УДК 004.8+004.93

В. Б. Мокін, д. т. н., проф.; М. А. Гораш; Є. М. Крижановський, к. т. н., доц.; Вуж Т. Є., к. т. н., доц.

ІНФОРМАЦІЙНА ІНТЕЛЕКТУАЛЬНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОЇ ГЕОПРИВ'ЯЗКИ ЕКОЛОГІЧНОЇ ТЕКСТОВОЇ ПРИРОДНО-МОВНОЇ ІНФОРМАЦІЇ

Статтю присвячено розробленню інформаційної інтелектуальної технології автоматизованої геоприв'язки екологічної текстової природно-мовної інформації за допомогою технології розпізнавання іменованих сутностей NER та технологій опрацювання природної мови NLP з прив'язкою до географічних об'єктів векторних карт. Навчальний набір формується шляхом розбиття розмічених сутностей-локацій та сутностей-організацій на окремі вибірки, які містять у певних спосіб скомбіновані сутності, що характеризують площинні об'єкти більшої площі, та, окремо, ті, що характеризують менші площинні об'єкти, лінійні та точкові об'єкти. Такий розподіл даних дозволяє організувати багатоетапне уточнення результатів ідентифікації та моделей, які використовуються і це дозволяє забезпечити одночасно підвищення повноти, точності та швидкості геоприв'язки заданої екологічної текстової інформації.

Розроблено рекомендації щодо застосування цієї технології для української, англійської та інших мов, а також щодо алгоритму підготовлення вхідних картографічних даних з використанням ГІС-пакету програм ArcGIS. Наведено приклади застосування окремих елементів запропонованої технології до реальних текстових даних про стан масивів вод басейну р. Південний Буг.

Ключові слова: *технологія розпізнавання іменованих сутностей, технологія опрацювання природної мови, NLP, геоприв'язка даних, просторові відношення, ГІС, машинне навчання, штучний інтелект, екологічна текстова інформація.*

Вступ

З кожним днем все більше формується електронної екологічної інформації про навколишній світ і все більш актуальним є її автоматичне опрацювання, структурування, формалізація для забезпечення пошуку у ній під час розв'язання різних задач.

До екологічної електронної інформації, як відомо, відноситься будь-яка інформація про стан складових навколишнього середовища (повітря, вода, ґрунт, земля, ландшафт і природні об'єкти, біологічне різноманіття тощо та взаємодію між цими складовими), про діяльність або заходи, включаючи адміністративні заходи, угоди в галузі навколишнього середовища, політику, законодавство, плани і програми, що впливають або можуть впливати на ці складові; про аналіз затрат, результати та припущення, використані в процесі прийняття рішень з питань, що стосуються навколишнього середовища; про стан здоров'я та безпеки людей, умови життя людей, стан об'єктів культури і споруд тією мірою, якою на них впливає або може вплинути стан складових навколишнього середовища або через ці складові, та ін. [1]. Важливою особливістю екологічної інформації є те, що вона аналізується лише щодо конкретних географічних площинних об'єктів – країн, областей, басейнів річок, населених пунктів, водойм, заповідних об'єктів тощо. У тексті можуть міститись як назви цих об'єктів, так і назви інших об'єктів, які там розташовані і стан чи наслідки впливу яких саме й є основним предметом аналізу (річки, коридори екомережі, місця скидання вод підприємствами-природокористувачами, місця видалення відходів, сільгоспугіддя тощо), але їх назви можуть, з певних причин, бути невідомими аналітику. Отже, виникає задача геоприв'язки інформації до об'єктів з наперед невідомими назвами, які лише знаходяться у певних просторових відношеннях з відомими об'єктами. Розв'язання такої задачі дозволило б швидше наповнювати спеціалізовані інформаційно-пошукові системи [2 – 4].

Протягом останніх років в Україні активно розробляються плани управління усіма річковими басейнами за підходами країн ЄС, в яких слід проаналізувати екологічний стан,

виявити основні задачі, які на нього впливають та розробити план заходів щодо їх вирішення. Але головна задача в тому, що усі дослідження мають стосуватись кожного з десятків тисяч масиву вод. Наприклад, у басейні р. Південний Буг їх – більше тисячі. В Україні відсутня така кількість постів та об'єктів постійного контролю. Один із виходів – пошук по усіх можливих джерелах екологічної інформації, які мають просторове відношення до цих масивів вод, тобто, можуть мати до них геоприв'язку. Аналогічні задачі виникають у разі аналізу та оптимізації елементів екомережі різних регіонів, де треба швидко знайти та проаналізувати інформацію про усі її коридори та ядра різного рівня. Важливо, що подібна задача має місце не тільки в Україні, а й за кордоном. І не тільки на етапі розроблення різних проєктів і планів, а й на етапах їх верифікації та контролю стану і ходу реалізації, особливо, коли є важливим аналіз максимально актуальної інформації із ЗМІ про різні задачі у заданих регіонах.

