DOI: 10.35681/1560-9189.2020.22.3.219002

УДК 519.816

М. Ю. Дубок, В. В. Циганок
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Метод частиномовної розмітки на основі квазіфлексій

Якість рекомендацій, що надаються як результат застосування систем підтримки прийняття рішень, значною мірою залежить від якості та достовірності знань, наданих експертами. Вирішення задачі автоматичного визначення неоднозначності у текстових формулюваннях експертів є беззаперечним кроком до підвищення достовірності знань і адекватності моделей, на основі яких здійснюється підтримка прийняття рішень. Більшість підходів до автоматичного визначення неоднозначності спираються на використання частиномовної розмітки як першого етапу аналізу при визначенні неоднозначності. Запропоновано метод автоматичної частиномовної розмітки на основі квазіфлексій (змінюваних складових слова), точність якого є співмірною із наявними реалізаціями підходу на основі правил. До переваг підходу на основі правил відносяться: значне зменшення необхідного обсягу інформації, нескладне впровадження вдосконалень аналізатора та висока ступінь портативності компонентів (правил, лексики, квазіфлексій, винятків).

***Ключові слова**: підтримка прийняття рішень, частиномовна розмітка, квазіфлексія, неоднозначність.*

Аналіз проблемної ситуації

Одним із етапів процесу прийняття рішень є групова побудова бази знань (БЗ) предметної області (ПО), «базуючись як на об'єктивній, так і на експертній інформації». Оскільки значна частка знань належить спеціалістам-експертам, використання експертних знань є дуже важливим «для повного та адекватного відображення всіх властивостей предметної області у БЗ» систем підтримки прийняття рішень (СППР) [1].

Для СППР усіх класів обов'язковим є етап отримання експертної інформації [2], і саме на цьому етапі може початися непорозуміння, результатом якого буде неправильна інтерпретація при побудові ієрархії критеріїв, взаємозв'язків і/або впливів на головну ціль і підцілі.

Формулювання головної цілі, підцілей, програм, проектів, заходів, як і будь-які експертні формулювання, використовують природну (людську) мову, що характеризується неоднозначністю на всіх мовних рівнях. Формулювання, які можуть

сприйматися по-різному різними зацікавленими сторонами внаслідок неоднозначності, становлять загрозу зниження адекватності моделей і спричиняють збільшення витрат і часу розробки у разі потреби виправлень і змін. Адекватність створених моделей ПО безпосередньо впливає на якість рекомендацій, що надаються системами підтримки прийняття рішень.

Для зниження неоднозначності текстової інформації застосовують наступні способи її обробки. Найпоширенішим способом є вирішення неоднозначності (ambiguity resolution) — автоматичний процес, у ході якого визначається, який сенс (тлумачення) є більш імовірним. Однак, враховуючи ризик неправильної автоматичної інтерпретації і кількість залучених у процес осіб, які оперуватимуть такою інтерпретацією, це не є прийнятним варіантом для систем підтримки прийняття рішень. Неавтоматичний спосіб обробки неоднозначності представлений чотирма різними техніками, за допомогою яких його можна реалізувати: уникнення неоднозначності (інструкції щодо написання), запобігання неоднозначності (написання у фіксованому форматі), виявлення неоднозначності (автоматичне виявлення у письмовому тексті) та виправлення неоднозначності (напівавтоматичні засоби корекції, що взаємодіють із користувачами) [3]. Уникнення неоднозначності є не дуже ефективним, оскільки навіть при наданих інструкціях користувачі вкрай рідко дійсно роблять речення менш неоднозначними [4]. Запобігання неоднозначності, безумовно, встановлює межі, що зменшують шанс писати неоднозначно, але це також досить обмежувальний прийом. Значну кількість наукових публікацій присвячено методам розпізнавання неоднозначності людьми, але автоматичних реалізацій виявлення неоднозначностей є набагато менше.

Метод визначення неоднозначності застосовуватиметься у двох функціональних підсистемах експертних СППР: підсистемі отримання знань від експертів про предметну область (під час написання експертних формулювань) і підсистемі обробки та узагальнення експертних знань [5]. До таких СППР зокрема належать Система підтримки прийняття рішень «Солон-3» [6] та Система розподіленого збору та обробки експертної інформації «Консенсус-2» [7].

