УДК 51.681.3

АЛГЕБРО-ЛОГІЧНИЙ ПІДХІД ДО АНАЛІЗУ ТА ОБРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ

О.В. Палагін, С.Л. Кривий, М.Г. Петренко, Д.С. Бібіков

Інститут кібернетики імені В.М. Глушкова НАН України,
03680, МСП, Київ-187, проспект Академика Глушкова, 40,
e-mail: palagin_a@ukr.net, факс: +38044 5263348
Київський національний університет імені Тараса Шевченка,
01601, МСП, Київ-01вул. Володимирівська, 64,
e-mail: krivoi@i.com.ua, факс: +38044 2590439

Пропонується формальна модель системи аналізу та обробки текстової інформації на основі алгебри множин і відношень та простої логічної мови. Описані алгебро-логічний підхід до аналізу природномовного тексту та алгебраїчна система спискових структур, орієнтована на обробку текстової інформації.

The formal model of system of the analysis and processing of the text information on the basis of algebra of sets and relations and simple logic language is offered. Are described algebro-logical the approach to the analysis of naturally-language text and the algebraic system of list structures focused on processing of the text information.

Вступ

Бурхливий розвиток науки і техніки протягом останніх десятиліть 20-го та початку 21-го століть призвів до величезного росту інформації, яку одній людині (навіть висококваліфікованому спеціалісту в галузі науки і техніки) не під силу освоїти, зрозуміти і скористатися нею з метою проведення наукових досліджень. В зв'язку з такою ситуацією виникає необхідність в автоматизації процесів пошуку та обробки потрібної інформації для її подальшого ефективного використання. При цьому виникає декілька проблем.

Першою проблемою і однією з основних є проблема аналізу природномовної текстової інформації (морфологічний, синтаксичний, семантичний та логічний аналіз) з метою добування знань [1].

Другою проблемою є проблема проектування системи пошуку та добування знань, побудова її архітектури та розробка інструментарію для користувача [1, 2].

Третя проблема – це інтеграція знань із декількох (зокрема двох) предметних областей для забезпечення ефективності проведення досліджень міждисциплінарного характеру, використання вже існуючих алгоритмів, фактів, теоретичних положень та практичних розв'язань [1].

В даній роботі пропонується деяка формалізація методів аналізу природномовних текстів (ПМТ), пошуку логічних наслідків з даних фактів та перевірка суперечності в цій формалізації, формальна мова, названа алгебраїчною системою спискових структур (АССС), яка орієнтована на обробку текстової інформації. Ця АССС розглядається як алгоритмічна мова високого рівня.

1. Короткий огляд методів дослідження природномовних текстів

Формалізація природної мови з метою автоматизації аналізу природномовних текстів була започаткована ще на початку тридцятих років 20-го століття в роботах А. Тарського та його учнів, хоча про таку необхідність говорили ще Арістотель, Лейбніц і Ейлер. Зокрема, Арістотель виділив чотири типи висловлювань:

A – «все *X* є *У*»; E – «все *X* не є *У*»; I – «деякі *X* є *У*»; O – «деякі *X* не є *У*».

Ці типи висловлювань були названі ***силогізмами***, а сам підхід Арістотеля – ***силогістикою***. Пізніше Ейлер в популярній формі виклав своє розуміння силогістики Арістотеля, скориставшись геометричною інтерпретацією силогістики у вигляді кругів (цю інтерпретацію стали називати ***кругами Ейлера***). Ідеї Ейлера далі були розвинуті в роботах французького математика і астронома Ж.Д. Жергона, який ввів типи відношень й інтерпретацію силогістики Арістотеля в термінах цих відношень. Основні типи відношень, введених Жергоном, є такими:

G^1 – «збігається або рівнозначно»; G^2 – «лівостороннє включення »; G^3 – «окремий випадок збігання»; G^4 – «правостороннє включення»; G^5 – «несумісність».

Жергон показав, що кожний тип силогізму Арістотеля можна виразити у вигляді певної множини можливих варіантів таких відношень. Зокрема, A:{G^1,G^2}, E:{G^5}, I:{G^1,G^2,G^3,G^4}, O:{G^3,G^4,G^5}. Наприклад, висловлювання типу I означає, що деяка непуста підмножина множини, чи клас X включається в У. Головна трудність у використанні жергонових відношень полягає в тому, що практично всі типи цих відношень при складному реченні вимагають перевірки великої кількості варіантів. Тому більш підходящим виявився підхід на основі використання формальної математичної логіки.

Якщо Арістотель, Лейбніц та Ейлер лише говорили про необхідність формалізації і робили лише перші кроки на цьому шляху, то більш серйозні спроби такої формалізації були зроблені логіком А. Тарським [3, 4], результатом яких стала поява поняття виконуваності (сатисфакції) формул – більш загального ніж поняття істинності. Це поняття Тарський застосував до відкритих і замкнутих формул (під замкнутою формулою стосовно речень природної мови розуміють фразу), що дало змогу сформулювати поняття істинності природномовного речення і накласти на кожну відкриту атомарну формулу, яка складається з примітивного предиката (тобто предметної константи) і стількох змінних, стільки відповідає арності предиката. Оскільки таких формул існує лише скінченна множина, то такий підхід стає конструктивним.

Наступна спроба вдосконалення формалізації А. Тарського була зроблена Д. Девідсоном [5]. Він запропонував додати до понять виконуваності та істини рекурсивне означення істини. Тоді теорія Т, яка включає рекурсивне означення істини, пояснює яким чином значення фраз залежить від значення (сенсу) слів цих фраз. Але, оскільки слово «значення» не є синонімом слова «істина», то означення істини не завжди буде означенням значення (сенсу). Отже, теза Девідсона не зовсім очевидна, зате її легко обґрунтувати. Звідси випливає, що необхідно співставляти змістові значення розповідних речень з умовами їх істинності.

Якщо приймається рівняння «значення розповідного речення = умова істинності цього речення», то потрібно накласти умову: якщо в означенні говориться про те, яким чином умови істинності складного речення залежать від умов істинності його складових простих речень, то в цьому означенні повинно говоритися про те, як значення складного речення залежить від значень його складових простих речень.

Монтегю теж вірив у те, що методи формальної семантики можна застосувати до дослідження семантики природної мови. Але він, на відміну від Девідсона, відмовився від застосування логіки предикатів першого порядку, віддавши перевагу ***категоріальним граматикам***. Ці граматики включають в себе ті категорії, які спеціалісти в області граматик традиційно використовують при означенні природної мови, наприклад, такі категорії, як підмет чи присудок. Цей підхід протилежний до підходу Девідсона. Це дало змогу Монтегю замінити поняття абсолютної істини на поняття відносної істини в моделі, тому що в одній моделі одне і теж речення може бути істинним, а в другій – хибним. Таке розширення дало змогу визначити поняття логічної істинності і логічного наслідку для більш широкого фрагмента природної мови [6]. Таким чином Монтегю виділив два елементи: інтенцію (сенс) і екстенцію (денотат) й застосував їх до підметів, присудків та фраз. Існують й інші підходи до аналізу природномовних текстів, які ґрунтуються на поняттях семантичної мережі, фрейма тощо.

Якщо коротко характеризувати ідеї, які домінували останні десятиліття 20-го століття, то вони зводяться до таких положень.

Починаючи з 70-х років в дослідженнях природних і штучних (формальних) мов домінували намагання побудувати теорію, яка б охвачувала як природні, так і штучні мови.

Синтаксис викликає інтерес тільки в зв’язку з семантикою.

Метою семантики є пояснення понять істини та логічного наслідування.

Метою синтаксису є характеристика синтаксичних категорій, з яких побудовані висловлювання.

Ці ідеї привели до того, що виробилися нові погляди на такі поняття як граматика та значення виразів. А саме:

а) граматика – це інтерпретоване числення виразів мови (генеративна лінгвістика);

б) значення виразів обчислюються інтерпретатором (інтерпретаціонізм);

в) композиційність категорій синтаксису, семантики і прагматики (категоріальні граматики).

