V. B. Mokin, Dc. Sc. (Eng.), Prof.; M. A. Horash; Y. M. Kryzhanovskyi, Cand. Sc. (Eng.), Ass. Prof.; T. E. Vuzh, Cand. Sc. (Eng.)

INFORMATION INTELLIGENT TECHNOLOGY OF THE AUTOMATIC GEOREFERENCING OF THE ECOLOGICAL TEXT NATURAL-LANGUAGE INFORMATION

Research is devoted to the development of the information intelligent technology of the automated georeferencing of text information-language information by means of the NER (named entity recognition) technology and NLP (natural language processing) technology with the reference to the geographical objects of the vector maps. The study kit is formed by means of partition of the labeled entities-locations and entities-organizations into separate samples, which contain combined in a certain way entities, that characterize the planar objects of larger surface and, separately, those that characterize smaller planar objects, linear and point objects. Such division of the data enables to organize multistage revision of the identification results and models used, this allows to provide simultaneously the increase of the completeness, accuracy and speed of georeferencing of the set ecological text information.

Recommendations, regarding the application of this technology for Ukrainian, English and other languages as well as for the algorithm for the preparation of the input cartographic data, using GIS-package of ArcGIS programs are developed. The examples of the application of the separate elements of the suggested technology to real text data about the state of water arrays of the South Bug basin are given.

Key words: *named entity recognition (NER) technology, natural language processing (NLP) technology, NLP, data georeferencing, spatial relations, GIS, machine learning, artificial intelligence, ecological text information.*

Introduction

Every day the increasing volume of the electronic ecological information about the world around us is formed and its automatic processing, structuring, formalization to provide searching during the solution of various problems becomes more and more urgent.

As it is known ecologic electronic information comprises any information about the state of the environment components (air, water, soil, landscape and natural objects, biological variety, etc and the interaction between these components), about the activity or measures, including administrative measures, agreements in the sphere of environment protection, policy, legislation, plans and programs, influencing or may influence these components; about the analysis of expenses, results and assumptions, used in the process of decision- making, regarding the problems, dealing with the ecological issues; about the state of health and security of people, conditions of life, state of the objects of culture and buildings, to the extent, that the state of the environment components influence or may influence them [1]. The important characteristic of the ecological information is that it is analyzed only regarding the specific geographical planar objects – countries, regions, river basins, settlements, water bodies, conservancy areas, etc. Text may contain both the names of these objects and the names of the other objects, located there and the state or consequences of their impact is the main subject of the analysis (rivers, corridors of the ecosystem, water discharge points of the enterprises-users of natural resources, landfill sites, farming lands, etc.), but their names the analyst, due to certain reasons, cannot know. Thus, there appears the problem of the data georeferencing to the objects with the *a priori* unknown names, which are in certain spatial relations with the known objects. The solution of this problem would enable to fill the specialized information searching systems more rapidly, for instance [2 – 4].

In recent years the plans of all the river basins management are developed in Ukraine applying the approaches, used by the EC countries, according to these plans the ecological state should be analyzed, main tasks influencing the state are to be identified, the list of measures, aimed at realization of these tasks is to be developed. But the main task is that all the studies must concern each of tens thousands of the water bodies. For instance, in the basin of the South Bug river there

are more than thousand water bodies. In Ukraine there are no sufficient amount of monitoring stations the objects of the permanent control. One of the way outs – is the search in all possible sources of the ecological information, which have spatial relation to these water bodies i. e., may have georeferencing to them. Similar problems appear as a result of the analysis and optimization of the elements of the ecological system of various regions, where it is necessary to find and analyze quickly the information, regarding its corridors and nuclei of different levels. It is important that the similar problem occurs not only in Ukraine but also abroad. And not only at the stage of development of various projects and plans but at the stages of their verification and state control, progress in the realization; especially when the analysis of maximum actual information from mass media regarding different problems in the preset regions is important.

Technological solution of such problem contains a complex of solutions both in the sphere of the location (geographical objects) allocation in ETI and its comparison with the data, about the objects of the environment and in the sphere of ETI classification and formation of the classifiers in the form of the ontologies and semantic grid, in the sphere of maintaining the hierarchically structured spatial relations, for instance, in the form of ontologies and semantic grid in the known formats XML, JSON or RDF [5 – 7]. As a rule, it is impossible to solve such problems without expert support. That is why, all the existing approaches, as a rule, differ by the level of the automation and experts participation at different stages, first of all, at the stage of the formation of geographical names which can be in ETI. In the paper [4] the technology of the synchronous classification of texts and formation of the ontological model of the classificatory according to the principle of the mesh-grid was suggested, but however, this process also needs human participation.

