DOI: 10.35681/1560-9189.2022.24.1.262928

УДК 519.816

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

# Виявлення анафоричної неоднозначності експертних формулювань у системах підтримки прийняття рішень

*Якість рекомендацій системи підтримки прийняття рішень прямо залежить від однозначної інтерпретації експертних формулювань. Оскільки формулювання надаються експертами природною мовою, якій притаманне явище неоднозначності на усіх мовних рівнях, критично важливим є зниження неоднозначності експертних формулювань. Використовуючи виявлення неоднозначності для його зниження, увагу приділено підтипу синтаксичної неоднозначності — анафоричній неоднозначності, яка становить значну частку усіх наявних неоднозначностей і проявляється у вигляді можливості посилання слова на одразу декілька попередніх слів. Запропоновано та розроблено простий метод для пошуку та класифікації анафор, а також знаходження її антецедентів для виявлення цього підтипу синтаксичної неоднозначності, що має абсолютну повноту, високу точність і задовільну влучність.*

***Ключові слова:*** *підтримка прийняття рішень, система підтримки прийняття рішень, неоднозначність, зниження неоднозначності, виявлення неоднозначності, анафорична неоднозначність.*

## Вступ

Здатність людини приймати рішення в умовах великої кількості критеріїв, цілей, підцілей і зв’язків між ними є обмеженою внаслідок психофізіологічних можливостей людини. Для того, щоб допомогти людині прийняти ефективне рішення, створено комп’ютерні системи підтримки прийняття рішень, що здатні оперувати необмеженою кількістю об’єктів і зв’язків між ними в моделі предметної області. Ця модель значною мірою будується зі знань експертів, які надають системі свої формулювання природною мовою.

Оскільки будь-якій природній мові притаманна неоднозначність, формулювання можуть спричинити різні трактування, що негативно впливає на адекватність побудованої моделі.

## Аналіз проблемної ситуації

Важливим етапом процесу прийняття рішень є групова побудова бази знань предметної області (ПО), що включає об'єктивну та експертну інформацію [1]. Це є обов'язковим для систем підтримки прийняття рішень (СППР) усіх класів для відображення властивостей предметної області у базі даних [2]. На цьому етапі існує ризик неправильної інтерпретації наданих експертами формулювань, що призводить до хибної побудови структури ієрархії цілей, підцілей, критеріїв, взаємозв'язків і впливів. Наслідком хибного розуміння формулювань є зниження адекватності моделей та збільшення фінансових і часових витрат. Якість рекомендацій, наданих СППР, безпосередньо залежить від адекватності створеної моделі ПО.

Для зниження рівня неоднозначності можуть використовуватись автоматичний підхід (вирішення неоднозначності), який передбачає автоматичний вибір одного трактування, та неавтоматичний, який передбачає взаємодію з експертом. Неавтоматичний спосіб зниження неоднозначності може бути реалізований за допомогою інструкцій з написання (уникнення неоднозначності), написання у фіксованому форматі (запобігання неоднозначності), напівавтоматичних засобів корекції (виправлення неоднозначності) та автоматичного виявлення неоднозначностей [3]. Дослідивши недоліки інших технік, пріоритетним визначено саме виявлення неоднозначності [4].

Відповідно до класифікації за мовним рівнем, неоднозначність буває лексичною (слова, словосполучення), синтаксичною (структура), семантичною (контекст у межах формулювання) та прагматичною (попередній або позамовний контекст) [5]. Поширеним підтипом як семантичної, так і прагматичної неоднозначності, є анафорична неоднозначність. Анафора — це слово, зазвичай займенник, що посилається на антецедент — попереднє слово, зазвичай іменник, у даному чи попередньому формулюванні. Анафорична неоднозначність виникає тоді, коли анафора може посилатися на кілька потенційних слів [6]. Наприклад, у формулюванні «особа зі зброєю, яка знаходиться на обліку» слово «яка» може посилатися як на слово «особа», так і на слово «зброєю».

