НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Касілов Олег Вікторович

УДК 681.3.01

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОЇ ПЕРЕРОБКИ ТЕКСТОВОЇ ІНФОРМАЦІЇ ПРИ СТВОРЕННІ ЕЛЕКТРОННИХ СЛОВНИКІВ-ТЕЗАУРУСІВ

Спеціальність 05.13.06 – Інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2008

Дисертацією є рукопис.

Робота виконана на кафедрі систем інформації в Національному технічному університеті «Харківський політехнічний інститут» Міністерства освіти і науки України, м. Харків.

Науковий керівник: кандидат технічних наук, професор
Кравець Валерій Олексійович,
Національний технічний університет
«Харківський політехнічний інститут»,
завідувач кафедри систем інформації

Офіційні опоненти: доктор технічних наук, професор
Шаронова Наталія Валеріївна,
Національний технічний університет
«Харківський політехнічний інститут»,
завідувач кафедри інтелектуальних комп'ютерних систем

доктор технічних наук, професор
Замаруєва Ірина Вікторівна,
Військовий інститут
Київського національного університету
ім. Тараса Шевченка,
професор кафедри інформаційно-психологічного
протиборства

Захист відбудеться 11 грудня 2008 р. о 14-30 годині на засіданні спеціалізованої вченої ради Д 64.050.07 у Національному технічному університеті «Харківський політехнічний інститут», за адресою: 61002, м. Харків, вул. Фрунзе, 21.

З дисертацією можна ознайомитись у бібліотеці Національного технічного університету «Харківський політехнічний інститут», за адресою: 61002, м. Харків, вул. Фрунзе, 21.

Автореферат розісланий « 7 » листопада 2008 р.

Вчений секретар
Спеціалізованої вченої ради І.П. Гамаюн

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність роботи. Інформація служить визначальним чинником розвитку економічної, технічної і наукової сфер людської діяльності. В сучасному світі жодну значну проблему зараз неможливо розв'язати без підключення лінгвістичного забезпечення в інформаційних процесах.

Інформація, яка циркулює в інформаційних системах, існує в конкретних формах: у вигляді даних, знань, фреймів, скриптів, текстів, гіпертекстів та ін. Для людей найбільш прийнятною формою інформації, що використовується, є природна мова. Тому при створенні сучасних інформаційних систем постійно зростають вимоги до дружнього інтерфейсу і, як наслідок, зростає потреба в дослідженнях прикладної лінгвістики, тобто існує об'єктивна необхідність у тісній взаємодії наук інформатики і лінгвістики.

Комп'ютерна лінгвістика займається вивченням формальних властивостей природної мови за допомогою ЕОМ і моделюванням процесів аналізу, синтезу і розуміння природномовних текстів на ЕОМ. Саме могутні комп'ютерні словники становлять основу всіх інформаційних систем, в яких використовується природна мова. Очевидно, що інформатизація лінгвістичних досліджень і особливо лексикографічних робіт в Україні – це необхідність часу, і від ефективного їх вирішення залежить темп просування України до інформаційного суспільства. Характерною особливістю сучасних програм, які працюють з об’єктами природної мови, є зростання рівня їх інтелектуальності. Розробка таких програмних засобів пов’язана з використанням методів, що традиційно належать до теорії і практики штучного інтелекту в комп'ютерній лінгвістиці.

Аналіз існуючих систем обробки текстової інформації, систем перекладу, пошукових систем, систем розуміння текстів, написаних природною мовою, показав, що існує необхідність у формуванні екстралінгвістичних знань, знань про навколишній світ, знань у конкретній області. Подібна інформація зосереджена в ідеографічних словниках. Тому розробка методики, математичного апарата опису структурованих текстів, алгоритмів і інформаційних технологій автоматизації процесу створення електронних словників є актуальною задачею та складає напрямок дисертаційного дослідження.

Зв’язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалася відповідно до плану наукових досліджень Національного технічного університету «ХПІ» у рамках державної теми «Створення експериментальної платформи з розробкою перспективних технологій інформаційного обміну між установами освіти і розробка методик впровадження сучасних інформаційних технологій в навчальному процесі» (№ ДР 0103U001541, МОН України), де здобувач був виконавцем окремих розділів.

Мета і задачі дослідження. Метою роботи є підвищення ефективності розробки електронних словників шляхом побудови інформаційної технології спеціалізованої обробки текстів природної мови через удосконалення

лінгвістичного процесора для перетворення словників у спеціальну електронну форму на базі модифікованої мови розмітки XML (eXtensible Markup Language). Для досягнення вказаної мети в роботі поставлено наступні завдання:

1) провести аналіз завдань обробки текстової інформації, що використовують для свого розв'язання електронні словники, а також засобів автоматизації створення електронних словників;

2) дослідити структуру представлення даних у паперових словниках і розробити засоби подання текстів у спеціальній електронній формі;

3) сформувати вимоги до мови розмітки даних на прикладі словника-тезауруса і надати рекомендації користувачеві з настройки системи;

4) дослідити структуру лексикографічного процесора і запропонувати модифікацію його модулів розмітки;

5) розробити інформаційну технологію перетворення текстів природної мови в спеціальну електронну форму з використанням удосконаленої мови розмітки;

6) провести перевірку працездатності і ефективності розробленої інформаційної технології на прикладі створення електронного словника-тезауруса.

Об'єкт дослідження: структуровані тексти природної мови, їх електронні версії і зв'язок між ними, процеси створення електронних словників-тезаурусів.

Предмет дослідження: моделі, методи, інформаційні технології перетворення структурованих текстів природної мови в електронну форму.

Методи дослідження: теорія формальних аналітичних граматик використовується для визначення за текстовою послідовністю належність її до певної мови. Синтаксичні діаграми використовуються для побудови відповідних скінченних автоматів, які застосовуються для розпізнавання текстових послідовностей в лексикографічному процесорі. Правила продукцій використовуються для опису послідовності перетворення вхідних текстових даних у вихідну структуровану послідовність даних. Апарат скінченних автоматів використовується для опису дій лексико-синтаксичного обробника текстових даних. Теорія графів використовується для опису роботи модуля розмітки лексикографічного процесора обробки текстової інформації.

Наукова новизна отриманих результатів. У дисертаційному дослідженні розв'язано науково-прикладне завдання удосконалення лінгвістичного процесора для перетворення словників у спеціальну електронну форму на базі модифікованої мови розмітки. Наукова новизна роботи полягає в наступному:

вперше розроблено спеціалізовану інформаційну технологію, що реалізує автоматизоване перетворення структурованих текстів природної мови в їх електронну форму, що дозволило підвищити ефективність лексикографічного процесора в порівнянні з напівавтоматичною обробкою: час обробки тексту при автоматичному режимі зменшується в порівнянні з ручним режимом у $5 \cdot 10^6$ разів, а кількість помилок на 1000 рядків зменшується в 6 разів;

удосконалено модифікацію мови відкритої розмітки тексту XML для представлення словників-тезаурусів в електронній формі за рахунок введення

спеціальних елементів розмітки, що дозволило уніфікувати і погоджувати документ, поданий у такій розмітці, з іншими системами розмітки і зберігання документів. Розроблено модуль XML розмітки лексикографічного процесора, що дозволило автоматизувати процес перетворення текстової інформації, яка міститься у словнику, в спеціальну електронну форму;

отримала подальший розвиток методика перетворення XML опису словника в базу даних різних форматів, що дозволяє провести правильне і однозначне портування даних в різні системи, які використовують електронні словники.