Ґлейх та ін. [8] використали лексичний і поверхневий синтаксичні аналізи для виявлення неоднозначностей усіх чотирьох типів, проте всі прикметники та прислівники розглядаються у методі як потенційна неоднозначність, що дає занадто багато помилково позитивних результатів.

Тому актуальним є створення власного методу лексичного аналізу при написанні експертних формулювань. Основу цього аналізу становить частиномовна розмітка.

Частиномовна розмітка або граматичне тегування — це процес приписування частиномовних тегів словам у тексті [9]. Процес тегування складається з трьох етапів: токенізації, морфологічного аналізу (приписуванні можливих тегів) та вирішенню неоднозначності (вибору найімовірнішого тегу) [10].

Традиційно виділяють 2 основні підходи до автоматичного морфологічного аналізу: підхід на основі правил (rule-based) або на основі обмежень (constraint-based) та статистичний (statistical) або ймовірнісний (probabilistic) підхід, більш відомий як стохастичний (stochastic) [11–13]. Наразі виокремлюють також трансформаційний підхід як третій із основних [9, 14]. Деякі науковці вважають, що третім основним підходом є дистрибутивний [15].

Підхід на основі правил використовує контекстну інформацію, щоб обмежити кількість можливих частиномовних тегів [16] або визначити частиномовний тег слова [17]. Наприклад, якщо в англомовному тексті слову передує артикль, а слідує іменник, то буде приписано тег «іменник». Окрім контекстної інформації часто використовується морфологічна інформація. Наприклад, якщо слову передує допоміжне дієслово, і слово має закінчення -ing, буде приписано тег «дієслово». Деякі аналізатори враховують регістр та пунктуацію [12]. Така інформація є корисною в різних мовах. Наприклад, в англійській мові регістр допомагає визначити, чим є слово «us»: «нас» або абревіатурою «Сполучені Штати» (US). У німецькій мові, наприклад, усі іменники починаються з літери у верхньому регістрі.

При статистичному (стохастичному) підході обирається найімовірніший тег на основі статистичних даних, отриманих при аналізі однозначно розміченого тексту. Як критерій максимізації використовується частота слів або ймовірності *N*-грам. Найбільш загальним алгоритмом для впровадження підходу на основі *N*-грам є алгоритм Вітербі. Складнішим є поєднання ймовірностей певних послідовностей тегів і частоти слів [13].

Трансформаційний підхід, розроблений Бріллом, використовує машинне навчання, поєднуючи підхід на основі правил і ймовірнісний. Як і підхід на основі правил, трансформаційне навчання засноване на правилах. Подібно ймовірнісному підходу, правила автоматично здобуваються із даних (корпусу текстів). Одним із найбільш широко використовуваних трансформаційних інструментів тегування є аналізатор, розроблений Бріллом [18].

Дистрибутивне тегування є підходом, метою якого є усунення необхідності використання ручних правил і розмічених навчальних корпусів, які можуть бути відсутні для певних мов або галузей. На відміну від підходу на основі правил і стохастичного підходу, дистрибутивне тегування виконується повністю без допомоги вчителя. Шютце [19] запропонував аналіз дистрибутивних шаблонів слів шляхом побудови матриці суміжності термінів із подальшим сингулярним розкладом цієї матриці для виявлення латентних розмірностей (latent dimensions). У просторі зниженої розмірності, передбаченій сингулярним розкладом матриці, лексеми дійсно інтуїтивно групуються за частинами мови. При врахуванні контексту можливо досягти результатів, подібних до частиномовної розмітки.

При порівнянні підходу на основі правил і стохастичного підходу часто обирається ймовірнісний через можливість автоматичного навчання та менші часові витрати [11]. Також стохастичний підхід ефективніший при аналізі довгих речень, ніж підхід на основі правил [20]. Більше того, ймовірнісний підхід, зокрема із використанням Марковської моделі, краще використовувати у випадках, коли набір можливих тегів (tagset) є незначним за обсягом. Якщо ж обсяг набору тегів великий, підхід на основі правил є ефективнішим [21].

Порівнюючи заявлену точність частиномовних аналізаторів, Марковські моделі та трансформаційний підхід досягають точності до 97 % [18, 22]. Водночас точність підходу на основі правил варіюється від 97 % до 100 % [23].

Е. Брілл виділив такі переваги підходу на основі правил над стохастичним підходом: невеликий обсяг інформації, що зберігається; чіткість набору змістовних правил; простота знаходження та впровадження покращень; більша ступінь портативності з одного набору тегів, жанру корпусу чи мови в іншу [17]. Втім,

ступінь портативності правил все ж є низьким, оскільки для кожної мови потрібно виводити інші правила.

