УДК 621.3: 811.161.2

В.Ю. Дудник
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ВИКОРИСТАННЯ СИСТЕМНОГО АНАЛІЗУ ДЛЯ РОЗВ'ЯЗКУ АНАФОРИ ПРИРОДОМОВНИХ ТЕКСТІВ ДЛЯ УКРАЇНСЬКОЇ МОВИ

У статті автори проводять дослідження існуючих алгоритмів системного аналізу з метою розв'язання питання анафори природомовних текстів з української мови. Наводиться детальне обґрунтування актуальності поставленої задачі, оскільки стрімкий розвиток інформаційних технологій дає потужний поштовх до проведення досліджень в галузі аналізу текстових масивів.

У роботі досліджені функціональні можливості відомих підходів системного аналізу до розв'язку питання анафори природомовних текстів для української мови. Наведений алгоритм для проведення обчислень семантико-синтаксичних критеріїв для наперед вибраної пари антецедент-анафора.

Ключові слова: *анафора, комп'ютерна лінгвістика, аналіз, системний підхід, автоматизація, алгоритм.*

В.Ю. Дудник
Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского»

ИСПОЛЬЗОВАНИЕ СИСТЕМНОГО АНАЛИЗА ДЛЯ РЕШЕНИЯ АНАФОРЫ ПРИРОДОМОВНИХ ТЕКСТОВ ДЛЯ УКРАИНСКОГО ЯЗИКА

В статье авторы проводят исследования существующих алгоритмов системного анализа с целью решения вопроса анафоры природомовних текстов с украинского языка. Приводится детальное обоснование актуальности поставленной задачи, поскольку стремительное развитие информационных технологий дает мощный толчок к проведению исследований в области анализа текстовых массивов.

В работе исследованы функциональные возможности известных подходов системного анализа к решению вопроса анафоры природомовних текстов для украинского языка. Приведенный алгоритм для проведения вычислений семантико-синтаксических критериев заранее выбранной пары антецедент-анафора.

Ключевые слова: *анафора, компьютерная лингвистика, анализ, системный подход, автоматизация, алгоритм.*

V.Y. Dudnik
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

USE OF SYSTEM ANALYSIS FOR THE SOLUTION OF THE ANAPHOR FOR NATURAL LANGUAGE TEXTS FOR THE UKRAINIAN LANGUAGE

In the article, the authors conduct research of existing algorithms of system analysis in order to solve the problem of the anaphora of natural language texts from the Ukrainian language. The detailed justification of the relevance of the task is given, because the rapid development of information technology gives a powerful impetus to research in the field of analysis of text arrays.

The analysis of objects taken from natural texts is a subject area of automatic processing of natural language NLP (Natural Language Processing), the essence of which work to achieve from the computer a full understanding of natural texts. In this regard, the issue of the anaphora is a central issue in cognitive linguistics, besides this, the anaphora serves as a means of communication in transformational grammar, as well as anaphoric type expressions contained in a numerical reference system, since they denote the coercive nature of the names, which means direct reference.

The development of literary works has shown that the task of automating processes in solving the issues of anaphora occupies a leading place among the scholars. This can be explained by the fact that in natural language texts, in most cases, speech expressions are presented, which are difficult to interpret without understanding the previous contents of the test array, as a rule they are anaphoric pronouns, for example, the pronouns belonging to the third person he, he.

In this work the functional possibilities of known approaches of system analysis to the solution of the question of the anaphora of natural texts for the Ukrainian language are explored. The algorithm for calculating the semantic-syntactic criteria for the pre-chosen pair of antecedent-anaphora is given.

In this regard, there is a research interest in addressing the issue of anaphora natural texts with the involvement of automated systems. The process of automating the detection of anaphoric links will enable a successful combination and approximation of the automated process for the processing of textual information into the model of the speech model of a person, which in turn will open new promising directions for the solution of applied tasks, referring to automatic translation, the search of the necessary information , machine annotation and referencing.

Keywords: *anaphora, computer linguistics, analysis, systematic approach, automation, algorithm.*

Постановка проблеми. Виникнення комп'ютерних технологій є універсальним інструментом, яким повинен володіти кожен науковець сьогодення, адже завдяки використанню ЕОМ, сучасний науковець збільшує свої здібності та має змогу розв'язувати ускладнені від звичних задач за порівняно менший часовий проміжок [3]. Найважливішим у даному поєднанні

«людина-машина» є висока точність отриманих результатів у порівнянні із звичайними рутинними методами обчислення.