Технологічне розв'язання такої задачі містить комплекс рішень і у сфері виділення локацій (географічних об'єктів) в ЕТІ та її співставленні з даними про об'єкти навколишнього середовища, і у сфері класифікації ЕТІ та формуванні класифікаторів у вигляді онтологій та семантичної мережі, і у сфері збереження ієрархічно структурованих просторових відношень, наприклад у вигляді онтологій та семантичної мережі у відомих форматах XML, JSON чи RDF [5 – 7]. Звичайно, без експертного опрацювання розв'язати такі задачі неможливо. Тому усі сучасні підходи, як правило, відрізняються рівнем автоматизації та участі експертів на різних етапах, перед усім на етапі формування географічних назв, які можуть бути в ЕТІ. У статті [4] була запропонована технологія синхронної класифікації текстів та формування онтологічної моделі класифікатора за принципом mesh-мереж, але все одно, цей процес теж потребує участі людини.

Сучасні рішення [2 – 4, 8 – 10] дозволяють або розв'язати задачу точно та повно, але довго і дорого, з використанням великої кількості експертів, що проаналізують усі тексти. Або – швидко, але не точно, просто шукаючи в тексті збіги з наданими експертами назвами, без урахування їх контексту та різних можливих значень. Або – швидко, але не повно, враховуючи контекст, але лише для наперед фіксованої зібраної експертами певної кількості географічних назв-сутностей, що описують об'єкт, який аналізується.

Метою цієї статті є підвищення повноти, точності та швидкості автоматизованої геоприв'язки екологічної текстової природно-мовної інформації за рахунок удосконалення відповідної інтелектуальної інформаційної технології.

Аналіз сучасних підходів до автоматизованої геоприв'язки ЕТІ та основні ідеї щодо їх удосконалення

Задача автоматизованої геоприв'язки традиційно розв'язується таким чином [5, 8 – 10]:

– на карті ГІС розмічаються задані сутності-географічні об'єкти, до яких слід здійснити геоприв'язку, тобто знайти з якими з них є просторові відношення (у разі масивів вод, ними можуть бути не тільки межі водозбірної площі цих масивів, а й – водотоки, водойми, населені пункти у них, підприємства-водокористувачі тощо) та формується множина назв цих сутностей (більш детально технологічна реалізація цього алгоритму буде наведена нижче);

– формуються усі словоформи кожної сутності заданою мовою (наприклад: Вінниця, Вінниці, Вінницю, Вінницею тощо);

– в ЕТІ здійснюється пошук кожної сутності і, у разі збігу, відповідний текст маркується як такий, що має геоприв'язку до відповідної сутності.

Такий простий, як на перший погляд, алгоритм має багато складнощів:

1. Важко знайти карту ГІС, на якій будуть усі цікаві об'єкти, наприклад карту країни, на якій нанесені усі потенційні водокористувачі, як точкові (підприємства), так і

дифузійні (сільськогосподарські угіддя, ліси тощо), що є важливим для аналізу факторів, що впливають на екологічний стан вод.

2. Важко знайти україномовні безкоштовні бібліотеки, які генерують усі можливі словоформи, особливо для назв, що складаються із різних слів, у т. ч. запозичених з інших мов.

3. Важко знайти саме релевантні географічні сутності (в технологіях опрацювання природної мови (з англ. – «Natural Language Processing» (NLP)) та розпізнавання імен сутностей (з англ. – «Name Entity Recognition» (NER) такі сутності більш прийнято називати «локації» [9, 10]) в ETI, особливо такі, що мають у назві багато слів і не завжди зрозуміло чи усі вони є частиною назви, або є омонімами (наприклад, Сільниця – це і річка, і населений пункт, і Кунка – це і річка, і населений пункт).

Як правило, для розв'язання задачі пошуку сутностей використовується один із двох типів підходів:

– «Ruled-based» підходи – за системою правил – працюють швидко, але не можуть знайти сутності, які наперед не були відомі;

– підходи на основі застосування технологій штучного інтелекту та машинного навчання, зокрема – NER як підвид NLP-технологій – дозволяють знаходити й наперед невідомі можливі варіанти сутностей та їх зв'язки із відомими, але можуть займати певний час і знаходити чимало нерелевантних варіантів, тому потребують постійного коригування.

