DOI https://doi.org/10.15407/usim.2019.05.048
УДК 004.83

С.Д. ПОГОРІЛИЙ, д-р. техн. наук, професор,
Київський національний університет імені Тараса Шевченка,
03022, Київ, просп. Академіка Глушкова, 4Г,
sdp@univ.net.ua

А.А. КРАМОВ, аспірант,
Київський національний університет імені Тараса Шевченка,
03022, Київ, просп. Академіка Глушкова, 4Г,
artemkramovphd@knu.ua

МЕТОД ВИЯВЛЕННЯ ІМЕННИХ ГРУП В УКРАЇНОМОВНИХ ТЕКСТАХ

Здійснено порівняльний аналіз основних автоматизованих методів пошуку іменних груп та іменованих сутностей в англомовних та україномовних текстах; обґрунтовано доцільність використання моделі Universal Dependencies. Запропоновано комплексний метод на основі аналізу деревовидної синтаксичної структури речення та моделі виявлення іменованих сутностей. Здійснено експериментальну перевірку ефективності пропонованого методу та показано доцільність його використання для пошуку іменних груп в україномовних текстах.

Ключові слова: *обробка природної мови, іменна група, модель Universal Dependencies, модель NER, деревовидна структура речення.*

Вступ

Постійна динаміка росту потужностей обчислювальних систем зумовлює використання методів машинного навчання для формалізації та розв'язання задач, подібних до дій людини. Задачі такого типу, що не можуть бути розв'язані за допомогою алгоритмічних дій, називають *AI-повними*. Зважаючи на постійне зростання обсягу текстової інформації, актуальною проблемою є автоматизований аналіз природної мови для отримання структурованих даних: розпізнавання мовлення, машинний переклад, подолання лексичних неоднозначностей тощо. Зазначені задачі варто зараховувати до завдань комп'ютерної лінгвістики та методології машинного навчання, а саме, до галузі обробки природної мови (*Natural language processing — NLP*). Попри відмінність поставлених цілей, задачі *NLP* містять спільний початковий етап, а саме, *попередню обробку вхідних даних* (текстової інформації). Попередня обробка тексту необхідна для формального представлення текстової інформації у вигляді структурованих даних. Засоби формалізації тексту можуть відрізнятися відповідно до поставленої задачі, однак варто виокремити такі кроки попередньої обробки тексту, які використовуються в більшості задач обробки природної мови:

- токенізація (*tokenization*) — процес розбиття тексту на речення, а речення на окремі слова;
- розмічання слів — зіставлення кожній атомарній одиниці тексту (слову) частини мови, роду, відмінку;
- лематизація — приведення слова до нормальної форми; наприклад, для української

мови нормальною формою іменників є його представлення в називному відмінку, а дієслово трансформується до інфінітивної форми;

- пошук сутностей (іменних груп) у тексті.

На відміну від попередніх кроків, які здійснюються через використання заздалегідь визначених правил і різнотипних словників, останній крок потребує детальнішого аналізу. Зважаючи на постійну зміну лексичного складу мови, виявлення сутностей потребує спільного використання методології машинного навчання та засобів комп'ютерної лінгвістики. Таким чином, пошук іменних груп у тексті варто зараховувати до класу *AI*-повних задач, що не можуть бути формалізовані визначеним алгоритмом.

Отже, завдання пошуку іменних груп є важливим етапом у процесі розв'язання інших задач обробки природної мови. Підвищення точності детектування іменних груп у тексті уможливить покращення ефективності застосування методів розв'язання задач, залежних від цього пошуку. Наявність актуальних праць щодо визначення іменних груп у різних мовах свідчить про важливість дослідження методів розв'язання цієї задачі. Попри активний розвиток досліджень у напрямку обробки природних мов, дослідження пошуку іменних груп для україномовних текстів перебуває на початковому етапі.

Мета роботи

- аналіз наявних методів автоматизованого пошуку іменних груп в англомовних та україномовних текстах;
- створення комплексного методу детектування іменних груп на основі дерева залежностей речення та моделі розпізнавання іменованих сутностей;
- здійснення експериментальної перевірки зазначеного методу для корпусу текстів української мови.

Концепт іменної групи

Термін *«іменна група»* запозичено з англомовного варіанту *noun phrase*. В українській мові цей термін трактується як іменникове (субстантивне) словосполучення — словосполучення з іменником у ролі головного слова [1]. Однак особовий займенник (я, ти, він), який вказує на конкретний об'єкт, також може використовуватися як окрема сутність у реченні, тому надалі будемо розглядати термін «іменна група» як іменникове чи займенникове словосполучення.

Розгляньмо детальніше варіанти формування іменної групи в українській мові.

Іменник у ролі головного слова може сполучатися:

- з прикметником (*червоний колір, смачний обід*);
- з іменником у непрямих відмінках з прийменником або без нього (*брат Петра, думки про майбутнє*);
- з займенником (*ця думка, моя мрія*);
- з дієприкметником (*зів'ялі квіти, пожовкла трава*);
- з прислівником (*читання вголос*);
- з числівником (*два кольори*);
- з інфінітивом (*бажання вчитися*).

Займенник у ролі головного слова може сполучатися:

- з іменником (*хтось зі звірів, когось із тварин*);
- з прикметником (*щось цікаве*);
- з займенником (*кожного з нас*).

Попри наявність зазначених правил, процес пошуку іменних груп не є тривіальним для української мови. Для мов, у яких існує клас артиклів (наприклад, англійська), індикатором іменної групи є *детермінатив* — словоформа чи морфема, яка супроводжує іменну групу та узагальнює інформацію про групу (рід, число тощо). Наприклад, у синтаксичних теоріях англійської мови вважається, що будь-яка іменна група містить детермінатив [2]. Українська мова належить до класу мов без артиклів. Наразі немає однозначної відповіді щодо наявності в цьому класі детермінативу в іменних групах. Питання пошуку іменної групи в ієрархічній структурі для мов без артиклів розглянуто в праці [3] на прикладі російської мови. У зазначеній праці розглядаються загальні принципи узгодження головного слова

іменної групи із залежними словами. Зокрема, доводиться *ієрархічна побудова іменної групи в російській мові* та аналізується узгодження слів іменної групи за числом і родом.

Алгоритм детектування головного слова та дочірніх слів має враховувати принаймні наступні особливості текстової інформації:

- відсутність артиклів в українській мові, які певною мірою ідентифікують іменні групи в романо-германських та деяких інших мовах;
- неструктурована будова речення (можливий зворотній порядок слів);
- наявність фразеологізмів, власних назв та слів іншомовного походження.

