Using Semantic Pattern Method for Evaluating Foreign Language Texts of Mineral Resources Profile

Yuliya Murzo

DOI: 10.18355/XL.2025.18.02.06

Abstract
The purpose of the article is to develop and evaluate the effectiveness of an algorithm for finding and isolating semantic text patterns to automate online searches and assess professionally oriented training texts used in the education process at foreign language departments in mineral resources universities. The paper presents an algorithm for text analysis with identification of domain semantic patterns, classification of texts relative to industrial subdomains, calculation of text proximity metrics to industrial subdomains and their assessment according to the degree of this proximity.
Key words: pattern, semantic web, artificial intelligence, professional communication, foreign languages, term glossary, mineral resources

1. Introduction

This paper is a continuation of a research presented in paper (Murzo et al., 2022) that analyses the possibility to use methods of web search and identification of semantic text patterns with a view to automatising the web search and assessment of professionally oriented texts used to train students at foreign language departments of higher education institutions of mineral resource development profile. The presented approach assumed the use of both Natural Language Processing (NLP) technology (Manning & Schütze, 1999) and Semantic Web (Hitzler et al., 2009) as part of the SWBE (Semantic Web Based Education) approach (Devedžić, 2006) for the analysis of scientific and research articles in the mineral resources domain. The presented approach makes it possible to build automated applications that will be able to automatically look for relevant text documents such as scientific articles, reviews, research results, etc. The paper has shown that a task solved by such applications can be decomposed into the following components:

- formation of domain templates for the studied texts,
- classification of texts according to domain templates,
- calculation of the degree to which the studied text is in proximity to a domain template,
- deciding whether to include a text into a training program based on the criteria of proximity to a domain template.

This paper is a continuation of a research presented in the paper (Murzo et al., 2022), namely it represents a practical implementation and assessment of the automated application algorithm to search and assess scientific and training texts from the point of view of their proximity to domain and subdomain templates. The developed prototype of the information system implementing the text analysis algorithm is presented as a case study and the results of its use together with the developed domain templates and training texts are given. The work considers the historical context of obtaining profile engineering education (Rudnik at al., 2023), branch peculiarities of mineral resources universities (Litvinenko et al., 2022), (Boyko at al., 2022), (Sveshnikova at al., 2022), peculiarities of educational tracks (Kretschmann et al., 2020), experience of using modern approaches to the development of professional foreign language and information technology courses (Pushmina et al., 2021), (Nikonova et al., 2023), as well as general approaches to assessing the applicability of

information technologies in the mineral resources sector and best practices for their use (Samylovskaya et al., 2022), (Koteleva et al., 2021), (Cherepovitsyn et al., 2021).

In this paper, the section "Materials and Methods" feature: Associated papers analysis, Task setting, Justification of selection of text preprocessing algorithms, Algorithm for text classification and metrics calculation of text proximity to domain templates, Case study. The "Conclusion" section provides a retrospect analysis of all the work and outlines the tasks to continue the research.

2. Literature review

The task of text processing with subsequent recognition of named entities (facts, real object, terms, definitions etc. in form of natural language) and submitting the results to the knowledge graph of the University Educational Knowledge Base (a university knowledge base that contains all digital education content used in the teaching and learning processes) requires, first, processing of the text with a word processor and, second, transferring the result into the knowledge graph according to the Teaching Ontology (Murzo et al., 2022).

Text Processing Intelligent text processing (ITP) is suitable to be used for text processing tasks in information systems that use semantic approach. ITP is a subset of artificial intelligence (AI) that uses Natural Language Processing (NLP) and machine learning (ML) algorithms to extract information from unstructured data. NLP – is a sphere of artificial intelligence dealing with interaction of computers and human language. Machine learning algorithms study huge volumes of data to recognize regularities and links between different elements. When applied to text data, ITP can extract important information, such as entities, moods, and themes. This extracted information may then be used for various applications, such as text classification and mood analysis. Natural Language Processing can be structured differently, using different models of machine learning, depending on what is being analyzed. Some of the key algorithms and analysis used in NLP include:

- Syntactic analysis that uses the structure of a sentence and rules of formal grammar to determine the exact meaning of a text (Cereda et al., 2018), (Aziz et al., 2015).
- Mood analysis that measures the tone and intentions is useful for analysis of messages in social media and online reviews (Kannappan et al., 2023), (Wankhade et al., 2022).
- Semantic analysis that helps the machine to learn about less literal meaning of words and phrases (Maulud et al., 2021), (Rodriguez-Cardos et al., 2019).
- Lexicons which are lists of words and emotions they evoke (Hirschberg & Manning, 2015), (Kannan et al., 2016).
- Elimination of ambiguity of a word meaning used in computer linguistics to determine which meaning is being used (Mente et al., 2022), (AlMousa et al., 2022).
- Lemmatization that determines the context of a sentence and refers to the basic word form as it appears in the dictionary (Akhmetov et al., 2020).
- Stemming that analyzes the ending or the beginning of a word to determine the common stem form of the word (Khurana et al., 2022), (Alshalabi et al., 2022).
- Summing that is a brief review of large volumes of text data and usually used to generalize news reports (Muniraj et al., 2023), (Abo-Bakr & Mohamed, 2023).

