the sentence "Dioxins are highly toxic, and health effects on humans include reproductive, developmental, immune and hormonal problems"

In Figure 1, two disconnected conceptual graphs correspond to the sentence above. By selecting the *agent* and *patient* relationships in graphs, we have the trigrams “effect – include - problem” and “dioxin

– be - toxic". When adding *attribute* relations, we get five 5-grams based on the first trigram with the concepts of *health*, *hormonal*, *immune*, *developmental*, *reproductive*. For example, it will be a 5-gram “health – effect - include – problem – hormonal”, which corresponds to the event template given above. All these conceptual *n*-grams may be considered as events.

Conceptual graphs used in our work constitute a special case of AMR graphs. Conceptual graphs represent the semantics of individual sentences by acquiring from text known semantic roles as *agent*, *patient*, *attribute*, etc. We do not use external information resources and train neural network on *n*-grams extracted from conceptual graphs.

Neural Network. We use a compact chain recurrent neural network of standard architecture. An example of a structure of such network for 5 classes is shown in Figure 2.

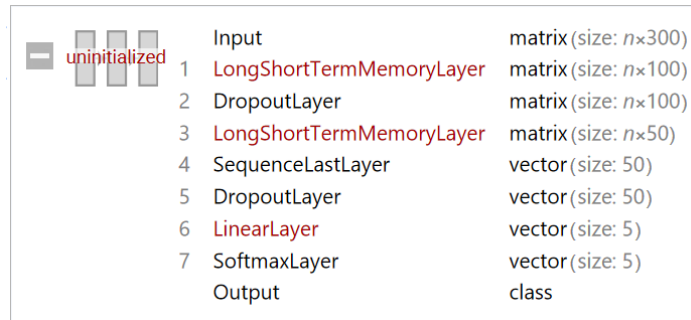


Figure 2: An example of a structure of neural network for 5 classes.

The network is trained on *n*-grams generated from the texts of the dataset. When training the network, several optimization methods were used: ADAM, RMSProp, SGD, Signed.

The input data in the network are phrases in natural language, usually in the form of a question. Conceptual *n*-grams are constructed corresponding to input phrases. If no *n*-grams correspond to a phrase, the first *n* words of the phrase come to the input layer. If several *n*-grams correspond to a phrase, then they consistently arrive at the input layer. The output of the network is the object-class corresponding to the text query. Training and network operation using conceptual *n*-grams were compared with the training method in which *n*-grams formed by using a sliding context window (SCW). The results are presented in the tables 1, 2 and on Figure 3.

The tables 1 and 2 contain some results of training and network operations for conceptual *n*-grams and for SCW *n*-grams constructed on the SQuAD dataset and on the KaggleTDC dataset.

Table 1

Results of training and network operations for SQuAD dataset

Parameter	Conceptual <i>n</i> -grams	SCW <i>n</i> -grams
Number of objects	636	30166
Time training (sec.)	0.2854565	5.2802665
Accuracy	0.796875	0.728207

Table 2

Results of training and network operations for KaggleTDC dataset

Parameter	Conceptual <i>n</i> -grams	SCW <i>n</i> -grams
Number of objects	1339	67076
Time training (sec.)	2.97245	163.408
Accuracy	0.802	0.811

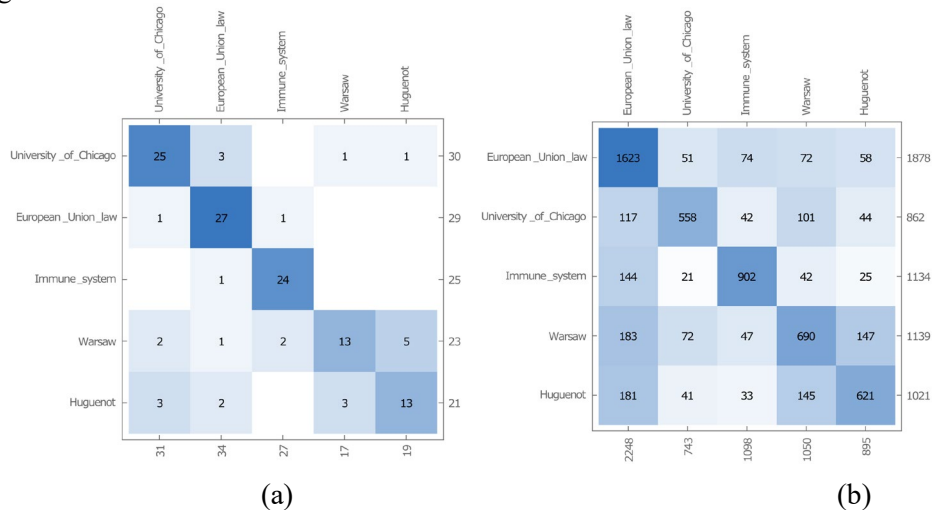
The table 3 contains information about the five most voluminous topics-classes from the SQuAD data set.