Порівняльний аналіз наявних методів пошуку іменних груп

Проблема пошуку іменних груп у тексті активно розв'язується для англомовних текстів, про що свідчить наявність праць [4–6]. Метод n-грам [4] полягає в пошуку всіх послідовностей слів, які зустрічаються в тексті, довжиною k ($1 \le k \le n$); послідовність має перебувати в межах одного речення. Такий підхід ефективно використовується для отримання ознак у задачі класифікації текстів, але з погляду семантичного значення групи метод n-грам має значний недолік: фіксований розмір послідовностей. Фіксований розмір групи призводить до втрати смислового навантаження набору слів, які входять до послідовності. Наприклад, іменна група «*Міністерство освіти і науки України*» може інтерпретуватися як послідовність «*Міністерство освіти і*», яка не відображає її семантичний зміст для користувача. Принцип роботи методу *NPFST* [5] полягає у використанні заздалегідь описаних шаблонів іменних груп. Шаблони представлено у вигляді рядків — регулярних виразів, у яких змінні елементи відображають різні частини мови. Далі наведено приклад такого регулярного виразу:

$$(A|N) * N(PD * (A|N) * N)* \qquad (1)$$

Після здійснення операцій токенізації та розмічення слів, кожному слову ставиться у відповідність текстова мітка частини мови (*A, N, P* тощо). Кожне слово замінюється на потрібну мітку, тобто речення тексту трансформуються в рядки, що містять не слова, а мітки. Далі до отриманого тексту застосовується набір шаблонів. У разі детектування відповідності шаблону частині тексту виконується екстракція знайденої частини з подальшим зворотнім перетворенням від мітки до слова. Недоліком такого підходу є залежність набору шаблонів від особливостей мови та стилістики тексту. Крім того, цей метод не є масштабованим, адже постійне збільшення кількості шаблонів підвищує ймовірність колізії регулярних виразів, що призведе до некоректної роботи методу.

У 2016 р. було запропоновано універсальний підхід (*Universal Dependencies — UD*) до перетворення текстової інформації на деревовидну структуру [6]. Універсальність підходу передбачає узагальнення різнотипних зв'язків між словами речення незалежно від мови.

У праці запропоновано загальну схему впорядкування слів речення залежно від частини мови, якою вони є: іменник і прикметник, іменник і займенник тощо. Внаслідок такої уніфікації формату для різних мов та за допомогою зусиль відкритої спільноти з різних країн вдалося створити набір моделей перетворення вхідного тексту на деревовидну структуру. Поточна версія *UD* 2.3 містить підтримку 76 мов. Для україномовних текстів також було підготовлено вхідні дані (розмічені тексти) і згодом було навчено відповідну модель [7]. Приклад такої структури наведено на рис. 1.

Іменник чи займенник, що є вершиною дерева та містить дочірні вузли, можна трактувати як потенційне головне слово своєї групи. Обхід дерева дає змогу поставити у відповідність до кожного потенційно головного слова іменної групи набір залежних слів, причому таке зіставлення може відбуватися і для глибших рівнів у рекурсивний спосіб.

Враховуючи належність української та російської мов до спільного класу мов без артиклів та ієрархічну структуру іменних груп у російськомовних текстах, доцільним є здійснення пошуку іменних груп в україномовних текстах за допомогою аналізу моделі *UD*.

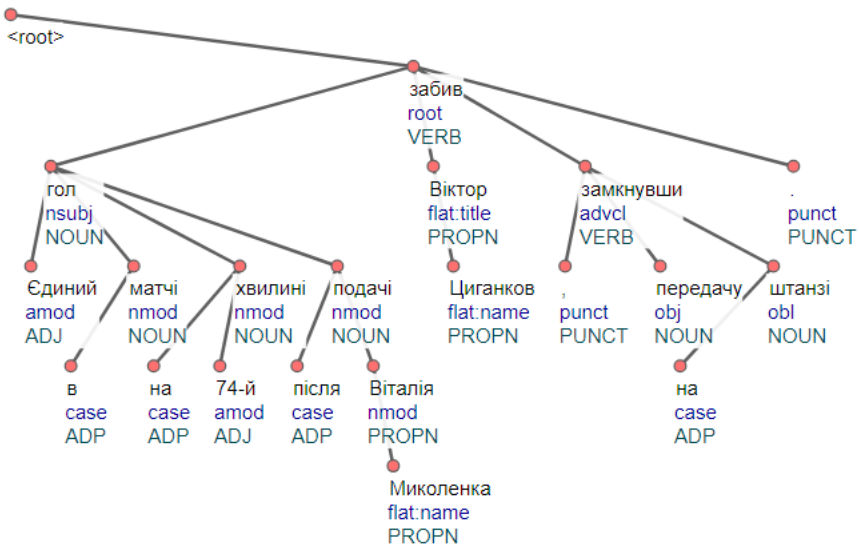

Рис. 1. Приклад представлення тексту в деревовидній структурі

Методи розпізнавання іменованих сутностей у тексті

Окремо варто розглянути питання виокремлення іменованих сутностей у тексті. Принцип встановлення порядку та узгодження слів в іменованій сутності може трохи відрізнятися від результату аналізу деревовидної структури речення. Така відмінність може виникнути через унікальну структуру іменованої сутності, що не підпорядковується загальним правилам побудови іменної групи, та некоректного перетворення вхідного тексту на деревовидну структуру. Розгляньмо для прикладу наступне речення: «**Група акціонерів компанії Facebook (1)** наполягає на тому, що **засновник соціальної мережі Марк Цукерберг (2)** повинен втратити **посаду голови правління (3)**». Напівжирним шрифтом виділено іменні групи речення, а в дужках вказано порядковий номер групи. Результат перетворення речення на деревовидну структуру зображено на рис. 2.

Групи (1) і (3) можуть бути ідентифіковані коректно, адже їхні елементи розташовано в рекурсивний спосіб відповідно до очікуваної структури цих груп. Розгляньмо групу (2). Батьківським словом відповідної групи в дереві є слово «засновник», яке помічено як іменник, тобто воно може бути головним словом групи. Виконавши обхід дочірніх елементів у рекурсивний спосіб, отримуємо наступну послідовність: «засновник соціальної мережі Марк». Порівнюючи з очікуваним результатом, відсутнє слово «Цукерберг».

Розглянувши детальніше відповідну область дерева, можна побачити, що слово «Цукерберг» не потрапляє до списку дочірніх елементів слова «засновник»; крім того, частина мови цього слова ідентифікована як дієслово.