These algorithms process incoming data to identify regularities and links between words, phrases, and sentences, and then use this information to determine the meaning of the text.

Tools and methods of Natural Language Processing provide the basis for implementation of these technologies in real applications. Various programming languages and libraries are available for NLP, every one of which has its strengths and weaknesses. The most popular NLP tool is Natural Language Toolkit (NLTK) (NLTK :: Natural Language Toolkit, n.d.) which includes a full set of libraries and programs written in Python that can be used for symbol and statistic Natural Language Processing in English. The set of tools provides functional possibilities for such tasks as tokenization or segmenting of words, marking of parts of speech and creation of text classification datasets. NLTK also provides an ample set of easy-to-use NLP tools for researchers and developers, which makes it one of the most popular NLP libraries. Also, NLP largely relies on statistic Natural Language Processing (Posekany & Dolezal, 2024), machine learning (Glaz et al., 2021) and deep learning methods (Lauriola et al., 2022). A combination of algorithms with models of machine and deep learning allows NLP to automatically extract, classify, and mark components of text data. After this process is finished, the algorithms determine statistical probability for every possible value of elements ensuring complex and effective solution for analysis of big sets of data.

Transformation of Data into Knowledge Graph Format. The process of knowledge acquisition for building a knowledge graph is a general process of element collection from multistructural data, including recognition of entities (facts, real object, terms, definitions etc.), coreferences resolution (finding all linguistic expressions that define the same entity), and determination of links (define links types). Entity Recognition detect mentions of entities in data. Coreference Resolution finds pairs of mentions close in meaning, after which the tasks of extraction of relationships follow that link the entities with their notional links. The following tasks are considered for building of knowledge graph:

1. **Entity Detection.** Detection of entities receives a subset of concepts for semi-structured or structured data which may be nodes of knowledge graph, including:
    a. Recognition of named objects based on semi-structured and not structured data (Adnan & Akbar, 2019), (Chen et al., 2022).
    b. Entity Typification. The task of entity typification (ET) provides detailed and super detailed information about the types of objects (Shimaoka et al., 2017), (Zhang et al., 2018).
    c. Entity Linking. The tasks of Entity Linking (EL), also called entity ambiguity elimination, link mentions of entities with objects corresponding to them in the knowledge graph (Khusro et al., 2014).
2. **Coreference Resolution** Coreference expressions often can be found in a non-structured text. Thus, Coreference Resolution (CR) detects mentions that refer to the same objects. The following models are used for this task:
    a. Statistic models (Martinez-Rodriguez et al., 2020), (Lata et al., 2023), (Zheng et al., 2018).
    b. Deep learning-based models. Deep learning models automatically transform documents into lexical representations to collect the required properties to detect coreference mention pairs (Xi et al., 2015).
3. **Relation Extraction.** The Relation Extraction tasks extract relations from not structured or semi-structured data to determine links between entities and their properties. Relation Extraction is often called Relation Classification. Extraction of binary relations extract relations between pairs of entities, while n-ary extraction extract relations for several entities. Relation Extraction fills the knowledge graph with semantic connections. The following Relation Extraction task types are considered:

a. Open Relation Extraction. The Open Relation Extraction detects facts in not structured data without a predetermined link type. These methods help detect nominal words (like subject or object) and verbal phrases (like subject or object) from a free text to form type knowledge triplets (subject, predicate, object) (Sukhobok et al., 2016).
b. Classification Relation relative to a domain. In this classification, entities are marked with tags from a predetermined set of relations, taking context into account (Augenstein et al., 2014), (Jin et al., 2020).
c. Extraction/classification of deleted controlled relations (Han et al., 2018), (Luo et al., 2017).
d. Relation classification in several stages (Xiong et al.,2018).
e. Extraction models of combined relations (Srikumar & Roth, 2023), (Li & Ji, 2014).
f. Relation extraction on a document level (Peng et al., 2017), (Song et al., 2018), (Sahu at al., 2018).

