НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ МАТЕМАТИЧНИХ МАШИН І СИСТЕМ

Вавіленкова Анастасія Ігорівна

УДК 004.421 (042.3)

МЕТОДИ ТА АЛГОРИТМИ АВТОМАТИЗОВАНОГО ФОРМУВАННЯ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ ТЕКСТОВОЇ ІНФОРМАЦІЇ

05.13.06 – Інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Київ – 2010

Дисертацією є рукопис.

Робота виконана на кафедрі комп'ютеризованих систем управління Національного авіаційного університету, м. Київ.

Науковий керівник — доктор технічних наук, професор
Литвиненко Олександр Євгенійович,
Національний авіаційний університет,
завідувач кафедри комп'ютеризованих систем управління

Офіційні опоненти: доктор технічних наук, професор
Литвинов Віталій Васильович,
Інститут проблем математичних машин і систем НАН України, завідувач відділу інтегрованих автоматизованих систем спеціального призначення

кандидат технічних наук
Сидорчук Надія Миколаївна,
Український мовно-інформаційний фонд НАН України, науковий співробітник відділу інформатики

Захист відбудеться "___" _____ 2011 р. о ___ годині на засіданні спеціалізованої вченої ради Д 26.204.01 в Інституті проблем математичних машин і систем НАН України за адресою: 03680, Київ-680, проспект Академіка Глушкова, 42.

З дисертацією можна ознайомитися у бібліотеці Інституту проблем математичних машин і систем НАН України за адресою: 03680, Київ-680, проспект Академіка Глушкова, 42.

Автореферат розісланий "___" _____ 2010 р.

Вчений секретар
спеціалізованої вченої ради В.І. Ходак

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Сучасний етап розвитку цивілізації привів до стрімкого росту інформаційних потоків. Людина у багатьох випадках уже не в змозі самостійно опрацювати всю необхідну для її діяльності інформацію. У зв'язку з цим широкого застосування набули комп'ютерні системи, що використовують методи автоматичного опрацювання природної мови. Створення великих сховищ даних та розвиток мережевих комунікацій робить актуальними задачі екстракції знань і формування коректно побудованих текстових документів, що вимагає створення засобів повнотекстового пошуку інформації, релевантної конкретному запиту користувача, а також порівняння текстів за їх змістом.

Задачі обробки текстової інформації виникли практично одразу після появи обчислювальної техніки. Проте, незважаючи на півстолітню історію досліджень у цій галузі штучного інтелекту (це роботи таких радянських, вітчизняних та зарубіжних вчених, як Попов Е.В., Поспєлов Д.А., Мельников Г.П., Гладкий А.В., Рубашкін В.Л., Грязнухіна Т.О., Дарчук Н.П., Кригін М.Ю., Ланде Д.В., Шабанов-Кушнаренко Ю.П., Широков В.А., Осуга С., Уено Х., Ісідзука М., Хомський Н.), розвиток інформаційних технологій та суміжних дисциплін, задовільного розв'язання більшості практичних задач аналітичної обробки текстової інформації поки що не існує. При опрацюванні цих задач виявилося, що комп'ютер не може вирішувати їх повністю, оскільки поки що не створено адекватних формалізованих моделей природномовних об'єктів, а розв'язання відповідних завдань містить неформальні, творчі елементи, властиві лише людині.

Існуючі на сьогодні засоби автоматичної обробки тексту для багатьох мов світу спроможні виконувати певні операції: здійснювати морфологічне маркування, виділяти частини мови, відмічати граматичні зв'язки, наприклад, дієслівні групи, певні синтаксичні відношення та ін. Більш глибока лінгвістична обробка ґрунтується на розв'язанні проблеми усунення лінгвістичних неоднозначностей, що зустрічаються в текстах. Подібні інструменти розглядаються як компоненти загальної системи розуміння природної мови. З їх допомогою тексти конвертуються в джерела інформації, доступні для опрацювання комп'ютером, що відкриває можливість для подальшої машинної обробки. Наприклад, ті з сучасних інформаційно-пошукових систем, які використовують ключові слова, надають релевантні – тією чи іншою мірою – документи. Проте, зазвичай, потрібні відповіді на запитання, а не документи, з яких тільки потенційно можна отримати відповіді.

Застосування універсальної системи підтримки лінгвістичних досліджень необхідне також для представлення текстових документів у формалізованому вигляді. З точки зору конструктивної семантики формалізація текстового документу передбачає перехід до оперування символами, при якому не потрібно додаткового аналізу речей об'єктивного світу і вся теорія розвивається в знаковій області, без звернення до досвіду, емпірії. Мова розробника моделі орієнтована на взаємодію конструктивної семантики з інженером по знаннях в процесі формування моделі.

В абсолютній більшості саме на користувача перекладається екстракція з отриманих документів корисної для нього інформації, тобто вилучення необхідних

знань. Для вирішення цієї задачі потрібні інструменти, що ґрунтуються на принципах розуміння природної мови, а це, у свою чергу, вимагає розроблення методів та засобів глибокої формалізації природномовних структур.

Зокрема, не вирішена проблема порівняльного аналізу електронних текстів, яка виникає щоразу, коли з'являється потреба у визначенні збігів або виявленні логічних протиріч у текстових документах. З проблемою визначення збігів у текстах людство зіштовхується у тих сферах своєї діяльності, де кінцевим результатом є текстовий документ. Це, в першу чергу: освіта, наука, законотворчість, патентування, інноваційна та інша діяльність, пов'язана з захистом інтелектуальної власності.

Друга проблема – виявлення логічних протиріч у текстових документах – лежить, головним чином, у площині професійних інтересів різних юридичних та інформаційно-аналітичних організацій і підрозділів. Особливу актуальність останнім часом вона набула у зв'язку з перспективою вступу України до Європейського співтовариства, що вимагає гармонізації українського законодавства та забезпечення його адекватності відносно загальноєвропейських нормативних актів.

Один із підходів до вирішення проблеми виявлення логічних протиріч у текстових документах є підхід, заснований на побудові логіко-лінгвістичної моделі тексту, що підлягає перевірці на логічну суперечливість відносно інших текстів. Основною проблемою на цьому шляху є автоматизація процесу побудови такої моделі.

Отже, актуальність теми дисертаційної роботи визначається необхідністю створення сучасних засобів автоматичної екстракції знань з природномовних текстів на основі побудови формальних моделей та алгоритмічної бази формалізованого опису структур природної мови.

Зв'язок роботи з науковими програмами, планами, темами. Основні дослідження за темою дисертації проводились у Національному авіаційному університеті в рамках виконання науково-дослідної роботи «Комп'ютерна технологія порівняльного аналізу електронних текстів» (№ 589-ДБ09, № ДР 0109U001771).

Мета і завдання дослідження. Метою дисертаційної роботи є створення логіко-лінгвістичних моделей текстової інформації, представленої у вигляді речень природної мови, та розробка алгоритмів автоматизованого формування таких моделей на основі використання механізмів синтаксичного й семантичного парсингу.

