Using Polyadic Formal Contexts for Information Extraction from Natural Language Texts

Mikhail Bogatyrev [1][0000-0001-8477-6006] and Olga Mitrofanova[2][0000-0002-3008-5514]

[1] Tula State University, 92 Lenin ave., Tula, Russia
[2] Saint-Petersburg State University, 11 Univeritetskaya emb., Saint-Petersburg, Russia
okkambo@mail.ru, o.mitrofanova@spbu.ru

Abstract. The paper considers the use of elements of Formal Concept Analysis – multidimensional or polyadic formal contexts – to extract information from natural language texts. We propose the method for constructing polyadic formal contexts by means of Semantic Role Labeling and Abstract Meaning Representation (AMR) of texts. Using semantic role labeling, a conceptual graph is created for each sentence of the text, and a specific scheme of abstract meaning representation of the sentence is developed based on its elements. The polyadic formal context is a multidimensional tensor, whose points are elements of an AMR scheme. To extract information from a polyadic formal context, data associations as sub-contexts of the original context are built. Each such subcontext is associated with a specific element of the AMR scheme. Queries to associations return responses that preserve the meaning of the phrases according to the AMR scheme. The method was tested in the task of finding dependencies between texts on the corpus of abstracts of scientific articles on biomedical subjects of the PubMed system.

Keywords: Information retrieval, Polyadic formal context, Abstract Meaning Representation.

Introduction

The current state of the Computational Linguistics is characterized by the active involvement of mathematical methods of Data Analysis: methods of machine learning, algebraic methods, and methods of graph theory. Such synthesis is doubly useful. On the one hand, it allows in some cases to define objects used in computational linguistics in a new way and also to offer new solutions in Natural Language Processing. On the other hand, the applications of these methods in specific tasks enrich the methods themselves, opening up new areas of development in them. These observations prove to be true in relation to experimental material involved in our study. In it we apply the Formal Concept Analysis (FCA), a mathematically rigorous theory of conceptual modeling, and its main object, the formal context, which in some sense generalizes the concept of context in linguistics. In this paper we prove that clustering used in the Formal Concept Analysis (FCA-clustering) is ineffective in the tasks of extracting information from our specific formal contexts built on texts.

The paper proposes another approach to the clustering of formal context data, based on the construction of data associations with a specific AMR scheme. Information Extraction (IE) from data is effective when the data models used for this purpose are sufficiently informative by themselves. This is especially true for information extraction from natural language texts. To extract information from text data, a common scheme «model + resource» is used.

The model reflects the structure and parameters of the information retrieval target. Forms of models are various. This can be a lexical-grammatical template in a fact extraction problem, or a matrix or graph in procedures based on mathematical models. Linguistic resources are used to train models: text corpora, ontologies, and thesauri. The peculiarity of multidimensional formal contexts used in this work is that they can simultaneously be models of objects, for example, in thesauri development, and information resources in question answering systems. In this paper, formal contexts are constructed using an abstract meaning representation of the text. This ensures that they are informative as models: the semantics of AMR schemes are preserved in the model and in the query results it delivers.

The method developed in this paper is tested on the texts of the AGAC corpus (Active Gene Annotation Corpus) which contains abstracts of scientific articles on biomedical topics of the PubMed system [42]. The efficiency of our approach applied to the task of information extraction is due to the preservation of sentence semantics in a multidimensional formal context.

1 Formal Concept Analysis and Polyadic Formal Contexts

Formal Concept Analysis (FCA) [1] is mathematically rigorous theory which formalizes the notion of concept and studies how concepts may be hierarchically organized. FCA has been applied in many modern areas of knowledge discovery, machine learning and information retrieval [2]. There are also increasing number of FCA applications in text mining and linguistics, bioinformatics and medicine, software engineering and databases [3].

Briefly consider the main issues of the FCA. Classical FCA deals with two basic notions: *formal context* and *concept lattice*. Formal context is a triple $\mathbf{K} = (G, M, I)$ where G is a set of objects, M – set of their attributes, $I \subseteq G \times M$ – binary relation which represents facts of belonging attributes to objects. Formal context may be represented by [0, 1] - matrix $\mathbf{K} = \{k_{i,j}\}$ in which units mark correspondence between objects $g_i \in G$ and attributes $m_j \in M$. The concepts in the formal context have been determined by the following way. If for subsets of objects $A \subseteq G$ and attributes $B \subseteq M$ there are exist mappings (which may be a functions also) $A' : A \rightarrow B$ and $B' : B \rightarrow A$ with the properties of $A' := \{m \in M \,|< g, m >\in I \text{ for all } g \in A\}$ and $B' := \{g \in G \,|< g, m >\in I \text{ for all } m \in B\}$ then the pair (A, B) that $A' = B$, $B' = A$ is named as *formal concept*. The composition of mappings demonstrates following properties of A and

B: $A'' = A$, $B''=B$; A and B is called the *extent* and the *intent* of a formal context $\mathbf{K} = (G, M, I)$ respectively.

By other words, a formal concept is a pair (A, B) of subsets of objects and attributes which are connected so that every object in A has every attribute in B, for every object in G that is not in A, there is an attribute in B that the object does not have and for every attribute in M that is not in B, there is an object in A that does not have that attribute.