Сучасний розвиток філологічних наук відбувається в часи коли інформаційні комунікаційні технології є невід'ємною частиною людського життя. Значна частина аудиторії, яка використовує ІТ-розробки у своїх наукових експериментах, дослідженнях не є спеціалістами даної галузі, а є представниками інших не менш важливих та цікавих наук, проте провести ґрунтовне дослідження потрібного напрямку, узагальнити отриманий результат експерименту, у більшості випадків неможливо без використання комп'ютера з метою автоматизації процесів у складних системах різної природи.

У зв'язку з цим якщо проглянути будову вікон програмних додатків, то можна прослідкувати досить простий та зрозумілий інтерфейс, який зосереджений на представників різноманітних професій. Найбільш доступним інтерфейсом для кожного з нас є засоби мовлення на рідній мові, тобто це можливість вимовляти в голос, тобто синтезувати необхідний звук із запропонованого тексту, а з іншої сторони це розробки засобів для здійснення аналізу та синтезу природомовних текстових масивів [6,7].

Ключовим питанням при роботі із інформацією є її опрацювання до необхідного виду, тобто вихідну інформацію необхідно подати у вигляді деякого масиву або вектору, інформаційної вибірки, якщо виникають складнощі у перетворенні то можливе подання інформації у графічному вигляді, головна мета при опрацюванні природомовних текстів – подати текстові дані таким чином, щоб фахівець зміг прийняти правильне рішення або використовувати інформацію для подальшого перетворення до потрібного виду. Успішна практична реалізація потребує залучення потужного апарату системного аналізу [5].

Системний аналіз має широкий спектр застосування, зокрема в філологічних науках, які останнім часом здійснюють перехід до автоматизації великої кількості рутинних процесів [3,4].

Для деталізації вивчення поставленого питання розв'язку анафори природомовних текстів системний аналіз у своєму арсеналі має різноманітні системні підходи, які в повній мірі вивчають об'єкт дослідження, явища та процеси які з ним пов'язані. Іншими словами системний підхід являє собою практичну реалізацію необхідних принципів цілісності досліджуваних процесів, які можуть мати різний ступінь складності, внутрішні зв'язки та інші характеристики їх виникнення та розвитку.

Сьогодення лінгвістичного аналізу складно уявити без використання методів та моделей, які дають змогу побудувати необхідну систему аналізу, та автоматизувати роботу значної кількості вчених, побудована система зможе здійснювати детальний аналіз та інтелектуальну обробку природомовних текстів, даний напрям займає вагоме місце у наукових працях науковців сьогодення.

Аналіз останніх досліджень і публікацій. Огляд літературних джерел показує, що сучасні системи аналізу мають високий показник ефективності роботи під час аналізу природомовних текстів, які полягають у тому що при обробці текстових масивів використовується білінгвістична семантична база, яка є досить об'ємною та оперує близько 120 000 словесними одиницями та внутрішніми зв'язками, які можуть носити асоціативний характер [6].

Розвиток інтелектуальних технологій для обробки текстової інформації займає перші місця у наукових дослідників та розробників програмного забезпечення. Це явище можна пояснити тим, що комбінування основних методів, базових підходів на перетині декількох наук завжди є актуальним, оскільки виникають перспективи для майбутнього розширення кола наукових інтересів кожного із представників досліджуваних галузей наук.

Розв'язанню питання анафори присвячені роботи таких вчених: А. Анісімова, Н.Д. Арутюнова, А. А. Кибрика, Е. В.Падучева, О. Палагіна, О. Литвиненка, Ю. Марчука, А. Д. Шмелева, K.W. Chang, C.J. Hsieh та ін.

У роботі Толпегіна П.В., Вєтрова Д.П., Кропотова Д.А. показано, яким чином можна сконструювати класифікатор, який використовуючи синтаксичну інформацію дозволяє встановити поєднання антецедент-анафора або навпаки доводить їхню несумісність [4].

На даному етапі вирішення питання синтаксичного аналізу тексту, науковці-лінгвісти мають успішні результати, проте вчені, в коло наукових інтересів, яких входить проведення семантичного аналізу природомовних текстів, приходять до багатьох проблем, вирішення яких не можливе без використання комп'ютерних технологій, а саме – питання вирішення анафори у природомовних текстах з української мови. Сучасні дослідники виокремили чотири типи анафори такі як: прислівникова, займенникова, іменникова, нульова [3,5].