Досягнення поставленої мети передбачає розв'язання таких завдань:

1) створення уніфікованої форми логіко-лінгвістичної моделі речення;
2) формування речень тексту у вигляді формальної системи;
3) розробка методу автоматизованого формування логіко-лінгвістичних моделей на базі автоматизованого синтаксичного аналізу речень;
4) верифікація розробленого методу шляхом створення на його основі інтелектуальної системи автоматизованого формування логіко-лінгвістичних моделей.

Об'єктом дослідження є текстова інформація, представлена у вигляді речень природної мови.

Предметом дослідження є автоматизація процесу побудови логіко-лінгвістичних моделей текстової інформації.

Методи дослідження. Вибір методів дослідження даної дисертаційної роботи базується на вивченні сучасних технологій інтелектуальної обробки текстової інформації, екстракції знань, створенні формальних моделей, здійсненні лінгвістичного аналізу текстів. Відповідно в дисертації розглядаються методи обробки текстової інформації, методи здійснення автоматизованого синтаксичного та семантичного розборів, апарат теорії множин, формальних граматик та логіки предикатів, теорія баз даних та знань, методи розробки інтелектуальних систем.

Наукова новизна одержаних результатів.

Вперше:

- запропоновано уніфіковану форму логіко-лінгвістичних моделей, яка, на відміну від існуючих моделей представлення знань, охоплює всі можливі концептуальні відношення і здатна відображати синтаксичну структуру довільного речення природної мови, що створює теоретичну основу для автоматизованого вилучення знань з текстової інформації;
- створено систему продукцій, яка відображає правила формування словосполучень, визначення синтаксичних ролей та типів речень природної мови, що дозволяє автоматизувати процес встановлювання характеристик структурних одиниць тексту;
- розроблено метод автоматизованого формування логіко-лінгвістичних моделей текстової інформації, в основу якого покладено відповідність між формулами логіки предикатів та концептами реального світу.

Удосконалено синтаксичний аналізатор шляхом введення до його складу бази знань, побудованої на основі продукційної моделі визначення характеристик структурних одиниць природної мови, що дало змогу автоматизувати формування концептуальних зв’язків між елементами формальної системи незалежно від предметної області, що розглядається.

Дістала подальшого розвитку комп’ютерна технологія порівняльного аналізу електронних текстів за рахунок автоматизованого формування та використання логіко-лінгвістичних моделей документів, що аналізуються з метою виявлення змістовних збігів та логічних протиріч.

Практичне значення одержаних результатів. Результати, отримані в процесі виконання роботи, носять як теоретичний, так і прикладний характер:

1) загальна форма логіко-лінгвістичної моделі текстової інформації створює методологічну основу для побудови бази знань експертних систем порівняльного аналізу текстів, екстракції знань, класифікації та пошуку релевантної інформації різних документів;

2) застосування методу автоматизованого перетворення речень у логіко-лінгвістичну модель при проектуванні інтелектуальних систем дозволяє обирати для досліджень будь-яку предметну область;

3) представлення текстової інформації у вигляді алгебраїчних форм зменшує час написання програм, дозволяє здійснити автоматизований порівняльний аналіз текстів за змістом, добування знань з великих об'ємів текстів.

Отримані результати дослідження призначені для використання в галузі комп'ютерної лінгвістики, системах обробки текстової інформації та інших лінгвістичних технологіях.

Верифікацією методу автоматизованого перетворення речення природної мови в логіко-лінгвістичну модель є українськомовна аналітична система екстракції знань з електронних текстів.

Особистий внесок здобувача. Дана робота є узагальненням результатів теоретичних та експериментальних досліджень, виконаних автором самостійно. В роботі автору належить ідея застосування лінгвістичних засад для визначення основних принципів побудови логіко-лінгвістичних моделей. Автором здійснено розробку алгоритму формування логіко-лінгвістичних моделей текстової інформації та налаштування даних моделей на опрацювання текстів науково-публіцистичного стилю.

Апробація результатів дисертації. Матеріали дисертації доповідалися та обговорювалися на таких наукових конференціях:

- Міжнародна науково-технічна конференція "Комп'ютерні системи та мережні технології" (м. Київ, 2008-2009 рр.);
- IX та X Міжнародні наукові конференції молодих учених, аспірантів та студентів «Політ-2008» та «Політ-2009» (м. Київ, 2008-2009 рр.);
- Третя та четверта науково-практичні конференції з міжнародною участю «Математичне та імітаційне моделювання систем. МОДС 2008» та «Математичне та імітаційне моделювання систем. МОДС 2009» (м. Київ, 2008-2009 рр.);
- Міжнародна конференція «Горизонти прикладної лінгвістики та лінгвістичних технологій – MegaLing'2008» (м. Партеніт, 2008 р.);
- Міжнародна науково-технічна конференція «Інтелектуальні технології лінгвістичного аналізу» (м. Київ, 2008-2009 рр.);
- Всеукраїнська науково-методична конференція студентів та молодих науковців «Прикладна лінгвістика 2009: проблеми і рішення» (м. Миколаїв, 2009 р.);
- П'ята дистанційна конференція з міжнародною участю «Системи підтримки прийняття рішень. Теорія та практика. СППР 2009» (м. Київ, 2009 р.);
- Міжнародна конференція «Горизонти прикладної лінгвістики та лінгвістичних технологій – MegaLing'2009» (м. Київ, 2009 р.).

Публікації. За результатами виконаних досліджень опубліковано 21 наукову роботу, серед яких 10 наукових статей, 6 з них надруковано у фахових спеціалізованих наукових виданнях і збірниках наукових праць згідно з переліком ВАК України, та 11 тез доповідей на науково-технічних конференціях.

Структура дисертації. Робота складається із вступу, трьох розділів, висновків, списку використаних джерел з 73 найменувань. Обсяг дисертації становить 130 сторінок основного тексту, ілюстрованих 53 рисунками та 5 таблицями.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** подано обґрунтування актуальності теми дослідження, теоретична та практична цінність розглянутої тематики, сформульовано цілі та завдання дослідження, описано основні наукові результати, їх новизна, практична значущість і місце в розглянутому напрямку досліджень.

У першому розділі представлено аналітичний огляд існуючих методів обробки текстової інформації, таких як класифікація, кластеризація, автоматичне реферування, досліджено закони, що використовуються при застосуванні цих методів. У результаті виявлено неефективність їх використання за умови, що предметною галуззю досліджень є вся природна мова.

Проведено дослідження моделей представлення знань та здійснено їх порівняльну характеристику. Виявлено, що всі розглянуті моделі представлення знань та системи, розроблені на їх основі, працюють з предметною галуззю, яка є конкретною сферою життєдіяльності людини. При цьому окремі речення (й ситуації, з яких необхідно екстрагувати знання для подальшої обробки системою) повинні подаватися на вхід системи у певному впорядкованому вигляді. Більшість інтелектуальних систем працюють із заздалегідь заданими шаблонами або зразками, а на основі введеного слова (або декількох слів), виводять наперед заданий шаблон з підстановкою в нього конкретних слів із запиту. Отже, всі існуючі системи не є універсальними по відношенню до речення довільного типу.

