Application of neural network language models based on distributive semantics for ontological modeling of the domain

M G Shishaev\textsuperscript{1,2}, V V Dikovitsky\textsuperscript{1} and V K Pimeshkov\textsuperscript{1}

\textsuperscript{1}Institute for Informatics and Mathematical Modeling, Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia
\textsuperscript{2}Murmansk Arctic State University, Murmansk, Russia

E-mail: shishaev@iimm.ru, dikovitsky@iimm.ru, pimeshkov@iimm.ru

Abstract. The article discusses the technology of automated formation of SKOS-ontologies for semantic modeling of the subject area, based on natural language texts analysis. The technology is based on neural network and distributive (vector) language models. A brief description of the content and formulation of the problem of extracting concepts and relations from natural language texts is given, the results of constructing a neural network classifier of SKOS relations using the Glove vector model, as well as an example of using the technology to construct a fragment of an applied SKOS ontology are given.

1. Introduction
Semantic (ontological) modeling of subject areas is one of the key tasks in the context of the implementation of the principles of "Industry 4.0" - a knowledge-based industry [1]. The main task of semantic models of subject areas in this context is to provide a unified, unambiguously interpreted conceptual base for complex heterogeneous components of production and technological systems. Examples of projects aimed at solving this problem by standardizing the names and identification of goods and services are the ECLASS electronic classifier [2], digital platform "Techexpert" [3], as well as a variety of standardized classifiers (OKVED, OKPD, etc.) [4]. However, in the design and manufacture of complex products from heterogeneous components, the problem of integrating data from different catalogs arises, as a result of which the data integrity may be compromised [5]. At the same time, an intensive expansion of the range of high-tech products is predicted in the coming decades [6]. This further actualizes the task of creating semantic models connecting various data sources by fixing (standardizing) the sets of concepts used and the relationships between them. The formation of such models is a labor intensive task, often the complexity is aggravated by the dynamism of ideas about the subject area (new concepts and connections appear, the structure of known ones changes, including when the subject areas are combined). A promising approach to solving this problem is the automation of the formation and populating of ontological models of the subject area using the methods of semantic analysis of natural language texts to extract concepts and relations.

The process of semantic analysis of a text \( T \) consists in its interpretation into a certain target semantic (ontological) model \( S \) and includes two stages: (1) the formation of a text model \( M(T) \), reflecting its properties that are significant for extracting the concepts of relations, and (2) the formation of a subset of the target semantic the model following from the given text:

\[ T \rightarrow M(T) \rightarrow S(M(T)). \]

In this work, distributive and lexico-syntactic text models are used as a model \( M(T) \), and the SKOS model is used as a target model. The main research question: Is it possible to efficiently extract facts
(concepts related by SKOS relations) from natural language texts based on the distribution models of
the latter? Despite the fact that the question is "what semantics in general can distributive models
give?" - has been studied for a long time (see, for example, [7]), the problem of extracting relations
that reflect the semantics of the subject area under consideration remains relevant.

2. The problem of concepts and relations extraction from natural language texts

2.1. Representation of the semantic model of a domain
Formalized knowledge provides machine reasoning and makes the basis for a variety of knowledge-
aware applications. The most actively used in recent years is the network model of knowledge
representation. Vertices in the network model correspond to concepts or entities and edges represent
relations between them. In general, the network model of knowledge can be represented in the form of
a set of triples specifying facts or true statements of the following form:
\[ \langle c_i, r^k, c_j \rangle \], where \( c_i, c_j \) – concepts or entities, \( r^k \) – some semantic relationship (not necessarily
symmetric) between them.

Note that the role of a semantic relation can be played by various links between concepts - the
attribution of a relation to the category of semantic is determined by its applicability for practical
problems that the model being formed is focused on. As a rule, the semantic category includes
relations that define a taxonomy or classification of entities, such as "equivalence", "class-subclass",
"part-whole", as well as a wide variety of domain-specific relationships that establish a logical system
suitable for machine inference in the context of the considering subject area.