Вивчаючи новітні наукові розробки в галузі системного аналізу можна виокремити такі перспективні напрями розвитку:

1. Теоретична частина системології у філологічних науках;
2. Практичне застосування системотехніки у задачах прикладного характеру;
3. Методологія системного аналізу природомовних текстів.

Можна із впевненістю стверджувати, що окреслені вище напрями розвитку системного аналізу, як науки прикладного значення, становлять концепцію розвитку системного підходу.

Опрацювання літературних джерел дає зрозуміти, що використання системного аналізу перш за все представляє можливість досліджувати процеси, які відбуваються у складних системах завдяки моделюванню функціонального стану розглядуваної системи та створює перспективи до зростання ефективності роботи складних систем та автоматизації її процесів на основі математичного моделювання та комп'ютерній реалізації побудованих алгоритмів [1,4].

Таким чином, автоматизація вирішення питання анафори можлива із залученням апарату системного аналізу, що в свою чергу потребує додаткових досліджень існуючих алгоритмів та створення перспектив для майбутніх досліджень.

Формулювання цілей статті (постановка завдання). Провести дослідження алгоритмів системного аналізу, результати роботи яких спрямовані на автоматизацію обробки природомовних текстів української мови, зокрема шляхи розв'язання питання анафори.

Виклад основного матеріалу дослідження. Розвиток наукового прогресу призводить до збільшення інформаційних обсягів, за помірними оцінками фахівців, накопичення інформації можна розглядати за принципами геометричної прогресії, але всі ми добре знаємо що кількість це не завжди якість, тобто для того щоб прийняти вдале рішення необхідно в повній мірі володіти необхідними інформаційними ресурсами, які будуть містити повні дані, достовірні та крім цього вони повинні бути доступні для осіб які займаються вирішенням поставленого питання.

У зв'язку з цим можна з впевненістю стверджувати, що на сьогоднішній день, інформація – це фундаментальний ресурс для прогресу суспільства, а засобом для вирішення питання продуктивності роботи сучасного спеціаліста є інформаційні системи та технології. Не секрет, що основний вид інформації, яку людина здатна сприймати та відразу опрацьовувати, залежно від розумових здібностей є текстова інформація.

Зародження кібернетики – науки про узагальнені принципи для управління громіздкими (множинними) системами різної природи виникнення, дало змогу зрозуміти кожному пересічному жителю планети, що всередині будь-яких живих організмів можна встановити інформаційні процеси які там відбуваються, а сконструйовані інженерами технічні винаходи теж можуть функціонувати завдяки інформаційному управлінні.

Розглянемо актуальне питання сьогодення – питання автоматизації обробки природомовних текстів. Якщо проводити аналіз текстової інформації на основі використання комп'ютерних технологій то такий аналіз перш за все полягає на первинних евристичних алгоритмах для опрацювання природомовних текстів, а також до цього списку додаються бази даних, які будуються за всіма лінгвістичними характеристиками [2].

На практиці філологи використовують наступні системи аналізу для розв'язання поставлених проблем у природомовних текстах:

1) Система проведення семантичного перебору текстів – дана система здійснює аналіз запропонованого тексту та може охарактеризувати документ з точки зору його приналежності до досліджуваної тематики.

2) Система перекладу «Вітамін Е» – дає змогу вдосконалити автоматизований машинний переклад, така система виникла на основі алгоритмів та системних підходів білінгвістичного аналізу, який носить семантичний характер.

3) Система сортування Інтернет-повідомлень – така система проводить ґрунтовний аналіз інформаційних текстових надходжень до глобальної мережі, що дає можливість захистити необхідний контент, трафік та перешкодити до його доступу.

Дані системи аналізу дають змогу виконати ґрунтовний аналіз текстових даних та відшукати відповіді на більшість питань лінгвістичного характеру.

Варто відмітити, що якщо необхідно розв'язати питання займенникової анафори то необхідно сконструювати класифікатор, підґрунтям до побудови якого будуть інформативні дані стосовно вивчених властивостей синтаксичного та семантичного характеру досліджуваних слів анафори та в свою чергу антецедента. Вказана інформація дає змогу сформулювати висновок відносно їхнього поєднання або несумісності [3].

