BRITISH VIEW

MULTIDISCIPLINARY JOURNAL

Anthropologie, Applied Linguistics, Applied Physics, Architecture, Artificial Intelligence, Astronomy, Biological Sciences, Botany, Chemistry, Communication studies, Computer Sciences, Computing technology, Cultural studies, Design, Earth Sciences, Ecology, Education, Electronics, Energy, Engineering Sciences, Environmental Sciences, Ethics, Ethnicity and Racism Studies, Fisheries, Forestry, Gender Studies, Geography, Health Sciences, History, Interdisciplinary Social Sciences, Labour studies, Languages and Linguistics, Law, Library Studies, Life sciences, Literature, Logic, Marine Sciences, Materials Engineering, Mathematics, Media Studies, Medical Sciences, Museum Studies, Music, Nanotechnology, Nuclear Physics, Optics, Philosophy, Physics, Political Science, Psychology, Publishing and editing, Religious Studies, Social Work, Sociology, Space Sciences, Statistics, Transportation, Visual and Performing Arts, Zoology and all other subject areas.

Manuscripts typed on our article template can be submitted through our website here. Alternatively, authors can send papers as an email attachment to editor@britishview.co.uk
Editor Multidisciplinary Journals
Website: http://britishview.co.uk
Email: editor@britishview.co.uk

British View ISSN 2041-3963 Volume 8 Issue 4 2023
Universal impact factor 8.528
SJIF 2022: 4.629

AUTOMATIC FORMATION OF DICTIONARIES OF THE SUBJECT AREA

M.M. Makhmudov PhD SUE "UNICON.UZ"
M.M. Mukhitdinov D.Sc. SUE "UNICON.UZ"
U.Y. Tuliev The National University of Uzbekistan

Abstract. The problem of creating dictionaries (groups) of subject areas from natural language words (concepts) based on data from text documents is considered. The identification of the semantic coherence of words based on the relationship between concepts is also investigated. Based on the ranking of features based on the results of splitting documents into clusters, semantically related words are selected by topic.

Keywords: semantic connectedness, word embedding, word2vec, ontology, stemming, lemmatization.

Language has become the most powerful means of communicating knowledge and ideas among humans as mankind has learned to speak and write. Nowadays, due to the extraordinary development of information technology, knowledge is collected and delivered in the form of electronic documents written in natural language. What is needed now are information systems that understand and analyze these documents. However, electronic systems have mastered many of the jobs that must be done directly by the human factor (robots used in factories, e-mail, autopilots, expert systems that diagnose patients, etc.). Significant work has been done in these areas to solve word processing and translation problems using machine learning algorithms. While natural language helps in communicating ideas, with its complex rules and features, it can be a serious obstacle to machine understanding. Such obstacles are related to synonymy and multiple meanings of words, and these problems are overcome with the help of ontological resources that make up the model of natural language. Ontological resources can be in the following forms:

- **Dictionary** - a list of unambiguous terms;
- **Glossary** - an explanatory dictionary containing the meanings of multi-valued terms;
- **Thesaurus** - is a glossary containing semantic relationships between terms.

Given that such resources play a large role in the automatic analysis of documents, one can understand the high demand for their automatic generation. In addition, the reason for the emergence of new terms is that today's news appears in rapidly changing pictures. This process creates the need for constant updating of ontological resources. For this purpose, it is more expedient to synthesize new terms from a set of documents by means of an automated system than by means of a human factor, and to organize their inclusion in dictionaries with the help of experts.

To automate the process of dictionary generation, the following steps must be performed:

- formation of a collection of text documents by subject areas;

To automate the process of forming dictionaries you need to perform the following actions:

