ДИБІНА А. В., ЛАЗАРЕНКО О. В.
Харківський гуманітарний університет "Народна українська академія"

ПОБУДОВА ТЕКСТОВОЇ БАЗИ В СИСТЕМІ АВТОМАТИЧНОГО РЕФЕРУВАННЯ НА ОСНОВІ СТРУКТУРНО-СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТУ

У роботі розглянуто процедуру автоматичної побудови текстової бази для заповнення моделі реферату в системі автоматичного реферування.
Ключові слова: текстова база, когезія, когерентність, автоматичне реферування.

В работе рассмотрена процедура автоматического построения текстовой базы для заполнения модели реферата в системе автоматического реферирования.
Ключевые слова: текстовая база, когезия, когерентность, автоматическое реферирование.

The article deals with the automatic construction of a text base to fill the model of the summary in the system of automatic summarization.
Key words*:* text base, cohesion, coherence, automatic summarization.

Мета. Подолання деяких недоліків сучасних систем автоматичного реферування (АР) за рахунок розробки методу структурно-семантичного аналізу тексту для побудови текстової бази в системі автоматичного реферування.

Актуальність наших досліджень визначається завданням побудови когнітивної або семантико-контекстної моделі, яка б забезпечила глибинне проникнення у смисл тексту та його трансформацію зі збереженням смислу.

Виконання автоматичної обробки тексту природною мовою на семантичному рівні є найважливішою складовою інтелектуальних систем лінгвістичного аналізу природномовних об'єктів. Інтелектуальні системи реферування, що забезпечують моделювання процесу розуміння та узагальнення інформації на семантичному рівні, залишаються найбільш гострою проблемою сьогодення.

Об'єктом дослідження є смислова організація наукових текстів для розробці процедури автоматичної побудови текстової бази з використанням моделі реферату та заголовку.

Існують різні підходи до аналізу смислу тексту.

В рамках нашого підходу до вирішення задачі розуміння тексту в системі автоматичного реферування ми спираємося на прагмастилістичний підхід, який передбачає встановлення того, які мовні

засоби, що зустрічаються в аналізованому тексті, найтісніше пов'язані з його ідеологією, а також виявлення "мета змісту", що міститься в тексті і створює певне організоване ціле. Щоб зрозуміти текст, недостатньо володіти набором еталонних лексичних одиниць і синтаксичних структур, потрібно мати в свідомості абстрактні інваріантні "суперструктури", які, реалізуючись в тексті, забезпечують зв'язок між фразами, абзацами, главами і т. п. [1:15–17]

Виділяють смислову та структурну зв'язність тексту. Смислова зв'язність (когерентність) визначається взаємозалежністю окремих речень, що забезпечує цілісність тексту. А структурна зв'язність (когезія) є внутрішньою організацією тексту за допомогою формально-граматичних аспектів зв'язку. Когезія не виявляє, що повідомляє текст, вона виявляє, як текст організований в семантичне ціле. Структурна і смислова зв'язність вивчається в рамках композиційно-тематичного підходу.

У роботі існуючих систем реферування найчастіше використовується метод формально-смислового аналізу текстів первинних документів, що є один з широко поширених підходів до складання рефератів [2]. Особливість цього підходу полягає в тому, що він враховує формальні текстові ознаки, так звані маркери – стійкі словесні звороти, що характеризують конкретні аспекти смислу.

Всі аспекти об'єднуються в три семантичні блоки:

1. Вступна частина (постановка проблеми).
2. Основний зміст (пропонований варіант вирішення проблеми).
3. Висновок (отримані результати, висновки).

Вступна частина найчастіше складається з таких аспектів, як постановка проблеми, відомий варіант вирішення, оцінка відомого варіанту вирішення. Ці аспекти знаходяться в першій частині тексту і готують читача до основної інформації. В основний смисл тексту входять пропонований варіант вирішення проблеми, її особливості і оцінка пропонованого варіанту вирішення. У висновку освітлюються результати, висновки, рекомендації і сфера застосування.

Кожному аспекту відповідає деякий фрагмент тексту, що характеризується смисловою закінченістю, зв'язністю і місцем в структурі тексту.

Реалізація даного підходу заснована на методі екстрагування. На думку авторів, що розділяють цей підхід, для побудови автоматичного реферату можна скласти список (словник) словесних кліше (маркерів,

індикаторів, коннекторів), що відображають позатематичну (метаінформаційну) лексику. За допомогою такого словника можна впізнати (маркувати) і вибирати з тексту первинного документу окремі речення, сукупності яких і утворюють реферати, своїми характеристиками істотно не поступаються традиційним рефератам, складеним фахівцями-референтами [3:35].

Проте практика показала, що одержувані при такому підході квазіреферати істотно поступаються традиційним (інтелектуальним) рефератам. По суті, такі екстракти правильніше було б називати текстовою базою, на матеріалі якої можна будувати реферати. І для цього необхідно, на наш погляд, мати формальний опис структури реферату, тобто його модель, і алгоритм заповнення цієї моделі необхідною інформацією зтекстової бази відповідно до розроблених правил семантичного конструювання реферату.

