УДК 81’322.2’33

Д. І. Панченко

МОДЕЛЮВАННЯ СЕМАНТИЧНИХ ЗВ’ЯЗКІВ «ТЕКСТ – РЕФЕРАТ» ПРИ АВТОМАТИЧНОМУ РЕФЕРУВАННІ ТЕКСТІВ

Резюме

С целью интеллектуализации процедуры индикативного реферирования при автоматической проработке текстов в статье предложено моделирование семантических связей «Текст-Реферат» с использованием текстовой базы, онтологии двух уровней и смысловых элементов заглавия. Описана онтология верхнего уровня и общенаучной лексики, и на основе предложенной методики проведена классификация элементов реферативных конструкций. Усовершенствован алгоритм процедуры реферирования и системы представления знаний на базе разработанных моделей реферата и заглавия.

Summary

The paper addresses the problem of modelling semantic links «Text-Summary» in automatic summarization with the usage of a word-oriented database, two level ontologies and the sense elements of the heading. The generalization in summarization has been defined and the peculiarities of syntactic and semantic summarizing sentences have been disclosed. The model of indicative summary has been built. The model of compression process has been developed. The title model has been constructed and investigations of title model and summary model connections have been realized. The possibility to use the heading as a starting point for automatic summarization has been under consideration.

Ключові слова: система автоматичного реферування, інтелектуальні системи, автоматизація процесів опрацювання інформації, індикативний реферат, автоматична компресія тексту, подання знань, онтологія.

Об’єктом дослідження є семантичні зв’язки тексту й реферату, їх специфіка й особливості.

Предмет становлять процедури автоматичного здобування знань із текстів природною мовою з використанням моделі реферату, моделі заголовка, текстової бази та онтологій двох рівнів.

Метою статті є опис семантико-контекстної моделі реферування та моделі подання знань у системі автоматичного реферування

з використанням онтологій, які оптимізують процедуру автоматичного аналізу та компресії текстів у процесі реферування.

Матеріалом дослідження обрано реферативні конструкції, які будують тексти індикативних рефератів. Реферати зібрано методом суцільної вибірки з текстів чотирьох предметних галузей: економіка (1000 рефератів), медицина (1000 рефератів), прикладна лінгвістика (2000 рефератів) та соціоніка (1000 рефератів). Загалом проаналізовано лінгвістичні характеристики приблизно 5000 рефератів і первинні тексти, що їм відповідають.

Актуальність роботи визначається проблемою моделювання семантичних зв'язків у природно мовних текстах з метою побудови ефективних систем автоматичного реферування з опорою на знання, що передбачає змістове опрацювання тексту в автоматичному режимі. Базовою моделлю опису знань обрано онтології як засіб класифікації терміносфери предметної галузі. Актуальність їх розробки уможливлює розвинене машинне опрацювання інформації за рахунок здійснення змістового аналізу текстів.

Підвищення якості автоматичної обробки текстової інформації постійно перебуває в центрі уваги розробників систем перекладу, реферування й пошуку інформації. На сьогодні роботи в цих галузях упритул підійшли до необхідності розробки семантичних технологій, спрямованих на автоматичне розпізнавання змісту текстів. Розробка таких технологій є одним із найскладніших завдань при проектуванні інтелектуальних систем. Окреме місце в цьому ряді займають системи, орієнтовані на автоматичне реферування текстів – побудову змістовних фрагментів, які стисло відображають зміст реферованих текстів. Роботи у сфері створення систем автоматичного реферування мають багате минуле. Понад півстоліття проводили активні дослідження, присвячені пошукам ефективних методів автоматичного реферування. Найбільш суттєвими вважаються класичний статистичний (Г. П. Лун, Р. Г. Піотровський) і позиційний (М. Л. Лосєва, М. С. Максудова) методи. Були започатковані вдалі спроби дослідження лінгвістичних і структурних характеристик реферату (В. І. Горькова, Е. А. Борохов, Х. Борко, С. Берньє, Р. Мейзел, Дж. Сміт), проводили дослідження з морфологічного й синтаксичного

аналізу текстів (В. С. Перебийніс, Н. Ф. Клименко, Г. Е. Мірам, Т. О. Грязнухіна, Н. П. Дарчук), формалізації понадфразових зв'язків у оригінальному тексті (І. П. Севбо, Л. В. Орлова) й використання інформаційно-логічних мов для побудови реферату автоматичним шляхом (Е. Ф. Скороходько).

Усі спроби автоматизації реферування, до яких удавалися протягом останніх років, одержали однозначний підсумок – залучення виключно статистичних і позиційних методів не дозволяє створити повноцінну систему реферування, яка б замість вилучення речень із тексту стисло викладала зміст. Тому протягом останніх років дослідники зосереджували увагу переважно на аналізі змісту текстової інформації (Г. С. Жданова, Г. Г. Бєлоногов, Р. Г. Котов, І. Мейні, М. Мейбюрі). Тут дослідження спираються на досягнення структурної і прикладної лінгвістики, когнітивної семантики, логічної семантики, математичної логіки та ряду інших фундаментальних і прикладних дисциплін. Результати цих теоретичних розробок знаходять застосування в системі автоматизації процесів опрацювання інформації та побудові інтелектуальних інформаційних систем у різних галузях науки (Д. В. Ланде, В. А. Широков).