Analysis of associated papers shows that enough methods and algorithms has been developed for analysis of natural language texts and for transformation of information extracted from a text into the knowledge graph format. However, the task of applying the methods of text semantic analysis to automated search and assessment of scientific and training texts from the point of view of their proximity to domain and subdomain templates has not yet been studied sufficiently.

3. Materials and Methods

3.1. Task setting

Fig. 1 shows a component diagram of an information system intended for automated search and assessment of training texts using textual semantic patterns and ontologies.

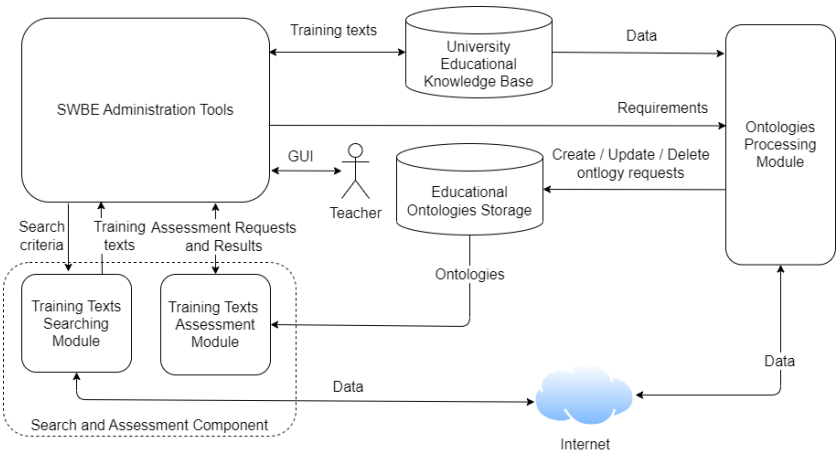


Figure 1: Component diagram of an information system intended for automated search and assessment of training texts [Compiled by the author]

Component description of the systems is represented in Table 1. One of the objectives of this paper is to develop and evaluate the assessment algorithm of the training texts found online from the point of view of proximity to subdomain training templates. This algorithm shall be implemented by the "Training Texts Assessment Module". The input data for the algorithm are the assessed text and subdomain templates in the

OWL-format referring to which the texts are being assessed. The output are the studied text proximity values to every one of the subdomain templates. A text can be classified by those proximity values, attributing it to a subdomain and assessing the text quality from the point of view covering the terminology apparatus defined in the subdomain template.

Table 1 System components

№	Component / Module	Description
1.	Search and Assessment component	Training text's Search and Assessment component, including "Educational Texts Searching Module" and "Training Texts Assessment Module"
2.	Training Texts Searching Module	The module is designed for automated search of educational texts on the Internet by set parameters.
3.	Training Texts Assessment Module	The module is designed to check whether the training texts correspond to certain ontologies containing domain semantic patterns. The assessment implies calculation of proximity of a training text to a certain ontology and, based on the comparison of this measure with a set threshold value, a decision is made about using the text in the educational process.
4.	WSBE Administration Tools	A set of administrative tools to manage an information system. Designed to provide an interaction of a "Teacher" user with the information system by using the graphic user interface (GUI).
5.	University Educational Knowledge Base	University Educational Knowledge Base contains all training materials used in the educational process in electronic format.
6.	Ontologies Processing Module	The module is designed to manage training ontologies. This module controlled by the user's requirements

3.2. Assessment Algorithm of Training Texts

The diagram of training text assessment process is presented at fig. 2.

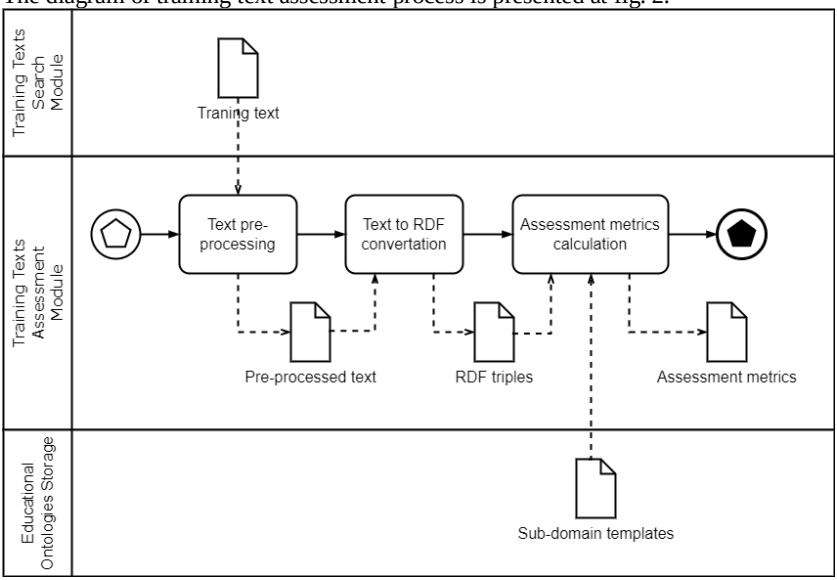


Figure 2: General diagram of training text assessment process [Compiled by the author]