У разі застосування технологій до великої кількості різноманітних текстів, як у нашому випадку, перший тип підходів лише оманливо є швидшим, оскільки для формування дійсно повної множини сутностей в усій заданій ETI, треба її спочатку всю проаналізувати експертним шляхом, а це – дуже багато часу.

Тому для таких різноманітних за просторовою прив'язкою задач, особливо для випадків, коли важливо аналізувати нові тексти в режимі он-лайн, використовується тільки друга група підходів. Але головною задачею, при цьому, є необхідність експертного корегування роботи NER-технології та суттєва невизначеність постановки задачі щодо того які просторові відношення і як саме слід враховувати.

Відповідно до поставленої мети, пропонується використовувати такий підхід:

1. Сформувати базову максимально достовірну множину U локацій з використанням геоінформаційних систем регіону, баз даних різних реєстрів та кадастрів, в яких можна зробити вибірку даних, яка точно стосується заданого регіону. Для задач управління водними ресурсами можна взяти, наприклад, ГІС басейну річки з шарами гідрографії, населених пунктів та водокористувачів, реєстр підприємств-водокористувачів, звітність 2-тп (водгосп) та ін. Із цих базових локацій сформувати базову тренувальну вибірку U_1, але, на відміну від наявних підходів, зробити її лише з порівняно великих площинних географічних об'єктів навколишнього середовища (басейни річок, водозбірні площі масивів вод, адміністративні області та райони, межі населених пунктів тощо), на території яких можуть розташовуватись просторові об'єкти, які на картах позначаються як лінійні і точкові (річки, місця водозабору або скидання підприємств-водокористувачів тощо) або як площинні об'єкти порівняно малої площі (території організацій, сільськогосподарські угіддя тощо). З другої групи сутностей-географічних об'єктів (лінійних, точкових і малих площинних) сформувати базовий тестовий датасет U_2. Базові датасети мають містити максимально достовірну інформацію на основі верифікованих ГІС, державних кадастрів, реєстрів просторових даних тощо.

2. Синтезувати словоформи (позначимо цю операцію функцією F_s) для назв усіх сутностей із п. 1 і додати їх в датасети, сформовані на тому етапі:

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

де X_1, X_2 – множини U_1 та U_2, відповідно, доповнені усіма можливими словоформами тією ж мовою.

3. Базову тренувальну вибірку використати для побудови моделі M для ідентифікації сутностей в ЕТІ, які відносяться саме до заданого географічного регіону (екосистеми області, басейну річки, масиву вод тощо).

4. У текстах T, які у п. 3 будуть визначені як такі, що містять сутності-локації з тренувальної вибірки, за допомогою моделі M ідентифікувати нову множину Y із сутностей-локацій і сутностей-організацій, пов'язаних (за критеріями NER-технології) з сутностями-локаціями цієї тренувальної вибірки:

$$\mathbf{Y} = M(T, X_1). \tag{2}$$

Порівняти ці нові сутності з базовим тестовим датасетом із п. 2. Налаштувати модель (або вибрати з декількох можливих моделей) таким чином, щоб максимізувати точність визначення сутностей за нею за F-критерієм, який враховує і точність, і повноту такої ідентифікації [11]:

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

Після чого (можливо після вибіркової експертної чи іншої перевірки), новий датасет додається до базового:

$$U_2 = U_2 + Y \tag{4}$$

і здійснюється перехід до п. 1 з наступним повторенням пп. 2 – 4, поки чергова вибіркова перевірка не покаже, що результат операції (3) не дає значення F-похибки, наприклад f1_score із бібліотеки sklearn на Python, вище певного мінімального значення, наприклад 0,7.

5. Якщо у п. 1 було сформовано достатньо великий тестовий датасет, який можна розбити на n вибірок U_{2i} $(i = 1, 2, \ldots n)$ за обсягом або за географічними критеріями, тоді пп. 3 – 4 можна застосовувати до них багатоетапно, поки будуть доступні достовірні для аналізу точності дані:

$$Y_1 i = \max \text{–F } [(M([T, X]_1 1),] \; F_1 s\,(U_1 2i)), \qquad Y \in Y_1 i, \qquad i = (1, n)^{-}. \tag{5}$$

Після завершення процесу геоприв'язки ЕТІ і до практичного використання ідентифікованого корпусу текстів дуже бажаною є повна перевірка коректності ідентифікованої множини сутностей-локацій та сутностей-організацій та їх зв'язків із сутностями-локаціями базових достовірних датасетів.

