УДК004.822

СТРУКТУРНИЙ ПІДХІД ДО ПОШУКУ ПРИРОДНО-МОВНОЇ ІНФОРМАЦІЇ

КИСЛЕНКО Ю.І., СЕРГЄЄВ Д.С.

Розглядаються можливості використання структурованого представлення природно-мовної інформації для покращення якості роботи пошуку у великому корпусі текстових даних, яким зокрема є мережа Інтернет. Як квант знань пропонується використовувати базову семантико-синтаксичну структуру, що постає похідною від структурно-функціонального рівня нейроорганізації зорового тракту. Теоретично обґрунтовуються переваги використання такої бази знань та приклади деяких прикладних проблем, вирішення яких може бути полегшене.

Ключові слова: база знань, пошук, інтернет, базова семантико-синтаксична структура, природно-мовна інформація, квант знань.

Keywords: knowledge base, search, Internet, basic semantic-syntactic structure, natural language information, quantum of knowledge.

1. Вступ

Питання текстового пошуку в мережі Інтернет вже не перший рік залишається однією з найбільш популярних тем для теоретичних досліджень та постійним об'єктом втілення прикладних інновацій. Це не дивно, враховуючи, що Інтернет бурхливо розвивається, а перед розробниками пошукових систем кожного дня постають нові виклики. Але, незважаючи на величезні зусилля та ресурси, що витрачаються на покращення релевантності пошуку, результати цих витрат наразі складно назвати задовільними [8]. На жаль, переважна більшість рішень щодо вдосконалення роботи пошукових систем дають досить невеликий приріст якості, оскільки приймають їх, в основному, через відсутність кращих пропозиції [5]. Основною причиною такого стану справ можемо справедливо назвати відносно молодий вік як самої мережі Інтернет, так і відповідних технологій пошуку. Хоча розвиток всесвітньої мережі розпочався близько 30 років тому, досі не можемо вважати його закінченим — а підходи, згідно з якими вона працює, не змінювалися з самого початку її існування. Ситуація дещо ускладнюється ще й тим, що технології веб-пошуку сильно залежать від досягнень у багатьох незалежних галузях науки, які за всі ці роки теж не стояли на місці. Як приклад наведемо представлення природно-мовного тексту як ізольованого об'єкта, без урахування процесів його синтезу — у наш час такий підхід вже вважається дещо наївним, все частіше зустрічається подання природної мови як самостійної структури збереження і передачі даних. І якщо такі розбіжності були допустимими на початку історії комп'ютерних технологій, то зараз, коли обсяг даних сягає сотень петабайт, вони є досить помітною перешкодою на шляху подальшого розвитку пошукових систем.

2. Постановка задачі

Для початку окреслимо ті задачі, які лежали в основі ідеології класичних пошукових систем (КПС) і які не вирішуються в повній мірі їх сучасними представниками.

Перше й найбільш очевидне, що спадає на думку – це можливість обробки природно-мовних (ПМ) запитів. Замість передбачених фантастами й теоретиками XX сторіччя пошукових систем, з якими користувач спілкується як з живою людиною — маємо користувача, що формує свій запит згідно з формальними системами правил пошукової системи.

Друга, не менш важлива, задача – це формування «загальної бази знань», тобто такої системи, яка б давала можливість знайти відповідь на будь-яке питання, або впевнитися, що ця відповідь людству не відома. На жаль, характерною рисою КПС є неминуча втрата інформації – риса, що є невід'ємною частиною будь-яких систем, які оперують не точними числами, а статистичними даними. Навіть якщо на кожному етапі роботи системи втрачається лише невелика частка інформації, це вже позбавляє усю систему детермінованості, тобто зменшує її надійність та погіршує якість результатів пошуку. Адже немає ніякої гарантії, що саме той результат, який шукав користувач, не було пропущено через недосконалість системи.

Можемо впевнено казати, що ці дві проблеми мають спільний корінь, а саме – відсутність моделі структурованого представлення природно-мовного повідомлення. Така модель автоматично вирішила б як проблему детермінованості результатів пошуку, так і проблему синтезу-аналізу природно-мовного повідомлення.