The training text assessment process is comprised of the following stages:

1. **Preprocessing.** At this stage, the studied texts are input in TXT, DOCX, PDF, of any other document format that enables extraction of textual data. The textual data undergoes the following subsequent transformations within the preprocessing stage:
   - Breaking the text down into a list of words.
   - Replacing all upper cases with lower cases.
   - Removing unnecessary spaces and special symbols.
   - Removing prepositions and other function words according to reference material.
   - Bringing the text to a CSV-file (RFC 4180: Common Format and MIME Type for Comma-Separated Values (CSV) Files, n.d.), looking like this:

Article_ID,Term,Value *<Training text ID>,<Term>,<Number of the term entries>* ...

2. **Transforming of Preprocessed Text to RDF Format.** At this stage, a preprocessed CSV-file is input. This file is transformed into RDF format (RDF - Semantic Web Standards, n.d.), (RDF Schema 1.1, 2014). The structure of RDF presentation of a preprocessed file is presented at fig. 3.

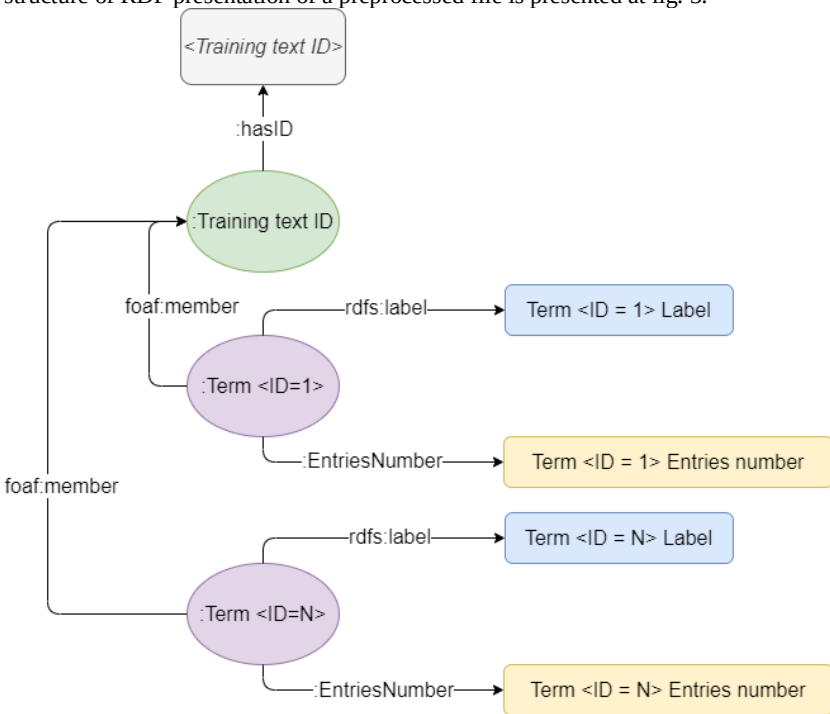


Figure 3: Structure of a training text in RDF format [Compiled by the author]

RDF presentation contains the following nodes, literals and links:

- **:Training text ID** – node containing data about a training text with an identifier <ID>. The value of this identifier is put in a literal connected with the node by predicate **:hasID**.
- **:Term <ID = 1, 2, …, N>** – multitude of nodes containing data about terms in the text. Every term in the text corresponds to a one such node. The value of a term (spelling of a term) is put in a literal linked to the node by the predicate **rdfs:label**. The number of term entries in a training text is contained in a literal linked to the node by the predicate **:EntriesNumber**.

The described presentation of a training text in RDF format is sent to the next stage as a text file in a format allowed to be included in the knowledge graph (e.g. RDF/XML or Turtle).

