Semantic Proximity Establishment in the Tasks of Knowledge Extraction and Named Entities Recognition

Kozerenko E.B.[1], **Kuznetsov K.I.** [1], **Morozova Yu.I.** [1], **and Romanov D.A.**[2]
[1] IIP FRC CSC RAS (Institute of Informatics Problems, Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences), Moscow, Russia
[2] HSE (National Research University Higher School of Economics), Moscow, Russia

Abstract - *The paper deals with the problem of establishing text segments containing the similar semantic units for the tasks of analytical text processing within the semantic technology platform. The methods and instruments presented in the paper provide the discovery of relevant content based on users' focused interests within a certain domain. The hybrid approach comprising linguistic rules and example-based learning techniques is employed. The legal and mass media texts are considered. In this paper a brief description of the NER task history is cited, the Pullenti-based engine is specified, the two-step Semantic Expansion Algorithm is presented, the Distributional Semantics methods for domain terms extraction are discussed as well as some technical challenges and the prospective directions of further research and development.*

Keywords: Semantic Analysis, Natural Language Processing, Named Entity Recognition, Rule-Based Approach, Hybrid Systems, Distributional Semantics

1 Introduction

The problem of establishing semantic similarity within different text documents is very urgent nowadays for a wide range of natural language processing tasks. In this paper the approach and instruments for discovery of text segments containing similar semantic units are presented for the tasks of analytical text processing within the semantic technology platform. The methods and tools described in the paper provide the discovery of relevant content based on users' focused interests within a certain domain. The hybrid approach comprising linguistic rules and example-based learning techniques is employed. The legal and mass media texts are considered.

The Named Entity Recognition (NER) is a key feature of the technology platform described in this paper. We discuss here the use of NER for establishing semantic similarity in Russian legal texts in connection with the regulations issued by the Constitutional Court and Ruling Legal Acts. The kernel of the technology platform is the Pullenti-based engine for selecting entities and structuring information. The feature set of the engine comprises common morphological algorithms and parsing. Modules are focused on the identification of specific entities. The system integrates the seed feature-value unification grammar which is extensible by creating new rule-based modules and incremental machine-learning components extracting phrase structures from texts under study.

In this paper a brief description of the NER task history is cited [1-35], the Pullenti-based engine is specified, the two-step Semantic Expansion Algorithm is presented, the Distributional Semantics methods for domain terms extraction are discussed as well as some technical challenges and the prospective directions of further research and development.

2 NER as a subtask of knowledge extraction

The term "Named Entity" which is now widely used in Natural Language Processing was introduced for the Sixth Message Understanding Conference (MUC-6) by R. Grishman and Sundheim [12]. The main efforts of the works presented at MUC were directed at Information Extraction (IE) tasks where information of company activities and defense related activities was extracted from unstructured text, such as newspaper articles. In specifying the task the researchers noticed that it was essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)". In the expression "Named Entity", the word "Named" is used to restrict the task to only those entities for which rigid designators, as defined by S. Kripke [15], stand for the referent. Rigid designators include proper names as well as certain terms like biological species and substances.

There is a general agreement in the NERC community about the inclusion of temporal expressions and some numerical expressions such as amounts of money and other types of units. Not all instances of these types are good examples of rigid designators there are also many not so rigid and incomplete ones which creates a major difficulty for establishing them in unstructured natural language text. In early works the NERC problem was formulated as recognizing "proper names" in general [5, 32]. The most studied types are three specializations of "proper names":

names of "persons", "locations" and "organizations". These types are collectively known as "enamex" since the MUC-6 competition. The type "location" can be subdivided into multiple subtypes of locations: city, state, country, etc. [9, 19]. The "fine-grained person" sub-categories like "politician" and "entertainer" were introduced in the work of M. Fleischman and Hovy in 2002 [10]. The type "person" which is common for all, in the work of O. Bodenreider and Zweigenbaum [3] was combined with other key words for extracting medication and disease names (e.g., "Parkinson disease"). The subsequent work did not limit the possible types for extraction and now is referred to as "open domain" NERC, as in E. Alfonseca and Manandhar [1], R. Evans [8]. In their research, S. Sekine and Nobata [29] defined a named entity hierarchy which includes many types of subcategories, such as museum, river or airport, and add a wide range of categories, such as product and event, as well as substance, animal, religion or color. The works in this direction try to cover most frequent types of names and designators appearing in Mass Media. The number of categories proposed in recent work is more than 200, and they are defining popular attributes for each category to create an ontology.

