КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

МАРЧЕНКО ОЛЕКСАНДР ОЛЕКСАНДРОВИЧ

УДК 681.3

АЛГОРИТМИ СЕМАНТИЧНОГО АНАЛІЗУ ПРИРОДНОМОВНИХ ТЕКСТІВ

Спеціальність 01.05.01 – теоретичні основи інформатики та кібернетики

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
кандидата фізико-математичних наук

Київ-2005

Дисертацією є рукопис.

Робота виконана на кафедрі математичної інформатики факультету кібернетики Київського національного університету імені Тараса Шевченка

Науковий керівник: доктор фізико-математичних наук, професор
Анісімов Анатолій Васильович,
Київський національний університет імені Тараса Шевченка, завідувач кафедри

Офіційні опоненти: доктор фізико-математичних наук, професор
Дорошенко Анатолій Юхимович,
заступник директора Інституту програмних систем НАН України, м. Київ

доктор технічних наук
Валькман Юрій Роландович,
зав. відділом расподілених інтелектуальних систем,
Міжнародний науково-учбовий центр ЮНЕСКО інформаційних технологій і систем, м. Київ

Провідна установа: Інститут кібернетики ім. В.М.Глушкова НАН України, відділ автоматизації програмування, м. Київ
Захист дисертації відбудеться “ 12 “ травня 2005 року о “14“ годині на засіданні спеціалізованої вченої ради Д 26.001.09 Київського національного університету імені Тараса Шевченка (03127, м. Київ, пр. Глушкова, 2, корп. 6, ф-т кібернетики, ауд. 40.Тел. 259-04-24. Факс 259-70-44. E-mail: rada1@unicyb.kiev.ua)

З дисертацією можна ознайомитися у Науковій бібліотеці Київського національного університету імені Тараса Шевченка (01033, м. Київ, вул. Володимирська, 58)

Автореферат розісланий “ 8 “ квітня 2005 року.

Вчений секретар спеціалізованої вченої ради,
кандидат фізико-математичних наук, доцент **В.П.Шевченко**

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Лінгвістична обробка природномовних текстів є однією з центральних проблем інтелектуалізації інформаційних технологій. Цій проблемі приділяється значна увага в розвинутих країнах Європи та США, свідченням чого є виділення величезних коштів на розробку лінгвістичного програмного забезпечення. Велика кількість науково-дослідних програм спрямовані на розвиток лінгвістичних інформаційних систем. В зв'язку з бурхливим розвитком Інтернет, інших комп'ютерно-комунікаційних технологій ця проблема набуває ще більшої значимості.
Ще з середини 50-х років минулого століття значні зусилля науковців були спрямовані на розробку математичних алгоритмів та комп'ютерних програм обробки текстів природною мовою. Для автоматизації аналізу та синтезу текстів створювалися різноманітні моделі процесів обробки тексту, а також відповідні алгоритми та структури представлення даних. Традиційно аналіз природномовних текстів представлявся як послідовність процесів – морфологічний аналіз, синтаксичний аналіз, семантичний аналіз. Для кожного з цих етапів були створені відповідні моделі та алгоритми. Для семантики тексту - класичні семантичні мережі та фреймові моделі Мінського, для синтаксису речення - граматики Хомського, системні граматики Холідея, дерева підпорядкування та системи складових Гладкого, розширенні мережі переходів; для морфолексичного аналізу розроблено багато різних моделей, орієнтованих на кокретні групи мов.
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 по мірі виникнення логічних задач, в інших випадках 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. Саме тому дана дисертаційна робота присвячена розробці алгоритмів асоціативно-семантичного аналізу з врахуванням контекстних зв'язків природномовного тексту.
Іншою актуальною проблемою є створення моделей представлення знань, подібних до тих, які використовуються людиною при аналізі та синтезі природної мови. В цих структурах дані різних рівнів обробки мови інтегровані функціонально, що робить сам процес аналізу зручнішим та прозорішим. Найбільш ефективними для роботи з природномовними базами знань є семантичні онтології, запропоновані В. Раскіним та С. Ніренбургом. Найвідомішою та найрозпосюдженішою в світі онтологічною системою є Wordnet– глобальна лінгвістична база знань, що розроблюється науковцями Принстонського університету впродовж 20 років. Багато розробників лінгвістичних комп'ютерних систем використовують її як базу для створення власних програм обробки текстів. Автором досліджені алгоритми мовної локалізації онтологій типу WordNet для створення мультилінгвістичних баз знань, а також розроблена оригінальна архітектура онтологічної бази знань, в якій інтегруються дані різних рівнів лінгвістичного аналізу.
Незважаючи на всі наукові досягнення, існуючі лінгвістичні алгоритми не можуть поки що зрівнятися по якості 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. В той же час 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.
Вивченню властивостей паралельних обчислень для різноманітних задач були присвячені роботи В.М.Глушкова, О.А.Летичевського, А.В.Анісімова та інших вчених. Саме результати математично-алгоритмічних досліджень паралельних розподілених процесів, отримані в 1980-90-і роки професором Анісімовим А.В., і знайшли своє ефективне застосування в задачах обробки природномовних текстів. Керуючі простори, що були запропоновані Анісімовим як універсальні засоби моделювання паралельних асинхронних розподілених процесів, були використані для привнесення в стандартні схеми обробки текстів елементів взаємодії між процесами при аналізі та синтезі природномовних текстів.
Поєднання послідовних та паралельних, формальних та евристичних підходів до вирішення описаних проблем дає можливість досягти значного підвищення ефективності програмної реалізації моделей обробки природномовних текстів, що і визначає актуальність роботи.
Зв'язок роботи з науковими програмами, планами, темами. Дана дисертаційна робота виконувалась в рамках наукової теми кафедри математичної інформатики факультету кібернетики Київського національного університету імені Тараса Шевченка "Розробка систем інтелектуалізації інформаційних технологій та дистанційного навчання" (державний реєстраційний номер 0101U002170) та господарчої теми Розробити інформаційну технологію синтезу, аналізу, реферування, пошуку і смислової інтерпретації текстової інформації та смислового перекладу текстової інформації з однієї мови на іншу" (державний номер реєстрації 0103U005499) в складі національної науково-дослідницької програми "Образний комп'ютер".
Мета і завдання дослідження. *Об'єктом дослідження* дисертаційної роботи є алгоритмічна модель семантичного аналізу природномовних текстів, *предметом дослідження* – структури даних онтологічних баз знань та алгоритми семантичного контекстного аналізу природномовних текстів та їх застосування для моделювання процесів смислової обробки текстів.
Методи дослідження. Методи та алгоритми теорії графів, теорії синтаксичного аналізу та штучного інтелекту, методики побудови онтологічних баз знань та розробки систем аналізу природномовних текстів.
Дослідження існуючих алгоритмів смислової обробки текстів, які коректно та ефективно працюють на онтологічних семантичних мережах довільного типу, показали, що найбільш доцільним для побудови потужних методів асоціативно-семантичного аналізу є використання кращих евристичних алгоритмів побудови найкоротших шляхів в графі.
Метою дисертаційної роботи є розробка математичних моделей та алгоритмів семантичного аналізу природномовного тексту.
Відповідно до сформульованої мети визначені основні *завдання* роботи:
дослідження структур семантичних онтологічних баз знань для визначення структур оптимального типу, в яких системно поєднуються дані різних етапів аналізу;
розробка паралельної моделі функціональної взаємодії процесів для ефективного лінгвістичного аналізу ;
побудова та дослідження оптимальних алгоритмів семантичного контекстного аналізу природномовних текстів на основі онтологічних баз знань ;
вивчення можливостей використання алгоритмів пошуку найкоротших шляхів та обходу графу для обробки онтологічних баз знань при асоціативно-семантичному аналізі текстів;
розробка програмно-алгоритмічних засобів семантичної обробки текстів.