Table 3

F1-Score for five most voluminous topics-classes from the SQuAD data set

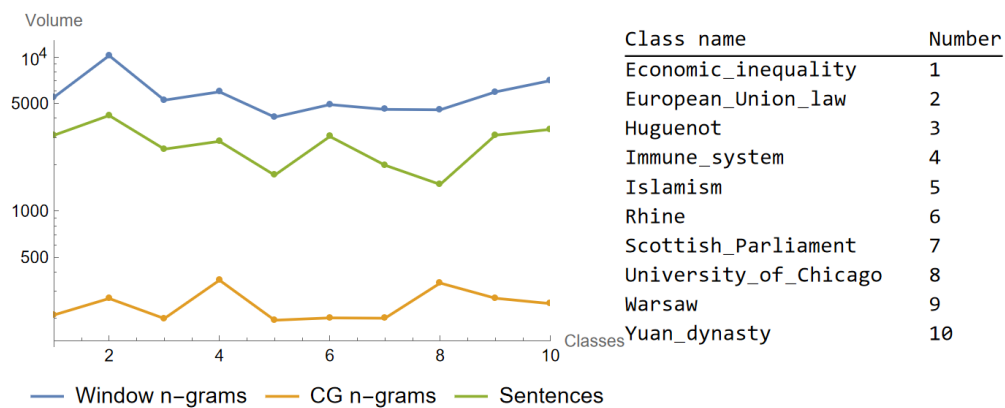
Class	Conceptual n -grams	SCW n -grams
European_Union_law	0.85714	0.78671
University_of_Chicago	0.81967	0.69532
Immune_system	0.92307	0.80824
Warsaw	0.65	0.63042
Huguenot	0.65	0.64822

Figure 3 shows confusion matrices for neural network learned by conceptual n -grams and by SCW n -grams.

**Figure 3:** Confusion matrices for neural network learned by conceptual n -grams (a) and by SCW n -grams

In general, from the most general positions, it can be argued that the more words are used when training a neural network, the better it works. The latter means, among other things, that the network responds to requests more meaningfully. Here we do not take into account the effect of overfitting.

Among others, there are two ways to train neural networks on textual data. The first one uses a bag of words – all the words of the text without stop words. The second uses n -grams obtained by applying a sliding context window. Both methods use a lot of words, but they clearly do not reflect the semantics of the text, since they are simple sequences of words. Figure 4 shows how the volumes of n -grams and sentences of the text are correlated.

**Figure 4:** N -gram volumes for different classes of the SQuAD dataset.

Conceptual n -grams that we form consist of a small number of words and their number is more than 10 times less than the number of n -grams formed by using SCW as can be seen in Figure 4. Based on this, it can be assumed that the use of conceptual n -grams will reduce the performance of the neural network. In reality, this does not happen, as can be seen from Table 2, Table 3 and Figure 3. In our opinion, this is due precisely to the presence of semantics in conceptual n -grams.

As the value of n in conceptual n -grams increases, the quality of the network increases too, as can be seen from Table 4.

Table 4
Variants of n -grams and results of their application

Type of n -gram	$F1$ -score
3-gram	0,704036
4-gram with agent role	0,67713
4-gram with patient role	0,737643
5-gram	0,817518
Complex 3-5-gram	0,8998

Joint use of neural net and multimodal clusters. The neural network is trained on n -grams, which are interpreted as events. The texts at the network input are also converted into n -grams. As a result, the network works only with n -grams. When solving a classification task, in response to an input query, the network returns a set of class names for the SQuAD dataset and a set of newsgroup names for the KaggleTDC dataset. It was found in experiments that in these sets there are often objects-classes with similar probabilities, which makes it difficult to choose between them. For these purposes, multimodal clusters are used. There are such clusters-formal concepts, where subsets of objects include objects-classes with similar probabilities. In such clusters, all n -grams are combined with all objects. As a result, we get a set of events corresponding to the request to the neural network.

Multimodal clustering is performed using an evolutionary algorithm that allows obtaining Pareto-optimal solutions for several criteria of cluster optimality. The framework [13] is used for multimodal clustering. The problem of multimodal clustering is formulated as a problem of multiobjective optimization. In the experiments, two clustering optimality criteria were used, the cluster volume and its density. These criteria contradict each other, so compromise Pareto-optimal solutions are needed. First of all, we were interested in absolutely dense clusters being formal concepts. In them, all objects are combined with all elements of n -grams. Large and dense clusters are interesting because combinations of elements of its subsets set a property that manifests itself on a large number of elements and, possibly, means a regularity. However, often the clustered data is sparse and the existence of large and dense clusters on them is unlikely. Therefore, when selecting clusters, a trade-off between density and volume is provided by the algorithm.

