Development of a method for extracting spatial data from texts for visualization and information decision-making support for territorial management

B M Pilecki², A V Vicentiy¹,²

¹Institute for Informatics and Mathematical Modeling – Subdivision of the Federal Research Centre “Kola Science Centre of the Russian Academy of Science”, Apatity, Russia
²Apatity branch of Murmansk Arctic State University, Apatity, Russia

gideon.stl@gmail.com, alx_2003@mail.ru

Abstract. Information support for decision making for planning activities in the Arctic and management of the Arctic territories of the Russian Federation using spatial information visualization tools is an important and urgent task. To effectively address this challenge, it is necessary to develop specialized information tools that focus on the analysis, processing, and visualization of large amounts of information to support decision-making. Such tools include methods and technologies for extracting spatial data from texts in natural language for their further visualization. This paper describes the first part of a large complex of research works aimed at creating methods and tools for visualization of spatial data extracted from texts in natural language. At this stage of the work the technology of geodata extraction from the Arctic texts written in Russian is being developed. The paper also substantiates the prospects of application of the created methods and tools to support decision making in the field of territory planning and regional management. The review of named entities recognition problems on the basis of the analysis of scientific works executed in various areas of processing and the analysis of texts is resulted. The main approaches to text analysis and the most significant results achieved so far in this area are considered. The technology developed by the authors to extract spatial data from texts to support decision-making in planning and management of territories is part of a larger system to support socio-economic development of the region. This system is primarily focused on the Arctic territories of the Russian Federation and includes modules for analysis and visualization of geodata identified in the text analysis. The paper describes the research problem, the methods and tools used for text analysis, as well as the main results.

1. Introduction
Managing large spatially distributed bio-socio-economic systems is a complex, non-trivial task. The complexity of planning and management of territories in the Arctic is even greater due to the need to take into account the specific natural and geographical characteristics of the High North [1, 2]. At the same time, the Arctic zone of the Russian Federation (AZRF) is one of the most promising sources of mineral resources needed for sustainable socio-economic development.

One way of reducing the complexity of territory management tasks is to develop various Decision Support Information Systems (DSS). For effective planning of territories, construction, location of engineering and energy systems, it is necessary to consider their geographical location. Therefore, the
The development of subsystems and separate modules for visualizing geographic data within decision support systems is an urgent task [3]. This paper describes the first stage of work devoted to the creation of methods and technologies of spatial data visualization for decision support. The peculiarity of the proposed methods and technologies is the extraction of geodata from the texts presented in the natural language. For these purposes, a special text analyzer (parser) is used. This parser is configured in such a way as to identify geo-referenced information in texts. The result of the work of the proposed methods and technologies is a set of words and phrases with a spatial reference. This set is an input data set for the following stages of work: geocoding and geo-visualization. This paper describes the problem of the text analysis stage, the main methods and technologies, the results of the analysis, as well as plans for further development.

Given the fact that currently more than half of all digital data are spatially linked [4], the use of geodata to support decision-making seems very promising. On the other hand, most of the information is presented in unstructured form. Most often these are texts in natural language (official documents, reports, projects, articles, posts in social networks, tweets, etc.). Therefore, additional efforts are needed to extract, store and process spatial data.

Spatial data are digital data about different spatial objects, which include information about their location and properties, spatial and non-spatial attributes. Such data consist of interrelated parts: position data (description of the spatial position) and non-position data (thematic data content). Thus, a complete description of spatial data consists of interrelated descriptions of the topology, geometry and attributes of geo-objects. Such data, given their semantic environment, form the basis of geographic information systems (GIS) information support. The concept of spatial data can be extended by taking into account their variability. In this case, so-called spatial and temporal data are considered.

