УДК 681.3

АЛГОРИТМІЧНА МОДЕЛЬ АСОЦІАТИВНО-СЕМАНТИЧНОГО КОНТЕКСТНОГО АНАЛІЗУ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ

А.В. Анісімов, О.О. Марченко, А.О. Никоненко

Київський національний університет імені Тараса Шевченка,
03680, Київ, проспект Академіка Глушкова, 2, корпус 6.
Тел.: факс 8 044 259 0427, rozenkrans@yandex.ru

Роботу присвячено розробці та обґрунтуванню ефективних методів семантичного аналізу та смислової обробки природномовних текстів. За результатами досліджень існуючих методів семантичної обробки текстів сформульовано концептуальний підхід до контекстного асоціативно-семантичного аналізу природної мови. В основі підходу лежить моделювання природномовного контексту за допомогою аналізу семантично близьких концептів в онтологічних семантичних базах знань із застосуванням алгоритмів пошуку найкоротших шляхів в графі. Процес семантичного аналізу стає більш потужним і ефективним та підпорядковує собі інші рівні лінгвістичної обробки тексту.

The paper deals with developing and basing efficient methods of the semantic analysis and semantic processing of natural language texts. Based on the results of analyses of the existing methods of semantic text processing, a conceptual approach to the context associative-semantic analysis of the natural language was worked out. It is realized through developing new modifications of existing methods and algorithms and creating the corresponding software. The approach is based on modeling the natural language context by means of analysis of semantically close concepts in ontological semantic databases. The algorithms of search of the shortest path in the graph render the process of the semantic analysis more powerful, efficient, with subordinating other levels of the linguistic text processing.

Вступ

Найбільш складні проблеми обробки природномовних текстів зумовлені явищами полісемії, омонімії, ононімії і т.д., які привносять неоднозначність в мову і значно ускладнюють задачу встановлення коректного відображення семантично-синтаксичної структури тексту в його формальне логічне представлення. Всі ці проблеми вирішуються на рівні семантичного аналізу.

З іншого боку застосування ресурсоємних функцій логічно-семантичного аналізу робить програми обробки тексту занадто складними та повільними. Людина в процесі розуміння тексту не так часто застосовує логіку – лише у випадку виникнення логічних задач, а інакше відбувається асоціативний пошук семантичного концепту, що відповідає даному слову та є контекстно близьким до свого оточення. При цьому асоціативний пошук є значно швидшим та економічнішим засобом розв'язання неоднозначності інтерпретації тексту [1]. Саме тому дана робота присвячена розробці алгоритмів асоціативно-семантичного аналізу з урахуванням контекстних зв'язків природномовного тексту.

Існуючі лінгвістичні алгоритми не можуть ще зрівнятися за якістю з можливостями людини. Однією з головних причин цього є інформаційна ізольованість процесів обробки на кожному етапі аналізу – під час роботи процесу обмін даними з іншими процесами не відбувається. Процеси обмінюються даними лише при переході від попереднього етапу до наступного – тобто вихід попереднього процесу є входом для наступного. Водночас семантичний, синтаксичний та морфологічний аналізи природної мови, що здійснюються людиною, є паралельними взаємодіючими процесами. При визначенні структури речення один процес використовує результати інших. Але напрямок такої взаємодії є не лише і не стільки "знизу-вгору" (морфологія визначає синтаксис, синтаксис – семантику), скільки "згори-донизу" – семантика керує синтаксисом та морфологією, синтаксис має вплив на морфологію. Коли в результаті аналізу один процес зустрічає неоднозначність, він підключає процес вищого рівня, який намагається ефективно розв'язати цю неоднозначність. Звідси випливає, що моделі процесів аналізу природномовних текстів доцільно розробляти як деякий простір паралельних розподілених процесів з заданим відношенням підпорядкування.

Семантичний аналіз текстів природною мовою

Задачею семантичного аналізу в класичному розумінні є визначення семантичної структури речень та тексту в цілому. Входом для цього етапу є розібрані синтаксичні структури речень тексту.

Процес семантичного аналізу можна умовно розкласти на три етапи:

перехід від слів та словосполучень до відповідних семантичних значень;
збірка семантичних фреймів окремих речень тексту;
злиття семантичних структур речень тексту в об'єднану семантичну мережу тексту.