3. **Training Text Assessment Metric Calculation.** At this stage, the input data are RDF presentation of a training text and subdomain templates to determine the training text quality indices presented in the university knowledge graph in RDF format. The following actions are required to assess the metric calculation for a training text:
   - Upload the RDF presentation of a training text to the university knowledge graph.
   - As all subdomain templates are already put in one knowledge graph with uploaded training text representation, for calculation of statistics by term mention frequencies from every subdomain template one has to run the following SPARQL (SPARQL 1.1 Query Language, n.d.) request:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX my: <http://127.0.0.1/bg/ont/test1#>

SELECT ?Domain_Template_label (COUNT(?term) as ?Terms_number)
(SUM(?term_text_number) as ?Terms_freq)

WHERE
{
    ?Domain_Template my:has_id ?Domain_Template_label .
    ?Domain_Template rdf:type "DomainTemplate" .
    ?term foaf:member ?Domain_Template .
    ?term rdfs:label ?term_label .
    ?Text my:has_id "texts\\1-OIL AND GAS.txt" .
    ?term_text foaf:member ?Text .
    ?term_text rdfs:label ?term_text_label .
    ?term_text my:EntriesNumber ?term_text_number .
   FILTER (?term_label = ?term_text_label) .
}
GROUP BY ?Domain_Template_label
```

The knowledge graph will produce the following table as a SPARQL response to the graph (Table 2).

Table 1 Response on SPARQL request

Domain Template Name (?Domain_Template_label)	The number of terms pertaining to the domain template that appears in the training text (?N_text)	How many times the terms listed in the domain template appear in the training text (?S_text)
OilAndGasDomainTemplate	34	77
….	…	…

- Using metrics and their definitions given in the paper [1], a metric calculation is performed for every subdomain template:
  - Term coverage (Cov). Term coverage is defined as:

$$Cov = \frac{N_{text}}{N} 100\%$$

  - Usage frequency of terms from a domain template ($Freq$). Usage frequency of terms from a domain template is defined as:

$$Freq = \frac{S_{text}}{S} 100\%$$

The following CSV file is produced as output data:

Template_ID,Cov,Freq
<Subdomain template ID>,<Terminology coverage (Cov)>,< Frequency of using the terms pertaining to a subdomain template (Freq)>
…

3.3. Assessment Algorithm of Training Texts

A case study was about assessing 18 training texts against the following subdomain templates:

1. Basic glossary for Mining Faculty.
2. Basic glossary for Geological Faculty.
3. Basic glossary for Mechanical Engineering Faculty.
4. Basic glossary for Oil and Gas Faculty.
5. Basic glossary for Civil Engineering Faculty.
6. Basic glossary for Faculty of Mineral Processing.
7. Basic glossary for Faculty of Fundamental Sciences and Humanities.
8. Basic glossary for Energy Faculty.
9. Basic glossary for Faculty of Economics.

Blazegraph (Blazegraph Database, n.d.) was used to solve the task. Also, there were program components developed on Python to process the training texts and calculate their assessment metrics. All initial data, developed software, and results, as well as SPARQL requests and responses to them are placed at a public repository (Yulia-Murzo, n.d.). The assessment results of the selected text named “The Rise of the Russian Oil Industry” (total words: 528) are shown below (Table 3). The study shows that the proposed algorithm of automated assessment of training texts reflects actual proximity of a text to subdomain templates. In case of the considered case study, it is proximity to the faculties’ basic glossaries.

Table 2 The assessment results of the selected text named "The Rise of the Russian Oil Industry"

#	Domain Template Name	Term coverage (Cov)	Usage frequency of terms from a domain template ($Freq$)
1.	OilAndGasDomainTemplate	2%	14%
2.	MiningDomainTemplate	0	0
3.	GeologicalDomainTemplate	1%	10%
4.	MechanicalEngineeringDomainTemplate	0	0
5.	CivilEngineeringDomainTemplate	0	0
6.	MineralProcessingDomainTemplate	0	0
7.	FundamentalSciencesAndHumanitiesDomainTemplate	1%	1%
8.	EnergyDomainTemplate	0	0
9.	EconomicsDomainTemplate	3%	4%

4. Discussion