Некоректне розмічення слів частинами мови може траплятися через відсутність слова в морфологічному словнику; найчастіше така ситуація може виникати для власних назв. Фрагменти тексту з власними назвами можливо знайти за допомогою додаткового використання моделі виокремлення іменованих сутностей (*named-entity recognition — NER*). У наведеному прикладі модель виокремлення іменованих сутностей може ідентифікувати пару слів «Марк Цукерберг» як особу. Подальше об'єднання множин «засновник соціальної мережі Марк» і «Марк Цукерберг», які мають спільні елементи (слово «Марк»), приводить до отримання очікуваного результату: «засновник соціальної мережі Марк Цукерберг». Отже, використання моделі виокремлен-

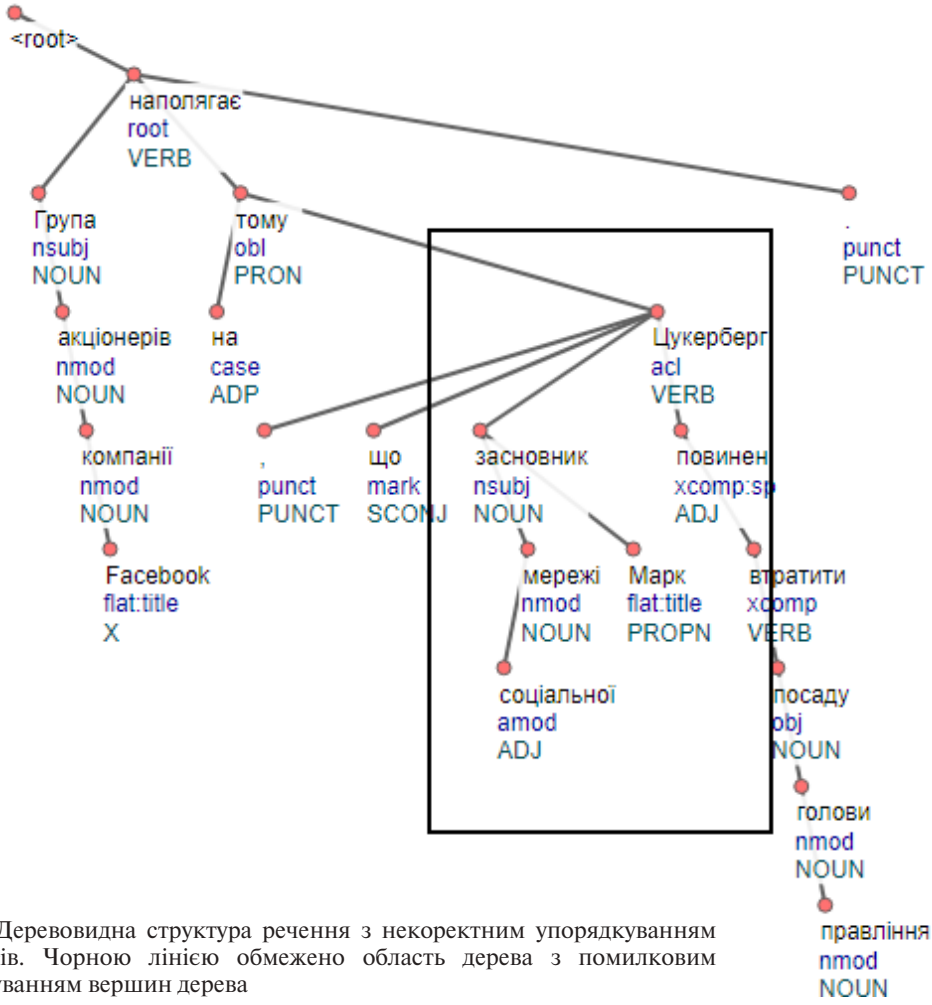

Рис. 2. Деревовидна структура речення з некоректним упорядкуванням елементів. Чорною лінією обмежено область дерева з помилковим розташуванням вершин дерева

ня іменованих сутностей може використовуватися як додатковий інструмент під час пошуку іменних груп у тексті. Ефективність застосування цієї моделі залежить від типу об'єктів, які вона здатна розпізнавати, та предметної області аналізованого тексту.

Розгляньмо наявні розв'язки та пропозиції щодо виокремлення іменованих сутностей в україномовних текстах. Відкритою спільнотою фахівців *lang-uk* було здійснено навчання моделі *NER* на проанотованому корпусі української мови [8]. Для пошуку іменованих сутностей використовувалася відкрита бібліотека *MITIE*, яка має інтерфейси для багатьох мов програмування: *C/C++*, *Python*, *Java*. Тренування моделі здійснювалося на вибірці з 229 текстів для розпізнавання сутностей, які належать до таких категорій:

персона; локація; організація; різне.

Також варто виокремити працю [9], в якій пропонується використовувати підхід пошуку сутностей за шаблонами. Для кожного типу сутності створюється окремий набір правил, який дає змогу однозначно ідентифікувати цей тип. Алгоритм виокремлення сутностей використовує *GLR*-парсер. У цій праці виокремлення іменованих сутностей здійснювалося для таких категорій: персона; організація; географічний об'єкт.

Для порівняння ефективності розглянутих методів доцільно розглянути їхні значення *F*-міри [10]. *F*-міра (*F*) — це середнє гармо-

нійне значення точності (*Precision*) і повноти (*Recall*):

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

$$F = 2\frac{Precision \times Recall}{Precision + Recall}, \quad (4)$$

де TP — кількість коректно розпізнаних сутностей; FP — кількість сутностей, які не було розпізнано; FN — кількість сутностей, які було ідентифіковано моделлю, але їх немає в експертній розмітці тексту. Значення *F*-міри досягає значення 0,8 для моделі, створеної спільнотою *lang-uk*; 0,54 — для моделі на основні пошуку шаблонів. Попри зазначений показник *F*-міри, модель на основі пошуку шаблонів може використовуватися як додатковий інструмент пошуку іменованих сутностей у текстах певної предметної області.

Метод пошуку іменних груп в україномовних текстах

Пошук іменних груп в україномовних текстах пропонується здійснювати через аналіз деревовидної структури речення, отриманої за допомогою підходу *Universal Dependencies* (*UD*) [11].Враховуючи, що головним словом іменної групи може бути іменник чи займенник, спочатку розгляньмо вершини з відповідними частинами мови. Відповідно до категорій розмітки тексту частинами мови, до потенційних головних слів іменної групи варто зарахувати слова з наступними категоріями:

- *NOUN* (іменник);
- *PRON* (займенник);
- *PROPN* (власна назва);
- *X* (інша частина мови).