В даній роботі розглядається підхід, який поєднує в собі алгебраїчні аспекти аналізу природної мови та логічні аспекти такого аналізу. Нижченаведений текст структурований таким чином.

Формулюється формальна постановка задачі добування знань з ПМТ. З цього формулювання випливає структуралізація текстів при накладанні на текст певних обмежень. В якості прикладів такого типу обмежень розглядається силогістика Арістотеля та тексти бібліографічного характеру. Зокрема для силогістики Арістотеля будується її теоретико-множинна інтерпретація з розширеною системою правил виведення та аналізом можливих ситуацій, які можуть виникати в процесі застосування цієї системи правил.

Для маніпуляції, аналізу та трансформації текстів вводиться поняття алгебраїчної системи спискових структур, за допомогою якої виконується оброблення текстів на рівні фізичного представлення.

На завершення розглядається алгебро-логічний підхід до дослідження семантичних властивостей природномовних текстів, наводиться проста логічна мова для дослідження властивостей природомовного тексту.

2. Формальна постановка задачі добування знань з ПМТ

Перш ніж перейти до розгляду системи оброблення та добування знань, що містяться в ПМТ, визначимо поняття ***добування знань з ПМТ***. З цією метою скористаємося поняттями, які використовуються в програмуванні з обмеженнями.

Нехай дана множина D, на якій визначена деяка скінченна сукупність n-арних відношень R на D, тобто $R_i \subseteq D^n$, де $R_i \in R \subseteq D$, $i = 1,\ldots, k$. Мовою обмежень L на D називається деяка непуста множина $L \subseteq R \subseteq D$. Проблема виконуваності обмежень формулюється таким чином.

Для довільної множини *D* і довільної мови обмежень *L* на *D* проблемою виконуваності обмежень *CSP*(*L*) є розв'язання такої комбінаторної проблеми:

Дано: трійка *P* = (*V, D, C*), де – *V* – скінченна множина змінних; – *C* – деяка множина обмежень $\{C_1,...,C_q\}$;

– кожне обмеження $C_i \in C$ – це пара (s_i, R_i), де s_i - *n*-ка, яка називається областю обмеження, $R_i \in L$ – n_i -арне обмеження на *D*, яке називається відношенням обмеження.

Знайти: функцію $\varphi: V \to D$ таку, що $\forall (s,R) \in C$, де $s = (v_1, v_2, ..., v_n)$, *n*-ка $(\varphi(v_1), \varphi(v_2), ..., \varphi(v_n)) \in R$ або переконатися в тому, що такої функції не існує.

Множина *D* в цьому випадку називається областю проблеми, а множина всіх розв'язків *CSP* вигляду *P=(V,D,C)* позначається *Sol*(*P*).

У випадку аналізу ПМТ з метою добування знань множина *D* інтерпретується як вхідний текст *T*, в якому задані («'закодовані'») відношення $R_i = (R_{i1}, R_{i2})$, де R_{i1} – синтаксичні обмеження, а R_{i2} – семантичні обмеження, $V = (v_1, v_2, ..., v_n)$ – лексико-граматичні розряди (іменники, дієслова, прикметники тощо).

Проблема добування знань з ПМТ полягає в пошуку інтерпретації $\varphi: V \to T$ і явній побудові відношення R_{i2} при умові істинності відношення R_{i1}, тобто необхідно знайти відображення φ таке, що $\varphi(V) = \varphi_2(\varphi_1(V)) = \varphi_1 * \varphi_2(V)$, де * означає суперпозицію відображень. Відображення φ_1 і φ_2 означають відповідно синтаксичну та семантичну правильність речення чи тексту в цілому.

Це означення досить загальне і потребує уточнення. Розглянемо деякі конкретизації цього означення. Уточнення можна виконувати в різних напрямках.

Одне із можливих уточнень є уточнення відображення φ_1. Це відображення, в свою чергу, теж можна розглядати як суперпозицію двох відображень, які реалізують морфологічний та синтаксичний аналіз речень ПМТ і разом з відображенням φ_2 утворюють цілісну систему класичного типу, схема якої показана на рис. 1 [7].

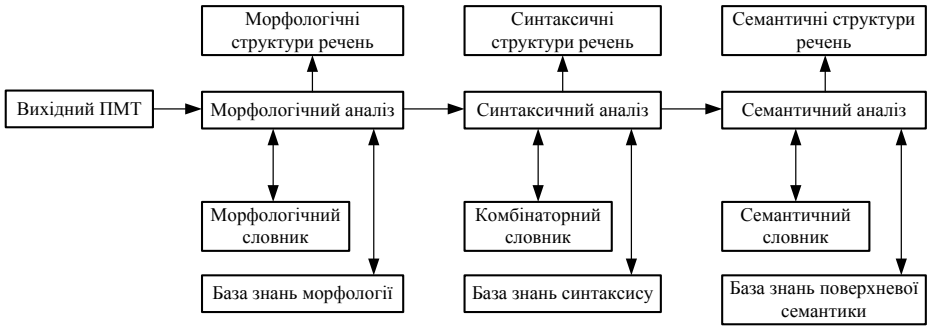

Рис. 1. Схема лінгвістичного аналізу класичного типу

Для перевірки коректності виконаного аналізу передбачається зворотний синтез речення до його запису в звичному орфографічному вигляді. Така перевірка може виконуватися в діалоговому режимі роботи системи з користувачем. Можлива структура словників, які використовуються в наведеній системі, та певне її обгрунтування описані в роботах [8–10].

З вищенаведеної формалізованої постановки проблеми аналізу ПМТ випливає, що однією з основних задач є задача побудови предметної моделі *A*. Ця задача є найбільш складною в зв'язку з тим, що предметна модель по суті є базою знань (побудова такої бази полягає в тому, щоб визначитися з об'єктами, які добуваються з тексту, з формальною логічною мовою, правилами виведення, аксіоматикою тощо).

3. Логічне слідування. Загальнозначущі та спеціальні відношення

При формалізації знань певної предметної області будемо розрізняти два типи відношень між поняттями цієї предметної області – загальнозначущі відношення та спеціальні. Перший тип відношень означає, що вони притаманні більшості або кожній предметній області (мають категоріальний характер), а другі – тільки певній конкретній предметній області. Наприклад, до першого типу відношень відносяться відношення «частина–ціле», «тип–підтип», «рід–вид», «розділ – параграф» тощо. Класичними типами загальнозначущих відношень є відношення часткового порядку та еквівалентності для подання існуючої ієрархії та синонімії в тексті: що аналізується. А до відношень другого типу відносяться специфічні, притаманні лише даній предметній області відношення. Наприклад, якщо предметна область відноситься до теорії автоматів, то специфічними відношеннями між поняттями цієї предметної області є відношення типу «скінченні автомати над скінченними

словами – регулярні мови, акцептовані цими автоматами», «автомати Б'юхі – регулярні надмови, акцептовані цими автоматами», «відношення еквівалентності на множині – розбиття цієї множини на класи еквівалентності» і т. п. Цей поділ є природним і його необхідність продиктована тим, що одна справа аналізувати текст на предмет існування в ньому необхідної нам інформації, а друга справа отримувати логічні наслідки з фактів, що знаходяться в такому тексті. Опишемо методи аналізу, які опираються на перший тип відношень, тобто на загальнозначущих відношеннях.

Арістотелева силогістика та її теоретико-множинна інтерпретація. Якщо детально проаналізувати типи загальнозначущих відношень, то можна помітити, що вони асоціюються перш за все з ***відношенням часткового порядку***. А такого типу відношення складають дистрибутивну гратку, яка має корисні властивості і ці властивості можна використати для генерації наслідків, тобто для пошуку (генерації) нових знань. Більше того, неважко помітити, що силогізми Арістотеля інтерпретуються в алгебрі множин і відношень. А з алгебри множин і відношень відомо, що відносно операцій об'єднання, перетину та доповнення ці алгебри є булевими кільцями, носії яких є частково упорядкованими множинами відносно теоретико-множинного включення. Закони цієї алгебри і властивості частково упорядкованих множин можна застосувати як правила виведення в такій формальній системі. Більш детально ці можливості розглянуто в [7]. Зокрема, це є закони алгебри множин і відношень (комутативності, асоціативності, дистрибутивності. ідемпотентності, поглинання і закони де Моргана), три основні властивості відношення включення (транзитивність, контрапозиція та асиметричність) та закон подвійного доповнення. При цьому останні дві можливості відіграють роль правил виведення. Нижче проілюструємо їх на прикладах.