The existing solutions [2 – 4, 8 – 10] allow either to solve the problem accurately and completely but it takes much time and it is rather expensive, because great number of experts are to be involved in the analysis of all the texts. Or quickly, but the experts, without taking into account the context and various possible meanings. Or quickly but not completely, taking into consideration the context but only for the prefixed, collected by the experts, certain amount geographical named entities, which describe the analyzed object.

Aim of the given research is to improve the completeness, accuracy and speed of the automated georeferencing of the ecological text natural-language information at the expense of the improvement of the corresponding intelligent information technology.

Analysis of the modern approaches to the automated georeferencing of ETI and basic ideas regarding their improvement

Conventionally the problem of the automated georeferencing is solved in the following way [5, 8 – 10]:

– on GIS map the preset entities- geographical objects , to which the georeferencing is to be made, are marked i. e., to find with which of them there exist spatial relations (in case of the water bodies, they may be not only the boundaries of the water intake areas of these bodies but also water flows, water bodies, residential areas enterprises -water resources users) and the set of the name entities is formed (more detailed technological realization of this algorithm will be presented below);

–all the word forms of each entity are formed by the preset language (for instance : Вінниця, Вінниці, Вінницю, Вінницею, etc.,);

– in ETI the search of each entity is carried out, and, in case of the coincidence, the corresponding text is marked as the text that has the georeferencing to a certain entity.

Such simple, at first view, algorithm has many complexities:

1. It is difficult to find GIS map, where all interesting objects are available, for instance, map of the country where all potential water users both point (enterprises) and diffusive (farming lands, forests) are plotted, it is important for the analysis of the factors, influencing the ecological state of the water.

2. It is difficult to find Ukrainian language free libraries, which generate all possible wordforms, especially for the names, consisting of different words, including words, borrowed from other languages.

3. It is difficult to find relevant geographical entities (in Natural Language Processing (NLP) and Name Entity Recognition (NER) such entities are usually called «locations» [9, 10]) in ETI, especially those which have in the name many words and sometimes it is not easy to understand if they all are the part of the name or they are homonyms (for instance, Silnytsia – it is a river and residential area, Kunka – it is a river and residential area).

As a rule, for the solution of the problem of the entities search one of two types of approaches is used:

– «Ruled-based» approaches – by the system of rules – operate quickly but can not find the entities, which were not known before;

– approaches on the base of the technologies of artificial intelligence and machine learning, in particular – NER as a subtype of NLP-technologies – enable to find the unknown possible variants of the entities and their connection with the known ones but the procedure may take some time and find numerous irrelevant variants, that is why, these approaches require constant correction.

In case of technologies application to a great number of various texts, as in our case, the first type of the approaches cannot be considered as the quick one, as for the formation of the complete set of entities in the whole ETI, first it must be analyzed in an expert way, and it requires much time.

That is why, for such various by the spatial referencing problems, especially for the cases, when it is important to analyze new texts in the on-line mode, the second group of approaches is used. But the main task in this case is the necessity of expert correction of NER-technology operation and considerable uncertainty of the problem set up, regarding what spatial relations and how should be taken into account.

According to the set aim, the following approach is suggested:

1. Form the basic maximally reliable set U of the locations, using geoformation systems of the region, data bases of different registers and cadasters, where the sample of data, regarding this region, can be made. For the problems of water resources management, GIS of the river basin with the hydrogeography layers, residential areas and water users, register of the enterprises-users of water resources, accounting of the 2^{hd} type may be taken. Form basic training sample U_1, from these basic locations, but unlike the existing approaches make it only from the comparatively large planar geographical objects of the environment (river basins, water intake area of the water bodies, administrative regions and districts, boundaries of the residential areas, etc.), on the territory of which spatial objects, which are marked on the maps as linear and points (rivers, water intake or discharge of the enterprises-users of water) or as planar objects of the comparatively small area (territories of the organizations, farming lands) may be located. From the second group of the entities-geographical objects (linear, points and small planar) basic test dataset U_2 should be formed. Basic datasets must contain maximally reliable information, based on the verified GIS, state cadasters, registers of the spatial data, etc.

2. Synthesize the word forms (we denote this operation by the function F_s) for the names of all the entities from the point 1 and add them to the datasets formed at that stage:

$$X_1 = F_s(U_1), \quad X_2 = F_s(U_2), \tag{1}$$

where X_1, X_2 – are sets U_1 and U_2, correspondingly, supplemented by all possible wordforms in the same language.

3. Basic training sample should be used for the construction of the M model for the identification of the entities in the ETI, which refer to the preset geographical region (ecosystem of the region, river basin, water bodies, etc).