In modern researches and applications, such network structures, with the advent of the appropriate
Google’s project [8], are often referred to as "knowledge graphs". The modern problems of
constructing and using knowledge graphs are presented in many reviews [9, 10]. The knowledge graph
(KG) is, in fact, an information management system that integrates all types of information and uses
graphs to describe data and knowledge. Initially, knowledge graphs were intended to improve the
efficiency of information search, but in recent years they have become increasingly used as a means of
semantic integration of heterogeneous data. [11, 12].

In general, a KG is a multigraph, where concepts (graph vertices) can be linked by relations of
arbitrary types. However, for practical use (first of all, effective machine reasoning), it is necessary to
limit the permissible set of types of relations between concepts. The imposition of restrictions on the
type of links, thus, defines the subspecies of knowledge graphs, for example, taxonomy (concepts are
linked by one or another hierarchical relationship) or thesauri (establish paradigmatic relationships
between concepts).

Technically, the W3C-standarized language RDF is most often used to represent the above triples
[13]. By itself, the RDF model specifies the minimum semantics, since it does not impose restrictions
on the allowed domains of concepts and relations. RDF provides a means for building information
models but does not deal with the semantics of what is described. Taken separately, an RDF graph can
only be understood as a graph. The interpretation of the meaning is based on the ability of RDF users
to interpret individual URIs, string literals and graph structure, and from them interpret the rest of the
URIs and data semantics [14].

Various knowledge representation models are used to describe the semantics of the domain. One of
the most common models is SKOS [15], also standardized by the W3C consortium and widely used in
various, including engineering, applications. The SKOS model includes three types of relationships —
'skos:broadser', 'skos:narrower' and 'skos:related' — making it an effective means of representing
taxonomies. The elements of SKOS are classes and their properties, which makes SKOS an efficient
means of conveniently and intuitively describing the structure of a domain.

2.2. Concepts and relationships extraction from natural language texts
There are many methods and approaches to automatic semantic analysis of natural language texts [16].
Within the framework of extracting knowledge from natural language texts, two interrelated tasks are
distinguished - (1) discovering meaningful concepts of the subject area and (2) identifying the
presence and type of relations between them.
The task of the automated extraction of concepts from texts (terminology extraction, automatic term recognition) is to find in the text lexical constructions denoting concepts that claim to be included in the ontological model of the subject area. Approaches to solving the problem, depending on the nature of the used features of lexical structures, are usually divided into linguistic and statistical, in practice, mixed approaches are also used. [17, 18]. For terminology extraction, both supervised machine learning methods based on labelled data and unsupervised learning are used. The latter is more preferable, but in general, they are inferior in efficiency (speed and accuracy of work) to methods using labelled data. The creation of methods for extracting terms using unsupervised learning models is one of the topical contemporary issues. For example, in [19] the authors propose unsupervised technology for extracting concepts, while considering the context of using the concept in terms of parts of speech (POS), which makes it possible to make a model that is universal for different subject areas.

In general, to extract meaningful concepts, two key questions to be answered: 1) what linguistic construction (word or phrase) denotes the concept and 2) whether this concept is significant in the context of the subject area under consideration. Within the second task, it is also necessary to separate specific (subject) and commonly used concepts. As a rule, the answer to the first question is sought by solving the problem of identifying noun phrases, including degenerate ones, consisting of one word. For this, there are a fairly well-developed methods based on the lexical and syntactic analysis of the text. One of the approaches based on the construction of a dependencies tree (including the one used in this work) is the use of the SyntaxNet parser based on artificial neural network [20]. In the practical part of this work, words filtered by lexical and syntactic features from an expertly selected fragment of a subject-specific text were used as meaningful concepts (see Section 3.2).

