



УДК 51.681.3

С.Л. КРИВИЙ

Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: sl.krivoi@gmail.com.

Н.П. ДАРЧУК

Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: nataliadarchuk@gmail.com.

Т.К. СКРИПНИК

Хмельницький національний університет, Хмельницький, Україна,
e-mail: tkskripnik1970@gmail.com.

АНАЛІЗ ПРИРОДНОМОВНИХ ТЕКСТІВ: АНТИМЕМИ, СУПЕРЕЧНІСТЬ, ОНТОЛОГІЇ

Анотація. Розглянуто проблеми аналізу природномовного тексту. Запропоновано підходи до часткового розв'язання проблем перевірки логічної суперечності/несуперечності фактів, добутих із тексту, пошуку антимем та побудови онтологій за результатами лінгвістичного (семантико-синтаксичного) аналізу тексту. Наведено логічний аналіз тексту на основі числення висловлювань та методів обґрунтування висновків у цьому численні.

Ключові слова: антимеми, онтологія, база знань.

ВСТУП

У процесі аналізу природномовного тексту, орієнтованого на перевірку правильності висновків (наслідків) із фактів, добутих з цього тексту, виникає проблема неповноти цих фактів для формального виведення висновків. Ця ситуація є цілком природною, оскільки в повсякденному житті часто буває так, що не всі факти певного міркування наводяться в явному вигляді. Було б помилкою критикувати за непевність наслідків того, хто говорить, що із $A_1, A_2, \dots, A_p$ випливає B , тільки тому, що при цьому він має на увазі посилення $A_{p+1}, \dots, A_m$. Доводи, які ґрунтуються на гіпотезах, що беруть до уваги, але явно не наводять, називаються антимемами. Останні відіграють важливу роль у процесі спілкування між людьми. Без них суттєво уповільнився б обмін думками і цей обмін був би занадто нудним. Існують такі посилення, які очевидні в певному контексті тому, що вони добре відомі і загальноприйняті і їх можна не наводити явно. Коли деяке міркування разом з іншими очевидними посиленнями для виведення наслідків висловлюють не прямо, то використовують натяки.

Отже, формальний логічний аналіз природномовного тексту повинен включати пошук посилення, яких бракує для обґрунтування висновків, у разі неповноти початкової множини посилення. В одних випадках вигляд явних посилення, які потрібно додати, не викликає сумнівів, в інших, — можна по-різному генерувати потрібні посилення $A_{p+1}, \dots, A_m$,¹ намагаючись отримати набір, який достатній для проведення повного доведення. Зрозуміло, що такий набір може бути не єдиним.

¹Зауважимо, що посилення $A_{p+1}, \dots, A_m$ у повсякденному житті є джерелом різного роду пліток.

НЕОБХІДНІ ОЗНАЧЕННЯ ТА ПОНЯТТЯ

Нехай маємо деяку множину фактів $A_1, A_2, \dots, A_p$, які добуті із заданого тексту T . Тоді постановка задачі набуває вигляду: знайти формальне логічне доведення (спростування) того, що із добутих фактів $A_1, A_2, \dots, A_p$ випливає (не випливає) як логічний наслідок факт B .

Обмежимося розглядом фактів, які представлені формулами мови числення висловлювань. З цього обмеження буде видно, у який спосіб отримані результати можна узагальнити на виразніші логічні мови. За такого обмеження для побудови доведення маємо три схеми аксіом числення висловлювань:

- A1) $A \rightarrow (B \rightarrow A)$,
- A2) $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$,
- A3) $(\neg B \rightarrow \neg A) \rightarrow ((\neg B \rightarrow A) \rightarrow B)$,²

і єдине правило виведення *modus ponens* (MP): із істинності посилянь A і $A \rightarrow B$ випливає істинність B .

Отже, розв'язання поставленої задачі зводиться до побудови формального логічного доведення факту B із посилянь $A_1, A_2, \dots, A_p$. Доведення типу «із $A_1, \dots, A_p$ випливає B » означає, що коли хтось говорить, що із $A_1, \dots, A_p$ випливає B , то він не лише хоче сказати, що B є наслідком $A_1, \dots, A_p$ (символьний запис $A_1, \dots, A_p \vdash B$), а й також, що $A_1, \dots, A_p$ істинні (або принаймні ними можна користуватися як істинними). Тому в строгому сенсі $A_1, \dots, A_p \vdash B$ означає, що:

- I) $A_1, \dots, A_p$ істинні,
- II) $A_1, \dots, A_p \vdash B$, а отже, і B істинне.

Таким чином, мета доведення полягає в тому, щоб переконати опонента, що B істинне на підставі пп. I та II.

Оскільки посилення $A_1, A_2, \dots, A_p$ добуті із заданого природномовного тексту T , то істинність чи хибність цих посилянь є питанням довіри або результатом емпіричних даних, або залежності від істинності попередніх гіпотез, на підставі яких отримано право користуватися $A_1, \dots, A_p$ у доведенні. Отже, надійність висновків відносна і залежить від тих критеріїв або мірил, які враховуються в твердженні про істинність висловлювань $A_1, \dots, A_p$. Повне викладення основ, на підставі яких деяке міркування вважається істинним, повинно включати посилення, які в повному обсязі вміщують всю необхідну інформацію. Які існують формальні (або, бодай, формалізовані) методи пошуку міркувань, яких бракує? Розглянемо це питання в наведеному численні.

Оскільки факти можуть мати різні ступені надійності, розіб'ємо розглядувану задачу на дві підзадачі. Для першої підзадачі припускається, що факти $A_1, A_2, \dots, A_p$ вважаються істинними і це не підлягає сумніву. Для другої підзадачі припускається, що істинність фактів сумнівна і вони можуть уточнюватися і змінюватися.

Далі будемо розглядати лише першу підзадачу. Для її розв'язання потрібно формалізувати такі поняття: посилення правдоподібне; посилення неоднозначне; висновок надійний; висновок правильний.

Означення 1. Нехай $P = \{A_1, A_2, \dots, A_p\}$ — множина висловлювань-фактів, добутих із тексту T . Еквівалентним замиканням $E(P)$ множини фактів P називається множина всіх формул, які еквівалентні формулам із множини P . Дедуктивним замиканням $D(E(P))$ еквівалентного замикання множини P називається множина всіх логічних наслідків множини формул із $E(P)$.

Побудова еквівалентного замикання, зокрема, ґрунтується на таких логічних еквівалентностях:

² У деяких аксіоматиках числення висловлювань схема A3 є такою $(\neg B \rightarrow \neg A) \rightarrow (A \rightarrow B)$ [1].