Приклад 1. Нехай задані такі факти (взяті з книги Л. Керола «Історія з вузликами»):

«Всі малі діти нерозумні»;

«Всі, хто приборкує змій, заслуговують на повагу»;

«Всі нерозумні люди не заслуговують на повагу».

Вияснимо, які наслідки випливають з цих фактів. Зазначимо, що такого типу факти в логіці часом називаються ***полісилогізмами*** або ***сорітами*** **.** А силогізмом називається система, яка має лише дві посилки.

Визначимо основні терміни, з яких складається система фактів, введемо для них позначення і виберемо універсум ***U***. В даному прикладі основними термінами є такі: *«малі діти»* (С), *«розумні люди»* (Р), *«ті, хто приборкує змій»* (Т), *«ті, хто заслуговує на повагу»* (П). Ясно, що ці терміни представляють якісь множини в універсумі «люди». Їх запереченнями будуть відповідно такі терміни: *«не малі діти»* ($\neg$С), *«нерозумні люди»* ($\neg$Р), *«ті, хто не приборкує змій»* ($\neg$Т), *«ті, хто не заслуговує на повагу»* ($\neg$П). Тепер наші факти приймають вигляд:

$$C \subseteq \neg P,\ T \subseteq П,\ \neg P \subseteq \neg П.$$

Таким чином, визначилася гратка (як універсальна множина), яка складається з елементів ($\varnothing$, *U, P, T, П,* $\neg C, \neg P, \neg T, \neg П$), де ***U*** – універсум. Отже, першими наслідками даних фактів є наступні наслідки на підставі правила контрапозиції (правило контрапозиції в даній інтерпретації має вигляд «з $A \subseteq B$ випливає $\neg B \subseteq \neg A$, де знак $\neg$ означає доповнення множини)**:**

$$(C_1)\colon P \subseteq \neg C,\quad (C_2)\colon \neg П \subseteq \neg T,\qquad (C_3)\colon П \subseteq P.$$

Якщо перевести отримані наслідки на природну мову, то вони відповідно означатимуть такі факти: *«всі розумні люди не є малими дітьми», «ті, хто не заслуговує на повагу, не приборкують змій», «заслуговує на повагу той, хто розумна людина».*

Користуючись правилом транзитивності, отримуємо такі наслідки:

$$(C_4)\colon C \subseteq \neg П,\qquad (C_5)\colon T \subseteq P,\qquad (C_6)\colon \neg P \subseteq \neg T,\qquad (C_7)\colon П \subseteq \neg C.$$

З цих наслідків за тим же правилом транзитивності отримуємо ще два наслідки:

$$(C_8)\colon C \subseteq \neg T,\quad (C_9)\colon T \subseteq \neg C.$$

Якщо перевести на природну мову останні наслідки, то вони звучатимуть так:*«всі малі діти не є приборкувачами змій», «всі, хто приборкує змій, не є малими дітьми».*♠ (цей знак означає кінець приклада).

Таким чином, розв'язання поставленої задачі в наведеному прикладі отримано, але методи генерації наслідків є досить складними процедурами і хотілося б мати які-небудь засоби спрощення процесу пошуку висновків. Якщо в вищенаведених позначеннях для формул замість знака включення $\subseteq$ використовувати знак імплікації $\rightarrow$, то одержуємо мову, подібну до мови математичної логіки (на жаль в цій логіці не працює правило ***modus ponens***), а в такому випадку одним із засобів покращення ситуації в процесі генерації висновків є використання концептуальних графів і семантичних мереж. При такому способі представлення фактів, пошук і генерація наслідків зводиться до проблеми досяжності вершин в графі чи семантичній мережі. Розглянемо детальніше процес пошуку наслідків.

Введемо деякі означення і нотацію, якою користуються в математичній логіці. Будемо вважати, що кожному терміну відповідає деяка множина або її доповнення, якщо термін вживається разом із запереченням.

Означення 1. ***Літералом*** *називається термін або його заперечення, а множину всіх літералів деякої множини термінів називатимемо* ***базовими літералами*****.**

Висловлюванням *називається відношення включення, що виражене за допомогою базових літералів, в лівій частині якого находиться єдиний базовий літерал, а в правій частині – перетин множин, що представлені базовими літералами.*

Множину висловлювань, яка зображує деяку множину силогізмів, називатимемо базовою множиною формул.

Означення 2. ***Силогістичною структурою*** *називається множина літералів, відношення між якими визначаються множиною висловлювань, які називатимемо* ***посилками.***

Отже, формалізація фактів, наведених в прикладі 1, виглядає так: сукупність літералів L = {C, T, P, П, $\neg$C, $\neg$P, $\neg$T, $\neg$П} з посилками $C \subseteq \neg P$, $T \subseteq П$, $\neg P \subseteq \neg П$. Крім того, з цього прикладу випливає, що проблема пошуку наслідків зводиться до проблеми побудови контрапозиційно-транзитивно-асиметричного (КТА) замикання деякої базової множини формул. При транзитивному замиканні можуть виникнути такі ситуації:

К1) отримана формула вигляду $A \to \neg A$ або $\neg A \to A$;

К2) в процесі побудови транзитивного замикання одержано принаймні один цикл.

Вияснимо, що означають ці ситуації в нашому випадку. Перша формула у випадку *К*1) відповідає ситуації $A' \subseteq A$. З властивості перетину множини та її доповнення маємо $A \cap A' = \varnothing$, а тому таке включення вірне лише тоді, коли *A* є пустою множиною. Друга формула у випадку *К1)* означає, що доповнення *A'* множини *A* повинно бути універсальною множиною. З точки зору алгебри множин такі ситуації не можна назвати суперечними, але в нашому випадку ця ситуація означає, що деякий об'єкт мусить існувати і в той же час не існувати. Така ситуація цікава принаймні з двох причин. Перша причина полягає в тому, що ситуація є катастрофічною, а друга причина дає можливість виключити з розгляду певні терміни, які призводять до появи суперечності. Проаналізуємо ці випадки детальніше. Першою причиною виникнення суперечності типу $A \to \neg A$є поява в множині наслідків формул вигляду $A \to B$ і $A \to \neg B$. Другою причиною появи суперечності $A \to \neg A$є поява двох формул $A \to B$ і $\neg A \to B$.

Приклад 1. (продовження). Якщо до фактів з прикладу 1 додати формулу $P \to \neg T$ («всі розумні люди не приборкують змій»). Такого типу висловлювання не здається дивним з точки зору здорового глузду, але воно приводить до появи катастрофічних наслідків. Якщо побудувати КТА-замикання отриманої множини фактів, то в цьому замиканні з'явиться формула $T \to \neg T$. Якщо приймається, що задана множина фактів правильна, то приходимо до висновку, що людей, котрі приборкують змій, не існує. ♠

Зауважимо, що наявність суперечності типу $A \to \neg A$ не завжди веде до катастрофічних наслідків. Часом поява такої суперечності дає можливість розпізнати факти, які приводять до суперечності і вилучити суперечності. Продемонструємо це на прикладі.

Приклад 2. Візьмемо такі факти (з книги Д. Свіфта «Подорожі Леймлюеля Гулівера»):

A: «всі члени парламенту мають здоровий глузд»;

B: «всі, хто носить титул пера, не розбиває яйце з тупого кінця»;

C: «всі члени палати лордів мають титул пера»;

D: «члени палати лордів є членами парламенту».