Завдання першого етапу – знаходження в семантичній мережі онтологічної бази знань концепту, який відповідає коректному значенню слова чи словосполучення. Ця задача розв'язується відшуканням того значення слова з множини можливих альтернатив концептів, яке семантично є найбільш близьким до значень слів-сусідів з локального оточення даного слова.

Другий етап – побудова семантичного фрейму поточного речення вхідного тексту. Він полягає у заповненні слотів фреймової структури речення. Вибір типу слоту для заповнення значенням концепту слова залежить від синтаксичної позиції даного слова в граматичній структурі речення. Заповнення слотів виконується шляхом аналізу дерева розбору речення та синтаксичних позицій слів і словосполучень для кожного концепту.

Третя фаза смислового аналізу – об'єднання ізольованих семантичних фреймів речень в зв'язну семантичну мережу тексту. Об'єднання двох структур в одну мережу відбувається згідно принципу об'єднання семантично тотожних вершин, тобто якщо в структурі $G1$ та $G2$ є вершини, що посилаються на один семантичний концепт, вони об'єднуються в одну вершину [2].

Таким чином на виході отримаємо семантичну мережу вхідного тексту, яка містить у вершинах концепти тексту, зв'язані дугами семантичних відношень.

Контекстний семантичний аналіз

Пропонується метод моделювання семантичного контексту та обчислення семантичної контекстної близькості слів з використанням онтологічної бази знань. Онтологія є основою семантичного аналізу, тим семантичним полем, у межах якого можна обчислювати смислову близькість семантичних інтерпретацій лексем тексту щодо найближчого оточення, тобто контексту. Це і є відправною точкою для моделювання такого ключового явища, як мовний контекст взагалі, та контекстного аналізу природномовних текстів зокрема.

Онтологія є ієрархічною семантичною мережею, вершинами якої є концепти (смислові одиниці), а дугами – семантичні відношення між концептами. Семантика (смисл) концепту описується його смисловими відношеннями до інших концептів мережі. Реляційна позиція концепту в онтологічній мережі позначає його семантичне значення, властивості, зв'язок з іншими концептами та всі інші характеристики, які можливо передати природною мовою. Онтологічні технології використовують лінгвістичні моделі представлення знань про оточуючий світ та предметні області для ефективного запису та обробки інформації природномовного типу [3].

Слова та словосполучення певної мови зберігаються в лексиконі системи. Кожна лексема в системі посилається на множину значень, яку вона має в даній мові. Слово вжите поза контекстом може мати будь-яке значення з множини концептів, які прописані йому в онтологічній базі знань. Якщо слово вжите в контексті певного речення, то його значення має узгоджуватися із значеннями слів, які стоять поряд. Семантичні значення слів речення мають утворювати смислову єдність в структурі семантичного фрейму. Тому значення концептів слів, які стоять поряд, мають знаходитись семантично якомога ближче один до одного.

На вхід блоку контекстного аналізу подається послідовність слів $w_1\ w_2 \ldots w_n$. Кожному слову послідовності відповідає множина значень-концептів з онтологічної мережі – $\{s_{1i}\}\{s_{2i}\}\ \ldots\ \{s_{ni}\}$. З кожної множини в процесі контекстного аналізу необхідно обрати по одному значенню таким чином, щоб вони знаходились на максимально близькій відстані один від одного [4]. Тобто щоб сума відстаней від кожного концепту до всіх інших була мінімальною (рис. 1).