If for formal concepts (A_1, B_1) and (A_2, B_2), $\mathrm{A}_1 \sqsubseteq \mathrm{A}_2$ and $\mathrm{B}_2 \Subset \mathrm{B}_1$ then (A_1, B_1) ≤ (A_2, B_2) and formal concept (A_1, B_1) is less general than (A_2, B_2). This order is represented by *concept lattice.* A lattice consists of a partially ordered set in which every two elements have a unique *supremum* (also called a least upper bound or *join*) and a unique *infimum* (also called a greatest lower bound or *meet*).

1.1 Applications FCA in Text Mining and Linguistics

First of all, it is necessary to define correlation between the notions of formal and linguistic contexts, the former being substantiated in algebraic theories, the latter being part of language representations.

Linguistic theories provide a variety of context types. In general, a context is regarded as an obligatory condition for actualization of basic relations within a language system, namely, syntagmatic and paradigmatic relations considered on morphological, syntactic and semantic levels. On the one hand, there are approaches which take into account the scope and size of linguistic contexts. This view is characteristic for distributional semantics based on the assumption that semantic similarity of lexical items arises from their contextual similarity. This assumption is well-grounded in the works of L. Wittgenstein, Z. Harris, J. Firth, etc. who are considered to be the founders of this trend in contemporary linguistics (cf. the survey [5]). The ideas of distributional semantics lay the foundations of the rules governing collocability of lexical items in non-compositional phrases (cf. the analysis given in [6]). In various distributional semantic models (from the early word space models – HAL, LSA, COALS, etc. – to contemporary count-based and predictive models – Distributional Memory, Word2Vec, Doc2Vec, etc., cf. the overview of the works in [7, 8, 9]) context window size is a crucial parameter for vector space model development and word embeddings training. On the other hand, cognitive interpretation of contextual relations constitute a basis of theories focused on construction analysis (Construction Grammar, Cognitive Grammar, Corpus Pattern Analysis, etc. [10, 11, 12, 13]).

Inspite of external differences, particular contexts considered in linguistic theories can be generalized as instances of formal contexts. Let's consider a certain class of context relations described as verbal constructions, or valency frames [11, 14, 15] thoroughly described in lexical databases, such as VerbNet, FrameNet, etc. for English, Lexicograph, FrameBank for Russian. Verbal valency frames are commonly treated in terms of syntactic relations: type of governance in pairs «head verb + dependencies»: cf. V(*prove*) → NP(*hypothesis*), V(*prove*) → PP(*in experiments*); and argument structures (Rel – *prove*; Arg0 – subject/agent (*researcher*); Arg1 – object/patient (*hypothesis*); Arg_M – modifier (*in experiments*)).

In Abstract Meaning Representation theory, the given frame is considered as an AMR scheme, a unified structural representation of AMR schemata set being a formal context. In Formal Concept Analysis distribution of formal context elements and their features over texts is visualized as an attribute-value matrix similar to term-document matrix in count-based vector space models. Parallel treatment of the notion of context in linguistic and algebraic theories proves the possibility of consistent combinations of contextual semantic approaches (frame analysis and distributional semantics) with FCA and AMR.

Interpretability of the basic notions of FCA from linguistic point of view explains its effectiveness in a wide range of applications in Text Mining [3, 4]. Being a competitive approach to representation of contextual relations, FCA is used in verb frame extraction and clustering [16], structuring lexical resources (thesauri and formal ontologies) [17], ontology development [18, 19], social network analysis and studying social communities organization [20, 21], fact extraction [22], named entity recognition [23], text clustering [24], duplicate detection [25], recommendation systems [26, 27], etc. In most cases FCA forms an ensemble with traditional NLP techniques: morphosyntactic annotation of corpora, collocation analysis, keyword extraction, common clustering and classification algorithms, similarity measures. In recent decades researchers witness a strong tendency to consider FCA as a theoretical platform for experiments within the framework of machine learning.

1.2 Multimodal Clustering in FCA

Formal Concept Analysis may be defined as «the paradigm of conceptual modeling which studies how objects can be hierarchically grouped together according to their common attributes» [2]. Such grouping of objects is really clustering of them. More accurately, this is biclustering: clustering of two sets simultaneously, the set of objects and the set of attributes. The output of FCA algorithms is concept lattice which contains hierarchically linked formal concepts which are biclusters.

Among the advanced issues of FCA there is the study of multidimensional formal contexts which can be represented as n-ary relations $R \subseteq D_1 \times D_2 \times \ldots \times D_n$ on data domains $D_1, D_2, \ldots, D_n$. For n = 3 these domains have the meanings of «objects», «attributes» and «conditions» and FCA on formal contexts of this dimension has been distinguished as Triadic Formal Concept Analysis [29]. Multidimensional formal contexts also generate corresponding lattices of concepts. Practical applications of polyadic formal contexts in FCA are limited to two- and three-dimensional formal contexts. At the same time, the transition from dimension two to dimension three with the subsequent finding of formal concepts is not a simple scaling, but is associated with the introduction of additional operators and analysis tools [30]. However, already starting from dimension three, their construction is a much more complicated task than in the classical two-dimensional case. The three-dimensional version of FCA is best studied, which allows us to distinguish the Triadic Formal Concept Analysis as a separate area of FCA [29]. The subject of research here is multimodal, in this case, three-dimensional clusters – triclusters.