Так, головний недолік представлення речення у вигляді семантичної мережі – це відсутність запитальних слів між службовими частинами мови та головними і другорядними членами речення; для фреймових структур характерна необмежена модифікація і знищення слотів, тому немає можливості створювати складні об’єкти на основі більш простих. У продукційних моделях представлення знань отриманий результат залежить лише від наперед заданих правил, і якщо система не знайде відповідного посилання, то результат не буде отримано; також відсутня стратегія управління, тобто розміщення правил у системі є випадковим, що уповільнює роботу та час пошуку необхідного посилання.

Для формалізації текстової інформації обрано логіко-лінгвістичні моделі як засіб відображення змісту речень природної мови та збереження в них всіх смислових зв’язків. Описано переваги та недоліки використання даного типу моделей для формування бази знань експертних систем управління.

Проаналізовано особливості аналітичних систем, що здійснюють автоматичний синтаксичний розбір (парсинг) та виявлено ряд перетворень для правильного функціонування механізму здійснення синтаксичного парсингу:

- пошук граматичних ідіом;
- лексико-граматичний аналіз речення з усуненням неоднозначності у визначенні частин мови;
- знаходження іменної групи об’єкта і суб’єкта;
- знаходження дієслівної групи;
- виділення головних та залежних речень.

Огляд стану проблем і досліджень за темою дисертації, зокрема аналіз існуючих систем обробки текстової інформації, принципів їх функціонування,

механізмів побудови та виявлення їх недоліків, обумовив необхідність вирішення наступних задач: проаналізувати існуючі методи інтелектуальної обробки текстової інформації, розробити єдиний принцип формування логіко-лінгвістичних моделей для речень природної мови; визначити загальну форму запису логіко-лінгвістичних моделей; розробити метод автоматизованого перетворення текстової інформації в логіко-лінгвістичну модель; створити аналітичну систему на базі вище вказаного методу, інструментом для створення якої буде синтаксичний аналізатор (так як саме синтаксичні зв'язки в реченні є ключем до вилучення змісту).

У другому розділі запропоновано загальну форму запису логіко-лінгвістичної моделі текстової інформації, яка охоплює концептуальні відношення, що можуть зустрітися в тексті, і є відображенням синтаксичної структури будь-якого речення природної мови.

Просте речення у формалізмі логіки предикатів – це атомарний предикат; складному реченню зіставляється складне логічне висловлювання, яке є сукупністю атомарних предикатів, поєднаних логічними зв'язками.

Нехай кожне речення S складається з множини слів M та множини простих речень $R(S)$. Тоді загальна форма логіко-лінгвістичної моделі набуває вигляду

$$(B_v \,\&\, C_v) \vee (B_v \to C_v) \vee (B_v \vee C_v) \vee (B_v \sim C_v) \vee A_v, \qquad (1)$$

де B_v і C_v - складні логічні висловлювання, які описують частину складного речення, що складається з p-тої кількості простих речень, $v = \overline{1, p}$, і може набувати вигляду (1), якщо множина простих речень $R(S)$ містить більше двох елементів. Якщо $R(S)$ містить два елементи, то вирази B_v і C_v представляють собою атомарні предикати;

A_v - просте логічне висловлювання, яке описує просте речення, для нього $R(S) = \{1\}$;

$(B_v \,\&\, C_v)$ - складний логічний вираз, в якому логічна зв'язка кон'юнкції $\&$ означає, що складові виразу B_v і C_v рівноправні за змістом;

$(B_v \to C_v)$ - складний логічний вираз, в якому логічна зв'язка імплікації $\to$ означає, що залежна частина речення C_v може уточнювати час, місце, причину, спосіб, про який йдеться в головній частині складнопідрядного речення B_v;

$(B_v \vee C_v)$ - складний логічний вираз, в якому логічна зв'язка диз'юнкції $\vee$ означає, що складові виразу B_v і C_v протиставляються або зіставляються;

$(B_v \sim C_v)$ - складний логічний вираз, в якому логічна зв'язка еквівалентності $\sim$ означає, що складові виразу B_v і C_v рівнозначні за змістом, тотожні.

Таким чином, логіко-лінгвістична модель (1) охоплює концептуальні відношення, які можуть зустрітися в текстовій інформації, і є відображенням синтаксичної структури будь-якого речення природної мови.

Якщо логічні висловлювання A_v, B_v і C_v являють собою атомарні предикати, їх можна представити за допомогою логічної формули, побудованої відповідно до функціональних відношень між об'єктами реального світу:

$$P(x_1\{\bigwedge_{d_1\in C_1(x_1)} c_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x_q\{\bigwedge_{d_2\in C_2(x_q)} c_{qd_2}\}]]), \quad (2)$$

де P - предикат, що відображає зміст речення;

x_1 - предикатна змінна (суб'єкт), знаходиться у предикативному відношенні з P;

c_{1d_1} - предикатна константа, що вказує на ознаку суб'єкта;

d_1 - номер предикатної константи, що вказує на ознаку суб'єкта;

$C_1(x_1)$ - множина предикатних констант суб'єкта x_1;

x_q - предикатна змінна (аргумент);

q - номер предикатної змінної (аргументу), початкове значення якого $q=2$;

$X(S)$ - множина предикатних змінних (аргументів)

c_{qd_2} - предикатна константа, що вказує на ознаку q-тої предикатної змінної (аргументу або об'єкта);

d_2 - номер предикатної константи, що вказує на ознаку предикатної змінної (аргументу);

$C_2(x_q)$ - множина предикатних констант предикатної змінної x_q;

$J(S)$ - множина предикатних змінних, які виконують у реченні рівнозначну роль, $J(S)\in X(S)$;

q_1 - номер предикатної змінної із множини $J(S)$; якщо речення не має ієрархічної будови або в ньому не зустрічаються аргументи, рівносильні за своєю роллю, то $J(S)=\varnothing$.

Логічна формула (2) є інтерпретацією синтаксичної структури тексту з урахуванням семантичних зв'язків, що є формальним засобом відображення змісту текстової інформації.

Сформовано основні принципи побудови логіко-лінгвістичних моделей, що базуються на синтаксичному парсингу речення, тобто визначенні зв'язків між усіма елементами формальної системи та встановлення їх синтаксичних ролей, що дає змогу зрозуміти зміст текстової інформації.

Принцип побудови логіко-лінгвістичної моделі (1)-(2) полягає в наступному:

1) визначити тип речення S, що розглядається, та множину простих речень $R(S)$, що входять до його складу;

2) проаналізувати множину простих речень $R(S)$ та концептуальні зв'язки між ними, що дає змогу визначити тип логічної зв'язки та кількість атомарних предикатів у формулі (1);

3) замість простих висловлювань у модель (1) підставити формулу (2);

4) встановити предикат P, що відображає зміст речення S, означає дію, стан або властивість суб'єкта і граматично йому підпорядкований;

5) встановити предикатну змінну (суб'єкт) x_1;

6) зафіксувати множину предикатних констант $C_1(x_1)$ для суб'єкта x_1;

7) визначити множину предикатних змінних (аргументів) $X(S)$;

8) визначити множину предикатних змінних $J(S)$, що виконують у реченні S рівносильну роль; число елементів цієї множини визначає кількість місць предиката P;

9) зафіксувати множину предикатних констант $C_2(x_q)$ для всіх аргументів з множини $X(S)$.