Перевагами такого підходу є швидкість його формування, оскільки можна автоматизувати усі операції, якщо проігнорувати етап вибіркової експертної перевірки і робити її один раз після завершення роботи технології. Недоліком є висока ймовірність помилки другого роду («False Negative»), коли сутність врахована неправильно. Наприклад, якийсь текст буде містити інформацію про екологічні проблеми усіх областей України, у т. ч. тієї області, яка аналізується, але, окрім неї, там будуть й інші області. Випадково такий алгоритм може їх додати теж, а експерт потім – пропустити такий випадок. Для мінімізації такої помилки можна, по-перше, робити попереднє опрацювання ЕТІ, наприклад, шукати сутності спочатку у змісті, а потім – лише на сторінках, які відповідають ідентифікованому пункту змісту, а по-друге, можна обмежити відстань у словах між сутністю із базової вибірки і новими ідентифікованими сутностями, наприклад у 1000 слів або 10% обсягу слів у документі, що зменшить ризик такої помилки.

Етапи розробленої інформаційної інтелектуальної технології автоматизованої геоприв'язки ЕТІ

Як було зазначено вище, першим етапом удосконаленої інформаційної інтелектуальної технології автоматизованої геоприв'язки ЕТІ, яка пропонується, є формування базової

множини сутностей *U*. Розглянемо цей процес на прикладі розв'язання задачі геоприв'язки ЕТІ до масивів вод заданого басейну річки. Цей етап складається з наступних кроків:

1) Векторизація водозбірних площ масивів вод. Можна це здійснювати в режимі автоматичної векторизації, наприклад з використанням ГІС ArcGIS Desktop. Для належного виконання цього етапу необхідно мати наступні ГІС-дані:
   – шар масивів вод досліджуваної території;
   – шар детальної гідрографії;
   – цифрову матрицю рельєфу (ЦМР).
2) Формування множини перетину водозбірних басейнів масивів вод з географічними об'єктами державних кадастрів: земельний, водний, кадастр лісового господарства, кадастр родовищ корисних копалин та інші. Таку множину перетину можна сформувати, використовуючи засоби оверлейного аналізу професійної ГІС (тієї ж ArcGIS Desktop, QGIS чи ін.). На рис. 1 наведено приклад множини перетину водозбірних басейнів масивів вод з населеними пунктами. Важливо, що внаслідок цього перетинання автоматично у таблиці атрибутів кожного географічного об'єкта, що потрапляє в межі водозбірного басейну, одразу записується офіційний код масиву вод.

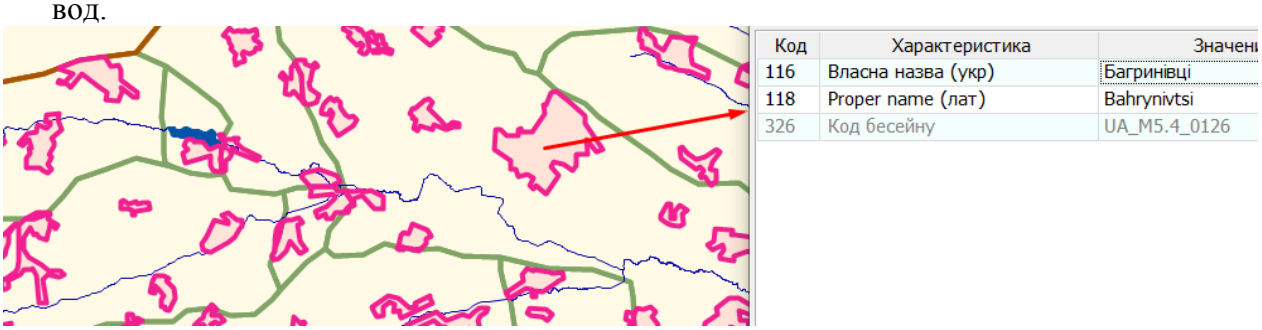

Рис. 1. Приклад множини перетину водозбірних басейнів масивів вод з географічними об'єктами державних кадастрів

3) Формування списку географічних назв для множини об'єктів перетину.

Для виконання цього кроку необхідно розробити модуль для роботи з текстовими даними, який забезпечуватиме:

– налаштування, в яких вказується, з яких саме полів таблиць атрибутів, і з яких саме шарів слід здійснювати збирання текстових даних;
– збирання текстових даних, відповідно до налаштувань;
– постопрацювання зібраних даних, їх форматування тощо.