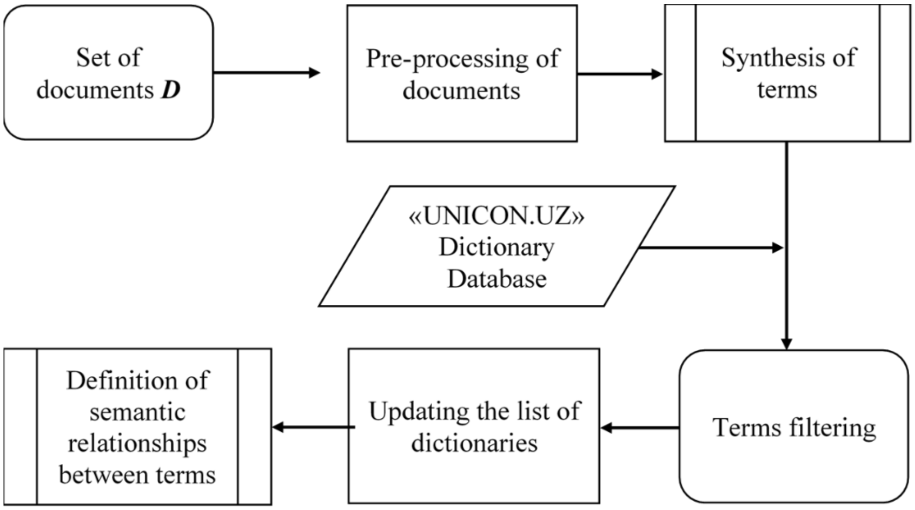
- formation of a collection of text documents by subject areas;
- preprocessing text documents in natural language;
- extraction of subject-oriented terms;
- filtering subject-oriented terms in dictionaries (UNICON.UZ)
- taking into account expert evaluations;
- clarification of semantic relationships, taking into account the inclusion of new terms, in subject area dictionaries.

Organizing a collection of text documents by subject area

Given that vocabulary resources play a large role in automatic document analysis, we can understand the need for their automatic generation. The reason for the emergence of new terms is that today's news is presented in rapidly changing pictures. This process creates the need for constant updating of ontological resources. For this purpose, it is more expedient to select new terms from document collections by means of an automated system without involving people, and to organize their inclusion in dictionaries, taking into account the opinion of experts.

Automatic organization of dictionaries can be realized by performing the following actions:

- organize the collection of text documents by subject areas;
- perform preprocessing of text documents in natural language;
- select specific terms by subject areas;
- update the set of terms using the existing dictionary (unicon.uz);
- perform expert analysis;
- re-form semantic relations taking into account subject area vocabularies update.


Set of documents ***D***
Pre-processing of documents
Synthesis of terms
«UNICON.UZ» Dictionary Database
Terms filtering
Updating the list of dictionaries
Definition of semantic relationships between terms

Organizing a collection of text documents by subject area

Text documents serve as the main source of updating terminological dictionaries on subject areas, taking into account the characteristics of natural language natural language. The main resource for machine learning algorithms used for term synthesis is the database [4]. The database contains normative documents, laws and scientific works related to the subject areas. Examples of such a database are dissertation abstracts at the disposal of the Higher Attestation Commission of the Republic of Uzbekistan. Since the text is considered as poorly structured data, before applying machine learning algorithms to it it is necessary to make their preprocessing. In other words, the texts should be presented in a tabular form for machine learning algorithms.

Pre-processing of text documents using natural language processing methods

The Uzbek language belongs to the Turkic language family and is an example of low-resource languages. In Uzbek, one word of the normal form can have dozens of word forms by adding suffixes. For a computer system, each of them is treated as a separate value. To avoid this, there are **stemming** [1,7] and **lemmatization** [6] technologies in natural language processing that are used to normalize words. Because word forms are equivalent to a single normal form. For example, in mathematics, the words *"differential"*, *"differentiallar*", *"differentiating"* are equivalent to the normal form "*differential"*. Therefore, a morphological analysis must be performed to clear the words from the complements and bring them to the normal form [2].

Extraction of single-valued terms by subject area

After forming a set of words W=(w1,...,wn), presented in normal form, from all documents in a subject domain, the frequency of occurrences of $w_i \in W$ is counted. We denote the number of occurrences of the word $w_i \in W$ by n_{w_i}in the set P, from the subject domain documents. The set W includes the set of terms (T) related to the subject

domain as well as the set of common words (U). Given that T=W\U, we must first define
the set U:

$$U = \{w_i \in W | n_{w_i} \geq p|P|\}. \qquad (1)$$

In (1) n_{w_i}) is the threshold coefficient of separation of frequently used words, the value of which is usually 0.7. That is, words occurring in more than 70% of all documents are commonly used and are considered terminologically irrelevant.