The above mentioned developments are mainly concerned with the English language, however the advanced in NER and knowledge discovery have been made for the Russian language as well by Osipov G.S., Smirnov I.V. et al.[24]; Ermakov A.E., [7]; Kuznetsov I.P., Kozerenko E.B. et al. [16,17]; Anisimovich K.V., et al. [2]; Bolshakova E., Loukachevitch N., et al.[4]; Efimenko I.V., Khoroshevsky V. F. [6]; Zolotarev O.V., Charnine M.M. [35].

The NER engine Pullenti and the technological platform presented in the paper are the developments of Kuznetsov Konstantin Igorevich and the research group of the Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences and National Research University Higher School of Economics

3 Pullenti Engine Features and Implementation

The semantic engine employed for the tasks described in the paper is based on the Pullenti technological platform by Semantic [30], the Named Entity Recognition (NER) being its kernel feature.

Pullenti is used as the engine for selecting entities and structuring information. Recognition of named entities from unstructured texts in the Russian language (Named Entity Recognition for Russian Language) is provided. The following types of entities are extracted: persons, organizations, dates, countries, rulings, legal acts, etc. The identification of entities is based on rules. Some entity types can be defined by means of external dictionary entries, if any (for example, prepared lists of employees or organizations). In a limited mode the system already works with Ukrainian texts.

The system is implemented as a *software development kit* (SDK) for information systems development dealing with unstructured data, i.e. texts in natural languages. This SDK is very convenient for use in the systems developed in .NET environment. For Mac and Linux systems it is running on Mono platform.

The SDK is completely written in C # and .NET, it is a set of assemblies of .NET Framework (2.0 and above). Third-party extensions are not used. If necessary, one can take the form of a standard service for unstructured information processing UIMA (Unstructured Information Management Architecture).

In Pullenti design a great attention is given to the quality of attributes associated with the entities and their case normalization. The emphasis is made on the Russian language, however, other languages can also be included. At present the work is underway to set up the Ukrainian language. The system is constantly being updated and improved basing on real data. An upgraded version is released once in 10-12 days.

The system architecture is determined by the demand for efficiency, extensibility and flexibility. The system consists of the kernel and dynamically attached modules. The core of the system contains common morphological and syntactic parsing algorithms. The modules are focused on the recognition of specific entities. The system is extended by creating new modules.

At the current stage of its development, the system does not allow third-party developers to create their own modules. The implementation of new modules is carried out by the team of Pullenti developers. However, it is easy to reach the level of full disclosure if required. Meanwhile, the third-party developers can upload their API-level external entities, if any.

The system development evolves towards a more sophisticated semantic analysis. Since version 2.24 for realization of semantic analysis, a semantic network based on parsing the text is generated. Partial visualization is implemented in the demo option, but it should be borne in mind here that we are only at the beginning. Since version 2.38 the works for the extraction of facts and relationships among entities based on semantic analysis have been carried out.

The system is free for non-commercial use and it can be downloaded. For commercial use the SDK is shipped without restrictions on the number of end users and installations.

The work speed can be very approximately estimated as 60.000-80.000 characters per second on a serial computer model. The maximum amount of text on a 32-bit computer that can be processed at one time is no more than 20 MB. On a 64-bit processor the result depends on the size of the operational memory.

Morphology processing is implemented in Assembly EP. Morphology.dll and can be used independently from the SDK. It comprises a POS-Tagger for the Russian, Ukrainian and English languages, the approximate speed of operation

is 2 MB per second. A very important feature is Morphological Conjecture: it builds the morphological parses for unknown words and establishes the normal form of the words (lemmas).

