УДК 681.3.06

СТІРЕНКО С.Г.

ФОРМАЛЬНИЙ ОПИС СТРУКТУРНО-ПАРАМЕТРИЧНИХ ХАРАКТЕРИСТИК ТЕХНІЧНОГО ТЕКСТУ

Запропоновані структурно-параметричні характеристики текстових документів, наведені їх обґрунтування і відповідність задачам формальної оцінки блоків технічного документу згідно з рівнем ієрархії. Показано, що при умові виконання вимог вузькоспеціалізованого застосування, дані характеристики дозволяють отримати оцінки семантичної близькості, порівнянності і міру семантичної відповідності для структурних одиниць технічної природної мови.

Proposed structural and parametric characteristics of text documents are the reasons for them, and compliance with an assessment of the challenges of formal blocks of a technical document in accordance with the level of the hierarchy. We show that, subject to compliance with the requirements of specialized applications, these characteristics will provide a semantic affinity evaluation, comparability and consistency to measure the semantic units of natural language technology.

Вступ

Достатньо актуальним завданням є опис структурних компонентів технічних текстів (наприклад, специфікації на програмні продукти). Ринок програмних продуктів розвивається так динамічно, оновлення версій відбувається так часто, що фірми-розробники просто не в змозі відволікати інтелектуальні ресурси, необхідні для складання задовільної документації супроводу. Оскільки витрати часу на опис можна порівняти з витратами на розробку продукту, але вимагають більшого об'єму висококваліфікованої ручної праці, деякі фірми прагнуть заощадити засоби на розробці документації і програють при необхідності розвитку програмного забезпечення. У такій ситуації розвиток засобів автоматизації обробки текстових блоків знижують інтелектуальне навантаження на розробників технічної документації.

При формальному розгляді технічного тексту як структурованого набору символьної інформації стає зрозуміло, що основні елементи його багаторівневої структури – загальновживані слова, а головне – односкладові та багатоскладові поняття проблемної області за своїми взаємозв’язками – можуть бути наведені на основі статистичного аналізу. Проведення структурного статистичного аналізу на базі великої кількості текстів дозволяє автоматично сформулювати декларативне представлення граматики мови, а на множині текстів, які описують проблемну область, автоматично побудувати описання її семантики у вигляді мережі понять та їх зв’язків [1-4]. Відображення нового тексту, який аналізується, на мережу дозволяє виділити в ньому концептуальні поняття та їх зв’язки, розкласифікувати ділянки (частини) тексту за темами на основі віднесення до відповідних понять семантичної мережі і, таким чином, визначити структуру змісту тексту.

Щоб отримати лексико-семантичну інформацію про технічну документацію (ТД) будемо використовувати просторово-векторну модель (ПВМ) [5]. Кожен технічний документ (ТДТ) D_n відображається вектором (w_{n1}, w_{n2},…, w_{nM}), де w_{nk} – вага (або важливість) ключового слова або словосполучення (КСС) t_k з розділу Q_t бази знань (БЗ) G_i для ТДТ D_n, а M – кількість елементів розділу Q_t. Вектор (w_{n1}, w_{n2},…, w_{nM}) будемо називати *лексико-семантичним профілем документу*[1] (ЛСПД).

Частіше за все вагу документу визначають як частоту його появи у тексті [6], але для наших цілей доцільно використовувати логарифмічну шкалу: $\log(N/df_i)$, де N – число документів, а df_i частота використання терміну t_i.

У відповідності до ПВМ, при кількості документів N отримаємо частотну матрицю зустрічальності $F=N\times M$, в якій кожен рядок є ЛСПД:

$$F = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1M} \\ w_{21} & w_{22} & \cdots & w_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NM} \end{pmatrix}. \quad (1)$$

[1] Оскільки різниця між синтаксисом та семантикою доволі розмита), іншими словами складно визначити, де закінчується синтаксис і починається семантика. Більше того, на наш погляд семантика починається вже в морфології.

Оцінка міри семантичної відповідності документів

Для контролю змістової відповідності [7] (семантичної близькості) *Sim* документів D_n будемо використовувати запити *qr*, кожен з яких представлений вектором (qr_1, qr_2,…, qr_M), де qr_i – вага КСС *i* в запиті *qr*.

Оцінювання виконується шляхом розрахунку міри змістової відповідності між документом D_n і запитом *qr* в призначеному змістовому просторі, вираженому вектором. Звичайно мірою близькості вважають косинус кута між двома векторами:

$$Sim_{\cos}(D_n, qr) = \frac{D_n \cdot qr}{\|D_n\| \, \|qr\|} = \frac{\sum_{i=1}^{M} w_{ni} qr_i}{\sqrt{\sum_{i=1}^{M} w_{ni}^2 \sum_{i=1}^{M} q_i^2}}. \quad (2)$$

Також можуть бути використані декілька інших варіантів розрахунку міри змістової відповідності (відстань χ^2, дивергенція Куллбека-Леблера). Їх опис представлений в роботах [8, 9].