Практичне значення отриманих результатів. Застосування розроблених інформаційних технологій перетворення природномовних текстів при створенні електронного словника-тезауруса дозволяє інтегрувати результати роботи (електронні словники-тезауруси) в різноманітні інформаційно-пошукові системи, гіпертекстові інформаційні системи, бібліотечні системи, системи класифікації та рубрикації текстів, системи машинного перекладу, бази знань, підсистеми обробки природномовних текстів та інші, що потребують обов'язкового включення словників до їх складу.

Результати роботи впроваджено в Інституті проблем машинобудування НАН України (м. Харків) при створенні бази даних технічних статей і довідників, які застосовуються у відділах інституту з використанням розробленого лексикографічного процесора (акт впровадження від 24.04.2008). Результати роботи використані у навчальному процесі на кафедрі «Системи інформації» НТУ «ХПІ» при викладанні курсу «Природномовні інтелектуальні системи» (акт впровадження від 15.05.2008). Розроблена комп'ютерна програма «Модуль XML-разметки лексикографического процессора обработки текстовой информации» (свідоцтво про реєстрацію авторського права № 22259, Державний департамент інтелектуальної власності (ДДІВ) від 05.10.2007).

Особистий внесок здобувача. Усі основні результати отримані здобувачем особисто. В роботі, яка написана в співавторстві, здобувачу належить огляд мов розмітки і вибір мови розмітки XML як елемента системи перетворення природномовних текстів, що розробляється, при створенні електронного словника-тезауруса, виділені основні модулі розробленої системи. Особистий внесок здобувача полягає у розробці інформаційної технології автоматизованого перетворення природномовних текстів при створенні електронного словника-тезауруса. Показана можливість використання скінченних автоматів в задачах розпізнавання та класифікації. Розроблено моделі лексикографічної системи, лексикографічного процесора та модель словника-тезауруса. На основі цих моделей було обґрунтовано вибір методів обробки структурованих даних, що містяться в словниках-тезаурусах, а також були запропоновані правила перетворення елементів структурованих текстів до модифікованої XML – розмітки. Запропонована модифікація мов розмітки структурованих текстів XML для словників-тезаурусів. Створено програмне забезпечення лексикографічного процесора, що дозволяє здійснювати синтаксичний аналіз текстової інформації, яку представлено в процедурній розмітці. Здійснено перевірку роботи програмного забезпечення.

Апробація результатів дисертації. Основні положення і результати дисертаційної роботи доповідались на міжнародних науково-технічних конференціях: «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я» (Харків, 2002, 2004 рр.); першій обласній конференції молодих учених «Тобі, Харківщино – пошук молодих» у рамках обласного форуму «Освіта, наука, виробництво – шляхи інтеграції» (Харків, 2002 р.); наукових семінарах НАН України «Ідентифікація і моделювання поведінки об’єктів в умовах електромагнітних випромінювань» (Харків, 2002–2007 рр.).

Публікації. Основний зміст дисертаційної роботи опублікований в 5 наукових працях, з яких 4 – статті у фахових наукових виданнях ВАК України.

Структура й обсяг дисертації. Робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Повний обсяг друкованого тексту дисертації становить 151 сторінку, у тому числі 20 рисунків та 5 таблиць за текстом, 6 рисунків та 3 таблиці на 10 сторінках, 3 додатки на 3 сторінках і список використаних джерел із 206 найменувань на 17 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

Вступ розкриває суть проблеми. У ньому обґрунтовується актуальність теми, формулюється мета роботи, завдання, об’єкт та предмет, визначено методичну базу дослідження, висвітлено наукову новизну та практичне значення.

У **першому** розділі проведено огляд існуючих систем автоматизації лексикографічних задач, сучасний стан, типи та формати електронних словників. Невирішеними є наступні задачі:

1. Задача автоматизації створення електронних словників, оскільки сучасні системи інформаційного пошуку, системи реферування, витягання знань, класифікації даних, системи машинного перекладу для поліпшення якості своєї роботи не можуть обходитися без використання спеціалізованих словників.

2. Необхідно створення лексикографічного процесора в цілому, або деяких його компонентів, для автоматизації процесу перетворення паперових словників в електронну форму і створення електронного словника.

3. Існуючих на даний момент електронних словників, які можуть бути інтегровані в інформаційно-пошукові системи, системи анотування, реферування і перекладу, недостатньо, а їх формат даних розрізнений, що робить важким їхнє використання в різних системах. Необхідно вирішити питання розробки універсального формату зберігання електронного словника, який дозволить інтегрувати цей словник в інші лінгвістичні системи.

4. Найбільш потрібними є словники-тезауруси, оскільки вони є основою для складніших словарних систем, дозволяючи обробляти смислові компоненти текстів.

5. Більшість існуючих методів розпізнавання складні для реалізації або їх програмна реалізація відсутня, це стосується, у тому числі, і питань, які пов’язані з обробкою текстової інформації в лінгвістичних задачах. Тому існує

необхідність використання методів розпізнавання, заснованих на простих, швидких і зручних в реалізації методах.

У **другому** розділі наведено узагальнене поняття моделювання лексикографічних систем, визначено проблеми й завдання моделювання таких систем, досліджено особливості моделювання словника-тезауруса, запропоновано механізм розмітки текстів у документах.

Сьогодні для формалізації теоретичних питань лексикографії широко використовується інформаційне моделювання лексикографічних ефектів. Під цим моделюванням розуміється наступне: інформаційна модель, що відповідає певній системі *S* довільної природи, повинна відображати лексикографічні аспекти системи *S*. В результаті такого моделювання формується лексикографічна система *LS*, що відображає елементи системи *S*, при цьому отримують структуровану лексикографічну систему, відповідну деякій початковій системі.

Словником є спеціальний вид тексту, в якому у систематизованому і структурованому вигляді представляється опис лексики певної мови в одному випадку, або лексики конкретної області – в іншому випадку. Але словник слід розглядати і як специфічний об'єкт техніки, а саме – інформаційну систему, в якій за допомогою поліграфічного виконання в паперовій копії або за допомогою елементів розмітки в електронному форматі позначені певні лінгвістичні ефекти. Ними можуть бути: шрифтові виділення, позиційні розміщення, спеціальні позначення і інше, які відіграють роль ідентифікаторів відповідних інформаційних змін.

Розробка програмних комплексів, орієнтованих на вирішення лексикографічних завдань, вимагає набору адекватних формальних інформаційних моделей словарних систем, які могли б відіграти роль початкової концептуальної бази для програмування лексикографічного процесора або його складових частин. Інтерпретація лексикографічних систем як інформаційних систем спеціального типу вимагає дотримання єдиної методології їх проектування. В першу чергу визначається загальна архітектура системи, розробляються складові частини цієї архітектури, зв'язки і відображення між ними. Відповідно до ANSI/X3/SPARK в архітектурі інформаційної системи виділяються три рівні опису даних: концептуальний, внутрішній і зовнішній.