Зосередимо свою увагу на вивченні основних підходів до розв'язання питання анафори у природомовних текстах з української мови. Найпопулярнішим та результативним методом відокремлення анафори, зокрема займенникового типу є метод Міткова, суть даного методу полягає у об'єднанні відомих евристичних підходів до єдиної зручної при роботі форми, яка дає змогу виокремити необхідні критерії аналізу антецедентів по належністю до анафори, суть якої полягає у використанні базової інформації, синтаксичного та морфологічного характеру, де кожному критерію присвоюється певна вага та кінцеве підсумовування ваг критеріїв, яким підходить по всім поставленим умовам розглядувана пара із текстового масиву.

Наукові розробки не стоять на місці, а це говорить проте, що метод Міткова отримав ряд модифікацій, доповнень роботи базового алгоритму, які в основному полягають у використання семантичної інформації в чистому вигляді та семантичної інформації, яка зображується у вигляді тензорної моделі.

Розглянемо критерії оцінювання, які беруть за основу синтаксичну інформацію. Вказані критерії опрацьовують інформацію різного походження, тобто вони здатні розпізнавати бінарні, дійсні, натуральні значення, де широкого застосування здобули бінарні критерії, які користуються законами дискретної математики, тобто процес кодування є цілком логічним: якщо маємо 0 – то стверджуємо про невиконання критерію, якщо 1 – то критерій успішно виконується.

Для кращого сприймання бінарних критеріїв наведемо перелік найпопулярніших: бінарні характеристики трьох родів, властивості множини анафори, а також антецеденту; підмет речення представляє собою – антецедент; якщо антецедент виступає як власна назва. Якщо говорити про критерії, які виділяють натуральне значення: враховується кількість входжень антецедента у досліджуване речення, текст, анафоричні зв'язки; береться до уваги значення власних назв, які зосереджені між анафорою та антецедентом [4].

Складними у реалізації є критерії з дійсними значеннями, основними з яких є: обчислення середнього значення двійки антецедента-анафора із використанням підходу опорних векторів, для них перед цим була вказана несумісність на досліджуваному інтервалі між антецедентом та анафорою; потрібний кандидат вступає у альфа-зв'язок від того слова, яке розташоване в ідентичній фактор-множині. Складність даного критерію можна пояснити занадто громіздкою практичною реалізацією, оскільки вони рідко використовуються незалежно від інших критеріїв, тобто для того щоб використати критерії з дійсними значеннями необхідно перед цим вже мати дієвий класифікатор [6].

Варто відмітити, що після перезавантаження необхідної кількості разів із використанням додаткових критеріїв, які необхідні залежно від властивостей текстового масиву, вид класифікатора буде іншим тобто буде необхідний додатковий перерахунок отриманих даних після цих критеріїв, які теж потрібно брати до уваги в процес тренування.

Розглянемо модифікований метод для проведення оцінки пари антецедент-анафора на основі використання тензорного аналізу [5]. Для цього будемо досліджувати слово або словосполучення антецеденту m та початкове речення якому воно належить n. Крім цього до розгляду вводиться безпосередньо речення n' яке містить анафоричний займенник m'. Розглянемо основні етапи для досліджуючої пари m та m' побудови критеріїв семантико-синтаксичних властивостей.

Перш за все необхідно сконструювати управляючі простори $\tilde{n}n$ та $\tilde{n}n'$ для заданих речень m та m', побудову такого простору можна виконати використовуючи метод який висвітлений в [1]. Синтаксичний простір будемо розглядати як систему, яка об'єднана довільними способами, різноманітних синтаксичних конструкцій, які взаємодіють між собою, дана система має у своєму складі засоби для вираження існуючих зв'язків та встановлення характеру між ними. Під синтаксичною конструкцією будемо розуміти довільну синтаксичну побудову.

На наступному етапі вказують та припускають існування two (двійки) та three (трійки), які задовольняють умову two in $\tilde{n}n$, three in $\tilde{n}n$; прийнято вважати, що m є базовим компонентом в two і three. Потім потрібно конкретно визначити значення двійки (two') та трійки (three') та накласти умови існування: two' in $\tilde{n}n'$, three' in $\tilde{n}n'$; m' аналогічно до m міститься в в two' і three'. Далі семантико-синтаксичний критерій полягає у виконанні обчислень на основі бінарного критерію:

$$\begin{cases} 1, \textit{коли виконується} \ \text{two.type} = \text{two'.type} \\ 0, \textit{у протилежному випадку}. \end{cases}$$

На наступному кроці потрібно провести обчислення критеріїв натуральних чисел беручи до уваги кінцеву інформацію та відшукання семантичної відстані, яка присутня між словами по фундаментальних характеристиках WordNet [2].