Враховуючи необхідність отримати коректні правила, які правильно визначають частиномовну приналежність, можливість автоматичного навчання не є пріоритетом. Текстові формулювання у системах підтримки прийняття рішень бувають різної, проте переважно незначної довжини. Розглядаючи обсяг наборів можливих тегів, окрім загальнопоширених частин у кожній мові (іменник, дієслово тощо), розмір кожного набору залежить від конкретної мови. Точність кожного із трьох підходів (на основі правил, статистичного та трансформаційного) є приблизно ідентичною, проте можливість досягнути дещо вищого показника є важливою при виборі підходу тегування, оскільки від частиномовної розмітки сильно залежить наступний етап глобального аналізу — синтаксичний. Отже, саме підхід на основі правил є пріоритетним для вирішення задачі в галузі експертної підтримки прийняття рішень. Утім, не виключається також гібридний підхід на певному етапі розробки, наприклад, використання ймовірності як останньої інстанції у випадках, коли кількість можливих тегів не скорочується до одного, незважаючи на використання правил. У дослідженнях [12] експериментально перевірено користь гібридного підходу: поєднуючи морфологічний аналізатор і статистичні дані з корпусу, було збільшено середню ефективність на 15,53 % порівняно з використанням лише морфологічного аналізатора.

Окрім використовуваного підходу, частиномовна розмітка поділяється на тегування з учителем (supervised), тобто автоматизоване, та без учителя (unsupervised) — автоматичне.

Система, представлена Бріллом у 1992 році, розглядається автором і деякими науковцями як вид підходу на основі правил [11]. Після того, як усім словам, наявним у словнику, приписується найчастотніший тег, модуль вгадує тег слова відповідно до його афіксів (префікс, суфікс). Далі застосовуються контекстні правила [17]. Цей метод є прийнятним у використанні як основа.

Окрім точного визначення правильного тегу, підхід на основі правил може використовуватися для виявлення переліку можливих тегів, тоді визначення конкретної частини мови здійснюється іншим модулем [12]. Утім, якщо надано інформацію про частини мови сусідніх слів, цей підхід здатен виконати обидва завдання. Таким чином, працює вже наявний тегер (аналізатор) на основі обмежень, зчитуючи послідовність слів та альтернативні теги і передаючи граматичному перетворювачу, який залишає лише один тег для кожного слова на основі контекстної інформації [10].

Не переходячи до синтаксичного аналізу, контекстний аналіз можливий шляхом аналізу послідовностей слів [13]. Наприклад, у реченні «у великих містах рівень забруднення повітря значно вищий» можна виокремити послідовність слів «у великих містах», яку можливо аналізувати незалежно від решти слів у реченні. Таким чином, є точна інформація, що слово «у» завжди є прийменником у всіх текстах, окрім розмовного та літературного стилів. Якщо відомо, що словоформа «великих» є прикметником, а словоформа «містах» є невідомою, то цій послідовності може відповідати послідовність тегів «прийменник + прикметник + іменник». Ця ж послідовність тегів може бути приписана, якщо відомо, що словоформа «містах» є іменником, а «великих» — невідомою. Такі послідовності слів називаються *N*-грамами.

Брілл продемонстрував, що підхід на основі правил без жодного знання про синтаксис може мати таку ж ефективність, що і стохастичний підхід. Розроблений ним аналізатор спочатку приписує кожному слову найбільш імовірний тег на основі аналізу великого розміченого корпусу без врахування контексту. Потім виконуються дві процедури: 1) слова, які відсутні в тренувальному корпусі та починаються з великої літери, вважаються власними назвами; 2) решті слів, які відсутні в тренувальному корпусі, приписується найчастотніший тег для останніх трьох літер слова. Цей простий алгоритм має низький коефіцієнт помилок — 7,9 %. Тренування полягає в укладанні списку слів і їхнього найчастотнішого тегу. Далі застосовуються шаблони правил, які значно зменшують кількість помилок, зроблених до застосування шаблонів, і призводять до появи незначної кількості нових помилок. Процес тестування нових шаблонів є легким, оскільки шаблони, які породжують правила, що дають переважно неправильні результати, не попадають до остаточного списку шаблонів. Застосувавши лише 71 шаблон, коефіцієнт помилок вдалося знизити до 5,1 %. Із цих шаблонів 66 знизили кількість помилок, 3 не дали результат, а 2 збільшили кількість помилок. Також Брілл відзначив здатність підходу на основі правил точніше робити частиномовну розмітку в ідіомах, завдяки автоматичному навчанню. Таким чином, єдине знання, здобуте неавтоматичним шляхом, — це процедура виявлення особових назв [17].