The results of this research have allowed us to develop an efficient algorithm for automatic analysis of technical texts with their proximity to resources templates. The conducted experiment has shown that the algorithm distinguishes texts by the frequency and coverage of special terms in the presented texts and determines the parameters of terminological set coverage from the template and the frequency of terms usage. On the basis of the given architectural solutions of the information system and the algorithm of text analysis the task of selection of educational texts for students of the direction 'Mineral Processing' with threshold values of proximity criteria Cov = 85% and Freq = 15% for each text was solved, which provided the coverage of seven selected texts (Cov=85%), which supplied the required saturation of educational materials with special terms. It is also worth noting that the proposed text analysis algorithm is able to classify the proximity of texts to domain patterns, which is provided by comparing the proximity criteria of the text to each of the analysed domain patterns. Taking into account that the field of knowledge of mineral resources when learning a professionally oriented foreign language depends to a very large extent on the knowledge of terminological apparatus in a foreign language, such a verifiable approach to analysing the quality of texts from the point of view of saturation with special terms makes a significant contribution to the quality of selection of teaching materials.

5. Conclusion

The paper shows analysis of existing methods of processing and analysis of texts and their presentation in formats of semantic networks and knowledge graphs. Based on the analysis, an algorithm is suggested for automated assessment of training texts found online via proximity assessment of the studied text to subdomain templates in OWL format. To calculate the metrics of the text proximity to subdomain templates, term use stats are extracted from the text. Those stats are then transformed into RDF format and placed on the university knowledge graph. Metrics calculation is done by running SPARQL requests and calculation tasks implemented on Python. Performance and efficiency of the suggested algorithm was confirmed by the performed case study; its materials are available at a public repository. Subsequently, this algorithm can be further developed by studying the possibilities of machine learning algorithms for classification of the found texts.

Bibliographic references

Abo-Bakr, H., & Mohamed, S.A. (2023). Automatic multi-documents text summarization by a large-scale sparse multi-objective optimization algorithm. In Complex & Intelligent Systems (Vol. 9, Issue 4, pp. 4629-4644). Springer Science and Business Media LLC. https://doi.org/10.1007/s40747-023-00967-y

Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. In Journal of Big Data (Vol. 6, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-019-0254-8

Akhmetov, I., Pak, A., Ualiyeva, I., & Gelbukh, A. (2020). Highly Language-Independent Word Lemmatization Using a Machine-Learning Classifier. Computación y Sistemas, 24(3). Instituto Politecnico Nacion-al/Centro de Investigacion en Computacion. https://doi.org/10.13053/cys-24-3-3775

AlMousa, M., Benlamri, R., & Khoury, R. (2022). A novel word sense disambiguation approach using Word-Net knowledge graph. Computer Speech & Language, 74, 101337. Elsevier BV. https://doi.org/10.1016/j.csl.2021.101337

Alshalabi, H., Tiun, S., Omar, N., AL-Aswadi, F. N., & Ali Alezabi, K. (2022). Arabic light-based stemmer using new rules. Journal of King Saud University - Computer and Information Sciences, 34(9), 6635–6642. Elsevier BV. https://doi.org/10.1016/j.jksuci.2021.08.017

Augenstein, I., Maynard, D., Ciravegna, F. (2014). Relation Extraction from the Web Using Distant Supervision. Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E. (eds) Knowledge Engineering and Knowledge Management. EKAW 2014. Lecture Notes in Computer Science, 8876. Springer, Cham. https://doi.org/10.1007/978-3-319-13704-9_3

Aziz, M. V. G., Prihatmanto, A. S., Henriyan, D., & Wijaya, R. (2015). Design and implementation of natural language processing with syntax and semantic analysis for extract traffic conditions from social media data. 5th IEEE International Conference on System Engineering and Technology (ICSET). IEEE. https://doi.org/10.1109/icsengt.2015.7412443 Blazegraph. https://blazegraph.com/

Boyko, S. A., Koltsova, E. A. & Valentina, A. (2022). Spiridonova Application of a corpus-based approach in teaching English for specific purposes to Master's degree students of engineering and technical majors. Global Journal of Engineering Education. 1, 40-45.

Chen, Z., Meng, W. & Dragut, E.. (2022). Web Record Extraction with Invariants. Proc. VLDB Endow. 16, 4 (December 2022), 959–972. https://doi.org/10.14778/3574245.3574276.

Cherepovitsyn A.E., & Tretyakov N.A. (2023). Development of New System for Assessing the Applicability of Digital Projects in the Oil and Gas Sector. Journal of Mining Institute, 262, 6ä28-642. EDN QYBHMC.

Common Format and MIME Type for Comma-Separated Values (CSV) Files. https://datatracker.ietf.org/doc/html/rfc4180