The task of relations extraction can be formulated in different ways, depending on the nature and method of using the available a priori information. The most general form is the task of identifying relationships - searching in the text for subsets of pairs of words (generally, lexical constructions) related by similar relationships. In this case, the solution is based on clustering a set of samples in a certain feature space with a metric which reflects a relational similarity [21]. Modern approaches use the representation of pairs of words in a vector space (embeddings), the metric of which is interpreted as the semantic similarity of relations between pairs [7, 22]. Despite the fact that this approach does not require preliminary labelling of samples for clustering (we are dealing with unsupervised learning) followingly, the problem of identifying the resulting clusters with one or another subject relation arises (in fact, this is an interpretation of the obtained relations into the target semantic model, which must be given a priori). In [11], for example, the problem of identifying the relationships between chemicals, genes and phenotypes is been solving, while to identify the type of subject relationships, existing manually created databases and thesauri defining subject concepts and relationships are used.

In this work, we a priori set the type of the relations to be extracted (these are the SKOS relations) and interpret the problem as a two-stage pattern recognition: (1) extraction of samples characterized by some features from the available text data, and (2) classification of these samples on the set of target relations:

1) Formation on the basis of available texts of a set of samples for classification, $O$, consisting of elements of the form:
\[ o_{ij} = \langle l_i, l_j, P^{ij} \rangle, \]
where $l_i, l_j$ – lexical units, which can be words, stable phrases, lemmas, synsets and other constructions that potentially denote a concept; $P^{ij}$ – a feature vector which components are both the individual properties of lexical units $l_i, l_j$ (for example, part of speech, frequency of use in a text, etc.), and their joint properties (for example, frequency of joint use in a text, cosine distance in vector space, etc.).

2) Formation of a training sample (dataset) in the form of a set of labeled samples of the form $\langle o_{ij}, r^k \rangle$, where $o_{ij} \in O$, $r^k$ – is the relationship between the concepts represented in the text by the lexical units $l_i, l_j$.

3) Classifier training.

4) Using the model to extract facts and assertions in the form of triples: $\langle c_i, r^k, c_j \rangle$.

Schematically, the process is shown in Fig. 1.
3. Experiments for extracting SKOS-relations from NL-texts using a neural network classifier

3.1. Dataset formation and training the network

To form the dataset, a labelled set of SKOS concepts from the DBpedia resource was used [14]. Relationships are represented by two classes: 'related' and 'broader' according to the SKOS classification. Only one-word concepts were included in the dataset. The original dataset was significantly disbalanced - elements of the 'broader' class prevailed. To obtain a balanced sample, the data was augmented by generating an artificial class 'empty' from concepts that have no relationship within the SKOS framework. An artificial class was formed by randomly combining concepts from the original set (related by the broader relationship) and then checking for not being in the broader or related class. The size of the resulting initial set of samples was 2,550,380 records.

Further, the samples were supplemented with labels of parts of speech, vectors (embeddings), morphological properties, and a cosine distance. The concepts have been tokenized with spaCy using the 'en_core_web_lg (3.0.0)' model. The dataset creation scheme is shown in Figure 2.

Tokens were used to take vectors of individual tokens and then transform them into a concept vector. The concept vector was taken as the average vector over all obtained vectors for the tokens of this concept. If at least one token of two concepts did not have a corresponding vector in the vector model, the corresponding record was not included in the sample set.

Figure 2. Generalized scheme for creating a dataset

In accordance with the described scheme, several datasets were formed using various models of vector representation of words. The following vector models were used:
• "Classic" W2V model of Google: 'GoogleNews-vectors-negative300'. Model trained on a portion of the Google News dataset. The model contains 300-dimensional vectors for 3 million words and phrases.

• Two models from the GloVe project [23]:
  o glove.42B. Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB);
  o glove.6B. Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 300d vectors, 822 MB);

• 5 models from the NLPL word embeddings repository [24].