Ще одним слабким місцем даного підходу є те, що автоматизоване екстрагування припускає після маркування першоджерела і подальшого редагування, що полягає в зменшенні надмірного обсягу екстрактів за рахунок виключення менш істотних пропозицій, побудова власне реферату-екстракту без яких-небудь заданих критеріїв відбору цих пропозицій. У пропонованому нами підході такі критерії спочатку закладаються існуючою моделлю реферату.

Наш підхід передбачаєподолання деяких недоліків сучасних систем автоматичного реферування (АР) за рахунок розробки методу структурно-семантичного аналізу тексту для побудови текстової бази в системі автоматичного реферування з використанням моделі реферату та заголовку.

В основі нашого підходу лежить гіпотеза, згідно з якою моделювання процесу реферування, який є сукупністю найскладніших процесів розуміння і компресії смислу, слід починати з вивчення не самих процесів, а з їх результату – реферату. Причому не розгорнутого, інформативного, а стислого, індикативного, перш за все, тому що розглядаємо його як відправну точку в дослідженні цього питання, як об’єкт найбільш простий за формою, але який відбиває усі особливості реферативного тексту.

Цілком усвідомлюючи всю складність задачі, основний акцент зроблено на спрощенні процедури реферування на початковому етапі досліджень за рахунок:

1) виключення побудови смислової структури тексту шляхом побудовою тематичної структури через індексацію ключовими словами, які досить повно розкривають тематику тексту;

2) заміни процедури укладання тексту реферату моделлю реферату з подальшим словниковим наповненням синтаксичних конструкцій, що входять до цієї моделі.

Дослідження смислової і синтаксичної структур реферату дозволили з'ясувати особливості у структурі реферативних речень, і на підставі виявлених особливостей семантико-синтаксичної структури цих речень була побудована модель індикативного реферату, а також заголовку, що дозволило нам підійти до питань автоматичної побудови текстової бази з використанням цих моделей.

Аналіз великого корпусу індикативних рефератів показав, що стиснення на смисловому рівні відбувається як в рамках окремого речення, так і всієї смислової структури реферату. Виявилось, що індикативний реферат, як правило, складається з двох речень із значеннями – об'єкт *дослідження* і результат *дослідження, які можуть містити додаткові смислові аспекти – мета і метод.*

Побудована нами інтенсіональна модель реферату, що описує його концептуальне значення, дозволила чіткіше визначити ті засоби представлення знань в системі автоматичного реферування, які необхідні для наповнення моделі реферату екстенсіональною семантикою:

- *текстова база*, яка є семантичним відображенням тексту;
- *онтології*: метаонтология для опису категорій реферативних конструкцій (категорій, за допомогою яких природна мова здійснює концептуальну структуризацію уявлень про дійсність) і онтології предметних областей, що містять поняття певної області знань або пов'язаних з нею областей, і що складаються з об'єктів і зв'язків між ними, описаних в термінології конкретної предметної області [4:91–98].

Текстова база повинна містити інформацію, виражену самим текстом, тобто референційне значення тексту, оскільки на відміну від художньої літератури "для такого жанрового різновиду тексту, як наукова і технічна література, характерна переважаюча роль референційних значень (тобто найбільш істотна інформація, що міститься в такому типі текстах, міститься саме в референційних значеннях, що входять в текст мовних одиниць)" [5].

Оскільки “процес розуміння припускає часткове планування (або очікування)” (у нашому випадку, очікування) “структур і значень пропозицій і цілих текстів” [5], саме із створення таких структур і слід починати розробку текстових баз.

Аналіз смислу тексту ми почали з аналізу його заголовку, який розглядаємояк реферат мінімального об’єму або як текст з максимальним рівнем узагальнення (стиснення) смислу. В результаті дослідження смислової і синтаксичної структури заголовку було виявлено його схожість зі структурою реферату. Як і в індикативному рефераті, смислова структура заголовку в загальному випадку складається з двох метазначень – *об’єкт* і *результат*. Така схожість смислових структур реферату та заголовку слугувала підставою для вивчення взаємозв’язку текстів і їх заголовків для того, щоб за допомогою інформації, яка міститься в заголовку, виявити в тексті ті лексичні одиниці, які необхідні для семантичного наповнення моделі реферату екстенсіональною семантикою даного тексту.

Для здійснення цієї процедури необхідно визначити ті мовні одиниці, що входять в текст, в яких і містяться референційні значення тексту. Що ж це за одиниці? Визначаючись з відповіддю на це питання, ми дотримувалися думки про вибірковий підхід до знань, необхідний для розуміння тексту. “Замість більш менш суцільної активації всього наявного знання, потрібного для розуміння слова, речення або конструкції з глобальною темою, має місце стратегічне використання знання, яке залежить від цілей користувача мови, об’єму знання, наявного в тексті і контексті, рівня переробки або ступеня зв’язності, що необхідного для розуміння і є критеріями стратегічного використання знання” [5:153–211].