The quality characteristics that are usually used in the classification problem (accuracy and $F1$ -score) are integral and do not reflect the quality of individual solutions, their semantics. In the task of event extraction, it is important to detect events as facts and present the composition of events. In our experiments, we still detect events as facts. In multimodal clusters, we count the number of n -grams and fix their composition. Table 5 shows the content of several multimodal clusters containing pairs of class objects. For every pair there are corresponding number of n -grams shown in the table.

Table 5
Contents of some multimodal clusters

Class objects	The number of n -grams
University of Chicago Warsaw	8
Immune_system European_Union_law	2
European_Union_law Scottish Parliament	28

In the Table 5, *University of Chicago* and *Warsaw* classes are linked by eight n -grams having mutual words in corresponding texts. *Immune_system* and *European_Union_law* classes are practically unlinked since their mutual words in two n -grams are commonly used. It is evident that *European_Union_law* and *Scottish Parliament* classes are closely linked by 28 n -grams.

As a result, in clusters we find events that have common class objects and class objects that have common events.

Obviously, in the task of event extraction, only n -grams are not enough for the end user and additional information is needed. The n -grams found in multimodal clusters correspond to text fragments that can be presented to the user in response to a request. In this case, events take the form of descriptions understandable to the user. Such a solution may be designed in the form of a user interface focused on a specific subject area.

4. Conclusion

In this paper, an approach to training neural networks on n -grams in the form of conceptual graphs is proposed, which allows using a recurrent network with a simple architecture. Another contribution is the use of multimodal clusters as an additional information resource in the task of event extraction. The results of the application of these solutions are the following.

1. The neural network learns on n -grams no worse than in words, but it works faster.
2. The use of an additional resource in the form of a multidimensional formal context and multimodal clusters makes it possible to improve the interpretability of the results of the neural network.

In the future, an analytical explanation of the experimental results obtained in this work is necessary.

5. Acknowledgements

The reported study was funded by Russian Foundation of Basic Research, the research project № 19-07-01178 and RFBR and Tula Region research project № 19-47-710007.

6. References

- [1] Ganter, Bernhard; Stumme, Gerd; Wille, Rudolf, eds., *Formal Concept Analysis: Foundations and Applications*, Lecture Notes in Artificial Intelligence, No. 3626. (2005). Springer-Verlag. Berlin
- [2] W. Xiang and B. Wang, "A Survey of Event Extraction from Text." *IEEE Access*, vol. 7, pp. 173111-173137, 2019, doi: 10.1109/ACCESS.2019.2956831.
- [3] Carpineto, Claudio, Romano, Giovanni. *A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval*. *Machine Learning*. 24. 95-122. (1996). doi: 10.1007/BF00058654.
- [4] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading Wikipedia to answer open-domain questions, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1870–1879. doi:10.18653/v1/P17-1171.
- [5] Parakal E. G., Kuznetsov S. Intrinsicly Interpretable Document Classification via Concept Lattices, in: *Proceedings of the 10th International Workshop "What can FCA do for Artificial Intelligence?"* / Ed. by S. Kuznetsov, A. Napoli, S. Rudolph. Vol. 3233. CEUR Workshop Proceedings, 2022. Ch. 2. P. 9-22.
- [6] Sudha Rao, Daniel Marcu, Kevin Knight, Hal Daum'e III. Biomedical Event Extraction using Abstract Meaning Representation. *Proc. of the BioNLP 2017 workshop*, Canada, August 4, 2017.
- [7] Ananiadou, S. DeepEventMine: End-to-end Neural Nested Event Extraction from Biomedical Texts. *Bioinformatics* (2020). <https://doi.org/10.1093/bioinformatics/btaa540>
- [8] Cerf, L., Besson, J., Robardet, C., Boulicaut, J.F.: Closed Patterns Meet N-ary Relations. In: *ACM Trans. Knowl. Discov. Data*. 3, 1, Article 3, (2009). 36 p.

- [9] Ignatov D. I., Gnatyshak D. V., Kuznetsov, S. O., Mirkin, B., G.: Triadic Formal Concept Analysis and triclustering: searching for optimal patterns. *Mach. Learn.* 101:271–302. (2015).
- [10] The Stanford Question Answering Dataset. URL: <https://rajpurkar.github.io/SQuAD-explorer/>
- [11] The Kaggle Dataset Text Document Classification. URL: <https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification>.
- [12] Sowa J.: *Conceptual Graphs: Draft Proposed American National Standard*, International Conference on Conceptual Structures ICCS-99, Springer (1999).
- [13] Mikhail Bogatyrev, Dmitry Orlov. Framework for Pareto-Optimal Multimodal Clustering. In Sergei O. Kuznetsov, Amedeo Napoli, Sebastian Rudolph (Eds.): *The 10th International Workshop "What can FCA do for Artificial Intelligence?" FCA4AI 2022*, co-located with IJCAI-ECAI 2022, July 23 2022, Vienna, Austria. CEUR Workshop Proceedings. Vol. 3233. Pp. 51–62.