З наведеного списку варто виокремити два пункти: *PROPN* і *X*. Власну назву (ім'я, прізвище, місто тощо) може бути розмічено як категорію *PROPN*, яка теж здатна формувати іменну групу або входити до складу наявної. Категорія *X* встановлюється для слова тоді, коли модель не може передбачити частину мови. Однак слова з такою категорією можуть мати додатковий параметр *Foreign = Yes*, який вказує, що це слово має іншомовне походження. Проаналізувавши 2500 різних текстів, написаних українською мовою, було виявлено, що 99 відсотків слів іншомовного походження є сутностями, які вказують на певний об'єкт. Таким чином, доцільно додатково розглядати слова з тегом *X* і додатковим параметром *Foreign =Yes* як потенційне головне слово групи чи складову частину іншої групи.

Визначивши тип вершин, які можуть розглядатися як головне слово іменної групи, визначмо правила приєднання дочірніх слів до іменної групи батьківського слова. Відповідно до деревовидної структури речення можна зробити припущення, що всі дочірні вершини головного слова логічно пов'язані з ним та входять до його іменної групи. Однак таке припущення є хибним, враховуючи наступні фактори:

- похибка попередньої обробки тексту, а саме процесу токенізації та розмічення слів частинами мови;
- похибка власне моделі побудови деревовидної структури;
- граматичні та пунктуаційні помилки у вхідному тексті.

Отже, потрібно сформувати набір правил приєднання дочірніх вершин до батьківської іменної групи. Розгляньмо окремо принципи входження потенційних головних слів, дієслів та інших дочірніх елементів до поточної батьківської групи.

Загальний підхід приєднання дочірнього елемента до іменної групи

Враховуючи правила формування іменних груп (субстантивних словосполучень) в українській мови, до складу іменної групи можуть входити слова з наступними частинами мови (в дужках вказуються відповідні теги моделі *UD*): прикметник (*ADJ*), прислівник (*ADV*), прийменник (*ADP, DET, AUX*), числівник (*NUM*), іменник (*NOUN, PROPN, X*), займенник (*PRON*), дієслово (*VERB*), знаки

пунктуації (*PUNCT*). Слова, розмічені як інші частини мови, чи додаткові символи (знаки арифметичних операцій, сполучники тощо) не включаються до іменної групи. Крім того, всі елементи групи мають розташовуватися в тексті послідовно. Якщо між дочірнім елементом, який може входити до групи, міститься заборонений елемент, входження зазначеного дочірнього елементу до групи відхиляється. Додатковою перевіркою умови приєднання дочірнього елемента до групи може бути уточнення його узгодженості з головним словом за числом і родом. Наприклад, у такий спосіб узгоджено іменні групи «дві медалі» (за числом) і «кваліфікований фахівець» (за родом). В українській мові є різнотипні варіанти такого узгодження, пов'язані з граматичною складовою мови.

Для прикладу розгляньмо словосполучення: «п'ятдесят один кілометр». Головне слово «кілометр» має число однини, хоча в цьому контексті складений числівник «п'ятдесят один» вказує на число множини. Використовується таке правило: після числівника «один», навіть якщо він входить до складених числівників, іменник вживається у формі однини. Застосування набору правил узгодження головного слова з підрядними дає змогу перевірити можливість їх приєднання до групи.

Однак необхідною умовою використання згаданого набору правил є врахування всіх аспектів формування словосполучення, що для веб-ресурсів більшості сучасних ЗМІ є малоймовірним. Помилкові вирази «два з половиною місяця», «заступниця Міністра культури» можуть зустрічатися в новинних текстах чи розмовній мові. Таким чином, було вирішено не здійснювати перевірку узгодження слів іменної групи за числом і родом для коректної обробки текстів з різною стилістикою.

Приєднання дієслова до іменної групи

Іменна група з іменником у ролі головного слова може містити дієслово у формі інфінітиву (наприклад, «бажання вчитися» чи «необхідність працювати»). Отже, необхідно уточнити, чи дочірнє слово (дієслово) має форму інфінітиву. Для цього можна скористатися додатковим параметром моделі *UD Verb Form*. Параметр *Verb Form* наявний лише для дієслів; у разі представлення дієслова у формі інфінітиву, атрибут набуває значення *Inf*.

Приєднання потенційного головного слова до іменної групи

Найскладнішим є рішення щодо входження дочірнього потенційного головного слова (ДПГС) до батьківської іменної групи, адже дочірній елемент може формувати окрему групу. Проаналізувавши деревовидні структури україномовних текстів та відповідні синтаксичні зв'язки моделі *UD* [12], було вирішено сформувати наступні критерії входження ДПГС до іменної групи:

- наявність відповідного типу синтаксичного зв'язку;
- відсутність заборонених елементів серед дочірніх вершин ДПГС.

Розгляньмо типи синтаксичного зв'язку між ДПГС та батьківською вершиною, необхідні для входження ДПГС до іменної групи. До таких типів належать:

- *flat* — встановлюється між словами, які входять до складу власних назв чи дат, тобто в тому разі, коли невідома внутрішня синтаксична структура виразу;
- *nmod* — зв'язок між елементами, один із яких модифікує інший; зазвичай, такий зв'язок передбачає представлення дочірнього елемента в родовому відмінку.

Наведені типи зв'язку можна відслідкувати у реченні: «Під час позачергових парламентських виборів 2014 р. майбутній міністр Лілія Гриневич потрапила до парламенту». На рис. 3 зображено деревовидну структуру цього речення.

У наведеному прикладі зв'язок *flat* дає змогу з'єднати власні іменники «Лілія» і «Гриневич» зі словом «міністр», у такий спосіб утворюючи іменну групу «майбутній міністр Лілія Гриневич». Зв'язок *nmod* простежується