An important result was obtained here, consisting in the fact that every three-dimensional concept of a conceptual lattice belongs to some tricluster. According to multimodal clustering, for any dimension of formal context, the purpose of its processing is to find *n*-sets $H =< X_1, X_2, ..., X_n >$ which have the closure property [30]:

$$\forall u = (x_1, x_2, ..., x_n) \in X_1, X_2, ..., X_n, u \in R, \tag{1}$$

$\forall j = 1,2,...,n, \forall x_j \in D_j \setminus X_j < X_1,...,X_j \cup \{x_j\}, ..., X_n >$ does not satisfy (1). The sets $H =< X_1, X_2, ..., X_n >$ constitute *multimodal clusters*.

As two-dimensional biclusters are built formally, as for the dimension $n \geq 3$ clustering is performed with the use of various measures of proximity. Accordingly, the problem of interpretation of multimodal clusters in the context of the selected proximity measure arises.

2 Constructing Polyadic Formal Contexts on Natural Language Texts

The central notion of Formal Concept Analysis, the notion of formal concept seems very attractive for applying it in the areas where the term «concept» is used naturally. Natural Language Processing (NLP) is just that area. The cherished goal in the NLP is computerized understanding of texts. One a way of such understanding is using concepts being acquired from texts. Formal contexts potentially contain concepts but it is evident that expressiveness of standard two-dimensional formal contexts is not enough for modeling all peculiarities of natural language texts. So we apply multidimensional or polyadic formal contexts constructed on texts.

Consider in general the process of constructing polyadic formal contexts on natural language texts. It includes the following steps.

Establishing the problems to solve. Determining the range of tasks that the developed model is oriented towards in the form of a multidimensional formal context. In a general setting, these are the tasks of extracting information. They come down to extracting named entities from the text, extracting relationships, facts, and events. These may be the results of a query to a system that uses formal contexts.

Choosing a semantic text model. There should be an intermediate link between the text and the formal context the link as semantic model which is the data source for the formal context. As such a model, we chose the abstract-semantic representation of the text. To construct AMR-schemes, conceptual graphs are used.

Formal context construction. At this stage, it is necessary to choose the dimension of the context, the composition of the sets, $D_1, D_2, ..., D_n$ and build the relation $R \subseteq D_1 \times D_2 \times ... \times D_n$. The constructed multidimensional formal context should be implemented as a storage object, for example, in a database in such a way as to ensure work with context by executing queries to it.

2.1 Conceptual Modeling Text Semantics

We apply conceptual graphs (CGs) [31] for modeling text semantics. There are several methods of acquiring conceptual graphs from natural language texts [32, 33]. Among them, the method based on Semantic Role Labeling [34] is most suitable for building formal contexts. Some peculiarities of conceptual graphs created with this method, and examples of applications CGs in knowledge discovery are illustrated in [35].

Certain problems arise when using conceptual graphs as input to formal contexts. Among them there is the problem of *redundancy of conceptual graphs*. A conceptual graph acquired from quite a long sentence may contain many various semantic roles, and it is difficult to represent the variants of connections they specify in the formal context, even when the dimension of a context is greater than two. The solution to this problem is to aggregate conceptual graphs. An aggregated conceptual graph is a smaller graph that summarizes the information contained in the original graph [36].

The method of aggregation that we apply is based on the construction of an Abstract Meaning Representation (AMR) on each conceptual graph.

AMR [37] «*is a rooted, directed acyclic graph that captures the certain notion in text, in a way that sentences that have the same basic meaning often have the same AMR*». The nodes in the AMR graph map to words in the sentence and the edges map to relations between the words. This definition of AMR graph demonstrates the similarity AMR graphs and conceptual graphs.

Let's call the *AMR schema* a template $\boldsymbol{T}\ (C, S)$ where C is a set of concepts, S is a set of semantic roles, they both are from conceptual graph. Concrete content of AMR schema is defined by certain values (meanings) of C and S and it has a certain meaning too. For example, the template

$$\boldsymbol{T}(C, S) = <\mathit{Concept\_1}> \leftarrow (\textbf{“Agent”}) \leftarrow <\textbf{Verb}> \rightarrow (\textbf{“Patient”}) \rightarrow <\mathit{Concept\_2}> \tag{2}$$

specifies the AMR schema with the meaning «*who did what to whom*». The template (2) defines a conceptual graph in which «Agent» and «Patient» are the names of semantic roles, *Concept_1*, *Concept_2* – words being its concepts, «Verb» is concept-verb from conceptual graph.

Figure 1 demonstrates an example of interpreting AMR schema as sub graph of conceptual graph.

Conceptual graph on the Fig. 1 derives AMR schema «who did what to whom» with the content «*SHP-2 attenuates function*» according with the template (2). This AMR schema represents the meaning of the whole sentence and, certainly, represents it very broadly. Using AMR schemes, semantic compression of text sentences is performed.

There are two propositions which we can formulate based on the analysis of works [37 - 39] and the essence of conceptual graphs.