Згідно з принципом побудови, наведеним вище, логіко-лінгвістичні моделі для різних типів речень будуть мати такий вигляд:

1) просте речення:

$$P^1(x_1^1\{\bigwedge_{d_1\in C_1(x_1)} c^1{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^1{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^1{}_{qd_2}\}]]); \quad (3)$$

2) складне безсполучникове або складносурядне речення, в різних частинах якого йдеться про одночасність виконання дій і що складається з двох простих речень:

$$P^1(x_1^1\{\bigwedge_{d_1\in C_1(x_1)} c^1{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^1{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^1{}_{qd_2}\}]]) \,\&\, P^2(x_1^2\{\bigwedge_{d_1\in C_1(x_1)} c^2{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^2{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^2{}_{qd_2}\}]]); \quad (4)$$

3) складнопідрядне або безсполучникове речення, що складається з двох простих, одне з яких уточнює, конкретизує інше:

$$P^1(x_1^1\{\bigwedge_{d_1\in C_1(x_1)} c^1{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^1{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^1{}_{qd_2}\}]]) \to P^2(x_1^2\{\bigwedge_{d_1\in C_1(x_1)} c^2{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^2{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^2{}_{qd_2}\}]]); \quad (5)$$

4) складне речення, що включає в себе три простих, два з яких описують одночасність дій і протиставляються третьому простому реченню:

$$P^1(x_1^1\{\bigwedge_{d_1\in C_1(x_1)} c^1{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^1{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^1{}_{qd_2}\}]]) \,\&\, P^2(x_1^2\{\bigwedge_{d_1\in C_1(x_1)} c^2{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^2{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^2{}_{qd_2}\}]]) \vee P^3(x_1^3\{\bigwedge_{d_1\in C_1(x_1)} c^3{}_{1d_1}\}, \bigwedge_{q1\in J(S)}[\bigwedge_{q\in X(S)}[x^3{}_q\{\bigwedge_{d_2\in C_2(x_q)} c^3{}_{qd_2}\}]]). \quad (6)$$

Розроблено метод автоматизованого формування логіко-лінгвістичних моделей, який включає в себе декілька етапів, кожен з яких представляє собою складний механізм роботи формальної системи (рис. 1), а її елементи відіграють важливу роль для вилучення знань із текстової інформації. В основу методу покладено відповідність між формулами логіки предикатів та концептами, що належать до реального світу.

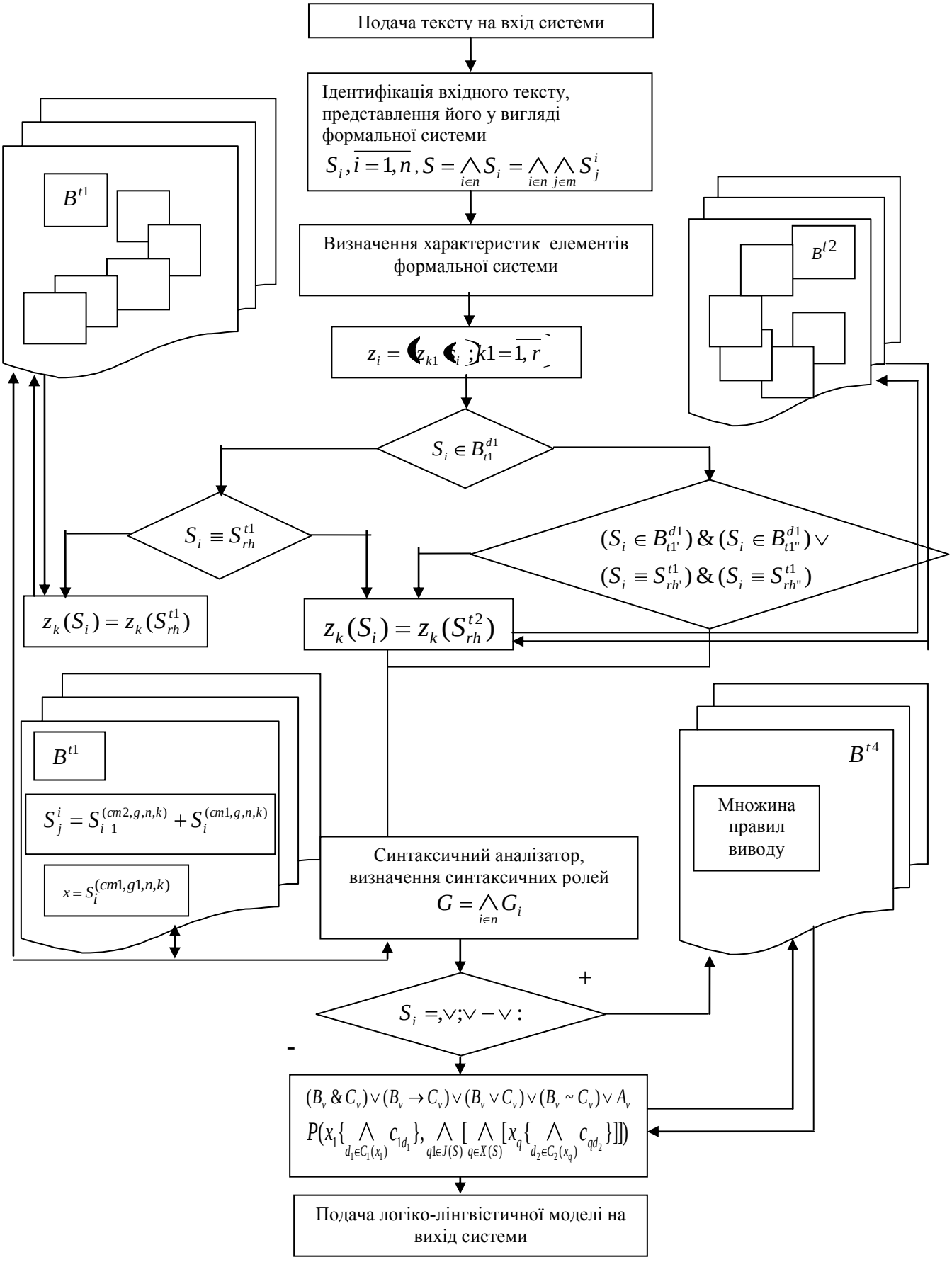

Рис. 1. Алгоритм автоматизованого формування логіко-лінгвістичних моделей

Основними етапами алгоритму є такі:

1. Ідентифікація вхідного тексту – розбиття вхідної текстової інформації на словоформи і представлення речення (об'єкта управління) у вигляді складної системи, що складається з n простих $S_i, \overline{i=1,n}$ та m складних взаємодіючих елементів $S_j^i, \overline{i=1,n}; \overline{j=1,m}$. Тобто на етапі ідентифікації вхідна текстова інформація, яка представляє собою сукупність синтаксичних одиниць природної мови, ототожнюється з формальною системою $S = \bigwedge_{i\in n} S_i = \bigwedge_{i\in n} \bigwedge_{j\in m} S_j^i$.

