МОДЕЛЮВАННЯ СЕМАНТИЧНИХ ЗВ'ЯЗКІВ «ТЕКСТ-РЕФЕРАТ» У НАУКОВИХ ТЕКСТАХ З МЕТОЮ ПОБУДОВИ ЕФЕКТИВНОЇ СИСТЕМИ АВТОМАТИЧНОГО РЕФЕРУВАННЯ

Панченко Д. І.
Харківський гуманітарний університет
«Народна українська академія»,
м. Харків, вул. Лермонтовська, 27, тел. 716–47–23
e-mail: panchenko_di@yahoo.fr

Розв'язувана проблема моделювання семантичних зв'язків «Текст-Реферат» у системах автоматичного реферування зосереджена на створенні системи з опертям на знання й передбачає змістове опрацювання тексту в автоматичному режимі.

Змістовий аналіз тексту є першим етапом процесу реферування, після якого йдуть не менш складні завдання – стиснення змісту й синтез реферату. Отже, пошук шляхів і методів автоматичної компресії тексту є, на наш погляд, дуже важливою складовою досліджень при розробці систем автоматичного реферування.

Моделювання процесу реферування як сукупності найскладніших процесів розуміння й компресії змісту ми починаємо з вивчення не самих процесів, а з їх результату – реферату. Причому не розгорнутого, інформативного, а стиснутого, індикативного, перш за все тому, що розглядаємо його як відправну точку в дослідженні цього питання, як об'єкт найбільш простий за формою, але такий, що відбиває всі особливості реферативного тексту. Відштовхуючись від цього, були проведені дослідження змістової та синтаксичної структур реферату, які дозволили з'ясувати природу компресії в реферуванні та її наслідки щодо структури реферативних речень, і на підставі виявлених особливостей семантико-синтаксичної структури цих речень було побудовано модель індикативного реферату, на базі якої створено перша версія системи реферування «АвтоРеферат». Аналіз результатів роботи програми продемонстрував правильність породження реферативних речень у відповідності до розробленої моделі реферування, але разом з тим вказав на змістову неповноту цих речень та необхідність більш глибинного змістового аналізу первинного тексту [4].

Загальну модель реферату було побудовано у вигляді типових для індикативних рефератів семантико-синтаксичних конструкцій і сформульовано правила породження реферативних речень.

$$\boldsymbol{R=\{}СК_i(s_1)\boldsymbol{,\ СК_i(s_2)\}}\text{, (i=1,2),}$$

де R – реферат; $СК$ – синтаксична конструкція; s_1 – об'єкт; s_2 – результат.

Індикативний реферат, як правило, складається з двох речень зі значеннями – «об'єкт дослідження» $СК_1(s_1)$ та $СК_2(s_1)$ і «результат дослідження» $СК_1(s_2)$ та $СК_2(s_2)$. При цьому першим реченням у рефераті є

речення зі значенням «об'єкта», а друге – зі значенням «результату». А правила синтезу реферативної конструкції представлені набором синтаксичних, семантичних та граматичних ознак, що характеризують ці конструкції.

Наповнення моделі реферату необхідною семантикою забезпечує семантико-контексна модель реферування, яка включає модель заголовка, текстову базу, онтології предметних галузей і словник категорій реферативних конструкцій [7].

На нашу думку, аналіз змісту тексту повинен містити аналіз заголовку, тому, що заголовки науково-технічних статей дають уявлення про основний напрямок змісту статті. Ми розглядаємо заголовок як реферат мінімального об'єму або як текст з максимальним рівнем узагальнення змісту [5].

У зв'язку з цим було проведено порівняльний аналіз компресії у заголовку та рефераті. У результаті дослідження змістової і синтаксичної структури заголовка було виявлено схожість зі структурою реферату. Як і в індикативному рефераті змістова структура заголовка складається з двох метазначень – «об'єкт» і «результат». Утім на відміну від реферату вони є елементами змістової структури одного речення і вживаються у зворотній послідовності: «результат» і «об'єкт».

Подібність змістових структур стала підставою для вивчення взаємозв'язку текстів і заголовків, аби за допомогою інформації, що міститься в заголовку, виявити в тексті ті лексичні одиниці, які необхідні для семантичного наповнення моделі реферату. Для цього було побудовано класифікацію лексем СК заголовка та сформульовано загальну модель заголовка:

$$\textbf{СКЗ} = [O(б) / K\ ]\ [\mathrm{Sr}]\ [\mathrm{V}(m_5)]\ \mathbf{A}(\mathbf{m_4})\ [\mathrm{A}(m_7)]\ [\mathrm{A}(m_9)]\ [\mathrm{A}(m_8)]$$

Жирним шрифтом виділено обов'язковий елемент заголовка, наявність всіх інших – можлива (напр., *О формировании семантических признаков; Об одной реализации языка запросив; К вопросу системы управления базой данных*).