Devedžić, V. (2006). Semantic web and education. Integrated series on information systems/Integrated series in information systems. https://doi.org/10.1007/978-0-387-35417-0

Han, X., Yu, P., Liu, Z., Sun, M. & P. Li. (2018). Hierarchical relation extraction with coarse-to-fine grained attention. Proc. EMNLP, 2236–2245.

Hitzler, P., Krotzsch, M., & Rudolph, S. (2009). Foundations of Semantic Web Technologies. Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9781420090512

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. In Science, 349 (6245), 261–266. American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.aaa8685
Hotho, A., Martinez-Rodriguez,J. L., Hogan, A. & Lopez-Arevalo, I. (2020). Information extraction meets the Semantic Web: A survey. Semant. web 11(2), 255–335. https://doi.org/10.3233/SW-180333.
Dastan Hussen Maulud, Subhi R. M. Zeebaree, Karwan Jacksi, Mohammed A. Mohammed Sadeeq, Karzan Hussein Sharif (2021). State of Art for Semantic Analysis of Natural Language Processing. In Qubahan Academic Journal (Vol. 1, Issue 2, pp. 21–28). QUBAHAN. https://doi.org/10.48161/qaj.v1n2a40
Jin, L., Song, L., Zhang, Y., Xu, K., Ma, W.-Y., & Yu, D. (2020). Relation Extraction Exploiting Full Dependency Forests. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8034–8041. Association for the Advancement of Artificial Intelligence (AAAI). https://doi.org/10.1609/aaai.v34i05.6313
Kannan, S., Karuppusamy, S., Nedunchezhian, A., Venkateshan, P., Wang, P., Bojja, N., & Kejariwal, A. (2016). Big Data Analytics for Social Media. Big Data, 63–94. Elsevier. https://doi.org/10.1016/b978-0-12-805394-2.00003-9
Kannappan, S. (2023). Sentiment analysis using natural language processing and machine learning. Shu Ju Cai Ji Yu Chu Li, Journal of Data Acquisition and Processing, 38, 520-526. 10.5281/zenodo.7766376.
Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744). Springer Science and Business Media LLC. https://doi.org/10.1007/s11042-022-13428-4
Khusro, S., Latif, A., & Ullah, I. (2014). On methods and tools of table detection, extraction and annotation in PDF documents. Journal of Information Science, 41(1), 41–57). SAGE Publications. https://doi.org/10.1177/0165551514551903
Koteleva N., Frenkel I. (2021). Digital Processing of Seismic Data from Open-Pit Mining Blasts. Applied Sciences, 11(1), 383. https://doi.org/10.3390/app11010383
Kretschmann, J., Plien, M., Nguyen, T. H. N., & Rudakov, M. (2020). Effective capacity building by empowerment teaching in the field of occupational safety and health management in mining. Journal of Mining Institute, 242-248. https://doi.org/10.31897/pmi.2020.2.248
Lata, K., Singh, P., & Dutta, K. (2023). SMDDH: Singleton Mention detection using Deep Learning in Hindi Text (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2301.09361.
Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. In Neurocomputing, 470, 443–456. Elsevier BV. https://doi.org/10.1016/j.neucom.2021.05.103
Le Glaz, A., Haralambous, Y., Kim-Dufor, D., Lenca, P., Billot, R., Ryan, T., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. In Journal of Medical Internet Research (Vol. 23, Issue 5, e15708). JMIR Publications Inc. https://doi.org/10.2196/15708
Li, Q., & Ji, H. (2014). Incremental joint extraction of entity mentions and relations. ACL, 1: Long Papers, 402–412.
Litvinenko, V., Bowbrick, I., Naumov, I., & Zaitseva, Z. (2022). Global guidelines and requirements for professional competencies of natural resource extraction engineers: Implications for ESG principles and sustainable development goals. Journal of Cleaner Production, 338, 130530.) Elsevier BV. https://doi.org/10.1016/j.jclepro.2022.130530.
Luo, B. Y. Feng, Wang, Z., Zhu, Z., Huang, S., Yan, R. & Zhao, D. (2017). Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. ACL, 1: Long Papers, 430–439.

Manning, C. D., & Schütze, H. (1999). Foundations of Statistical natural language processing. https://doi.org/10.1017/S1351324902212851
Massa Cereda, P. R., Miura, N. K., & Neto, J. J. (2018). Syntactic analysis of natural language sentences based on rewriting systems and adaptivity. Procedia Computer Science, 130, 1102–1107. Elsevier BV. https://doi.org/10.1016/j.procs.2018.04.164
Rajivkumar Mente, Someshekhar Aland, Bapu Chendage (2022). Review of Word Sense Disambiguation and It'S Approaches, April 30. Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022, Available at SSRN: https://ssrn.com/abstract=4097221 or http://dx.doi.org/10.2139/ssrn.4097221
Muniraj, P., Sabarmathi, K. R., Leelavathi, R., & Balaji B., S. (2023). HNTSumm: Hybrid text summarization of transliterated news articles. International Journal of Intelligent Networks, 4, 53–61. Elsevier BV. https://doi.org/10.1016/j.ijin.2023.03.001