2. Концептуалізація (визначення характеристик елементів формальної системи) – відбувається експліціювання ключових понять, відношень та зв'язків між елементами, про які йшлося на етапі ідентифікації, а також визначення характеристик елементів, необхідних для опису подальшого процесу розв'язку поставленої задачі. Кожен простий елемент системи описується вектором значень характеристик $z_i = (z_{k1}(s_i); k1 = \overline{1, r})$, де r - кількість граматичних характеристик i-го елемента системи, $i = \overline{1, n}$. Процес визначення характеристик кожного елемента формальної системи представляє собою концептуальну схему, що задає множину можливих способів роботи на теоретичному рівні, припущень про природу і властивості елементів, які досліджуються.

3. Синтаксичний аналізатор, визначення ролей – на вхід аналізатора надходить масив простих елементів формальної системи та їх граматичних характеристик SZ. Кожен елемент цього масиву по черзі заноситься до робочої пам'яті бази знань, де за допомогою механізму логічного виводу зіставляється зі зразками бази правил. Робота синтаксичного аналізатора здійснюється за допомогою використання продукційної моделі як бази знань. Завдяки цьому стає можливим створення бази правил формування зв'язків та їх типів між елементами формальної системи незалежно від предметної галузі. Саме це забезпечує особливість методу автоматизованого формування логіко-лінгвістичних моделей і здійснює зв'язок методологічної бази лінгвістичних досліджень та принципів формування автоматизованих систем управління. База знань формальної системи описується у формі конкретних фактів та правил логічного виводу над базами даних та процедурами обробки інформації (зокрема, за допомогою правила «модус поненс»), що представляють собою відомості про синтаксичну будову речень природної мови в логічній формі.

4. Формалізація процесу побудови логіко-лінгвістичної моделі – відбувається безпосереднє формування логіко-лінгвістичної моделі (1)-(2), яка подається у вигляді сукупності одновимірних масивів елементів формальної системи, впорядкованих за певними правилами. При побудові моделі використовуються правила побудови складних елементів системи, а також правила визначення концептуальних зв'язків між атомарними предикатами моделі (1).

Третій розділ присвячено створенню інтелектуальної системи, що є програмною реалізацією методу автоматизованого формування логіко-лінгвістичних моделей. Підґрунтям для її створення є основні принципи проектування експертних систем управління.

Предметною галуззю цієї системи є природна мова. Тому система не базується на стандартних шаблонах, які містять основні слова – терміни певної предметної галузі, і користувач не обмежений у використанні лексики.

Оскільки зазначена інтелектуальна система є програмною реалізацією методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації, кожен її блок представляє собою реалізацію певних етапів згаданого методу. Саме такий підхід робить систему універсальною, можливою для використання спеціалістами різних галузей. Система де факто виконує верифікацію розробленого методу, отже вона є практичним підтвердженням правильності загальної ідеї методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації.

Середовищем розробки інтелектуальної системи автоматизованого формування логіко-лінгвістичних моделей текстової інформації є NetBeans IDE 6.7. Основний текст програми написано на об'єктно-орієнтованій мові програмування Java, що обумовлено простотою та потужністю, об'єктною орієнтованістю, інтерактивністю, архітектурною незалежністю, можливістю роботи з базами даних, написаних в SQL та Access, а також написанням запитів до бази даних на мові MySQL. Базою даних інтелектуальної системи автоматизованого формування логіко-лінгвістичних моделей (САФЛЛМ) є словник українських слів, представлений у вигляді реляційних таблиць.

Продемонстровано роботу системи автоматизованого формування логіко-лінгвістичних моделей. На основі отриманих результатів проведено аналіз логіко-лінгвістичних моделей різних типів речень української мови, написаних у науково-публіцистичному стилі (табл. 1).

Таблиця 1

Приклади роботи САФЛЛМ для різних типів речень природної мови

Речення природної мови	Тип речення	Логіко-лінгвістична модель
Неформалізовану задачу можна характеризувати двома способами	Просте, безособове	*можна_характеризувати*(, *задачу*{*неформалізовану*}, *способами*[*двома*]) $P_1\_P_2(, x_2\{c_{21}\}, x_3[x_4])$
Така організація управління дозволяє вирішувати неформалізовані задачі	Просте, особове	*дозволяє_вирішувати*(*організація*{*така*}[*управління*], *задачі*{*неформалізовані*}) $P_1\_P_2(x_1\{c_{11}\}[x_2], x_3[c_{31}])$
Для подолання труднощів, викликаних змінами проблемної області, використовуються методи пошуку в динамічному просторі	Просте, особове, ускладнене дієприкметниковим зворотом	*використовуються*(*методи*[*пошуку*[*просторі*{*динамічному*}], *подолання*[*труднощів*{*викликаних_змінами_проблемної_області*}]) $P(x_1[x_2[x_3[x_4\{c_{41}\}]]], x_5[x_6\{c_{61}\}])$

Продовж. табл. 1

Вчені знайшли вихід із ситуації: необхідно збільшити тиск, зменшивши температуру	Складне, безсполучникове	$знайшли(вчені, вихід[ситуації]) \rightarrow необхідно\_збільшити(, тиск, зменшивши\_температуру)$ $P^1(x_1^1, x_2^1[x_3^1]) \rightarrow P_1^2 \_ P_2^2(, x_2^2, C_1^2)$
Експерти знайшли вихід, щоб запобігти проблемі перевантаження ресурсів	Складно-підрядне	$знайшли(експерти, вихід) \rightarrow запобігти(, проблемі [перевантаження[ресурсів]])$ $P^1(x_1^1, x_2^1) \rightarrow P^2(, x_2^2[x_3^2[x_4^2]])$
Інженер по знаннях отримував дані, а експерт експериментував	Складне, сполучникове	$отримував(інженер\_знаннях, дані)$ & $експериментував(експерт)$ $P^1(x_{11}^1 \_ x_{12}^1, x_2^1)$ & $P^2(x_1^2)$
Кабінетом міністрів затверджено закон про те, що навчальні заклади закриваються на карантин	Складно-підрядне	$затверджено(Кабінетом[міністрів], закон) \rightarrow закриваються(заклади\{навчальні\}, карантин)$ $P^1(x_1^1[x_2^1], x_3^1]) \rightarrow P^2(x_1^2\{c_{11}^2\}, x_2^2)$
Життя без книги – хата без вікна	Просте	$хата(життя[книги], вікна)$ $P(x_1[x_2], x_3)$
Написано дисертацію – проведено величезну дослідницьку роботу	Складне безсполучникове	$написано(, дисертацію) \sim проведено(, роботу\{величезну, дослідницьку\})$ $P^1(, x_2^1) \sim P^2(, x_2^2\{c_{21}^2, c_{22}^2\})$
Єдина концепція побудови таких моделей дає змогу надалі виробити алгоритми порівняння текстів за змістом	Просте	$дає\_змогу\_виробити(концепція\{єдина\}[побудови[моделей\{таких\}]], алгоритми[порівняння[текстів[змістом]]])$ $P_1 \_ P_2 \_ P_3(x_1\{c_{11}\}[x_2[x_3\{c_{31}\}]], x_4[x_5[x_6[x_7]]])$

З табл. 1 видно, що всі отримані моделі побудовані за одним і тим самим принципом, вони відображають зміст поданих на вхід системи речень і кожен елемент формальної системи займає в моделі місце у відповідності до того, яку синтаксичну роль він виконує.

