УДК 004

ВЫДЕЛЕНИЕ АББРЕВИАТУР ДЛЯ ПОСТРОЕНИЯ СЛОВАРЯ ПРЕДМЕТНОЙ ОБЛАСТИ

Липинская А.Г, Поточняк Я.В.
д. т. н., проф. Кунгурцев А.Б.
Одесский Национальный Политехнический Университет, УКРАИНА

АННОТАЦИЯ. В данной работе описан процесс выделения аббревиатур для построения словаря предметной области.

Введение. Словарь предметной области (СПО) – документ, необходимый для разработки информационной систем, проектирования базы данных, сопровождения программных продуктов [1]. Для построения СПО выделяют термины из документов, представляющих некоторую предметную область, и определяют частоту их вхождения в тексты. Термины с высокой частотой появления в документах включаются в словарь [2]. Некоторые термины в документах представляются аббревиатурами. Обычно они относятся к категории наиболее часто употребляемых. Для представления таких терминов в СПО необходимо привести не только аббревиатуру, но и её расшифровку. Существуют словари сокращений [3], однако их использование для конкретной предметной области невозможно, ввиду необходимости выбора из десятков и сотен предлагаемых вариантов.

Цель работы. Целью работы является выделение из текста двух типов аббревиатур – инициальных и сложносокращенных.

Математическая модель выделения аббревиатур. Слова, определяющие имена в формулах: Punctuation mark Letter Lowercase letter Cursive letter capital letter The text consists of sentences Size Attribute Image of a letter

Пусть Sn представляет некоторое предложение:

$Sn = e_1 e_2 ... e_i ... e_n$,

где e_i - элемент предложения (слово, либо знак препинания).

Определим слово W как последовательность символов

$W = s_1 s_2 ... s_j ... s_k$, (1)

Определим операцию выделения символа из слова $s_j = W[j]$

и отношение принадлежности некоторого символа слову $s_j \in W$

Каждая буква характеризуется изображением li, размером is (строчная, прописная) и алфавитом al (кириллица, латиница)

$l =< li, ls, al >$,

где ls может принимать два значения low и cap;

- al может принимать значения la и ki.

Будем считать, что слово должно начинаться с буквы.

Определим знак препинания, используемый внутри предложения как

$Pm = \{",",":",";","(",")","–"\}$

Полагая, что аббревиатура может содержать только буквы, запишем условие первого появления аббревиатуры в тексте. Пусть $W_m = s_1 s_2 ... s_j ... s_k$ - некоторое слово, где m - номер слова, как элемента предложения. Если

$e_{m-1} = "(" \wedge e_{m+1} = ")" \wedge \forall s_j \mid s_j.ls = cap$, (2)

То можно считать, что W_m представляет собой аббревиатуру. Обозначим её Ab

Следующей задачей является определение текста Ta, соответствующего аббревиатуре. Полагаем, что Ta расположен слева от открывающей скобки аббревиатуры и между его элементами отсутствуют какие-либо знаки препинания. Тогда

$$Ta = e_p ... e_{p+(k-1)}, \quad (3)$$

где $p = m - (k+1)$

Тогда условием успешной операции по определению Ta является

$$\forall e_i (i = p, p+(k-1)) \mid e_i[1].li = Ab[j].li(j = 1, k) \quad (4)$$

Если условие (4) не соблюдается, то можно предположить, что в аббревиатуре присутствует сложносокращенное слово. Например: «…система автоматизированного проектирования (САПР)…». Здесь «ПР» представляет собой сокращение слова «проектирование».

На первом этапе предлагается определить первую букву аббревиатуры: $l = W_m[1].li$. Определить количество повторений первой буквы в аббревиатуре n. В соответствии с п.1 исключить из текста предложения слева от аббревиатуры знаки препинания и слова, заключенные в круглые скобки. Начиная от слова с номером $m-1$ искать n слов, таких что $l = e_j[1].li(j = m-1, m-k)$

Если n таких слов не обнаружено, то можно предположить, что одна из букв l в аббревиатуре W_m входит в сокращение не на первой позиции. В этом случае уменьшаем число $n := n-1$ и продолжаем поиск слов. Если в результате оказалось что $n = 0$, то поиск аббревиатуры прекращается. В противном случае переходим ко второму этапу.

На втором этапе необходимо убедиться, что имеется возможность определить Ta в соответствии с W_m .

На рис. 1 изображена упрощенная схема выделения аббревиатур.

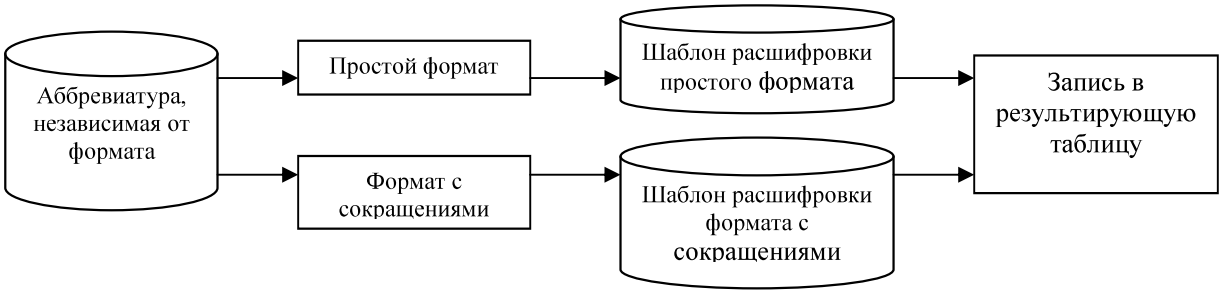

Рисунок 1 – Схема автоматизированной обработки аббревиатур разных форматов

Выводы. Разработанные модели, алгоритмы и их реализация в виде программного продукта позволили выделять основные типы аббревиатур из анализируемых текстов, что позволило усовершенствовать технологию создания словарей предметной области.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Кунгурцев А. Б. Метод автоматизированного построения толкового словаря предметной области/ А.Б. Кунгурцев, Я.В. Поточняк, Д.А. Силяев//Технологический аудит и резервы производства — № 2/2(22), 2015. – С. 58 – 63.

2. Кунгурцев, О. Б. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою [Текст] / О. Б. Кунгурцев, С. В. Ковальчук, Я. В. Поточняк, М. В. Широкоступ // Технічні науки та технології – №3 (5), 2016. – С. 164-174.

3. Словарь сокращений русского языка [Електронний ресурс]. - Режим доступа: URL http://www.sokr.ru/