Наслідками з цих фактів, зокрема, будуть такі висновки:

«всі, хто не в здоровому глузді, не є членами палати лордів» і «всі, хто розбиває яйце з тупого кінця, не є членами парламенту».

Якщо хто-небудь (а це на думку Д. Свіфта повинні бути лікар і аптекар) вирішив перевірити розумові здібності членів парламенту і отримав результат: *«всі члени палати лордів не мають здорового глузду»*. Коли отриманий факт додати до фактів A, B, C, D, то отримуємо суперечність типу: «всі, хто не в здоровому глузді, мають здоровий глузд». Така суперечність, як про це було сказано вище, може бути у випадку пустоти множини *A*. Для нашого прикладу це означає, що в парламенті немає членів, які не в здоровому глузді. Цей висновок дає можливість вилучити з розгляду термін «ті, хто не мають здорового глузду». ♠

Ситуація у випадку *К*2) означає (на підставі правила антисиметричності ПА), що всі об'єкти, які знаходяться в ланцюжку виведення, еквівалентні між собою. Звідси випливає, що в ієрархії об'єктів, які складають дану інформаційну систему, діють два загальнозначущі фундаментальні відношення – відношення часткового порядку і відношення еквівалентності. Відношення еквівалентності дає змогу, в разі необхідності, факторизувати об'єкти в такій інформаційній системі і цим досягати більшої компактності в представленні її об'єктів.

З розглянутих прикладів ситуацій для К1) і К2) випливає ще один надзвичайно важливий момент. Суперечності типу К1), К2) носять формальний характер, оскільки вони появляються тільки в результаті логічного аналізу заданої множини фактів, але існує іще один тип суперечності, який суттєво відрізняється від суперечностей К1) і К2). Припустимо, що в результаті побудови КТА-замикання з добре обґрунтованих і перевірених фактів, що не є суперечними, наприклад, відомих і обґрунтованих теорій, одержані суперечні наслідки, тобто наслідки, які суперечать фактам початкових теорій. Це говорить про те, що ***між початковими теоріями існують суперечності, а це і є ознакою появи нового знання*** або, принаймні стимулом для пошуку розв'язання отриманої суперечності. Все вищесказане дає право ввести таке означення.

Означення 3. Інформаційна система називається коректною, якщо в ній не виникають суперечності типу К1) чи К2).

Відомо, що в процесі побудови КТА-замикання, виходячи з різних початкових множин фактів, можна отримати одну і ту саму множину наслідків. Це дає можливість ввести таке відношення еквівалентності на множинах фактів: **дві множини фактів Ф і Ф' називаються еквівалентними, якщо КТА(Ф) = КТА(Ф').** На підставі цього відношення можна спрощувати (шляхом елімінації) структуру інформаційної системи, а також початкові множини фактів.

Розширення множини правил виведення силогістичної структури. Вищенаведені правила виведення в силогістичній структурі можна розширити шляхом використання інших властивостей відношення включення і законів алгебри множин і відношень. Розглянемо деякі з них.

(ЗПД) Закон подвійного доповнення (подвійного заперечення): $\neg(\neg A) = A$;

(ЗО) Закон об'єднання для включення (закон диз'юнкції): якщо $A \subseteq C$ і $B \subseteq P$, то $A \cup B \subseteq C \cup P$, зокрема якщо $A \subseteq C$ і $B \subseteq C$, то $A \cup B \subseteq C$;

(ЗП) Закон перетину для включення (закон кон'юнкції)**:** якщо $A \subseteq C$ і $B \subseteq P$, то $A \cap B \subseteq C \cap P$ і $A \cap B \subseteq C$, $A \cap B \subseteq P$;

А коли до результатів застосування цих правил застосувати правило контрапозиції, то приходимо до необхідності використання законів де Моргана:

ЗДМ1) Якщо $A \cup B \subseteq C \cup P$, то $\neg(C \cup P) = \neg C \cap \neg P \subseteq \neg(A \cup B) = \neg A \cap \neg B$;

ЗДМ2) Якщо $A \cap B \subseteq C \cap P$, то $\neg(C \cap P) = \neg C \cup \neg P \subseteq \neg(A \cap B) = \neg A \cup \neg B$.

Ці правила дають можливість розширити вищенаведене поняття висловлювання та силогістичної структури шляхом використання правил *ЗПД, ЗО, ЗП, ЗМД1* і *ЗМД2*. Застосовуючи ці правила до базової множини формул з прикладу 1, можна отримати висловлювання типу $T \cap \neg P = \varnothing$ і $C \cap T = \varnothing$, які говорять про те, що «серед приборкувачів змій немає нерозумних людей», а «серед малих дітей немає приборкувачів змій».

Розширена таким чином множина правил виведення дає змогу конструювати дескриптори, які включають кон'юнкції та диз'юнкції (подібна можливість описується й в інших системах [8]). Слід зауважити, що при застосуванні правил *ЗО* і *ЗП*, як і *ЗМД1* і *ЗМД2* необхідно слідкувати за певними обставинами збереження сенсу. Це пов'язано з тим, що при застосуванні правила *ЗП* необхідна узгодженість, яка полягає в тому щоб з'єднувалися родове і конкретне коректним чином в лексичному і семантичному значенні. Звідси випливає, що необхідно вказувати відмінності в різних моделях, в лексичних функціях, в полях стійких словосполучень і т. п. Як правило, в реальних комп'ютерних системах приймаються обмеження типу: слово представляється не більше ніж k словоформами, де k – стала величина для даної комп'ютерної системи [9, 10].

Чи є розширена таким чином силогістична структура булевою алгеброю і чи є додана множина правил такою, яка насправді розширює потужність силогістичної структури, потребує додаткового дослідження.

Що стосується спеціальних відношень, то на закінчення даного підрозділу зазначимо, що оскільки спеціальні відношення залежать від конкретної предметної області, то сформулювати їх в загальному вигляді неможливо. Формулювання цієї множини відношень цілком лежить на відповідальності експерта в даній предметній області, який ці відношення декларує і вносить в систему.

Перейдемо тепер до розгляду питання про засоби маніпуляції з текстовою інформацією. Представимо формальну алгебро-логічну мову, орієнтовану на обробку такого типу інформації.

4. Алгебраїчна система спискових структур

Побудова та опис такої мови виконується в два етапи. На першому описується алгебра спискових структур, а на другому – доповнення цієї алгебри до алгебраїчної системи спискових структур.

Алгебра спискових структур. Нехай $F(X)$ – вільна напівгрупа з одиницею над деяким скінченним алфавітом $X = \{x_1, x_2, ..., x_n\}$. Роль одиниці відіграє пусте слово e. Нагадаємо, що словом в алфавіті X називається довільна скінченна послідовність символів цього алфавіту. Довільне слово $p = y_1y_2 \dots y_m$ із $F(X)$ будемо називати **списком** елементів $y_1y_2 \dots y_m$, а самі елементи $y_i \in X$, $i = 1, 2, ..., m$, – складовими цього списку. При цьому елемент y_1 називається *початком*, а елемент y_m – *кінцем списку*. Якщо $p \in F(X)$, то число складових списку p називається його довжиною і позначається через $l(p)$. Якщо p, q – два списки, то список (слово) q називається *початком* (*кінцем*) списку (слова) p, коли існує таке слово p', що $p = qp'$ ($p = p'q$). Два списки $p = s_1s_2...s_k$ і $q = t_1t_2...t_l$ рівні між собою, якщо $l = k$ і $s_i = t_i$, $i = 1, 2, ..., k$.

З теорії відомо, що $F(X)$ є алгеброю з однією бінарною операцією конкатенації (*conc*) і однією нульарною операцією (пусте слово e). Введемо в розгляд ще декілька функцій і операцій над списками, тобто над елементами множини $F(X)$ [7].