Дійсно, вирішенню цієї проблеми присвячено багато досліджень, і в результаті деяких з них навіть з'явилися дієздатні продукти – в тому числі такі гіганти як *Wordnet* та *Wolfram Alpha*. Були і спроби інтегрувати структуровані дані у результати текстового пошуку КПС [7, 10]. Ці системи працюють в певній мірі успішно, але в основі їх лежать виключно штучні структури, що не розраховані на всеосяжне охоплення всієї множини ПМ структур. Відповідно, якість їх роботи з наповненням бази катастрофічно падає, що сильно обмежує сферу їх можливого використання. Більш того, деякі з проблем КПС так само актуальні і для структурованих БЗ на основі штучних структур, у чому зможемо переконатися пізніше.

У роботах [2,6] запропоновано новий інтегральний підхід до моделювання мовленнєвої діяльності, в основі якого лежать сучасні досягнення у багатьох суміжних напрямах, зокрема нейрофізіології, психо-

логії та кібернетики. Дана стаття присвячена аналізу можливості використання цього підходу для покращення якості роботи текстового пошуку у мережі Інтернет.

3. Теоретичний аспект

Почнемо з визначення термінології, що буде використовуватись надалі в цьому тексті.

Індивідуальна мовна система (ІМС) — це модель мовленнєвої діяльності людини, побудована на принципах, запропонованих Л. Щербою [3]. Модель ІМС (рис. 1) складається з двох структурних частин — лінгвістичного процесора (ЛП), який виконує функції синтезу та аналізу природно-мовних повідомлень, та бази знань (БЗ), яка містить дані про навколишній світ у впорядкованому вигляді.

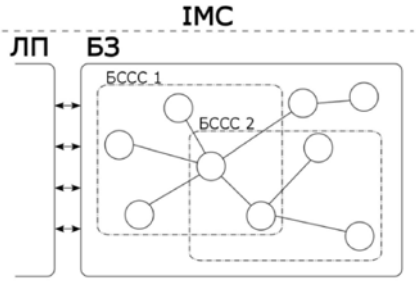

Рис. 1. Індивідуальна мовна система та її складові

Найменшою частиною (квантом знань) у моделі БЗ є БССС – структуроване представлення однієї ситуації зовнішнього світу, що існує в триєдності часу, простору та дії. Кожна така структура формується з елементів: об'єкти (*Obj*), суб'єкти (*Subj*) та дія (*Mov*), що доповнюються атрибутами *Attr* та мірою атрибута *Attr*(*Attr*) (рис. 2).

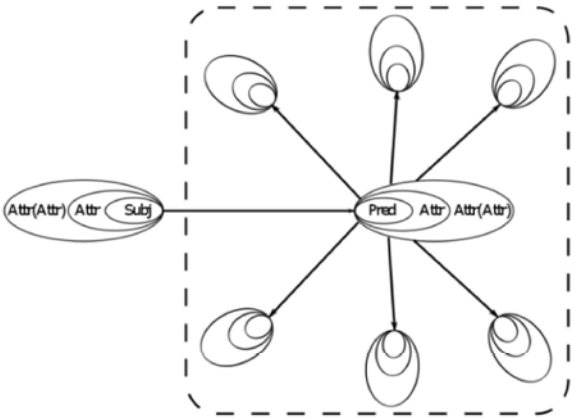

Рис. 2. Структура БССС

Структура БССС не суперечить канонам класичної лінгвістики; більш того, спираючись на досягнення досліджень нейрофізіологів, узагальнених Семіром Зекі [9], можемо засвідчити, що структура окремого повідомлення постає похідною від структурно-функціонального рівня нейроорганізації зорового тракту [2]. Головні етапи формування такої структури гарно простежуються в дослідженнях процесу опанування дитиною мовного ладу [1]. Головний висновок запропонованого бачення структурної організації мовного повідомлення досить чіткий та категоричний – довільний текст, як множина окремих повідомлень, формується з окремих стандартних структур, визначених на моно/поліпредикатному рівнях, отже, окремим квантом знань на мовному рівні постає базова семантико-синтаксична структура.