Недоліком підходу Брілла є низька швидкість роботи порівняно зі стохастичним підходом. Підвищення швидкості за умови збереження ефективності можливе шляхом перетворення програми на основі трансформацій у детермінований скінченний перетворювач [11]. Утім, це нівелює важливу перевагу аналізатора — простоту його побудови, на якій наголошував Брілл [17].

Підхід Брілла, пізніше названий трансформаційним, все ж є гібридом підходів на основі правил і стохастичного та досягає точності, що є ідентичною ймовірнісному підходу — 97 % [18].

Оскільки визначення неоднозначності формулювань значною мірою залежить від точності частиномовного аналізу, прийнято рішення використати підхід на основі правил з автоматизованим здобуттям правил.

Виходячи із вище описаного аналізу, існує нагальна необхідність розробити метод, що має високу точність. Оскільки статистичний і трансформаційний підходи досягають точність, не вищу за 97 %, перспективним є підхід на основі правил.

Мета дослідження

Підвищення достовірності знань та адекватності моделей, на основі яких здійснюється підтримка прийняття рішень завдяки автоматичному визначенню неоднозначності у текстових формулюваннях експертів. Розробка високоточного методу частиномовної розмітки є однією із складових визначення неоднозначності текстових формулювань.

Формальна постановка задачі

Дано: $W = \{w_i\}$, $i = (1, n)$ — множина слів експертних формулювань; $P = \{p_j\}$, $j = (1, m)$ — множина назв частин мови; $D = \{d_k\}$, $k = (1, l)$ — множина словників, де словам ставиться відповідність граматики; n, m та l — потужності відповідних множин (кількість елементів у множинах).

Потрібно визначити: Відображення $W \rightarrow P$.

Запропонований метод

Для вирішення задачі пропонується метод, який. на відміну від наявних методів, базується на використанні виключно списків (словників) правил і квазіфлексій. Завдяки використанню в методиці побудови словників автоматичного здобуття елементів множини W, що не мають відображення $W \rightarrow P$, автоматичного інверсійного (зворотного) алфавітного сортування невідображених елементів та автоматичного здобуття усіх елементів множини W, що містять квазіфлексію, чия частиномовна відповідність перевіряється, досягається високий рівень відповідності у визначеному відображенні.

Методика побудови словників правил і квазіфлексій полягає в процедурі, яка передбачає автоматизовану перевірку достовірності правил і квазіфлексій. Вона може виконуватися довільно або системно. При довільному виконанні процедури:

1) укладач правил (УП) автоматично здобуває перший елемент з множини W, що не має відображення $W \rightarrow P$;

2) УП власноруч визначає відповідний елемент множини P;

3) УП власноруч висуває гіпотезу, що усім елементам W, які містять квазіфлексію, виділену з даного елемента W, потрібно поставити у відповідність такий самий елемент із P;

4) УП власноруч створює відповідне правило типу *Умова* → *Результат* у вигляді «*if ТипПравила текст тег*», записуючи у текстовий файл з правилами (описано нижче);

5) УП автоматично за допомогою програмного забезпечення здобуває текстовий файл з усіма неповторюваними елементами W, що відповідають умові правила, відсортованими в інверсійному алфавітному порядку;

6) УП власноруч переглядає усі отримані елементи W на випадок невідповідності елементу множини P;

7) якщо виявлено невідповідність, УП власноруч:

а) уточнює правило, збільшуючи межу квазіфлексії до межі, де підмножина W не міститиме елементи, що не відповідають вказаному в правилі елементу множини P, або

б) додає перед цим правилом додаткові правила, що мають охопити усі елементи підмножини W, які не відповідають вказаному в правилі елементу множини P;

8) УП повторює процедуру спочатку для наступного елемента із W, що не має відображення $W \rightarrow P$.