У символьній формі архітектура лексикографічної системи представлена у такому вигляді

$$ALS = \{KM, EXM, INM, F, G, H, \Sigma\},$$

де *KM* – концептуальна модель лексикографічної системи *LS*;

EXM = {*exm*} – множина зовнішніх моделей *LS*, які відповідають даній концептуальній моделі *KM*;

INM = {*inm*} – відповідна множина внутрішніх моделей *LS*.

F = {*f*} – множина відображень *KM* в *EXM*; $f : KM \rightarrow exm$, де $exm \in EXM$;

G = {*g*} – множина відображень *KM* в *INM*: $g : KM \rightarrow inm$, де $inm \in INM$;

H = {*h*} – множина відображень *INM* в *EXM*;

$\Sigma = \{\sigma\}$ – множина відображень *EXM* в *INM*.

Одній концептуальній моделі може відповідати декілька внутрішніх і зовнішніх моделей, тобто відображення *f* і *g* є неоднозначними. За визначенням, на множині моделей $(\forall\ inm \in INM\ \forall exm \in EXM)\ \{\exists\ h \in H: h\ (inm) = exm\}$.

При цьому відображення *f*, *g*, *h* будуються так, що діаграма (рис. 1):

Рис. 1

є комутативною: $g \circ h = f,\ f \circ y = g$. Вимога комутативності цієї діаграми є суттєвою, оскільки вона гарантує узгодженість між всіма рівнями архітектури системи.

Концептуальну модель можна записати

$$KM = \{Ob(LS); PrOb(LS); ConOb(LS), FunOb(LS); ConDiscrOb(LS)\},$$

де $Ob(LS) = \{AL(L),\ S(L),\ T(L),\ \ldots\}$ – множина об'єктів лексикографічної системи;

PrOb(*LS*) – множина представлень об'єктів лексикографічної системи;

Con(*LS*) – множина зв'язків між об'єктами лексикографічної системи;

FunOb(*LS*) – множина функцій (відношень) між об'єктами лексикографічної системи;

ConDiscrOb(*LS*) – множина обмежень цілісності лексикографічної системи.

Внутрішня модель лексикографічної системи має таку структуру

$$INM = \{DS, ALG, OS, PL\},$$

де *DS* – множина типів і структур даних (у даній роботі розглядається текстовий тип);

ALG – множина алгоритмів обробки даних;

OS – операційне середовище;

PL – множина мов програмування.

Зовнішня модель лексикографічної системи представляється у вигляді

$$EXM = \{SC, APR\},$$

де *SC* – множина сценаріїв;

APR – множина прикладних програм.

Файлом настройки є сценарій виконання прикладної програми.

Прикладна програма – це інструментальний засіб кінцевого користувача лінгвіста-лексикографа, що формує конкретний лексикографічний комплекс програм. Очевидно, що для такого користувача при такій архітектурі *LS* не потрібні знання в області об'єктів внутрішньої моделі *INM*.

Словник як абстрактна лексикографічна система має структуру, яка містить дві необхідні частини: ліва (реєстрова) і права (інтерпретаційна). Будь-яка лексикографічна система (або її реалізація) в інформаційно-лексикографічній моделі набуває такого вигляду

$$V(\lambda) = \{\Lambda(\lambda); P(\lambda), H\},$$

де через V(λ) позначена лексикографічна система, як множина словарних статей; Λ(λ) – множина лівих частин словарних статей словника V(λ); $P(\lambda)$ – множина правих частин цього ж словника; $H : \Lambda(\lambda) \rightarrow P(\lambda)$.

Елементарною структурною одиницею тезауруса є словарна стаття дескриптора, яка будується за алфавітно-структурним принципом

$$d_i < M_{i1}, M_{i2}, M_{i3}, M_{i4} >,$$

де d_i – заголовний дескриптор; M_{i1} – упорядкована за абеткою множина умовних синонімів даного заголовного дескриптора, створюючи разом з ним клас умовної еквівалентності; M_{i2} – упорядкована за абеткою множина дескрипторів, кожен з яких пов'язаний із заголовним дескриптором відношенням «рід – вид»; M_{i3} – упорядкована за абеткою множина дескрипторів, кожен з яких пов'язаний із заголовним дескриптором відношенням «вид – рід»; M_{i4} – упорядкована за абеткою множина дескрипторів, кожен з яких пов'язаний із заголовним дескриптором принаймні одним із таких парадигматичних відношень: ціле – частина, частина – ціле, причина – слідство, слідство – причина, функціональна схожість (асоціативні зв'язки). Будь-яка з перерахованих множин може бути одноелементною і навіть порожньою, тобто може бути відсутньою в словарній статті. Множина M_{i1} у сукупності з дескрипторами d_i утворює клас умовної еквівалентності, який також є дескриптором. Ця множина M_{i1} виконує функцію номінального визначення, яке уточнює сенс дескриптора d_i, вибраного для позначення цього класу умовної еквівалентності.

Розглянувши структуру словарної статті словника-тезауруса і її запис на мові розмітки структурованих текстів XML, сформулюємо набір правил для перетворення вхідних даних (словарна стаття) у вихідні дані (словарна стаття в XML запису). На рис. 2 представлений запис елементів словарної статті і відповідний запис на мові XML на прикладі словника-тезауруса.

$$\left.\begin{array}{l} ПР1\,T_0^j = R1, \\ ПР2\,T_1^j = R2, \\ ПР2\,T_2^j = R3, \\ ПР4\,T_3^j = R4, \\ ПР5\,T_4^j = R5 \end{array}\right\} \; j = 1,\overline{N}\,,$$

де T_0 – дескрипторна група; T_1 – родовий дескриптор; T_2 – видовий дескриптор; T_3 – дескриптор; T_4 – умовний синонім; T_5 – асоціативний дескриптор; j – індекс словарної статті словника; $ПР_{\text{п}}$ – програма, що виконує перетворення; $R_{\text{п}}$ – результат перетворення. Досліджено існуючі варіанти розмітки документів. Розмітка документа має на меті: виділення смислових частин (логічних елементів) документа і зв'язків між ними (структурна розмітка); визначення дій, які повинні бути здійснені з цими елементами.

Мова розмітки повинна визначати ряд спеціальних інструкцій, правил і угод для опису структури елементів документа і відносин між елементами цієї структури. Спеціальні інструкції, їх ще називають маркерами або тегами, в структурованих документах повинні певним чином кодуватися, тобто виділятися серед основного тексту. Їх головне призначення – служити інструкціями, що управляють, для програмних засобів обробки структурованих текстів.