On the obtained datasets, various versions of the feed-forward ANN with fully connected layers were trained. The best results were shown by a neural network based on the biggest vector model - 'glove.42B. Common Crawl'. The network architecture included 5 layers with a 631-500-350-250-150-3 funnel, the ReLU activation function was used, and training was carried out in 15 epochs. The results of testing the resulting model on verification data are shown in Figure 3.

Thus, a neural network classifier was obtained that associates an input pair of concepts with a confidence score of finding them in relation of 'broader', 'related' or 'empty'. Such a classifier can be
used both for the formation of probabilistic semantic networks (knowledge graphs) [25], when semantic connections have a weight corresponding to the probability of the existence of a relationship between concepts and for deterministic knowledge graphs. In the latter case, it is necessary to establish the minimum probability threshold at which the corresponding connection is included in the graph. The ability of the classifier to extract the SKOS relation 'skos: broader' makes it useful for building taxonomies and classifications. It should be noted, however, that the classifier does not provide complete automation of the taxonomy formation process, since the 'skos: broader' relation can be interpreted in various ways and denote, in addition to hyper-/hyponymy, which define the hierarchy of classes of concepts, also other semantic relations, in particular, meronymy ("part-whole") [26]. Nevertheless, the formation of many candidate triples for inclusion in the final taxonomy greatly facilitates the task of experts and allows, in particular, to address the problem of the dynamism of a natural language, when over time the same concepts receive new lexical designations.

3.2. Using the classifier to identify SKOS relationships within a subject area

To evaluate the efficiency of the obtained neural network classifier, a set of candidate relations was formed for inclusion in the SKOS ontology of concepts describing the field of mechanics of hydraulic motors, and the accuracy of the classifier operation was expertly evaluated.

The set of candidate concepts were formed on the basis of the Wikipedia dictionary entry on hydraulic motors [27]. Embeddings of candidate words were formed using the 'glove.42B.CommonCrawl' vectorization library. Pairs of words were formed on the basis of the syntactic tree of each sentence, using a filter for syntactic roles and morphology (words in a pair differ; they are 3 or more characters long; they are not prepositions, punctuation, pronouns, defining words). A formal record of the filter:

\[(token_{text}! = token_{head}) \& (\text{len}(token_{head}) > 2) \& (token_{dep} != "det") \& (token_{dep} != "prep") \& (token_{dep} != "pobj") \& (token_{pos} = "PUNCT") \& (token_{pos} = "DET")\]

The SpaCy library was used for the lemmatisation of words. Words missing in a Glove vector model were skipped. The output vector consists of 3 real numbers \(<related; broader; none>\), indicating the probability of the existence of a corresponding SKOS relationship between a pair of input words. The number of word pairs that fall into the set of candidates to include in a taxonomy depends on the used threshold of having a 'skos:broader' relationship. Thus, for a threshold of 0.4, 76 pairs were obtained from an initial set of 350 words, classified as being in the 'skos:broader' relationship.

The resulting pairs were subjected to expert evaluation, during which fully correct relationships were marked ('Exact matching'), as well as relationships, the correctness of which depends on the context of the potential use of the taxonomy and requires clarification by an expert ('Soft matching'). Depending on the threshold value of the estimation of the 'skos:broader' relationship existence, the classification accuracy reaches 67% for 'Exact matching' and 100% for 'Soft matching'. The completeness of the classification in this case, as expected, decreases with an increase in the threshold (to assess the completeness, the number of relations labelled with a probability of 0.4 and higher was taken as a reference value).

| Table 1. Accuracy and Completeness of SKOS Relations Extraction |
|---------------------------------------------------------------|
| **Threshold** | 0.4 | 0.7 | 0.9 | 0.95 | 0.99 |
| **Exact matching** | **Accuracy** | 53% | 64% | 62% | 57% | 67% |
| | **Completeness** | 100% | 44% | 25% | 13% | 6% |
| **Soft matching** | **Accuracy** | 70% | 86% | 85% | 86% | 100% |
| | **Completeness** | 100% | 45% | 26% | 14% | 7% |
In the considered example, the classifier does not provide 100% accuracy and, therefore, cannot be used for fully automatic construction of the SKOS ontology. Nevertheless, the automatic selection of a set of candidate relations from the texts greatly facilitates the task of constructing an ontology by experts. A fragment of the SKOS ontology, built from the results of the experiment, is shown in figure 4.