При системному виконанні процедури:

1) УП автоматично здобуває усі неповторювані елементи множини W, що не мають відображення $W \rightarrow P$, відсортовані в інверсійному алфавітному порядку;

2) УП власноруч по порядку виділяє нерозривну послідовність елементів W, яким відповідає 1 елемент множини P;

3) УП знаходить спільну квазіфлексію у всіх елементів виділеної послідовності;

4) УП власноруч укладає правило на основі спільної квазіфлексії записуючи у текстовий файл з правилами (описано нижче);

5) УП повторює процедуру з кроку 2 для наступної послідовності елементів множини W.

Для укладання словника квазіфлексій процедура та способи її виконання ідентичні до укладання словника правил. Відмінність тільки в тому, що для словника квазіфлексій єдиною умовою є лише закінчення слова на певну квазіфлексію.

Метод вирішення:

1) попередній етап (описано вище);

2) перевірити елемент множини W на повний збіг. У випадку збігу перейти до кроку 5;

3) шляхом аналізу правилами отримати елемент множини P, тобто аналізований клас (частину мови);

4) шляхом аналізу отримати аналізований підклас, тобто граматичні характеристики, наявні у кожного елемента множини P;

5) шляхом поступового усічення кінцевих літер спробувати знайти базову форму елемента множини W в усіх елементах множини D. У випадку відсутності збігу перейти до кроку 6;

6) додати інформацію з відповідних словників до елемента відображення $W \rightarrow P$;

7) повторити кроки 1–5 для кожного елемента множини W;

8) повернути відображення $W \rightarrow P$.

Метод реалізовано у вигляді автоматичного частиномовного аналізатора на основі квазіфлексій, використовуючи підхід на основі правил.

При використанні аналізатора користувач вводить текст у діалогове вікно, запускає автоматичний аналіз та отримує інформаційні картки для кожного слова. Кожна інформаційна картка містить: 1) введене слово; 2) його аналізовані клас і підклас, якщо їх вдалося визначити; 3) вихідну форму; 4) перелік можливих словникових варіантів, якщо їх було знайдено. Кожен словниковий варіант містить власні клас і підклас. Наприклад, для введеного слова «вулиць» інформаційна картка міститиме аналізований клас «іменник», аналізований підклас «мнЖ» (множина від жіночого роду), неповний збіг з вихідною формою «вулиця» та один словниковий варіант, що передбачає іменник жіночого роду однини.

Для коректної роботи аналізатора описаним чином, слід виконати попередній етап. Він полягає у створенні бази правил для визначення основних класів (частин мови), що здійснюється укладачем правил, і заповненні словника, який укладається лексикографом або імпортується. Перелік, назви та структура файлів, що містять словники та правила, не є фіксованими та можуть змінюватися відповідно до потреб. На попередньому етапі також укладається список квазіфлексій для визначення підкласів (граматичних характеристик).

Порядок укладання правил може бути довільним, наприклад при випадковому виявленні нового нерозпізнаного слова, або системним. При системному підході автоматично здобувається перелік усіх нерозпізнаних слів, з якого в автоматизованому режимі має можливість аналізу та подальшого виводу правил.

В обох випадках підставою для укладання правила є невідома словоформа, яка в режимі розробки демонструється людині, яка укладає правила (укладач). Укладач створює правило типу «*if ТипПравила текст тег*», де *ТипПравила* — це тип перевірки з визначеного переліку («закінчується», «дорівнює», «не дорівнює», «починається»), *текст* — це частина словоформи, на яку діє правило, *тег* — це приписана частина мови з визначеного переліку («іменник», «прикметник, «дієслово» тощо). Наприклад, відповідно до правила «*if EndsWith ба іменник*», якщо

словоформа закінчується на «*ба*», то приписується тег «*іменник*». Таким чином висувається гіпотеза, що усі слова, які закінчуються на літери «ба» є іменниками.

Кожну гіпотезу слід перевіряти, перш ніж повноцінно використовувати нове правило. Для підтвердження або спростування гіпотези при перевірці правила «*if*» змінюється на «*ifLog*» для відстеження, після чого укладач запускає автоматичний аналіз корпусу текстів. Усі випадки спрацювання правила фіксуються. Для фільтрації тексту повторювані словоформи усуваються з остаточного звіту. Після аналізу результати виводяться в інверсійному до алфавітного порядку. Також, застосовується розширюваний список винятків, які не фіксуються у звіті.