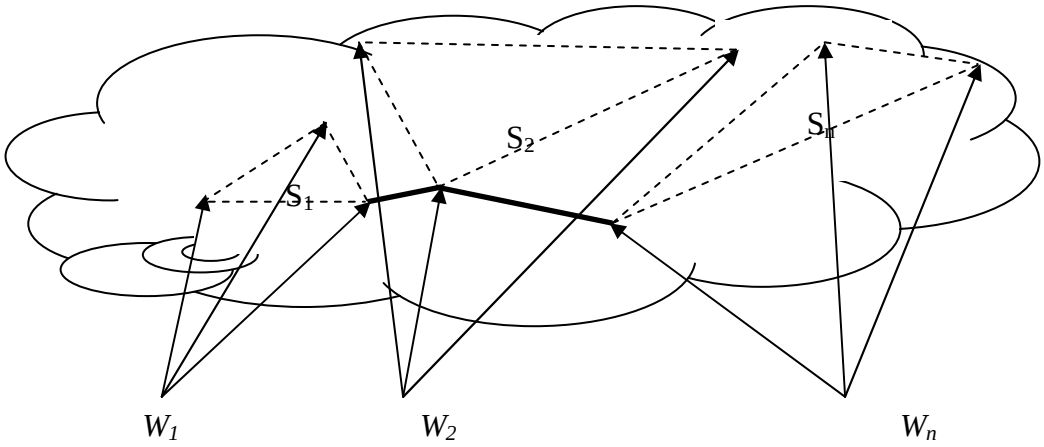


Рис. 1. Контекстний аналіз. Відбір значень концептів, найближче розташованих один до одного

Семантична відстань

Семантична відстань між двома концептами може бути проінтерпретована як довжина найкоротшого шляху між відповідними вершинами у графі онтологічної мережі. Окремого розгляду заслуговує питання пошуку найкоротшого шляху в онтологічному графі між вузлами концептів слів. Чи можна будувати шляхи не враховуючи типи зв'язків-відношень між вузлами та вважаючи їх однотипними? Якщо ні, то які послідовності типів зв'язків-відношень у шляху вважатимуться коректними, а які ні? В залежності від відповідей на ці питання можна запропонувати два підходи до визначення семантичної відстані.

1. Простий пошук шляху. Тоді вирішується класична задача пошуку найкоротшого шляху у графі. Типи зв'язків-відношень не враховуються. Вважається що всі дуги одного типу. Ще одним варіантом цього підходу є числове ранжування зв'язків-відношень, де дугам різного типу присвоюються різні вагові коефіцієнти, але сам алгоритм пошуку найкоротшого шляху залишається без змін.
2. Евристичний пошук. При побудові найкоротшого шляху дозволяються лише деякі послідовності типів зв'язків-відношень (наприклад, в ланцюгу шляху дозволяється послідовність типів зв'язків *гіпернімія-мелонімія-гіпонімія*, і не дозволена *гіпернімія-антонімія-мелонімія-гіпонімія*) Такі послідовності пропонуємо називати евристиками шляхів. Процедура пошуку найкоротших шляхів керується автоматом евристик, що в якості фільтра відбирає лише ті зв'язки-відношення, які відповідають закладеним евристикам.

Коли найкоротший шлях знайдено, його довжина буде прийнята як семантична відстань між даною парою концептів.

Семантико-синтаксичний аналіз

Коли на вхід блоку контекстного семантичного аналізу подається пара слів W_1 та W_2 , потрібно з множин їх семантичних значень S_1 та S_2 обрати відповідно пару значень концептів, якій буде відповідати мінімальна семантична відстань, тобто мінімальна довжина найкоротшого шляху в мережі онтології. Якщо побудувати найкоротший шлях в онтології між лексемами W_1 та W_2, він пройде через дану пару концептів, розташованих найближче один до одного. Якщо на вхід подається послідовність з n лексем, то для кожної з них слід виконати (n - 1) операцій пошуку найкоротшого шляху до лексем-сусідів за вхідним контекстом. Тобто при розв'язанні задачі аналізу вхідної послідовності довжиною n лексем необхідно виконати $n(n - 1)/2$ операцій пошуку. Пошук найкоротшого шляху між вершинами в графі є алгоритмічно дуже складною операцією, тому попередня оцінка є очевидно неприйнятною.

Але насправді нема необхідності будувати найкоротші шляхи в онтологічній мережі між всіма лексемами вхідного речення. Здійснювати контекстну прив'язку в онтології із визначенням значень концептів лексем потрібно, якщо ці лексеми зв'язані синтаксичними відношеннями в структурах дерева виведення речення. У разі існування правила, яке зв'язує деяку пару слів вхідної послідовності в єдину синтаксичну групу, дані лексеми зв'язуються побудовою найкоротшого шляху між ними в онтологічній мережі. Серед множин значень даних лексем обираються ті вершини-концепти, через які знайдено найкоротший шлях в онтології. Таким чином відпадає необхідність побудови надлишкових ланцюжків найкоротших шляхів між всіма словами речення. Перевіряються лише ті пари лексем, які зв'язані відношеннями в синтаксичній структурі вхідного речення.