Grammar formalisms underlying the system are based on morphological-syntactic feature-value structures, regular expressions, context-free and mildly context-sensitive grammars and the unification. Since the module of semantic networks construction being a novel one, the design choice was made in favour of the flat flexible phrase-based presentations which provide sufficiently robust and speedy performance for mass applications. The previous developments dealing with the Extended Semantic Networks (ESN) (Kuznetsov I.P. et al., 2008) are not used in the technology presented in this paper

4 Pullenti-based projects and their evaluation

A number of projects have been implemented on the Pullenti engine platform, consider some of them.

Legal Expertise: "Legal Expertise" is the legal documents examination system designed to automate the process of examination of ruling legal acts (RLA) projects, organizational administrative documents, contracts and other documents. The analysis of the documents is produced directly in MS Word, the instances of possible incorrectness are indicated. Text links to external documents that are present in the database are processed automatically and hyperlinked.

AIS of the State Duma: the engine is used in the "Automated Information System" of the State Duma (the Russian Parliament) for the analysis of meetings minutes and extraction of the meaningful information from them to automatically populate a database.

Experts: the system "Expert Search" allows HR and other users to make search for competent staff among the employees of an Organization basing on issues and topics which are defined in a free form (for example, a text query or a document).

Dr. Watson: "Dr. Watson" is a system designed to study textual information arrays in order to identify entities and relationships between them. The result is given out in the form of the generated report on the investigated object. The program is designed for analysts working with excessive text data, professional security services, competitive intelligence, marketing, PR, etc.

Government purchases: the module is processing procurement information performing the extraction of the data on positions, quantities, prices, suppliers; normalization of trade names. The module is used in the system for loading and analyzing the solicitation information.

Lawyer (Aktion): the "Lawyer" system provides the online service to check the texts of contracts the analysis of the legal risks. The service allows to recognize the type of contract, determine significant conditions (date, time, amount, liability risks, counterparty details, etc.).

Particular attention in the developments for legal texts processing is given to the establishment of semantic similarity. The Pullenti-based technology is a novel development designed about four years ago. However, depending on the objectives of each particular project, the subject area and the complicity of extracted entities the Precision (P), Recall (R) and F-measure demonstrated by Pullenti for the NER tasks are competitive: P = 0.7957-0.9663; R = 0.7091-0.8675; F-measure = 0.8083-0.8570. We shall not go in deep analysis of the numbers in this paper, since the Pullenti-based engine participated in the NER competition organized for the Dialog 2016 Conference, and showed the best results in the two tracks and the second best result in the third track.

5 Semantics Expansion Algorithm

An example of a non-trivial use of the NER feature for the task of similar semantic objects discovery is the Semantic Expansion Algorithm (SEA). We discuss here the use of NER for establishing semantic similarity in Russian legal texts in connection with the regulations issued by the Constitutional Court and Ruling Legal Acts.

The two steps of the Semantics Expansion Algorithm (SEA) are presented in Fig. 1 and Fig. 2. Consider Step 1: in the top left window there is a text of a newly received application to the Constitutional Court (CC). At the bottom left inside there is a list of links to the profile of the applications traced by links to Ruling Legal Acts (RLA) and legal norms (LN) (meaningful terms and bigrams). In the middle there is a ranked list of earlier regulations and determinations of the Constitutional Court (i.e. the "precedents").