Питання виявлення анафоричної неоднозначності висвітлено науковцями Х. Янгом та ін. [6, 7], метою дослідження яких стало виявлення анафоричної неоднозначності та представлення лише «шкідливих», тобто тих, які з високою ймовірністю спричиняють різночитання, анафоричних неоднозначностей на прикладі англомовних текстів технічних вимог. Залежно від критичності неправильного трактування у конкретній галузі, дослідники використали різний поріг імовірності неоднозначного трактування, при якому виявлена неоднозначність представляється користувачеві для перегляду. Оскільки неправильне інтерпретування формулювання призводить до побудови моделі системи, що не відповідає уявленням особи, що приймає рішення у СППР, доцільним є лише процес знаходження анафоричних неоднозначностей.

Для знаходження анафоричних неоднозначностей Х. Янг та ін. використали попередню обробку тексту як перший етап, під час якого текст поділяється на окремі речення. У кожному реченні частиномовний аналізатор визначає слова, їхні леми та частини мови, а також позначає межі фраз. Іменникові фрази (англ. Noun

phrase) вважаються антецедентами. Далі в іншому модулі відбувається пошук анафор, тобто займенників третьої особи, шляхом повного збігу, а їхніми можливими антецедентами стають дві та більше іменникові фрази, що їм передують. Якщо займеннику передує лише одна іменникова фраза, інша береться із попереднього речення [6]. Теоретично використання фрагмента попереднього речення дозволяє виявляти прагматичну анафоричну неоднозначність, проте також теоретично більш імовірним є посилання лише на антецедент у тому ж реченні, де знаходиться і анафора. Також для експертних формулювань не характерне посилання на попереднє формулювання.

На жаль, праця Х. Янга та інших зазначає повноту (частка загального числа неоднозначних формулювань, які було знайдено) та влучність (частка релевантних формулювань серед знайдених) виявлення лише «шкідливих» анафоричних неоднозначностей [6], а не всіх анафоричних неоднозначностей, що не дає змогу порівняти результати дослідників з результатами даного дослідження.

Оскільки наступні етапи пов'язані зі звуженням загальної кількості анафоричних неоднозначностей до виключно потенційно «шкідливих» [6], що може призвести до зменшення повноти, лише перші два етапи представляють інтерес для виявлення неоднозначності в експертних формулюваннях.

## Мета дослідження

Підвищення адекватності моделей, на основі яких надається підтримка прийняття рішень, використовуючи автоматичне виявлення анафоричних неоднозначностей в експертних формулюваннях з абсолютною повнотою та задовільною влучністю.

## Запропонований метод

Для вирішення задачі пропонується метод, що не вимагає синтаксичного аналізу та використовує мінімальну необхідну інформацію для знаходження анафор і їхніх антецедентів, досягнувши при цьому абсолютної повноти (100 %) та ненульової влучності згідно з принципом, встановленим Тжон і Беррі [8], відповідно до якого повнота має пріоритет над влучністю. У дослідженні поділяється ця думка, оскільки важливіше відображати усі потенційно неоднозначні формулювання, ніж лише ті, що з високою ймовірністю є неоднозначними. Таким чином, робота експерта буде покращена в тому сенсі, що експерт отримає повний перелік потенційно неоднозначних формулювань, які потрібно переглянути. Якщо зробити пріоритет на влучності, а не повноті, то експерт повинен буде переглянути усі формулювання, оскільки не матиме впевненості, що усі неоднозначні формулювання були відображені.

Подібно до методу Х. Янга та інших [6] спочатку відбувається поділ тексту на формулювання, якщо виявлення буде проводитися в уже наданому переліку формулювань. Далі кожне формулювання поділяється на окремі слова, кожному з яких власний частиномовний аналізатор [4] визначає частину мови та деякі граматичні характеристики. У запропонованому методі необхідними є лише категорії роду та числа іменників. Важливим є використання переліку можливих варіантів вказаних категорій, оскільки кожен з них порівнюється з категоріями анафори, за винятком

універсальних анафор, таких як «що» та «це», та множинних, таких як «які», «вони», «їх» тощо.

На відміну від методу Х. Янга та ін. [6], не відбувається групування слів у іменникові фрази, оскільки іменникові фрази містять ті самі іменники, які слугують антецедентами.