4) Автоматизована перевірка географічних назв на актуальність. Цей крок включає звірку сформованої вибірки географічних назв з офіційними переліками назв, які зазнали змін внаслідок імплементації законодавства про декомунізацію в Україні чи різні перейменування з інших причин.
5) Формування результуючого набору даних, де про кожну сутність буде збережена така важлива для наступних етапів інформація:
   – про назву (бажано різними мовами, які є в кадастрах – часто назви є і українською, і англійською мовами);
   – про тип подання об'єкта на карті (площинний, лінійний, точковий),
   – якщо об'єкт – площинний, то – й про його площу;

Якщо ГІС містить інформацію про просторові відношення між об'єктами, тоді важливо їх зберігати теж, наприклад, про те, що певне місто розташоване на певній річці чи певний район входить до складу певної області. Інформацію про просторові відношення між об'єктами доречно зберігати у JSON-форматі, який представляє собою набори пар типу «ключ: значення», та може містити в тому числі посилання на інші об'єкти. Також перевагою

цього формату є гнучкість структури даних при достатньому рівні формалізації, що дозволяє досить легко програмно зчитувати дані для подальшого аналізу, наприклад з використанням з використанням мови Python.

Для формування словоформ X_1, X_2 можна використати бібліотеку pymorphy2 на Python, яка містить готові алгоритми для англійської, німецької, італійської, французькою та інших мов [12]. Для української мови можна використати словник ВЕСУМ, який був розроблений командою БрУК [13].

Для побудови моделі M можна використати бібліотеку spaCy на Python, яка містить готові алгоритми для англійської, китайської, французької, італійської, польської, іспанської мов [14]. Для української мови можна використати stanza-lang-uk — технологію для роботи з українськими текстами [15].

Блок-схема алгоритму запропонованої технології наведена на рис. 2.

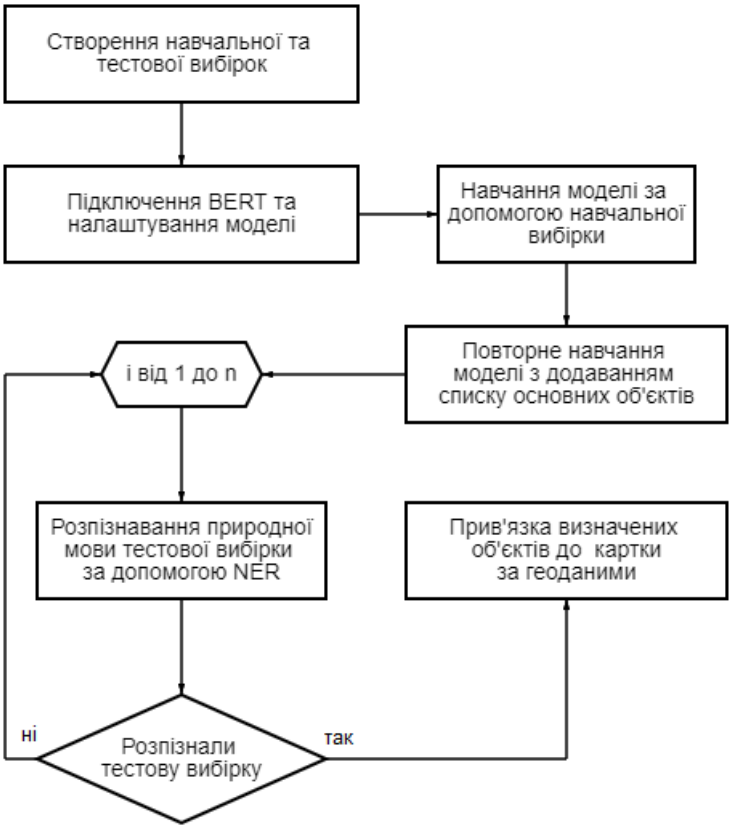

Рис. 2. Блок-схема алгоритму запропонованої технології інформаційної інтелектуальної технології автоматизованої геоприв'язки ЕТІ

Приклад застосування розробленої інформаційної інтелектуальної технології автоматизованої геоприв'язки ЕТІ

Розглянемо приклад застосування запропонованої технології на прикладі масивів вод басейну р. Південний Буг. Розглянемо водогосподарську ділянку (ВГД) «р. Південний Буг від гирла р. Іква до г/п Селище», яка у Державному водному Кадастрі України має офіційний номер UA_M5.4.0.02. На рис. 3 наведено приклад векторизованих водозбірних басейнів масивів вод цієї ВГД.

