Research and development of linguo-statistical methods for forming a portrait of a subject area

Oleg V. Zolotarev
ol-zolot@yandex.ru
ANO HE «Russian New University», Moscow, Russia

The project aims to solve the fundamental scientific problem of semantic modeling, within the framework of which a methodology is developed for the automated identification of translation links (translation correspondences), as well as hierarchical, synonymous and associative links from Internet texts and the construction of multilingual associative hierarchical portraits of subject area (MAHPSA), in particular, on autonomous uninhabited underwater vehicles (UUV). Accounting for multilingual and heterogeneous resources allows you to get a more complete picture of what is happening in the subject area, to identify the sources of the origin of ideas, the speed and directions of their distribution, to identify significant documents and promising directions. The solution to the problem is based on an integrated approach that combines the methods of statistics, corpus linguistics and distributive semantics, and is implemented in technology that involves the development of linguo-statistical mechanisms for the formation of a multilingual associative hierarchical portrait of a subject area, which is a dictionary of significant terms of the subject area, the elements of which organized in synonymous series (synsets), including translational correspondences, as well as associative and hierarchical relationships.

***Keywords**: Linguo-statistical methods, associative and hierarchical portrait of the subject area, multilingual integrated ontology, forecasting the spread of ideas, multilingual body of the subject area.*

1. Introduction

The growth of volumes on the Internet significantly complicates the search for information. Using semantic search, comparing multilingual documents will allow you to find new interesting trends and ideas, which will significantly reduce the cost of developing and popularizing new areas in science. Using a multilingual associative hierarchical portrait of a subject area when comparing documents will allow us to compare texts not only on the basis of matching phrases included in these documents, but also on the matching of the described objects and processes. MAHPSA allows you to determine the semantic similarity of documents even if the documents do not have common words that are included in both documents. MAHPSA allows you to calculate the integrated statistics of a multilingual collection, determine significant documents and promising areas without translating documents into one of the languages. This is important for the automatic processing of a large number of documents (Big Data). The construction of MAHPSA will provide an opportunity not only to compare documents and search for new ideas, but also to solve other problems associated with the rapid analysis of a large amount of information.

2. Technique of automatic formation of a multilingual associative-hierarchical portrait of a subject area

The essence of the proposed method for the formation of a multilingual associative-hierarchical portrait of a subject domain consists in iteratively expanding the initial multilingual dictionary of significant phrases to the hierarchy of multilingual synonymous series (synsets). The method can be stated as the following algorithm:

1) Compiling a collection of multilingual texts by means of a directed search in the databases of scientific documents (for example, Dimensions) by keywords;
2) Word processing by means of the Pullenti program, tokenization and metatoke nization;
3) Automatic generation of glossaries of terms and megalemms; expert quality control of generated dictionaries;
4) Automatic selection of topics on the basis of thematic modeling methods, the formation of a dictionary of subject areas, the selection of many keywords of subject areas, expert control, topic correction;
5) The formation of a dictionary of key terms mapped to topics;
6) Compilation of frequency dictionaries of domain terms (using statistical methods);
7) Compilation of frequency dictionaries of subject domain megalemmas;
8) Building multilingual synsets by combining BabelNet resources and a megalemma dictionary;
9) Building SVPs using a neural network model (a combination of Word2Vec with multilingual recurrent neural networks RNN) for texts that have undergone preprocessing;
10) Performing hierarchical clustering using Word2Vec and RNN, taking into account the hierarchical relationships of synsets;
11) The construction of an ordered list of candidates for hierarchical relationships from associative connections of the neural network model; viewing and correction of hierarchical relations is implemented on the basis of the Keywen Knowledge Architect resource [1].

3. Methodology for calculating integral statistics based on MAHPSA

MAHPSA is created automatically on the basis of statistical analysis of large volumes of texts from the Internet. The hierarchical connections that make up the MAHPSA form a hierarchy and classifier that facilitate the search and navigation in the multilingual subject area of the UUV.

The proposed methodology also includes the integration of various MAHPSA s with multilingual linguistic resources (WordNet, Wikipedia, BabelNet, etc.) to obtain the largest multilingual ontology with relevant knowledge and improved coverage of terminology in the subject areas under consideration. The combined (integral) ontology contains a hierarchy of synonymic series (synsets) of multilingual terms, including Russian, and

serves as the basis for constructing a single multilingual vector space that allows us to evaluate the semantic proximity of multilingual texts, synsets and terms, similar to NASARI and MAFFIN methods. The translation correspondences between the multilingual synsets of MAHPSA are built using Word2Vec technology. Integral ontology allows you to calculate integrated multilingual statistics and trends in the use of terms and ideas, which allows you to predict the distribution of ideas between languages and determine promising directions. A measure of the semantic proximity of multilingual documents allows you to identify implicit links between documents and determine significant documents, which is necessary to collect high-quality information from the open Internet and build large relevant multilingual corpuses of texts for the subject area. Thus, increasing the size and quality of integral ontology will allow us to build a better similarity measure and subject corpus of texts, extracting knowledge from which in turn will further increase the size and quality of integral ontology.