Результати такого експериментального дослідження можна застосовувати як базу знань для експертних систем порівняльного аналізу текстів, екстракції знань та пошуку протиріч у текстовій інформації.

Система автоматизованого формування логіко-лінгвістичних моделей представляє собою прикладну програму, яка дозволяє автоматизувати процес пошуку збігів та виявлення логічних протиріч у текстових документах.

Один із підходів до вирішення проблеми виявлення логічних протиріч у текстових документах є підхід, заснований на побудові логіко-лінгвістичної моделі тексту, що підлягає перевірці на логічну суперечливість відносно інших текстів. Основною проблемою на цьому шляху є автоматизація процесу побудови такої моделі.

САФЛЛМ являється програмною реалізацією методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації, розробленого в дисертаційній роботі, і дозволяє застосувати для виявлення логічних протиріч у текстах алгоритми доказу логічної суперечливості формальних моделей, представлених предикатами першого порядку. Також САФЛЛМ вирішує проблему екстракції знань із текстової інформації, представленої у вигляді речень природної мови.

Під знаннями, які дозволяє вилучити САФЛЛМ, розуміється форма представлення інформації, що пройшла трансформацію в соціальному середовищі. В кожному конкретному випадку трансформація здійснюється користувачем програми; інформація подається на вхід системи САФЛЛМ у вигляді речень природної мови, а знання вилучаються в рамках предметної області, про яку йдеться в цих реченнях.

Система автоматизованого формування логіко-лінгвістичних моделей представляє собою прикладну програму, що виступає засобом формування бази знань аналітичної системи порівняльного аналізу. Тобто система САФЛЛМ виконує функції допоміжного механізму, що використовується для визначення набору логіко-лінгвістичних моделей ключових речень, які максимально точно відображають зміст тексту, що підлягає подальшому аналізу, і вирішує проблему екстракції знань з текстової інформації.

Дослідження показали, що ні коди бібліотечних класифікаторів, ні назва текстового документу, ані множина слів, які найчастіше зустрічаються у тексті, у більшості випадків недостатньо адекватні або зовсім неадекватні його змісту. Тому при їх використанні у якості критерію добору текстів стандартний пошуковий сервер видає величезний обсяг інформації, більша частина якої немає ніякого відношення до тематики тексту, що підлягає аналізу. Тому відсоток отримання релевантної інформації в таких системах на запит користувача низький і складає порядку 79-80%.

Існуючі відкриті системи порівняльного аналізу текстової інформації, а також системи, що здійснюють повнотекстовий пошук та аналітичну обробку текстової інформації базуються на спільних механізмах вилучення знань з текстової інформації. Результати порівнянь функціонування таких систем дають змогу визначити переваги застосування САФЛЛМ, зокрема, її впровадження в систему порівняльного аналізу електронних текстів забезпечує підвищення відсотку відшукання збігів з врахуванням змісту на 10%.

Отже, система САФЛЛМ є прикладним засобом, що використовується в системі порівняльного аналізу електронних текстів та забезпечує порівняння текстів за змістом, що стає можливим завдяки використанню методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації.

ВИСНОВКИ

Суспільна потреба у розробці ефективних лінгвістичних технологій, на яких базуватимуться технології оперування знаннями, вимагає створення універсальної системи підтримки лінгвістичних досліджень та розробок. Зокрема, з метою представлення текстової інформації у формалізованій формі для можливості подальшого порівняння текстів, необхідне створення системи відшукання протиріч. Один із підходів створення такої системи базується на побудові логіко-лінгвістичної моделі текстової інформації, який підлягає перевірці на логічне протиріччя відносно других текстів. Основною проблемою на цьому шляху є автоматизація процесу побудови такої моделі.

Дисертаційна робота є закінченим науковим дослідженням, результатом якої стала розробка нової методології , направленої на вивчення проблем обробки інформації, представленої в мовній формі і пов’язаної з розробкою математичного та програмного забезпечення обчислювальних машин та систем.

В результаті виконання роботи було вирішено такі завдання:

1. Проаналізовано існуючі методи обробки текстової інформації та законы, що використовуються при їх застосуванні. Проведено дослідження моделей представлення знань та здійснено їх порівняльну характеристику. Для подальшої роботи обрано логіко-лінгвістичні моделі як засіб відображення змісту речень природної мови та збереження в них всіх смислових зв’язків. Описано переваги та недоліки використання даного типу моделей для формування бази знань експертних систем управління.

2. Сформовано загальну форму запису логіко-лінгвістичної моделі текстової інформації, що охоплює всі концептуальні відношення, які можуть зустрітися в тексті і являється відображенням синтаксичної структури як завгодно складного речення природної мови, що дозволяє вилучити з інформації знання; сформульовано основні принципи побудови таких моделей, що дозволяє будувати ЛЛМ як завгодно складних типів речень природної флективної мови;

3. Розроблено метод автоматизованого перетворення речень в логіко-лінгвістичну модель, який включає в себе декілька етапів, кожен з яких представляє собою складний механізм роботи формальної системи, а її елементи відіграють важливу роль для вилучення знань із тестової інформації. В основу методу покладено відповідність між формулами логіки предикатів та концептами, що належать реальному світу. При проектуванні інтелектуальних систем метод автоматизованого перетворення речення в логіко-лінгвістичну модель дозволяє вирішити проблему багатозначності в флективних мовах;

4. Розроблено систему автоматизованого формування логіко-лінгвістичних моделей текстової інформації, предметною областю якої являється вся природна, флективна мова. Тому дана система не базується на стандартних шаблонах, в які включені основні слова, що стосуються певної предметної області (наприклад, медицини, діагностування, продажу тощо), і користувач не повинен обмежувати себе у використанні певної термінології. Дана інтелектуальна система являється програмною реалізацією методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації, тому кожен її блок представляє собою реалізацію

певних етапів вище згаданого методу. Саме ця методика робить систему універсальною, можливою для користування спеціалістами різних галузей. Верифікацією методу автоматизованого перетворення речення природної мови в логіко-лінгвістичну модель являється україномовна аналітична система вилучення знань з електронних текстів.