4. In texts *T*, which will be defined as those which contain entities- locations from the training sample, by means of the model *M* new set *Y* from the entities-locations and entities-organizations, connected (by the criteria of NER-technologies) with entities-locations of this training sample should be identified:

$$\mathrm{Y} = M(T, X_1). \quad (2)$$

Compare these new entities with the basic test dataset from p. 2. Arrange the model (or select from some possible models) in order to maximize the accuracy of entities definition by F-criterion, which takes into account both the accuracy and completeness of the identification [11]:

$$\mathrm{Y} = \max_F(M(T, X_1), X_2). \quad (3)$$

After that (probably after the random expert or other type of verification), new dataset is added to the basic one:

$$U_2 = U_2 + Y \quad (4)$$

and the transition to p. 1 with further iteration of the pp. 2 – 4 is carried out, until the next random inspection shows that the result of the operation (3) does not give the value of F-error, for instance, f1_score from the library sklearn on Python, is higher than certain minimal value, for instance 0.7.

5. If in p. 1 rather large test dataset was formed, which can be divided into *n* samples U_{2i} $(i = 1, 2, \dots n)$ by the volume or geographical criteria, then pp. 3 – 4 can be applied to them on multistage basis, until reliable for the accuracy analysis data are available:

$$Y_i = \max_F(M(T, X_1), F_s(U_{2i})), \quad Y \in Y_i, \ i = \overline{1, n}. \quad (5)$$

After the completion of the georeferencing of ETI process and to practical usage of the identified bulk of texts the complete verification of the correctness of the identified set of entities-locations and entities -organizations and their connections with the entities-locations of the basic reliable datasets is desirable.

The advantage of such approach is the speed of its formation as all the operations can be automated if the stage of the random expert verification is neglected and the verification is performed one time after the completion of the technology operation. The drawback is high probability of the second type error («False Negative»), when the entity is taken into account not correctly. For instance, certain text will contain information regarding the ecological problems of all the regions of Ukraine. Including the region, being analyzed, but besides this region, other regions will be included in the text. Randomly, such algorithm may add then, and the expert – miss the case. For the minimization of such an error, first the preliminary processing of ET Ican be done, for instance, entities are searched first in the content and then – only on the pages, which correspond to identified point of the contents, secondly distance in words between the entity from the basic sample and new identified entities can be limited, for instance, by 1000 words or 10% of the volume of words in the document, to decrease the risk of such error.

Stages of the developed information intelligent technology of the automatic georeferencing of ETI

As it was mentioned above, the first stage of the improved information intelligent technology of the automated georeferencing of ETI is the formation of the basic set of *U* entities. We will consider this process on the example of the solution of the problem of georeferencing ETI to the water bodies of the given river basin. This stage consists of the following steps:

1) Vectorization of the water intake areas of the water bodies. This can be performed in the mode of the automatic vectorization, applying GIS ArcGIS Desktop. To perform this stage the following GIS-data must be available:
   – layer of the water bodies of the studied territory;
   – layer of the detailed hydrography;
   – digital array of the terrain (DAT).

2) Formation of the set of the intercrossing of the water intake basins of the water bodies with the geographical objects of the state cadasters: land, water, water registers, forestry inventory, mineral deposits inventory, etc. Such set of the intercrossing can be formed, using the facilities of the overlay analysis of the professional GIS (ArcGIS Desktop, QGIS or others). Fig. 1 shows the example of the intercrossing set of the water intake basins of the water arrays with the residential areas. It is important to note, that as a result of this intercrossing in the Table of attributes of each geographical object, entering the boundaries of the water intake basin the official code of water array is written down automatically.

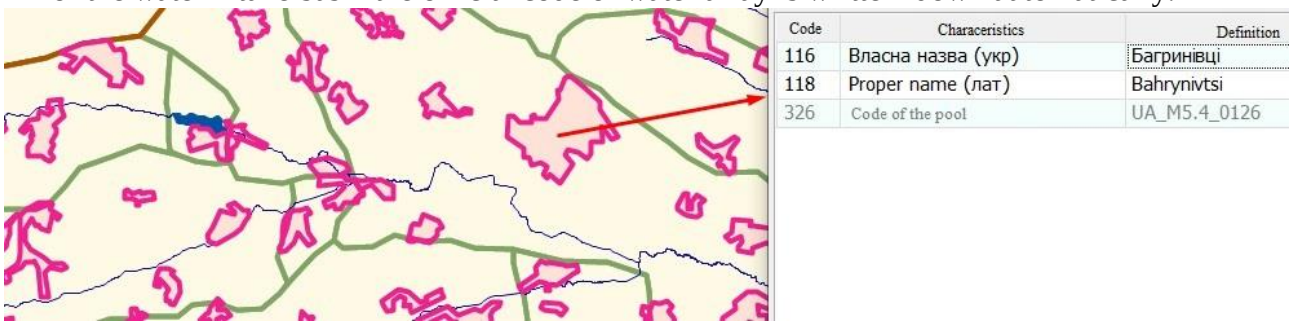

Fig. 1. Example of the set of the intercrossing of water intake basins of the water arrays with the geographical objects of the state cadasters

3) Formation of the list of the geographical names for the set of the intercrossing objects.