The methodology includes not only the identification of significant documents, but also the identification of trends and the identification of promising areas for the development of science.

To develop the first version of the integrated statistics methodology based on MAHPSA, it is necessary to do the following:

1) Conduct morphological, syntactic and partially semantic analysis of the text;
2) Select typed objects - named entities;
3) Identify formal elements for the presentation of concepts;
4) Develop a structure and software for storing a multilingual collection of documents;
5) Create dictionaries for storing structured information;
6) Develop neural network algorithms for calculating integrated statistics based on MAHPSA.

The first version of the program has been developed for highlighting interlingual implicit connections and assessing the semantic similarity of phrases in different languages.

Text processing is carried out using the program PullEnti [2]. This is a unique product that wins the computer linguistics competitions held as part of the Dialogue conference.

Pullenti is a linguistic processor developed at the Institute of Informatics Problems, which is constantly being refined and allows morphological, syntactic and partially semantic analysis of the text, distinguishing typed objects - named entities.

Pullenti SDK includes the following main blocks:

1) Tokenization: breakdown into words (tokens) as adjusted (Fig. 1 [2-12]);
2) Morphological analysis: definition for tokens of parts of speech (this is a POS-tagger - Part of Speech, which gives out all possible options for a word form regardless of its surrounding context). Languages are Russian, Ukrainian and English. There is normalization, reduction of the word form to the desired case \ gender \ number, and there is also processing of unknown and new words, and there is also a mode for correcting errors (Fig. 2 [2-12]);
3) Selection of named entities [13] (NER - Names Entity Recognition): a lot of so-called analyzers that find entities of the corresponding type (person, organization, geographical objects, etc.) in sequences of tokens (Fig. 3 [2-12]);
4) A lot of tools for working with numerical data, nominal and verb groups, brackets and quotation marks, dictionaries of terms and abbreviations, various checks (for example, equivalence of strings in Latin and Cyrillic letters) and other useful features that appeared during the solution of practical problems (Fig. . 4 [2-12]);
5) Derivative dictionary: a dictionary of the so-called derivative groups (many same-root words, but different parts of speech, and one group contains words in different languages), group management model (what can come after a group), synonymy, etc.;
6) Semantic representation: tokens are structured in the form of a graph with semantic connections to solve more complex problems related to meaning [14].