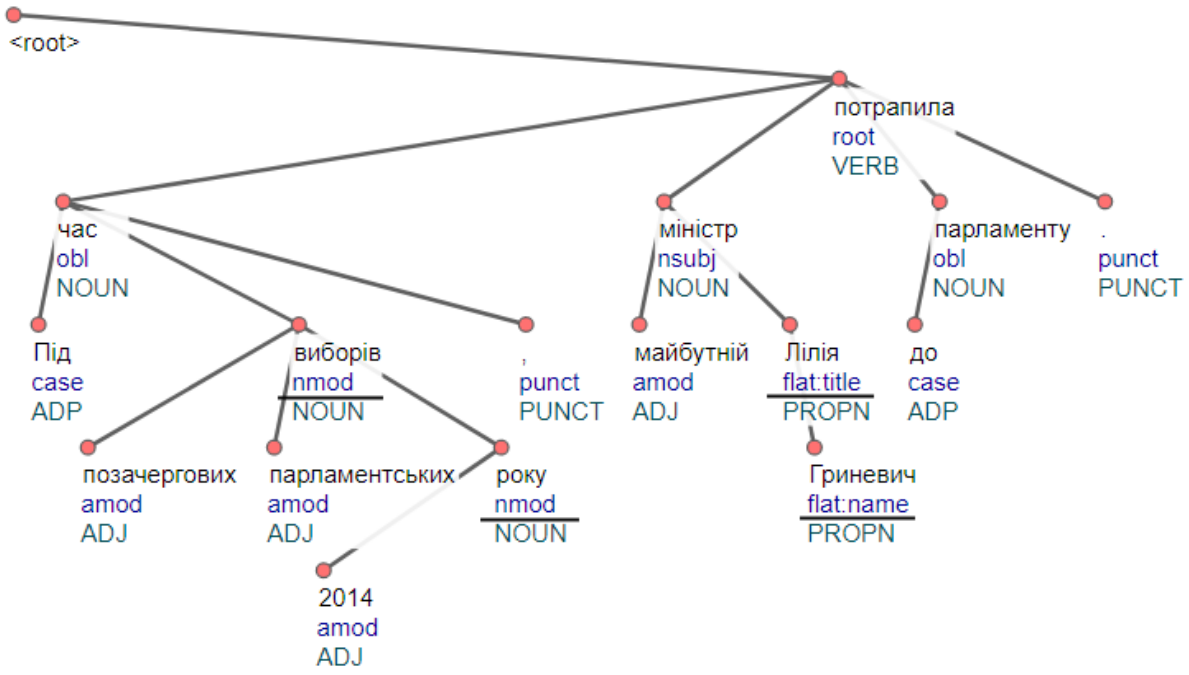

Рис. 3. Деревовидна структура речення, що містить іменні групи, утворені зв'язками *flati nmod*

в першій частині реченням між словами «час» і «виборів»,«виборів» і «року». Поєднуючи всі відповідні слова у рекурсивний спосіб відповідно до їхнього порядку розташування в тексті, отримуємо іменну групу «під час позачергових парламентських виборів 2016 р.».

Варто згадати ще два типи зв'язків, які разом зі *flat* належать до типу *MWE* (*multiword expressions* — багатослівні вирази): *fixed* і *compound*. Зв'язок *fixed* вказує на стійке словосполучення: «до того ж», «мало не сто років» тощо. Щодо *compound,* то цей тип зв'язку зазвичай використовується для композицій із числами. Вказані зв'язки можуть використовуватися для формування різних структурних одиниць в тексті, але застосування їх не є доцільним для відстеження зв'язків в іменних групах.

Обхід деревовидної структури речення

Розглянувши критерії відбору потенційного головного слова групи та правила приєднання дочірніх елементів до батьківської групи, варто звернути увагу на порядок обходу деревовидної структури. Зрозуміло, що обхід структури такого типу здійснюється у рекурсивний спосіб (використовується центрований порядок). Необхідно зазначити, що елементи іменної групи розташовуються в тексті послідовно, тобто між цими елементами немає сторонніх слів, які не належать до групи. Таким чином, у разі перевірки входження дочірнього елемента до батьківської групи потрібно здійснювати додатковий аналіз того, чи входять до групи елементи, які розташовані в реченні між поточним дочірнім елементом та батьківською вершиною. Для уникнення зазначених перевірок пропонується здійснювати обхід дочірніх вершин у наступний спосіб:

- від найближчого дочірнього елемента, розташованого ліворуч від головного слова в тексті, до крайнього лівого дочірнього елемента;
- від найближчого дочірнього елемента, розташованого праворуч від головного слова в тексті, до крайнього правого дочірнього елемента.

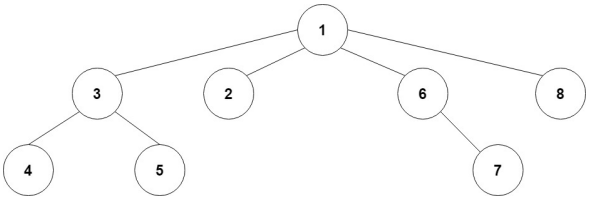

Рис. 4. Приклад порядку обходу деревовидної структури речення

Такий порядок обходу дочірніх вершин уможливлює уникнення зазначеної додаткової перевірки, адже у разі виявлення несумісності дочірнього елемента з батьківською вершиною всі наступні вершини можуть розглядатися як об'єкти, незалежні від головного слова. Приклад порядку обходу деревовидної структури речення зображено на рис. 4.

Пошук іменованих сутностей у тексті

Додатково до пропонованого пошуку іменних груп е тексті варто застосовувати виокремлення іменованих сутностей. Як було зазначено раніше, виокремлення іменованих сутностей дає змогу виявити сполучення слів, які неможливо ідентифікувати за допомогою аналізу отриманої деревовидної структури (через некоректне розмічення слів чи похибки роботи моделі синтаксичного розбору речення). Пошук іменованих сутностей варто розпочинати із застосування *газетирів* — словників, які містять перелік географічних назв із додатковою інформацією про них. Із погляду автоматизованої обробки тексту під терміном «газетир» зазвичай розглядається список власних назв відповідно до предметної області дослідження. Результатом застосування газетирів до вхідного тексту є набір груп — іменованих сутностей, кожна з яких містить індекси-вказівники на певні слова тексту. Для формування газетиру було використано такі бази даних:

- перелік найпопулярніших прізвищ та імен жінок і чоловіків (форма «прізвище, ім'я»);
- перелік країн;
- перелік міст.

Після отримання результату застосування газетиру до тексту наступним кроком є запуск навченої моделі виокремлення іменованих сутностей. Як модель виокремлення іменованих сутностей було обрано *NER*-модель спільноти *lang-uk*; для застосування моделі було використано відкриту бібліотеку *MITIE*. Вихідним результатом роботи моделі є набір об'єктів, кожен із яких відповідає розпізнаній іменованій сутності та має наступні атрибути:

- діапазон індексів слів, які входять до іменованої сутності;
- категорія іменованої сутності;
- оцінка «впевненості» моделі в тому, що поточну іменовану сутність розпізнано коректно.