- а) $A \rightarrow (B \rightarrow C) \leftrightarrow B \rightarrow (A \rightarrow C)$ — правило транспозиції;
- б) $\neg\neg B \leftrightarrow B$ — правило подвійного заперечення;
- в) $A \rightarrow (\neg A \rightarrow B) \leftrightarrow A \rightarrow (A \vee B)$ — правило диз'юнкції;
- г) $(\neg B \rightarrow \neg A) \leftrightarrow (A \rightarrow B)$ — правило контрапозиції;
- д) $A \vee B \leftrightarrow \neg(\neg A \wedge \neg B)$ — правило де Моргана;
- е) $A \wedge B \leftrightarrow \neg(\neg A \vee \neg B)$ — правило де Моргана.

Проілюструємо це означення прикладом.

Приклад 1.1. Нехай $P = \{p \rightarrow \neg q\}$, тоді еквівалентне замикання $E(P) = \{p \rightarrow \neg q, q \rightarrow \neg p\}$ отримуємо на підставі правила контрапозиції. Дедуктивне замикання $D(E(P))$ у цьому випадку збігається з еквівалентним замиканням $E(P)$, оскільки жодний логічний наслідок не виводиться з множини $E(P)$ за правилом виведення МР.

Приклад 1.2. Нехай $P = \{q \rightarrow p, \neg q \rightarrow (c \rightarrow \neg p)\}$, тоді, застосовуючи еквівалентності а)–е), отримуємо:

$$E(P) = \{q \rightarrow p, \neg q \rightarrow (c \rightarrow \neg p), \neg p \rightarrow \neg q, \neg(c \rightarrow \neg p) \rightarrow q, \\ \neg q \rightarrow (p \rightarrow \neg c), \neg(p \rightarrow \neg c) \rightarrow q\}.$$

Застосовуючи правило виведення МР, транзитивність імплікації, правила контрапозиції та транспозиції, одержуємо таке дедуктивне замикання множини $E(P)$:

$$D(E(P)) = E(P) \cup \{\neg p \rightarrow (p \rightarrow \neg c), p \rightarrow (\neg p \rightarrow \neg c), p \rightarrow (c \rightarrow p), \\ q \rightarrow (\neg p \rightarrow \neg c), q \rightarrow (c \rightarrow p)\}.$$

ПОШУК АНТИМЕМ

Розглянемо, як можна шукати антими, користуючись введеними замиканнями. Як впливає з викладеного, в процесі пошуку фактів-посилань, яких бракує, генеруються проміжні факти $A_{p+1}, \dots, A_m$. Виникає запитання: якими з цих фактів слід поповнювати множину посилань $P = \{A_1, \dots, A_p\}$? Очевидна відповідь є такою: потрібно поповнювати множину P такими фактами, які не спричиняють суперечності.

Розглянемо приклад, який є продовженням прикладів 1.1 та 1.2.

Приклад 2.1. Для множини $P = \{p \rightarrow \neg q\}$ було побудовано таке еквівалентне замикання $E(P) = \{p \rightarrow \neg q, q \rightarrow \neg p\}$, дедуктивне замикання якого збігається з його еквівалентним замиканням.

Нехай потрібно перевірити, чи є формула $\neg p$ наслідком множини формул $E(P)$. Очевидно, що з $E(P)$ вивести $\neg p$ неможливо, тому що МР немає до чого застосувати. Якого факту не вистачає для доведення? У цьому випадку підказка міститься у множині $E(P)$, а точніше — у формулі $q \rightarrow \neg p$. Ця формула показує, що для застосування правила виведення МР для отримання $\neg p$ бракує посилання q . Перед тим як додати q до $C(P)$, потрібно перевірити несуперечність отриманої множини $C(P) \cup \{q\} = \{p \rightarrow \neg q, q \rightarrow \neg p, q\}$. (Про таку перевірку йдеться далі.) Неважко перекоонатися, що ця множина несуперечна. Отже, якщо цю формулу додати до посилань, то отримаємо таке формальне доведення:

- 1) $q \rightarrow \neg p$ — факт;
- 2) q — факт;
- 3) $\neg p$ — наслідок.

Підказкою стала та обставина, що серед формул $E(P)$ виявилася формула $\neg p$, яку потрібно було довести. Отже, еквівалентне замикання може містити інформацію про відсутні посилання.

Приклад 2.2. Для множини $P = \{q \rightarrow p, \neg q \rightarrow (c \rightarrow \neg p)\}$ було побудовано замикання множини P :

$$\begin{aligned} E(P) &= \{q \rightarrow p, \neg q \rightarrow (c \rightarrow \neg p), \neg p \rightarrow \neg q, \neg(c \rightarrow \neg p) \rightarrow q, \\ &\quad \neg q \rightarrow (p \rightarrow \neg c), \neg(p \rightarrow \neg c) \rightarrow q\}, \\ D(E(P)) &= E(P) \cup \{\neg p \rightarrow (p \rightarrow \neg c), p \rightarrow (\neg p \rightarrow \neg c), p \rightarrow (c \rightarrow p), \\ &\quad q \rightarrow (\neg p \rightarrow \neg c), q \rightarrow (c \rightarrow p)\}. \end{aligned}$$

Припустимо, що потрібно перевірити справедливість наслідку $c \rightarrow p \leftrightarrow \neg c \vee p$ з наведених посилань. Використаємо для цієї перевірки метод резолюцій, який має певні переваги над методом виведення з аксіом.

Неважко переконатися, що справедливість наслідку не виводиться методом резолюцій із множини фактів P . Записавши множину формул із P у вигляді фактів-гіпотез та додавши до отриманої множини заперечення наслідку, тобто множину диз'юнктивів c і $\neg p$, отримаємо таке виведення:

- 1) $\neg q \vee p$ — факт;
- 2) $q \vee \neg c \vee \neg p$ — факт;
- 3) c — факт;
- 4) $\neg p$ — факт;
- 5) $\neg q$ — пп. 1, 4;
- 6) $\neg c \vee \neg p$ — пп. 5, 2;
- 7) $\neg p$ — пп. 3, 6.

Для отримання порожнього диз'юнкта в цьому виведенні бракує посилань. Але з нього випливає, що множину P можна поповнити двома способами так, щоб вона була несуперечною: додати посилання p або додати посилання q . Якщо додати посилання p , то відразу отримаємо суперечну множину, оскільки серед диз'юнктивів є факт $\neg p$. Якщо поповнимо множину P фактом q , то поповнена у такий спосіб множина посилань P дає змогу отримати таку множину фактів:

- 1) $\neg q \vee p$ — факт;
- 2) $q \vee \neg c \vee \neg p$ — факт;
- 3) c — факт;
- 4) $\neg p$ — факт;
- 5) q — факт.

Звідси одержуємо виведення порожнього диз'юнкта:

- 6) p — пп. 5, 1;
- 7) $\emptyset$ — пп. 4, 6.

Отже, формула $c \rightarrow p$ дійсно є наслідком несуперечної множини фактів $P' = \{q \rightarrow p, \neg q \rightarrow (c \rightarrow \neg p), q\} = P \cup \{q\}$ або множини $P'' = \{q \rightarrow p, \neg q \rightarrow (c \rightarrow \neg p), p\} = P \cup \{p\}$.