Наукова новизна одержаних результатів. При виконанні роботи одержані такі нові результати:

Створені нові алгоритми семантичного контекстного аналізу текстів основані на принципі побудови асоціативно-семантичних шляхів в онтологічному графі бази знань.
На основі онтологічного семантичного графу, розроблена оригінальна архітектура бази знань в якій структурно інтегровані дані різних рівнів лінгвістичного аналізу.

Вперше застосовано алгоритми пошуку найкоротшого шляху в графі для вирішення проблеми обчислення смислової контекстної близькості в семантичному аналізі текстів.
На базі методів онтологічного семантичного аналізу текстів природною мовою створено ефективний алгоритм семантичного білінгвістичного аналізу, що дало змогу розробити систему покращення якості машинного перекладу "VitaminE".
Розроблена нова методика автоматизованої адаптації онтологій типу Wordnet до інших мов, яка була використана для локалізації до української та російської мов існуючої онтологічної семантичної бази знань та створення білінгвістичної системи аналізу.
Обґрунтована коректність та досліджені оцінки часової складності розроблених алгоритмів семантичного аналізу текстів, що підтвердило їх ефективність у порівнянні з існуючими алгоритмами.
Ефективність та коректність створених алгоритмів підтверджені здійсненою автором програмною реалізацією.
Практичне значення отриманих результатів. Розроблені алгоритми онтологічного семантичного аналізу текстів природною мовою на основі обробки семантичного контексту та побудови асоціативно-семантичних шляхів в мережах онтологічних лінгвістичних систем є базою для створення математично-програмних моделей багатьох процесів смислової обробки текстів, таких як семантичне реферування текстів, індексування, смисловий переклад текстів з однієї мови на іншу, підтримка семантично зв'язного природномовного діалогу та багатьох інших.
На основі розроблених алгоритмів була створена система покращення якості машинного перекладу VitaminE, що по семантичному контексту обирає коректні переклади слів. Експериментальне застосування системи свідчить про вірність ідей та гіпотез, що були покладені в основу моделі, описаної в даній роботі.