Для підвищення точності правил, тобто обмеження їхнього застосування, укладач може поєднувати їх, використовуючи оператор логічного «І» («AND»). Наприклад, правило «if EndsWith аві AND NotEquals праві іменник» приписує тег «іменник» усім словам, які мають квазіфлексію «аві», окрім словоформи «праві».

Частина мови більшості слів, а саме тих, які не можуть належати до кількох частин мов або до однієї частини мови з різними граматичними характеристиками (наприклад, різний відмінок), однозначно встановлюється, використовуючи лише квазіфлексії, тобто останні літери слів, які можуть бути меншими, рівними або більшими за морфемне закінчення, зокрема квазіфлексія може бути рівна всьому слову. Тобто, квазіфлексія є змінною частиною слова, починаючи з кінця.

Для зменшення кількості правил певний клас розглядається як основний, а решта — як винятки. Наприклад, за замовчанням слова, що закінчуються на літеру «г» в українській мові вважаються іменниками. Але словам, що мають квазіфлексію «міг» або «сяг», приписується тег «дієслово». Основне правило розміщується останнім в списку, щоби правила-винятки перевірялися першими. Таким чином, на момент публікації у розробленій системі використовується 804 правила, за якими визначається частина мови.

Окрім графемного аналізу із використанням правил, використовується неповний збіг для слів, що вжиті не у базовій формі, наприклад «слова» (базова форма «слово»). У рамках даного дослідження реалізовано алгоритм, що передбачає поступове усічення останніх літер словоформ у текстовому формулюванні та словнику, що є реалізацією стемінгу (обрізання) змінюваних частин слів [24], який використовується для багатьох потреб обробки природної мови, зокрема у системах контент-моніторингу [25]. Для уникнення неправильних збігів, наприклад іменник «раз» у словнику для прислівника «разом» у формулюванні, забезпечено порівняння аналізованої частини мови з частиною мови у словнику. Також, для підвищення кількості правильних збігів порівнюються граматичні характеристики. Наприклад, якщо у текстовому формулюванні зустрічається іменник «президент», і в словнику ця словоформа порівнюється з наявним іменником «президентство», обов'язково потрібно порівнювати рід іменників. Таким чином, збіг не відбудеться, оскільки шляхом аналізу словоформи «президента» буде отримано чоловічий рід, а в словнику іменник «президентство» має середній рід. Неповний збіг може охоплювати до 6 останніх символів.

У цій реалізації як для повного, так і для неповного збігу використовується 16 словників загальним обсягом у 5030 реєстрових слів. У випадку, якщо у словнику є слово з експертного текстового формулювання, вжите саме у базовій формі, аналіз правилами та неповний збіг не здійснюються. За таким самим принципом діє і підхід Брілла [17]. Недоліком цього принципу є неправильне визначення

тегу у випадках, коли у тексті зустрічається словоформа не в базовій формі, проте є формальний збіг з іншим словом. Наприклад, у словосполученні «промислового робота» словоформі «робота» буде приписано тег жіночого роду замість чоловічого, оскільки не здійснюватиметься неповний збіг.

Розпізнавання числа та роду іменників у формулюванні реалізовано також шляхом аналізу квазіфлексії. Для кращого збігу множини з правильною базовою формою виділяються окремі квазіфлексії для кожного із трьох типів множини: 1) множина, базовою формою якої є чоловічий рід; 2) множина, базовою формою якої є жіночий рід; 3) множина, базовою формою якої є середній рід. При первинному розпізнаванні множини іменник перевіряється на наявність квазіфлексії будь-якого типу множини.

Для укладання правил і списків квазіфлексій було укладено власний україномовний корпус текстів офіційно-ділового стилю обсягом у понад 137 тис. слів. До корпусу увійшли текстові формулювання у 7 файлах структур ієрархії цілей, які вносились у Систему підтримки прийняття рішень «Солон-2», тексти 14 конвенцій ООН (з морського права, про договори міжнародної купівлі-продажу товарів, про права дитини, проти корупції, проти транснаціональної організованої злочинності, про захист прав людини і основоположних свобод, про ліквідацію всіх форм дискримінації щодо жінок, про охорону біологічного різноманіття від 1992 року, про права осіб з інвалідністю, проти катувань та інших жорстоких, нелюдських або таких, що принижують гідність, видів поводження та покарання тощо), один контракт, судова ухвала, понад 40 уривків текстів.