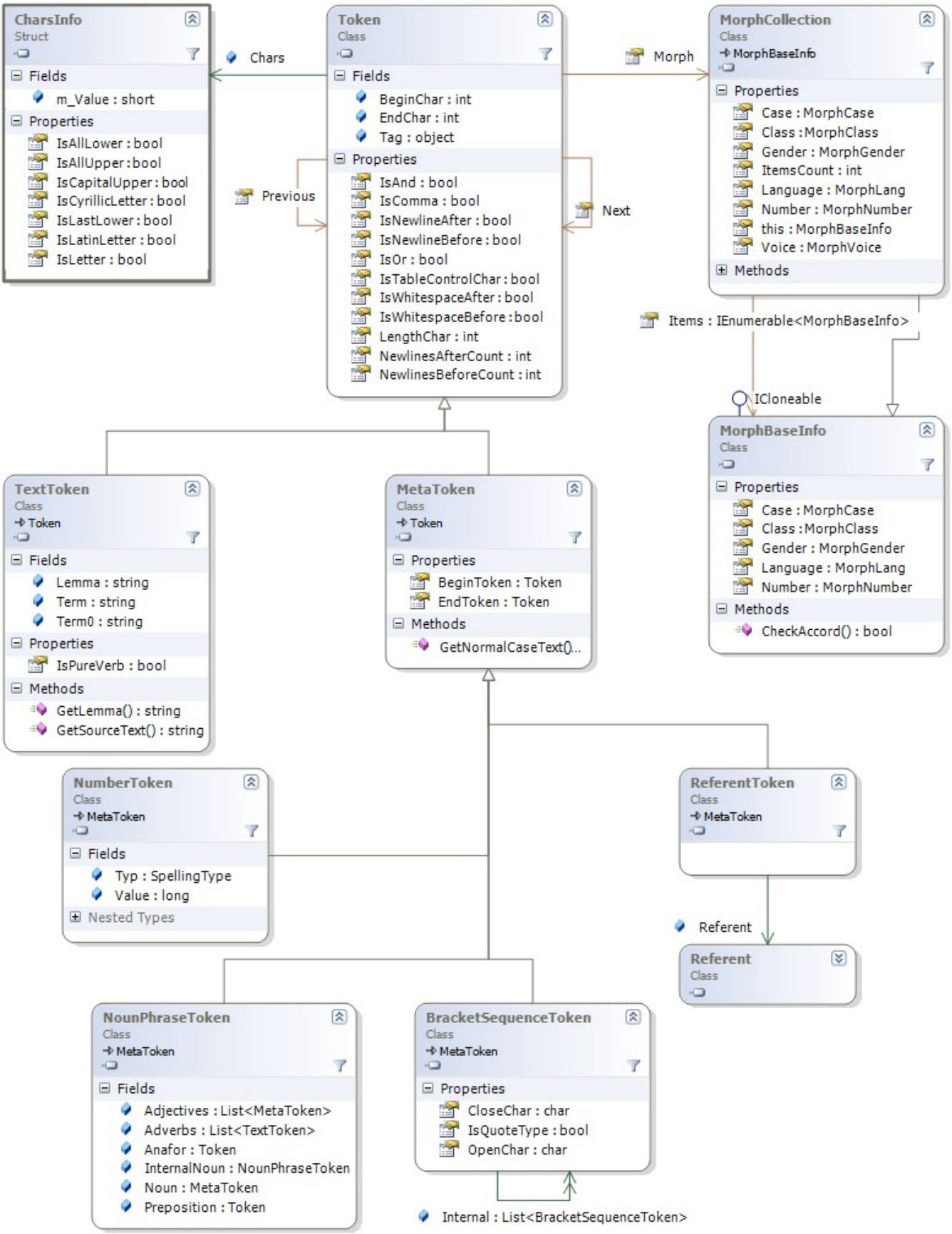

Fig. 1. Tokenization

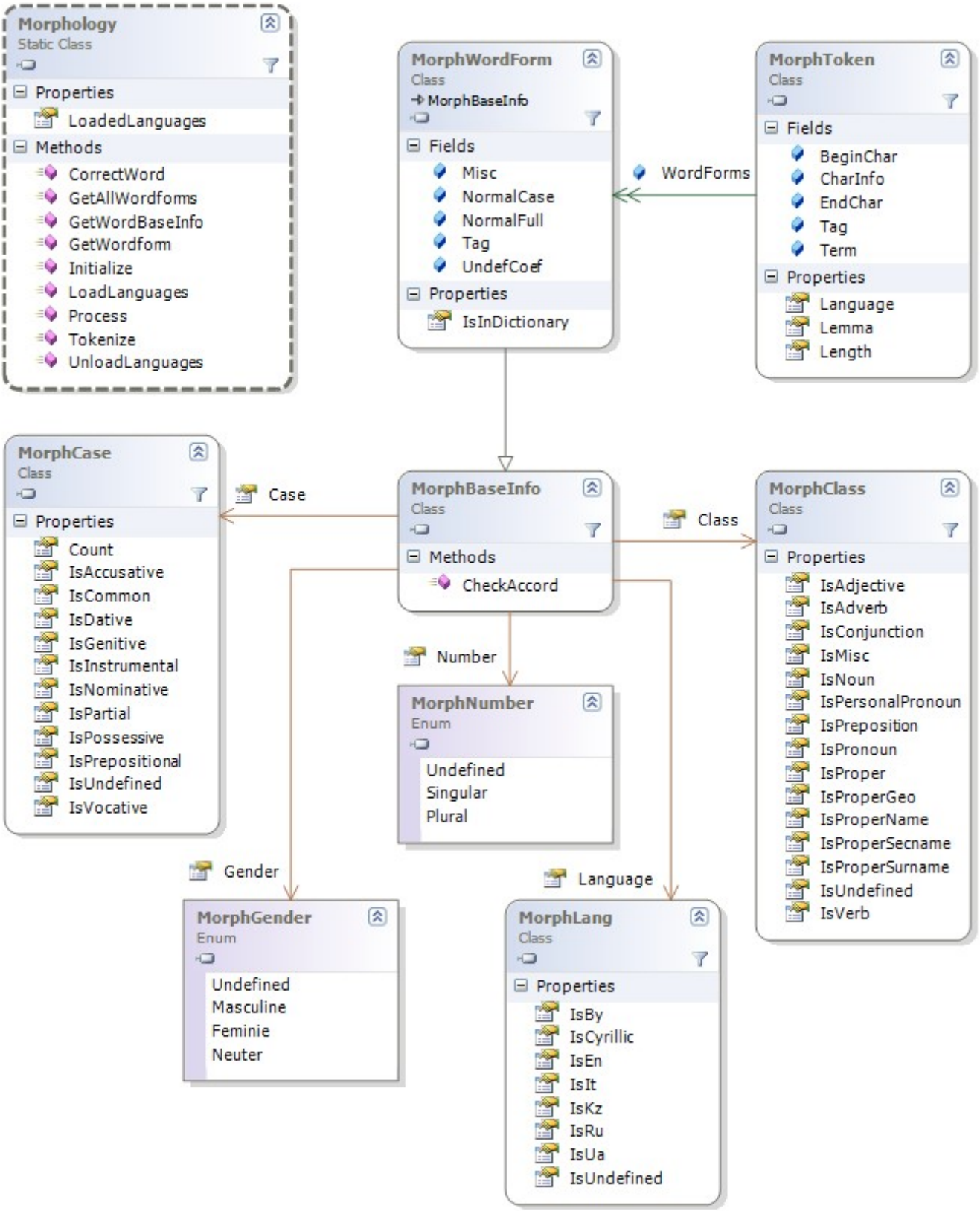

Fig. 2. Morphological analysis

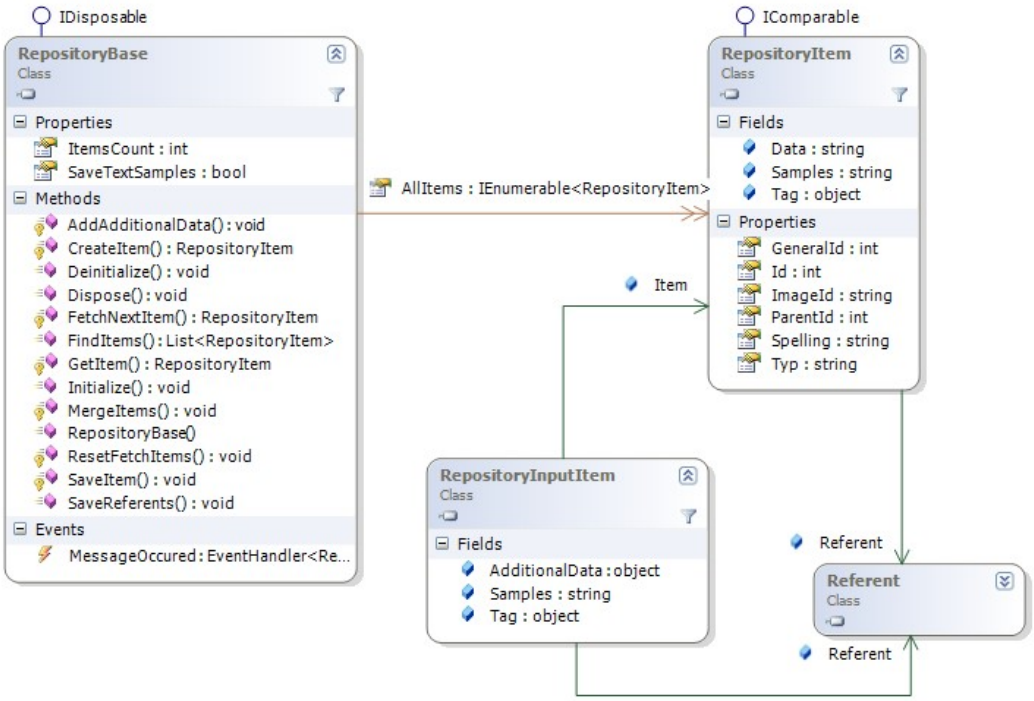

Fig. 3. Highlighting the named entities

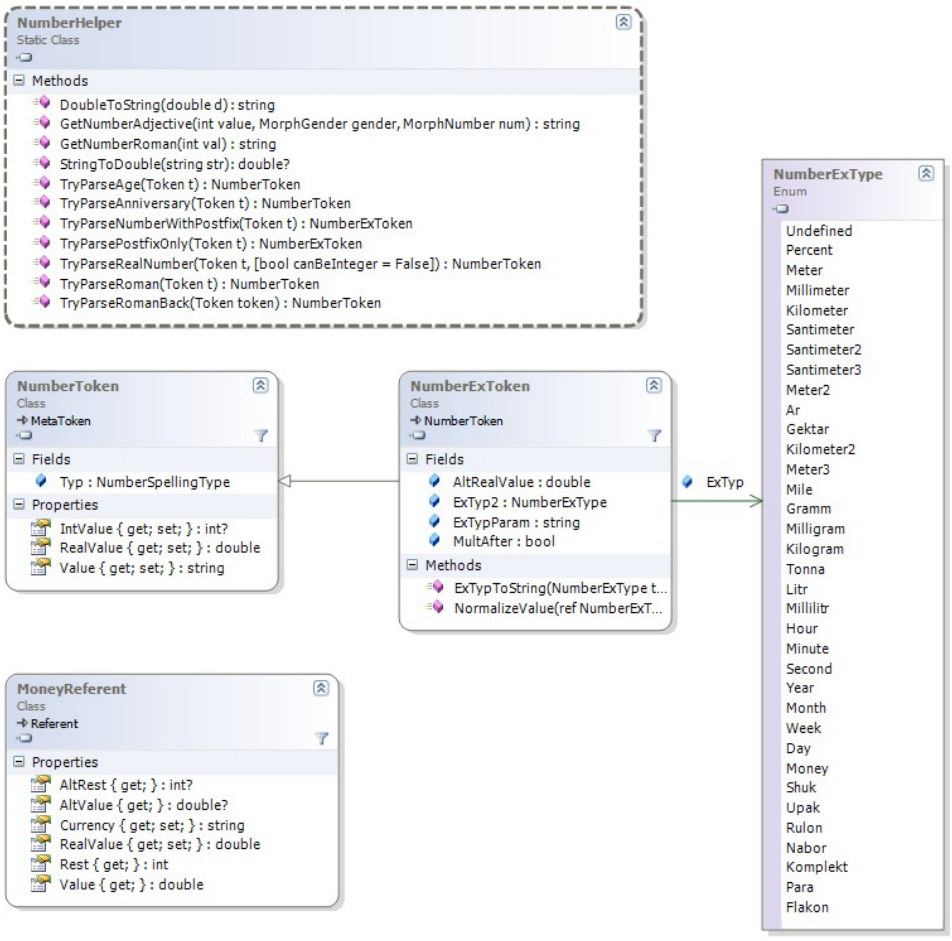

Fig. 4. Numeric Tools

Specially for this project, the linguistic processor has been modified so that it is possible to more accurately highlight implicit links in documents

The concept of a token (Token base class) is at the heart of the Pullenti SDK model. Each token refers to a merged fragment of the source text (BeginChar and EndChar positions). First, the text is divided into a sequence of text tokens (TextToken), and then during processing they are converted - merging into meta-tokens (MetaToken). A metatoken is a token that has "absorbed" a fused sequence of other tokens. Metatokens, for example, represent places of occurrence of named entities (ReferentToken) in the text. Metatokens can represent various numerical data (lowercase spelling of numbers), name groups (in the example, NounPhraseToken is the inherited class from MetaToken), etc. Most of the elements received and used during the analysis are metatokens.