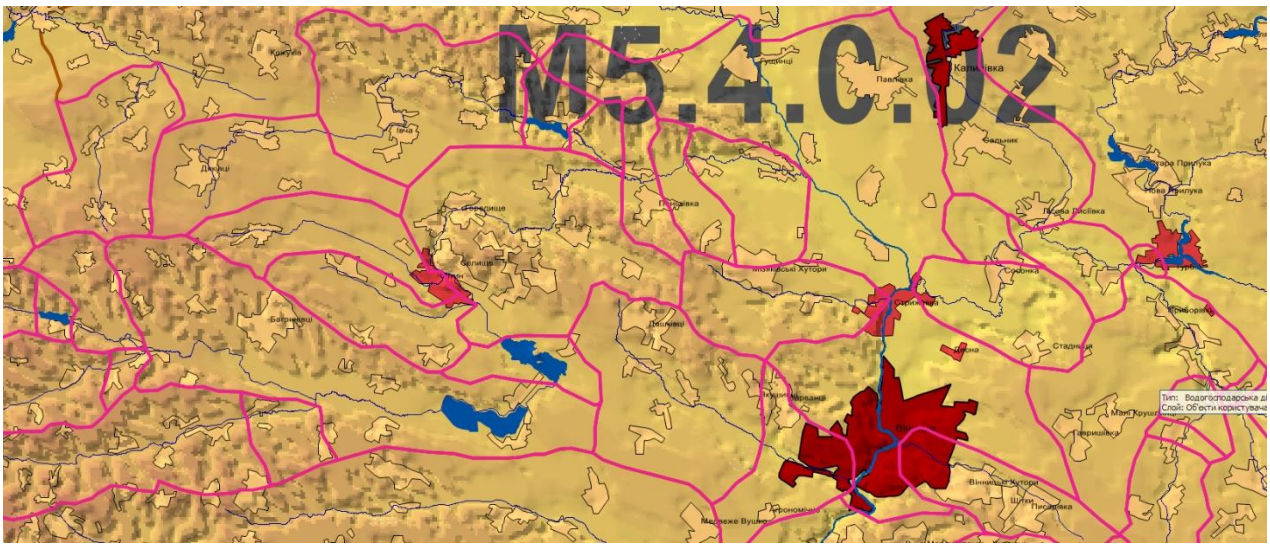

Рис. 3. Приклад векторизованих водозбірних басейнів масивів вод водогосподарської ділянки UA_M5.4.0.02 «р. Південний Буг від гирла р. Іква до г/п Селище»

Використовуючи запропоновані підходи для збирання та постопрацювання даних ГІС здійснено формування набору даних (рис. 4) для кожного масиву вод.

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

Рис. 4. Фрагмент результуючого набору даних

Для випробування технології були відібрані 3 масиви вод ВГД UA_M5.4, які містять міста Вінниця (UA_M5.4_0013), Калинівка (UA_M5.4_0181) та Хмільник (UA_M5.4_0011), відповідно. Для них була сформована базова тренувальна множина U_1 даних із площинних об'єктів більшого розміру (ВГД UA_M5.4.0.02, Вінницька область, райони Вінницької області) та тестова множина U_2 із площинних об'єктів меншого розміру (межі населених пунктів), лінійних (річки) та точкових об'єктів (місця скидання вод, водозабори підприємств-водокористувачів, пости моніторингу якості або кількості вод). На рис. 5 приведено приклад суміщення водозбірних басейнів водних масивів з іншими точковими об'єктами на карті.