Результати досліджень включені до програми курсу "Штучний інтелект", який читається для студентів спеціальності „Інформатика" Київського національного університету імені Тараса Шевченка.

Особистий внесок здобувача. Всі основні результати дисертації одержані автором самостійно. В працях, опублікованих в співпраці з науковим керівником, професору Анісімову А.В. належить постановка задачі та участь в обговоренні результатів.
В роботі [1] Дерев'янченку О.В. належить розробка базового алгоритму розпізнавання текстів. В роботі [3] Нагорному В.А. належить ідея створення мультиагентного середовища для моделювання синтаксичних структур тексту.
Апробація результатів дисертації. Основні результати роботи доповідались на:
Міжнародній конференції "Моделювання та оптимізація складних систем" (МОСС-2001), Київ, 2001 р.
Міжнародній конференції "Обчислювальна та прикладна математика" 09.09-10.09.2002 р., Київ, 2002 р.
Міжнародній науково-практичній конференції "Штучний інтелект-2002", Кацивелі, Крим, 16-20 вересня 2002 р.
а також на наукових семінарах факультету кібернетики Київського національного університету імені Тараса Шевченка, Міжнародного науково-навчального центру інформаційних технологій і систем (ЮНЕСКО) НАН України та Інституту кібернетики ім. В.М. Глушкова НАН України.
Публікації. Основні результати роботи опубліковано у 7 роботах, які наведені у списку використаних джерел, з яких 4 – статті у фахових збірниках наукових праць, 3 – тези міжнародних конференцій.
Об'єм та структура роботи. Робота складається з вступу, п'яти розділів, висновків, списку використаних джерел – 110 найменувань. Загальний обсяг роботи складає – 150 сторінок, обсяг основного тексту – 125 сторінки.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтована актуальність теми дисертації, сформульовані мета та задачі роботи, наведено загальну характеристику роботи та одержані в ній наукові результати

В першій главі міститься огляд систем аналізу природномовного тексту, що використовують семантичний аналіз. Одні з них були програмно реалізовані (LinkParser, Mikrokosmos), інші мають оригінальні семантичні апарати (Formal Semantics, Generative Lexicon).
Для формального представлення семантичного значення речення дуже часто використовуються формули лямбда-числення. Цей підхід є дуже розповсюдженим в англомовно орієнтованих системах. Як приклад можна назвати програму підтримки діалогу Тері Віноградa та багатофункціональну систему семантичної обробки природної мови Lolita, що були реалізовані за допомогою мови Lisp та ліспоподібної мови Haskel, які базуються на фундаменті лямбда-числення. Практичний успіх цих проектів цілком доводить адекватність та зручність використання лямбда-числення для формального запису семантики природномовного речення.
В першій главі також проаналізовано відомий проект Мікрокосмос. В ньому дослідники використали нову семантичну природномовну технологію створення баз знань –"Семантичну Онтологію". На базі створеної онтології були розроблені семантичні алгоритми обробки текстів. В рамках проекту дослідники намагалися поєднати ряд існуючих розрізнених теоретичних лінгвістичних теорій (тобто мікротеорій) в єдину систему.
В другій главі описана загальна архітектура та основні принципи роботи системи автоматичної обробки природномовних текстів.
Система складається з блоку морфологічного лексичного аналізу, блоку синтаксичного аналізу та блоку семантичного аналізу.
Блок морфологічного лексичного аналізу.
Вхідні дані: текст природною мовою
———Обробка: блок розкладає текст на речення, речення на слова, потім обробляє кожне слово окремо. Функції морфологічного аналізу мають привести кожне слово до нормальної форми, та знайти його морфологічні характеристики
———Вихід: текст, що складається із нормалізованих слів із визначеними морфологічними характеристиками.
Блок синтаксичного аналізу
Вхідні дані: розібраний морфологічно-нормалізований текст природною мовою.
Обробка: блок розбирає речення тексту, генеруючи дерева синтаксичного виводу речень.
Вихід: синтаксичні дерева речень тексту.
Блок семантичного аналізу
Вхідні дані: синтаксичні дерева речень тексту.
Обробка: функції семантичного аналізу розбирають синтаксичні дерева тексту за допомогою онтологічної бази знань та генерують семантичну структуру речень. Потім семантичні структури речень інтегруються в семантичний граф тексту. Блок семантичного аналізу вирішує проблеми полісемії, метонімії, заміни займенників та інші складні лінгвістичні проблеми неоднозначності.
Вихід: семантичний граф тексту, який обробляється функціями постсемантичного аналізу, що виконують такі прикладні задачі, як визначення тематики тексту, генерація реферату, смисловий переклад з однієї мови на іншу, підтримка природномовного діалогу з користувачем, забезпечення природномовного інтерфейсу для БД та багато інших.
В третій главі описана онтологічна семантична база знань, що використовується функціями семантичного аналізу; дана організація її структури та методи кодування знань.
Онтологія є графом G(V,E), де вершинами є концепти – смислові одиниці, а дугами - семантичні відношення між ними, що описують смислове значення концептів. Саме релятивна позиція концепту в графі визначає його семантику.
Для запису концептів онтології використовується фреймова модель. Формально, концепт – це множина слотів (slots). Слот - множина пар виду <Поле, Значення> (<facet, filler>), де Поле (facet) служить для позначення ідентифікатора типу семантичного відношення, а Значення (filler) використовується для позначення того концепту, з яким, власно, встановлюється цей тип відношення. Як ідентифікатор типу відношення використовуються імена концептів, тому що відношення та атрибути також є смисловими одиницями. Це утворює універсальний засіб запису, який дозволяє виражати смислові структури будь-якої складності.