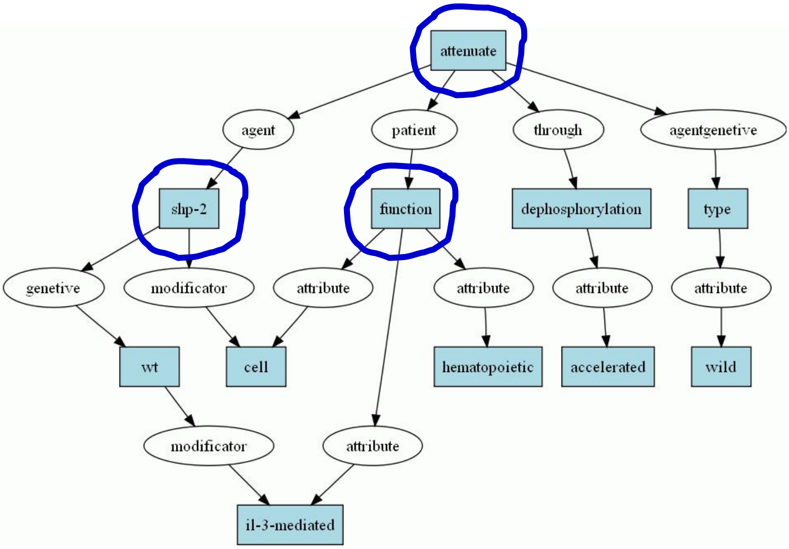


Fig. 1. Fragment of conceptual graph for the sentence «*SHP-2 attenuates IL-3-mediated hematopoietic cell function through accelerated dephosphorylation of STAT5*»

1. In many applications, particularly in the field of Bioinformatics, the expressiveness of AMR schemata is sufficient to represent the meaning of sentences.
2. Conceptual graphs allow implementing a variety of AMR schemata, including more complex ones that reflect the meaning of the sentence more fully.

Based on these propositions, consider the method for constructing formal contexts on a set of conceptual graphs.

2.2 Acquiring Polyadic Formal Contexts

The polyadic formal context is constructed as follows. By semantic role labeling for each sentence of the text, a conceptual graph is constructed, on the elements of which a concrete AMR scheme is created. The formal context $\mathbf{K} \subseteq D_1 \times D_2 \times ... \times D_n$ is a multidimensional tensor whose points are the elements of the AMR scheme for each representation, $k_{i,j,...,n} = \{c_i, c_j, ..., c_n\}$ where c_k, k = 1, 2,..., N are the concepts of the concept graphs, N is the total number of concepts obtained on the processed text. The number of points in the formal context matches the number of AMR schemata found in the text. The vast majority of points in a formal context are meaningful phrases, for example, the phrase «*SHP-2 attenuates function*» from figure 1 is a point <SHP-2, attach, function> in a three-dimensional context.

Query Support on Polyadic Formal Context. After creating a polyadic formal context, it is necessary to organize its storage and access to its content in order to solve the problems of extracting information. Information is extracted by querying a polyadic formal context.

There is the following idea concerned with Conceptual Modeling. If a query to a conceptual model can be represented as an element of this model itself − for example, as its concept, then the refinement of this query or even the answer to it is contained in concepts adjacent to the concept-query. This idea also holds for concept lattices and has been tested in several papers [40]. In general, selecting data that matches the query is a solution to the clustering problem. On polyadic formal contexts, solving the clustering problem requires determining the proximity measure for the points that make up the context. When using clustering algorithms used in FCA [29], a Boolean value is used as the proximity measure − the fact that clustering objects fall into a relation $R \subseteq D_1 \times D_2 \times ... \times D_n$ that sets the context. When using this measure in a context consisting of AMR schema points, clustering will result in subsets containing subsets of words $< X_1, X_2, ..., X_n >$ found according to the *R* relation. For example, the point considered in Figure 1 may appear in the following cluster of three points (the maximum number of elements is three; they are in the first subset):

<{SHP-2, val174del, ephrin-b2}, {attenuate, cause}, {function, dysplasia}>.

Although this cluster can be useful by demonstrating the relationship of objects from the first subset through words from the second and third subset, it is impossible to extract information from it, for example, about what exactly causes dysplasia. It turns out that individual elements of multidimensional formal context, its points, contain specific information in the form of an AMR scheme, but after processing the context this information is lost.

Thus, in order to extract information from multidimensional formal contexts based on AMR schemes, a different than FCA-clustering method is needed.

In our method, specific clusters − *associations* are built on formal contexts. An Association is a set of points ordered relative to the selected word position in the AMR scheme for a point. This corresponds to the logic of the AMR scheme: certain semantic elements are selected in it. The Association includes all words in the selected position of the AMR scheme. Therefore, an Association is a cluster built on the basis of the proximity measure «belong to a certain position of the AMR scheme». On the other hand, the Association is a function $A(x_1,\dots, x_p)$ whose argument can be a given word or a set of *p* words belonging to the *k*-th position of the AMR scheme.

The meaning of highlighting such associations is closely related to the logic of queries to the formal context. These queries usually correspond to the structures of the AMR charts.