Результати, отримані в процесі виконання даної роботи носять як теоретичний, так і прикладний характер:

1) загальна форма логіко-лінгвістичної моделі текстової інформації являється методологічною основою для бази знань експертних систем порівняльного аналізу текстів, вилучення знань, класифікації та пошуку релевантної інформації різних документів;

2) застосування методу автоматизованого перетворення речень в логіко-лінгвістичну модель при проектуванні інтелектуальних систем дозволяє вирішити проблему багатозначності в флективних мовах;

3) представлення текстової інформації у вигляді алгебраїчних форм зменшить час написання програм, дозволить здійснення автоматизованого порівняльного аналізу та перекладу текстів за змістом, вилучення знань з великого об'єму інформації.

Отримані результати дослідження призначені для використання в галузі комп'ютерної лінгвістики та системах обробки текстової інформації. Подальші дослідження, визначені запропонованим в дисертації напрямком, передбачають можливість застосування побудованих схем для формалізації задач перетворення інформації, пов'язаної з побудовою комп'ютерного та програмного забезпечення систем різних класів. Вони створюють новий інструмент дослідження цих складних проблем. Так логіко-лінгвістичні моделі можуть являтися базою знань для експертних систем в будь-якій предметній області, до яких можна застосовувати алгоритми порівняння та відшукання протиріч, що можливе завдяки використанню логіки предикатів при побудові ЛЛМ.

Загальна форма логіко-лінгвістичної моделі (1)-(2) – це формалізована форма запису будь-якого речення природної флективної мови. Автоматизована побудова таких моделей дає змогу простежити ланцюжок дій, за допомогою яких людиною формуються речення природної мови і відповідно зрозуміти їх зміст. Таким чином результати дослідження даної дисертаційної роботи являються засобом вирішення проблеми вилучення знань з текстової інформації.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Вавіленкова А.І. Аналіз моделей представлення знань в експертних системах управління /А.І. Вавіленкова // Проблеми інформатизації та управління: зб. наук. праць. – К.: НАУ, 2007. – Вип. 4(22). – С. 14–17.

2. Вавіленкова А.І. Побудова логіко-лінгвістичних моделей для складних речень /А.І. Вавіленкова // Наука і молодь. Прикладна серія: зб. наук. праць. – К.: НАУ, 2008. – С. 83-87.

3. Вавіленкова А.І. Особливості опитування експертів при побудові логічних моделей управління складними об’єктами /А.І. Вавіленкова // Математичні машини та системи. – 2009. – № 2. – С. 107-112.

4. Вавіленкова А.І. Побудова логіко-лінгвістичної моделі управління на основі результатів експертного опитування /А.І. Вавіленкова // Математичні машини та системи. – 2009. – № 3. – С. 157–163.

5. Вавіленкова А.І. Формалізація процесу формування логіко-лінгвістичної моделі представлення знань на основі синтаксичного парсингу речення /А.І. Вавіленкова // Прикладна лінгвістика та лінгвістичні технології MegaLing’2008: зб. наук. праць. – К.: Довіра, 2009. – С. 30–36.

6. Вавіленкова А.І. Автоматизація процесу побудови логіко-лінгвістичної моделі простого речення /А.І. Вавіленкова // Проблеми інформатизації та управління: зб. наук. праць. – К.: НАУ, 2008. – Вип. 2(24). – С. 28–31.

7. Вавіленкова А.І. Обробка текстової інформації через призму аналізу та інтерпретації елементів формальної системи /А.І. Вавіленкова // Системи підтримки прийняття рішень. Теорія і практика: зб. доп. наук.-практ. конф. з міжнар. участю.– Київ: ІПММС НАНУ, 2009. – С. 198–201.

8. Вавіленкова А.І. Розробка експертної системи на базі методу автоматизованого формування логіко-лінгвістичної моделі текстової інформації /А.І. Вавіленкова // Проблеми інформатизації та управління: зб. наук. праць. – К.: НАУ, 2009. – Вип. 3. – С. 14 –19.

9. Вавіленкова А.І. Логіко-лінгвістична модель як засіб відображення синтаксичних особливостей текстової інформації /А.І. Вавіленкова // Математичні машини та системи. – 2010. – № 2. – С. 134–137.

10. Вавіленкова А.І. Метод автоматизованого формування логіко-лінгвістичних моделей текстової інформації як об’єднуюча ланка інформаційних технологій та лінгвістики на шляху розуміння комп’ютером природної мови /А.І. Вавіленкова // Прикладна лінгвістика та лінгвістичні технології MegaLing’2009: зб. наук. праць. – К.: Довіра, 2009. – С. 100–103.

11. Вавіленкова А.І. Методика перетворення речення в логіко-лінгвістичну модель /А.І. Вавіленкова // Міжнар. наук.-техн. конф. «Комп'ютерні системи та мережні технології»: зб. тез. – К.: НАУ, 2008. – С. 49-53.

12. Вавіленкова А.І. Побудова логіко-лінгвістичних моделей для складних речень /А.І. Вавіленкова // Політ: зб. тез VIII міжнар. наук. конф. студентів та молодих учених. – К.: НАУ, 2008. – Т. 2.– С. 273.

13. Вавіленкова А.І. Особливості опитування експертів при побудові логічних моделей управління складними об’єктами /А.І. Вавіленкова // Третя наук.-практ. конф. з міжнар. участю «Математичне та імітаційне моделювання систем. МОДС 2008»: тези доповідей. – Київ, 2008. – С. 212–215.

14. Вавіленкова А.І. Формалізація процесу формування логіко-лінгвістичної моделі представлення знань на основі синтаксичного парсингу речення /А.І. Вавіленкова // Наукова конф. «Горизонти прикладної лінгвістики та лінгвістичних технологій»: доповіді міжнар. конф. – Сімферополь: ДИАЙПИ, 2008. – С. 148–149.

15. Вавіленкова А.І. Автоматизація процесу побудови логіко-лінгвістичної моделі простого речення /А.І. Вавіленкова // Міжнар. наук. конф. «Інтелектуальні технології лінгвістичного аналізу»: тези доповідей. – К.: НАУ, 2008. – С. 15.

16. Вавіленкова А.І. Аналіз речення як формальної системи /А.І. Вавіленкова// IX міжнар. наук. конф. студентів та молодих учених «Політ»: зб. тез. – К.: Нау-друк, 2009. – С. 230.

17. Вавіленкова А.І. Засоби та механізми перетворення текстової інформації /А.І. Вавіленкова // Тези всеукр. наук.-метод. конф. студентів та молодих науковців «Прикладна лінгвістика 2009: проблеми і рішення». – Миколаїв: НУК, 2009. – С. 8–10.

18. Вавіленкова А.І. Логіко-лінгвістична модель як засіб відображення синтаксичних особливостей текстової інформації /А.І. Вавіленкова // Четверта наук.-практ. конф. з міжнар. участю «Математичне та імітаційне моделювання систем. МОДС 2009 »: тези доповідей. – Київ, 2009. – С. 188–190.

19. Вавіленкова А.І. Розробка експертної системи на базі методу автоматизованого формування логіко-лінгвістичної моделі текстової інформації /А.І. Вавіленкова // Збірник тез II Міжнар. наук.-техн. конф. «Комп'ютерні системи та мережні технології» (CSNT–2009). – К.: НАУ, 2009. – С. 22.