Таким чином, природно-мовна база знань (ПМБЗ) представлена відповідною частиною ІМС — БЗ, тобто сукупністю квантів знань (окремих БССС), пов'язаних між собою на семантичному рівні. Мережа БССС, у свою чергу, використовує мережу слів, що є текстовими ярликами об'єктів та процесів реального світу. Зазначимо також, що запропонована модель даних викликає цілком обґрунтовані асоціації зі способом збереження знань у нейромережі людини — нейронами (або їх ансамблями), пов'язаними системою асоціативних зв'язків.

Звісно, для повноцінного використання цієї моделі ІМС у пошукових системах необхідно за допомогою ЛП заповнити ПМБЗ з бази вхідних текстів. Така задача виглядає поки що досить складною, оскільки до реалізації ЛП ще далеко; утім, розробка повноцінного ЛП є окремою темою для дослідження. Для створення базового прототипу БЗ достатньо і досить простої емуляції ЛП. Наприклад, ЛП, здатний відокремлювати *Obj*/*Subj* та ідентифікувати *Attr*, дозволить частково виділити моделі БССС на як завгодно великому корпусі вхідних даних. Крім того, оскільки складові БССС добре розкладаються по майже незалежних рівнях, подальше вдосконалення моделі може бути виконано на основі вже заповненої бази, без спотворення або оновлення вже внесених даних.

Маючи уявлення про ІМС, ПМБЗ та БССС, можемо тепер приступити до розгляду конкретних проблем КПС — тих випадків, коли інтуїтивно правильний запит обробляється некоректно (не таким чином, як на нього реагував би живий співрозмовник). Задача полягає в тому, щоб визначити: по-перше, наскільки ці проблеми актуальні — наскільки якісно вони вирішуються на рівні КПС; по-друге, чи можливе їх вирішення засобами ПМБЗ.

4. Структура речення та слова

Природно-мовна інформація — це досить складний об'єкт дослідження; особливо сильно це впливає на задачі синтезу та аналізу природно-мовного повідомлення. Оскільки для пошуку необхідно проаналізувати запит користувача та порівняти його з інформацією у БЗ, проблеми виникають уже на найнижчих рівнях мовної організації – абзац, речення, слово. Хоча в основі ідеології КПС передбачена саме обробка ПМІ на структурному рівні, часто вони не можуть в достатній мірі адекватно розпізнати структуру запиту. Розглянемо це на прикладах.

Так, якщо у пошуковому запиті представлено словоформу, яка є омоформом (належить до різних слів з різним значенням — наприклад, «*три*» як чисель-

ник та «*три*» як форма дієслова «*терти*»), в ідеальному випадку повинна відбуватись однозначна ідентифікація і словоформи, і слова, до якого вона в даному випадку належить (рис. 3). У КПС взагалі відсутня можливість вказати форму слова у запиті, тобто інформація про частини мови може бути виділена (з певною ймовірністю) тільки за допомогою аналізу запиту. Якщо запит складається з кількох слів, результати пошуку так само можуть бути спотворені, навіть якщо запит – це повноцінна ПМ конструкція, частини якої пов'язані семантичними зв'язками.

У БЗ кожна словоформа входить у лексему — сукупність усіх можливих словоформ даного слова. Це дає змогу не тільки вказати, до якого саме слова відноситься дана словоформа у контексті запиту (і, відповідно, виконати пошук саме за цим словом), але й вказати її роль у реченні. Приклад подібної неоднозначності представлено на рис. 3. Звісно, так само можна й автоматично визначити роль слова, але це більше не є єдиним доступним способом. На класичному прикладі неоднозначності *«души прекрасные порывы»* визначити ролі слів можливо виключно в ручному режимі.