Варто звернути увагу на останній атрибут. Відповідно до документації бібліотеки чим більшим є значення оцінки «впевненості», тим вищою є ймовірність коректного передбачення. Зважаючи на відсутність еталонного порогового значення зазначеної оцінки, було вирішено встановити це значення експериментальним шляхом за допомогою розрахунку *F*-міри моделі на множині україномовних текстів. Відповідно до розміченої тестової вибірки текстів отримане оптимальне порогове значення оцінки «впевненості» моделі дорівнює 0,8. Вказане значення оцінки моделі було використано в подальших експериментальних перевірках цієї праці.

Експериментальна перевірка методу

Для здійснення експериментальної перевірки ефективності пропонованого методу було створено відповідне застосування; серверна мова програмування — *Python* 3.6. Відповідно до послідовних етапів здійснення перевірки роботи методу, застосування складається з трьох компонентів:

- веб-сторінка розмітки іменних груп у тексті;
- модуль пошуку іменних груп в україномовних текстах;
- утиліта розрахунку оцінки ефективності роботи методу.

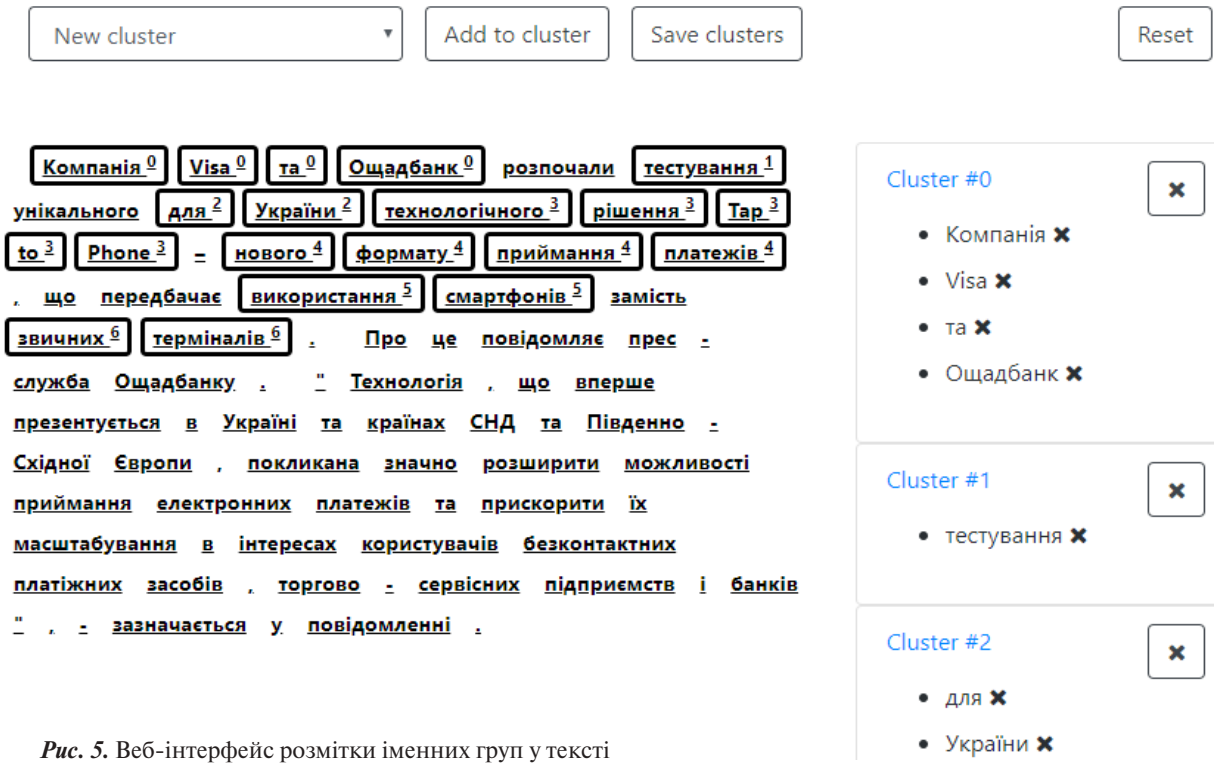

Рис. 5. Веб-інтерфейс розмітки іменних груп у тексті

Веб-сторінка розмітки іменних груп

Оцінка *F*-міри методу можлива за наявності попередньо розміченого тексту — комбінацій слів і символів, які експерт позначив як іменні групи. Перевірочну вибірку текстів було сформовано зі статей новинних порталів різної тематики. Протягом дослідження було оброблено 100 різних документів; загальна кількість знайдених іменних груп — 1488. Для формування перевірочної вибірки було створено веб-сторінку, яка здійснює графічне відображення результату токенізації вхідного україномовного тексту та дає змогу виконувати групування слів та символів в іменні групи. Приклад обробки вхідного тексту за допомогою зазначеної веб-сторінки зображено на рис. 5.

Розпізнані об'єкти підкреслено лінією, іменні групи, позначені користувачем, додатково виділено суцільної рамкою; індекс у правому верхньому куті об'єкта вказує на номер іменної групи, до якої він входить. Принцип роботи веб-сторінки є таким: користувач копіює україномовний текст із зовнішнього джерела та вставляє його в текстове поле сторінки. Далі він натискає на кнопку «*Recognize*», після чого виконується токенізація вхідного тексту. Користувач вибирає об'єкти, які належать до спільної іменної групи, та формує відповідний кластер. Після закінчення процесу розмітки тексту користувач натискає кнопку «*Saveclusters*», зберігаючи створені кластери в базі даних. Для створення інтерактивного режиму формування іменних груп використано фреймворк *Marionette.js*. Збереження сформованих кластерів для подальшої оцінки ефективності роботи методу здійснено за допомогою реляційної бази даних *MySQL*.

Модуль пошуку іменних груп

Модуль пошуку іменних груп в україномовних текстах реалізовано мовою програмування *Python*. Створений модуль розміщено на платформі *The Python Package Index* (*PyPI*), що

дає змогу виконувати імпорт модуля у сторонні проекти. Інструкції щодо встановлення та використання модуля доступні за посиланням [13]. Модуль містить сторонні пакети, які необхідно попередньо встановити для коректної роботи модуля. Передбачено додаткове підключення моделі пошуку іменованих сутностей в україномовному тексті, а також використання сторонніх газетирів.