Головним відношенням онтології є відношення ЦЕ (is-a), яке також записується в слотах концепту. Коренем всієї онтології (по відношенню ЦЕ) є концепт ВСЕ. Його безпосередніми нащадками – концепти ДІЯ, ОБ'ЄКТ та ВЛАСТИВІСТЬ. Всі інші концепти наслідуються від них. За допомогою дуг-відношень в онтологічному графі концепти-об'єкти описуються як множина атрибутів, властивостей, та посилань на концепти типу дія, до яких вони мають деяке "рольове" відношення. Концепти типу "дія" описуються множиною модальних "рольових" відношень з концептами типів "об'єкт" та "властивість", які задають модально-рольову логічну схему концепту-"дії" – тобто аргументну структуру відповідної функції. Концепти типу "властивість-атрибут" задаються як деякі точки на шкалі властивостей або елементи певної множини.
Концепт може успадковувати семантичні відношення від батьківських концептів, якщо вони не мають суперечних слотів – тому в системі передбачений механізм контролю перекриття батьківських слотів.
Модально-рольові відношення, за допомогою яких концепти типу "дія" відображаються в онтології, в мові виражаються за допомогою синтаксису (наприклад, відношення **ІНСТРУМЕНТ** англійською мовою виражається за допомогою прийменникової групи з прийменниками "with" та "by", українською мовою – за допомогою орудного відмінку іменника (наприклад "вдарити *молотком*"). Тому необхідним стає введення додаткового типу зв'язку, що прив'язує потрібні синтаксичні шаблони до відповідних рольових семантичних відношень в онтології. Таким чином, отримаємо структурний зв'язок між синтаксисом та семантикою в межах інтегрованої онтологічної бази.
Словник системи містить лексеми конкретної природної мови, а онтологія - концепти, що є загальними для всіх мов. Онтологія і лексикон зв'язані відношенням *реалізація* (instance), по якому можна сказати, який концепт яким словом в якій мові може позначуватися.
В одне слово лексикону може входити багато відношень реалізації - у випадку омонімії. Тоді при аналізі такого слова в тексті постає питання про вибір одного з відношень.
Наприкінці розділу наводиться алгоритм автоматизованої локалізації одномовних онтологій до інших мов для створення мультилінгвістичних баз знань на прикладі українізації та русифікації всесвітньо відомої семантичної онтологічної бази WordNet.
В четвертій главі описані розроблені автором алгоритми семантичного аналізу природномовних текстів, а також проведено аналіз їх ефективності. Детально досліджується взаємодія процедур семантичного та синтаксичного аналізу.
На початку розділу розглядається алгоритм синтаксичного аналізу, що виконується паралельно та під керуванням семантичного аналізу. За рахунок цього ще на етапі синтаксичного аналізу вдається вирішити проблеми неоднозначності семантичної інтерпретації слів тексту та неоднозначного визначення синтаксичної структури речень.
В якості базового алгоритму синтаксичного аналізу розглядається Алгоритм А.
Алгоритм А. *Вхід*. Контекстно-вільна граматика G= (N,Σ, P, S) в нормальній формі Хомського без *e* - правил і вхідний ланцюжок $\omega = a1a2an \in \Sigma^+$ *Вихід*. Таблиця розбору Т Для ланцюжка $\omega \in L(G)$, така, що $A \in t_{ij}$ тоді і тільки тоді, коли $A \Rightarrow^+ a_i a_{i+1} a_{i+j-1}$.
Алгоритм.
1. Покласти $t_{i1}=\{A \mid A \to a_i \in P \ \forall i=1..n\}$. Після цього кроку з $A \in t_{i1}$ слідує, очевидно, $A \Rightarrow^+ a_i$.
2. Припустимо, що вже обчислені t_{ij} для всіх $1 \le i \le n$ і всіх $1 \le j' < j$. Покласти $t_{ij} = \{A$ для деякого $1<k \le j$ правило $A \to BC$ належить P, $B \in t_{ik}$ и $C \in t_{i+k, j-k}\}$. 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 $A \Rightarrow^+ a_i a_{i+1} a_{i+j-1}$.Після цього кроку з $A \in t_{ij}$ слідує $A \Rightarrow BC \Rightarrow^+ a_i a_{i+k-1} C \Rightarrow^+ a_i a_{i+k-1}\, a_{i+k} a_{i+j-1}$.