Цей приклад показує, що посилання, яких бракує, можуть міститися в еквівалентному замиканні заданої множини фактів, а тим більше, — в її дедуктивному замиканні.

Але в цьому разі виникає низка запитань:

- 1) як можна виконати перевірку несуперечності заданої множини формул та несуперечності її поповнення?
- 2) якими фактами поповнювати задану множину фактів, якщо кандидатів декілька?
- 3) як перевірити узгодженість поповненої множини фактів з синтактико-семантичним сенсом тексту T .

Відповідь на запитання 1 впливає з несуперечності і повноти числення висловлювань. На підставі цієї властивості маємо: множина фактів несуперечна тоді і тільки тоді, коли вона має модель. Дійсно, в прикладі 2.1 поповнена множина має модель $p=0, q=1$, а в прикладі 2.2 поповнення множини P фактом q (як і фактом p) має дві моделі: $p=q=1, c=0$ та $p=q=1, c=1$. Отже, таке поповнення має право на існування.

Відповідь на запитання 2 не однозначна, оскільки вибір може бути недетермінованим. Але, як показують приклади, слід вибирати додаткові факти найменшої можливої довжини. У прикладах 2.1 та 2.2 додаткові факти (диз'юнкти) мали довжину 1. У тому випадку, коли додаткових фактів довжини 1 немає, перевіряються факти найменшої можливої довжини і, якщо вони не спричиняють суперечності, то ними поповнюються наявні факти. Цього достатньо для розв'язання першої підзадачі, а під час розв'язання другої підзадачі потрібно провести додатковий аналіз для визначення причин суперечності і їхнього вилучення.

Відповідь на запитання 3 є значною мірою лінгвістичним аспектом аналізу, де застосовуються системи синтактико-семантичного аналізу природномовних текстів. Про деякі елементи цього аналізу йдеться далі.

Інколи початковий текст T може мати вигляд, як у наступному прикладі. Цей текст потрібно тільки спростити, тобто сформулювати сенс найлаконічніше. Для такого типу текстів розглядувана задача спрощується, оскільки вони містять надлишок інформації.

Приклад 3. В університеті існує клуб з такими правилами:

- 1) члени наглядової ради університету повинні вибиратися з-поміж членів фінансового комітету;
- 2) не можна бути одночасно членом великої вченої ради і членом наглядової ради, не будучи членом фінансового комітету;
- 3) жоден член великої вченої ради не може бути членом фінансового комітету.

Потрібно спростити правила.

Розв'язання. Нехай x — довільна особа, яка є членом клубу, тоді P означає « x є членом фінансового комітету», Q — « x є членом наглядової ради університету», а R — « x є членом великої вченої ради університету». Тоді правила 1–3 можна записати так:

$$(P \rightarrow Q) \wedge (\neg(Q \wedge R) \vee P) \wedge \neg(R \wedge P).$$

Перетворюючи цю формулу в кон'юнктивну нормальну форму, отримаємо таку еквівалентну формулу:

$$(Q \rightarrow \neg R) \wedge (P \rightarrow Q).$$

Отже, правилами насправді є правило 1 і дещо змінене правило 2: «жоден член великої вченої ради не може бути членом наглядової ради університету».

Для таких не зовсім простих текстів запропонований підхід є застосовним, до того ж без суттєвих обмежень. Потрібно лише коректно перевести на формальну мову задані факти і застосувати еквівалентні перетворення отриманих фактів, орієнтованих на їхнє спрощення.

Розглянемо деякі формальні означення необхідних понять.

Означення 2. Висновок B називається правильним відносно множини формул P , якщо B є наслідком множини формул P . Інакше кажучи, існує доведення формули B із множини формул P у такому численні. Посилання A називається правдоподібним відносно множини формул P , якщо множина $P \cup \{A\}$ несупе-

речна. Висновок B називається надійним відносно множини формул P , якщо виконуються умови пп. I та II. Інакше кажучи, якщо ці умови виконуються, то говорять, що міркування про факт B на підставі множини фактів P не тільки правильне, але й надійне.

З наведеної відповіді на запитання 2 не випливає загального правила пошуку антимем. У загальному випадку це запитання формулюється так: «як можна розв'язати проблему антимем у системах аналізу природномовних текстів?». Часто в доведеннях, які не можна закінчити за браком інформації, не завжди наявні підказки про додаткові факти. Тому єдиним способом відновлення потрібних антимем має бути використання систем штучного інтелекту, в яких міститься певна додаткова інформація. Таку інформацію повинні включати бази знань (БЗ). Але, чи можна забезпечити повноту інформації БЗ? Відповіді на це запитання наразі не існує.

Отже, очевидно, що для відновлення антимем потрібно вводити проміжні посилання $A_{p+1}, \dots, A_m$ і називати висновок B правдоподібним, якщо він правильний, а $A_{p+1}, \dots, A_m$ правдоподібні. І знову виникає запитання: «який ступінь правдоподібності отриманих висновків?». Відповідь на нього можна отримати за допомогою n -значної логіки (зокрема, 3-значної) і P -стилю виведення [2]. Наприклад, якщо перше посилання $q \rightarrow \neg p$ з прикладу 1.1 градувати значенням 1 (тобто, що воно істинне), а друге посилання q градувати значенням $1/2$, то отримаємо висновок $\neg p$ зі значенням $1/2$. Це означає, що ступінь впевненості в істинності висновку лише 50 %. А коли перше посилання градуйоване значенням $1/2$, то висновок буде мати значення 0 на підставі правил градування МР.

Означення 3. Факт B називається істинним ступеня істинності k , якщо він має виведення із множини посилань P з градуйованим значенням k . Факт B називається логічним наслідком множини фактів P у 3-значній логіці, якщо не існує інтерпретації, за якої формули із P набувають значення 1 або $1/2$, а B набуває значення 0.

У 3-значній логіці факти, які ніколи не набувають значення 0 (хибності), називаються квазітавтологіями, тобто є посиланнями, правдоподібність яких не менша 50 %.

У тому випадку, коли міркування сформульовано чітко, неважко впевнитися в правильності кожного окремого його етапу. Для перекладу на формальну логічну мову потрібно вилучення всіх нечітких формулювань або багатозначності природної мови. Таким чином, проблемою є правильність перекладу понять і фактів, які містить природномовний текст, на логічну мову. Вичерпної відповіді щодо способів розв'язання цієї проблеми наразі немає, але можна зазначити деякі її аспекти.

Задача перекладу спрощується, якщо проведений синтактико-семантичний аналіз тексту T і його результати представлені у вигляді анотованого тексту [3]. Крім того, від цього аналізу залежить відповідь на запитання 3: «як додаткові факти впливають на зміст аналізованого природномовного початкового тексту T ?».

Лінгвістичний аналіз тексту та його результати. Розглянемо задачу аналізу і перекладу природномовного тексту з лінгвістичної точки зору. Аналіз природномовних текстів представлено двома компонентами: лінгвістичним і логічним.