The concept of PullEnti megatokens served as the basis for building dictionaries of megalemmas, each of which can consist of several tokens or megatokens. The megalemma is the basis for comparing meaningful phrases from different languages, i.e. the concept of megalemma is broader than the concept of megatoken, since it additionally includes identifying connections between different languages.

Megalemma dictionaries are constructed using the method for determining the proximity of terms [11]. It is this method that allows us to form megalemmas on the basis of statistical patterns of occurrence of terms in the framework of the formation of an associative-hierarchical portrait of a subject area.

Thematic dictionaries of megalemmas are formed by subject areas and serve as the basis for the classification of texts. Megalemma dictionaries are also used to represent knowledge in ontologies and automatically supplement them with relevant vocabulary.

The formal element for the presentation of concepts was chosen synset. This is the basis of knowledge representation in systems such as Wordnet, Babelnet and others. This is a well-established and generally accepted concept [15]. Synsets can chain together (megalemmas include synsets).

Thus megalemmas are presented - these are chains of synsets. The concept of synset is initially oriented toward multilingualism.

The work was carried out in two subject areas - “computer graphics and visualization” and “autonomous uninhabited underwater vehicles”.

Algorithms for the semantic analysis of information have been developed [2-11, 15]. Prototypes of software components for semantic analysis of textual information have been developed too.

Implicit links are searched using the megalemma dictionary. First, the text is processed using the PullEnti program, normalization of words in the text, selection of named entities (NER - named entity recognition), formation of dictionaries of tokens and megatokens for the text are performed. Next, a thematic analysis of the text is carried out using megalemma dictionaries. In the dictionaries of megalemmas, as already mentioned, there is a correlation of each megalemma with a specific document and with a specific subject area. This allows the classification of texts in subject areas and a statistical analysis of documents for the presence of implicit references. According to the publication date of the document, the source document of the megalemma and the document that has a link to the megalemma are determined.

To control the quality of automatic detection of implicit links, methods of collective intelligence and crowdsourcing were used [17]. It was proposed to conduct a quality check for the detection of implicit links using an expert approach.