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

Після виконання алгоритму в кожному t_{ij} зберігаються нетермінальні символи, що утворюються шляхом злиття нетерміналів нижніх рівнів. Але кожен такий нетермінал має семантичне навантаження. В синтаксичних правилах типу A → B C завжди на рівні синтаксису можна вказати, який з нетерміналів є головним. Наприклад, в NG → ADJ NG (іменникова група → прикметник

+іменникова група) головною частиною буде іменникова група. Новий нетермінал А наслідує семантичне навантаження (концепт) головного нетерміналу з пари (B,C). На всьому просторі синтаксичних правил вводимо функцію F_1: S → {1,2}, що по елементу множини синтаксичних правил S визначає номер головного нетерміналу.

Крім цього, коли ми "зливаємо" два нетермінали в один нетермінал вищого рівня, ми можемо оцінити в цей момент семантичну адекватність такого злиття. Якщо нетермінал B, що представляє цілісний синтаксично (тобто і семантично) об'єкт, зливається з іншим нетерміналом C, щоб утворити нетермінал вищого рівня A, ми можемо виміряти "семантичну відстань" між концептами, що містяться в B та C та помістити це значення як семантичну вагу нетерміналу A. Введемо функцію F_2: Sem^2 → N, де Sem – множина концептів онтології. F_2 по парі концептів визначає довжину семантичного шляху в онтології між цими концептами.

Таким чином визначаються дві функції за допомогою яких в таблиці розбору до нетерміналів прикріплюється їх семантичне навантаження (концепти) та семантична вага. В результаті алгоритми обходу таблиць розбору, що будують синтаксичні дерева, можуть розглядати не всі можливі варіанти розбору, а лише найлегші дерева, що значно спрощує алгоритм та робить його ефективнішим. Крім цього в вершинах синтаксичних дерев після роботи алгоритму стоять концепти, вирішується проблема семантичної інтерпретації слів тексту, що робить алгоритм **семантико-синтаксичним**.

Результати дослідження ефективності та коректності роботи семантико-синтаксичного **Алгоритму A** зведені в наступну теорему:

Теорема 4.1. Часова складність алгоритму генерації таблиці розбору $O(n^3)$, де n – довжина вхідного ланцюжка. Алгоритм вимагає $O(n^2)$ елементів пам'яті.

Далі в роботі розглядаються алгоритми семантичного аналізу, що базуються на принципі побудови найкоротших шляхів в онтології між концептами-варіантами інтерпретації лексем. Ланцюжок найкоротших шляхів відповідає набору вірних варіантів концептів з можливих альтернатив семантичних інтерпретацій слів тексту. Додаткові обмеження на побудову шляхів в онтологічному графі виникають через вживання тих чи інших синтаксичних шаблонів в реченні. Таким чином скорочується перебір в алгоритмі побудови найкоротших шляхів і це робить його більш швидким та інтелектуальним. Синтаксична розмітка онтології дала змогу використовувати спеціалізовані оптимізовані під структуру онтологічної бази алгоритми семантичного аналізу, що розглядаються в даному розділі .

Ідеальним базовим алгоритмом для розробки методів пошуку найкоротшого шляху в досить специфічному за структурою графі онтології виявився локальний алгоритм для задачі про найкоротший шлях з одного джерела. Це, певною мірою, універсальний алгоритм, що в залежності від ситуації вибирає найоптимальнішу з евристик обходу графу і мінімізує час побудови найкоротших шляхів.

Наприкінці розділу розглядається рекурсивна функція семантичного аналізу analyze(node), що обходить дерево синтаксичного виводу, яке є виходом вищенаведеного алгоритму **A**. Алгоритм **B** обходить це дерево зліва-направо, будуючи найкоротші шляхи в онтології між концептами сусідніх вершин із врахуванням синтаксичного навантаження дуги, що їх з'єднує в дереві синтаксичного аналізу. Побудова найкоротшого шляху в онтології закінчується заповненням відповідного слоту у фреймі речення.

```
Function analyze(node)
1  analyze(node.left); /заходить в ліве піддерево/
2 Rule:=(node.nonterm←node.left.nonterm node.right.nonterm); /визначає правило граматики/
concept1:=node.left.concept;
term1:=node.left.term;/ запам'ятовує концепти та термінали/
concept2.= node.right.concept;
term2:=node.right.term;
```

/ функція interpretate знаходить найкоротший шлях в онтології між концептами concept1 та concept2, що проходить по відношенню, навантаженому синтаксичним шаблоном Rule з відповідними терміналами term1 та term2
Після цього заповнює відповідний слот фрейму даного речення/
interpretate(concept1,concept2, rule,term1, term2);
analyze(node.right); /заходить в праве піддерево/