Нехай N – множина натуральних чисел і $p = y_1y_2 \dots y_m$ – довільне слово із $F(X)$, тоді

$$head(p) = y_1 (head: F(X) \to F(X)). \tag{1}$$

Іншими словами, функція $head(p)$ дає перший символ слова p. Безпосередньо з визначення цієї функції випливають такі її властивості:

$$head(e) = e,\ head(y) = y, \text{ якщо } y \in X,\ head(head(p)) = head(p),$$
$$tail(p) = y_2 \dots y_m \ (tail: F(X) \to F(X)). \tag{2}$$

Очевидно, що $tail(e) = e$, $tail(y) = e$, якщо $y \neq X$.

Зміст наведених нижче функцій випливає з їх визначення.

$$add(p, i, x) = y_1 \ldots y_i x y_{i+1} \ldots y_m,\ 0 \le i \le l(p). \tag{3}$$

$$sub(p, i) = y_1 \ldots y_{i-1} y_{i+1} \ldots y_m,\ 1 \le i \le l(p). \tag{4}$$

$$dist(p, i) = (p_1, p_2), \tag{5}$$

де $p_1 = y_1 \ldots y_i$, $p_2 = y_{i+1} \ldots y_m$, $0 \le i \le l(p)$

$$hl(p, i) = y_1 \ldots y_i,\ 0 \le i \le l(p). \tag{6}$$

$$tr(p, i) = y_{i+1} \ldots y_m,\ 0 \le i \le l(p). \tag{7}$$

$$push(p, x) = px = add(p, l(p), x). \tag{8}$$

$$pop(p) = y_1 \ldots y_{m-1} = sub(p, l(p)). \tag{9}$$

Виявляється, що базовими операціями, тобто такими, через які виражаються всі останні із перелічених функцій, є операції *e*, *conc*, *head*, *tail*, *рекурсії* та *суперпозиції*. Іншими словами має місце таке просте твердження.

Теорема 2. Усі функції (3)–(9) представляються у вигляді термів за допомогою операцій e, conc, head, tail і операторів рекурсії та суперпозиції.

Доведення. Розглянемо послідовно випадки:

add$(p, i, x) = hl(p, i)\ xtr\ (p, i)$; ***sub***$(p, i) = hl(p, i - 1)\ tr\ (p, i)$; ***dist***$(p, i) = (hl(p, i), tr(p, i))$;

$$hl(p,i) = \begin{cases} e, & \text{якщо } i = 0, \\ head(p), & \text{якщо } i = 1, \\ head(p)\,hl(tail(p), i-1), & \text{якщо } i > 1. \end{cases}; \quad tr(p,i) = \begin{cases} p, & \text{якщо } i = 0, \\ tail(p), & \text{якщо } i = 1, \\ tr(tail(p), i-1), & \text{якщо } i > 1. \end{cases}$$

З представлення функцій *hl* і *tr* випливає представлення функцій *add*, *sub* і *dist*, а, отже, і представлення функцій *push* і *pop*.

Слід зазначити, що функції *head* і *tail* (*push* і *pop*) дійсно є операціями, у той час як решта функцій (3)–(7) операціями не являються. Це дозволяє ввести таке

Означення 4. Універсальна алгебра G = (F(X), Ω = {conc, head, tail, e}) розширена операторами рекурсії та суперпозиції називається **алгеброю спискових структур (АСС).**

У цій алгебрі легко встановити справедливість таких тотожностей.

(*а*) $sub(p,1) = tail(p) = tr(p,1)$; **(*б*)** $sub(add(p,i,x),i+1) = p$; **(*в*)** $hl(p,l(p)) = p$; **(*г*)** $tr(p,l(p)) = e$; **(*д*)** $pop(push(p,x)) = p$.

Має місце і ряд інших співвідношень у цій алгебрі.

Інколи разом з множинами функцій (3)–(9) розглядаються й інші корисні функції, такі як:

$$substr(p, i, j) = x_i \ldots x_{i+j-1}, \tag{10}$$

тобто це підслово слова *p*, яке починається з *i*-го символу і має довжину *j*;

$$conv(p) = x_m x_{m-1} \ldots x_2 x_1, \tag{11}$$

тобто це перестановка символів, які складають список *p* у зворотному порядку. Очевидно, що $conv(e) = e$ і $conv(y) = y$, якщо $y \in X$, а представлення функцій *substr* і *conv* у вигляді термів алгебри спискових структур має вигляд $substr(p,i,j) = hl(tr(p,i),j)$ і, $conv(p) = conv(tail(p))head(p)$.

Тепер, користуючись операціями і функціями (1)–(11), можна вводити й інші функції. Розглянемо деякі з таких функцій.

Нехай задано деякий алфавіт $X = \{x_1, x_2, \ldots, x_n\}$. Множина $F(X)$, як було зазначено вище, є повністю впорядкованою відносно лексикографічного порядку, мінімальним елементом якої є пусте слово *e*. Користуючись цим порядком, визначимо за індукцією такі функції:

а) $l(p) : F(X) \to N$ – функція довжини слова, означення якої було дане вище. Індуктивне означення цієї функції є таким:

$$l(p) = \begin{cases} 0, & \text{коли } p = e \\ 1 + l(tail(p)), & \text{інакше} \end{cases};$$

б) $subword(p, q) : F(X) \times F(X) \to \{0, 1\}$. Ця функція дорівнює 1, якщо слово *q* є підсловом слова *p*, і дорівнює 0 – у протилежному випадку. Індуктивне означення цієї функції таке: для довільних слів $p, q \in F(X)$

$$subword(p,q) = \begin{cases} 0, & \text{якщо } l(p) < l(q), \\ 1, & \text{якщо } hl(p, l(q)) = q, \\ subword(tail(p), q), & \text{інакше.} \end{cases}$$

Безпосередньо з означення цієї функції випливають такі її прості властивості: оскільки пусте слово є підсловом довільного слова і довільне непусте слово є підсловом самого себе, то

$$subword(p,e) = 1,\ subword(p,p) = 1;$$

в) *substit*(*p*, *q*, *r*) : *F*(*X*) × *F*(*X*) × *F*(*X*) → *F*(*X*). Результатом цієї операції є підстановка слова *r* замість першого входження слова *q* в слово *p*. Визначення цієї функції таке: для довільних слів *p*, *q*, *r* = *F*(*X*)

$$substit(p,q,r)=\begin{cases} e, & \text{якщо } p=e,\\ r\cdot tr(p,l(q)), & \text{якщо } hl(p,l(q))=q,\\ head(p)\,substit(tail(p),q,r), & \text{інакше.}\end{cases}$$

Безпосередньо з означення випливають такі очевидні властивості:
substit(*p*, *e*, *r*) = *rp*, *substit*(*p*, *q*, *q*) = *p*.

Приклад 3. Знайти

а) subword(abbcda, cd); б) subword(abbc, ax); в) substit(abbcda, cd, a); г) substit(abcd, xa, da).

Розв'язок:

а) subword(abbcda,cd) = subword(bbcda,cd) = subword(bcda,cd) = subword(cda,cd) = 1.

б) subword(abbc,ax) = subword(bbc,ax) = subword(bc,ax) = subword(c,ax) = subword(e,ax) = 0.

в) substit(abbcda,cd,a) = asubstit(bbcda,cd,a) = absubstit(bcda,cd,a) = = abbsubstit(cda,cd,a) = abbasubstit(a,cd,a) = =abbaasubstit(e,cd,a) = abbaa.

г) substit(abcd,xa,da) = asubstit(bcd,xa,da) = absubstit(cd,xa,da) = abcsubstit(d,xa,da) = abcdsubstit(e,xa,da) = abcd. ♠

Алгебраїчна система спискових структур. Розширимо АСС предикатом рівності та умовним оператором. Розширену таким чином АСС будемо називати **а**лгебраїчною системою спискових структур (АССС). Для цієї алгебраїчної системи має місце

Теорема 3. АССС є алгоритмічно повною системою, тобто системою, в якій можна обчислити довільну частково рекурсивну функцію.

