Зоя Куделько
Чернівецький національний університет

АВТОМАТИЧНЕ РЕФЕРУВАННЯ ТЕКСТІВ

У даній роботі йдеться про проблеми та методи автоматичного реферування текстів. Описується метод структурно-семантичного реферування тексту. Його головними параметрами є: функціональна вага речень, функціональна вага слова, лінійний коефіцієнт. Механізм реферування показано на економічному тексті.

This article highlights the problems and methods of the automatic text abstracting. The method of the structural-semantic abstracting of the text is described. The main parameters of it are functional value of the text, functional value of the word, linear coefficient. The abstracting mechanism is shown on the marketing text.

Термінологія (сукупність термінів) існує у 2-х вимірах: як результат фіксації наукового пізнання (термінологічні словники) та функціонування як у контексті наукової і навчальної літератури, так і в комунікативному аспекті [3, с.3].

У бурхливий час нових технологій, нових методик, інформаційного буму ми постаємо перед проблемою обробки та засвоєння нових знань. Причому обсяг їх настільки великий, що вчений чи дослідник не в змозі ознайомитись з усіма новинками. Тому складання реферату (від лат. refero – повідомляю), який є коротким викладом у письмовому вигляді або у формі публічної доповіді змісту наукової праці (праць) літератури за темою є надзвичайно необхідним.

Важливим джерелом наукової інформації є науковий документ. За формою книги, журнали, статті тощо належать до письмових наукових документів. Вони можуть бути первинними і вторинними.

До первинних документів відносять монографії, збірники, матеріали наукових конференцій, симпозиумів, конгресів тощо, підручники, журнали, газети й інші видання.

Реферат і анотація належать до вторинних документів [1, с. 5].

Реферат має наступні функції:

1) передає основну інформацію, що міститься в реферованому документі;
2) описує первинний документ;
3) є джерелом для отримання довідкових даних.

Реферати мають певні спільні риси:

1) в них відсутні роздуми, докази, історичні екскурси;
2) матеріал не подається у формі консультації чи опису фактів;
3) інформація викладена стисло, точно, без викривлень та суб'єктивних оцінок.

Всі методики автоматичного реферування засновані на відборі ключових слів, якими в науково-технічних текстах виступають термінологічні одиниці. При цьому терміни служать основою формального аналізу семантики тексту [2, с.88].

Існують кілька методик формального аналізу тексту: лексична, семантична та синтаксична методики. Семантична методика є найбільш досконалою, але надійнішою і

найбільш прийнятною є методика спрощеного аналізу. Прикладом такої методики спрощеного аналізу може служити алгоритм компресії інформації тексту.

Під компресією тексту ми розуміємо один з можливих варіантів його квазіреферування, тобто складання реферату на основі вибору та комбінування готових фрагментів тексту.

В основі більшості методик реферування тексту лежить відносна частотність слів, з яких потім і складається реферат, але у процесі відбору найбільш частотних лексичних одиниць дуже часто випадають інші, семантично вагоміші слова, крім того складання реферату з найбільш частотних одиниць є завданням не з легких, а іноді й неможливим.

Тому, на нашу думку, найбільш прийнятними і досконалими є методи структурно-семантичного реферування.

В нашому дослідженні був застосований метод, розроблений в Інституті кібернетики АН УРСР. Він визначає кількісні параметри семантичної сітки тексту і окремих його елементів – речень і слів. Головними параметрами є: функціональна вага речень, функціональна вага слова, лінійний коефіцієнт [2, с.89].

В основі даного методу лежить гіпотеза, що найбільш інформаційно важлива фраза тексту має найбільше число лексичних зв'язків з іншими фразами тексту. Лексичний зв'язок – це наявність однакових основ в реченнях, що розглядаються. Застосуємо даний метод до тексту, поданого нижче.

The True Impact of Corruption

1. For years, it was believed that bribery and other forms of corruption were effective and even necessary tools for doing business in developing countries **(12)**. 2. By greasing the right palms, so the thinking went, firms achieved a competitive advantage **(0)**. 3. Not so, research undertaken by the World Bank and others shows that far from lubricating business activity, bribery actually fuels the growth of excessive and discretionary regulations **(5)**. 4. Bribery, in short, feeds on itself, producing layer upon layer of bureaucracy eager to get in on the action **(3)**. 5. The fact of the matter is that in countries where corruption is recognized to be high, firms spend more time with bureaucrats and public officials negotiating licenses, permits, and taxes **(8)**.

6. The evidence also shows that countries with notoriously high levels of corruption risk marginalization in a world of rapid economic integration **(9)**. 7. Many of the challenges we face today can be traced, in part, to cronyism, shallow disclosure requirements, and opaque record-keeping **(0)**. 8. Open markets cannot work behind closed doors **(0)**. 9. Both private capital flows and official development assistance are increasingly discriminating with regard to policy performance and institutional integrity **(2)**. 10. Investors today have too many options, and they are better able to move their money to where the risks of corruption are less pronounced **(6)**. 11. And official donors, with shrinking aid budgets, have also drawn the line **(0)**. 12. Well-informed publics and wary aid agencies and development institutions are seeking returns on their aid investments – in the form of poverty reduction and social development – with the same rigor that private investors look for financial returns **(1)**. 13. Perceptions in donor countries that corruption in recipient countries sends their aid assistance down a black hole is one of the greatest threats to future aid **(8)**.