Доведено теорему, що визначає часову складність даного алгоритму:
Теорема 4.2. Часова складність алгоритму **B** семантичного аналізу дорівнює O(N), де N – кількість термінальних символів.
П'ята глава присвячена алгоритмам постсемантичної обробки структури, що генерується процедурами семантичного аналізу, тобто семантичного графу, що є виходом блоку семантичного аналізу. Розглянуто алгоритми семантичного білінгвістичного аналізу для покращення якості текстів машинного перекладу, алгоритми реферування тексту та деякі інші.
Задача покращення якості тексту машинного перекладу вирішується за допомогою алгоритму контекстного семантичного аналізу. З метою вирішення неоднозначності перекладу деякого слова тексту для кожної з альтернатив перекладу розраховуються найкоротші шляхи в онтології до слів, що стоять поряд та переклалися однозначно. Так як алгоритм працює з білінгвістичною онтологією, то він може використовувати також, звичайно, і текст оригіналу для уточнення та покращення якості перекладу. Тому алгоритм є контекстно-семантичним та білінгвістичним. Із кількох варіантів перекладу обирається той, що відповідає найменшій сумі довжин найкоротших шляхів в онтології до вершин-концептів з найближчого оточення. Тим самим формалізується поняття контексту.
Розглянемо $l_1 l_n$ – вхідний текст, l_i – і-те слово тексту. Нехай при машинному перекладі і-те слово переклалося неоднозначно. Вихідний текст $w_1\{w_i\}w_n$, де $\{w_i\}$- альтернативи перекладу слова l_i. Алгоритм перебирає альтернативи перекладу і для кожного елементу з $\{w_i\}$ розраховує суму довжин найкоротших шляхів в онтології до слів з деякого околу і-тої позиції, що переклалися однозначно.
На вході системі подається масив ***a*** з ***n*** однозначно перекладеними словами та масив b, що містить k варіантів перекладу неоднозначного слова.
Для того, щоб обрати вірний варіант з масиву b, треба для кожного з варіантів перекладу неоднозначного слова побудувати множину найкоротших шляхів до однозначно перекладених сусідів і підрахувати сумарну довжину цих шляхів. З масиву b вибирається той варіант, якому відповідає найменша сума довжин найкоротших шляхів.
Алгоритм має наступний вигляд:

```
MinSum:=+∞;
For i:=1 to k {Для кожного варіанту перекладу}
Begin Sum:=0;{Обнуляємо поточну суму довжин найкоротших шляхів}
For j:=1 to n do Sum:= Sum+Findpathword(b[i],a[j]);
{Підраховуємо поточну суму довжин найкоротших шляхів}
If Sum<MinSum then{Якщо поточна сума довжин найкоротших шляхів}
        Begin                   {менша за попередню}
        MinSum:=Sum;{Фіксуємо поточну суму довжин найкоротших
        Rs:=b[i];{шляхів як мінімальну. Фіксуємо поточний варіант перекладу b[i]}
End;
End;
Result:=Rs;{Результат}
```

Базовою функцією алгоритму є функція Findpathword(a,b), що знаходить найкоротший шлях в онтології між словами a та b та визначає його довжину. В якості алгоритму використовується один із варіантів локального алгоритму пошуку найкоротших шляхів, що використовує оптимальну в

даному випадку стратегію обходу вершин. Доведена теорема про коректність побудови найкоротшого шляху в онтології:

Теорема 5.1. (правильність алгоритму побудови найкоротшого шляху в онтології). *Нехай G=*(V, *E) онтологічний* граф з невід'ємною ваговою функцією w: E→ *R* і вхідною вершиною s. Тоді після застосування алгоритму побудови найкоротших шляхів в онтології для всіх вершин u $\in$ V будуть виконуватися рівності d [u]=$\delta(s, u)$. Тобто виконується умова коректності знайдених найкоротших шляхів.

Описаний алгоритм покращення якості текстів машинного перекладу є прозорим та ефективним. Час роботи алгоритму пропорційний розміру графу(кількості ребер та вершин):

Теорема 5.2. Час роботи алгоритму семантичного контекстного білінгвістичного аналізу дорівнює $0 (V^2)$.

Було проведений ряд експериментів з розробленою на базі описаних алгоритмів системою покращення якості текстів машинного перекладу. Результати цих експериментів довели ефективність та коректність семантичного контекстного аналізу.

Одним з найбільш актуальних на сьогодні алгоритмів постсемантичної обробки є алгоритми реферування текстів. Алгоритми частотного реферування природномовних текстів мають ряд недоліків, через нехтування семантичною структурою тексту. В роботі розглянуто алгоритм асоціативного реферування, який працює з семантичною мережею тексту, що є виходом блоку семантичного аналізу. Вводиться поняття асоціативної ваги. Асоціативною вагою концепту є кількість дуг, інцидентних до вершини концепту в семантичній мережі тексту. Припускається, що найбільш важливими в тексті є слова-концепти з найбільшою асоціативною вагою. Алгоритм обходить семантичну мережу тексту, для кожного концепту-вершини обчислює асоціативну вагу, а потім оптимізує граф тексту, видаляючи вершини з малою вагою. Видаляючи вершини, алгоритм стежить за виконанням умови збереження зв'язності графу тексту. Зв'язність графу гарантує зв'язність тексту, тому алгоритм уникає ситуації створення ізольованих компонент графу. Алгоритм виконується за два обходи по вершинах семантичної мережі тексту. Перший прохід – зважування вершин, другий - видалення вершин з малою вагою, якщо це не суперечить умові зв'язності. Обидва проходи є рівними по часовій складності.