To perform this step it is necessary to develop the module for the work with the text data, the module will provide:

– settings, where it is indicated from what fields of the Table of attributes and from what layers the text data are to be collected;
– collection of the text data according to the settings;
– post processing of the collected data, their formatting, etc.

4) Automated verification of the geographical names on the subjects of their actuality. This step includes the revision of the formed sample of the geographical names with the official lists of names, being object of changes as a result of the implementation of the decommunisation legislation in Ukraine or various renaming for reasons other than those above.
5) Formation of the resulting data set, where for each entity the important information for the next stage will be saved:
– about the name (it is desirable , in different languages, available in cadasters – often the names are both in Ukrainian and English languages);
– about the type of the objects presentation on the map (planar, linear, point),
– if the object is planar, then – the information about its area is needed.

If GIS contains the information regarding the spatial relations between the objects, then it is important to save these relations, for instance, that certain city is located on certain river or certain district is a part of a certain region. The information, regarding spatial relations between the objects is expedient to save in JSON-format, which is the sets of the pairs of «key: meaning» type and may contain the references to other objects. The advantage of the given format is the flexibility of the data structure at the sufficient level of the formalization, that enables to read out the data by means of the program for further analysis, for instance, using Python language.

For the formation of the wordforms X_1, X_2 library pymorphy2 on Python can be used, it contains the algorithms for English, German, Italian,French and other languages [12]. For Ukrainian language ВЕСУМ dictionary, developed by the team БрУК [13] can be used.

For the construction of the model M spaCy library in Python can be used, it contains the algorithms for English, Chinese, French, Italian, Polish, Spanish languages [14]. For Ukrainian language stanza-lang-uk csan be used — it is the technology for the work with Ukrainian texts [15].

Block-diagram of the algorithm of the suggested technology is shown in Fig. 2.

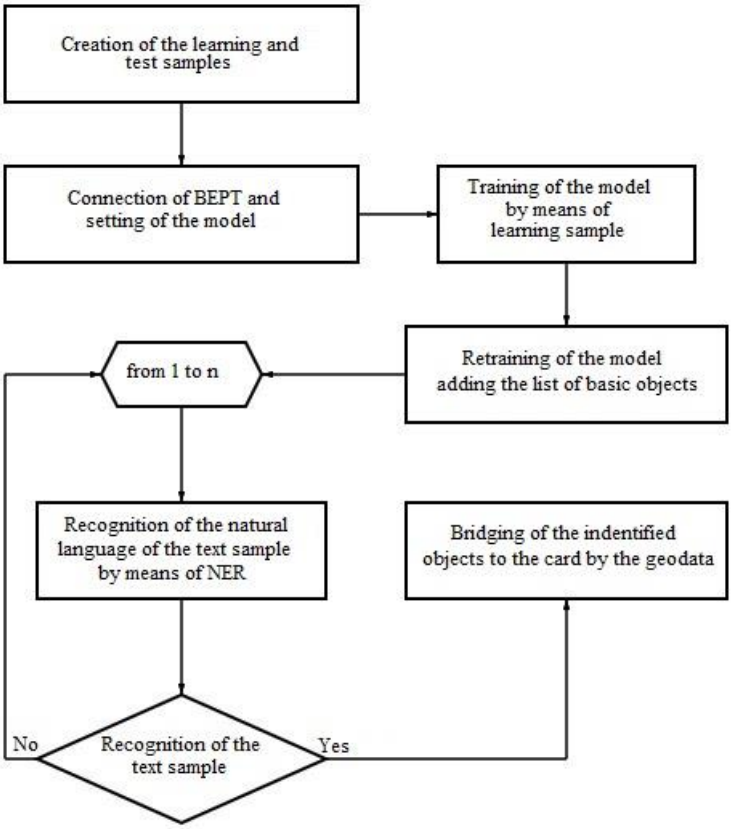

Fig. 2. Block-diagram of the algorithm of the suggested information intelligent technology of the automatic georeferencing of ETI

Example of the application of the developed information intelligent technology of the automated georeferencing of ETI

Below we will consider the example of the application of the suggested technology on the example of the water arrays of the south Bug River. Water-resource region (WRR) «The South Bug river from the mouth of the river Ikva to residential area Selysche», which in the Water Register of Ukraine has the official number UA_M5.4.0.02. Fig. 3 shows the example of the vectorized water intake basins of the arrays of this WRR.