3 Applications in Information Extraction

3.1 State of the Art

Let's discuss an example of applying FCA to data analysis in a specific area. One of the areas where NLP applications become more in demand is Bioinformatics. The Biomedical Natural Language Processing (BioNLP) [38] is the new area of research in Bioinformatics which appearance was due to the avalanche-like growth of publications in the field of biomedicine. The main purpose of BioNLP is to obtain new knowledge from published texts, not completely contained in each individual publication. Initially, the main area of application of BioNLP methods was genomic studies. Over time, the subject matter of texts processed by BioNLP has expanded to other areas and BioNLP was formed as a research area with its own data, tasks and methods [37-39]. All the BioNLP tasks may be classified as more or less general. The general tasks of fact extraction and event extraction usually transform to the standard tasks of Named Entity Recognition (NER) and Relation Extraction (RE). NER consists in automatically identifying occurrences of biological or medical terms in unstructured text. As named entities, there are the names of genes, proteins, living organisms or diseases – it depends on the domain to which processed text belongs to.

RE is another standard task of BioNLP. Relations are associations among biomedical entities. The simplest relations are binary, involving only the pair-wise associations between two entities. But biomedical relationships can involve more than just two entities. This kind of relationship is actual in the task of event extraction. In our time, named as genomic era, much of BioNLP work has focused on automatically extracting interactions between genes and proteins. Other associations include interactions between proteins and mutations, proteins and their binding sites, genes and diseases, genes and phenotypic context.

Leading BioNLP research groups are mainly interested in processing English data, although Russian biomedical texts attract growing attention.

Researchers collected and prepared for distribution a great amount of textual data. BioNLP competitions inspired creation of richly annotated corpora for NER, RE, Semantic Role Labeling (SRL), etc. The given empirical data is a great asset to computational linguists working in the field of Bioinformatics.

3.2 Experimental Data

Experiments aimed at the empirical verification of our approach were carried out for texts of the AGAC corpus (Active Gene Annotation Corpus) which contains abstracts of scientific articles on biomedical topics of the PubMed system. The corpus was created for BioNLP Shared Tasks 2019 competition [42] and was proposed as a dataset for NER and RE tasks. The corpus contains 250 annotated abstracts and 1000 raw abstracts, it size being about 300 000 tokens. Conceptual graphs were built for separate sentences from annotated abstracts; experiments with distributional semantic models were carried out for the whole dataset.

3.3 Finding Dependencies Between Texts

The problem of finding the relationships of texts is well known in the field of IE and has a variety of options. In our experiments, we studied a variant in which it is not known in advance by what attributes the links between texts are established. These attributes, which are ultimately reduced to subsets of words, are determined by analyzing the contents of texts, which in this case are replaced by a formal context built on them.

In our experiments, we compared the informativeness of two formal contexts built on texts in accordance with 3 and 5 element AMR schemes. The three-element scheme has the form (2), and the five-element AMR scheme has the following form:

$$\begin{tikzcd}
<\mathit{Concept\_1}> \arrow[d] & (\text{«}\mathbf{Agent}\text{»}) \arrow[l] & <\mathbf{Verb}> \arrow[l] \arrow[r] & (\text{«}\mathbf{Patient}\text{»}) \arrow[r] & <\mathit{Concept\_2}> \arrow[d] \\
(\text{«}\mathbf{Attribute}\text{»}) \arrow[d] & & & & (\text{«}\mathbf{Attribute}\text{»}) \arrow[d] \\
<\mathit{Concept\_3}> & & & & <\mathit{Concept\_4}>
\end{tikzcd} \tag{3}$$

An additional dimension was included in each context to fix the number of the text to which this point belongs. As a result, contexts of dimensions 4 and 6 were subject to processing.

Obviously, multi-element AMR schemata allow more detailed modeling of the semantics of a sentence. Formal contexts built on their basis are more informative. This position was checked in experiments. The experiments included the following steps.

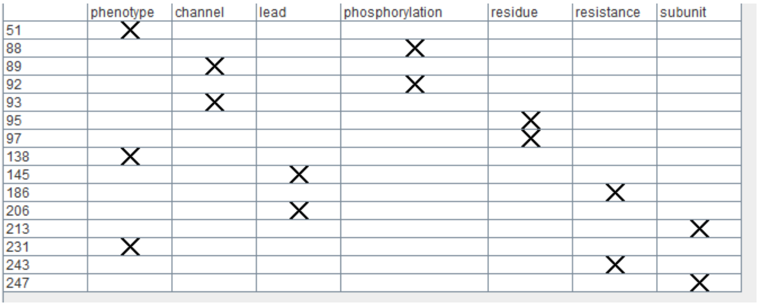
1. Building associations on selected positions of the AMR-scheme of the formal context.
2. Generating queries for associations based on query words
3. Obtaining query results in the form of clusters containing formal context points.
4. Interpretation of clusters.

Consider some experimental results. Associations were created on both formal contexts regarding the position of the subject of the action − the first position for the three-element AMR scheme and the second position for the five-element one. Next, the size and content of each association were estimated.