Була досліджена часова складність алгоритму асоціативного реферування.

Теорема 5.3. Час роботи процедури асоціативного зважування графу *G* пропорційний розміру представлення графу в виді списків суміжних вершин.

В заключному розділі також були детально розглянуті алгоритми підрахунку оптимальної ваги для вершин семантичного графу тексту та доведена їх коректність і ефективність.

Наприкінці описана система покращення якості текстів машинного перекладу "VitaminE", що використовує наведені вище алгоритми. Розглянуто приклади роботи програми. З системою було проведено ряд експериментів, результати яких подані у вигляді таблиць та графіків і свідчать про коректність та ефективність розроблених алгоритмів.

ВИСНОВКИ

В дисертації розроблені алгоритми онтологічного семантичного аналізу текстів природною мовою на основі обробки семантичного контексту та побудови асоціативно-семантичних шляхів в онтологічних мережах лінгвістичних систем, які розв'язують важливу задачу створення математично-програмних моделей багатьох процесів смислової обробки текстів, таких як семантичне реферування, індексування, смисловий переклад з однієї мови на іншу, підтримка семантично зв'язного природномовного діалогу, які мають істотне значення для розробки нових систем інтелектуального програмного забезпечення.

При виконанні роботи були одержані такі результати:

На основі принципу побудови асоціативно-семантичних шляхів в онтологічному графі бази знань розроблені нові алгоритми семантичного контекстного аналізу текстів.

На базі класичного онтологічного семантичного графу розроблена оригінальна архітектура бази знань в якій структурно інтегровані дані різних рівнів лінгвістичного аналізу.
Досліджена модель взаємодії паралельних процесів лінгвістичного аналізу, зокрема семантичного та синтаксичного аналізу.
За допомогою методів семантичного контекстного природномовного аналізу створений оригінальний алгоритм семантичного білінгвістичного аналізу текстів, що дало змогу розробити програмні засоби покращення якості текстів машинного перекладу "VitaminE".
Завдяки застосуванню алгоритмів пошуку найкоротшого шляху в графі при моделюванні смислової контекстної близькості в семантичному аналізі вдалося ефективно вирішити проблеми семантичної неоднозначності при обробці природномовних текстів.
В результаті створення мультилінгвістичної бази знань та білінгвістичної семантичної системи аналізу розроблена нова методика автоматизованої адаптації онтологій типу Wordnet до інших мов, яка була використана для локалізації до української та російської мов існуючої онтологічної семантичної бази знань.
Математично досліджені оцінки часової складності розроблених алгоритмів семантичного аналізу текстів, що підтвердило їх ефективність.

ОСНОВНІ РЕЗУЛЬТАТИ ДИСЕРТАЦІЇ ОПУБЛІКОВАНІ В ТАКИХ ПРАЦЯХ:

1. Марченко О.О. Дерев'янченко О.В. Застосування семантико-синтаксичної моделі для поліпшення розпізнавання рукописних текстів // Вісник Київського університету. Сер. фіз.-мат. науки.- 1999.-Вип. 4- С. 200-205.
2. Марченко О.О. Система обробки тексту природною мовою // Вісник Київського університету. Сер. фіз.-мат. науки.- 2001.-Вип. 5- С. 125-130.
3. Анісімов А.В. Марченко О.О. Нагорний В.А Створення керуючого простору синтаксичних структур природної мови // Вісник Київського університету. Сер. фіз.-мат. науки.- 2002.-Вип. 1- С. 159-169.
4. Анисимов А.В. Марченко А.А. Система обработки текстов на естественном языке // "Искусственный интеллект" НАНУ и ІПШІ- 2002. – Вип. 4 - С.157-164.

5. Анісімов А.В. Марченко О.О. Побудова керуючого простору синтаксичних структур речень природної мови // Міжнародна конференція "Моделювання та оптимізація складних систем-2001". Тези доповідей. –Київ – 2001 –С. 60-61
6. Марченко О.О. Система обробки природної мови // Міжнародна конференція "Обчислювальна та прикладна математика-2002". Тези доповідей. –Київ – 2002- С. 67
7. **Анисимов А.В. Марченко А.А. Система обработки текстов на естественном языке// Міжнародна конференція "Штучний інтелект-2002" Тези доповідей. - Кацивелі,Крим – 2002-С.157**

АНОТАЦІЇ

Марченко О.О. *Алгоритми семантичного аналізу природномовних текстів. – Рукопис.*

Дисертація на здобуття наукового ступеня кандидата фізико-математичних наук за спеціальністю 01.05.01 – теоретичні основи інформатики та кібернетики. – Київський національний університет імені Тараса Шевченка. – Київ, 2005.

Дисертацію 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. Він реалізований розробкою нових і модифікацією існуючих методів та алгоритмів, створенням відповідного програмного забезпечення. 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.