Лінгвістичний компонент системи має такий склад: лінгвістичне опрацювання вхідного тексту на рівні морфологічного і синтаксичного аналізу, керувального алгоритму, який здійснює всю стратегію автоматичного оброблення

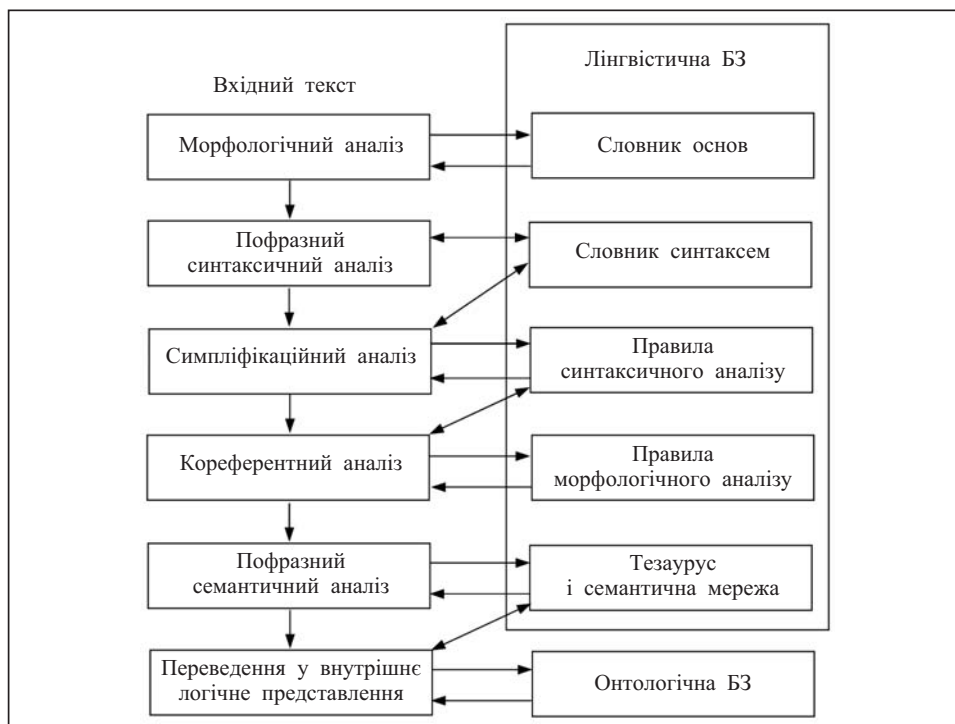


Рис. 1. Структура лінгвістичного і логічного оброблення природномовного тексту

тексту; правила симпліфікації тексту, за якими спрощується складна синтаксична структура з сурядністю і підрядністю до простих речень, позбавлених вставних слів і конструкцій, прикладок; правила кореферентності з відновленням референта (рис. 1).

Семантичний аналіз тексту має такий склад: метамова семантичних структур, якою перекладаються всі джерельні тексти зі спрощеною структурою (після етапів симпліфікації і встановлення кореферентних відношень); тезаурус онтологічних класів концептів, вербалізованих словами-термами певної галузі знань і словником відношень між ними. Метамова представлення структур — це логічна мова-посередник, головною одиницею якої є формула $P(A, B)$ як елементарне висловлювання. Про одиницю A повідомляється, що вона є у відношенні P до одиниці B . Приклади семантичних відношень: ПРИЧИНА; ЛОКАЛІЗАЦІЯ; ЧАС тощо. Посередником вона названа тому, що є мовою спілкування між текстами, частинами текстів, між текстом й адресатом, текстом і різними предметними областями, аналізованим текстом і БЗ. Кожна лексична одиниця, яка входить до БЗ, одержує опис на морфологічному, синтаксичному, семантичному рівнях у вигляді відповідного набору характеристик. Синтаксичні й морфологічні характеристики потрібні для визначення значення текстової словоформи і встановлення синтаксичної функції у реченні. Семантичні характеристики описують значення лексичної одиниці.

Уведення даних тезаурусного типу у словникову базу зумовлене тим, що під час здійснення автоматичного аналізу тексту конче потрібно залучення не тільки мовних, а й позамовних знань. Через це в сучасних системах дедалі частіше з'являється компонент, який називається онтологічним, тому що основою будь-якої семантичної класифікації одиниць є класифікація предметів і явищ дійсності [4].

Для пояснення деталей розглянемо приклад.

Приклад 4. Проаналізуємо текст і проілюструємо етапи аналізу природномовного тексту такого юридичного змісту:

T = «Казначейство здійснює платежі на оплату придбання медикаментів та перев'язувальних матеріалів; придбання води питної (зокрема, бутильованої); забезпечення продуктами харчування; надання підтримки внутрішньо переміщеним та/або евакуйованим особам, придбання мийних та дезінфекційних засобів (зокрема, для закладів соціального захисту); оплату ритуальних послуг (зокрема послуг з перевезення, поховання); оплату послуг із благоустрою населених пунктів (зокрема, оплату енергоносіїв та інших комунальних послуг, охорону об'єктів благоустрою), оплату експлуатаційних послуг, пов'язаних з утриманням будинків і споруд та прибудинкових територій (зокрема, вивіз сміття)». (Фрагмент Постанови Кабінету Міністрів України № 346 від 21.03.2022.)

На рис. 2 представлено результати лінгвістичного аналізу наведеного тексту *T* з таблицею підпорядкування, яку програма лінгвістичного аналізу будує автоматично за правилами комп'ютерної граматики АГАТ, розробленої для українськомовного тексту в лабораторії комп'ютерної лінгвістики Навчально-наукового центру Київського національного університету імені Тараса Шевченка. Ця програма виконує розмітку (кодування) вхідного тексту так, як це показано на рис. 2. Тут представлено тегування словоформ тексту кодами онтологічних класів (латинські літери) відповідно до концептів, розроблених в корпусі української мови для кожної частини мови. Кожну словоформу речення або словосполучення можна замінити кодом онтологічного класу, і навпаки, онтологічний клас представити набором слів, що вербалізують його.