Результати роботи синтаксичного аналізу враховуються асоціативно-семантичним контекстним аналізом для оптимізації процесу побудови асоціативних зв'язків контексту між словами та словосполученнями речення всередині ієрархічної мережі онтологічної бази знань. Синтаксична структура вхідного речення є фундаментом та каркасом для наступного семантичного аналізу.

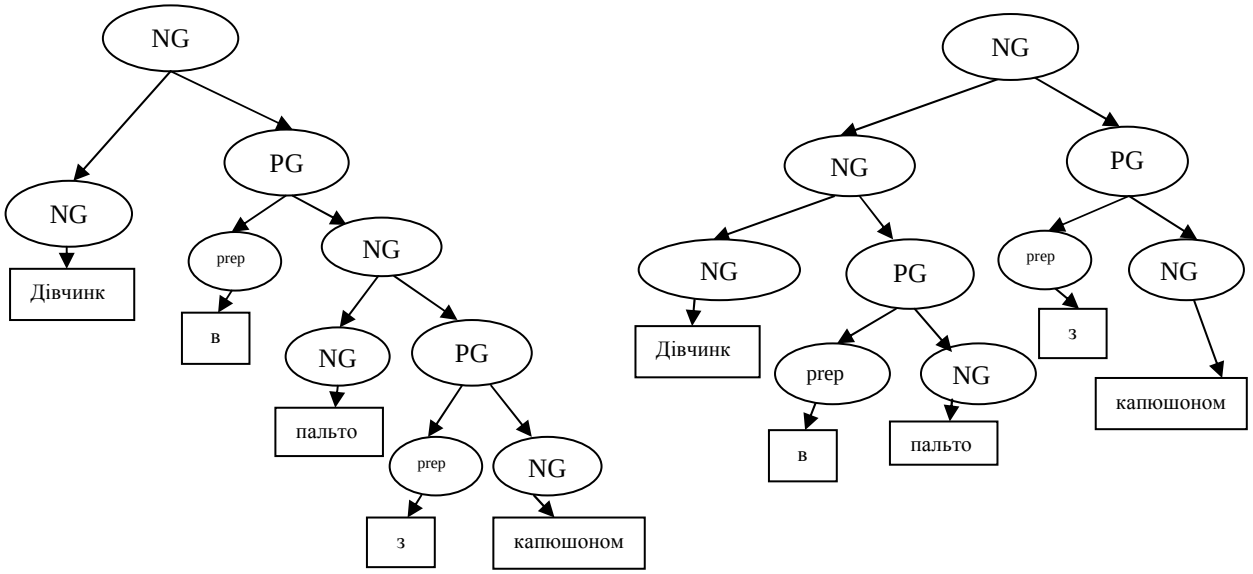
Але синтаксичний аналіз, як правило, не в змозі визначити на рівні граматики однозначну синтаксичну структуру вхідного речення. Завжди існує декілька варіантів дерев виведення вхідної послідовності речення, але лише один є найбільш адекватним з точки зору семантичної інтерпретації синтаксичного дерева [5]. Розглянемо приклад на рис. 2.

З точки зору семантики коректним є варіант А, тому що концепт «капюшон» є частиною та атрибутом концепту «пальто», і саме тому прийменникова група «з капюшоном» приєднується до іменникової групи «пальто». У варіанті В прийменникова група «з капюшоном» приєднується до іменникової групи «дівчинка в пальто» на верхньому рівні абсолютно без прив'язки до іменникової групи «пальто», тобто «капюшон» в такому разі претендує на позицію атрибута концепту «дівчинка», що є неадекватним. Якщо замінити приклад на «дівчинка в пальто з валізою», то коректним з точки зору семантичної інтерпретації буде варіант В. Отже, семантична інтерпретація є фільтром для відбору коректного варіанта синтаксичної структури речення.