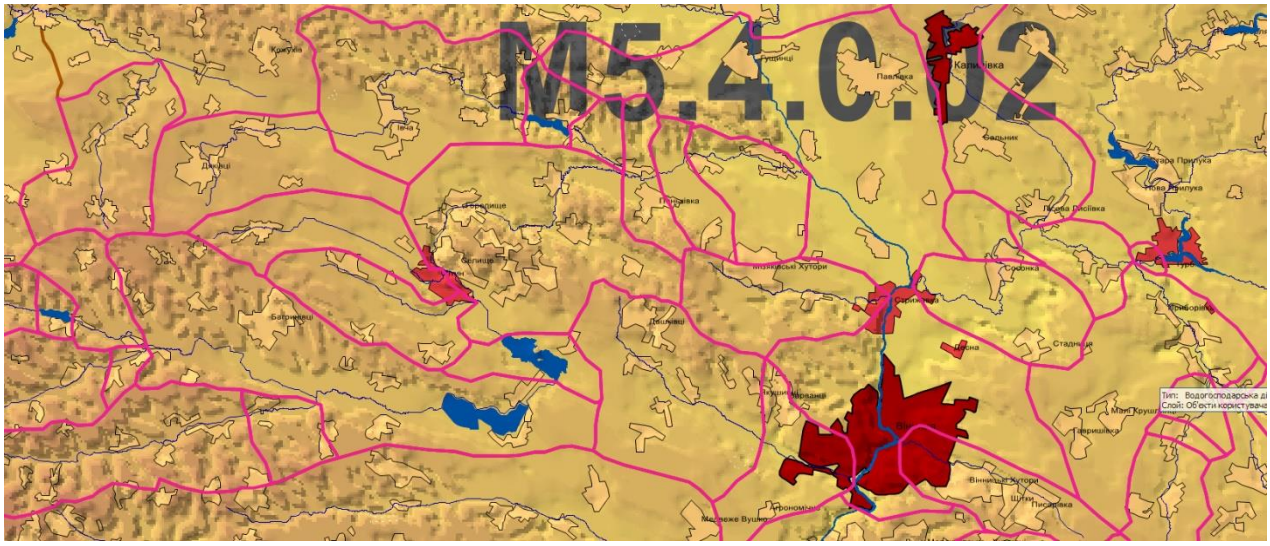

Fig. 3. Example of the vectorized water intake basins of the water arrays of the water resources region

UA_M5.4.0.02 «South Bug River from the mouth of the river Ikva to residential area Selysche»

Using the suggested approaches for the collection and postprocessing of the GIG data the formation of the data set (Fig. 4) for each water array is carried out.

Name of WB (UA)	Name of WB (EN)	Code of WB	Type of WB (UA)	Type of WB (EN)	UA-name entities	EN-name entities
1	2	3	4	5	6	7
Згар	Zgar	UA_M5.4_0123	річка	river	с. Городище, с. Старий Майдан, с. Осикове, с. Козачки, с. Варенка, с. Грушківці	Horodysche, Staryi Maidan, Osykove, Kozachky, Varenka, Hrushkivtsi
Згар	Zgar	UA_M5.4_0125	річка	river	с. Голенищеве, с. Буцні, с. Сахни, с. Білецьке	Holenyscheve, Butsni, Sakhny, Bilets'ke
Згар	Zgar	UA_M5.4_0126	річка	river	с. Лисогірка, с. Ріжок, с. Зоринці, с. Микулинці, с. Залужне, с. Соколівка, с. Кільянівка, с. Голенищеве, с. Українка, с. Багринівці, с. Лозни, с. Гончарівка, с. Сахни, с. Майдан-Сахнівський, с. Яблунівка	Lysohirka, Rizhok, Zoryntsi, Mykulyntsi, Zaluzhne, Sokolivka, Kil'ianivka, Holenyscheve, Ukrainka, Bahrynivtsi ,Lozny, Honcharivka, Sakhny, Maidan-Sakhnivs'kyi, Yablunivka
Сандракське водосховище	Sandrakske reservoir	UA_M5.4_0011	водосховище	reservoir	м. Хмільник, с. Стара Гута, с. Широка Гребля, с. Голодьки, с. Вугли, с. Вербівка, с. Лелітка, с. Крутнів, с. Березна, с. Соколова	Khmil'nyk, Stara Huta, Shyroka Hreblia, Holod'ky, Vuhly, Verbivka, Lelitka, Krutniv, Berezna, Sokolova
Сабарівське водосховище	Sabarivske reservoir	UA_M5.4_0013	водосховище	reservoir	м. Вінниця, м. Калинівка, смт Десна, смт Стрижавка, с. Зарванці, с. Стадниця, с. Тютюнники, с. Лаврівка, с. Дорожнє, с. Медвідка, с. Мізяків, с. Мізяківська Слобідка, с. Павлівка, с. Майдан-Бобрик, с. Гущинці, с. Кам'яногірка, с. Калинівка Друга, с. Іванів	Vinnytsia, Kalynivka, Desna, Stryzhavka, Zarvantsi, Stadnytsia, Tiutiunnyky, Lavrivka, Dorozhnie, Medvidka, Miziakiv, Miziakivs'ka Slobidka, Pavlivka, Maidan-Bobryk, Huschyntsi, Kamianohirka, Kalynivka Druha, Ivaniv
Сутиське водосховище	Sutiske reservoir	UA_M5.4_0014	водосховище	reservoir	м. Вінниця, с. Іванівка, с. Яришівка, с. Селище, с. Студениця, с. Урожайне, с. Лани, с. Бохоники, с. Парпурівці, с. Лука-Мелешківська, с. Хижинці, с. Прибузьке, с. Тютьки, с. Майдан-Чапельський	Vinnytsia, Ivanivka, Yaryshivka, Selysche, Studenytsia, Urozhaine, Lany, Bokhonyky, Parpurivtsi, Luka-Meleshkivs'ka, Khyzhyntsi, Prybuz'ke, Tiut'ky, Maidan-Chapel's'kyi