Наявність одних і тих самих семантичних компонентів у заголовку, рефераті й тексті, що є різними формами вираження одного й того ж поняття, дозволяє описати смислові структури словосполучень на різних рівнях згортання інформації, для чого необхідна наявність онтологій предметних галузей та загальнонаукової лексики. І лише тоді можна побудувати низку переходів по цих структурах від заголовка до тексту й далі до реферату при його змістовому конструюванні. Для здійснення цього переходу необхідно побудувати текстову базу, до якої входять речення, що містять слова із заголовка або ж їх смислові еквіваленти з тексту.

Текстова база складається з фактів і тверджень, пов'язаних із певною ситуацією (конкретним текстом). І на противагу онтології, яка містить не залежну від ситуації і стану інформацію, є «інформаційним ядром», що містить залежну від ситуації і стану інформацію [2].

Для побудови текстової бази знань ми відштовхувалися від понять, які містяться в заголовку документа. За ключовими словами, знайденими у заголовку відшукуються відповідні їм іменникові групи в тексті, після чого

формуються ланцюжки іменникових груп для реферативних конструкцій відповідно до наявної моделі реферату [6].

На сьогодні ведеться робота над побудовою схеми, яка забезпечує швидкий аналіз поверхневих структур тексту за рахунок використання слів-вказівників на змістові аспекти в тексті, необхідних для побудови реферату (об'єкт, результат, мета, засіб).

Однак не завжди виділення речень за допомогою слів-вказівників дозволяє здійснити оптимальний вибір речень із тексту для текстової бази. Для того щоб бути впевненим у правильності вибраних речень, необхідна наявність у системі автоматичного реферування онтологій, у яких достеменно зафіксовані всі концепти відповідної предметної галузі.

Під моделлю онтології ми розуміємо множину концептів (понять, термінів) предметної галузі, множину відношень між концептами, множину функцій інтерпретації, заданих на концептах і відношеннях онтології [1].

Для опису онтологій обрано мову OWL з метою забезпечення можливості застосування розробленої системи реферування в мережі Інтернет.

У системі автоматичного реферування було створено декілька онтологій, які сукупно дають можливість змістового конструювання реферату:

– онтологія верхнього рівня перебуває над онтологіями предметних галузей (ПГ) і є самостійною, не залежить від ПГ і конкретної задачі, оскільки описує загальні поняття (простір, час, матерія, об'єкт, подія, дія, результат тощо; в нашому випадку – об'єкт, результат, мета й засіб). У запропонованій нами моделі такий словник використовується для опису категорій реферативних конструкцій. Він є виродженою онтологією і являє собою кінцеву множину понять верхнього рівня, що відображають змістову структуру рефератів.

– проміжна онтологія загальнонаукової лексики, що містить загальні поняття й відношення для різних ПГ. Певною мірою вона запроваджується як посередня ланка між онтологіями ПГ.

– онтології предметних галузей містять поняття певної сфери знань або сфер, які входять до неї, і складаються з об'єктів та зв'язків між ними, описаних у термінології конкретної ПГ.

Зв'язок категорій з онтології верхнього рівня з об'єктами з онтології ПГ описується формалізмами, які задають принципи віднесення до цих категорій об'єктів світу.

Запропонований підхід до опису семантичних зв'язків «Текст-Реферат» у вигляді семантико-контекстної моделі реферування забезпечує аналіз змісту тексту і його трансформацію зі збереженням смислу та дозволяє суттєво поліпшити роботу системи «АвтоРеферат».

Список літератури

1. *Гаврилова Т. А.* Базы знаний интеллектуальных систем / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб.: Питер, 2000. – 384 с.

2. *Дейк ван Т. А.* Стратегии понимания связного текста / Т. А. ван Дейк, В. Кинч // Новое в зарубежной лингвистике. – Вып. 23: Когнитивные аспекты языка. – М., 1988. – С. 153–211.

3. *Клещев А. С.* Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология» / А. С. Клещев, И. Л. Артемьева // НТИ. – 2001. – Сер. 2. – № 2. – С. 20–27.

4. *Лазаренко О. В.* Моделювання процесу узагальнення в системі автоматичного реферування / О. В. Лазаренко, А. А. Яковенко. – Х.: Изд-во НУА, 2007. – 136 с.

5. *Лазаренко О. В.* Аналіз смислової структури заголовка як тексту з максимальним рівнем узагальнення / О. В. Лазаренко, Т. В. Попова // Проблеми семантики слова, речення та тексту: Збірник наукових праць. – К.: КНЛУ, 2004. – Вип. 12. – С. 143–149.

6. *Лазаренко О. В.* Классификация понятий в системе автоматического реферирования / О. В. Лазаренко // Wiek XXI. – THE 21st CENTURE.– Варшава, PWSBiA, 2002. – № 4 (6) – С.189–196.

7. *Панченко Д. І.* Моделювання семантичних зв'язків «Текст-Реферат» у системах автоматичного реферування / Автореферат дисертації на здобуття наукового ступеня кандидата філологічних наук за спеціальністю 10.02.21 – структурна, прикладна та математична лінгвістика. – Х.: Вид-во НУА, 2012.