Очевидним стає те, що не лише синтаксичний аналіз підтримує процес семантичного аналізу та забезпечує його вхідними даними у вигляді синтаксичних дерев вхідного речення, але семантичний аналіз може і має брати участь у процесі синтаксичного аналізу для розв'язання синтаксичних неоднозначностей структури речення. Саме так можна уявити взаємодію різнорівневих процесів лінгвістичного аналізу, де обмін даними відбувається не лише у напрямку «знизу-вверх», але і у напрямку «зверху-вниз», де більш високорівневі процеси коректують процеси нижнього рівня та керують ними.

Дівчинка в пальто з капюшоном

G: NG→NG PG
PG → prep NG
NG → Дівчинка | пальто | капюшон
Prep → в | з



A *B*

Рис. 2. Два варіанти побудови дерева виведення словосполучення «дівчинка в пальто з капюшоном» за правилами описаної граматики

Очевидним стає те, що не лише синтаксичний аналіз підтримує процес семантичного аналізу та забезпечує його вхідними даними у вигляді синтаксичних дерев вхідного речення, але семантичний аналіз може і має брати участь у процесі синтаксичного аналізу для розв'язання синтаксичних неоднозначностей структури речення. Саме так можна уявити взаємодію різнорівневих процесів лінгвістичного аналізу, де обмін даними відбувається не лише у напрямку «знизу-вверх», але і у напрямку «зверху-вниз», де більш високорівневі процеси коректують процеси нижнього рівня та керують ними.

Опишемо алгоритмічну модель взаємодії процесів синтаксичного та семантичного аналізів. В якості синтаксичного аналізу використаємо стандартний класичний алгоритм Кока – Янгера – Косамі [6].

Вхід. Контекстно-вільна граматика $G=(N, \Sigma, P, S)$ в нормальній формі Хомського без e – правил і вхідний ланцюжок $\omega= a_1 a_2 \ldots a_n \in \Sigma^+$ *Вихід*. Таблиця розбору T для ланцюжка $\omega \in L(G)$, така, що $A \in t_{ij}$ тоді і тільки тоді, коли $A \Rightarrow^+ a_i a_{i+1} \ldots a_{i+j-1}$

Алгоритм

Візьмемо $t_{i1}=\{A | A \rightarrow a_i \in P \ \forall i=1..n\}$. Після цього кроку з $A \in t_{i1}$ випливає, очевидно, $A \Rightarrow^+ a_i$

Припустимо, що вже обчислені t_{ij} для всіх $1 \le i \le n$ і всіх $1 \le j' < j$. Візьмемо $t_{ij} = \{A$ для деякого $1 < k \le j$ правило $A \rightarrow BC$ належить P, $B \in t_{ik}$ і $C \in t_{i+k, j-k}$ Оскільки $1 < k \le j$, то k та $j-k$ менші за j. Таким чином, t_{ik} і $t_{i+k,j-k}$ обчислюються раніше, ніж $A \Rightarrow^+ a_i a_{i+1} \ldots a_{i+j-1}$. Після цього кроку з $A \in t_{ij}$ випливає $A \Rightarrow BC \Rightarrow^+ a_i \ldots a_{i+k-1}$ $C \Rightarrow^+ a_i \ldots a_{i+k-1}\, a_{i+k} \ldots a_{i+j-1}$.

Повторювати крок (2) до тих пір, доки не стануть відомі t_{ij} для всіх $1 \le i \le n$ і $1 \le j \le n - i + 1$.