Domain terms were highlighted in the corpus. The term «mutation» has extensive connections, it organizes one of the most voluminous associations. Indeed, most of the texts of the corpus are devoted to the study of various manifestations of mutation and its influence on organisms. Therefore, our queries to associations were performed using the keyword «mutation». The results of the query are clusters. The question that determines further actions with the resulting clusters is: «What does the mutation manifest itself on?» The implementation of this request on clusters was carried out by building associations with respect to the position of the action object − the third position for the three-element AMR scheme and the fourth position for the five-element one. The query words obtained in the constructed associations were compared with text numbers and then presented for analysis.

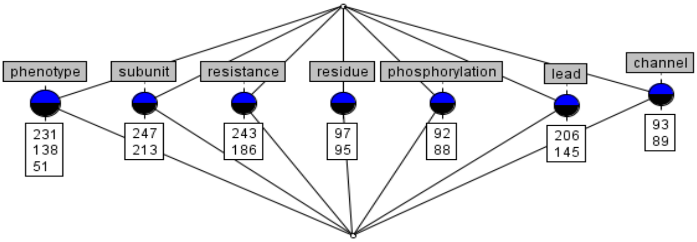
Responses to association requests are generated in tabular form. If the result of a two-element query to associations is presented as a cross-table, it is interpreted as a two-dimensional formal context. In this case, it can be visualized as a concept lattice according to the classical version of FCA.

Fig. 2 (a) shows the sub context as a cross-table of the four-dimensional formal context constructed for three-element AMR scheme (2), Fig. 2 (b) shows concept lattice.



	phenotype	channel	lead	phosphorylation	residue	resistance	subunit
51	X						
88				X			
89		X					
92				X			
93		X					
95					X		
97					X		
138	X						
145			X				
186						X	
206			X				
213							X
231	X						
243						X	
247							X

(a)



(b)

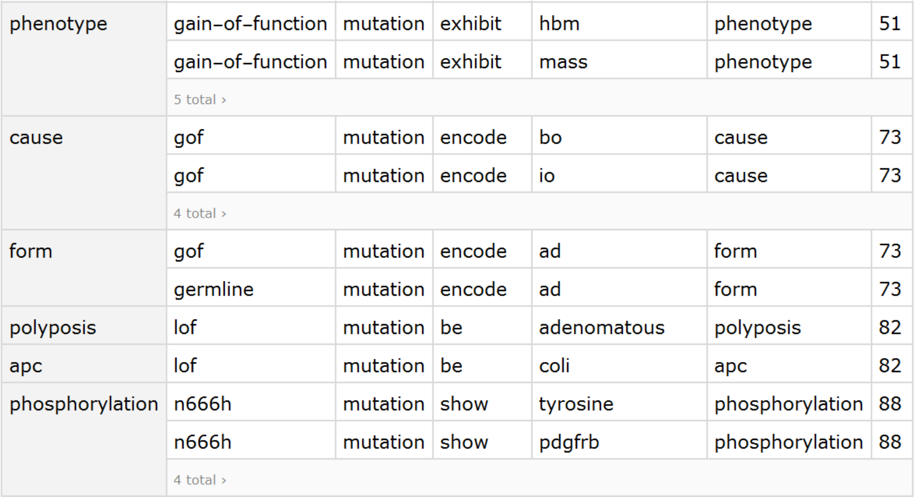
Fig. 2. The subcontext of the formal context built for the three-element AMR-scheme and its visualization in the form of concept lattice

The query that generates the result in Fig. 2, can be made in the form of «How are texts related in the context of the word «mutation» through its manifestations?» Texts with numbers in the left column of the sub context on Fig. 2 a) are linked in the context of the word «mutation» by means of the words indicated in grey rectangles in the concept lattice.

The lattice in Fig. 2 (b) is trivial. It has only one layer and all concepts are independent. The three-element AMR scheme does not reveal the connections of texts in sufficient detail. For comparison, the same request was processed on a formal context built for the five-element AMR scheme (3).

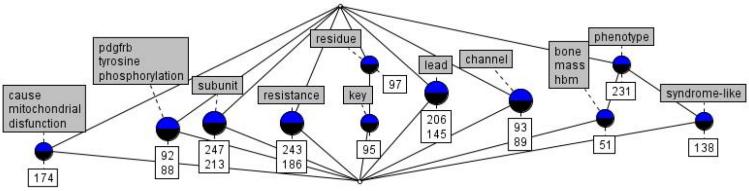
Fig. 3 shows the fragment of the association built on the «mutation» cluster for the fifth element of the five-element AMR scheme. The numbers of occurrences of certain words in context points are shown in the summary table in Fig. 3 a). So the word «phenotype» occurs in 5 points and in some documents, including document No 51.

Processing a query in a six-dimensional context reveals a larger number of words that link texts. The corresponding two-dimensional formal sub-context has a larger size and its concept lattice shown in Fig. 3 b), is not trivial: it has a hierarchy of concepts.



phenotype	gain–of–function	mutation	exhibit	hbm	phenotype	51
	gain–of–function	mutation	exhibit	mass	phenotype	51
	5 total ›					
cause	gof	mutation	encode	bo	cause	73
	gof	mutation	encode	io	cause	73
	4 total ›					
form	gof	mutation	encode	ad	form	73
	germline	mutation	encode	ad	form	73
polyposis	lof	mutation	be	adenomatous	polyposis	82
apc	lof	mutation	be	coli	apc	82
phosphorylation	n666h	mutation	show	tyrosine	phosphorylation	88
	n666h	mutation	show	pdgfrb	phosphorylation	88
	4 total ›					

a)



b)

Fig. 3. A fragment of association built on the «mutation» cluster the five-element AMR scheme and its concept lattice

Comparing the lattices in Fig.2 (b) and Fig. 3 (b), we see that, for example, in Fig. 2 (b) texts 231, 138, 51 are included in the same concept with the word «phenotype» combining them, and in the lattice in Fig. 3 these texts form three different concepts with a large number of unifying words.