Fig. 4. Fragment of the resulting data set

For testing the technology water arrays ВГД UA_M5.4 were selected, they contain cities Vinnytsia (UA_M5.4_0013), Kalynivka (UA_M5.4_0181) and Khmilnyk (UA_M5.4_0011), correspondingly. For these cities basic training set U_1 of the data from planar objects of larger size (WRR UA_M5.4.0.02, Vinnytsia Region, districts of Vinnytsia Region) and test set U_2 from planar objects of the smaller size (boundaries of the residential area) linear (rivers) and point objects (sites of the water discharge, water intakes of the enterprises-users of water resources, monitoring stations of water quality or amount) were formed. Fig. 5 presents the example of the combination of the water intakes basins of the water arrays with other point objects on the map.

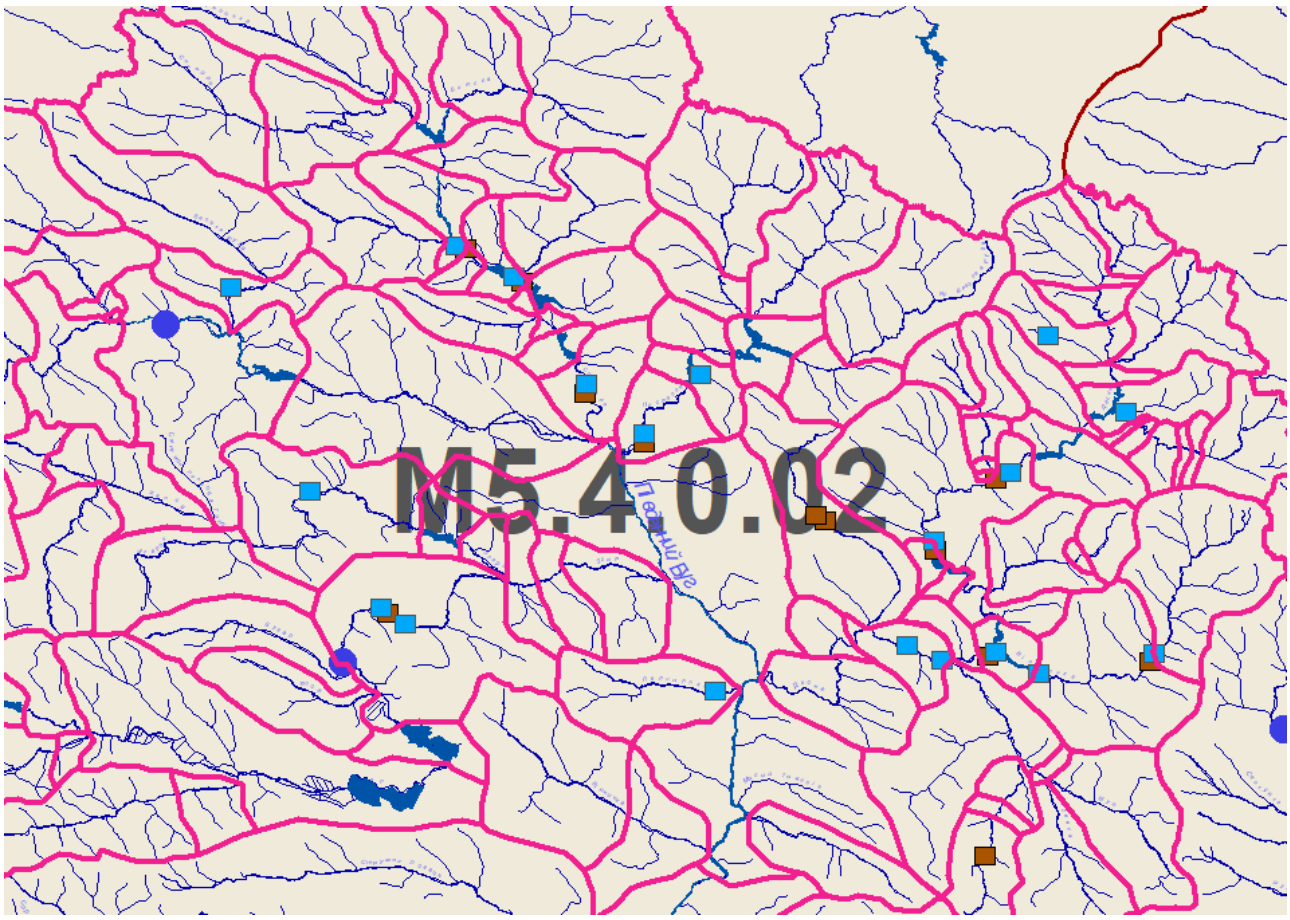

Fig. 5. Water collecting areas of the water arrays are combined with the stations of the hydrologic control (blue circles), water intakes (blue rectangulars) and water discharge (brown rectangulars)

Further, for these sets with the relation (1) the sets with all word forms X_1, X_2 are formed. Test set X_2 is divided into 3 – separately for each water array.

For testing the technology one of the co-authors of the paper Mokin V. B.> with the participation of the other co-author Gorash M. A. and students of Vinnytsia National Technical University Pasichnik D. V. and Radetskyi O. V. created the public dataset of the sentences about the state of the water resources of the South Bug river basin, the dataset is installed on the platform Kaggle [16]. This dataset contains sentences from the monograph, published by the co-authors Mokin V. B. and Kryzhanivsyi E. M., the monograph has published English [17] and Ukrainian [18] versions. For the experiments English part of the dataset was used. The application of the suggested technology allowed to form the list of the entities locations, error f1_score for which on average was 0.78, the operation of separate solutions, suggested in the paper, was tested. Further full scale experiment for the determination of the completeness, accuracy and speed of the georeferencing of the set Ukrainian and English ecological text natural-language information for a number of water array area and for the comparison of these indices with the indices regarding the realization of such georeferencing in the traditional for such problems way is planned to perform.

Conclusions

The given paper describes the development of the intelligent information technology of the automated georeferencing of the ecological text natural-language information. New approach to the information of the training set of data by means of the division of the entities-locations and entities-organizations into separate samples, which contain entities, combined in a certain manner, which characterize the planar objects of the larger area and, separately, those which characterize smaller