Зробимо наступні модифікації алгоритму. Нетермінали позначають синтаксичні групи. Синтаксичні групи в мові є носіями цілісного смислового значення – простого або складного концепту. Тому перша модифікація алгоритму полягає у тому, що за нетерміналом закріплюється семантичне значення. Коли на першому етапі використовуються правила типу $A \rightarrow a_i$, в t_{i1}, в якості семантичного значення заносяться лексеми a_i. Коли в t_{ij} заноситься деякий нетермінал A згідно знайденого правила $A \rightarrow B\ C$, то перевіряється наскільки семантично близько в онтологічній мережі розташовані слова-концепти, що є семантичними значеннями нетерміналів B та C. Побудова найкоротшого шляху між двома словами дає можливість визначити оцінку семантичної

адекватності поєднання нетермінальних груп *B* та *C*, що буде дорівнювати довжині знайденого шляху. Крім того, після побудови найкоротшого шляху між лексемами відбувається заміна цих слів на їх значення-концепти. З можливих альтернатив значень лексем обираються ті, що відповідають вузлам, через які пройде найкоротший шлях. Таким чином, разом з синтаксичним аналізом виконується перший етап семантичного аналізу – перехід «слово → концепт». Оцінка адекватності утвореного синтаксичного відношення, яка дорівнює довжині знайденого шляху, присвоюється нетерміналу *A*. Крім того, вирішується питання щодо визначення семантичного значення утвореного нетерміналу *A*. Воно може бути успадковане з нетермінілу *B* чи нетерміналу *C*, або може бути отримане через інтерпретацію утвореного сполучення (наприклад, «білий» та «дім» разом дає сполучення «Білий дім», яке вимагає окремої інтерпретації, яке не може були отримане перетином значень концептів «білий» та «дім»). Тому семантичне значення обчислюється та присвоюється нетерміналу *A* разом з оцінкою адекватності утвореного зв'язку. Далі із всіх альтернатив нетерміналів, що потрапили до t_{ij}, є сенс при подальшій побудові нетермінальних груп більш високого рівня враховувати лише один варіант із найкращою оцінкою, що відчутно спрощує алгоритм побудови структури. Оцінка адекватності утвореного нетерміналу *A* обчислюється як сума довжини найкоротшого шляху в онтологічній мережі між концептами- значеннями *C* та *B* та оцінок адекватності *C* та *B* відповідно. Якщо вверху таблиці в t_{1n} утворюється декілька варіантів нетермінальних груп, обирається група з найкращою оцінкою. Вона буде коренем побудованого найбільш легкого, найбільш коректного синтаксичного дерева вхідного речення, у вузлах якого знаходяться семантичні концепти-значення (рис. 3). Ще раз підкреслимо, що разом з синтаксичним аналізом виконано перший етап семантичного аналізу – перехід «слово → концепт».

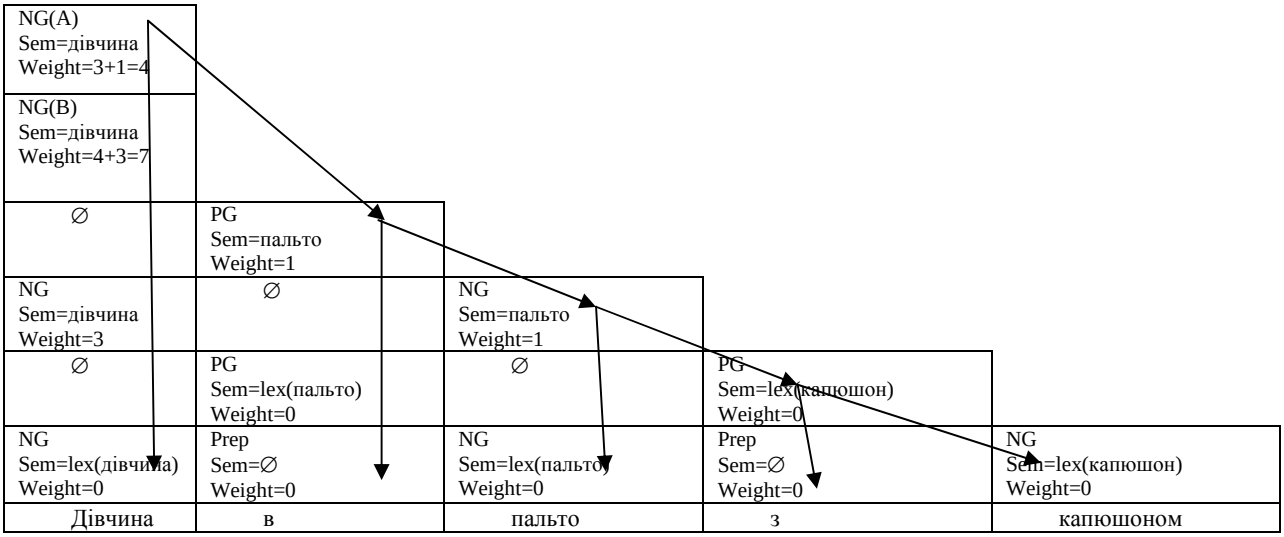


Рис. 3. Таблиця розбору вхідного ланцюжка модифікованим семантико-синтаксичним алгоритмом