Концепція розподіленої семантики

Основна ідея концепції розподіленої семантики (РС) полягає у твердженні лінгвістів, що існує сильна кореляція між розподіленою характеристикою слова (словосполучення) W_{Hd}, яка є об'єктом спостереження, та її значенням W_m. Іншими словами семантика слів Sem_w – це набір контекстів S_c, в яких вони використовуються. Розглянемо три вирази з однієї предметної області G_i:

1. *Вузол X знаходиться в гнізді системної плати.*

2. *Вузол X потребує хорошого охолодження.*

3. *Адміністратор апаратно збільшив тактову частоту вузла X.*

Множина слів W_{si} = {*вузол, знаходиться, гніздо, системна плата, потребує, охолодження, адміністратор, апаратно, тактова частота*} створює спрощене уявлення про сукупний контекст у даному дискурсі поняття *X*. Отримано загальне представлення про ідентифікацію вузла X (мікропроцесор).

Таким чином, резюмуючий тезис можна сформулювати наступним чином: два КСС є семантично близькими, якщо їх розширений контекст також є близьким. Звідси можна зробити висновок, що вищенаведене твердження можна застосувати и до цілих документів: два документи є семантично близькими, якщо складові їх КСС також семантично близькі.

Оцінка семантичних характеристик документу

Поняття *контекст слова* визначає модель зустрічальності. *Спільна зустрічальність*[2] [10, 11] між двома словами визначається як зустрічальність двох слів у даному текстовому модулі у вибраному змістовому просторі. Частоти парної зустрічальності граматичних слів розглядаються різними авторами як істотні характеристики формальної структури тексту.

Для обробки семантичних конструкцій пропонується ієрархія понять *G*. [68] В якості найбільш загального поняття або категорії береться те, яке має найбільший об'єм G_g і, відповідно, найменший зміст. Це найвищий рівень абстракції для даної ієрархії. Потім дане загальне поняття конкретизується, тим самим зменшується його об'єм і збільшується зміст G_{gi} (де *i* – номер рівня ієрархії). З'являється менш загальне поняття, яке на схемі ієрархії буде розташоване на рівень нижче обраного поняття. Цей процес конкретизації понять може продовжуватись доти, доки на самому нижньому рівні не буде отримане поняття, подальша конкретизація якого в даному контексті або неможлива, або недоцільна.

Отже, у якості текстового модуля використовуємо вікно з *k* слів, речень, параграфів, розділів або цілих документів, згідно з восьмирівневою моделлю [12]. Для конкретного КСС *профіль спільної зустрічальності* (ПСЗ) визначається як вектор спільної зустрічальності між цим КСС та кожним елементом, який апріорно належить до заданої множини КСС, створеної як множина *семантичних індексних характеристик* (СІХ).

Для множини *M* КСС будується сочастотна матриця $C = M \times P$, кожний рядок якої представляє собою ПСВ КСС t_i:

$$C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1M} \\ c_{21} & c_{22} & \cdots & c_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \cdots & c_{MP} \end{pmatrix}. \quad (3)$$

Весь документ D_n оцінюється за допомогою середньозваженого ПСВ для КСС і має вигляд:

$$D_n = \sum_{i=1}^{M} w_{ni} c_i \,. \quad (4)$$

Вага w_{ni} задана для кожного ПСВ c_i і має той же КСС t_i як було показано вище. Отже, ком-

[2] або *созустрічальність, сочастотність.*

плект документів представляється за допомогою добутку матриць $D = FC$.

Потрібно звернути увагу, що розмір вектора представлення Vr_i документу D_n – це розмір масиву СІХ, який може бути менше, ніж розмір масиву КСС. Проте ЗС має такий же розмір масиву КСС як і ПВМ, тому що всі КСС у масиві представлені середньозваженою величиною за допомогою їх ПСЗ в СІХ.

Модель ЗС має деяку перевагу, з точки зору оптимізації (скорочення) розмірності шляхом отримання об'єднаного вектора представлення Vc_i документу D_n через представлення КСС двох: спільної зустрічальності та СІХ.

Проблема вибору СІХ в ПВМ вирішується шляхом індексування масиву КСС на основі частоти їх зустрічальності в документі D_n. Для того, щоб не втратити, або не зашумлювати інформаційність КСС, що знаходиться в документі, будемо комбінувати два підходи РС і ПВМ.

Якщо F^* реструктурована середньозважена матриця тільки для КСС, які індексуються, а α скалярний комбінований параметр у діапазоні від 0 до 1, то комбіноване представлення документу D_n буде мати наступний вигляд:

$$D_n = (1-\alpha)\, F^* + \alpha FC. \qquad (5)$$

Уточнення розподіленої семантичної моделі

Зробимо деякі уточнення, які стосуються обчислення матриць КСС, які спільно зустрічаються, оскільки проста модель спільної зустрічальності (в основі якої лежить спільно зустрічні КСС без визначеного (будь-якого) текстового модуля з додатковими обмеженнями) може мати незадовільні результати обчислень. Розглянемо текстовий фрагмент: «*Потужний мікропроцесор швидко обробляє запити з робочих станцій»*

Візьмемо в якості текстового модуля ціле речення. Тоді КСС: *робоча станція*, *потужний мікропроцесор* з точки зору спільної зустрічальності є релевантними, тоді як *мікропроцесор запитів*, або *станція запитів* є ілюзорною (або найменш передбачуваною).