planar objects, linear and point objects. Such division of data enables to organize multistage verification of the identification results and models used, this allows to provide simultaneously the increase of the completeness, accuracy and speed of the georeferencing of the set ecological text information.

Recommendations, regarding the application of this technology for Ukrainian, English and other languages as well as the algorithm of the preparation of the input cartographic data, using GIS-package of ArcGIS programmes are developed. The examples of the application of the separate elements of the suggested technology to real text data, concerning the state of the water arrays of the South Bug river are given:

1. By means of GIS, for instance ArcGIS, applying the facilities of the overlay analysis it is possible to form the set of the water arrays of the water collecting basins of the water arrays with geographical objects of the state cadasters: land, water, forestry, cadaster, natural resources cadaster, etc.

2. For testing and creation of ETI such water arrays of the South Bug river were used UA_M5.4_0013, UA_M5.4_0181, UA_M5.4_0011. The suggested technology was realized on Python using technologies Named Entity Recognition, BERT, also libraries spaCy and TensorFlow were used.

3. For the given water arrays the data with the information, concerning the problems and rivers description, located there, were selected. The monograph, containing the description of the South Bug river was used. This monograph was published in two languages (English and Ukrainian, within the frame of Swedish-Ukrainian project, sponsored by SIDA) that enables to teach and train information technology both in English and Ukrainian. Bilingual dataset was formed from this monogram, but for testing only English text was used.

4. Successful testing of the separate elements of the suggested technology was carried out. The error was 0.78. Further the development of this technology is planned to work with Ukrainian language.

REFERENCES

1. Convention on the access to the information, participation of the civil society in the decision-making process and access to the justice on the problems, concerning the environment [Electronic resource] / Access mode : https://zakon.rada.gov.ua/laws/show/994_015#Text. (Ukr).

2. Kuo Chiao-Ling Kuo Interoperable cross-domain semantic and geospatial framework for automatic change detection» / Chiao-Ling Kuo, Jung-Hong Hong // Journal Computers & Geosciences. – 2016. – Issue C, Volume 86. – P. 109 – 119. – DOI 10.1016/j.cageo.2015.10.011.