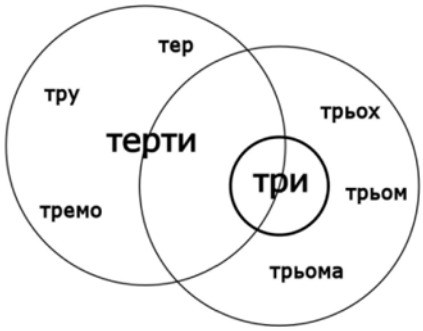

Рис. 3. Нерівнозначний перетин лексем на прикладі словоформи «терти»

Власне, пошукові запити, що складаються з кількох слів, варті окремого розгляду. Як правило, такі запити формують повноцінне ПМ повідомлення, а отже, при їх обробці мають значення не лише форми слів, але й їх порядок та знаки пунктуації. У КПС для цих цілей повсюдно використовується оператор «лапки», що відповідає пошуку за точним порядком слів, але при цьому ніяк не враховуються знаки пунктуації. Отримані таким чином результати, хоча й не є зовсім некоректними, містять велику кількість нерелевантних записів.

Розглянемо пошук за запитом «*зробити згодом*» (слова у запиті явно пов'язані). У результатах присутні не лише релевантний «Його успіх не тільки дозволив *зробити згодом* кілька тематичних продовжень», а й очевидно зайві: «Однак, ми заборонили це *зробити. Згодом* виявилося, ... » та «Що і як ми можемо для цього *зробити — згодом*». Як бачимо, тут ігноруються не тільки знаки пунктуації, але й сама структура речення взагалі («… *зробити. Згодом* …» — два слова з запиту явно належать до різних речень).

У ПМБЗ вхідний текст зберігається таким чином, що усі знаки пунктуації (а також інші елементи тексту, що не містять смислового навантаження — вставні слова, вигуки тощо) зберігаються у їх початковому вигляді. Більш того, розділені за змістом окремі речення трансформуються у окремі БССС, тобто помилки з належністю різних частин запиту до різних БССС взагалі ліквідуються.

Зазначимо, що *Google* повертає 830,000 документів, де зустрічаються обидва слова «*зробити*» та «*згодом*». Це, за мірками сучасних БЗ — порівняно мала кількість запитів, тобто для подальшого уточнення результатів пошуку вже не є критичною швидкодія аналізатора.

5. Смислові (семантичні) зв'язки

Від структурних проблем, які вирішуються на рівні структури об'єкта, перейдемо до проблем семантики – більш складних випадків обробки ПМІ. Ці проблеми за своєю природою скоріше філософські, ніж технічні, але вони впливають на роботу пошукових систем в тій же – якщо не більшій – мірі. В широкому сенсі семантичні проблеми –це проблеми неточного відображення засобами природної мови відповідного явища або об'єкта реального світу. Більш формально, семантичні проблеми пошукових систем – це такі випадки, коли в запиті не міститься достатньо інформації для виконання якісного пошуку (тобто, для адекватної обробки запиту необхідно мати певний рівень базових знань). Власне, задача зводиться до вибору окремої БССС з усього лексично-семантичного оточення її складових, як це показано на рис. 4.

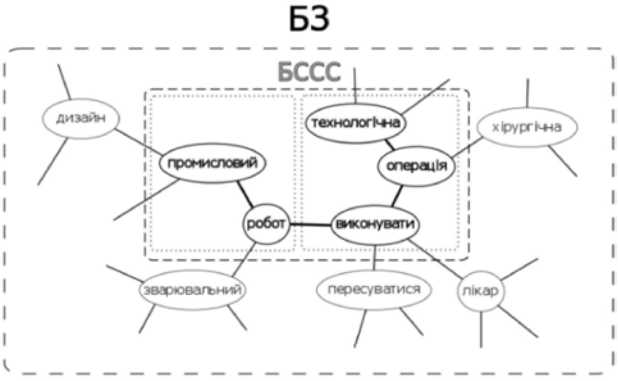

Рис. 4. Окрема БССС та семантичне оточення її складових

Першим за популярністю серед них є, безсумнівно, питання обробки синонімів — тобто такої ситуації, коли під час пошуку необхідно враховувати не лише саме слово, але й можливі варіанти його заміни (синоніми). Тут слід згадати розподіл синонімів на абсолютні (ті, які завжди мають подібні значення, як-то «обман» та «омана») та контекстні (ті, які мають