Казначейство (ЛИ)	51), <i>t0inst</i> , Державний фінансовий орган, що відає касовим виконанням державного бюджету
здійснює (ГЯ)	150), Виконувати, робити і таке інше що-небудь
платежі (ЙУ)	1), <i>t0abst0finance</i> , Розрахунок за працю, за щось придбане
на (ПВ)	61), <i>j0prep0obs0dlj</i> , Обставинні відношення, уживається для позначення <i>c</i>
оплату (КВ)	51), <i>t0activity,der0v</i> , Дія за значенням оплатити, оплачувати
придбання (ЛР)	51), <i>t0activity,der0v</i> , Дія за значенням придбати, придбавати
медикаментів (ЙЕ)	51), <i>t0stuff</i> , Лікувальні засоби, ліки
та (СС)	51), <i>s0scnj0sur0jed</i> , Єднальний, уживається для приєднання двох рівноправних синтаксичних
перев'язувальних (АЕ)	51), <i>r0rel,dr0dlj</i> , Те саме, що перев'язний
матеріалів (ЙЕ)	1), <i>t0stuff</i> , Те саме, що сировина
придбання (ЛР)	51), <i>t0activity,der0v</i> , Дія за значенням придбати, придбавати
води (КР)	150), Прозора безбарвна рідина, що становить найпростішу хімічну сполуку
питної (АЗ)	51), <i>r0rel,dr0lmen</i> , Придатний для вживання, придатний для пиття
зокрема (НО)	150), Вживається для підкреслення, виділення чого-небудь з-поміж однотипного
бутильованої (АЗ)	150), Розфасована у пляшках, природна, мінеральна
забезпечення (ЛР)	51), <i>t0activity,der0v</i> , Дія за значенням створювати надійні умови для здійснення чого-небудь, гарантувати щось
продуктами харчування (ЙЮ)	5), <i>class</i> , (множина) їстівні припаси, харчі
надання (ЛР)	51), <i>t0activity,der0v</i> , Дія за значенням надати послуги
.....	

Рис. 2. Фрагмент семантичного аналізу тексту

Таблиця тексту T містить морфологічну анотацію, інформацію про частини мови й категорійні характеристики слів. Таблиця має три стовпчики: у першому — головний член бінарної сполуки («хазяїн»), у другому — підпорядкований їй член сполуки («слуга»), а в третьому — синтаксична інформація про тип синтаксичного зв'язку сполуки. Це уможливило після завершення роботи програми утворення алфавітно-частотного словника словосполучень тексту. Вся ця інформація складає лінгвістичну БЗ, яка є основою для виконання логічного аналізу тексту та побудови онтологічної БЗ.

На рис. 3 наведено фрагмент автоматичної побудови словосполучень, а на рис. 4 — граф, що зображує залежності між об'єктами цього тексту.

Використовуючи результати лінгвістичного аналізу наведеного тексту T , представлені залежностями, наведеними на рис. 4, знаходимо такий універсум D ,

здійснює	казначейство	координаційний зв'язок
здійснює	платежі	дієслівна безприйменикова сполука
здійснює	придбання	дієслівна безприйменикова сполука
здійснює	надання	дієслівна безприйменикова сполука
здійснює	забезпечення	дієслівна безприйменикова сполука
платежі	на	іменникова прийменникова сполука
на	оплату	прийменикова сполука
на	придбання	прийменикова сполука
на	оплату	прийменикова сполука

Рис. 3. Фрагмент автоматичної побудови словосполучень

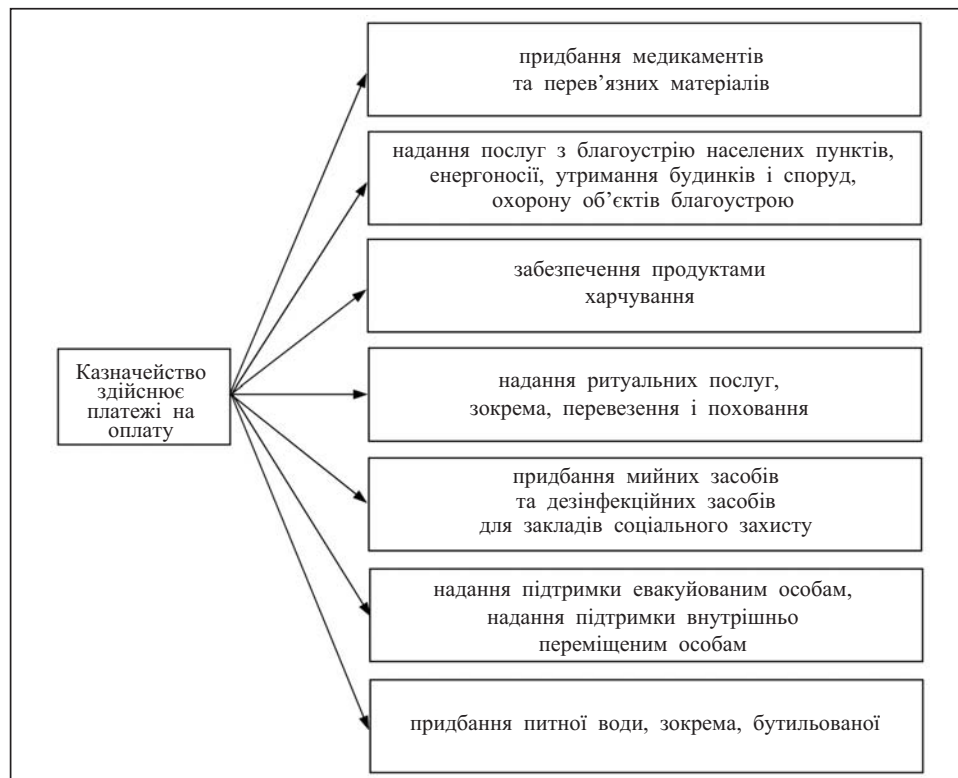


Рис. 4. Граф залежностей між об'єктами

над яким буде виконуватися логічний аналіз:

$$D = \{K\} \cup D_1 \cup D_2 \cup D_3 \cup D_4 \cup D_{41} \cup D_5 \cup D_6 \cup D_7 \cup D_{71} \cup D_8 \cup D_9 \cup D_{10}, D_{11},$$

де

$\{K\}$ — одноелементна множина, яка означає «Казначейство»;

$D_1 = \{m = \text{медикаменти}, m_1 = \text{перев'язні-матеріали}\};$

$D_2 = \{m_1 = \text{перев'язні-матеріали}\};$

$D_3 = \{x_1 = \text{харчові-продукти}, x_2 = \text{питна-вода}, v_1 = \text{вода-бутильована}\};$

$D_4 = \{v_1 = \text{вода-бутильована}\};$

$D_{41} = \{o_1 = \text{внутрішньо-переміщені-особи}, o_2 = \text{евакуйовані-особи}\};$

$D_5 = \{mz = \text{мийні-засоби}\};$

$D_6 = \{d = \text{дезінфекційні-засоби}\};$

$D_7 = \{p_1 = \text{ритуальні-послуги} (p_{71} = \text{перевезення}, p_{72} = \text{поховання})\};$

$D_{71} = \{p_{71} = \text{перевезення}, p_{72} = \text{поховання}\};$

$D_8 = \{e = \text{благоустрій-населених-пунктів} (e_1 = \text{енергоносії}, e_2 = \text{інші-комунальні-послуги}, e_3 = \text{охорона-об'єктів-благоустрою})\};$

$D_9 = \{s_1 = \text{утримання-будинків}, s_2 = \text{утримання-споруд}, s_3 = \text{утримання-прибудинкових-територій} (s_4 = \text{вивіз-сміття})\};$

$D_{10} = \{c_1 = \text{дезінфекційні засоби для закладів-соціального-захисту}\};$

$D_{11} = \{s_4 = \text{вивіз-сміття}\}.$

Нехай КЗ(дія, об'єкт) — бінарне відношення (на лінгвістичному рівні це відношення підмет–присудок), яке означає, що «Казначейство здійснює платежі», тобто деяку дію по відношенню до об'єкта. Отже, маємо бінарні відношення, які діють на множині *actions* і об'єктах області $D^f = \Delta$.