Доведення. Для доведення необхідно показати, що довільний нормальний алгорифм Маркова можна представити і обчислити його значення в цій системі. Нехай Ф – деякий нормальний алгорифм Маркова, заданий системою формул підстановки *R* в алфавіті *X*:

$$\begin{cases} p_1 & \to & [.]q_1\\ p_2 & \to & [.]q_2\\ \dots & \dots & \dots\\ p_m & \to & [.]q_m \end{cases} \quad \begin{cases} p_1 & \to & q_1'\\ p_2 & \to & q_2'\\ \dots & \dots & \dots\\ p_m & \to & q_m' \end{cases}$$

де $q_i' = .q_i$, якщо формула підстановки заключна і $q_i' = q_i$, якщо формула підстановки проста. Далі, нехай $p \in F(X)$ і *G* – АССС над алфавітом $X' = X \cup \{.\}$. Тоді, використовуючи операції і предикати АССС, а також функції *subword*, *substit* і функцію довжини, можемо записати для довільного $p \in F(X)$:

Ф(*p*) = якщо *subword*(*p*, p_1) = 1 то
якщо *head*(q_1') = «.» то *substit*(*p*, p_1, *tail*(q_1'))
інакше Ф(*substit*(*p*, p_1, q_1')
інакше якщо *subword*(*p*, p_2) = 1 то
якщо *head*(q_2') = «.» то *substit*(*p*, p_2, *tail*(q_2'))
інакше Ф(*substit*(*p*, p_2, q_2')
інакше ……..
інакше якщо *subword*(*p*, p_m) = 1 то
якщо *head*(q_m') = «.» то *substit*(*p*, p_m, *tail*(q_m'))
інакше Ф(*substit*(*p*, p_m, q_m')
інакше *p*.

Покажемо індукцією за числом *n* застосованих підстановок до слова *p*, що функція *f*(*p*), яка отримана за допомогою системи *R*, і функція *f'*(*p*), яка отримана за допомогою системи Ф(*p*), збігаються.

При *n* = 0 маємо *f*(*p*) = *p*, оскільки жодна із підстановок системи *R* не застосовна до слова *p*. Із системи Ф(*p*) випливає, що *f'*(*p*) = *p*, оскільки всі умови вигляду *subword*(*p*, p_i) = 0. Отже, у цьому випадку *f* (*p*) = *f'* (*p*).

Припустимо, що рівність виконується для всіх *m* < *n* і нехай *n*-ою формулою підстановки є формула $p_i \to q_i'$. Можливі такі випадки: формула $p_i \to q_i'$ – заключна і формула $p_i \to q_i'$ не заключна.

Нехай у першому випадку *m* = *n* – 1 і після виконання *m*-ї підстановки одержано слово *p'*. За припущенням індукції *f*(*p'*) = *f'*(*p'*). З того, що $p_i \to q_i'$ застосовна до *p'* випливає, що жодна з підстановок, які їй передують у системі *R*, не застосовні до *p'*, або що те саме, що всі умови вигляду *subword*(*p'*, p_j) = 0, де *j* < *i*, а умова *subword*(*p'*, p_i) = 1 і *head*(q_i') = «.». що *f*(*p'*) = *substit*(*p'*, p_i, *tail*(q_i')) = *substit*(*p'*, p_i, q_i) = *f'* (*p'*).

Другий випадок аналогічний першому, з тією лише різницею, що в першому випадку обчислення закінчуються, а в другому – продовжуються.

5. Алгебро-логічний підхід до аналізу природномовного тексту

Формалізовану мову для аналізу природномовного тексту будемо нарощувати поступово, розширюючи синтаксис і семантику цієї мови. Почнемо з означення синтаксису та семантики мови висловлювань, які відрізняються від традиційних синтаксису і семантики числення висловлювань [5].

Синтаксис і семантика мови висловлювань. Розглянемо мову висловлювань, в якій висловлювання подаються у вигляді константних предикатів. Нехай L^0 означає цю мову. Кожне висловлювання мови L^0 представляється предикатом, аргументи якого приймають значення в множині предметних констант. Це означає, що L^0 не включає ні змінних, ні функціональних констант, відмінних від предметних констант. Крім того, висловлювання вже не є атомами мови L^0 на відміну від класичної мови висловлювань, оскільки воно будується за допомогою предикатних та предметних констант, які стають новими атомами мови L^0. Це нововведення дає змогу ввести *семантичне поняття виконуваності* для формул як узагальнення поняття істини. Для пояснення сказаного розглянемо приклад.

Приклад 4. Нехай множину предметних констант складають співробітники деякого відділу та статті, авторами яких вони є. Висловлюваннями виступають відношення між авторами і написаними ними статтями.

Нехай c_i, *i=1,2,…,n,* означають предметні константи, $s(c_i)$ – семантичне значення предметної константи. Для нашого прикладу розглянемо такі предметні константи:

КОНСТАНТИ	СЕМАНТИЧНЕ ЗНАЧЕННЯ
автор1	*s*(автор1) = Палагін
автор2	*s*(автор2) = Яковлєв
автор3	*s*(автор3) = Кривий
автор4	*s*(автор4) = Петренко
автор5	*s*(автор5) = Опанасенко
автор6	*s*(автор6) = Кургаєв

Предметні константи складають алфавіт об'єктної мови, а семантичні значення – інтерпретацію предметних констант або їх сутність. Мова L^0 включає предикатні константи (відношення) P_j, *j = 1,2,…,m,* семантичні значення яких є множинами. Семантичне значення предикатної константи P_j позначатимемо через $s(P_j)$. Для нашого прикладу визначимо такі предикати:

ПРЕДИКАТНА КОНСТАНТА	СЕМАНТИЧНЕ ЗНАЧЕННЯ
$P_{ав}$	$s(P_{ав})$ = множина авторів
$P_{ст}$	$s(P_{ст})$ = множина статей
$P_{ас}$	$s(P_{ас})$ = множина пар *(x,y)*, де *x* – автор, *y* – стаття, автором якої є *x*
$P_{арс}$	$s(P_{арс})$ = множина трійок *(a,p,c)*, де *a* – автор, *p* – рецензент, *c* – стаття
$P_{чс}$	$s(P_{чс})$ = множина пар (*x,y*), де *x* – читач, *y* – стаття. ♠

Синтаксис і семантика мови L^0. Нехай *X* – алфавіт, символами якого є предметні та предикатні константи, тобто $X = \{c_1, c_2, \ldots, c_n, P_1, \ldots, P_m\}$.

Синтаксис мови L^0 визначається так:

довільний предикат є формулою мови L^0;

якщо *A, B* – формули мови L^0, то $\neg A$, $A \vee B$, $A \wedge B$, $A \rightarrow B$, $A \Leftrightarrow B$ теж формули мови L^0;

ніякі інші вирази не є формулами мови L^0.

Семантичні значення предикатних і предметних констант дають можливість обчислити семантичні значення висловлювань, а отже, на їх підставі і значення формул.

Означення семантики мови L^0, як випливає з означення синтаксису і того, що говорилося вище про висловлювання, повинно включати традиційні для мови висловлювань кроки і специфічні кроки для предикатних і предметних констант.

Семантика мови L^0 визначається так:

довільна формула з L^0 інтерпретується як істинна або хибна;

якщо $A \in L^0$, то $\neg A$ істинна тоді і тільки тоді, коли *A* хибна;

якщо $A, B \in L^0$, то $A \vee B$ істинна тоді і тільки тоді, коли *A* або *B* істинна;

$A \wedge B$ істинна тоді і тільки тоді, коли *A* і *B* істинні;

$A \rightarrow B$ істинна тоді і тільки тоді, коли *A* істинна, а *B* хибна;

$A \Leftrightarrow B$ істинна тоді і тільки тоді, коли *A* і *B* або обидві хибні, або обидві істинні;

якщо P_j – предикатна константа арності *k* і $c_1, c_2, \ldots, c_k$ – предметні константи, то формула $P_j(c_1, c_2, \ldots, c_k)$ істинна тоді і тільки тоді, коли $(s(c_1), s(c_2), \ldots, s(c_k)) \in s(P_j)$.