At the same time, the concept that includes text 231 is more general for the concepts that include texts 51 and 138. The lattice in Fig. 3 b) can be named as "What is affected by mutation in different texts".

Based on these results, the following conclusions can be drawn.

1. The problem of finding dependencies between texts can be solved by clustering data of polyadic formal context using data associations.

2. The informativeness of a five-element AMR scheme is qualitatively higher than that of a three-element AMR scheme.

It is obvious that due to the universality of this text analysis tool, it can be used in various other tasks of relation extraction.

Conclusion

In this paper, we propose a method for constructing and applying polyadic formal contexts on natural language tests. The method uses conceptual graphs acquired from texts and, together with AMR-schemata, these graphs constitute a data source for polyadic formal contexts.

Polyadic formal contexts constructed by this way may be used as a tool for multimodal clustering. This tool was tested here on the problem of finding dependencies between texts.

It should be noted that the use of conceptual graphs makes it possible to construct AMR schemata of greater length than those considered in this paper. This will allow for the implementation of polyadic formal contexts that reflect the content of the modeled text more fully and, accordingly, to extract more complete information from it. The method can be applied in Question-answering systems, in which natural language queries correspond to the logic of AMR schemes.

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

References

1. Ganter, B., Stumme, G., Wille, R. (eds.): Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, № 3626, Springer-Verlag, Berlin, (2005).
2. Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal concept analysis in knowledge processing: A survey on models and techniques. In: Expert Systems with Applications, 40(16), pp. 6601-6623 (2013).
3. Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Formal concept analysis in knowledge processing: A survey on applications. In: Expert Systems with Applications, 40(16), pp. 6538-6560 (2013).
4. Ignatov, D.I.: Formal Concept Analysis: from Theory to Practice. In: AIST 2012 Proceedings, pp. 3-15. Ekaterinburg (2012).
5. Sahlgren, M.: The Distributional Hypothesis. In: Rivista di Linguistica. 20 (1), pp. 33-53 (2008).
6. Apresjan, Ju.D.: On the Rule of Lexical Meaning Composition. In: The Problems of Structural Linguistics 1971. Moscow, Nauka (1972) [In Russian].
7. Baroni, M., Dinu, G., Kruszewski, G.: Don’t Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In: 52nd Annual Meeting of

the Association for Computational Linguistics, ACL 2014, Proceedings of the Conference, vol. 1, pp 238-247 (2014).

8. Clark, S.: Vector Space Models of Lexical Meaning. In: Lappin, Sh., Fox, Ch. (eds.) The Handbook of Contemporary Semantic Theory, pp. 493-522. Blackwell Publishing, Ltd. (2015).
9. Rohde, D.L.T., Gonnerman, L.M., Plaut, D.C.: An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. In: Cognitive Science, vol. 8 (2006).
10. Goldberg, A.: Constructions at Work: The Nature of Generalization in Language. New York, Oxford University Press (2006).
11. Fillmore, Ch.J., Kay, P.: A Construction Grammar Coursebook. University of California, Berkeley (1995).
12. Hanks, P.: Corpus Pattern Analysis. In: Williams G., Vessier, S. (eds.) Proceedings of the XI Euralex International Congress, Lorient, Université de Bretagne-Sud (2004).
13. Rakhilina E. (ed.) Construction Linguistics. Moscow (2010) [In Russian].
14. Apresjan, Ju.D.: Integral Description of Language and Systemic Lexicography. In: Selected works, vol. 2. Moscow (1995) [In Russian].
15. Melchuk, I.A.: Experience in the Theory of Linguistic Models «Sense <=> Text». Moscow (1974/1999) [In Russian].
16. Falk, I., Gardent, C.: Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Grids to French Verbs. In: Napoli, A., Vychodil, V. (eds.): CLA 2011, pp. 223-238, INRIA Nancy Grand Est and LORIA (2011).
17. Priss, U.: Modeling lexical databases with formal concept analysis. In: Journal of Universal Computer Science, vol. 10(8), pp. 967-984 (2004).
18. Zhou, W., Liu, Z.T., Zhao, Y.: Ontology Learning by Clustering Based on Fuzzy Formal Concept Analysis. In: Proceedings of the 31st Annual International Computer Software and Applications Conference COMPSAC'07, vol. 1, pp. 204-210. IEEE Computer Society, Washington, DC, USA, (2007).
19. Gamallo, P., Lopes, G.P., Agustini, A.: Inducing Classes of Terms from Text. In: Matoušek V., Mautner P. (eds.) Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science, vol 4629, pp. 31-38. Springer, Berlin, Heidelberg (2007).
20. Kuznetsov, S.O., Ignatov, D.I.: Concept Stability for Constructing Taxonomies of Website Users. In: Proc. Satellite Workshop «Social Network Analysis and Conceptual Structures: Exploring Opportunities» at the 5th International Conference Formal Concept Analysis (ICFCA'07), pp. 19-24. Clermont-Ferrand, France (2007).
21. Roth, C., Obiedkov, S., Kourie, D.G.: On Succint Representation of Knowledge Community Taxonomies with Formal Concept Analysis. In: Int. J. of Foundations of Computer Science, vol. 19, № 2, pp. 383-404. World Scientific Publishing Company (2008).
22. Elzinga, P., Poelmans, J., Viaene, S., Dedene, G., Morsing, S.: Terrorist Threat Assessment with Formal Concept Analysis. In: Proc. IEEE International Conference on Intelligence and Security Informatics, pp. 77-82. Vancouver, Canada, 77-82 (2010).
23. Girault, T.: Concept Lattice Mining for Unsupervised Named Entity Annotation. In: Belohlavek, R., Kuznetsov, S.O. (eds.): Proc. CLA 2008, Palacký University, Olomouc, pp. 35-46 (2008).
24. Maille, N., Statler, I., Chaudron, L.: An Application of FCA to the Analysis of Aeronautical Incidents. In: Ganter, B. et al. (eds.): ICFCA, LNAI 3403, pp. 145-161. Springer (2005).
25. Ignatov, D.I., Kuznetsov, S.O. Frequent Itemset Mining for Clustering Near Duplicate Web Documents. In: Rudolph, S., Dau, F., Kuznetsov, S.O. (eds.) Proceedings of the 17th