З наведеного прикладу тексту T отримуємо таку множину дій:

$$actions = \{a_1 = \text{придбання}, a_2 = \text{забезпечення}, a_3 = \text{надання}\},$$

а також множину бінарних відношень $R = \{KZ(a_i, v_j)\}$, де $v_j \in D$,

$$\begin{aligned} RN = \{ & KZ(a_1, m), KZ(a_1, m_1), KZ(a_2, x_1), KZ(a_1, x_2), KZ(a_1, x_3), \\ & KZ(a_3, o_1), KZ(a_3, o_2), KZ(a_1, mz), KZ(a_1, d), KZ(a_3, p_1), KZ(a_3, p_2), \\ & KZ(a_3, p_3), KZ(a_1, e), KZ(a_1, e_1), KZ(a_1, e_2), KZ(a_1, s_1), KZ(a_1, s_2), \\ & KZ(a_1, s_3), KZ(a_1, s_4), KZ(a_1, s_5), KZ(a_1, s_6), KZ(a_1, c_1)\}. \end{aligned}$$

ПОБУДОВА ОНТОЛОГІЙ

За означенням, онтологією називають упорядковану трійку [5]:

$$O = (D, \mathcal{R}, f),$$

де D — скінченна множина концептів (універсум); $\mathcal{R}$ — скінченна множина (семантичних) відношень, визначених на D ; f — функція інтерпретації на деякій області Δ елементів із X і $\mathcal{R}$, тобто $f : D \cup \mathcal{R} \rightarrow \Delta$. Причому множини концептів і відношень поділяються на атомарні (CN і DN) і похідні, які будуються з атомарних за допомогою конструкторів.

Із викладеного випливає, що універсум (з лінгвістичної точки зору це словники і природномовні тексти) є підставою для побудови онтології.

Дійсно, області D_j , $j=1, \dots, 11$, можна трактувати як атомарні концепти, а бінарні відношення $KZ(a_i, v_j)$ — як множину атомарних відношень. Важливим фактом у цьому випадку є те, що синтактико-семантичний аналіз дає змогу структурувати концепти на атомарні та похідні. У наведеному прикладі це про-

являється в тому, що деякі концепти виокремлюються з множини однотипних об'єктів (за допомогою слів «зокрема», та «а також» і т.п.).

Далі побудова онтології зводиться до формулювання термінологічних аксіом і фактів, які записуються з використанням цієї термінології.

Приклад 5. Розглянемо елементи знайденого універсуму як атомарні концепти:

$CN = \{m = \text{медикаменти}, m_1 = \text{перев'язні-матеріали},$
 $x_1 = \text{харчові-продукти}, x_2 = \text{питна-вода}, x_3 = \text{вода-бутильована},$
 $o_1 = \text{внутрішньо-переміщені-особи}, o_2 = \text{евакуйовані-особи},$
 $mz = \text{мийні-засоби}, d = \text{дезінфекційні-засоби},$
 $p_1 = \text{ритуальні-послуги}, p_2 = \text{перевезення}, p_3 = \text{поховання},$
 $e = \text{благоустрій-населених-пунктів}, e_1 = \text{енергоносії},$
 $e_2 = \text{інші-комунальні-послуги},$
 $s_1 = \text{охорона-об'єктів-благоустрою}, s_2 = \text{утримання-будинків},$
 $s_3 = \text{утримання-споруд},$
 $s_4 = \text{утримання-прибудинкових-територій},$
 $s_5 = \text{вивіз-сміття}, s_6 = \text{експлуатаційні-послуги},$
 $c_1 = \text{дезінфекційні-засоби-закладів-соціального-захисту}\}.$

У прийнятих позначеннях атомарні відношення набувають вигляду:

$RN = \{K3(a_1, m), K3(a_1, m_1), K3(a_2, x_1), K3(a_1, x_2), K3(a_1, x_3),$
 $K3(a_3, o_1), K3(a_3, o_2), K3(a_1, mz), K3(a_1, d), K3(a_3, p_1), K3(a_3, p_2),$
 $K3(a_3, p_3), K3(a_1, e), K3(a_1, e_1), K3(a_1, e_2), K3(a_1, s_1), K3(a_1, s_2),$
 $K3(a_1, s_3), K3(a_1, s_4), K3(a_1, s_5), K3(a_1, s_6), K3(a_1, c_1)\}.$

Термінологічні аксіоми $TBox$:

медпрепарати $\equiv$ медикаменти $\cup$ перев'язні-матеріали
 вода $\equiv$ питна-вода $\cup$ бутильована
 біженці $\equiv$ внутрішньо-переміщені-особи $\cup$ евакуйовані-особи
 дезінфекційні-засоби $\equiv$ дезінфекційні-засоби $\cup$ дезінфекційні-засоби-закладів-соціального-захисту
 пральні-засоби $\equiv$ мийні-засоби $\cup$ дезінфекційні-засоби
 ритуальні-послуги $\equiv$ перевезення $\cup$ поховання
 комунальні-послуги $\equiv$ енергоносії $\cup$ експлуатаційні послуги $\cup$ інші-послуги
 благоустрій $\equiv$ охорона-об'єктів-благоустрою $\cup$ утримання-будинків $\cup$ утримання-споруд $\cup$ утримання-прибудинкових-територій

Нехай задана інтерпретація f на універсумі $D^f = \Delta$, яка має такий вигляд:

медикаменти $\equiv \{m_1 = \text{серцеві}, m_2 = \text{антибіотики}, m_3 = \text{діабетичні}\}$
 перев'язні-матеріали $\equiv \{p_1 = \text{вата}, p_2 = \text{бинт}, p_3 = \text{пластир}, p_4 = \text{жгут}\}$
 харчові-продукти $\equiv \{x_1 = \text{хліб}, x_2 = \text{м'ясо}, x_3 = \text{молоко}, x_4 = \text{риба},$
 $x_5 = \text{чай}, v_1 = \text{вода-бутильована}\}$
 питна-вода $\equiv \{v_1 = \text{вода-бутильована}, v_2 = \text{вода-газована},$
 $v_3 = \text{вода-мінеральна}\}$
 мийні-засоби $\equiv \{mz_1 = \text{мило-туалетне}, mz_2 = \text{мило-господарче},$
 $mz_3 = \text{мило-рідке}, mz_4 = \text{шампунь},$
 $mz_5 = \text{пральний-порошок}\}$