6. Означення моделі та істини

Коли мова заходить про семантичне значення речень природної мови, то це значення може залишатися сталим, а може змінюватися в залежності від інтерпретації в тій чи іншій моделі. В зв'язку з такою можливістю розрізнятимемо абсолютну і відносну істинність простого речення. В теорії моделей формули логічної мови інтерпретуються в ***універсумі суджень,*** який будемо позначати через $\tilde{U}$.

Під моделлю для мови будемо розуміти деякий стан світу, яка включає в себе непусту множину $\tilde{U}$ (універсум) і функції інтерпретації $V: X \to \tilde{U} \cup B(\tilde{U}) \cup B(\tilde{U} \times \tilde{U}) \cup \ldots$, яка ставить у відповідність кожній предметній чи предикатній константі мови L^0 відповідний об'єкт з $\tilde{U} \cup B(\tilde{U}) \cup B(\tilde{U} \times \tilde{U}) \cup \ldots$, де $B(A)$ означає булеан множини A.

Приклад 5. Для універсума з введених вище констант, предметних констант та їх семантичних значень маємо:

$\tilde{U}$ = {Палагін, Яковлєв, Кривий, Петренко, Опанасенко, Кургаєв, сПЯ, сПО, сКр, сПП, сКуО, сПКП};

Функція інтерпретації визначається таким чином:

V(автор1) = s(автор1) = Палагін,
V (автор2) = s(автор2) = Яковлєв,
V (автор3) = s(автор3) = Кривий,
V (автор4) = s(автор4) = Петренко,
V (автор5) = s(автор5) = Опанасенко,
V (автор6) = s(автор6) = Кургаєв.

$V(P_{ав}) = s(P_{ав})$ = {Палагін, Яковлєв, Кривий, Петренко, Опанасенко, Кургаєв},

$V(P_{ст}) = s(P_{ст})$ = {сПЯ, сПО, сКр, сПП, сКуО, сПКП},

$V(P_{ас}) = s(P_{ас})$ = {(Палагін, сПЯ), (Яковлєв, сПЯ), (Кривий, сКр), (Палагін, сПП),
(Петренко,сПП), (Палагін, сПКП), (Кривий, сПКП), (Петренко, сПКП),
(Палагін, сПО), (Опанасенко, сПО), (Кургаєв, сКуО), (Опанасенко, сКуО)};

$V(P_{арс}) = s(P_{арс})$ = {(Кривий, Палагін, сКр), (Кургаєв, Яковлєв, сПЯ), (Яковлєв, Опанасенко, сПЯ)};

$V(P_{чс}) = s(P_{чс})$ = (Кривий, сПО), (Петренко, сПЯ), (Кургаєв, сКр), (Петренко, сКуО), (Кривий, сПЯ)}. ♠

Формальне означення моделі для мови L^0 має вигляд:

Означення 3. Моделлю мови L_0 називається пара M = ($\tilde{U}$, V), де $\tilde{U}$ – деякий універсум суджень, а $V: X \to \tilde{U} \cup B(\tilde{U}) \cup B(\tilde{U} \times \tilde{U}) \cup \ldots$, причому

а) $V(c_i) \in \tilde{U}$ для всіх $c_i \in X$;

б) $V(P_i) \subset \tilde{U}^n$ для всіх n-арних предикатних констант P_i.

Тепер семантичне значення предметної і предикатної константи визначається відносно моделі *M*. Якщо $c_i, P_j \in X$, то $s_M(c_i), s_M(P_j)$ буде означати значення констант c_i і P_j в моделі *M*. Отже, поняття моделі дає можливість замінити поняття *істини* на поняття *істини в моделі*. Оскільки функція інтерпретації *V* може приписувати формулам мови L_0 різні значення істинності відносно вибраної моделі $M = (\tilde{U}, V)$, то формула може виявитися істинною в одній моделі і хибною в іншій.

Ця обставина дає можливість на формальному рівні розв'язати проблеми омонімії та синонімії шляхом використання поняття моделі й відношення еквівалентності на розширенні одноелементної множини значень предметних констант. Дійсно, для розв'язання проблеми омонімії необхідно перевірити істинність даного висловлювання в різних моделях і вирішити яке з них має сенс.

Означення 4. Формула мови L^0 називається загальнозначущою, якщо вона істинна в довільній моделі мови L^0, і називається хибною, якщо вона хибна у всіх моделях цієї мови.

Дві формули мови L^0 називаються логічно еквівалентними, якщо вони одночасно істинні в одних і тих же моделях. Формула А називається логічним наслідком множини формул Г тоді і тільки тоді, коли формула А істинна в кожній моделі, в якій істинні всі формули з множини формул Г.

Аналіз природномовних текстів з метою добування знань (необхідної інформації) зводиться до виявлення таких властивостей як істинність в моделі та логічного слідування.

7. Збагачення мови L^0 – мова L^1

Розглянемо розширення мови L^0 шляхом додавання до L^0 предметних змінних і кванторів. Отриману таким чином мову позначатимемо L^1.

Алфавіт мови L^1 складається з множини $X = \{c_1, \ldots, c_n, x_1, \ldots, x_k, P_1, \ldots, P_m\}$, де предметні константи і змінні називаються *термами.*

Символи алфавіту *X* називаються нелогічними константами або атомами.

Означення 5. Формальне означення синтаксису мови L^1 має вигляд:

1) довільний атом є формулою мови L^1;

2) якщо A і B – формули мови L^1 і x – предметна змінна, то $\neg A$, $A \vee B$, $A \wedge B$, $A \to B$, $A \Leftrightarrow B$, $\forall xA$ і $\exists xA$ теж є формулами мови L^1;

3) інших формул в мові L^1, крім означених в пунктах 1), 2), немає.

Семантика мови L^1. Означення семантики мови L^1 виконується за допомогою функції інтерпретації. Інтерпретація мови $L1$ визначається так само як і у випадку мови L^0. А саме, задається універсум $\tilde{U}$, який складається з об'єктів (індивідів), і функція інтерпретації V для констант. В зв'язку з появою предметних змінних появляється функція g, яка присвоює значення змінним в множині $\tilde{U}$, тобто g: $\{x_1, ..., x_n\} \to \tilde{U}$.

Приклад 6. Якщо мова L^0 така як в вищенаведених прикладах, то синтаксис мови L^1 породжується формулами вигляду $P_{ac}(x_1, x_2)$, де x_1, x_2 – предметні змінні (на відміну від мови L^0, яка породжувалась тільки фразами вигляду $P_{ac}(c_1, c_2)$), семантичне значення якої можна обчислити як значення функції семантичних значень констант P_{ac}, c_1, c_2. А тому необхідно мати процедуру, яка присвоює значення істинності формулам типу $P_{ac}(x_1, x_2)$. Ця формула повинна бути істинна тоді і тільки тоді, коли стаття x_2 написана автором x_1. Отже, виникає необхідність присвоювання семантичних значень істинності предикатам, які мають вільні входження змінних, а отже, самим вільним змінним.

Таким чином, необхідно до правил семантики, які були описані вище, додати функцію, що присвоює кожній змінній x, мови L_1, деяке значення з універсума $\tilde{U}$. Ось для цієї мети і служить функція g: $\{x_1, ..., x_n\} \to \tilde{U}$. ♠

Визначимо спочатку поняття формули, істинної в моделі відносно функції присвоювання g, використовуючи рекурсію. Слід зазначити, що функція g не приймає участі в інтерпретації предметних і предикатних констант даної мови, а тому g не є частиною моделі M.

Нехай g – яка-небудь функція присвоювання значень предметним змінним мови L_1. Назвемо x-вибором функцію g', яка збігається з функцією g на всіх значеннях змінних x, крім змінної x. Символом $s_{M,g}(t)$ позначимо семантичне значення виразу t в моделі M з функцією присвоювання g.