Результати оцінки ефективності роботи методу

Метрикою оцінки ефективності роботи методу було обрано три параметри: точність, повнота та *F*-міра. Здійснення розрахунку параметрів виконано для двох режимів: повної та часткової відповідностей. Для детектування повної відповідності необхідно, щоб прогнозована та розмічена іменні групи вповні збігалися (порівняння за символами та позиціями в тексті); іменні групи вважаються частково відповідними одна одній, якщо хоча б одна межа груп збігається (початкове чи кінцеве слово). Крім того, було вирішено розрахувати метрику для трьох різних варіантів використання моделей:

- з моделями аналізу деревовидної структури речення і пошуку іменованих сутностей (*UD+NER*);
- з моделлю аналізу деревовидної структури речення, але без використання моделі пошуку іменованих сутностей (*UD*);
- без використання моделей, зазначених в попередніх пунктах; у цьому разі окремими іменними групами вважаються іменники та особові займенники (–).

У табл. 1 наведено оцінки ефективності роботи різних варіантів використання моделей відповідно до розрахованої метрики для режиму повної відповідності. Значення всіх метрик варіантів *UD* і *UD+NER* відрізняються в межах 0,01, що вказує на низьку ефективність додаткового застосування поточної моделі пошуку іменованих сутностей. Значення *F*1-міри для варіанту без використання моделей *UD* і *NER* є меншим від 0,1, тобто представлення іменних груп як окремих іменників і особових займенників є малоефективним та недоцільним у задачах, що потребують попереднього виявлення іменних груп у тексті.

У табл. 2 наведено оцінки ефективності роботи розглянутих варіантів застосування моделей для режиму часткової відповідності. Аналогічно до режиму повної відповідності значення метрик варіантів моделей *UD* і *UD+NER* є рівними в межах похибки 0,001, що підкреслює низьку ефективність застосування моделі пошуку іменованих сутностей. Значення *F*1-міри для обох варіантів дорівнює 0,902, що вказує на доцільність використання варіанта моделі *UD* для знаходження іменних груп в україномовних текстах.

Висновки

Проаналізовано головні методи пошуку іменних груп та іменованих сутностей для англомовних та україномовних текстів. Методи аналізу англомовного тексту не можуть

Табл. 1. Оцінка ефективності роботи методу з різними варіантами моделей для режиму повної відповідності

Метрика / Варіанти моделей	Точність	Повнота	*F*-міра
UD+NER	0,552	0,573	0,559
UD	0,555	0,572	0,560
–	0,175	0,039	0,062

Табл. 2. Оцінка ефективності роботи методу з різними варіантами моделей для режиму часткової відповідності

Метрика / Варіанти моделей	Точність	Повнота	*F*-міра
UD+NER	0,973	0,844	0,902
UD	0,974	0,843	0.902
–	0,948	0,201	0,320

бути використані для україномовних документів, адже їх створено з урахуванням особливостей структури побудови речень лише в англомовних текстах. Для виявлення іменованих сутностей в україномовних текстах доцільно використовувати попередньо навчену модель відповідно до предметної області вхідних текстів; можливим є використання додаткових регулярних виразів для екстракції іменованих сутностей із фіксованою структурою. Проаналізовано результати застосування технології *Universal Dependencies* для україномовних текстів з метою здійснення перетворення вхідної текстової інформації на деревовидну структуру. На основі аналізу деревовидної структури запропоновано метод пошуку іменованих груп в україномовних документах. Отримано експериментальні результати застосування пропонованого методу з різними варіаціями його використанням: окремо та разом із моделлю пошуку іменованих сутностей у тексті. Розраховані метрики ефективності роботи методу вказують на доцільність його використання для пошуку іменованих груп в україномовних текстах. Для підвищення точності роботи методу можуть бути застосовані наступні підходи:

- використання навченої моделі пошуку іменованих сутностей та набору газетирів відповідно до предметної області;
- застосування набору регулярних виразів для виявлення іменних груп із фіксованою структурою;
- використання сторонніх моделей токенізації тексту для зменшення похибки виявлення частин мови слів тексту.

ЛІТЕРАТУРА

1. *Шкуратяна Н. Г.* Сучасна українська літературна мова: Модульний курс / Н. Г. Шкуратяна, С. В. Шевчук. Київ: Арій, 2010. 824 с.

2. *Bo kovi* . Whatwillyouhave, DP or NP. Proceedingsof NELS.P. 101–114.

3. *Лютикова Е.А.* Согласование, признаки и структура именной группы в русском языке. Русскийязык в научном освещении. 2015. № 2 (30). С. 44–74.

4. *Su Nam K., Baldwin T., Kan M.* Evaluating N-gram based evaluation metrics for automatic keyphrase extraction. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2010. P. 572–580.

5. *Handler A., Denny M., Wallach H., O'Connor B.* Bag of what? Simple noun phrase extraction for text analysis. Proceedings of the First Workshop on NLP and Computational Social Science. 2016. P. 114–124.

6. *Nivre J., de Marneffe M., Ginter F., Goldberg Y., Haji J., Manning C., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D.* Universal Dependencies v1: A Multilingual Treebank Collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). P. 1659–1666.

7. UniversalDependencies/UD_Ukrainian-IU. URL: https://github.com/UniversalDependencies/UD_Ukrainian-IU (дата звернення: 18.10.2019).

8. Models: lang-uk. URL: http://lang.org.ua/en/models (дата звернення: 18.10.2019).

9. *Глибовець А.М.* Автоматизований пошук іменованих сутностей у нерозмічених текстах українською мовою. Штучний інтелект. №2. С. 45–52.

10. *Powers D. M.* Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies. 2011. Vol. 2, No 1. P. 37–63.

11. Universal Dependencies. URL: https://universaldependencies.org/guidelines.html (дата звернення: 18.10.2019).

12. Лабораторія української. URL: https://mova.institute/%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D0%B8%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82 (дата звернення: 18.10.2019).

13. Pythonpackagetoextract NP fromtheUkrainianlanguage. URL: https://github.com/artemkramov/np-extractor-ua (дата звернення: 18.10.2019).

Надійшла 29.10.2019

REFERENCES

1. *Shkuratjana, N. and Shevchuk, S.* (2010).Modern Ukrainian literary language. Modular course. [Suchasnaukrayins`ka literaturnamova. Modul`ny`jkurs].Kyiv: Arij,p.824.

2. *Bo kovi* , . (2008). Whatwillyouhave, DP or NP. In: Proceedingsof NELS. pp.101–114.

3. *Lyutikova, E.* (2015). Coordination, features and structure of the nounphrasein Russian [Soglasovanie, priznaki I struktura imennoy gruppyi v russkom yazyike]. Russkiyyazyik v nauchnomosveschenii, 2(30), pp.44–74.

4. *Su Nam, K., Baldwin, T. and Kan, M.* (2010). Evaluating N-gram based evaluation metrics for automatic keyphrase extraction. In: Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics, pp.572–580.