дезінфекційні-засоби $\equiv \{d_1 = \text{спирт}, d_2 = \text{одеколон}, d_3 = \text{йод}, d_4 = \text{хлор}, d_5 = \text{перекис водню}\}$
 ритуальні-послуги $\equiv \{p_1 = \text{перевезення}, p_2 = \text{поховання}, p_3 = \text{розпорядник}, p_4 = \text{обслуговування}\}$
 внутрішньо-переміщені-особи $\equiv \{oc_1 = \text{Петренко}, oc_2 = \text{Степаненко}, oc_3 = \text{Харченко}, oc_4 = \text{Остапенко}\}$
 евакуйовані-особи $\equiv \{ec_1 = \text{Верченко}, ec_2 = \text{Степовий}, ec_3 = \text{Харламов}, ec_4 = \text{Потапенко}\}$
 ритуальні-послуги $\equiv \{p_1 = \text{послуги-моргу}, p_2 = \text{перевезення}, p_3 = \text{поховання}, p_4 = \text{обслуговування}\}$
 комунальні-послуги $\equiv \{e = \text{благоустрій}, e_1 = \text{енергоносії}, e_2 = \text{інші-комунальні-послуги}\}$
 охорона $\equiv \{s_1 = \text{охорона-об'єктів-благоустрою}, s_2 = \text{утримання-будинків}, s_3 = \text{утримання-споруд}, s_4 = \text{вивіз-сміття}\}$
 інші-комунальні-послуги $\equiv \{\text{водопостачання, обслуговування водолічильників, електропостачання}\}$
 дезінфекційні-засоби $\equiv \{d_3 = \text{йод}, d_4 = \text{хлор}, d_5 = \text{перекис водню}\}$
 -закладів-соціального захисту
 благоустрій $\equiv \text{охорона-об'єктів-благоустрою} \cup \text{утримання-будинків} \cup \text{утримання-споруд} \cup \text{утримання-прибудинкових-територій}$

Факти $ABox$:

$$\begin{aligned}
 &K3(a_1, m) \rightarrow K3(a_1, m_1), \quad K3(a_2, x_1) \rightarrow K3(a_2, x_2) \wedge K3(a_2, x_3), \\
 &K3(a_3, o_1) \rightarrow K3(a_3, o_2), \quad K3(a_1, mz) \rightarrow K3(a_1, c_1), \\
 &K3(a_1, p_1) \rightarrow K3(a_1, p_2) \wedge K3(a_1, p_3), \quad K3(a_1, d) \rightarrow K3(a_1, d_3) \wedge K3(a_1, d_4), \\
 &K3(a_1, c_1) \rightarrow K3(a_1, d_3) \wedge K3(a_1, d_4) \wedge K3(a_1, d_5), \\
 &K3(a_1, e) \rightarrow K3(a_1, e_1) \wedge K3(a_1, e_2), \\
 &K3(a_3, s_1) \wedge K3(a_3, s_2) \wedge K3(a_3, s_3) \wedge K3(a_3, s_4).
 \end{aligned}$$

Застосовуючи метод резолюцій, неважко переконатися в тому, що наведені факти не є суперечними. Для цього потрібно побудувати онтологію цієї предметної області. Дійсно, побудувавши онтологію, як прототип БЗ, можна перевірити отримані факти на суперечність/несуперечність (використовуючи деякий стандартний або побудований ризонер) та формулювати запити до БЗ, наприклад, такого типу:

- 1) «Якими Казначейство забезпечує медпрепаратами?»;
- 2) «Які Казначейство закуповує дезінфекційні засоби для закладів соцзахисту?»;
- 3) «Чи Казначейство оплачує охорону об'єктів благоустрою і вивіз сміття?»;
- 4) «Чи можна придбати через Казначейство автомобіль?».

У наведених раніше позначеннях ці запити набувають вигляду:

- 1) «Print(D_1)», де Print — оператор друку вмісту D_1 ;
- 2) «Print(D_{10})»;
- 3) « $K3(a_2, s_2) \wedge K3(a_2, s_4)$?»;
- 4) « $K3(a_1, a)$ », де a — автомобіль.

У результаті генерації відповіді на запити отримуємо:

- 1) « $\{m_1 = \text{серцеві}, m_2 = \text{антибіотики}, m_3 = \text{діабетичні}, m_4 = \text{перев'язні-матеріали}\}$ »;
- 2) « $\{d_1 = \text{спирт}, d_2 = \text{одеколон}, d_3 = \text{йод}, d_4 = \text{хлор}, d_5 = \text{перекис водню}\}$ »;
- 3) «Так, оплачує»;
- 4) «Невідомо, що таке a ?».

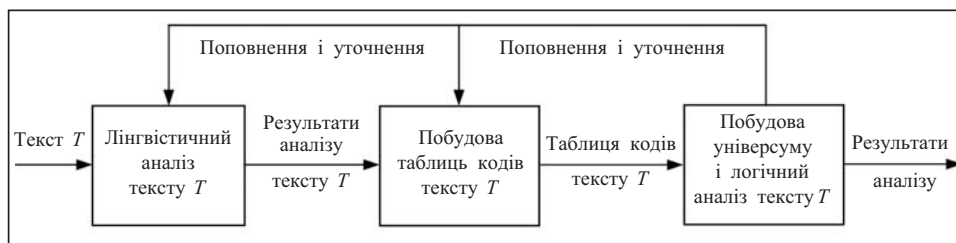


Рис. 5. Етапи логіко-лінгвістичного аналізу тексту T

З відповіді на останній запит випливає, що потрібно поповнити посилання і висновки новими фактами, а саме фактами про об'єкти, яких у цьому універсумі немає.

Найпростішим способом поповнити універсум є відповідь на всі запити, які не стосуються об'єктів універсума, такого типу: «Якщо об'єкт або відношення відсутнє в універсумі, то згенерувати відповідь НІ».

На рис. 5 наведено схему побудови моделі вхідного тексту T (у вигляді онтології) та перевірки коректності поповнень множини фактів новими фактами.

За етапами логіко-лінгвістичного аналізу, представленими на рис. 5, формується алгоритм роботи всієї системи в цілому.

Логіко-лінгвістичний аналіз (T)

Вхід: Початковий текст T .

Вихід: Результати запитів до БЗ тексту T .

Метод:

початок

1. Ввести початковий текст T ;
2. Виконати синтактико-семантичний аналіз T ;
3. За результатами аналізу тексту T побудувати таблицю(і) кодів класів тексту T ;
4. За таблицею(ми) кодів побудувати універсум D для тексту T ;
5. Виконати логічний аналіз універсуму.
 - 5.1. Якщо результати аналізу потребують поповнення (антимемами), то внести поповнення і перейти до п. 2, інакше перевірити отримані факти на суперечність.
 - 5.2. Якщо факти несуперечні, то занести їх у БЗ і згенерувати відповіді на запити до цієї БЗ.

кінець

Пояснимо деякі деталі щодо кроків наведеного алгоритму. Почнемо з розгляду побудови універсуму за таблицею кодів.

У результаті синтактико-семантичного аналізу отримано такий фрагмент таблиці цього аналізу (табл. 1): хазяїном вибрано «Казначейство».