![Figure 4. The resulting fragment of an SKOS ontology](image)

4. Conclusion
Obtained results allow us to conclude that, in general, the use of neural network models of distributive semantics for the construction of semantic models of subject areas in the form of SKOS ontologies is promising. The experiments carried out show that vector models of general vocabulary contain the SKOS semantics and, with sufficient size the model, make it possible to build on their basis neural network classifiers that provide the extraction of SKOS relations from natural language texts.

The proposed technology, due to the non 100% classification accuracy, does not provide full automation of the SKOS ontology construction, however, it can be used as an additional tool for experts, which significantly reduces the labour intensity of the process of ontology construction and populating. An expert assessment of the performance and efficiency of the technology on the example of the subject area 'mechanics of hydraulic motors' gave good results in terms of the accuracy of recognition of SKOS relations.

One of the directions for further development of the technology is the expansion of the feature vector with components that more fully reflect the lexico-syntactic context of the joint use of concepts. The promise of this approach has been demonstrated by many studies and applied developments [11, 28]. At the current stage of this work, the syntactic dependency between concepts was taken into account indirectly, by including in the set of samples only those lexical units that are simultaneously present in the natural language sentence. The explicit inclusion of syntactic dependence between the lexical units denoting concepts in the feature vector will allow taking into account the specifics of the
use of concepts in the context of the considered subject area and, potentially, increasing the efficiency of recognition of relations.

5. References

[1] Yahya M, Breslin J G and Ali M I 2021 Semantic Web and Knowledge Graphs for Industry 4.0 Applied Sciences vol. 11 no. 11 Art. no 11, Jan. doi: 10.3390/app11115110

[2] ECLASS: Introduction, Jun. 30, 2021. https://www.eclass.eu/en/standard/introduction.html (accessed Jun. 30, 2021)

[3] Cifrovara platforma «Tekhekspert» - Tekhekspert, Jun. 30, 2021. https://cntd.ru/techexpert?utm_source=cntd (accessed Jun. 30, 2021).

[4] Klassifikatory - docs.cntd.ru, 2021. http://docs.cntd.ru/classifications (accessed Jun. 30, 2021).

[5] Fimmers C et al. 2019 An Industry 4.0 Engineering Workflow Approach: From Product Catalogs to Product Instances IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society vol 1 pp 2922–2927 doi: 10.1109/IECON.2019.8926979.

[6] Ivanov V V and Malineckij G G 2010 Bol’shie vyzovy XXI veka Innovacii no 2 pp. 3–14 2020

[7] Turney P D and Pantel P 2010 From Frequency to Meaning: Vector Space Models of Semantics, jair, vol. 37, pp 141–188, Feb. 2010, doi: 10.1613/jair.2934.

[8] Singhal A Introducing the Knowledge Graph: things, not strings, Google Official Blog, May 16, 2012. https://blog.google/products/search/introducing/using-knowledge-graph-things-not/ (accessed Jun. 24, 2021).

[9] Gurin V S, Kostrov E V, Gavrlenko Y Y, Saada D F, Ilyushin E A and Chizhov I V 2019 Knowledge Graph Essentials and Key Technologies’, International Scientific Journal “Modern Information Technology and IT Education vol. 15, no 4 Art. no. 4 doi: 10.25559/SITITO.15.201904.932-944.

[10] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu 2021 A Survey on Knowledge Graphs: Representation, Acquisition and Applications IEEE Trans. Neural Netw. Learning Syst pp. 1–21 doi: 10.1109/TNNLS.2021.3070843.