спільне значення тільки за певних умов або обмежень). З абсолютними синонімами КПС наразі працюють з використанням статистичних методів — хоча й не ідеально, але досить успішно. Робота ж з контекстними синонімами у КПС представлена слабо. Це пояснюється тим, що для коректної обробки контекстних синонімів треба спочатку коректно враховувати контекст, а питання виділення контексту у КПС фактично не реалізовано.

У ПМБЗ контекстні синоніми визначаються на рівні субструктури або структури — через подібність БССС. Іншими словами, можемо вважати слова контекстними синонімами, якщо вони з'являються в одній й тій самій ролі у одній й тій самій структурі. Такий підхід дозволяє досить точно підтвердити, що деякі слова є синонімами в даному контексті, або ж показати, що використання їх як синонімів за даних умов недопустимо. Також зазначимо, що такий підхід дозволяє використовувати нові, набуті значення слів — наприклад, сленг або усталені помилки. Так, суто сленгове слово «*гуглити*», що означає «шукати в пошуковій системі *Google*», або більш широко «шукати в інтернеті», використовувалося задовго до першої його появи у словниках. Оскільки сценарії його використання переважно збігаються з використанням «шукати» в контексті «шукати в інтернеті», у ПМБЗ ці слова були б контекстними синонімами. Водночас, результати зі словом «нишпорити», яке також є синонімом до «шукати», були б серед останніх, оскільки його сценарії використання значно відрізняються. Інший приклад наведено на рис. 5, хоча слова «йти» та «крокувати є синонімами, заміна їх допустима лише в одному випадку з двох представлених.

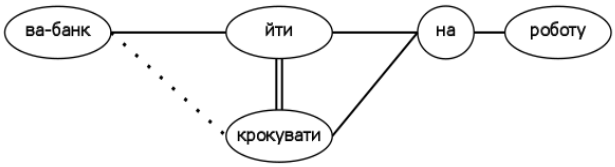

Рис. 5. Контекстна синонімія на прикладі слів *«йти»* та «*крокувати*»

Розкривши проблему синонімів, не можемо оминути й зворотній випадок — проблему багатозначних слів: одне й те ж саме слово, що може мати різні значення залежно від контексту. Зазначимо, що тут маються на увазі не тільки й не стільки омографи (різні словоформи різних слів, що мають однакове написання), а саме слова, які змінюють смислове навантаження залежно від контексту їх використання. Прикладом цього є слово «*поле*» — «електромагнітне поле» у фізиці, «засіяне поле» у сільському господарстві, «поле дослідження» в науці тощо.

Багато в чому ця ситуація подібна до аналізу запиту, що складається з кількох слів. Так само в результаті не враховується порядок та роль слів у реченні, але до цього додається ще й спотворення значення слова. Дійсно, при пошуку за запитом «*поле дослідження*» отримуємо як релевантний результат «Понятійно-термінологічне *поле дослідження*», так і нерелевантний «Українські офісні працівники готові змінити роботу в офісі на город чи *поле — дослідження* ». На даному прикладі ця проблема проявляється так само, як і при звичайному складному запиті — тільки у появі зайвих результатів. Але, на відміну від складного запиту, при семантичному спотворенні набагато складніше визначити, наскільки результат відповідає початковому задуму користувача — адже в багатьох випадках, особливо коли це стосується спеціалізованої термінології, різні значення слова можуть бути досить близькі за смисловим навантаженням.