Таблиця 1. Фрагмент таблиці з кодами синтаксичних класів

хазяїн	слуга	синтаксичний зв'язок	відношення	семантичний клас хазяїна	семантичний клас служ
охорону	об'єктів	—	—	$t0activity, der0v$	$t0concept$
здійснює	платежі	—	—	—	$t0abstr0finance$
оплату	послуг	—	—	$t0activity, der0v$	$t0activity, t0inter$
оплату	послуг	—	—	$t0activity, der0v$	$t0activity, t0inter$
оплату	послуг	—	—	$t0activity, der0v$	$t0activity, t0inter$
оплату	об'єктів	—	—	$t0activity, der0v$	$t0activity, t0inter$

Зазначимо, що маючи лінгвістичну та логічну БЗ, побудовані у вигляді онтології, отримуємо додаткові можливості пошуку антимем. Дійсно, деякої семантичної інформації може не бути в онтологічній БЗ, але вона може міститися в лінгвістичній БЗ, і навпаки. Це пояснюється тим, що крім відношень між словами, у реченні лінгвістична БЗ має ще один, більш важливий, вид впорядкованих відношень, а саме відношення між групами слів. Таким чином, ці дві БЗ, використані разом, взаємно доповнюють одна одну. А це полегшує аналіз текстів загалом.

Маючи онтології, можна сформулювати такі формальні означення, які випливають з викладеного.

Означення 4. Моделлю заданого тексту T відносно логічної теорії L називається четвірка: $M = (D, R, f, \Delta)$, де D — універсум тексту T ; $R = \{R_i\}$ ($i = 1, 2, \dots, m$) — множина відношень, визначених на універсумі D ; f — функція інтерпретації; Δ — область інтерпретації. Теорія L — логічне числення, в якому записуються і досліджуються властивості моделі M . Дві моделі: $M = (D, R = \{R_1, \dots, R_m\}, f, \Delta)$ і $M' = (D', R' = \{R'_1, \dots, R'_m\}, f', \Delta)$ відносно однієї і тієї самої логічної теорії L та області інтерпретації Δ називаються ізоморфними, якщо існує бієкція $\varphi: D \rightarrow D'$ така, що для $m \in D$, $m' \in D'$ таких, що $\varphi(m) = m'$ мають місце умови $(\varphi(m))^f = (m')^{f'}$ і для довільних $(a, b \in m^f, a', b' \in m'^{f'})$ $(a, b) \in R_i^f$ тоді і тільки тоді, коли $(a', b') \in R'_i{}^{f'}$ для всіх $R_i \in R$ і $R'_i \in R'$.

Наведене означення моделі тексту дещо відрізняється від означень з [6, 7], але сенс незмінний. Для успішного розв'язання задач логічного аналізу поняття моделі необхідне принаймні з двох причин.

Перша з них полягає в тому, що поняття моделі дає можливість визначити поняття смислової однозначності тексту, а саме текст називається смислово однозначним, якщо він має (з точністю до ізоморфізму) єдину модель. Текст називається смислово неоднозначним, якщо він має неізоморфні моделі. Друга причина полягає в тому, що за допомогою моделі формалізуються поняття міркування і тезауруса.

Таким чином, з огляду на наведені приклади можемо зробити такі висновки. Маючи результати лінгвістичного аналізу вхідного тексту та синтактико-семантичні залежності, які надаються цим аналізом, побудову універсуму і відповідно моделі можна автоматизувати. Така побудова виконується з використанням лінгвістичних кодів синтаксичних категорій типів складників природної мови. Крім того, правильність згенерованого універсуму можна перевірити повторним застосуванням засобів як лінгвістичного, так і логічного аналізу. Приклади 4 і 5 показують, що в результаті лінгвістичного аналізу можна отримати досить детальну інформацію про зв'язки окремих частин мови, які в повному обсязі не були використані під час побудови універсуму.

ВИСНОВКИ

Використання результатів синтактико-семантичного аналізу природномовних текстів суттєво допомагає будувати універсум міркувань, вони містять додаткову інформацію, якою можна послуговуватись для уточнення семантики. Крім того, з наведених прикладів випливає, що результати цього аналізу суттєво спрощують побудову онтологій.

Точність логічного аналізу фактів, добутих з природномовного тексту, залежить від правильності перекладу. Зазначимо, що юридичні тексти і особливо законодавчі тексти повинні підлягати логіко-лінгвістичному аналізу для вилучення невизначеності та неоднозначності трактування положень у цих документах. Логічний аналіз як раз і призначений для виявлення найменших невизначе-

ностей та неточностей у таких текстах. У класичному численні висловлювань зазвичай вважають, що кожне аналізоване висловлювання є або істинним, або хибним, але не тим і другим одночасно. У повсякденному житті, як відомо, висловлювання не так добре вкладаються в категорію тільки істинних або хибних. Можна навести чимало висловлювань, які не можна кваліфікувати тільки двома способами. Але класичне числення висловлювань досить корисне, оскільки воно вимагає бути точними в своїй логіці, навіть, якщо доводиться визнавати, що в посиланнях є елемент свавілля. Не будучи впевненими в правильності цих посилань, можна висловлювати твердження, які описують з максимально доступним ступенем істинності деякі обставини, або можна почати розглядати сукупність альтернативних наборів гіпотез. Цей напрямок досліджень буде розглядатися у майбутніх роботах.

СПИСОК ЛІТЕРАТУРИ

1. Клини С. Математическая логика. Москва: Мир, 1973. 480 с.
2. Bergmann M. An introduction to many-valued and fuzzy logic. Cambridge University Press, 2008. 329 p.
3. Дарчук Н.П. Комп'ютерне анотування українського тексту: результати і перспективи. Київ: Освіта України, 2013. 543 с.
4. Кривий С.Л., Дарчук Н.П., Проватар О.І. Онтологоподібні системи аналізу природномовних текстів. *Проблеми програмування*. 2018. № 2–3. С. 132–139.
5. The description logic handbook. Baader F., Calvanese D., McGuinness D.L., Nardi D., Patel-Schneider P.F. (Eds.). Cambridge: University Press, 2010. 601 p.
6. Шрейдер Ю.А. Логика знаковых систем. Москва: Знание, 1974. 64 с.
7. Кулик Б.А. Логика естественных рассуждений. СПб.: Невский Диалект, 2001. 128 с.

S. Kryvyi, N. Darchuk, T. Skrypnyk

NATURAL LANGUAGES ANALYSIS: ANTIMEMS, CONTRADICTIONS, ONTOLOGIES

Abstract. Problems of the analysis of natural-language texts are considered. In particular, algorithms for the partial solution of problems of logical consistency/inconsistency of facts derived from the text, search for antimems, and construction of ontologies based on the results of linguistic (semantic-syntactic) analysis of the text are proposed. Logical analysis of the text is performed on the basis of propositional calculus and methods of proof in this calculus.

Keywords: antimems, ontologies, knowledge base.

Надійшла до редакції 22.11.2022