До теперішнього часу не існує стандарту на мову розмітки інформації, представлену в словниках. Розробки, що існують в даній області, спираються на рекомендації TEI і CES, які повинні враховуватися розробниками електронних словників. Велика частина існуючих словників мають власні формати зберігання і представлення даних, що не дозволяє інтегрувати їх в інші інформаційні системи. Враховуючи вимоги TEI і CES, що висуваються до розробників лінгвістичних корпусів текстів, запишемо з урахуванням трьох рівнів стандартизації (рівень метамови, синтаксичний рівень, семантичний рівень) вимоги до розмітки даних словника-тезауруса. Кожен подальший рівень містить детальнішу інформацію про розмітку і вимагає стандартизації на попередньому рівні.

На рівні метамови визначаються внутрішні атрибути розмітки, базові набори символів, правила привласнення імен, зарезервовані слова, допустимі відхилення від стандарту (відсутність кінцевого тегу) відповідно до XML-стандарту. Оптимальним є дотримання вимог стандарту XML. Кодування даних задається за бажанням користувача. Доцільно використовувати Unicode/ISO 10646 для запису даних.

На синтаксичному рівні визначаємо точні назви тегів і синтаксичні правила їх уявлення. На цьому рівні відповідність синтаксичному стандарту може бути перевірена за допомогою аналізу тексту, тобто формальним способом. На семантичному рівні необхідно гарантувати однозначність розмітки і її інтерпретаційної частини.

Структура словарної статті дескриптора		*XML - запис*
НАВИГАЦИОННЫЕ СИСТЕМЫ 1402	Дескриптор і номер дескрипторної групи	<descript id="НАВИГАЦИОННЫЕ СИСТЕМЫ" area="1402">
(Системы, обеспечивающие навигацию подвижных боевых средств и средств сообщения)	Пояснення значення дескриптора	<explan>(Системы, обеспечивающие навигацию подвижных боевых средств и средств сообщения)</explan>
ИВ Морские навигационные системы Системы аэронавигации Системы космической навигации	Умовний синонім	<syn id="Морские навигационные системы"/> <syn id="Системы аэронавигации"/> <syn id="Системы космической навигации"/>
РД Навигационные средства	Родовий дескриптор	<parent id="Навигационные средства"/>
ВД Астронавигационные системы Инерциальные навигационные системы Комбинированные навигационные системы Радионавигационные системы Системы ближней навигации Системы дальней навигации	Видові дескриптори	<child id="Астронавигационные системы"/> <child id="Инерциальные навигационные системы"/> <child id="Комбинированные навигационные системы"/> <child id="Радионавигационные системы"/> <child id="Системы ближней навигации"/> <child id="Системы дальней навигации"/>
АД Навигация	Асоціативний дескриптор	<assoc id="Навигация"/> </descript>
Структура запису умовного синоніма		
Приклад запису умовного синоніма, що замінюється комбінацією дескрипторів:		
Системы космической навигации	Умовний синонім	<syn id="Системы космической навигации"> <forgroup> <for id="Космическая навигация"/> <for id="Навигационные системы"/> </forgroup> </syn>
ИСП Космическая навигация **И** Навигационные системы	Дескриптори, які необхідно використовувати замість умовного синоніма	
Приклад запису умовного синоніма, що замінюється одиничним дескриптором:		
Двигатели Ванкеля	Умовний синонім	<syn id="Двигатели Ванкеля"> <for id="Роторно-поршневые двигатели"/> </syn>
ИСП Роторно-поршневые двигатели	Дескриптор	
Структура запису дескрипторної групи		
0905 Программирование для вычислительных машин		<area id="0905" name="Программирование для вычислительных машин"/>

Рис. 2. Структура словарних статей словника-тезауруса і їх XML – подання

Для зняття неоднозначностей при перенесенні текстової інформації між різними системами необхідно вказати для користувача належність використовуваних тегів частинам документа. Такі семантичні правила в основному визначаються в супроводжуючих довідниках користувача.

У **третьому** розділі розглядається апарат скінченних автоматів для розв'язання лексикографічних завдань комп'ютерного аналізу тексту, зокрема розпізнавання та ідентифікації (лексичний аналіз). Побудова лексичного аналізатора значно спрощується завдяки тому, що словник має певну структуру, яка може бути представлена у вигляді правил обробки.

Робота лексичного аналізатора описується формалізмом скінченних автоматів. У даній роботі апарат скінченних автоматів використовується для опису процесу обробки природномовних текстів, зокрема, словників-тезаурусів.

Скінченні автомати моделюють поведінку, при якій реакції на майбутні події залежать від попередніх подій. Автомати найбільш корисні в ситуаціях, коли поведінка керується багатьма різними типами подій, а реакція на певну подію залежить від послідовності попередніх подій.

Побудова скінченного автомата за регулярним виразом відбувається в такій послідовності: побудова системи переходів за регулярним виразом; побудова діаграми станів за системою переходів; побудова недетермінованого скінченного автомата (НСА) за діаграмою станів; побудова скінченного автомата (СА) за НСА.

Текст на природній мові можна представити як цілісний об'єкт, елементами якого є знаки, організовані певним чином у рядки: ТЕКСТ = {знак}, {рядок}. Словник є одним із різновидів текстів на природній мові, отже, можна записати: Словарна стаття = {знак}, {рядок}.

Знаки і рядки розглядаються як граничні одиниці моделі. При цьому множина елементів тексту впорядкована таким чином, що інформація, яка передається текстом словарної статті, кодується графічними засобами оформлення. Щоб набути графемного значення тексту, необхідно змоделювати правила кодування одиниць графемного відображення. Для цього треба виділити закономірності взаємодії одиниць графемного відображення природної мови в аспекті їх знакової природи і описати правила виділення смислових одиниць тексту і відношень, що існують між ними.

Під знаком розуміють задану апріорі величину, що є символом (будь-якого роду) або відсутністю символу. Стаття тезауруса задається алфавітом будь-якої природної мови, пропусками, синтаксичними або логічними роздільниками. Під рядком розуміють певну послідовність знаків, а під порожнім рядком розуміють відсутність знаків.

У процесі формування тексту знаки вступають у синтагматичні відношення, утворюючи послідовність символів, які можна інтерпретувати, приписавши їм певне графемне значення.

Графемне значення тексту включає такі компоненти:

1. Основні одиниці тексту, такі як фрагмент, речення і лексема;

2. Типи названих вище класів елементів, тобто співвідношення елементів з певним класом, який відображає однакові властивості виділених елементів;

3. Зв'язки між цими елементами, тобто зазначення їх контекстних (синтагматичних) відношень.

Для опису алгоритму витягання графемного значення тексту треба записати правила, що дозволяють витягувати семантичну інформацію в тексті словарної статті за організації її знакової системи, виходячи з оформлення тексту. Дані правила описують усі термінальні вершини наведеної вище ієрархічної схеми.

У процесі графемного аналізу словника-тезауруса кожній виділеній графемній одиниці тексту приписується код. Використання графемного аналізу на першому етапі аналітико-синтаксичної обробки тексту дозволяє вирішити такі завдання: виділити структурно оформлені фрагменти тексту, що мають певне значення; виділити графічні елементи оформлення; виділити лексико-граматичні класи, що мають різне графемне оформлення. На етапі лексичного аналізу виділяються лексеми, що надходять на вхід синтаксичного аналізатора, який проводить остаточну обробку потоку текстових даних.