Означення семантики мови L^1. Означення семантики мови L^1 виконується індуктивно в кілька кроків,

Означення 5. а) інтерпретація термів:

- якщо x – предметна змінна із L^1, то $s_{M,g}(x) = g(x)$;

- якщо c – предметна чи предикатна константа із L^1, то $s_{M,g}(c) = V(c)$, де V – функція інтерпретації моделі $M = (\tilde{U}, V)$;

б) інтерпретація термів:

- якщо t_1, t_2 - терми із L^1, то $s_{M,g}(t_1 = t_2)$ істинна тоді і тільки тоді, коли $s_{M,g}(t_1) = s_{M,g}(t_2)$;

- якщо $t_1, ..., t_n$ - терми і P - n-арна предикатна константа із L_1, то $s_{M,g}(P(t_1, ..., t_n))$ істинна тоді і тільки тоді, коли $(s_{M,g}(t_1), ..., s_{M,g}(t_n)) \in s_{M,g}(P)$;

в) інтерпретація формул:

- якщо A і B – формули із L_1, то $\neg A$, $A \vee B$, $A \wedge B$, $A \to B$, $A \Leftrightarrow B$ стандартним способом так, як це визначалося вище в означенні 4.

- якщо A – формула і x – змінна мови L^1, то $s_{M,g}(\forall xA)$ істинна тоді і тільки тоді, коли $s_{M,g}(A)$ істинна для всіх x-виборів g' функції g;

- якщо A – формула і x – змінна мови L^1, то $s_{M,g}(\exists xA)$ істинна тоді і тільки тоді, коли $s_{M,g}(A)$ істинна хоча б для одного x-вибору g' функції g.

Користуючись цим означенням семантики для мови L^1, можемо формально означити поняття формули, істинної в моделі й загальнозначущої формули.

Означення 6. Формула A називається істинною в моделі M, якщо $s_{M,g}(A)$ істинна для довільної функції присвоювання g.

Формула A називається загальнозначущою, якщо $s_{M,g}(A)$ істинна в довільній моделі M і для довільної функції присвоювання g.

Зазначимо, що поняття виконуваності формул відносно деякої функції присвоювання g, в точності збігається з поняттям істинності формули відносно деякої функції присвоювання g. Ці поняття зустрічаються і у Монтегю і теж збігаються.

8. Підхід до реалізації

Проблема представлення, пошуку та обробки природномовних текстів є однією з найбільш інтригуючих та складних проблем. Складність цієї проблеми полягає в тому, що вона погано формалізується і в наслідок цього погано піддається автоматизації. Використання формальних методів перевірки несуперечності чи виконуваності множини формул в системах гільбертовського типу (тобто, систем в яких доведення будуються формальним способом з аксіом шляхом застосування правил виведення) не мають якого-небудь задовільного розв'язання. Справа в тому, що системи гільбертовського типу не є структурованими, а це означає, що для збору необхідної інформації по одному єдиному об'єкту необхідно переглянути всю множину логічних формул, яка знаходиться в системі (як правило такою системою є база даних). На жаль, цим недоліком страждають всі формальні логічні системи гільбертовського типу. З метою ліквідації цього недоліку було запропоновано графічне представлення формул і їх аргументів, яке служить глобалізації й структурованості інформації. Основою графічного представлення є ***концептуальні графи (КГ)*** та їх родові структури – ***семантичні мережі (СМ)***. Таке представлення дає можливість візуалізувати модель природномовної картини світу, до якої відноситься проблема, що розглядається. Крім того, ця візуалізація дозволяє отримувати, в разі необхідності, весь процес доведення.

Подання семантичної моделі ПМО в залежності від типів задач, що потребують розв'язання (і відповідних КГ) може мати різний ступінь деталізації. При цьому розрізняють дві задачі формально-логічного подання ПМО в процесі комп'ютерного аналізу та інтерпретації. Перша з них відноситься до внутрішньомовної обробки, при якій результат семантичного аналізу представляється найбільш повними логічними виразами у відповідному логічному базисі. Їх формування виконується паралельно з процедурою зняття лексичної неоднозначності, яка потребує деталізації КГ і експліцитного представлення відповідних контекстних залежностей. Друга задача відноситься до позамовної обробки, до етапу побудови бази знань предметної області, а точніше – до побудови бази правил логічного виведення. Така база знань повинна мати короткі правила, які дають можливість реалізувати ефективне виведення (з точки зору швидкодії та пам'яті). Зі сказаного вище випливає, що для першої задачі слід використовувати деталізовані КГ, а для другої – максимально спрощені КГ і відповідні їм СМ.

Висновки

У роботі запропоновано формальну постановку задачі добування знань з природномовних об'єктів, з якої випливає структуралізація текстів при накладанні на них певних обмежень. Наведено приклади такого типу обмежень, такі як силогістика Арістотеля та тексти бібліографічного характеру. Для силогістики Арістотеля побудована її теоретико-множинна інтерпретація з розширеною системою правил виведення та аналізом можливих ситуацій, які можуть виникати в процесі застосування цієї системи правил. Для маніпуляції, аналізу та трансформації текстів введено поняття алгебраїчної системи спискових структур, за допомогою якої виконується оброблення текстів на рівні фізичного представлення. Розроблено алгебро-логічний підхід та просту логічну мову до дослідження семантичних властивостей ПМО.

Зрозуміло, що розроблені формальні засади комп'ютерного опрацювання ПМО далеко не вичерпують всіх теоретичних засад обробки природної мови, зокрема засад семантичного аналізу. Запропонована формалізація семантики та підхід, що її реалізує, потребують об'ємних спеціалізованих словників для різних етапів лінгвістичного аналізу ПМО. Іншим недоліком системи, що пропонується, є суттєве обмеження на стиль природномовних текстів.

Наступним кроком вдосконалення та розширення формально-логічних засад та підходу до оброблення ПМО повинен бути перехід до обробки неструктурованих текстів, тобто таких, стиль яких не вичерпується силогізмами Арістотеля та бібліографічним характером. Крім того, для природномовної обробки слід розглянути доцільність переходу від множини різного роду граматичних словників до єдиної, системно-онтологічної, природномовної бази знань, або мовно-онтологічної картини світу.

1. *Палагин А.В., Крывый С.Л., Петренко Н.Г.* Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация // УСиМ. – 2009. – № 3. – С. 42–55.
2. *Палагин А.В., Петренко Н.Г.* Системно-онтологический анализ предметной области // УСиМ. – 2009. – № 4. – С. 3–14.
3. *Tarski A.* The semantic conception of truth. Philosophy and phenomenological Research. – v.4. – 1944. – P. 241–375.
4. *Tarski A.* Logique, Semantique and Metamathematique (1923-1944). Colin. – Paris. – 1972.
5. *Devidson D.* Proceedings of Philosofical Logic. – Reidel. – Dordrecht. – 1969.
6. *Montague R.* Universal grammars. Theoria. Formal Phylisophy: Selected Papers of R. Montague. – Yale University Press. -1974. – vol. 36. – P. 222–246.
7. *Кривий С.Л.* Дискретна математика: вибрані питання. Київ: Видавничий дім «Києво-Могилянська академія». – 2007. – 570 с.
8. *Тейз А,, Грибомон П., Луи Ж. и др*. Логический подход к искусственному интеллекту. От классической логики к логическому программированию. – М.: Мир. – 1990. – 429 с.
9. *Тейз А,, Грибомон П., Юлен Г. и др.* Логический подход к искусственному интеллекту. От модальной логики к логике баз данных. – М.: Мир. – 1998. – 494 с.
10. *Леонтьева Н.Н., Семенова С.Ю.* Семантический словарь РУСЛАН как инструментарий компьютерного понимания. – М.: МГГИИ. – 2003. – С. 41–46.
11. *Леонтьева Н.Н.* К теории автоматического понимания естественных текстов. Часть 1. Моделирование системы "мягкого понимания" текста: информационно-лингвистическая модель. – М., МГУ, 2000. – 43 с.
12. *Леонтьева Н.Н.* К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. – М., МГУ, 2001. – 41 с.
13. *Апресян Ю.Д.* Лингвистический процессор для сложных информационных систем. М.: Наука. –1992. – 324 с.
14. *Ахо А., Хопкрофт Дж., Ульман Дж.* Построение и анализ вычислительных алгоритмов. М.: Мир. – 1979. – 535 с.