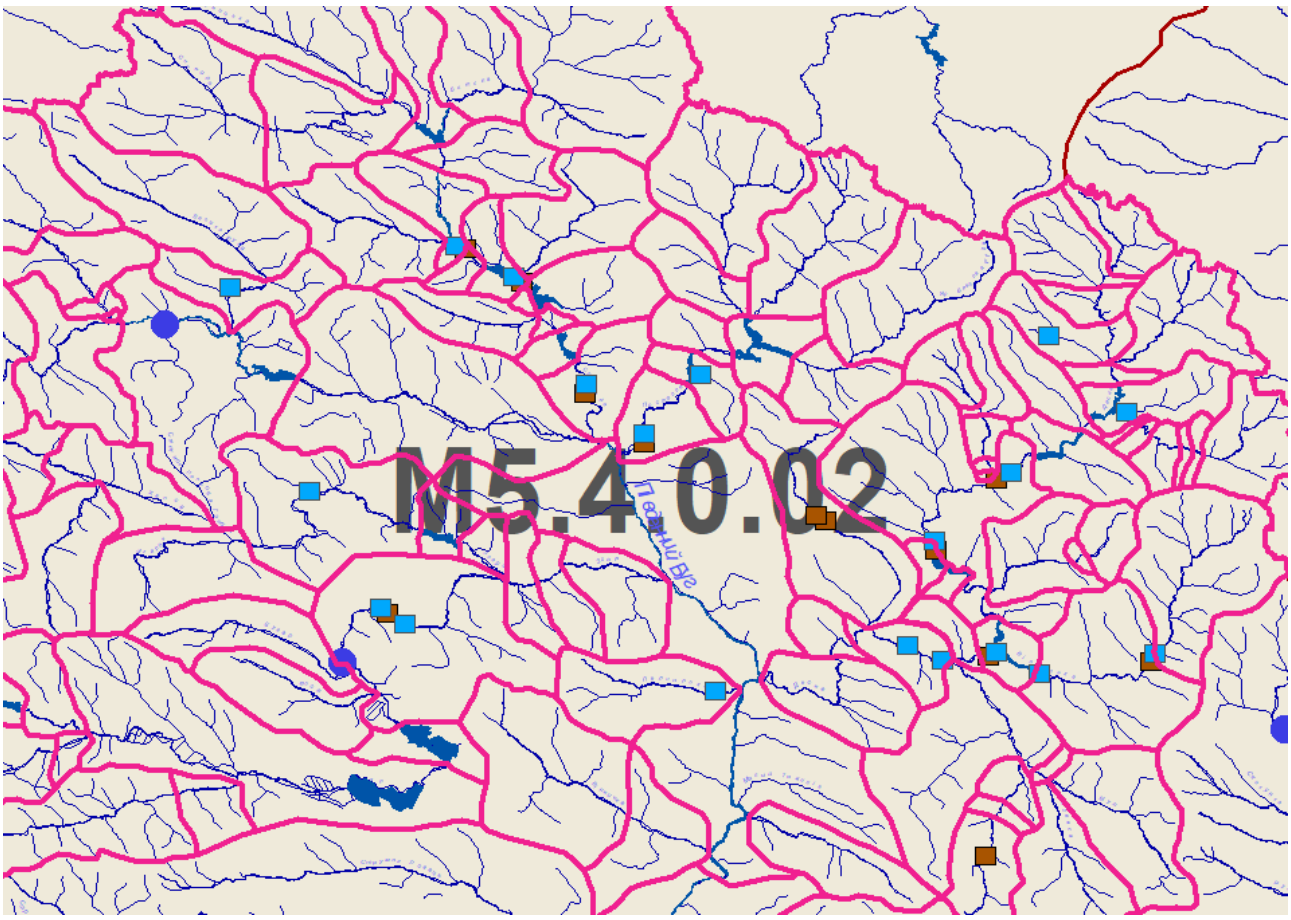

Рис. 5. Водозбірні площі водних масивів суміщені з постами гідрологічного контролю (сині кола), водозаборами (блакитні прямокутники) та скиди стічних вод (коричневі прямокутники)

Далі для цих множин за співвідношенням (1) сформовано множини з усіма словоформами X_1, X_2. Тестову множину X_2 розбито на 3 – окремо для кожного масиву вод.

Для випробування технології одним із співавторів статті Мокіним В. Б. за участі іншого співавтора Гораша М. А. та студентів Вінницького національного технічного університету (ВНТУ) Пасічнюка Д. В. і Радецького О. В. створено публічний датасет речень про стан водних ресурсів басейну р. Південний Буг, який розміщено на платформі Kaggle [16]. Цей датасет містить речення із монографії, співавторами якої є Мокін В. Б. та Крижановський Є. М. і яка має опубліковану і англомовну [17], і україномовну [18] версії. Для експериментів було використано саме англомовну частину датасету. Застосування запропонованої технології дозволило сформувати список сутностей-локацій, похибка f1_score для якого склала в середньому 0,78, під час якого була перевірена працездатність окремих запропонованих у роботі рішень. В подальшому планується провести більш повномасштабний експеримент для визначення повноти, точності та швидкості геоприв'язки заданої україномовної та англомовної екологічної текстової природно-мовної інформації для ряду масивів вод ділянки, також, для порівняння цих показників із показниками щодо реалізації такої геоприв'язки у традиційний для таких задач спосіб.

Висновки

У цій статі було описано розроблення інтелектуальної інформаційної технології автоматизованої геоприв'язки екологічної текстової природно-мовної інформації. Запропоновано новий підхід формування навчального набору даних шляхом розбиття розмічених сутностей-локацій та сутностей-організацій на окремі вибірки, які містять у

певний спосіб скомбіновані сутності, що характеризують площинні об'єкти більшої площі, та, окремо, ті, що характеризують менші площинні об'єкти, лінійні та точкові об'єкти. Такий розподіл даних дозволяє організувати багатоетапне уточнення результатів ідентифікації та моделей, які використовуються і це дозволяє забезпечити одночасно підвищення повноти, точності та швидкості геоприв'язки заданої екологічної текстової інформації.