Зробивши аналіз словарних статей ряду словників-тезаурусів, можна уніфікувати правила обробки елементів словарної статті словників-тезаурусів. Для цього необхідно провести перетворення вхідного тексту до верхнього регістра перед початком лексичного аналізу, що дозволить спростити правила обробки текстових даних. З урахуванням цього запишемо набір правил обробки словарної статті. У деяких випадках різні за сенсом рядки словарної статті описуються однаково. Отже, лексичний аналіз таких рядків буде утруднений. Для виключення неоднозначностей в процесі аналізу даних, необхідно ввести додаткові змінні «прапори станів», які допоможуть однозначно інтерпретувати розпізнавані дані. Введемо такі прапори станів (ІВ = 0, РД = 0, ВД = 0, АД = 0), які за умовчанням рівні «0». Прапор стану змінюється на «1», якщо, розпізнаний аналізатором, рядок є відповідно умовним синонімом, родовим, видовим або асоціативним дескриптором. Для однозначного розпізнавання рядків необхідно перевірити значення «Прапор стану» і залежно від їх значення ухвалити рішення про тип розпізнаваного рядка. При цьому правила матимуть наступний вигляд:

ПРАВИЛО 0: ІВ=0, РД=0, ВД=0, АД=0
ПРАВИЛО 1 (*Дескриптор Номер дескрипторної групи*)
IF {букви}, {пробіл}, {цифри} **THEN** Код_1
ПРАВИЛО 2 (*Пояснення значення дескриптора*)
IF [{Відкрита дужка}, {букви}, {закрита дужка}] **THEN** Код_1.1
ПРАВИЛО 3 (*Умовні синоніми й заміна умовного синоніму комбінацією дескрипторів (одиничним дескриптором*))
IF [ІВ, {пробіл}], {знак плюс}, {пробіл}, {букви} **OR** {букви} **AND** (ІВ=1, РД=0, ВД=0, АД=0) **THEN** Код_1.2, ІВ=1, РД=0, ВД=0, АД=0
ПРАВИЛО 4 (*Родовий дескриптор*)
IF РД, {пробіл}, {букви} **OR** {букви} **AND** (ІВ=0, РД=1, ВД=0, АД=0) **THEN** Код_1.3, ІВ=0, РД=1, ВД=0, АД=0

ПРАВИЛО 5 (*Видові дескриптори*)
IF [ВД, {пробіл}, {знак мінус}, {пробіл}], {букви} **OR** {букви} **AND** (ІВ=0, РД=0, ВД=1, АД=0) **THEN** Код_1.4, ІВ=0, РД=0, ВД=1, АД=0
ПРАВИЛО 6 (*Асоціативний дескриптор*)
IF [АД, {пробіл},] {букви} **OR** {букви} **AND** (ІВ=0, РД=0, ВД=0, АД=1) **THEN** Код_1.5, ІВ=0, РД=0, ВД=0, АД=1
ПРАВИЛО 7
IF ІСП, {пробіл}, {букви} **THEN** Код_1.2.1
ПРАВИЛО 8
IF І, {пробіл}, {букви} **THEN** Код_1.2.2
ПРАВИЛО 9 (Символ кінця абзацу)
{символ кінця абзацу} **THEN** Код_1.6

Складемо граф станів і переходів для обробки словарної статті словника-тезауруса. Детальну інформацію про роботу програми подано у розділі 4. Функції обробки словарної статті словника-тезауруса представлені на сумісному графі переходів у вигляді словесного опису виконуваних дій та графі з повною формою логічних умовних подій модуля розмітки, реалізованих в лексикографічному процесорі (рис. 3), відповідності код – дія наведені в табл. 1., та таблиця переходів (табл. 2).

Таблиця 1

Код – дія для графа рис. 3

№	Дія
0	Нічого не робить
1	Читання заголовка першої словарної статті
2	Збереження коментаря
3	Читання заголовка словарної статті
4	Виділення помітки
5	Збереження помітки
6	Читання чергового рядка
7	Запис збереженого асоціативного дескриптора
8	Запис заголовка асоціативного дескриптора
9	Запис заголовка
10	Запис чергового рядка
11	Запис збереженого коментаря
12	Запис збереженої помітки
13	Запис закінчення словарної статті