5. *Handler, A., Denny, M., Wallach, H. and O'Connor, B.* (2016). Bag of what? Simple noun phrase extraction for text analysis. In: Proceedings of the First Workshop on NLP and Computational Social Science. pp.114–124.

6. *Nivre, J., de Marneffe, M., Ginter, F., Goldberg, Y., Haji , J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D.* (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). [online] European Language Resources Association (ELRA), pp.1659–1666. Available at: https://www.aclweb.org/anthology/L16-1262.pdf [Accessed 18 Oct. 2019].

7. *GitHub*. (2019). UniversalDependencies/UD_Ukrainian-IU. [online] Available at: https://github.com/UniversalDependencies/UD_Ukrainian-IU [Accessed 18 Oct. 2019].

8. Lang.org.ua. (2019). Models: lang-uk. [online] Availableat: http://lang.org.ua/en/models [Accessed 18 Oct. 2019].

9. *Glybovets, A.* (2017). AutomatedsearchofnamedentitiesinunmarkedUkrainiantexts. [Avtomaty`zovany`jposhukimenov any`xsutnostej u nerozmicheny`xtekstaxukrayins`koyumovoyu].Shtuchny`jintelekt, 2(76), pp.45–52.

10. *Powers, D.* (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), pp.37–63.

11. Universaldependencies.org. (2019). UniversalDependencies. [online] Availableat: https://universaldependencies.org/guidelines.html [Accessed 18 Oct. 2019].

12. Laboratoriyaukrayins`koyi. (2019). Zoloty`jstandart. [online] Availableat: https://mova.institute/%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D0%B8%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82 [Accessed 18 Oct. 2019].

13. GitHub. (2019). Pythonpackagetoextract NP fromtheUkrainianlanguage. [online] Availableat: https://github.com/artemkramov/np-extractor-ua [Accessed 18 Oct. 2019].

Received 29.10.2019.

*S.D. Pogorilyy***,** Doctor of technical sciences, professor,
Taras Shevchenko National University of Kyiv,
03022, Kyiv, Glushkov ave., 4G, Ukraine,
sdp@univ.net.ua

A.A. Kramov, Postgraduate student,
Taras Shevchenko National University of Kyiv,
03022, Kyiv, Glushkov ave., 4G, Ukraine,
artemkramovphd@knu.ua

METHOD OF NOUN PHRASE DETECTION IN UKRAINIAN TEXTS

Introduction. The area of natural language processing considers *AI*-complete tasks that cannot be solved using traditional algorithmic actions. Such tasks are commonly implemented with the usage of machine learning methodology and means of computer linguistics. One of the preprocessing tasks of a text is the search of noun phrases. The accuracy of this task has implications for the effectiveness of many other tasks in the area of natural language processing. In spite of the active development of research in the area of natural language processing, the investigation of the search for noun phrases within Ukrainian texts are still at an early stage.

Purpose. Comparative analysis of the main methods of noun phrases detection in English and Ukrainian texts. The creation of a complex method for the detection of noun phrases in texts according to the features of the Ukrainian language. The performing of experimental examination of the suggested method on the corpus of Ukrainian articles.

Results. The different methods of noun phrases detection have been analyzed. The expediency of the representation of sentences as a tree structure has been justified. The key disadvantage of many methods of noun phrase detection is the severe

dependence of the effectiveness of their detection from the features of a certain language. Taking into account the unified format of sentence processing and the availability of the trained model for the building of sentence trees for Ukrainian texts, the Universal Dependency model has been chosen. The complex method of noun phrases detection in Ukrainian texts utilizing Universal Dependencies means and named-entity recognition model has been suggested. Experimental verification of the effectiveness of the suggested method on the corpus of Ukrainian news has been performed. Different metrics of method accuracy have been calculated.

Conclusions. The results obtained can indicate that the suggested method can be used to find noun phrases in Ukrainian texts. An accuracy increase of the method can be made with the usage of appropriate named-entity recognition models according to a subject area.

***Keywords**: natural language processing, noun phrase, Universal Dependencies model, NER model, tree structure of a sentence.*

*С.Д. Погорелый***,** д-р. техн. наук, профессор,
Киевский национальный университет имени Тараса Шевченко,
03022, Киев, просп. Академика Глушкова, 4Г,
sdp@univ.net.ua

*А.А. Крамов***,** аспирант,
Киевский национальный университет имени Тараса Шевченко,
03022, Киев, просп. Академика Глушкова, 4Г,
artemkramovphd@knu.ua

МЕТОД ОПРЕДЕЛЕНИЯ ИМЕННЫХ ГРУПП В УКРАИНОЯЗЫЧНЫХ ТЕКСТАХ

Введение. Отрасль обработки естественного языка рассматривает *AI*-полные задачи, которые не могут быть решены с помощью алгоритмических действий. Задачи такого типа решаются с использованием методологии машинного обучения и методов компьютерной лингвистики. Одной из задач предварительной обработки текста является поиск именных групп; точность их определения существенно влияет на эффективность решения многих задач обработки естественного языка. Несмотря на активное развитие исследований в направлении обработки естественного языка, исследование поиска именных групп в украиноязычных текстах находится на начальном этапе.

Цель статьи. Сравнительный анализ основных методов поиска именных групп в англоязычных и украиноязычных текстах. Создание комплексного метода определения именных групп в текстах соответственно с особенностями украинского языка. Осуществление экспериментальной проверки предложенного метода на корпусе украиноязычных статей.

Результаты. Проанализированы методы поиска именных групп в тексте и обоснована целесообразность использования древовидной синтаксической структуры предложения. Недостатком многих методов поиска именных групп в тексте является зависимость эффективности их определения от свойств конкретного языка. Решено использовать модель *Universal Dependencies* в связи с унифицированным форматом обработки предложения для разных языков и наличием обученной модели построение древовидной структуры предложений украиноязычных текстов. Предложен комплексный метод определения именных групп в украиноязычных текстах с использованием средств *Universal Dependencies* и модели распознавания именованных сущностей. Осуществлена экспериментальная проверка эффективности предложенного метода на корпусе украиноязычных новостей и рассчитаны метрики точности метода.

Выводы. Полученные результаты рассчитанных метрик точности предложенного метода могут свидетельствовать о целесообразности применения метода для поиска именных групп в украиноязычных текстах. Улучшение точности метода возможно с помощью применения моделей и шаблонов распознавания именованных сущностей в соответствии с рассматриваемой предметной областью.

***Ключевые слова**: обработка естественного языка, именная группа, модель Universal Dependencies, модель NER, древовидная структура предложения.*