Були створені алгоритми семантичного білінгвістичного аналізу, що працюють на базі мультилінгвістичної онтології та дають змогу здійснювати смисловий переклад природномовних текстів.

Ключові слова: семантичний аналіз текстів природною мовою, контекстно-асоціативний аналіз текстів природною мовою, семантична онтологія, комп'ютерна лінгвістика, штучний інтелект.

Марченко А.А. *Алгоритмы семантического анализа естественно-языковых текстов. – Рукопись.* Диссертация на соискание ученой степени кандидата физико-математических наук по специальности 01.05.01 – теоретические основы информатики и кибернетики. – Киевский национальный университет имени Тараса Шевченко. – Киев, 2005.

Диссертация посвящена разработке и обоснованию эффективных методов семантического анализа и смысловой обработки текстов на естественном языке. В работе проведен анализ существующих методов семантической обработки текстов.

Детально изучена типичная архитектура современных программных систем лингвистического анализа. Исследуется роль блока семантического анализа в программах лингвистической обработки. Во всех масштабных лингвистических проектах блок семантической обработки вместе с семантическими базами знаний образует ядро, на основе которого разрабатываются прикладные пакеты реферирования текстов, индексирования, смыслового перевода, поддержки естественноязыкового диалога и многие другие программные приложения. Подавляющее большинство лингвистических задач завязаны на проблемах языковых неоднозначностей, которые решаются только на уровне семантики. Таким образом, блок семантического анализа берет на себя управление другими процессами анализа.

Особое внимание в работе уделено исследованию принципов взаимодействия процессов лингвистического анализа, которые нецелесообразно рассматривать исключительно как последовательные этапы обработки текста. Модели процессов анализа рассматриваются как некоторое пространство параллельных распределенных процессов с заданным отношением управления. Структуры данных разных уровней лингвистической обработки текстов интегрированы функционально и структурно в единой базе знаний, что делает схемы работы и взаимодействия алгоритмов анализа более эффективными. В работе детально описана оригинальная архитектура онтологической базы знаний, в которой семантические, синтаксические и лексические структуры интегрированы с помощью функциональных связей в единый ресурс онтологических лингвистических данных.

Разработана методика автоматизированной языковой локализации онтологических лингвистических баз знаний. С ее помощью была создана мультилингвистическая онтологическая база знаний на основе всемирно известной глобальной системы Wordnet.

Автором сформулирован концептуальный подход к ассоциативно-семантическому контекстному анализу естественного языка. Рассматриваются методы решения полисемических неоднозначностей в тексте на основе анализа семантических и ассоциативных связей между концептами в онтологической базе знаний. Особое внимание уделено проблеме формализации естественноязыкового контекста и его использование для решения проблем неоднозначной интерпретации текста. Был создан метод контекстного анализа, в основе которого лежит

моделирование естественно-языкового контекста с помощью разбора семантически близких концептов в онтологических базах знаний с использованием алгоритмов поиска кратчайших путей в графе. Исследуются алгоритмы поиска кратчайших путей, оптимизированные для структуры онтологического графа семантической базы знаний.

На основе алгоритмов ассоциативного контекстного анализа текстов на естественном языке разработан метод билингвистического анализа, который использует созданную мультилингвистическую онтологическую базу знаний. Созданный алгоритм программно реализован в системе улучшения качества текстов машинного перевода VitaminE. С программой был проведен ряд экспериментов, результаты которых доказывают корректность и эффективность разработанных алгоритмов естественноязыкового анализа.

Также исследован ряд алгоритмов для решения задач постсемантической обработки естественноязыковых текстов, таких как семантическое реферирование текстов, индексация, генерация семантического поискового образа. Алгоритмы постсемантической обработки анализируют семантическую сеть текста, полученную на выходе блока семантического анализа. Задача генерации реферата сводится к оптимизации полного графа текста удалением второстепенных вершин, не несущих основной смысловой нагрузки. При этом необходимо учитывать критерий сохранения целостности текста.

В рамках исследуемого ассоциативно-контекстного подхода процесс семантического анализа становится более прозрачным и эффективным. Ассоциативно-семантический анализ в системе управляет другими процессами лингвистической обработки текста и выходит на качественно новый уровень.

Ключевые слова: семантический анализ текстов на естественном языке, контекстно-ассоциативный анализ текстов, семантическая онтология, компьютерная лингвистика, искусственный интеллект.

Marchenko O.O. *Algorithms of the Semantic Analysis of Natural Language Text.—Manuscript.* Thesis for the degree of the Candidate of Physics and Mathematics. Science in speciality 01.05.01. – theoretical foundation for the informatics and cybernetics.- Taras Shevchenko Kyiv National University. - Kyiv, 2005.

The thesis deals with developing and basing efficient methods of the semantic analysis and semantic procession of natural language texts. Based on the results of analyses of the existing methods of semantic text procession, 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 procession.

The key words: semantic analysis of natural language texts, contest-associative analysis of natural language texts, semantic ontology, semantic text procession, computational linguistics, artificial intelligence.