Як і в методі Х. Янга та ін. [6], наступним етапом є знаходження анафор, тобто займенників різного виду («він», «її», «яке» тощо), шляхом повного збігу. Використання неповного збігу в даному випадку є недоцільним, оскільки ключове значення має визначення категорій роду та числа анафори, крім універсальних і множинних анафор.

У розробленому методі відбувається пошук як універсальних анафор («що», «це»), так і множинних («яких», «їхньої», «свої» тощо) та певного роду: чоловічого («який», «він», «той» тощо), жіночого («яка», «вона», «тієї» тощо) та середнього («яке», «воно», «його» тощо). Універсальні та множинні анафори передбачають антецеденти будь-якого числа та роду. Наприклад, у формулюванні «працівники та відвідувачі, які зареєстровані за місяць до заходу» анафора «які» може посилатися як на слово «працівники», так і на слово «відвідувачі».

Оскільки одна анафора у різних родах може збігатись у певних відмінках, анафорі приписується кожен можливий рід. Наприклад, слова «який» чоловічого роду та «яке» середнього роду збігаються у родовому, давальному, орудному та місцевому відмінках, тому анафора може посилатися як на слова чоловічого роду, так і середнього.

На відміну від методу Х. Янга [6], анафорична неоднозначність фіксується лише за наявності щонайменше двох антецедентів, які передують анафорі в межах одного формулювання, оскільки в усіх наявних формулюваннях не виявлено жодного випадку прагматичної анафоричної неоднозначності, тобто посилання анафори на антецедент у попередньому формулюванні.

Відповідність роду та числа анафор і антецедентів впроваджена для підвищення влучності, не втрачаючи при цьому повноту.

Метод також фіксує у кожному формулюванні анафори, число їхніх можливих антецедентів, самі антецеденти та зберігає текст формулювання.

Дано: множина експертних формулювань.

Потрібно визначити: підмножину експертних формулювань, що містять анафоричну неоднозначність.

Послідовність кроків у запропонованому методі така:

1) поділ тексту на формулювання;

2) поділ кожного формулювання на окремі слова;

3) визначення частини мови та деяких граматичних характеристик кожного слова;

4) знаходження анафор шляхом повного збігу;

5) ідентифікація роду (родів) слів, на які може посилатися анафора;

6) пошук антецедентів; перевірка кількості антецедентів і відповідності роду та числа анафор і антецедентів;

7) у випадку наявності щонайменше двох антецедентів, які передують анафорі в межах одного формулювання, фіксується анафорична неоднозначність.

Метод може бути застосовано до інших мов шляхом використання анафор відповідної мови та частиномовної розмітки для потрібної мови.

Метод розроблено на основі дослідження 80,24 % формулювань із 9202 україномовних експертних формулювань, наявних на момент дослідження.

## Порівняльне дослідження

Новизна методу полягає у використанні незначної кількості інформації для виявлення анафоричної неоднозначності у порівнянні з методом Х. Янга та інших.

Маючи на момент проведення дослідження 9202 україномовні формулювання, випробування методу проведено на матеріалі 1818 експертних формулювань, що складають 19,76 % формулювань.

У досліджуваному матеріалі шляхом ручного аналізу кожного речення виявлено 73 формулювання, що містять анафоричну неоднозначність, включаючи ті, які лише формально можуть викликати різночитання. Таким чином, 4,02 % усіх досліджуваних формулювань містять анафоричну неоднозначність. В усіх знайдених випадках антецеденти та анафори знаходяться в межах одного формулювання, що є характерним для типу досліджуваного матеріалу на відміну від багатьох інших стилів і типів тексту.

Весь досліджуваний матеріал містить 159 анафор і 438 антецедентів (наведені кількості можуть бути нерепрезентативними для будь-яких інших типів текстів). Відмінність кількості неоднозначних формулювань від кількості анафор має дві причини. По-перше, деякі анафори мають лише один антецедент, що робить їх однозначними. По-друге, одне формулювання може містити кілька анафор, кожна з яких матиме свої антецеденти, кількість яких може бути великою.

Сам метод був перевірений виконанням такого ж аналізу автоматично, використовуючи імітаційну модель, тобто імітуючи діяльність дослідника.