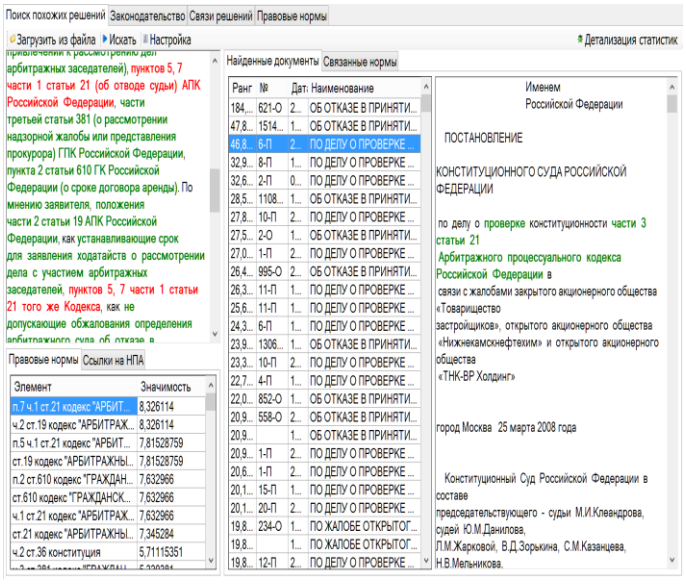

Figure 1. Step 1 - Significant terms identification

With the cursor one can select the necessary document. The right panel is the text of the application selected with the cursor. A special semantic algorithm is used to expand the meaningful terms from a user request (the newly received application) with the strongly-related terms and concepts (associated legal norms) identified when analyzing the whole array of decisions of the Constitutional Court.

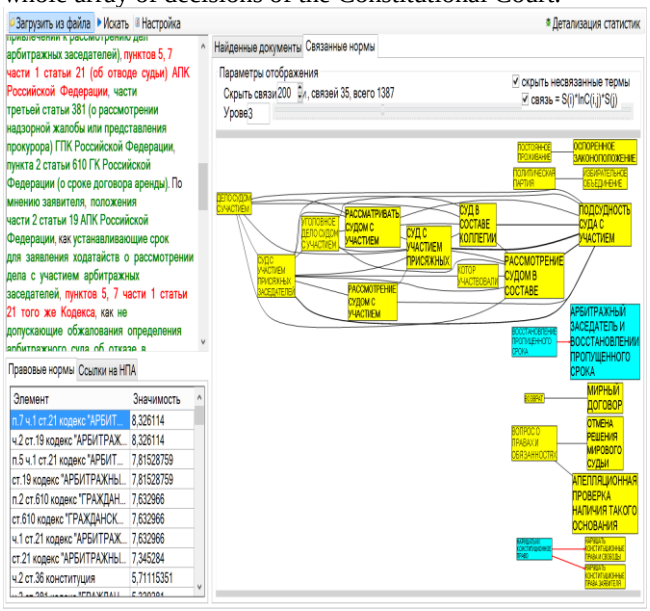

Figure 2. Step 2 - Semantic Expansion results

Here is an example of a ruling of the Constitutional Court. The identified entities are Persons, Organizations (their significance is cleared away by the algorithm - the null value is assigned - to shift the focus on the legal norms) and links to Normative Legal Acts (NLA) (their significance is forcibly increased).

The support for external ontologies (dictionaries) is provided. The extraction of entities is based on rules. However, for some types of entities one can load external vocabularies (ontologies), containing descriptions of the existing entities, and then in the allocation system it is possible to link to external entities by setting the OntologyItems Referent instances field. Dictionary class implements ExtOntology, containing a list of items that one wants to add, for that the ExtOntologyItem function Add (...) is employed. The dictionary can contain elements of any of the supported types.

6 Distributional semantics methods in establishing semantic similarity

One of the methods to be developed within the NER technology platform is connected with the distributional semantics employment for semantic similarity establishment as stated in the works by Sparck Jones K. [31].

Distributional semantics is a field of scientific research that aims at calculation of semantic proximity between different linguistic units using their distributional properties in large linguistic corpora. The distributional models are used in numerous research projects dealing with semantics of natural language and have a diverse range of potential and working applications. The main application areas of distributional semantics models are: lexical ambiguity resolution, information retrieval, document clustering, automatic extraction of lexicographic information (dictionaries of semantic relations, multilingual dictionaries), semantic maps of different domains, modeling of synonymy, document topic detection, sentiment analysis, bioinformatics.

The theoretical foundations of distributional semantics go back to the distributional methodology of Z. Harris [13. 14]. Similar ideas were expressed by the founders of structural linguistics F. de Saussure and L. Vitgenstein. The theoretical basis of distributional models is the distributional hypothesis stating that linguistic units with similar distributions have similar meanings, e.g. in Sahlgren M., 2008 [27]; Turney P. D., 2010 [33].

Linear algebra is used as the computational instrument and as the means of model representation. First the information on linguistic units distribution is represented in the form of multidimensional vectors. These vectors constitute a matrix, in which vectors correspond to linguistic units (words or word combinations) and dimensions correspond to contexts of different sizes (documents, paragraphs, sentences, word combinations, words). When the matrix is populated from texts, semantic proximity between linguistic units can be calculated as the distance between vectors.