Оскільки ми вже побудували модель реферату і провели класифікацію лексем, що беруть участь в заповненні реферативних конструкцій, а також побудували класифікацію лексем заголовка [6], ми підійшли до завдання побудови текстової бази, обмеживши на даному етапі її зміст виділенням невеликої кількості смислових аспектів тексту – *об’єкту, результату, мети і методу.* З цією метою в тексті були виділені слова-покажчики на ці аспекти. Як приклад розглянемо деякі з них.

Об’єкт: *розглянемо, розглядаються; (будемо, можна, доцільно) розглядати; у статті аналізуються, вивчаються, використовуються, реалізуються, описуються; стаття присвячена і ін.*

Об'єкт + результат: *мета (статті), (основною, нашою) метою, метою даної роботи (є, було); зроблена спроба, представляє спробу та ін.;*

Результат: *у висновку слід зазначити, відзначимо, що; (був) описаний, використаний, запропонований, одержаний; одержані в результаті ... ; аналіз, результати, експерименти показали (свідчать) та ін.*

За допомогою цих покажчиків в текстах були виділені речення, які увійшли до текстової бази відповідного тексту як головні (але не всі) смислові аспекти.

Стаття 1. *Математическое моделирование семантических закономерностей процесса терминологизации.*

Целью данной работы является построение математической модели процесса терминологизации (формальное описание его семантических закономерностей) на материале лексических единиц цветообозначения общелитературного английского языка.

Анализ признаков процесса терминологизации показал наличие определенной зависимости одних признаков от других, что отражает дерево структурно-семантических зависимостей.

Стаття 2. *Исследование свойств предиката дифункциональности.*

В статье изучаются некоторые свойства предиката дифункциональности.

Построенные черты предиката свидетельствуют об оптимальности найденных нами условий.

Стаття 3. *Метод формирования грамматических категорий по смысловым оттенкам морфем.*

В этой статье описывается метод формирования грамматических категорий (в том числе словообразовательных).

В результате получаем разбиение множества всех оставшихся оттенков на классы эквивалентности.

Стаття 4. *О математическом описании именного приставочного словообразования.*

Основной целью математического описания языка является описание всех языковых процессов в виде формализованных правил, по которым ЭВМ могла бы произвести обработку текстовой информации.

Данная статья посвящена проблемам описания процессов словообразования путем префиксации.

В данной статье математически описывается формирование временных и пространственных приставок в именах существительных и прилагательных.

Решая полученные уравнения, мы можем сформировать необходимую приставку из числа пространственных и временных именных префиксов.

Стаття 5. *О математическом описании смысла текста.*

Эта статья продолжает исследования, начатые в работе 1, в которой развит общий подход к математическому описанию смысла текста и отношения между текстом и смыслом.

Стаття 6. *О математическом моделировании явлений чередования.*

В данной работе рассмотрим один из возможных подходов к математическому моделированию явлений чередования в именах существительных русского языка при словоизменении.

З цих прикладів видно, що побудова текстової бази дозволяє не тільки точніше вибрати необхідну інформацію для смислового конструювання реферату [5], але і зробити свій внесок в розробку моделі розуміння, оскільки текстова база є семантичним представленням тексту і спільно із заголовком і ключовими словами дозволяє зрозуміти його тему.

Перспективи дослідження. Проведене дослідження показує, що стратегія аналізу тексту, яка оперує різними видами текстової інформації, кожний з яких окремо може бути недостатній для розуміння тексту, при їх сумісному використанні може забезпечити його інтерпретацію. У цьому напрямку і відбуваються наші наступні дослідження.

ЛІТЕРАТУРА

1. *Волков В. В.* Прагмастилистика и филологическая герменевтика / Валерий Вячеславович Волков // Материалы VII Междунар. конф. по проблемам семантических исследований [“Текст и методика его анализа”], (Харьков, 1994). – Ч. I : Теоретические основы лингвистики текста. – Харьков, 1994. – С. 15–17. 2. *Гендина Н. И.* Лингвистическое обеспечение автоматизированных библиотечных систем / Наталья Ивановна Гендина. – Алма-Ата : Гылым, 1991. – 221 с. 3. *Берзон В. Е.* Синтаксические сверхфразовые связи и их инженерно-лингвистичекое моделирование / Виктор Ефимович Берзон. – Кишинев : Штиинца, 1984. – С. 35. 4. *Лазаренко О. В.* Модель представлення знань в системі автоматичного реферування на базі онтологій / Ольга Владимировна Лазаренко, Дмитрий Игоревич Паченко // Проблеми семантики слова, речення та тексту : зб. наук. праць / відп. ред. Ніна Миколаївна Корбозерова. – К. : Вид. центр КНЛУ, 2005. – Вип. 15. – С. 91–98. 5. *Дейк ван Т. А.* Стратегии понимания связного текста / Тён Адрианус ван Дейк, Вальтер Кинч // Новое в зарубежной лингвистике. – Вып. 23 : Когнитивные аспекты языка. – М., 1988. – С. 153–211. 6. *Лазаренко О. В.* Моделювання узагальнення в системі автоматичного реферування : [монография] / Ольга Владимировна Лазаренко, Анастасия Анатольевна Яковенко. – Харьков : Изд-во НУА, 2007. – 136 с.