In databases, spatial data attributes can be represented as a relational model. Such a model is usually called a geo-relational data model. Such a model includes all views based on the support of the attributive part of the data. The following criteria are used to determine the quality of spatial data: accuracy, reliability, validity, completeness, and consistency. It should also be noted that a multitude of such data can be used to define various operations. In particular, these are export, import, input, exchange, pre-processing, processing, analysis, output, visualization and other operations implemented within the functionality of modern GIS.

The high importance of using spatial data for territory management is noted by many modern authors. Thus, the papers [5, 6] analyze the use of spatial information as a factor of management and decision support, and describe the existing economic and static methods of modeling, taking into account the geodata. In these papers special attention is paid to the synergetic effect of using spatial information to manage economic activity.

The geoinformational approach is widely used in the tasks of controlling different modes of transport. In works [7-9] the processes of control of the distributed transport streams and systems are analyzed, approaches to the solution of transport routing control tasks are offered. Various approaches to multimodal transportation in AZRF conditions are considered and the main components of transport spatial management are described.

In papers [10, 11] geodata are considered as the main resource of cadastre maintenance, organization of cadastre systems and different methods of land assessment. Stratification of geodata used for solving cadastre tasks is described. Digital maps are considered as the main tool of geodata visualization during cadastre maintenance.

Thus, it can be stated that the application of spatial data allows successfully solving various applied and scientific tasks related to management of spatially distributed objects and systems. At the same time, many modern geographic information systems and decision support systems do not have standard text processing tools for geodata identification. As a rule, in GIS and DSS there are no tools of automatic identification, structuring, storage, and visualization of the geodata received at the analysis of texts in natural language. In this regard, the creation of methods, technology, algorithms and software to extract geodata from texts in natural language for their subsequent geo-visualization is an actual scientific and applied task.
2. Background

The problem of extracting geodata from texts belongs to the class of problems with Natural Language Processing (NLP). In [12] NLP is described as a set of artificial intelligence and linguistics designed for computers to understand statements or words written in natural language. Processing the natural language makes it easier for the user to work and communicate with the computer in natural language. In NLP review [13] the process of natural language analysis is considered as a sequence of several processing levels. The main levels are: syntax, morphology, semantics, pragmatics.

More common problems with NLP are Named-entity Recognition (NER) and Automatic Content Extraction (ACE) problems. Different Named-Entity recognition systems are designed to search for and classify mentioned named entities in unstructured or semi-structured texts. re-defined entity categories are used to improve the recognition quality. These categories may include, for example, names of organizations, geo-objects, various toponyms, etc. [14].

One of the main problems arising in the identification of toponyms and other geo-objects is the problem of ambiguity. Different approaches are offered to solve this problem. In particular, in the works [15, 16] the use of so-called "influential terms" is proposed. To detect influential terms in the text it is necessary to take into account both type of geo-object and type of description of geo-referencing of an object. The type of geo-object may be, for example, a city or region. When analyzing texts, a hierarchy of geo-object types is used. Postcodes, phone numbers, or place names can be used to define a geo-referencing location. It should be noted that with this approach, an influential term can refer to several geo-objects at once.

It should be noted that efficiency of one or another text analysis method often depends on the type of texts. Thus, methods that work well for relatively long texts (laws, decrees, books, articles, etc.) may work badly for texts obtained from social networks and messengers. Social networks, messengers and blogs are characterized by texts of small volume written using informal abbreviations, neologisms and other elements of informal language. When working with such texts there are difficulties associated with low volume of context and recognition of individual words [17]. Therefore, the analysis of texts from social networks and microblogs is often highlighted as a separate problem.

The Automatic Content Extraction approach solves the problems of extracting information from various text sources as well as multimedia. ACE identifies three main problems: 1) identification of entities mentioned in the text, 2) identification of relationships between entities, 3) identification of events mentioned in the text. Works [18, 19] describe a method for extracting events from social media news texts based on the use of dictionaries and ontology of the subject area. The authors suggest using the structure of Subject-Verb-Object (SVO) as the main event structure. The paper [20] is devoted to the extraction of temporal expressions from texts. A special mention should be made of the algorithm for solving the problem of binding named entities into chains described in [21]. This algorithm showed better results in the process of testing than such systems as DBpedia Spotlight, Dexter, Babelfly. However, the algorithm's speed in some tasks turned out to be much lower.