Як було вказано вище, у ПМБЗ в такому випадку можемо визначити контекст або вручну, або за допомогою автоматичного аналізу запиту. Саме у випадку багатозначних слів дуже доречним виглядає використання такого потужного аспекту ПМІ як *когнітивний зворотній зв'язок*. Ідея його використання полягає у тому, що за оточенням слова передбачаються можливі варіанти закінчення запиту; таким чином, використання правильно побудованого запиту дозволяє істотно зменшити кількість релевантних результатів. Так, пошук за запитом «*король*» поверне дуже велику кількість результатів, тобто практично не буде корисним. Уточнення «*король Франції*» не тільки додасть до запиту новий елемент «*Франція*», але і визначить структурні (члени речення, частини мови) й навіть семантичні («*король*» як титул) зв'язки між ними. Уточнення «*король Франції у XIV сторіччі*» обмежить поле пошуку декількома сутностями (королями Франції з 1400 до 1500 років) і пов'язаними з ними результатами.

Ще раз окремо зазначимо, що обчислювальна складність кожного уточнення буде падати зі зростанням складності запиту, оскільки фактично від пошуку слова у БЗ фокус буде переходити до пошуку структури у ПМБЗ за повною відповідністю. Так, навіть на матеріалах пошуку у *Google* бачимо, що при переході від запиту «*поле*» (майже 90 000 000 результатів) через «*поле дослідження*» (близько 3 000 000 результатів) до «*поле наукового дослідження*» (1 000 000 результатів) – всього за 3 ітерації – кількість результатів зменшується майже на 2 порядки (у 90 разів).

В останню чергу розглянемо технічні проблеми КПС – тобто ті проблеми, які зумовлені не стільки складністю природної мови як об'єкта досліджень, скільки особливостями реалізації КПС у їх сучасному вигляді.

По-перше, ще раз підкреслимо характерний недолік усіх систем, що працюють з ключовими словами: недостатнє охоплення вхідних матеріалів. Очевидно, що при створенні індексу над документом пошукові роботи обробляють лише малий відсоток його змісту – це зумовлено самим визначенням індексу. Отже, велика частина вхідних даних просто ігнорується, в результаті чого при роботі з пошуковими системами

доводиться орієнтуватися саме на популярні ключові слова, а якість пошуку прямо залежить від якості алгоритму побудови індексу.

По-друге, КПС у більшості випадків повертають неймовірно велику кількість результатів, в той час як доля релевантних серед них дуже мала. Як правило, запит загального характеру повертає мільйони результатів, спеціалізований запит – тисячі, а більша частина релевантних знаходиться на перших 10..20 сторінках (еквівалент 100-200 позицій), і навіть з них лише 10-15% дійсно дають відповідь на запит.

Цю проблему частково вирішує уточнення запиту, але, як правило, ця операція використовується досить рідко — адже при уточненні кожного разу втрачається певна частина значущої інформації, і ця втрата є мультиплікативною. Іншими словами, користувачі намагаються змінити формулювання свого запиту замість використання функції пошуку у знайденому.

І, нарешті, знову повернемося до використання ключових слів. Хоча сама по собі ця технологія не є чимось поганим, її повсюдне використання у текстовому пошуку призвело до вкрай негативних наслідків. Замість адаптації пошукових систем до природномовних запитів користувача бачимо адаптацію користувачів до запропонованої мови введення запитів; замість покращення рейтингу документа шляхом розміщення нової або актуальної інформації бачимо «накрутку» штучних показників пошукових систем; замість використання потужної мови пошукових запитів бачимо спроби через довгі ланцюжки посередників вийти до першоджерела шуканих даних [4]. Хоча пошуком у такому вигляді цілком можливо користуватись, питання про його якість та оптимальність, кажучи обережно, залишається відкритим.

6. Висновки

Незважаючи на те, що тематика текстового пошуку у великих базах природно-мовних даних (однією з яких є мережа Інтернет) вже багато років є популярним напрямком досліджень, якість відповідних систем залишається не дуже задовільною. Більшість класичних пошукових систем, так само як і найбільш популярні бази структурованої природно-мовної інформації, досі мають істотні недоліки як на рівні архітектури, так і на рівні реалізації.