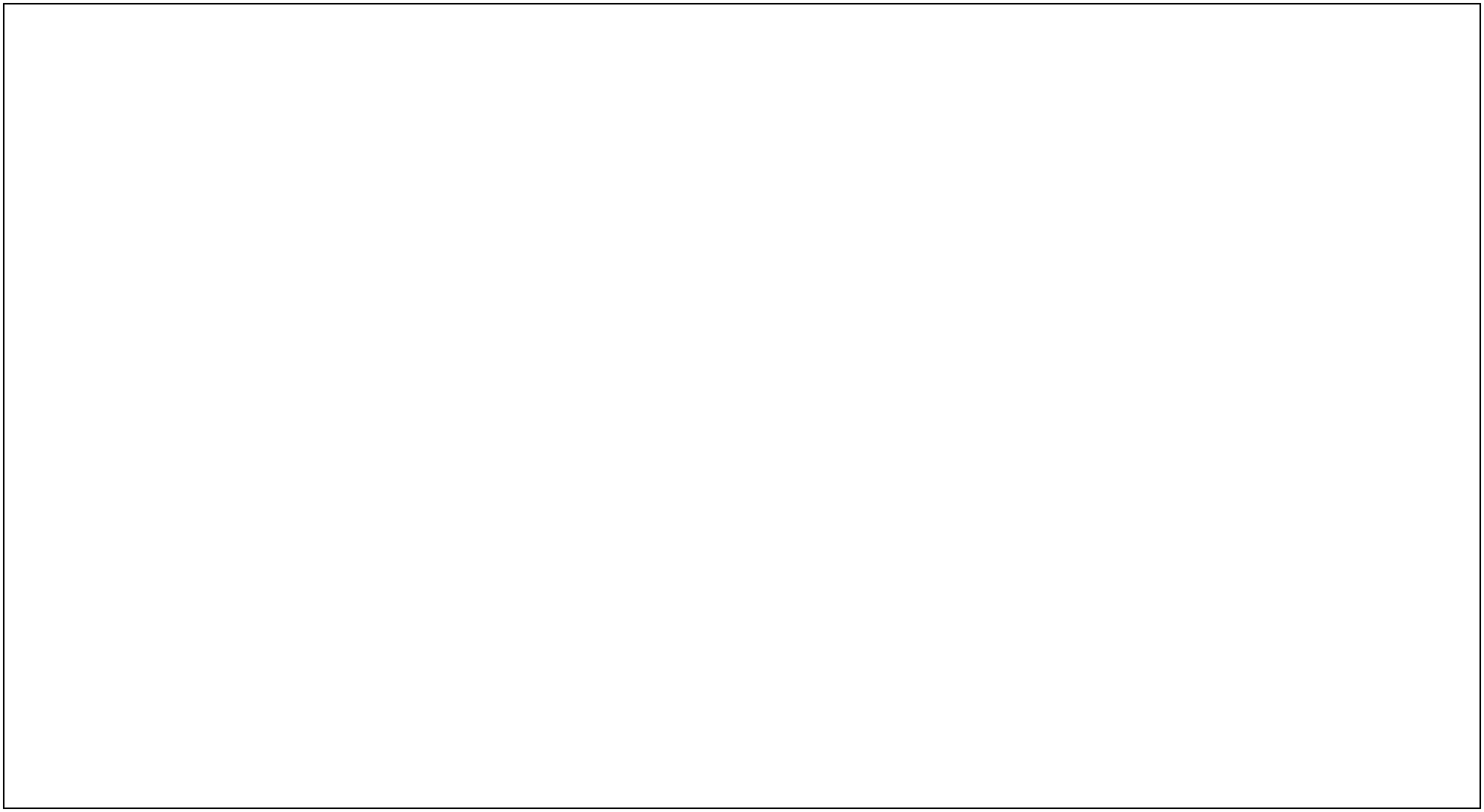

Рис. 3. Сумісний граф переходів лексикографічного процесора та
граф з повною формою логічних умовних подій модуля розмітки

Таблиця 2

Таблиця переходів для графа на рис. 3

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1/1	1/1	1/1	1/1	1/1	1/1	1/1	1/1
1	2/2	3/0	3/0	3/0	3/0	3/0	3/0	3/0
2	3/3	3/3	3/3	3/3	3/3	3/3	3/3	3/3
3	4/4	4/4	4/4	4/4	4/4	4/4	4/4	4/4
4	5/6	5/5	5/6	5/6	5/6	5/6	5/6	5/6
5	6/9	6/9	6/7	6/8	6/9	6/9	6/9	6/9
6	7/10	7/10	7/10	7/10	7/10	7/10	7/10	7/10
7	8/0	8/0	8/0	8/0	8/11	8/0	8/0	8/0
8	11/9	11/9	11/9	11/9	11/9	11/13	11/9	11/9
9	10/0	10/0	10/0	10/0	10/0	10/0	10/12	10/0
10	11/10	11/10	11/10	11/10	11/10	11/10	11/10	11/10
11	1/3	1/3	1/3	1/3	1/3	1/3	1/3	12/0
12	12/0	12/0	12/0	12/0	12/0	12/0	12/0	12/0

У **четвертому** розділі наводиться показник ефективності розробки модуля розмітки лексикографічного процесора. Критерій оцінки якості перетворення текстової інформації вимірюється індексом якості обробки Total Quality Index (TQI). За прийнятими стандартами, у документах для друку (відображення в Web), до яких належить дана розробка, цей індекс не може бути меншим за 99,9%. Кількість помилок не повинна бути більш за одну на тисячу слів перетворення. Проте при обробці текстової документації редактором цей результат відрізняється від необхідного.

Для досягнення необхідної точності ε і необхідної надійності P отриманих результатів експерименту було визначено необхідну кількість вимірювань. У роботі показано, що для досягнення надійності довірчої оцінки 0,995, у разі невідомої величини σ, слід провести 30 вимірювань.

У експерименті було проведено вимірювання показників кількості помилок і час, що витрачається оператором при розмітці документа. Також фіксувався час розпізнавання рядка і загальний час, витрачений на розпізнавання і розмітку всієї запропонованої послідовності рядків в експерименті.

Тестова послідовність становила 34 рядки, які пред’являлися операторові для визначення і встановлення розмітки в поточному аналізованому рядку. При цьому фіксувалися час і кількість помилок, що допускаються.

На рис. 4 наведені значення максимального, мінімального і середнього значення часу, що витрачається оператором на розпізнавання чергового рядка тестової послідовності. При цьому мінімальний показник часу, витраченого на розпізнавання рядка, становить 2,5 с., максимальне значення 20,5 с., середнє значення розпізнавання рядка 8,93 с.

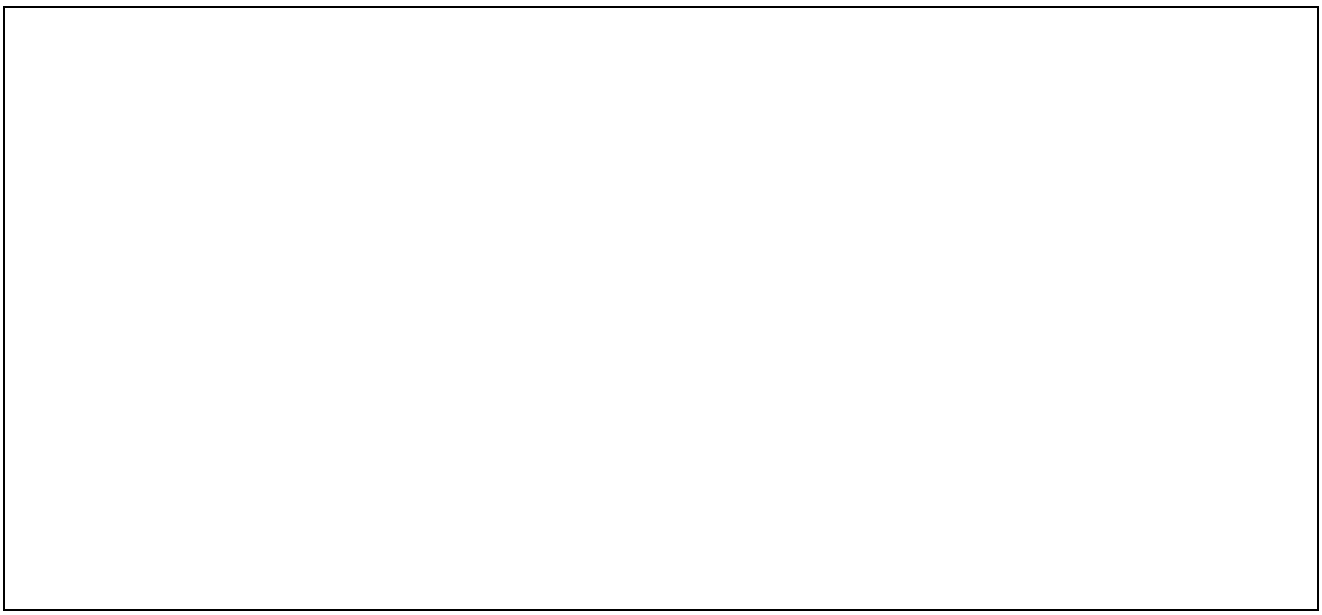

Рис. 4. Час розпізнавання рядка в експерименті

На рис. 5 приведені значення часу розпізнавання тестової послідовності (34 рядки) в експерименті. При цьому тестова послідовність була повністю розпізнана (розмічена) за максимально 455,50 с., мінімально – за 219,00 с., а середнє значення при цьому становило 303,52 с.