To compute the distance between vectors one can use various formulas: Minkowski distance, Manhattan distance, Euclidean distance, Chebyshev distance, scalar product, cosine measure. The most widely used formula (1) is the cosine measure:

$$\frac{x \bullet y}{|x| \bullet |y|} = \frac{\sum_{i=1}^{n} x_i \bullet y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \bullet \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

There are many different types of distributional semantics models which differ according to the following parameters:

- type of the context (its size, left or right, ranking);
- measure to calculate frequency of a word in a given context (absolute frequency, entropy, mutual information etc.);
- method used to compute the distance between vectors (cosine measure, scalar product, Minkowski distance etc.);
- method of reducing matrix dimensionality (Random Projection, Singular Value Decomposition etc.).

The most popular distributional semantics models are latent semantic analysis which was designed to solve the synonymy problem in information retrieval as shown by Landauer Th. K., [18], and the model of hyperspace

analogue to language thought as the model of human semantic memory stated by Lund K. [20].

The idea of semantic vector spaces was first realized in the information system SMART by Salton G. M. [28]. Documents from a text collection are represented as vectors in a vector space. A user inquiry is viewed as a pseudodocument and is also represented as a vector in the same vector space. The system finds *n* vectors of documents which are closest to the vector of the inquiry. The results are sorted by distance between vectors which reflects semantic proximity and are displayed to the user.

Later, the idea of semantic vector spaces was applied successfully for other semantic tasks. For example, in the research by Rapp R. [25] a vector space was used to evaluate semantic proximity of words. The system reached the accuracy level of 92.5% on TOEFL tests to choose a synonym of a set of words, mean human result for this test being 64.5%.

Comprising the techniques of distributional semantics into the technological platform described in this paper is under way now, and the results will be tested and reported to the NLP community.

7 Conclusions

Since the NER task is just a subtask of Information Extraction and Knowledge Discovery, the natural way to evolve the algorithms and data structures of the development presented here is to introduce more powerful features into the linguistic processor, i.e. include additional languages and services. At present the efforts are aimed at the design of the bilingual Russian-English language engineering environment.

8 References

[1] Alfonseca, E.; Manandhar, S. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. Proceedings of the International Conference on General WordNet. 2002.

[2] Anisimovich K.V., Druzhkin K. Ju., Minlos F.R., Petrova M. A., Selegey V. P., Zuev K.A. Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"], Bekasovo. Vol. 2, P. 91-103. 2012.

[3] Bodenreider, O.; Zweigenbaum, P. Identifying Proper Names in Parallel Medical Terminologies. Stud Health Technol Inform 77.443-447, Amsterdam: IOS Press. 2000.

[4] Bolshakova E., Loukachevitch N., Nokel M. Models Can Improve Domain Term Extraction // ECIR Proceedings. LNCS. Ed. SPRINGER HEIDELBERG. Vol. 7814. P. 684–687. 2013.

[5] Coates-Stephens, S. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. Computers and the Humanities 26.441-456, San Francisco: Morgan Kaufmann Publishers. 1992.

[6] Efimenko I.V., Khoroshevsky V. F. New Technology Trends Watch: An Approach and Case Study. Lecture Notes in Computer Science. No. 8722. P. 170-177. 2014.

[7] Ermakov A.E. Automatization of an Ontological Engineering for Systems of Knowledge Mining in Text [Avtomatizatsiya ontologicheskogo inginiringa v sistemakh izvlecheniya znaniy iz teksta] Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Bekasovo. 2008.

[8] Evans, R. A Framework for Named Entity Recognition in the Open Domain. Proceedings of the Recent Advances in Natural Language Processing. 2003.

[9] Fleischman, M. Automated Subcategorization of Named Entities. Proceedings of the Conference of the European Chapter of Association for Computational Linguistic. 2001.

[10] Fleischman, M.; Hovy. E. Fine Grained Classification of Named Entities. Proceedings of the Conference on Computational Linguistics. 2002.

[11] Ganitkevitch J., B. V. Durme, and Callison-Burch C. PPDB: The paraphrase database. In Human Language Technology and North American Association for Computational Linguistics (HLT/NAACL), pp.758–764. 2013.