Filtering the term set T and updating the dictionary

The next step is to extract pre-existing terms from the synthesized terms. We suggest using the UNICON.UZ vocabulary database. This database contains more than 40 dictionaries created by employees in the organization. The incoming set *T* is filtered using these dictionaries to pass it for further examination. Based on the results of expert reviews, the dictionaries are updated.

Determining the semantic relationship between terms

It is known that subject areas are divided into several subdomains, which are treated as topics. Within each topic there are semantic relations between terms. Determining such relationships in mathematical linguistics is called **word embedding** [8,3]. One of the most common algorithms used for **word embedding** is the **word2vec** algorithm [5]. In addition, it is possible to use the works of Uzbek scientists, who conduct research on creating a model of natural language, taking into account the specific features of the Uzbek language. The method of calculating the content authenticity of documents [4] developed by the scientists of the National University of Uzbekistan is aimed at determining the semantic relationships between the terms related to the same topic. Let $G_1,...,G_h$, i=1, ...,h, h≥2 partition into non-overlapping groups of documents from D by a set of latent features Y(group). For each group G_i, we define the value of the class membership function of objects K_1 over G_i as $\lambda_i(K_1)=d_{i1}/|G_i|$, where di1 is the number of objects of class K_1 in G_i. The content authenticity of the documents from D when dividing them into h groups will be calculated as

$$F\big(h, Y(guruh)\big) = \frac{1}{m}\sum_{j=1}^{h}\begin{cases} |G_j|\lambda_j(K_1), \lambda_j(K_1) > 0.5; \\ |G_j|(1-\lambda_j(K_1)), \lambda_j(K_1) < 0.5. \end{cases} \qquad (2)$$

With the help of such methods it is possible to solve the problems of synonymy and polysemy (polysemy) of natural language.

Computational experiment

The computational experiment was conducted in the authored abstracts of VAK of the Republic of Uzbekistan, relating to 12 subject areas. The number of all documents - 1634 and their distribution by subject areas are presented in Table 1.

Table 1. Distribution of the documents by subject areas

Subject area	Number of	Subject area	Number of

	documents		documents
Biology	109	Law	83
Physics	162	Economics	189
Geography	28	Chemistry	120
Geology	44	Culture	11
Mathematics	95	Technology	380
Pedagogy	137	Medicine	277

The results of the preprocessing of these text documents and the synthesis of terms from them are shown in Table 2.

Table 2. Number of words and extracted terms in documents by subject area

Subject area	Number of words (|W|)	Number of terms extracted (|*T*|)
Biology	231835	16236
Physics	244620	14497
Geography	123915	2203
Geology	215025	9609
Mathematics	107050	4967
Pedagogy	102734	18435
Law	214739	16066
Economics	138170	31556
Chemistry	224378	17188
Culture	72734	4891
Technology	148426	39391
Medicine	224749	45087

Analysis of Tables 1 and 2 shows that the weight of terms synthesized by subject area depends on the volume of the collected corpus. Determining the semantic relationships between the terms listed in Table 2 remains a work in progress.

Literature

1. A. Ismailov, N. Abdurakhmonova (2022). The development of Alisher stemmer for Uzbek Language. Science and Education. 3.
2. Abdurakhmonova N., Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation / X International Journal of Systems

Engineering 2018; 2(1): 26-28 http://www.sciencepublishinggroup.com/j/ijsedoi:10.11648/j.ijse.20180201.16

3. Dubin, David (2004). "The most influential paper Gerard Salton never wrote". Retrieved 18 October 2020.
4. Ignatev N.A., Tuliev U.Y. Semantic structuring of text documents based on patterns of natural language entities // Computer Research and Modeling, 2022, vol. 14, no. 5, pp. 1185-1197. DOI: 10.20537/2076-7633-2022-14-5-1185-1197
5. J. Nay (21 December 2017). "Gov2Vec: Learning Distributed Representations of Institutions and Their Legal Text". SSRN. SSRN 3087278
6. M. Sharipov, O. Sobirov (2022). Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language.
7. M. Sharipov, O. Yuldashov UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language// The International Conference on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 6, 2022, Koper, Slovenia.
8. Reisinger, Joseph; Mooney, Raymond J. (2010). Multi-Prototype Vector-Space Models of Word Meaning. Vol. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California: Association for Computational Linguistics. pp. 109–117. ISBN 978-1-932432-65-7. Retrieved October 25, 2019.