Вихідним пунктом нашого дослідження є положення про те, що моделювання процесу реферування як сукупності найскладніших процесів розуміння й компресії змісту 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 такий, що відображає всі особливості реферативного тексту. Тому на першому етапі дослідження передбачають, що змістова й синтаксична структури реферату дозволять з'ясувати природу компресії в реферуванні та її можливі наслідки щодо структури реферативних речень, і на підставі виявлених особливостей семантико-синтаксичної структури цих речень побудувати модель індикативного реферату.

Наступний крок – перехід до розробки процедури здобування знань із тексту й заповнення моделі реферату відібраними з тексту іменниковими групами. При цьому передбачається, що відправною точкою

для змістового аналізу тексту є заголовок (назва статті), що дозволить знайти найважливіші змістові елементи для пошуку іменникових груп у тексті й побудови на їх основі текстової бази.

На сьогодні для розробки систем автоматичної обробки текстів широко використовують технології семантичного аналізу, що ґрунтуються на формуванні формального опису змістового навантаження тексту у вигляді фреймів, семантичних мереж або інших засобів представлення знань. Одним із шляхів у розв'язанні цього завдання є побудова когнітивної або семантико-контекстної моделі, яка забезпечила б глибинне проникнення в зміст тексту і його трансформацію зі збереженням цього змісту.

У нашому дослідженні для формального опису змісту тексту й формального опису смислових перетворень у процесі реферування використано:

1) заголовок, що презентує зміст вихідного тексту в концентрованому вигляді;

2) текстова база, яка містить речення, що є «інформаційним ядром» тексту, котре утримує інформацію, залежну від ситуації (тематики тексту);

3) онтології, що містять не залежну від тематики тексту інформацію: онтології верхнього рівня (набір змістових категорій, котрі входять до реферату), онтології загальнонаукової лексики й онтології предметних галузей.

На нашу думку, аналіз змісту тексту повинен містити аналіз заголовка, тому що заголовки науково-технічних статей дають уявлення про основний напрямок змісту статті. Важлива риса заголовка – відображення змісту тексту, котрий автор намагався донести до читача. При цьому заголовки, звичайно, пишуться максимально стисло, надто лаконічно, в них опущено всі семантично другорядні елементи. Отже, йдеться про компресію змісту тексту. Більше того, ми розглядаємо заголовок як реферат мінімального обсягу або як текст із максимальним рівнем згортання змісту.

У зв'язку з цим був проведений порівняльний аналіз компресії в заголовку й рефераті. В результаті дослідження змістової і синтаксичної структури заголовка виявилася його 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, порівняно зі змістовою структурою реферату (об'єкт – результат), послідовності: результат – об'єкт. Така подібність змістових структур реферату й заголовка, що містять однакові змістові аспекти, 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 здійснення переходу від заголовка до тексту й далі до реферату, необхідно побудувати текстову базу, до якої входять речення, що містять слова із заголовка або ж їх змістові еквіваленти з тексту. Для побудови текстової бази знань на цьому етапі досліджень ми відштовхувалися від понять, які містяться в заголовку документа. За ключовими словами, знайденими в заголовку відшукуються відповідні їм іменникові групи в тексті (будується текстова база знань), після чого формуються ланцюжки іменникових груп для реферативних конструкцій відповідно до наявної моделі реферату.

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

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

Необхідність використання в системі автоматичного реферування онтологій, як одного з найбільш перспективних способів подання знань, поставила перед нами завдання: по-перше, аналіз чинних онтологій та вибір того визначення, що відповідає аналізу змісту тексту й подальшого його подання у вигляді реферату; по-друге, було розглянуто варіанти використання онтологій в сучасних інформаційних системах.

Ми погоджуємося з точкою зору авторів [5], за якою онтологія – це експліцитна специфікація концептуалізації, де в ролі концептуалізації виступає опис значної кількості об'єктів предметної галузі та зв'язків між ними. Також вважаємо, що модель онтології кожної предметної галузі повинна містити як формальні елементи, так і змістове тлумачення в термінах, зрозумілих фахівцям, тобто поняття, визначені в словнику, повинні бути прийнятими в даній предметній галузі термінології [3; 4].

Побудовані нами онтології загальнонаукової лексики, метазначення для використання в системі автоматичного реферування та запропонований алгоритм побудови текстової бази є лише першим кроком до створення відповідної системи подання знань у системі автоматичного реферування, що зумовлює *перспективу* подальшого дослідження.

Взагалі моделювання семантичних зв'язків «Текст-Реферат» у процесі інтелектуального реферування – внесок у розв'язання проблеми автоматичного опрацювання текстової інформації, що дозволило з'ясувати, як відбувається змістове згортання в процесі реферування і які специфічні ознаки у структурі реферативних речень і заголовку воно унаочнює. Результати роботи можуть бути використані в системах автоматичного опрацювання інформації для покращення якості процесів реферування.

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. Кальченко Д. Интеллектуальные агенты семантического Web'а / Д. Кальченко // Компьютер Пресс. – 2004. – № 10. – С. 26–32.

4. 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.

5. Gruber T. R. A translation approach to portable ontology specifications / T. R. Gruber // Knowledge Acquisition. – 1993. – № 5. – P. 199–220.