[11] B. Percha and R. Altman 2018 A global network of biomedical relationships derived from text’, Bioinformatics (Oxford, England), vol. 34 doi: 10.1093/bioinformatics/bty114.

[12] Y. Tang, T. Liu, G. Liu, J. Li, R. Dai, and C. Yuan 2019 Enhancement of Power Equipment Management Using Knowledge Graph IEEE Innovative Smart Grid Technologies pp 905–910 doi: 10.1109/ISGT-Asia.2019.8881348.

[13] ‘RDF Primer’, World Wide Web Consortium (W3C). https://www.w3.org/TR/rdf-primer/ (accessed Jul. 02, 2021).

[14] ‘Resource Description Framework’, Википедия. May 25, 2021. Accessed: Jul. 21, 2021. [Online]. Available: https://ru.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=114422665

[15] Ameri F, Yoder R, and Zandbiglari K 2020 SKOS Tool: A Tool for Creating Knowledge Graphs to Support Semantic Text Classification Advances in Production Management Systems. Towards Smart and Digital Manufacturing, Cham pp 263–271 doi: 10.1007/978-3-030-57997-5_31.

[16] Shishavae M G 2020 Nejrosetevye modeli v zadachah semantichestkogo analiza tekstov na estestvennom yazyke Trudy Kol’skogo nauchnogo centra RAN vol. 11, no. 8 (11) pp 91–00 doi: 10.37614/2307-5252.2020.8.11.008.

[17] Astrakhan'tsev N A, Fedoreenko D G and Turdakov D Yu 2015 Methods for automatic term recognition in domain-specific text collections: A survey Programming and Computer Software vol. 41 no. 6 pp 336–349 doi: 10.1134/S036176881506002X.

[18] Pazienza M T, Pennacchiotti M and Zanzotto F M 2005 Terminology extraction: an analysis of linguistic and statistical approaches Knowledge mining (Berlin, Heidelberg: Springer) pp 255–279

[19] Shao W, Hua B, Ma Q, Liu J, He H and Chen K 2020 A Unsupervised Method for Terminology Extraction from Scientific Text

8
[20] Weiss D and Petrov S 2017 An Upgrade to SyntaxNet, New Models and a Parsing Competition Google AI Blog URL:http://ai.googleblog.com/2017/03/an-upgrade-to-syntaxnet-new-models-and.html (accessed Dec. 08, 2020).
[21] Turney P D 2006 Similarity of Semantic Relations Computational Linguistics vol. 32, no. 3, pp 379–416 doi: 10.1162/coli.2006.32.3.379.
[22] Washio K and Kato T 2018 Neural Latent Relational Analysis to Capture Lexical Semantic Relations in a Vector Space Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium) pp 594–600 doi: 10.18653/v1/D18-1058.
[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, ‘GloVe: Global Vectors for Word Representation’. https://nlp.stanford.edu/projects/glove/ (accessed Jul. 23, 2021).
[24] ‘NLPL word embeddings repository’, Vector explorers. http://vectors.nlpl.eu/repository/ (accessed Jul. 22, 2021).
[25] X. Chen, M. Chen, W. Shi, Y. Sun, and C. Zaniolo 2018 Embedding Uncertain Knowledge Graphs
[26] Revenko A and Mireles V 2017 Discrimination of Word Senses with Hypernyms Proceedings of the 5th International Workshop on Linked Data for Information Extraction (LD4IE) (Aachen) pp 50–61 [Online]. Available: http://ceur-ws.org/Vol-1946/#paper-06
[27] ‘Hydraulic motor’, Wikipedia. Jun. 07, 2021. Accessed: Jul. 08, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Hydraulic_motor&oldid=1027427666
[28] Peirsman Y, Kris H and Speelman D 2007 Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models p 16

Acknowledgments
The work is supported by the Russian Foundation for Basic Research (RFBR), project number 20-07-00754.