The probability of a positive decision is determined by the mathematical model:

$$K_0 = \sum_{i=0}^{M=1/2} C_M^i G_R^{M=i} (1 - G_R)^i$$

In accordance with this formula, the probability K0 of a positive decision by a group of M experts with the probability of the correct GR solution for one expert is determined by this formula. The analysis of expert estimates showed a rather high level of revealing implicit links and determining the semantic similarity of phrases and documents.

There was developed software for storing a multilingual collection of documents. A software implementation of thematic modeling methods using dictionaries of megalemmas in subject areas has been developed [18].

As a result of processing collections of documents, dictionaries of terms and dictionaries of megalemmas are built. Statistics is collected for the use of terms and megalemmas by articles.

BabelNet is an integration resource based on the following resources: WordNet, Wikipedia, OmegaWiki, Wiktionary, Wikidata, Wikiquote, VerbNet, Microsoft Terminology, GeoNames, ImageNet, FrameNet, WN-Map, Open Multilingual WordNet, WoNeF, Albanet, Arabic WordNet (AWN v2), BulTreeBank WordNet (BTB-WN), Chinese Open WordNet, Chinese WordNet (Taiwan), DanNet, Greek WordNet, Princeton WordNet, Persian WordNet, FinnWordNet, WOLF (WordNet Libre du Français), Hebrew WordNet, Croatian WordNet, IceWordNet , MultiWordNet, ItalWordNet, Japanese WordNet, Multilingual Central Repository, WordNet Bahasa, Open Dutch WordNet, Norwegian WordNet, plWordNet, OpenWN-PT, Romanian WordNet, Lithua.

BabelNet is fully integrated with BabelFly's multilingual lexical ambiguity and entity binding system. BabelNet is also integrated with Wikipedia's bitaxonomy [20], which is built around two hierarchies: page hierarchies and category hierarchies [15].

Integration with BabelNet will be carried out by analogy with the approach that BabelNet uses to integrate with other (described above) resources, using automatic display and filling of lexical gaps in languages with limited resources using statistical machine translation. The result is an “encyclopedic dictionary” that provides concepts and named entities lexicalized in many languages and associated with a large number of semantic relations [21]. Additional vocabulary and definitions are added by reference to free networks such as WordNet, OmegaWiki, English Wiktionary, Wikidata, FrameNet, VerbNet and others. Like WordNet, BabelNet groups words in different languages into sets of synonyms called Babel synsets. For each Babel syntax, BabelNet provides short definitions (called glosses) in many languages, taken from both WordNet and Wikipedia.

In the future, it is planned to use the Babelscape product [22], which allows us to analyze documents, perform semantic markup of texts, build semantic knowledge graphs in several languages, etc., but this issue requires additional careful study [15].

The dictionaries of terms and megalemmas proposed within the framework of the project allow not only to classify texts, but also to define implicit links between articles.

The structure of the glossary is represented by a tuple:

$$Dterm = < IDterm, Term>, \quad (1)$$

where Dterm is a glossary of terms, IDterm is a term identifier in a dictionary, Term is a term.

The structure of the megalemma dictionary is represented by a tuple:

$$Dmeg = < IDmeg, MegL>, \quad (2)$$

where Dmeg is the megalemma dictionary, IDmeg is the megalemma identifier in the dictionary, MegL is the megalemma.

The structure of the document dictionary is represented by a tuple:

$$Ddoc = <IDdoc, NAMEdoc, SRCdoc, YEARdoc, NUMwrd>, \quad (3)$$

where Ddoc is the document dictionary, IDdoc is the document identifier in the dictionary, NAMEdoc is the document name, SRCdoc is the publication source, YEARdoc is the publication year, NUMwrd is the total number of terms in the document.

The structure of the domain dictionary is represented by a tuple:

$$Dsa = < IDsa, SA>, \quad (4)$$

where Dsa is the domain dictionary, IDsa is the domain identifier in the dictionary, SA is the domain name.

While the Dterm dictionary is a general glossary of terms, dictionaries of documents contain the terms of the document and the frequency of occurrence of the term in the document. The same thing applies to the dictionary of megalemmas. These two dictionaries are associative tables in the database. An associative table in the database implements a relationship between many-to-many entities.

The structure of the dictionary of terms of the document is represented by a tuple:

$$Dtd = < IDterm, IDdoc, Fterm>, \quad (5)$$

where Dtd is the dictionary of terms of the document, Fterm is the relative frequency of occurrence of the term in the document, calculated as follows: first, all insignificant words are removed from the document (stop words, rare words, etc.), only the terms remain, then the frequency of occurrence of the term is divided by the total number of terms in the document.