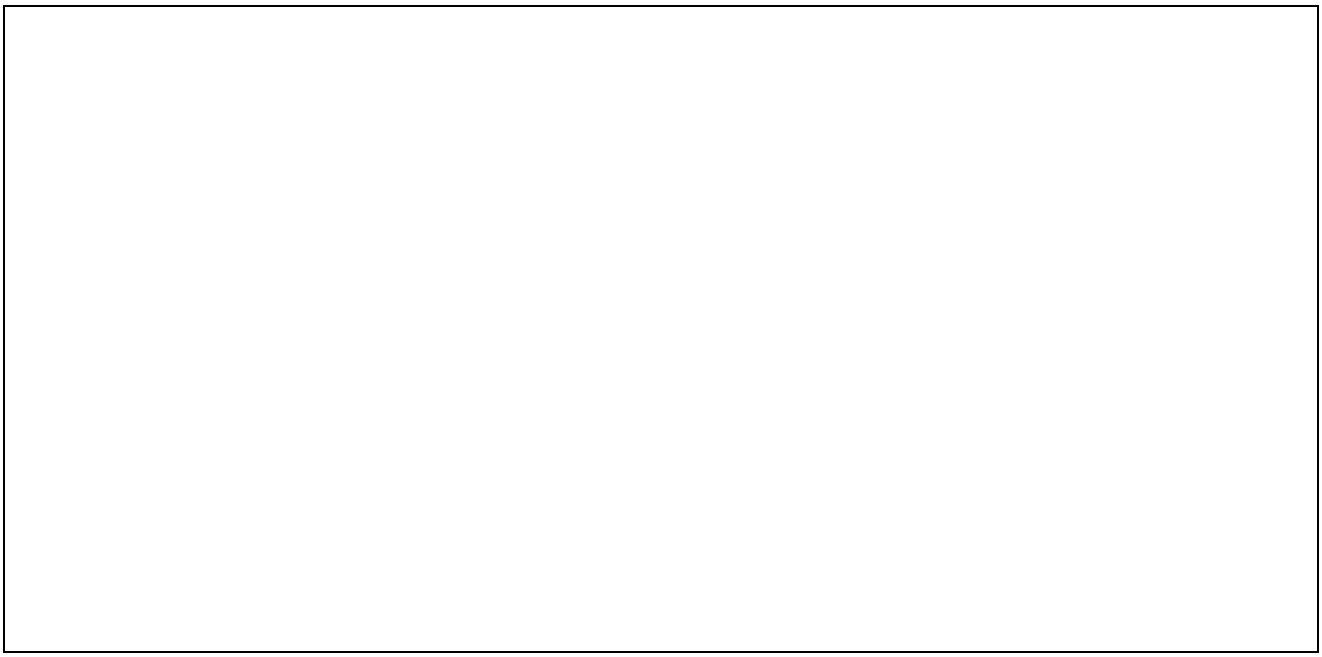

Рис. 5. Час розпізнавання тестової послідовності в експерименті

Таким чином, у результаті проведеного експерименту встановлено, що реальний середній час на розпізнавання рядка становив 8,93 с., середній час розпізнавання тестової послідовності з 34 рядків 303,52 с.; кількість помилок на 1000 рядків – 6. При цьому показник довірчої оцінки отриманих результатів 0,995%.

Для визначення ефективності розробленого програмного забезпечення необхідно порівняти час розпізнавання і кількість помилок, отриманих у результаті експериментів, які проводилися в ручному режимі з аналогічними показниками, отриманими при автоматизованій обробці тексту.

Для оцінки ефективності розробленого програмного забезпечення необхідно порівняння часу і кількості помилок при ручному розпізнаванні тексту, і в результаті проведеного експерименту (автоматична обробка тексту).

За результатами автоматичної обробки, кількість рядків у файлі для розпізнавання – 1000; час виконання програми – 0,0017 с.; помилок розпізнавання (не більше) – 1. При автоматизованій обробці кількість помилок зменшилася в 6 разів, час обробки зменшився в $5 \cdot 10^6$ разів.

У цьому ж розділі розглянуто розроблену програму «Модуль XML розмітки лексикографічного процесора», її побудову, основні модулі та їх призначення, функціонування та структуру файлу настройки. Програма здійснює синтаксичний аналіз текстової інформації, представленої у процедурній розмітці, і перетворює її в описову розмітку на розширеній мові розмітки XML, або довільну розмітку, що задається користувачем. Застосовується для структуризації даних, що містяться в словниках.

За результатами роботи програмного комплексу здійснено впровадження в бібліотеці Інституту проблем машинобудування НАН України для створення та роботи з базою даних технічних статей і довідників. У додатку В наведено документи щодо практичного застосування одержаних результатів, а також впровадження в навчальному процесі на кафедрі «Системи інформації» НТУ «ХПІ» при вивченні курсу «Природномовні інтелектуальні системи».

ВИСНОВКИ

У дисертаційній роботі вирішено науково-практичне завдання створення і використання інформаційної технології для автоматизованої переробки інформації у процесі створення електронних словників-тезаурусів. Основні результати роботи полягають в наступному:

1. Проведено аналіз завдань обробки текстової інформації, що використовують для свого вирішення електронні словники, а також засобів автоматизації створення електронних словників. Вирішення задач розробки інформаційних технологій автоматизованого створення електронних словників дозволяє максимально ефективно інтегрувати електронні словники в спеціалізовані інформаційні системи (пошуку, класифікації, інформаційного стиску текстів та ін.).

2. Досліджено структуру представлення даних у паперових словниках і розроблені засоби перетворення текстів у спеціальну електронну форму. Показано, що структура словарної статті будь-якого словника складається з двох частин: лівої реєстрової (дескрипторної) частини та правої, що містить заголовний ряд і позначає парадигматичні відношення дескриптора з іншими елементами словникової статті. Побудовано концептуальну математичну модель словника-тезауруса.

3. Розроблено вимоги до мови розмітки даних на прикладі словника-тезауруса і вироблено рекомендації користувачеві з настройки системи. Спираючись на рекомендації TEI і CES, розвинуто можливості XML з метою застосування її для розмітки структурованих текстів словників-тезаурусів. Визначені внутрішні атрибути розмітки, базові набори символів, правила привласнення імен, зарезервовані слова. Визначені назви тегів і синтаксичні правила їх представлення, що дозволило забезпечити гарантування однозначності розмітки та її інтерпретаційної частини. Вказано належність використовуваних тегів певним частинам документа, що дозволило зняти неоднозначність при обміні текстовою інформацією між різними системами.

4. Досліджено структуру лексикографічного процесора (ЛП) і запропоновано модифікацію його модулів розмітки. Розроблено інформаційну технологію та програмне забезпечення ЛП, що дозволило провести автоматичне форматування службової інформації до фрагментів словникової статті. Система дозволяє виконувати автоматичну перевірку структури одно типових словникових статей та забезпечує створення індексних файлів словників.

5. Розроблено інформаційну технологію перетворення структурованих текстів природної мови в спеціальну електронну форму з використанням вдосконаленої мови розмітки. Запропонований набір тегів розмітки для словників-тезаурусів сприяє виділенню смислових елементів та їх зв'язків в основному та зв'язаних документах, що дозволяє використати їх як інструкції, які управляють програмними засобами обробки структурованих текстів.