In the majority of modern approaches to extraction of the information from texts in natural language at the first stage of text processing the syntactic analysis (parsing) of the text occurs. Parsing, as a rule, is performed with the help of a special program called parser. In more detail with possibilities of various parsers and software products for the analysis of texts in the natural language it is possible to familiarize in work [22]. Among others, the authors consider possibilities of Tomita Parser, [23] which we use in our work.

3. Problem statement

This article describes the first part of a large complex of research works aimed at creating methods, technologies, algorithms and software for visualizing spatial data extracted from texts in natural language. At this stage of the work the problem of creating a technology for extracting geodata from texts on the Arctic theme was raised. A set of texts in Russian was created as a source of information for analysis. Other languages are not supported by the current version of the technology. Support for
analysis of texts in other languages will be implemented in the next version of the technology, if necessary.

Several requirements and limitations for technology implementation were formulated. One of the most essential requirements is the necessity to use ready-made parsers or specialized program libraries for analysis, which allow us to quickly create and set up a text analyzer. Such requirement is connected with limited financial and time resources allocated for the project realization. Another requirement is related to the language and format of text representation. It means that all the texts should be presented in the natural Russian language in a format that allows their processing by the parser.

4. Materials and methods
The method of syntactic text analysis (parsing) was chosen as the main method in this paper. The ready-made software product Tomita parser by the Yandex company was chosen as the parsing tool. For today Tomita parser code has been opened and uploaded for public access in GitHub [24]. The parser is distributed under MPL 2.0 and can be used for scientific and educational purposes without any restrictions.

A corpus of Arctic texts was used as materials for analysis. A significant part of these texts was collected at the previous stage of research work aimed at creating methods, technologies, algorithms and software for visualizing spatial data extracted from natural language texts. [9] Within the framework of this work, the corpus was supplemented with various texts related to the Murmansk region of the Russian Federation. When selecting texts, preference was given to texts related to the description of transport and logistics infrastructure of the region, plans, projects and programs of socio-economic development of the Murmansk region and other similar texts.

To work with Tomita parser, keyword dictionaries (gazetteers) as well as context-free grammars (CFG) must be created. The creation of adequate gazetteers and grammars is the main task of the researcher. In general, context-free grammar is a system of the following kinds:

\[ G = \langle V_T, V_N, R, s \rangle, \]

where

- \( V_T \) is a set of terminal grammar symbols;
- \( V_N \) is a set of nonterminal grammar symbols;
- \( R \) is a set of output rules on elements of grammar;
- \( s \) is a starting axiom of a set of nonterminal grammar symbols.

It is important to note that the overall efficiency of the parser depends on the quality of the gazetteers and grammars created by the researcher.

5. Results
In order to apply Tomita parser to extract spatial data from texts, a general scheme of interrelationship of files was created, which are necessary to perform the procedure of syntactic analysis of texts (Figure 1).
The files necessary for text analysis include directly the parser file itself (tomitaparser.exe). This file implements the text analysis procedure according to the connected dictionaries and grammars. Gazetteer (dic.gzt) is a special dictionary, which contains keywords for analysis. Key words from the gazetteer are used by context-free grammars (event.cxx). The entries that make up this gazetteer define different sets of words (word combinations) with common properties. For example, one of these sets may be a set which may be called "toponyms of the Murmansk region". Common properties of words are used in text analysis. For example, when using grammar, we can use the property "is toponym of Murmansk region". Words (word combinations) included in such sets can be defined explicitly in the list format or by using special functions. To use the functional approach, it is necessary to create a grammar that describes the necessary word chain. Such chains can describe, for example, addresses of geo-objects. After creating a chain, we can use it as a marker to identify various city events, that is, specific facts described in the text.