Ця проблема вирішується шляхом введення додаткової синтаксичної інформації в процес обчислень. Інакше кажучи, спільну зустрічальність потрібно вираховувати тільки:

1. Між КСС в однакових синтаксичних групах;

2. Між головними СКК в різних групах.

Зробимо уточнення для попереднього прикладу, в якому розглядалась спільна зустрічальність тільки між іменниками, дієсловами та прикметниками. У прикладі на малюнку нижче жирним курсивом виділені головні КСС у даній синтаксичній групі.

Підхід до верифікації фрагментів технічних документів у цільовій предметній області

У зв'язку з вищесказаним розглянемо методику визначення приналежності ТДТ (або його контексту) до конкретної проблемної області:

1. Для кожного з наявних документів отримуємо матрицю А частот парної зустрічальності.

2. Аналізуючи кожну матрицю, виділяємо для кожного тексту сукупність зв'язків з високими (тобто такими, що перевищують деяке порогове значення) частотами.

3. При порівнянні отриманої сукупності "істотні" зв'язків тексту, що досліджується, з іншими визначається відповідність, який текст характеризується найбільш близькою по деякому критерію сукупністю „істотних" семантико-синтаксичних зв'язків (ССЗ).

У тому випадку, коли передбачається введення в аналіз „загальну частину", або „ядро", в методику вбудовуються два додаткових пункти:

а) Порівнюючи отримані сукупності ССЗ, виділяємо загальномовне ядро, тобто набір таких зв'язків, які містяться в усіх (або майже усіх) текстах.

б) Сформоване "загальномовне ядро" видаляється з кожної сукупності відібраних ССЗ з високими частотами; "істотні" ССЗ кожної сукупності, що залишилися після цього уже більше характеризують приналежність до контексту.

Загальномовне ядро має різні рівні аналізу. При дослідженні приналежності одного документу – це найбільш характерні особливості його стилю. При дослідженні документів певної проблемної області, але різних типів документів – це риси, властиві перш за все проблемній області.

Висновки

В роботі були визначені основні підходи до логічної організації моделей текстових докуме-

нтів, реалізація яких в СПРТД створює передумови до зменшення часу розробки типових технічних документів, а також надає потенційну можливість підвищення рівня інтелектуалізації систем даного класу. Отримані результати можна виразити наступним чином:

1. Виділені компоненти моделі предметної області, які необхідно розробити, і визначені базові формалізми дослідження.

2. Запропоновані структурно-параметричні характеристики текстових документів, наведені їх обґрунтування і відповідність задачам формальної оцінки блоків ТДТ згідно з рівнем ієрархії.

3. Показано, що при умові виконання вимог вузькоспеціалізованого застосування, дані характеристики дозволяють отримати оцінки семантичної близькості, порівнянності і міру семантичної відповідності для структурних одиниць технічної природної мови.

Список литературы

1. Beeferman D., Berger A., Lafferty J. Statistical Models for Text Segmentation // Machine learning. – 1999. – Vol. 34. – P. 1-34.
2. Berry M.W., Dumais S.T., O'Brein G.W. Using Linear Algebra for Intelligent Information Retrieval // SIAM Review. – 1995. – Vol. 37, № 4, P. 573-595.
3. Rijsbergen C.J. van. Information Retrieval, Butterworths // London. – 1979. – 358 p.
4. Yang Y. An Evaluation of Statistical Approaches to Text Catigorization // 1999. –Vol. 1, № 1. – P. 69-90.
5. Salton G., Buckley C. Term weighting approaches in automatic text retrieval // Information Processing and Management. – 1988. – № 24. – P. 513-523.
6. Salton G., Yang C., and Yu C. A theory of term importance in automatic text analysis // Journal of the American Society for Information Science. – 1975. – №12. – P. 248-251.
7. Стиренко С.Г., Пикуза О.В. Оценка стилистических характеристик текстовой информации // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. – К.: Век+. – 2000. – № 35. – С. 127-131.
8. Lee L. Similarity–Based Approaches to Natural Language Processing. PhD thesis, Harvard University. – 1997.
9. Rajman M., L. Lebart. Similarités pour données textuelles // In 4th International Conference on Statistical Analysis of Textual Data (JADT'98), Nice. 1998. – P. 336-342.
10. Фукс В. По всем правилам искусства (точные методы в исследованиях литературы, музыки и изобразительного искусства. Искусство и ЭВМ. М.: Мир, 1975. – 320 с.
11. Lebanon G. Metric Learning for Text Documents // IEEE Transaction on Pattern and Machine Intelligence. – 2006. – Vol. 28, № 4. – P. 497-508.
12. Стиренко С.Г. Автоматизована розробка технічної документації з використанням засобів інтелектуальної підтримки // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. - К.: Век+, – 2006. – № 45. – С. 173-179.