Murzo, Y., Chuvileva, N., & Schetinina, A. (2022). Using semantic patterns in web search and assessment of professionally oriented texts in a foreign language for training students in higher education institutions of mineral resource profile. International Journal of Emerging Technologies in Learning/International Journal: Emerging Technologies in Learning, 17(05), 238-251. https://doi.org/10.3991/ijet.v17i05.28231
Yulia Murzo, Case study GitHub open access repository. https://github.com/yulia-murzo/texts-assessment/
Nikonova, E., Yakhyaeva, K., Pivkina, N., & Schetinina, A. (2023). Using Artificial Intelligence Tools in Teaching a Foreign Language in Higher Technical Institutions. European Journal of Contemporary Education, 12(2), 578-589. http://dx.doi.org/10.13187/ejced.2023.2.578 NLTK: https://www.nltk.org/
N. Peng, H. Poon, C. Quirk, K. Toutanova, & W. Yih. (2017). Cross-sentence n-ary relation extraction with graph lstms. Trans. Assoc. Comput. Linguistics, 5, 101–115.
Posekany, A., & Dolezal, D. (2024). Providing a Natural Language Processing App for Language Teachers. Towards a Hybrid, Flexible and Socially Engaged Higher Education, 467–473. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51979-6_48
Pushmina S. A. (2021). English for specific purposes (ESP) on-line teaching for engineering students World Transactions on Engineering and Technology Education, 19(2), 215–219.
RDF Schema 1.1 Standard: https://www.w3.org/TR/rdf-schema/
Resource Description Framework (RDF). https://www.w3.org/RDF/
Rodriguez-Cardos, R., Chicharro, F. P. R., Varela, J. A. O., & Serrano-Guerrero, J. (2019). A method for the semantic analysis of documents in a business context in Spanish. Procedia Computer Science, 162, 803-810. Elsevier BV. https://doi.org/10.1016/j.procs.2019.12.053
Rudnik, S. N., Afanasev , V. G., & Samylovskaya , E. A. (2023). 250 years in the service of the Fatherland: Empress Catherine II Saint Petersburg Mining University in facts and figures. Journal of Mining Institute, 263, 810-830.
Sahu, S. K., Christopoulou, F., Miwa, M. & Ananiadou, S. (2019). Intersentence relation extraction with document-level graph convolutional neural network. Proc. ACL, 1: Long Papers, 4309-4316.
Samylovskaya E., Makhovikov A., Lutonin A., Medvedev D., & Kudryavtseva R.-E. (2022). Digital Technologies in Arctic Oil and Gas Resources Extraction: Global Trends and Russian Experience. Resources. 11(3), 29. https://doi.org/10.3390/resources11030029
Shimaoka, S., Stenetorp, P., Inui, K. & Riedel, S. (2017). Neural architectures for fine-grained entity type classification. Proc. EACL, 1: Long Papers, 1271-1280.
Song, L., Zhang, Y., Wang, Z., & Gildea, D. (2018). N-ary relation extraction using graph-state LSTM. Proc. EMNLP, 2226–2235.

SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/
Srikumar, V., & Roth, D. (2023). The Integer Linear Programming Inference Cookbook (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.00171
Sveshnikova S. A. , Skornyakova E. R., Troitskaya M. A., Rogova I. S. (2022). Development of Engineering Students' Motivation and Independent Learning Skills. European Journal of Contemporary Education, 11, 555-569. https://ejce.cherkasgu.press/journals_n/1656514745.pdf.
Sukhobok, D., Nikolov, N., Pultier, A., Ye, X., Berre, A., Moynihan, R., Roberts, B., Elvesæter, B., Mahasivam, N., & Roman, D. (2016). Tabular Data Cleaning and Linked Data Generation with Grafterizer. The Semantic Web, 134-139. Springer International Publishing. https://doi.org/10.1007/978-3-319-47602-5_27.
Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780. Springer Science and Business Media LLC. https://doi.org/10.1007/s10462-022-10144-1
Xi, X., Zhou, G., Hu, F. & Fu, B. (2015). A convolutional deep neural network for coreference resolution via modeling hierarchical features. IScIDE, 2015, Revised Selected Papers, Proceedings, Part II, 9243, 361-372.
Xiong, W., Yu, M., Chang, S., Guo, X., & Wang, W. Y. (2018). One-shot relational learning for knowledge graphs. Proc. EMNLP, 1980–1990.
Zhang, S., Duh, K., & Durme, B. V. (2018). Fine-grained entity typing through increased discourse context and adaptive classification thresholds. Proc. *SEM@NAACL-HLT, 2018, 173–179.
Zheng, Q., Diao, X., Cao, J., Zhou, X., Liu, Y., & Li, H. (2018). Multi-Modal Coreference Resolution with the Correlation between Space Structures (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1804.08010

Words: *4847*
Characters: *34 110 (19 standard pages)*

Yuliya Murzo
Department of Foreign Languages of Saint Petersburg Mining University
21st Line of Vasilevskiy Island 2
199106, Saint-Petersburg,
Russia
Murzo_YuE@pers.spmi.ru
ORCID ID: 0000-0002-5373-5604