It should be noted that in addition to the key words, special lemmas can be specified in the gazetteer. The use of lemmas facilitates the output of text analysis results and increases parsing efficiency. In particular, using lemmas, synonyms of words are accounted for, as well as various forms of word recording are brought to a single record standard. Another example of using lemmas is processing abbreviations in texts. For example, the abbreviation "MTN" can be derived as "Murmansk Transport Node".

Grammar files are presented in the format of rules written in the context-free grammar language. These rules are collected in sets. The sets of rules describe the syntactic structure of word chains that are identified during text analysis. While analyzing the sets of terminal grammar symbols, the grammar is projected onto a sentence or individual words in the text. In this case, one and the same word can be mapped to several terminal symbols. During the text analysis, a set of terminal symbols is entered at the parser input. For example, when using three terms verb, adjective and noun, the sentence "Chibis visited Kola NPP" will be parsed as a list: [noun - verb - adjective - noun]. In this case, such a chain of words derived from the recognition with such grammar is output.

The facts identified in the text are described in tabular form. Each column of this table is a field of fact. These columns are filled in by the parser during the analysis of sentences. In each individual case it is necessary to indicate the specificity of filling the facts table in grammar. After receiving the data in the table form, we can convert the file with the table into a database file or another format that is convenient for further processing.

Interpretation is the last process of text analysis. This process allows to distribute word subchains from a grammar-chain by fact fields. As a result of the interpretation procedure, the tree syntactic structure is displayed in many linearly organized facts.
The main files that were used when analyzing texts with Tomita parser are described in Table 1. The table contains the name of the file, its format and a brief comment about what this file is for.

Table 1. Table of main files for text analysis in Tomita Parser

| File name / file description                                                                 | File format         | Commentary                                                                 |
|------------------------------------------------------------------------------------------------|---------------------|---------------------------------------------------------------------------|
| Kwtypes.my.proto / contains a description of keyword types                                     | protobuf            | This file is necessary to create new key types                             |
| Facttypes.proto / contains a description of the types of facts used in the project             | protobuf            | This file is necessary to create facts while the parser is working         |
| Dic.gzt / contains all the dictionaries and grammars used in the project                        | protobuf / gazetteer| This is the root dictionary - mandatory project file                       |
| Event.cxx / contains grammar description                                                      | grammar description language | This file is necessary to use grammar. There may be several such files |
| Config.proto / contains information for the parser about paths to other files and other parser settings | protobuf            | This file is necessary to save the settings of the parser                 |

6. Conclusion and future work

Spatial data extracted from natural language texts can be used to visualise and support spatial management decisions. The use of geo-visualization tools in solving management tasks can be an effective tool for decision-making in the implementation of projects in the field of energy, construction, development of mineral deposits, and other tasks. Spatial data also allow to better understand the level of socio-economic development of a region, distribution of population and industry by territory.

As a result of consideration of the problems of extracting named entities from texts in natural language and analysis of the toolkit for the direct obtaining of geodata from the texts, Tomita parser was chosen, which formed the basis of the developed technology. Thus, the main results of this research phase are as follows:

1. The file structure for using Tomit Parser to extract spatial data from texts to support decision making on territory management was formed. The proposed structure allows analysis of Arctic texts. For the analysis it is necessary to enter the corpus of the analyzed texts to the parser input. Further analysis is carried out in automatic mode.

2. Dictionary of key words (gazetteer) on the example of the Murmansk region was created. The gazetteer allows identifying toponyms when analyzing texts in natural language. It supports correct identification of synonyms, abbreviations and informal names of geo-objects in the Murmansk region.

3. The grammar is created to identify facts from texts. Several typical structures of facts are formed, which can be used in further research on the analysis of texts in natural language.

4. Several tests of the proposed structure, gazetteer and grammar on limited sets of texts were conducted. The accuracy of geo-objects identification is currently at about 80-85%.