Реалізувавши метод у вигляді комп'ютерної програми, було отримано 134 експертні формулювання. У дану кількість увійшли усі 73 власноруч визначені формулювання, які містять анафоричну неоднозначність, що задовольняє поставлену мету дослідження — отримати абсолютну повноту $R$:

$$R = \frac{TP}{TP + FN} = \frac{73}{73 + 0} = 100\ \%,$$

де $TP$ — кількість правильно визначених неоднозначних формулювань; $FN$ — кількість хибно визначених однозначних формулювань.

Кількість хибно визначених формулювань, що містять анафоричну неоднозначність, складає 61. Таким чином, влучність $P$ становить:

$$R = \frac{TP}{TP + FP} = \frac{73}{73 + 61} = 0,54477 = 54,48\ \%,$$

де $FP$ — кількість хибно визначених неоднозначних формулювань.

Точність (частка правильно виявлених як однозначних, так і неоднозначних формулювань) $A$ методу складає:

$$A = \frac{TP + TN}{TP + TN + FP + FN} = \frac{73 + 1684}{73 + 1684 + 61 + 0} = 0,9664 = 96,64\ \%,$$

де *TN* — кількість правильно визначених однозначних формулювань.

Невисока влучність порівняно з повнотою та точністю пояснюється тим, що множинні анафори ігнорують рід і число антецедентів, зважаючи лише на їхню кількість. Наприклад, у формулюванні «використання новітніх джерел енергії, крім тих, які шкодять навколишньому середовищу» анафора «які» однозначно посилається на слово «джерел», проте наявність ще одного іменника «енергії» формально задовольняє умову наявності щонайменше двох антецедентів, що призводить до фіксування хибно позитивного неоднозначного формулювання.

## Висновки

Представлено метод виявлення анафоричної неоднозначності в експертних формулюваннях, використовуваних у системах підтримки прийняття рішень. Метод має абсолютну повноту, високу точність і задовільну влучність.

Запропонований метод потребує лише точний перелік анафор для конкретної мови та частиномовний аналізатор, здатний визначити категорії роду та числа іменників, що дає змогу використовувати незначну кількість інформації для виявлення анафоричної неоднозначності.

Метод буде використано з іншими розробками, які виявляють неоднозначність різних типів з метою знизити неоднозначність експертних формулювань у комп'ютерних системах підтримки прийняття рішень.

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., Elazhary H. Implementation of a formal software requirements ambiguity prevention tool. *International journal of advanced computer science and applications (ijacsa)*. 2018. No. 9(8). P. 424–432. URL: http://dx.doi.org/10.14569/ijacsa.2018.090854 (Last accesses: 23.02.2022).

4. Дубок М.Ю., Циганок В.В. Метод частиномовної розмітки на основі квазіфлексій. *Реєстрація, зберігання і оброб. даних*. 2020. Т. 22. № 3. С. 96–106. Doi: 10.35681/1560-9189.2020.22.3.219002.

5. Berry D.M., Kamsties E., Krieger M.M. From contract drafting to software specification: linguistic sources of ambiguity — a handbook. 2003. URL: https://cs.uwaterloo.ca/~dberry/handbook/ambiguityhandbook.pdf (Last accesses: 23.02.2022).

6. Yang H., Willis A., Roeck A.D., Nuseibeh B., Gervasi V. Analysing anaphoric ambiguity in natural language requirements. *Requirements engineering*. 2011 № 16 (3). P. 163–189. Issu 0947–3602. URL: https://doi.org/10.1007/s00766-011-0119-y (Last accesses: 23.02.2022).

7. Yang H., Roeck A.D., Gervasi V., Willis A., Nuseibeh B. Extending nocuous ambiguity analysis for anaphora in natural language requirements. *18th IEEE International requirements engineering conference*. 27 sept.–1 oct. 2010. URL: https://doi.org/10.1109/re.2010.14 (Last accesses: 23.02.2022).

8. Tjong S.F., Berry D.M. The design of sree — a prototype potential ambiguity finder for requirements specifications and lessons learned. Proceedings of the 19th international conference on requirements engineering: foundation for software quality. April 2013. P. 80–95. URL: https://dl.acm.org/doi/10.1007/978-3-642-37422-7_6 (Last accesses: 23.02.2022).

Надійшла до редакції 05.05.2022