Оскільки мовлення є динамічним явищем, усі правила укладаються на основі фактичних випадків, а не теоретичних принципів. Наприклад, незважаючи на теоретичні відомості, що в текстах офіційного стилю більшість слів є іменниками, не для всіх квазіфлексій створюються основні правила, згідно яких словоформі приписується тег «іменник». В окремих випадках більшість слів, що закінчуються на певну квазіфлексію, можуть бути прикметниками або дієсловами. Більше того, є літери, на які слова не можуть закінчуватися. У тренувальному корпусі не виявлено словоформи, які б закінчувалися на літери «ґ» або «ц». Таким чином, дотримуючись фактичного принципу укладання правил, можливо уникнути надмірності.

Результати

Новизна методу полягає у застосуванні квазіфлексій як основного і єдиного методу для визначення частини мови та граматичних характеристик слова.

Для перевірки ефективності запропонованого методу тестування проводилося на основі текстових формулювань у 7 файлах структур ієрархії цілей, що надавались Системі підтримки прийняття рішень «Солон-3» [6] та застосовувалися при колективній побудові моделей предметних областей у Системі розподіленого збору та обробки експертної інформації — «Консенсус-2» [7], загальним обсягом у 4378 слів. Результати ручного тегування порівнювалися з результатами автоматичного тегування, вважаючи ручне тегування як правильне. В розмітці використано 17 основних класових тегів: іменник, прикметник, прийменник, сполучник, абревіатура, дієслово, прислівник, займенник, дієприкметник, дієприслівник, частка, число, скорочення, числівник, помилка (неправильне написання), сполучне-Слово (наприклад, «як»), інше (наприклад, іншомовні слова). Класові теги слів при ручній розмітці узгоджено з інформацією, що надається ресурсом «Словники

України online» українського мовно-інформаційного фонду НАН України. Кожній текстовій одиниці приписувався тег з урахуванням контексту.

При тестуванні було виявлено 59 помилок автоматичної розмітки, при цьому в ручній розмітці було 2 слова з тегом «помилка», тобто це слова, що мали неправильне написання у тексті. Для нівелювання їхнього впливу, кількість текстових слів, які містять помилки, слід відняти від кількості помилок у автоматичній розмітці та від загальної кількості слів тексту.

Для визначення точності A використано наступну формулу:

$$A = 100\ \% - \frac{x - y}{n - y} \times 100\ \%,$$

де A — точність; x — кількість помилок в автоматичній розмітці; y — кількість застосувань класу «помилка» в ручній розмітці; n — кількість слів.

Підставивши наявні значення, маємо:

$$A = 100\ \% - \frac{59 - 2}{4378 - 2} \times 100\ \% = 98{,}697\ \% \approx 98{,}70\ \%.$$

Таким чином, аналізатор продемонстрував точність у 98,70 % при частиномовному аналізі текстових формулювань у тестовому корпусі. Слова, які при ручній розмітці отримали тег «помилка» не враховано при визначенні точності, оскільки частиномовний аналізатор не є програмою для виявлення та виправлення помилок. Пріоритетним було обрано тег, отриманий шляхом аналізу. Якщо аналізований тег відсутній через повний збіг або неможливість визначення, частиномовний тег отримувався зі словника.

Більшість слів, клас яких не вдалося визначити, є скороченнями.

Продемонстрована висока точність методу може бути досягнута за умови чіткого дотримання описаної методики укладання словників правил та квазіфлексій.

Висновки

Представлено метод частиномовної розмітки на основі квазіфлексій, що має точність 98,70 %, співмірну з наявними сучасними методами, які використовують підхід на основі правил, але водночас відрізняється послідовністю використовуваних засобів, зокрема квазіфлексій і неповного збігу.

Запропонований метод та аналізатор, розроблений на його основі, є значною мірою портативними. Портативність полягає у можливості використати розроблену систему словників, винятків, правил і їхні типи та число-родові списки квазіфлексій для інших мов, утім їхні переліки відрізнятимуться для кожної мови. Заповнення кожного компоненту не вимагає внесення змін до програмного коду.

Іншою перевагою запропонованого методу, як і кожної реалізації підходу на основі правил, є відсутність потреби зберігати значні за обсягом статистичні таблиці.

Подальше дослідження планується виконати, використовуючи запропонований метод на наступному етапі автоматичного визначення неоднозначності — синтаксичному, який спирається на інформацію, отриману на лексичному етапі, зокрема частиномовну розмітку.