Розроблено рекомендації щодо застосування цієї технології для української, англійської та інших мов, а також щодо алгоритму підготовлення вхідних картографічних даних з використанням ГІС-пакету програм ArcGIS. Наведено приклади застосування окремих елементів запропонованої технології до реальних текстових даних про стан масивів вод басейну р. Південний Буг:

1. За допомогою ГІС, наприклад ArcGIS, засобами оверлейного аналізу можна сформувати множину перетину водозбірних басейнів масивів вод з географічними об'єктами державних кадастрів: земельний, водний, кадастр лісового господарства, кадастр родовищ корисних копалин та інші.

2. Для тестування та створення ETI були використані такі масиви вод басейну р. Південний Буг UA_M5.4_0013, UA_M5.4_0181, UA_M5.4_0011. Запропонована технологія реалізована на Python з використанням технологій Named Entity Recognation, BERT, також були використані бібліотеки spaCy та TensorFlow.

3. По даних масивах вод було відібрано дані із інформацією про проблеми та описом річок, які знаходяться там. Також було використано монографію, в якій є опис річки Південий Буг. Ця монографія опублікована двома мовами (англійською та українською, у межах шведсько-українського проєкту за фінансування SIDA), що дозволяє навчати і тренувати інформаційну технології і на англійській, і на українській мовах. З цієї монографії було сформовано двомовний датасет, але для випробувань використано тільки англомовний текст.

4. Проведено успішні випробування окремих елементів запропонованої технології. Похибка становила 0,78. Далі планується розвивати цю технологію для роботи з українською мовою.

СПИСОК ЛІТЕРАТУРИ

1. Конвенція про доступ до інформації, участь громадськості в процесі прийняття рішень та доступ до правосуддя з питань, що стосуються довкілля [Electronic resource] / Access mode : https://zakon.rada.gov.ua/laws/show/994_015#Text.

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. Побудова масштабованої інформаційно-пошукової системи для управління річковим басейном на основі реєстрів та онтологічних моделей / В. Б. Мокін, І. І. Овчаренко, А. М. Лучко [та ін.] // Математичне моделювання в економіці. – Київ, 2019. – № 2 (15). – С. 45 – 56.

5. Концепція інтелектуальної NLP технології для геоприв'язки та класифікації відкритої текстової інформації про масиви вод [Електронний ресурс] / В. Б. Мокін, М. А. Гораш, Д. Пасічнюк, О. Радецький // Матеріали XV міжнародної конференції "Контроль і управління в складних системах (КУСС-2020)", м. Вінниця, 8-10 жовтня 2020 р. – Вінниця, 2020. – Режим доступу : http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/30607/KUSS%202020%20MHPR%20-%20NLP.pdf?sequence=1.

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

7. Стрижак О. Є. Засоби онтологічної інтеграції і супроводу розподілених просторових та семантичних інформаційних ресурсів / О. Є. Стрижак // Екологічна безпека та природокористування. – 2013. – № 12. – С. 166 – 177.

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 : Посібник з розпізнавання іменованих організацій 2021 року. Огляд розпізнавання іменованих сутностей (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. Морфологический анализатор pymorphy2 [Електронний ресурс] / Режим доступу : https://pymorphy2.readthedocs.io/en/stable/.

13. Словник ВЕСУМ та інші пов'язані засоби NLP для української мови [Електронний ресурс] / Режим доступу: https://r2u.org.ua/articles/vesum.

14. 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. План управління річковим басейном Південного Бугу : аналіз стану та першочергові заходи / [С. Афанасьєв, А. Петерс, О. Ярошевич, В. Мокін та ін.]. – К. : ТОВ «НВП «Інтерсервіс», 2014. – 188 с.

Стаття надійшла до редакції 22.12.2020.

Стаття пройшла рецензування 26.12.2020.

Мокін Віталій Борисович – д. т. н., професор, завідувач кафедри системного аналізу та інформаційних технологій.

Гораш Микола Андрійович – аспірант кафедри системного аналізу та інформаційних технологій.

Крижановський Євгеній Миколайович – к. т. н., доцент кафедри системного аналізу та інформаційних технологій.

Вінницький національний технічний університет.

Вуж Тетяна Євгенівна – к. т. н., доцент кафедри біологічної фізики, медичної апаратури та інформатики.

Вінницький національний медичний університет імені М. І. Пирогова.