14. We at the World Bank, as with other multilateral organizations, are fully aware that despite continued vigilance and state-of-the-art auditing and investigative measures, the projects that we support are not immune from the pressures of corruption **(7)**. 15. There is simply no way to fully isolate individual projects and program lending from fraud if it is pervasive throughout the environment in which they function **(0)**. 16. This calls for continued efforts on the part of the Bank to pursue and prosecute fraud wherever we find it, while simultaneously strengthening the institutional structures that will ultimately help stop corruption at its source **(7)**. 17. It will be a difficult, long-term struggle **(0)**. 18. But make no mistake, it is a winnable fight, and one that must be fought **(0)** [4, p.15–16]

Текст було поділено на окремі речення й пораховано кількість лексичних зв'язків для кожного з них (цифра у дужках опісля кожного речення). Звідси випливає, що найбільш важливими реченнями тексту є речення № 1, 5, 6, 13, 14, 16.

Для отримання точніших результатів доцільно також брати до уваги функціональну вагу слова, яку можна визначити за допомогою дерева семантичних зв'язків речення, запропонованої Кияком Т.Р. [2, с.89], де враховується ступінь віддаленості слова від вершини дерева всього речення:

$$F= t: l (n-1),$$

де F – функціональна вага речення;

t – кількість лексичних зв'язків в реченні;

l – кількість лексичних основ в реченні;

n – кількість речень в тексті.

Визначимо функціональну вагу першого речення:

F = 12 : 11 (years, believed, bribery, forms, corruption, effective, necessary, tools, business, developing, countries) (18-1) = 0, 06

Таким же чином визначаємо функціональну вагу кожного речення в межах досліджуваного тексту і виділяємо речення з найбільшою функціональною вагою.

Потім лишаємо поза увагою речення з найменшою функціональною вагою і так поступово залишаємося з реченнями, кількість яких відповідає розмірам реферату, який ми хочемо скласти. В нашому випадку це, скажімо, 4 речення:

Реферат

1. For years, it was believed that bribery and other forms of corruption were effective and even necessary tools for doing business in developing countries. 5. The fact of the matter is that in countries where corruption is recognized to be high, firms spend more time with bureaucrats and public officials negotiating licenses, permits, and taxes. 6. The evidence also shows that countries with notoriously high levels of corruption risk marginalization in a world of rapid economic integration. 13. Perceptions in donor countries that corruption in recipient countries sends their aid assistance down a black hole is one of the greatest threats to future aid.

Отриманий реферат досить задовільно передає зміст поданої вище статті.

Даний метод був також застосований для аналізу ще 10 текстів економічної тематики. Результатом є реферати з 6–7 речень, які досить добре, стисло передають головну ідею статей, текстів. Крім того за доцільність цього методу свідчить той факт, що до рефератів увійшли найбільш частотні лексичні одиниці досліджуваних текстів, що говорить про семантичну вагу не лише речень, а і слів у тексті.

Слід зазначити, що темі автоматизації роботи з термінами присвячено багато робіт як на терені України, так і за кордоном. Розроблені комп'ютерні програми добору термінів з текстів, виділення ключових слів та виразів, програми визначення частотності термінів, їх перекладу. Це програми: DANTERM, CDS/ISIS, QUIRK [5, p.374].

На основі вище поданої методики нами була розроблена власна комп'ютерна програма автоматичного реферування текстів.

Схема роботи програми

І. Поділ тексту на речення.

- Кожне окреме речення є окремим елементом масиву і позначається A [i], де [i] порядковий номер речення в тексті.
- Кожне речення поділяється на окремі слова B [i, j], де і – порядковий номер речення в тексті, а j – порядковий номер слова в даному реченні.

1. Окремо формується масив службових слів, які не розглядаються в процесі реферування тексту.
2. Підраховуємо для кожного B [i, j] кількість повторів у тексті, за виключенням речення A [i]. І число цих повторів заносимо в масив sum [i] (для кожного A [i] – це свій sum[i]).
3. Сортуємо отримані sum [i] у порядку зменшення і виділяємо кількість речень, потрібну для реферату.
4. Обрані реченння і будуть нашим рефератом. Зауважимо, що розглядаються лише основи слів.

Програма виконана на мові об'єктного орієнтування DELPHI 6.0. Інтерфейс програми виконаний у зручній для користувача формі, в самій програмі подані правила користування нею.

Отже,

1) в основі автоматичного реферування тексту лежить методика спрощеного аналізу, а саме алгоритм компресії інформації тексту;

2) головними параметрами викладеного вище методу є функціональна вага речень, функціональна вага слова, лінійний коефіцієнт;

3) запропонована програма буде надзвичайно корисна для широкого кола дослідників даної проблеми, оскільки дає можливість опрацювати велику кількість фактичного матеріалу за відносно короткий період часу, може бути вдосконалена і застосовуватись у різних галузях знань.

1. Аннотирование и реферирование. Пособие по английскому языку / Г.И.Славина, З.С.Харьковский, Е.А.Антонова, М.А.Рыбакина. – М.: Высшая школа, 1991. – 156с. 2. КиякТ.Р. Лингвистические аспекты терминоведения: Учеб. пособие. – К.: УМК ВО, 1989. – 104с. 3. Симоненко Л.О. Українська наукова термінологія: стан та перспективи розвитку // Українська термінологія і сучасність: Зб. наук. праць. Вип. IV / Відп. ред. Л.О.Симоненко. – К.: КНЕУ, 2001. – С.3–8. 4. Economic Perspectives. Corruption: An Impediment to Development. A USIA Electronic Journal. – U.S. Information Agency, 1998 – 52c. 5. Terminology in Advanced Microcomputer Applications: Proceedings of the 4th TermNet Symposium; Tools for Multilingual Communication / TAMA '98. – Vienna: TermNet, 1998. – 374c.