1. Циганок В.В., Андрійчук О.В. Експериментальне дослідження методу визначення змістової подібності об’єктів баз знань систем підтримки прийняття рішень. *Реєстрація, зберігання і оброб. даних.* 2014. Т. 16. № 4. С. 64–75.

2. Тоценко В.Г. Методы и системы поддержки принятия решений. Алгоритмический аспект. ИПРИ НАНУ. Киев: Наук. думка, 2002. 382 с.

3. Alomari R. and Elazhary H. Implementation of a Formal Software Requirements Ambiguity Prevention Tool. 2018.

4. Winkler S. Ambiguity: Language and Communication. 2015.

5. Циганок В.В. Концепція створення систем підтримки прийняття рішень, що адаптивні до рівня компетентності експертів. *Реєстрація, зберігання і оброб. даних*. 2011. Т. 13. 2. С. 106–114.

6. Свідоцтво про державну реєстрацію авторського права на твір № 8669. Міністерство освіти і науки України державний департамент інтелектуальної власності. Комп'ютерна програма «Система підтримки прийняття рішень СОЛОН-3» (СППР СОЛОН-3) / Тоценко В.Г., Качанов П.Т., Циганок В.В. зареєстровано 31.10.2003.

7. Свідоцтво про реєстрацію авторського права на твір № 75023. Комп'ютерна програма «Система розподіленого збору та обробки експертної інформації для систем підтримки прийняття рішень – «Консенсус-2» / Циганок В.В., Роїк П.Д., Андрійчук О.В., Каденко С.В. Від 27/11/2017.

8. Gleich B., Creighton O. and Kof L. Ambiguity Detection: Towards a Tool Explaining Ambiguity Sources. 2010.

9. Pisceldo F., Adriani M., and Manurung R. Probabilistic Part Of Speech Tagging for Bahasa Indonesia. 2009.

10. Chanod J., and Tapanainen P. Tagging French — Comparing a Statistical and a Constraint-based Method. 1995.

11. Roche E. and Schabes Y. Deterministic Part-of-Speech Tagging with Finite-State Transducers. 1995.

12. Altunyurt L., Orhan Z., and Güngör T. Towards Combining Rule-Based and Statistical Part of Speech Tagging in Agglutinative Languages. 2007.

13. Altunyurt L., Orhan Z., and Güngör T. A Composite Approach for Part of Speech Tagging in Turkish. 2006.

14. Sajjad H. Statistical Part of Speech Tagger for Urdu. 2007.

15. Chew P.A., Bader B.W., and Rozovskaya A. Using DEDICOM for Completely Unsupervised Part-of-speech Tagging. 2009.

16. Karlsson F., Voutilainen A., Heikkila J., Anttilla A. Constraint Grammar: a Language Independent System for Parsing Unrestricted Text. 1995.

17. Brill E. A Simple Rule-based Part of Speech Tagger. Proceedings of 3rd Conference on ANLP. 1992. P. 152–155. Trento.

18. Brill E., Transformation-based Error-driven Learning and Natural Language Processing: a Case Study in Part-of-speech Tagging. *Computational Linguistics*. 1995. **21**(4). P. 543–565.

19. Schütze H., Distributional Part-of-Speech Tagging. In Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics. 1995. P. 141–148.

20. Weischedel R., Meteer M., Schwartz R., Ramshaw L. and Palmuzzi J. Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics*. 1993. **19**(2). P. 359–382.

21. Tapanainen P., and Voutilainen A. Tagging Accurately — don't Guess if you Know. In: Proceedings of the Fourth Conference on Applied Natural Language Processing. Stuttgart. 1994.

22. Hardie A. The Computational Analysis of Morphosyntactic Categories in Urdu. PhD thesis, Lancaster University, 2003.

23. Voutilainen A. Morphological Disambiguation. In: Karlsson et all. 1995.

24. Ландэ Д.В., Дармохвал А.Т., Жигало В.В. Матричные критерии качества выявления подобных документов в информационных потоках: зб. наук. праць Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України. Київ: ІПМЕ ім. Г.Є. Пухова НАН України, 2009. Вип. 53. Бібліогр.: 13 назв. — рос.

25. Ланде Д.В. Елементи комп'ютерної лінгвістики в правовій інформатиці. Київ: НДІІП НАПрН України, 2014. 168 с. ISBN 978-966-2344-33-2.

Надійшла до редакції 09.10.2020