Останній етап даного алгоритму полягає у обчисленні критерію дійсних чисел використання у вжитку антецедента у розумінні анафори на основі формування запитів до утвореної тензорної моделі за такими векторами:

1. Якщо two'.first = m',
то (m, two'.second)
інакше (two'.first, m)
2. Якщо three'.first= m'
то (m, three'.second, three'.third)
інакше якщо three'.two= m'
то (three.first, m, three'.third)
інакше (three'.first, three'.second, m).

Виходячи з вище наведеного алгоритму, відмітимо, що бінарний критерій за своєю структурою повністю нагадує критерій синтаксичного паралелізму, це можна пояснити тим що відбувається перевірка типів зв'язку на входження антецедента та анафори.

Використання натуральнозначних критеріїв дає змогу встановити синтаксично-семантичні паралельні значення та виконати перевірку на ідентичність за своїм змістом слів, які пребувають у одних і тих самих відношеннях синтаксичного характеру стосовно анафори та антецедента.

Використання дійснозначних критеріїв із залученням тензорної моделі дає змогу встановити можливість використання у даному реченні антецедента на місці анафори. Процедура заміни відбувається наступним чином: коли виконавши заміну анафори антецедента у двійці та трійці з'являються семантично необґрунтовані вислови, то утворену пару потрібно відкинути [7]. Заслуговує на увагу не менш важливий метод, який знаходить успішну реалізацію в системному аналізі, а саме тренувальне використання класифікатора на основі методу опорних векторів. Для практичного застосування вказаного методу потрібно визначити розмічену вибірку тренувального характеру, беручи до уваги вказану вибірку необхідно вказати позитивні та негативні задачі для тренувального навчання, щоб виконати поставлені вимоги треба обчислити перераховані раніше характеристичні ознаки всіх можливих пар антецедентів та анафор.

Наукові доробки показують, що число пар у негативних задачах буде наближатися до $k * l$, коефіцієнт k відповідає за чисельність анафоричних займенників у вибірці для тренування, а l показує чисельність кандидатів у антецеденти. У зв'язку з цим число негативних задач буде на порядок більшим від числа позитивних задач.

З метою прискорення роботи описаного алгоритму вчені прийшли до того, що потрібно зменшувати чисельність негативних задач, шляхом нехтування негативних пар антецедент-анафора, у досліджуваній вибірці тренувального характеру, вибрати для відкидання потрібно тільки ті пари у яких встановлено, що відстань між ними значно більша ніж встановлене стале число, яке вибране на основі інформації стосовно ступеня зв'язності початкової нашої вибірки для тренування. Аналітичне подання описаного методу має наступний вигляд:

$$D = \left\{ \left(x_i, y_i \right) \middle| x_i \in \Re^{33}, y_i \in \{-1,1\}_{i=1}^{n} \right\}, \quad (1)$$

x_i потрібно розуміти як і-ту характеристичну точку вибірки тренувального характеру у досліджуваному просторі ознак

$$y_i = \begin{cases} 1, \textit{точка являє собою сумісну пару} \\ -1, \textit{у протилежному випадку}, \end{cases}$$

k показує кількість двійок (пар) у досліджуваній вибірці.

Якщо реалізація вище наведеного алгоритму відбувається у лінійному випадку цього ж методу опорних векторів то дослідник матиме справу із гіперплощиною, яку можна задати наступним чином:

$$m \cdot x - b = 0,$$

де $\cdot -$ представляє скалярний добуток.

Більшість відомих алгоритмів системного аналізу, які представляють собою системи штучного аналізу не розвинені настільки, щоб безпомилково проаналізувати та зрозуміти природомовні тексти українською мовою. Побудований вище простір, які має вказані характеристики також відноситься до таких систем, у зв'язку з цим може трапитися так що не всі точки у просторовій реалізації будуть чітко вказувати на сумісність чи несумісність пар в тому чи іншому сенсі, в якому потрібно для конкретного дослідження. Це пояснюється тим, що новостворена поверхня ніяк не поділить простір на дві частини, де будуть зосереджені точки, які відносяться до свого класу у кожному із цих частин простору. При практичній реалізації інколи виникають труднощі у зв'язку із неправильним вибором роздільної поверхні.

Для детального дослідження характеристик отриманої поверхні потрібно її візуалізувати, адже геометрична інтерпретація дасть можливість використати виміри та ін. наявні результати.

Якщо все ж таки маємо справу із нероздільними класами поверхні, то потрібно брати до уваги деякі припущення тобто пом'якшувати допустимі норми роздільності, а це вже є модифікацією методу опорних векторів.