20. Вавіленкова А.І. Система автоматизованого формування логіко-лінгвістичних моделей текстової інформації /А.І. Вавіленкова // Міжнар. наук.-техн. конф. «Інтелектуальні технології лінгвістичного аналізу»: тези доповідей. – К.: НАУ-друк, 2009. – С. 14.

21. Вавіленкова А.І. Метод автоматизованого формування логіко-лінгвістичних моделей текстової інформації як об’єднуюча ланка інформаційних технологій та лінгвістики на шляху розуміння комп’ютером природної мови /А.І. Вавіленкова // Наукова конф. «Горизонти прикладної лінгвістики та лінгвістичних технологій». Матеріали міжнар. наук. конф. MegaLing’2009. – К.: Довіра, 2009. – С. 55.

АНОТАЦІЯ

Вавіленкова А.І. Методи та алгоритми автоматизованого формування логіко-лінгвістичних моделей текстової інформації. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – Інформаційні технології. – Інститут проблем математичних машин і систем НАН України, Київ, 2010.

Метою даної дисертації є розробка загальної методології вирішення проблем обробки текстової інформації та вилучення знань на основі використання логіки предикатів і механізмів здійснення автоматизованого синтаксичного й семантичного розбору, перетворення інформації, представленої у вигляді речень природної мови, до алгебраїчних форм.

Сформовано загальну форму запису логіко-лінгвістичної моделі текстової інформації, яка охоплює всі концептуальні відношення. Розроблено метод автоматизованого перетворення речень у логіко-лінгвістичну модель, який включає

в себе декілька етапів, кожен з яких являє собою складний механізм роботи формальної системи, а її елементи відіграють важливу роль для вилучення знань із тестової інформації. Створено інтелектуальну систему, що являється програмною реалізацією методу автоматизованого формування логіко-лінгвістичних моделей текстової інформації, кожен її блок представляє собою реалізацію певних етапів вище вказаного методу.

Ключові слова: логіко-лінгвістичні моделі, метод, інтелектуальна система, текстова інформація, природна мова.

АННОТАЦИЯ

Вавиленкова А.И. Методы и алгоритмы автоматизированного формирования логико-лингвистических моделей текстовой информации. – Рукопись.

Диссертация на соискание научной степени кандидата технических наук по специальности 05.13.06 – Информационные технологии. – Институт проблем математических машин и систем НАН Украины, Киев, 2010.

Общественная проблема в разработке эффективных лингвистических технологий, на которых будут базироваться технологии оперирования знаниями, требует создания универсальной системы поддержки лингвистических исследований и разработок. Основной проблемой на этом пути является автоматизация процесса построения такой модели.

Целью данной диссертации является разработка общей методологии решения проблем обработки информации, а также извлечения знаний на основании использования логики предикатов и механизма осуществления автоматизированного синтаксического и семантического разбора информации, представленной в виде предложений на естественном языке, к алгебраическим формам.

Достижение поставленной цели возможно при решении следующих задач:

1) создание общей формы логико-лингвистической модели предложений текстовых документов;
2) представление текста в виде формальной системы;
3) формирование алгоритмов решения задачи извлечения знаний из текстовой информации;
4) разработка метода автоматизированного формирования логико-лингвистических моделей на основе синтаксического разбора предложений;
5) создание системы автоматизированного формирования логико-лингвистических моделей на основе разработанного метода.

Проведен анализ существующих методов обработки информации, а также законов, которые используются при их применении. Исследованы модели представления знаний и сделана их сравнительная характеристика.

Сформирована общая схема записи логико-лингвистической модели текстовой информации, которая включает в себя все концептуальные отношения, возможные в тексте, и является отображением синтаксической структуры как угодно сложного

предложения естественного языка, что позволяет извлечь из информации знания. Сформулированы основные принципы построения таких моделей.

Разработан метод автоматизированного преобразования предложений в логико-лингвистическую модель, который включает в себя несколько этапов. Каждый из них представляет собой сложный механизм работы формальной системы, а ее элементы играют важную роль для извлечения знаний из текстовой информации.

Основу метода составляет соответствие между формулами логики предикатов и концептами, которые принадлежат реальному миру:

1) идентификация входящего текста – разбитие входной информации на словоформы и представление предложения (объекта управления) в виде сложной организационной системы;

2) концептуализация, определение характеристик элементов формальной системы – осуществляется экспицирование ключевых понятий, отношений и связей между элементами формальной системы;

3) синтаксический анализатор, определение ролей – на вход анализатора подается массив простых элементов формальной системы и их грамматических характеристик, после чего с помощью механизма логического вывода базы знаний определяются синтаксические роли каждого элемента;

4) формализация процесса построения логико-лингвистической модели – непосредственное формирование логико-лингвистической модели, которая физически представляет собой одномерные массивы элементов формальной системы, упорядоченные по определенным правилам.

Создана интеллектуальная система, которая является программной реализацией метода автоматизированного формирования логико-лингвистических моделей текстовой информации, каждый блок которой представляет собой реализацию конкретных этапов выше упомянутого метода. Именно эта методика делает систему универсальной и возможной для использования специалистами различных отраслей. Аналитическая система извлечения знаний их электронных текстов для украинского языка является верификацией метода автоматизированного преобразования предложения естественного языка в логико-лингвистическую модель.

Полученные результаты исследования предназначены для использования в отрасли компьютерной лингвистики и систем обработки текстовой информации. Дальнейшие исследования, определенные предложенным в диссертации направлением, подразумевают возможность использования построенных схем для формализации задач преобразования информации, связанные с построением компьютерного и программного обеспечения систем разных классов.

Автоматизированное построение таких моделей дает возможность проследить цепочку действий, с помощью которой человеком формируются предложения естественного языка и соответственно их смысл. Таким образом, результаты исследования данной диссертационной работы являются одним из способов решения проблемы извлечения знаний из текстовой информации.

Ключевые слова: логико-лингвистические модели, метод, интеллектуальная система, текстовая информация, естественный язык.

ABSTRACT

Vavilenkova A.I. Methods and algorithms of the automated forming logico-linguistic models of text information. – Manuscript.

Dissertation on the receipt of scientific degree of candidate of engineerings sciences after speciality 05.13.06 – Information technologies. – Institute of problems of mathematical machines and systems of NAC of Ukraine, Kyiv, 2010.

The purpose of this dissertation is development of general methodology of decision of problems of treatment of text information and exception of knowledges on the basis of the use of quantificational and mechanisms of realization of the automated syntactic and semantic analysis, transformation of information, human language presented as suggestions logic, to the forms of algebra.

The general form of record is formed logiko-linguistic to the model of text information, which engulfs all conceptual relations. The method of the automated transformation of suggestions is developed in logico-linguistic model, which includes for itself a few stages, each of which is a difficult mechanism of work of the formal system, and its elements play an important role for the exception of knowledges from test information. The intellectual system which appears programmatic realization of method of the automated forming logico-linguistic models of text information is created, that is why every its block is realization of the certain stages of the higher mentioned method.

Key words: logico-linguistic models, method, intellectual system, text information, human language.