International Conference on Conceptual Structures, ICCS 2009, LNCS (LNAI) 5662, pp. 185-200. Springer-Verlag Berlin Heidelberg (2009).

26. Ignatov, D.I., Kuznetsov, S.O. Concept-based Recommendations for Internet Advertisement. In: Belohlavek R., Sergei O. Kuznetsov, S.O. (eds.): Proceedings of The Sixth International Conference Concept Lattices and Their Applications (CLA'08), CLA2008, pp. 157-166. Palacky University, Olomouc (2008).
27. Ebner, M., Mühlburger, H., Schaffert, S., Schiefner, M., Reinhardt, W., Wheeler, S.: Getting Granular on Twitter: Tweets from a Conference and Their Limited Usefulness for Non-participants. In: IFIP Advances in Information and Communication Technology, vol. 324, pp. 102-113. Springer Berlin Heidelberg (2010).
28. Kaytoue, M., Kuznetsov, S.O., Macko, J., Napoli, A.: Biclustering Meets Triadic Concept Analysis. In: Annals of Mathematics and Artificial Intelligence, vol. 70, pp. 55-79. Springer Verlag, Germany (2014).
29. Ignatov, D.I., Gnatyshak, D.V., Kuznetsov, S.O., Mirkin, B.G.: Triadic Formal Concept Analysis and Triclustering: Searching for Optimal Patterns. In: Machine Learning, April, 2015, pp. 1-32 (2015).
30. 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, 36 p. (2009).
31. Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA (2000).
32. Hensman, S.: Construction of Conceptual Graph Representation of Texts. In: Proceedings of Student Research Workshop at HLT-NAACL, pp. 49-54. Boston (2004).
33. Bogatyrev, M.Y., Mitrofanova, O.A., Tuhtin, V.V.: Building Conceptual Graphs for Articles Abstracts in Digital Libraries. In: Proceedings of the Conceptual Structures Tool Interoperability Workshop (CS-TIW 2009) at 17th International Conference on Conceptual Structures (ICCS'09), pp. 50-57 (2009).
34. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. In: Computational Linguistics, 2002, vol. 28, pp. 245-288 (2002).
35. Bogatyrev, M.: Fact Extraction from Natural Language Texts with Conceptual Modeling. In: Communications in Computer and Information Science, vol. 706, pp. 89-102. Springer-Verlag (2017).
36. Chein, M., Mugnier, M.-L.: Graph-based Knowledge Representation. Computational Foundations of Conceptual Graphs. Springer-Verlag, London (2009).
37. Rao, S., Marcu, D., Knight, K., Daumé III, H.: Biomedical Event Extraction using Abstract Meaning Representation. In: Proceedings of the BioNLP 2017 workshop, Vancouver, Canada, Association for Computational Linguistics, pp. 126-135 (2017).
38. Cohen, K.B., Demner-Fushman, D.: Biomedical Natural Language Processing. John Benjamins Publishing Company, Philadelphia (2014).
39. Simpson, M.S., Demner-Fushman, D.: Biomedical Text Mining: A Survey of Recent Progress. In: Aggarwal, Ch.C., Zhai, Ch.X. (eds.): Mining Text Data. Springer (2011).
40. Carpineto, C., Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. In: ACM Computing Surveys, 44(1), Article 1 (January 2012), 50 p. (2012).
41. Henriques, R, Madeira, S.C.: Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey. In: ACM Computing Surveys, 51(5), pp. 1-43 (2019).
42. BioNLP Open Shared Tasks (BioNLP-OST), https://2019.bionlp-ost.org/home, last accessed 2020/05/15.