[12] Grishman, R.; Sundheim, B. Message Understanding Conference - 6: A Brief History. Proceedings of the International Conference on Computational Linguistics. 1996.

[13] Harris Z.S. Papers in Structural and Transformational Linguistics. Dordrecht, Reidel. 1954.

[14] Harris Z.S. Mathematical Structures of Language. New York,. 1968.

[15] Kripke, S. Naming and Necessity. Boston: Harvard University Press.1982.

[16] Kuznetsov I.P., Kozerenko E.B. Linguistic Processor "Semantix" for Knowledge extraction from natural texts in Russia and English. Proceedings of International Conference on Artificial Intelligence, ICAI'2008, 14-18 July, 2008, Las Vegas, USA, CSREA Press. Vol. II, P. 835-841. 2008.

[17] Kuznetsov I.P., Kozerenko E.B., Kuznetsov K.I. and Timonina N.O. Intelligent System for Entities Extractions (ISEE) from Natural Language Texts. Proceedings of the SENSE Workshop on conceptual Structures for Extracting Natural language SEmantics, Moscow, Russia, July 2009. Edited by Uta Priss, Galia Angelov. Available at: http://ceur-ws.org/Vol-476/paper3.pdf. 2009.

[18] Landauer Th. K., McNamara D. S., Dennis S., Kintsch W. Handbook of Latent Semantic Analysis. Mahwah New Jersey. 2007.

[19] Lee, S.; Geunbae Lee, G. Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping. Proceedings of the International Joint Conference on Natural Language Processing. 2005.

[20] Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence // Behavior Research Methods, Instruments & Computers, 1996, 28(2). p. 203-208.

[21] Mikheev, A., C. Grover, and M. Moens. (1998) Description of the LTG system used for MUC-7. In Proceedings of the 7th Message Understanding Conference (MUC-7), pages 1–12, Fairfax, VA.

[22] Mikheev, A. (1999) A Knowledge-free Method for Capitalized Word Disambiguation. Proceedings of the Conference of Association for Computational Linguistics.

[23] Mikheev, A.; Moens, M.; Grover, C. Named Entity Recognition without Gazetteers. Proceedings of the Conference of European Chapter of the Association for Computational Linguistics. 1999.

[24] Osipov G.S., Smirnov I.V., Tikhomirov I.A., Zavjalova O. Application of linguistic knowledge to search precision improvement. 2008 4th International IEEE Conference Intelligent Systems. 2008.

[25] Rapp R. Word sense discovery based on sense descriptor dissimilarity // Proceedings of the 9th MT Summit. New Orleans, LA, P. 315–322. 2003.

[26] Reddy S., M. Lapata, and Steedman M. Large-scale semantic parsing without question-answer pairs. Transactions of the Association for Computational Linguistics (TACL), 2(10):377–392. 2014.

[27] Sahlgren M. (2008). The Distributional Hypothesis. From context to meaning // Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), volume 20, numero 1, p. 33-53.

[28] Salton G. M. The SMART Retrieval System: Experiments in Automatic Document Processing. - Prentice-Hall, 1971.

[29] Sekine, S.; Nobata, C. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. Proceedings of the Conference on Language Resources and Evaluation. 2004.

[30] Semantic: Technology Platform available at: http://semantick.ru/ .2013-2017.

[31] Sparck Jones K. A Statistical Interpretation of Term Specificity and its application in retrieval. Journal of Documentation. No 28. P. 11-21. 1972.

[32] Thielen, C. An Approach to Proper Name Tagging for German. Proceedings of the Conference of European Chapter of the Association for Computational Linguistics. SIGDAT. 1995.

[33] Turney P. D., Pantel P. From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR), №37, p. 141-188. 2010.

[34] Zettlemoyer L. S. and M. Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In Uncertainty in Artificial Intelligence (UAI), pp. 658–666. 2005.

[35] Zolotarev O., Charnine M., Matskevich A., Kuznetsov K. "Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet", Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), WORLDCOMP'15, July 27-30, Las Vegas Nevada, USA, v.I, pp.295-299. 2015.