3. WISE - Water Information System for Europe is the European information gateway to water issues [Electronic resource] / Access mode : https://water.europa.eu/.

4. Construction of the scaled information-searching system for the river basin management on the base of registers and ontological models / V. B. Mokin, I. I. Ovcharenko, A. M. Luchko [et al.] // Mathematical modeling in economics. – Kyiv, 2019 – № 2 (15). – P. 45 – 56. (Ukr).

5. Concept of the intelligent NLP technology for georeferencing and classification of the open text information about water arrays [Electronic resource] / V. B. Mokin, M. A. Gurash, D. Pasichniuk, O. Radetskyi // Materials of XV International conference “Control and management in complex systems (KYCC-2020)”, Vinnytsia, October 8-10, 2020. – Vinnytsia, 2020. – Access mode : http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/30607/KUSS%202020%20MHPR%20-%20NLP.pdf?sequence=1. (Ukr).

6. ‘Chapter One - Fast execution of RDF queries using Apache Hadoop [Electronic resource] / Somnath Mazumdara, Alberto Scionti // Advances in Computers. – 2020. – Volume 119. – P. 1 – 33. – Access mode : https://www.sciencedirect.com/science/article/pii/S0065245820300401.

7. Stryzhak O. E. Means of ontological integration and support of the distributed spatial and semantic information resources / O. E. Stryzhak // Ecological safety and environment management. – 2013. – № 12. – P. 166 – 177. (Ukr).

8. A deeply annotated testbed for geographical text analysis : The Corpus of Lake District Writing [Electronic resource] / Paul Rayson, Alex Reinhold, James Butler [et al.] // GeoHumanities'17 : Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities. – November 2017. – P. 9 – 15. – Access mode : https://doi.org/10.1145/3149858.3149865.

9. A 2021 Guide to Named Entity Recognition : Manual on recognition of the entity-organization of 2021. Survey

of the recognition of named entities (NER) [Electronic resource] / Access mode : https://nanonets.com/blog/named-entity-recognition-2020-guide/.

10. Semi-Supervised Disentangled Framework for Transferable Named Entity Recognition [Electronic resource] / Zhifeng Hao, Di Lv, Zijian Li [et al.] // Computation and Language (cs.CL); Machine Learning (cs.LG). – 22 Dec. 2020. – DOI:10.1016/j.neunet.2020.11.017. – Access mode : https://paperswithcode.com/paper/semi-supervised-disentangled-framework-for.

11. Scikit-learn Machine Learning in Python. Metrics and scoring: quantifying the quality of predictions [Electronic resource] / Access mode : https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics.

12. Morphological analyzer pymorphy2 [Electronic resource] / Access mode : https://pymorphy2.readthedocs.io/en/stable/. (Rus).

13. Dictionary BECYM and other connected means of NLP for Ukrainian language. [Electronic resource] / Access mode: https://r2u.org.ua/articles/vesum. (Ukr).

14. Electronic Text-Book Models & Languages [Electronic resource] / Access mode : https://spacy.io/usage/models.

15. Ukrainian NER data set conversion to be used by Stanza (Stanford NLP Library) [Electronic resource] / Access mode : https://github.com/gawy/stanza-lang-uk.

16. Kaggle Dataset «NLP : Reports & News Classification. ENG & UKR Automatic Environmental Reports & News Classification» [Electronic resource] / V. Mokin, D. Pasichniuk, O. Radetskyi, M. Horash. – 2020. – Access mode : https://www.kaggle.com/vbmokin/nlp-reports-news-classification.

17. Pivdenny Bug River Basin Management Plan: River Basin Analysis and Measures (Summary) / [S. Afanasiev, A. Peters, O. Iarochevitch, V. Mokin et al.]. – K. : Interservice publishing house, 2014. – 188 p.

18. Plan of the river basin of the South Bug management: analysis of the state and urgent measures / [S. Afanasiev, A. Peters, O. Iarochevitch, V. Mokin et al.]. – K. : LLC “SIP Interservice”, 2014. – P. 188. (Ukr).

Editorial office received the paper 22.12.2020.

The paper was reviewed 26.12.2020.

Mokin Vitalii – Dc. Sc (Eng.), Prof., Head of the Department of System Analysis and Information Technologies.

Horash Mykola – Graduate student with the Department of System Analysis and Information Technologies.

Kryzhanovskyi Yevheniy – Cand. Sc. (Eng.), Ass. Prof. with the Department of System Analysis and Information Technologies.

Vinnytsia National Technical University.

Vuzh Tetiana – Cand. Sc. (Eng.), Ass. Prof. with the Department of Biological Physics, Medical Equipment and Computer Science.

Vinnytsia M. I. Pyrohov National Medical University.