Тренувальне використання класифікатора на основі використання методу опорних векторів полягає у розв'язанні задачі, яка відноситься до оптимізаційних методів:

$$\begin{aligned} &\arg\min_{m,b,\xi}\left(\frac{1}{2}\|m\|^2 + C\sum_{i=1}^{n}\xi_i\right) \\ &y_i\cdot\left(m\cdot x_i - b\right)\geq 1-\xi_i,\ 1\leq i\leq n \\ &\xi_i\geq 0 \end{aligned} \tag{2}$$

де вважається $C-$ лінійний штрафний коефіцієнт. Наведену задачу можна розв'язати будь-яким методом оптимізації.

Стає очевидним, що першочерговим завданням сучасної комп'ютерної лінгвістики є правильне розпізнавання у текстових масивах згадок різноманітної природи: особи, події, явища та ін., а на наступному етапі потрібно встановити існуючі між ними зв'язки та описати характеристичні властивості. Властивості таких об'єктів залежать від розглядуваної предметної області. Отримані дані проходять необхідний формальний опис для збереження у загальній інформаційній базі даних. Якщо говорити про інформаційний об'єкт то він завжди відповідає певному відношенню до деякої предметної області із заданою структурою. Дослідники, як правило припускають, що аналіз та опрацювання тексту відбувається у межах деякої системи аналізу, яка має обмежену предметну область та представлена на мові формальних алгоритмів.

Висновки. У статі описана концепція сучасного розвитку системного аналізу на основі використання системних підходів при розв'язанні задач прикладного характеру, зокрема для розв'язку анафори природомовних текстів та встановлення міжпредметних зв'язків, з метою окреслення перспективних шляхів подальшого розвитку фундаментальних наук.

Досліджені основі критерії оцінювання, які побудовані на базі синтаксичної інформації, які здатні обробляти інформацію різної природи походження.

Встановлено, що в процесі проведення аналізу текстової інформації природною мовою, головною умовою є якісне виконання процедури ототожнення об'єктів, які неодноразово зустрічаються в тексті.

Розглянутий модифікований метод тензорного аналізу спрямований на проведення оцінки пари антецедент-анафора. Описаний процес вибору класифікатора на базі морфологічних ознак та синтаксично-семантичних властивостей. Для вибору антецеденту, відбувається морфологічний, синтаксичний фільтр для представників в кандидати.

Проведене дослідження основних етапів побудови класифікатора, який бере до уваги синтаксично-семантичну інформацію, яка представлена у вигляді тензорної моделі мови. Даний метод є дієвим при вирішенні задачі анафори завдяки тренуванню класифікатора.

Виходячи з вище сказаного, питання розв'язку анафори представляє інтерес для майбутніх досліджень, оскільки це дасть змогу автоматизувати роботу великої кількості вчених по виконанню аналізу текстових масивів.

Список використаних джерел:

1. Вознюк Т.Г. Алгоритм побудови керуючого простору синтаксичних структур природномовних текстів. Вісник Київського національного університету імені Тараса Шевченка. Фізико-математичні науки. 2014. № 1. С. 122- 127.
2. Марченко О.О. Методи оцінювання семантичної близькості–зв'язності слів природної мови. Штучний інтелект. 2012. №4. С. 213-219.
3. Павленко М.А. Анализ методов решения задачи извлечения информации из текстов. *Системи обробки інформації*. Харків. Харків, 2013. Вип. 8 (115). С.158-162.
4. Толпегин П.В. Автоматическое разрешение кореференции местоимений третьего лица русскоязычных текстов : дисс. ... канд. техн. наук / Вычисл. центр РАН. Москва, 2008. 238 с.
5. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. *International Conference on Analysis of Images, Social Networks and Texts*. Springer, Cham, 2015. С. 320-332.
6. Nivre J., Hall J., Nilsson J. et al. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*. 2007. Vol. 13. No. 2. P. 95–135.
7. Protopopova E.V., Bodrova A.A., Volskaya S.A., Krylo- va I.V., Chuchunkov A.S., Alexeeva S.V., Bocharov V.V., Granovsky D.V. Anaphoric Annotation and Corpus-Based Anaphora Resolution: An Experiment, Computational Linguistics and Intellec-tual Technologies. *Диалог-2014*: сб. тр. Междунар. науч. конф. по компьютер. лингвистике. 2014. Вып. 13. С. 562–571.

Стаття надійшла до редакції 07.03.2019