We consider that the results obtained at this stage are satisfactory, although additional efforts are required to increase the accuracy of geo-objects identification. To improve the accuracy of geo-objects identification in text analysis, we plan, first of all, to expand the gazetteer and complement the
grammar. We also plan to consider the possibility of analyzing texts in English. This may require the use of another parser or other natural language text analysis methods.

References

[1] Bolsunovskaya Y et al 2016 IOP Conf. Ser.: Earth Environ. Sci. 43 012099
[2] Vicenti A 2019 Development of methods and tools to support regional management in the Arctic zone of the Russian Federation based on cognitive interfaces IOP Conf. Ser.: Earth Environ. Sci. 302 012139
[3] Vicenti A, Vicenti I 2019 The method of dynamic visualization of spatial data for cognitive interfaces of information systems supporting regional management, Int. Mult. Sc. GeoConf Surveying Geology and Mining Ecology Management, SGEM 19 (2.1) 667-672
[4] Hahmann S, Burghardt D 2013 How much information is geospatially referenced? Networks and cognition, International Journal of Geographical Information Science 27 1171-1189
[5] Bakhareva N 2014 Synergetics of spatial information, Slavic Forum 1 25-32
[6] Markelov V 2013 Spatial information as a management factor, State Advisor 4 34-38
[7] Tsvetkov V, Alpatov A 2014 Management of distributed transport streams, State Advisor 3 55-60
[8] Rosenberg I 2015 Spatial management in the field of transport, Slavic Forum 2 268-274
[9] Vicenti A, Dikovitsky V and Shishaev M 2019 Automated Extraction and Visualization of Spatial Data Obtained by Analyzing Texts About Projects of Arctic Transport Logistics Development, Advances in Intelligent Systems and Computing 1046 419-433
[10] Bakhareva N 2015 Decision support in land valuation, State Advisor 1 50-56
[11] Bakhareva N 2016 Geodata in the land registry, Educational resources and technologies 3 69-79
[12] Khurana D et al 2017 Natural Language Processing: State of The Art, Current Trends and Challenges 1-2
[13] Reshamwala A, Pawar P 2013 Review on natural language processing, Engineering Science and Technology: An International Journal 3 113-115
[14] Huang H et al 2016 European Handbook of Crowdsourced Geographic Information 196-197
[15] Campelo E, Baptista C 2009 Model for Geographic Knowledge Extraction on Web Documents, Advances in Conceptual Modeling - Challenging Perspectives 9 317-326
[16] Campelo E, Baptista C 2016 Mining Influential Terms for Toponym Recognition and Resolution, Brazilian Journal of Cartography 68 1123-1132
[17] Marrero M et al 2013 Named Entity Recognition: Fallacies, Challenges and Opportunities, Computer Standards & Interfaces 35 482-489
[18] Wu S, Tsai T and Hsu W 2003 Domain event extraction and representation with domain ontology, In Proceedings of the IJCAI-03 Workshop on Information Integration on the Web 33-38
[19] Hanova A, Grigoryeva I 2009 Subject ontology as a way of forming a semantic model of knowledge of a cargo port, Vestnik of Astrakhan State Technical University 1 2-3
[20] Lando N 2015 Modern methods of automatic analysis of temporal expressions in natural language texts, Program systems: theory and applications 6 419–439
[21] Panteleev M, Samarev R 2017 An Algorithm for Binding Named Entities, Science and Education, Moscow State Technical University named after N.E. Bauman. electron. Journal 7 248-261
[22] Bolshakova E, Efremova N and Sharikov G 2015 Tools for developing systems for extracting information from Russian-language texts, New information technologies in the automated systems 533-543
[23] Official Tomita parser website Available from: https://yandex.ru/dev/tomita/ [Accessed 23 February 2020]
[24] Official Tomita parser github page Available from: https://github.com/yandex/tomita-parser/
[Accessed 23 February 2020]