Використання запропонованого Л. Щербою підходу з чіткою ідентифікацією функціонального навантаження ІМС через складові ЛП/ПМБЗ з формально визначеною структурною організацією кванта знань – БССС дозволяє вже на етапі моделювання вирішити деякі проблеми класичних пошукових систем. Хоча цей підхід все ще потребує як більш детального теоретичного вивчення, так і експериментальної перевірки, видається очевидним, що спроби надалі покращувати тільки кількісні показники роботи пошукових систем вичерпали свій потенціал, і поява докорінно нових підходів у цій галузі – це лише питання часу.

Наразі неможливо сказати, наскільки якісно практична реалізація ІМС буде вирішувати усі освітлені у даній статті проблеми, але вже можемо впевнено стверджувати, що використання навіть окремих її елементів у пошукових системах може позитивно вплинути на якість їх роботи, а сама модель ІМС є досить цікавим і перспективним полем для подальших досліджень.

Література: 1. *Гвоздев А.Н.* От первых слов до первого класса / Александр Гвоздев. Саратов: Изд-во Саратовского университета, 1981. **2.** *Кисленко Ю.І.* Архітектура мови (лінгвістичне забезпечення інтелектуальних інтегрованих систем) : Учбовий посібник. К.: Віпол, 1998. 343 с. **3.** *Щерба Л.В.* Языковая система и речевая деятельность. Л.: Наука, 1974. **4.** *Hųlscher C.* Web search behavior of Internet experts and newbies / C. Hųlscher, G. Strube. // Computer Networks. 2000. №33. P. 337–346. **5.** *Hsinchun C.* Internet Categorization and Search: A Self-Organizing Approach / C. Hsinchun, C. Schuffels, R. Orwig // Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries. 1996. №7. P. 88–102. **6**. *Kyslenko Y.* Cognitive architecture of speech activity and modelling thereof / Y. Kyslenko, D. Sergeiev // Biologically Inspired Cognitive Architectures. 2015. №12. P. 134–143. **7.** *Kruse P.* Clever Search: A WordNet Based Wrapper for Internet Search Engines / P.Kruse, A. Naujoks, M. Kunze, D. Roesner // Proceedings of 2nd GermaNet Workshop 2005. 2005. **8.** *Tirri H.* Search in vain, challenges for Internet search / Tirri. // Computer. 2003. №36. P. 115–116. **9.** *Zeki S.* A visual image in mind and brain: Collection of papers / Semir Zeki. // The World of Science. 1992. №11. P. 33–41. **10.** Google Knowledge Graph [Електронний ресурс] // Wikipedia, the free encyclopedia. 2012. Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Knowledge_Graph.

Кисленко Юрій Іванович, канд. техн. наук, доцент кафедри технічної кібернетики ФІОТ НТУУ «КПІ». Наукові інтереси: сенсорика біологічних та технічних систем, робототехніка, штучний інтелект, інформаційні природномовні технології. Адреса: Україна, 03187, Київ, вул. Теремківська, 11, кв 13. Тел. +38(097)522-35-22, email: y.i.kislenko@gmail.com

Сергеєв Данило Сергійович, аспірант кафедри технічної кібернетики ФІОТ НТУУ «КПІ». Наукові інтереси: інформаційні природно-мовні технології, бази знань, обєктно-орієнтовані бази даних. Адреса: Україна, 02192, Київ, вул. Космічна, 12, кв. 16. Тел. +38(095)402-97-40, email: d.sergeiev@gmail.com

Kyslenko Yuri Ivanovich, Ph.D., associate professor at the Department of Technical Cybernetics, FICT NTUU «KPI». Research interests: sensorics of biological and technical systems, robotics, artificial intelligence, natural language information technology. Address: Ukraine, 03187, Kiev, Teremkovskaya str. 11, apt. 13, tel. +38(097)522-35-22, email: y.i.kislenko@gmail.com

Sergeiev Danylo Sergiyovich, postgraduate student at the Department of Technical Cybernetics, FICT NTUU «KPI». Research interests: natural language information technology, knowledge bases, object-oriented databases. Address: Ukraine, 02192, Kiev, Kosmichna str. 12 apt. 16. tel. +38(095)402-97-40, email: d.sergeiev@gmail.com