6. Проведено перевірку працездатності й ефективності розробленої інформаційної технології на прикладі створення електронного словника-тезауруса. Розроблене програмне забезпечення дозволило зменшити кількість помилок на тисячу слів перетворення в 6 разів, час обробки при цьому зменшився в $5 \cdot 10^6$ разів.

7. Розроблені модулі лексикографічного процесора впроваджені в Інституті проблем машинобудування НАН України для створення бази даних технічних статей і довідників, в навчальному процесі на кафедрі «Системи інформації» НТУ «ХПІ» при вивченні курсу «Природномовні інтелектуальні системи», розроблена комп'ютерна програма «Модель XML-разметки лексикографического процессора обработки текстовой информации», що підтверджується відповідними актами про впровадження та реєстрацію.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Касилов О. В. Методы представления структурированных текстов естественного языка в XML-описании / О. В. Касилов // Вісник Національного технічного університету «Харківський політехнічний інститут». – Харків: НТУ «ХПІ». – 2002. – № 6. – Т. 2. – С. 3–8.

2. Касилов О. В. Основы разметки текстов / О. В. Касилов, А. Н. Самойлов, А. С. Шраер // Вісник Національного технічного університету

«Харківський політехнічний інститут». – Харків: НТУ «ХПІ». – 2002. – № 9. – Т. 7. – С. 191–195.

Здобувач виконав огляд існуючих методів розмітки структурованих текстів, сформулював вимоги до мови розмітки, яка застосована в лінгвістичному процесорі.

3. Касилов О. В. Моделирование лексикографических систем / О. В. Касилов // Вісник Міжнародного Слов'янського університету. Сер. «Технічні науки». – Харків.:– 2004. – Т. 7. – № 1. – С. 13–15.

4. Касилов О. В. Моделирование словаря-тезауруса. / О. В. Касилов // Вісник Національного технічного університету «Харківський політехнічний інститут». – Харків: НТУ «ХПІ». – 2004. – № 34. – С. 88–93.

5. Касілов О. В. Компьютерная программа «Модуль XML-разметки лексикографического процессора обработки текстовой информации». / О. В. Касілов, І. А. Конопльов // Свідоцтво про реєстрацію № 22259. ДДІВ від 05.10.2007.

Здобувач виконав теоретичне обґрунтування розробки концепції і створення програми, розробку модулів програми, тестування і перевірку працездатності програми.

АНОТАЦІЇ

Касілов О. В. Інформаційна технологія автоматизованої переробки текстової інформації при створенні електронних словників-тезаурусів. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – Інформаційні технології. – Національний технічний університет «Харківський політехнічний інститут», Харків, 2008.

У дисертаційній роботі розглянуто системи обробки текстової інформації, системи електронних словників, їх структуру та можливості застосування в різноманітних інформаційних системах. Розглянуті можливості сучасних засобів автоматизації лінгвістичних задач. Обґрунтовано застосування модифікованих скінченних автоматів для задач розпізнавання і класифікації. Розроблена інформаційна технологія, що реалізує автоматизоване перетворення структурованих текстів природної мови до їх електронної форми, здійснено практичну перевірку її працездатності на прикладі словників-тезаурусів. Шляхом введення додаткових елементів у розмітку вдосконалена модифікація мови відкритої розмітки тексту XML для представлення словників-тезаурусів в електронній формі. Розроблено лексикографічний процесор для перетворення словників-тезаурусів до електронної форми. Одержала подальший розвиток методика перетворення XML-опису словника в базу даних різних форматів.

Ключові слова: лексикографічний процесор, модель словника-тезауруса, скінченні автомати, мови розмітки, природномовні тексти, елктронний-словник, XML.

Касилов О. В. Информационная технология автоматизированной обработки текстовой информации при создании электронных словарей-тезаурусов. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 – Информационные технологии. – Национальный технический университет «Харьковский политехнический институт», Харьков, 2008.

В работе решена важная научно-практическая задача построения информационной технологии для автоматизированной обработки структурированных текстов естественного языка с целью построения электронных словарей-тезаурусов.

Анализ современного состояния и тенденций в области информатизации лингвистических исследований, проведенный в работе, в частности, обзор лексикографических работ в Украине, позволил выделить наиболее важные и актуальные задачи в этой области. В работе показана необходимость разработки эффективных средств автоматизации исследований, связанных с применением различного вида словарных структур, лежащих в основе большинства интеллектуальных информационных систем.

В диссертационной работе рассмотрены системы обработки текстовой информации, системы электронных словарей, их структура и возможности применения в разнообразных информационных системах. Рассмотрены возможности современных средств автоматизации решения лингвистических задач, в состав которых входят различные словари. Обосновано применение модифицированных конечных автоматов для задач распознавания и классификации при построении словарей. Разработана информационная технология, реализующая автоматизированное преобразование структурированных текстов естественного языка в электронную форму, осуществлена практическая проверка работоспособности модуля лексикографического процессора на примере словарей-тезаурусов. Путем введения дополнительных элементов в разметку усовершенствована модификация языка открытой разметки текста XML для представления словарей-тезаурусов в электронной форме. Разработан лексикографический процессор для преобразования словарей-тезаурусов в электронную форму. Получила дальнейшее развитие методика преобразования XML-описания словаря в базы данных разных форматов.

Результаты диссертационной работы внедрены в Институте проблем машиностроения НАН Украины при создании базы данных технических статей и справочников, а также в учебном процессе на кафедре «Системы информации» НТУ «ХПИ» при изучении курса «Естественноязыковые интеллектуальные системы».

Ключевые слова: лексикографический процессор, модель словаря-тезауруса, конечные автоматы, языки разметки, естественноязыковые тексты, электронный - словарь, XML.

Kasilov O. V. Information technologies of text processing at creation of electronic thesauruses. – Manuscript.

Thesis for a candidate's degree by speciality 05.13.06 – Information technologies. – National technical university «Kharkov polytechnic institute», Kharkov, 2008.

Thesis for a candidate's degree is devoted to the systems of treatment of text information, electronic thesauruses, their structure and possibilities of application in the various information systems. In this work are considered the capabilities of modern tools of automatization of linguistic tasks. Application of the modified finite-state automatons for the tasks of recognition and classification is grounded.

There is developed the information technology that implements the automated transformation of structured texts of natural language in electronic form, also there is carried out the practical verification of operability of the module of lexicographic processor on the example of thesaurus. By introduction of additional elements to markup, modification of language of opened markup of XML text for presentation of thesaurus in electronic form is improved. A lexicographic processor for transformation of thesaurus in electronic form is developed. The method of conversion of XML description of dictionary in the data-bases that have different formats have got further development.

Keywords: lexicographic processor, model of thesaurus, finite- state automaton, languages of markup, electronic dictionary, XML.

Підписано до друку 06.11.2008 р. Формат 60х90/16.
Папір офсетний. Друк - ризографія. Гарнітура Times New Roman.
Умовн. друк. арк. 0,9. Наклад 100 прим. Зам. № 262744

Надруковано у СПДФО Ізрайлєв Є.М.
Свідоцтво № 24800170000040432 від 21.03.2001р.
61002, м. Харків, вул. Фрунзе, 16