Далі виконується побудова семантичного фрейму речення тексту. Вона полягає у заповненні слотів фреймової структури під час обходу отриманого синтаксичного дерева речення. Вибір типу слоту для заповнення значенням концепту залежить від синтаксичної позиції відповідного слова в граматичній структурі речення з використанням таблиць відмінків Філмора [7]. Тобто заповнення слотів виконується шляхом аналізу дерева розбору речення та синтаксичних позицій для кожного концепту.

Останній етап семантичного аналізу – об'єднання ізольованих семантичних фреймів речень в зв'язну семантичну мережу тексту. Об'єднання двох структур в одну мережу відбувається згідно принципу злиття семантично тотожних вершин. Якщо в структурах *G*1 та *G*2 є вершини, що посилаються на один семантичний концепт, вони об'єднуються в одну вершину.

Таким чином на виході згенеровано семантичну мережу вхідного тексту, яка містить у вершинах концепти тексту, які пов'язані дугами семантичних відношень.

Висновки

Описується алгоритмічна модель асоціативно-семантичного аналізу текстів природною мовою. Основою даної моделі є принцип асоціативного контекстного пошуку найкоротших шляхів в онтологічній мережі між семантичними значеннями лексем тексту, які зв'язують коректні концепти-значення цих лексем в рамках даного локального контексту. Таким чином, реалізовано асоціативний аналіз контексту в семантичному розборі речень тексту. На основі даної моделі побудована технологія асоціативно-семантичної обробки текстів. Ядром

даної технології є лінгвістичні бази знань (лексикони, синтаксичні бази даних, онтологічні бази знань) та алгоритмічні блоки, які реалізують різні фази лінгвістичного аналізу та різні елементи смислової обробки структур тексту. Використовуючи розроблену технологію розроблено ряд систем автоматичної обробки текстів природною мовою:

- система "Рефератор" призначена для обробки текстів природної мови. Дана система автоматично генерує реферати текстів та проводить їх індексацію (визначення за тематикою);
- система "VitaminE", що призначена для поліпшення якості машинного перекладу. Побудована на базі алгоритмів білінгвістичного семантичного аналізу;
- система семантичної фільтрації текстів. Система аналізує текст та визначає, чи є документ семантично приналежним до заданих користувачем тем;
- система фільтрації Internet-повідомлень з використанням лінгвістичних методів аналізу текстів. Програма аналізує потоки текстової інформації у комп'ютерній мережі, з можливістю заборони доступу до визначеного контенту та аналізу трафіка;
- система тематичної рубрикації та кластеризації текстів природною мовою.

1. *Анисимов А.В*. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. – Киев: Наук. думка, 1991. – С. 208.
2. *Анисимов А.В., Марченко А.А*. Система обработки текстов на естественном языке.// Искусственный интеллект.–2002.–№ 4. – С. 157–163.
3. *Nirenburg S., Raskin V.* Ontological Semantics, 2001,University of New Mexico.
4. *Марченко О.О.* Моделювання семантичного контексту при аналізі текстів на природній мові. Вісник Київського університету. Сер. фіз.-мат. Науки.– 2006. – № 3. – С. 230–235.
5. *Jurafsky D., Martin J. H.* (2000) Speech and Language Processing Prentice Hall, Englewood Cliffs, New Jersey 07632.
6. *Ахо А., Ульман Дж.* Теория синтаксического анализа, компиляции и перевода.– М.: Мир.– 1978.– 2 тома, Том. 1. – 612 с., Том. 2. – 487 с.
7. *Fillmore Ch. J.* The case for case // Universals in linguistic theory.// Ed. By E.Bach and B.Halms.– N.Y.– 1968.