The structure of the dictionary of megalemmas of the document is represented by a tuple:

$$Dmd = < IDmeg, IDdoc, Fmeg>, \quad (6)$$

where Dmd is the dictionary of megalemmas in the document, Fmeg is the relative frequency of megalemma in the document, calculated as follows: the frequency of megalemma is divided by the total number of megalemmas in the document.

The structure of the keyword dictionary is represented by a tuple:

$$Dkeywrd = < IDterm, IDsa>, \quad (7)$$

Keywords are taken from a general vocabulary of terms and compared with the subject area. This is also an associative table.

The structure of the dictionary of document correlation with a subject area is presented below.

$$Ddsa = <IDdoc, IDsa>, \quad (8)$$

where Ddsa is a dictionary of subject areas of a document. One document can belong to several subject areas.

4. Results

A program was developed to implement methods for modeling topics and to identify implicit links between documents [23]. The megalemmas' dictionary is used to determine implicit references. The task is to determine the source of the megalemma and link to it. A storage structure and methods for constructing a multilingual collection of synsets - synonymous series are developed.

A neural network algorithm was developed using tags / tokens (flagging) and the Word2vec method modified by the team of authors, already described, to identify Russian-speaking terms in texts that are similar in context of lexical meaning [24].

The methodology for constructing forecasts for the development of new directions includes the ratio of the relative frequencies of occurrence of the same megalemmas calculated over adjacent years. This approach eliminates the problem of retraining neural networks in connection with the accumulation of information.

The analysis of clustering methods and thematic modeling to assess the quality / significance of texts carried out [25]. Various thematic modeling methods are considered, including the vector model, latent semantic analysis, latent Dirichlet placement, and others. The basis of these methods is a probabilistic approach, i.e. correlation of a term or document with several topics with a certain degree of probability. The disadvantage of this approach is the automatic formation of a list of topics.

5. Conclusion

As a result of this scientific research, a number of results will be obtained that have high scientific and applied significance:

1. The updated actual multilingual collection of scientific texts in various languages, containing more than 60 thousand scientific documents and having more than 6 thousand internal bibliographic references. This collection will allow us to accurately calculate the significance of documents using the scientific citation index (SCI) by the number of bibliographic references, as well as using the context scientific citation index (CSCI), calculated by the number of implicit references identified through the semantic similarity of texts.
2. The developed technique for the automatic formation of a multilingual associative-hierarchical portrait of a subject area (MAHPSA) containing a hierarchy of multilingual synonymous series (synsets). With the help of MAHPSA, it is possible to solve a wide range of problems, including calculating the semantic similarity of texts, identifying multilingual plagiarism, expanding queries in multilingual search.
3. The developed methodology and algorithms for calculating integrated multilingual statistics based on MAHPSA, including the identification of significant documents, trends and promising areas. Because of applying the technique to a multilingual collection, new concepts will be revealed, the dynamics of their

development over time will be considered, and promising areas for the development of the subject area will be constructed. Based on this, it will be possible to build forecasts of promising areas of research.

4. The developed methodology for integrating MAHPSA with other ontologies and linguistic resources, including BabelNet, which contains millions of multilingual synsets. As a result, the shortcomings of BabelNet related to the low level of coverage of Russian terms will be overcome. For integrated resources, updated ratings of the significance of documents will be calculated and updated forecasts of promising areas of research in selected subject areas will be constructed.

Acknowledgment

The reported study was funded by RFBR according to the research projects № 18-07-00225, 18-07-00909, 18-07-01111, 19-07-00455 and 20-04-60185.

References

[1] J. Galbraith, and R. Thayer, SECSH Public Key File Format, draft-ietf-secsh-publickeyfile-01.txt, March 2001, work in progress material.

[2] Zolotarev O.V., Sharnin M.M., Klimenko S.V., Kuznetsov K.I. PullEnty system - information extraction from natural language texts and automated building of information systems. In the collection: Situational centers and information-analytical systems of class 4i for monitoring and security tasks (SCVRT2015-16). Proceedings of the International Scientific Conference: in 2 volumes. 2016. P. 28-35.

[3] Zolotarev O.V., Kozerenko E.B., Sharnin M.M. The principles of constructing models of business processes in the subject area based on natural language text processing. Bulletin of the Russian New University. Series: Complex systems: models, analysis and control. 2014. No. 4. P. 82-88.

[4] Zolotarev O.V. Methods and tools for domain modeling. In the collection: The Civilization of Knowledge: Problems and Prospects of Social Communications Proceedings of the XIII International Scientific Conference. 2012. P. 71-72.

[5] Zolotareva V.P., Yashkova N.V., Zolotarev O.V. Project management. Educational-methodical manual / Nizhny Novgorod, 2016.

[6] Zolotarev O.V. Formalization of knowledge about the subject area based on the analysis of natural language structures. In the collection: The civilization of knowledge: the problem of man in science of the XXI century. Proceedings of the XII International Scientific Conference. 2011. P. 78-80.

[7] Zolotarev O.V., Sharnin M.M. Methods of extracting knowledge from natural language texts and building business process models based on the allocation of processes, objects, their relationships and characteristics. In the collection: Proceedings of the International Scientific Conference CPT2014. Institute of Computing for Physics and Technology. 2015.P. 92-98.

[8] Sharnin M.M., Zolotarev O.V., Somin N.V. Extracting and processing knowledge from unstructured texts of the business sphere and social networks. In the collection: Social computing: fundamentals, development technologies, social and humanitarian effects Materials of the Fourth International Scientific and Practical Conference. 2015. P. 364-371.

[9] Zolotarev O.V., Kozerenko E.B., Sharnin M.M. Analytical intelligence based on the analysis of unstructured information from various sources, including the Internet and the media. Bulletin of the Russian New University. Series: Complex systems: models, analysis and control. 2015. No 1. P. 49-54.

[10] Zolotarev O.V. New approaches in constructing the functional structure of the subject area. In the collection: Twenty Years of Post-Soviet Russia: crisis phenomena and modernization mechanisms materials of the XIV All-Russian Scientific and Practical Conference of the Humanitarian University: in 2 volumes. Humanitarian University. Ekaterinburg, 2011. P. 639-643.

[11] Zolotarev O.V., Sharnin M.M., Klimenko S.V. A semantic approach to the analysis of terrorist activity on the Internet based on thematic modeling methods.

[12] Zolotarev O.V., Sharnin M.M., Klimenko S.V. Bulletin of the Russian New University. Series: Complex systems: models, analysis and control. 2016. No. 3. P. 64-71.

[13] Kozerenko E. B., Kuznetsov K. I. Romanov D. A. Semantic processing of unstructured textual data based on the linguistic processor PullEnti Informatics and applications 2018 volume 12 issue 3. DOI: 10.14357/19922264180313, pp. 91–98

[14] Chiu, J.P. and Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308.

[15] Peters M. E. et al. Deep contextualized word representations //arXiv preprint arXiv:1802.05365. - 2018.

[16] Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic net-work.Artificial Intelligence, 193:217-250.

[17] John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez. Semantic Web Programming. - John Wiley & Sons, 2009. - 648 c.

[18] V.I.Protasov, Z.E.Potapova, R.O.Mirakhmedov, M.M. Sharnin, Minasyan V.B. Methods for finding solutions by a group actor with a low probability of error. In the collection of CPT2019. Materials of the international scientific conference of the Nizhny Novgorod State University of Architecture and Civil Engineering and the Scientific and Research Center for Information in Physics and Technique. 2019, Nizhny Novgorod. P. 284-291.

[19] Brickley D., Guha R.V. RDF vocabulary description language 1.0: RDF schema W3C working draft. 2002. http://www.w3.org/TR/2002/WD-rdf-schema-20020430/.

[20] Ehrmann M., Cecconi F., Vannella D., McCrae J.P., Cimiano P., Navigli R. Representing Multilingual

Data as Linked Data: the Case of BabelNet 2.0. - LREC (2014). - 2014. - URL: http://wwwusers.di.uniroma1.it/~navigli/pubs/ LREC_2014_Ehrmannetal.pdf.
[21] T. Flati, D. Vannella, T. Pasini, R. Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA, June 22-27, 2014, pp. 945-955.
[22] Ustalov, D., & Panchenko, A. (2017). A tool for effective extraction of synsets and semantic relations from BabelNet. B Proceedings - 2017 Siberian Symposium on Data Science and Engineering, SSDSE 2017 (стр. 10-13). [8071954] Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/SSDSE.2017.8071954
[23] R. Navigli, S.P. Ponzetto, BabelNetXplorer: a platform for multilingual lexical knowledge base access and exploration, in: Companion Volume totheProceedings of the 21st World Wide Web Conference, Lyon, France, 16–20 April 2012, pp. 393–396.
[24] Lau J.H., Newman D., Karimi S., Baldwin T. Best Topic Word Selection for Topic Labelling // COLING'10 Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010. Pp. 605-613.
[25] Google Cloud Machine Learning [CD] - https://cloud.google.com/ml-engine/docs/tutorials/python-guide.
[26] Xie Pengtao, Xing Eric P. Integrating document clustering and topic modeling. arXiv preprint, arXiv:1309.6874. 2013.

About the autors

Zolotarev Oleg V., Ph.D., Docent, ANO HE «Russian New University» (Moscow, Russia), E-mail: ol-zolot@yandex.ru