УДК 002.53/55:001.814

А. В. Жарковский, А. А. Лямкин, С. А. Тревгода
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Алгоритм автоматического реферирования научно-технических текстов на основе структурного подхода

Предлагается алгоритм автоматического реферирования научно-технического текста на основе учета структуры текста. Представлены алгоритмы реализации процессов анализа и построения структуры текста на различных этапах подготовки реферата.

Автоматизация, алгоритм, формализация, автоматическое реферирование, функциональные отношения, структура текста

Стремительный рост объемов научно-технической информации, представленной в электронном виде, приводит к необходимости постоянного совершенствования методов и средств ее автоматизированного приема, хранения и обработки, описания и анализа, в частности – методов получения сжатого представления текстовых документов – рефератов.

Согласно исследованиям в области компьютерной лингвистики, текст по природе своей нелинеен. Его структура определяется особенностями внутренней организации единиц текста и закономерностями взаимосвязи этих единиц в рамках текста как цельного сообщения [1], [2]. Как показала практика, различные статистические методы автоматического реферирования недостаточно эффективны. Они интерпретируют текст в виде набора линейно упорядоченных слов, словосочетаний и предложений, игнорируя при этом лингвистическую взаимосвязанность различных элементов естественного языка, что приводит к потере значимой информации.

Для преодоления этого недостатка был разработан метод формализованного описания структуры научно-технического текста [3]. Метод основан на использовании теории риторической структуры, согласно которой любой текст может быть представлен в виде дерева, узлы которого представляют собой элементарные текстовые элементы (ЭТЭ) или группы таких единиц, связанных между собой функциональными отношениями. Текстовый элемент, вступающий в функциональное отношение, может играть в нем различные роли. Более значимый его компонент называется ядром, менее значимый – сателлитом. Примерами таких отношений служат: условие, предпосылка, противопоставление, последовательность, обстоятельство.

Формализация описания структуры текста в соответствии с этим методом включает в себя следующие три этапа:

1) разработка критерия корректности структуры текста;

2) определение характеристик, описывающих структуру текста;

3) определение ограничений на корректные структуры текста.

Критерий корректности структуры текста формулируется следующим образом: если функциональное отношение лежит между двумя текстовыми элементами структуры текста, тогда оно же лежит между по крайней мере двумя ключевыми ЭТЭ-потомками этих элементов.

Основная идея введенного критерия корректности структуры текста заключается в том, что элементы-ядра играют бóльшую роль в тексте, нежели элементы-сателлиты и, в принципе, при удалении всех сателлитов смысл текста должен

сохраниться. Если применить этот принцип рекурсивно ко всему тексту, представляя его в виде дерева, получим дерево, удовлетворяющее критерию.

В качестве характеристик структуры текста выбраны следующие:

– $S(l, h, status)$, где l – левый индекс ЭТЭ; h – правый индекс ЭТЭ; *status* – значение статуса узла – показывает статус ЭТЭ. Он может иметь значения *NUCLEUS* (ядро), *SATELLITE* (сателлит) или *NONE* (не определен);

– $T(l, h, relation\_name)$, где $relation\_name$ – функциональное отношение – показывает имя функционального отношения, которое лежит между своими прямыми потомками;

– $P(l, h, unit\_name)$, где $unit\_name$ – название или индекс ЭТЭ – показывает имя ключевого ЭТЭ (отражающего значимую информацию).

Статус, тип и множество важных потомков, которые связаны с каждым узлом дерева, дают достаточную информацию для полного описания текстовой структуры.

Для сужения множества генерируемых корректных структур в алгоритме автоматического построения структуры текста разработаны ограничения на корректные структуры текста, представленные в виде системы предикатов [3].

В соответствии с приведенной формализацией предлагается обобщенный алгоритм автоматического реферирования, который может быть описан как совокупность процедур (рис. 1).

Основными составляющими данного алгоритма являются: определение ключевых фраз и построения функциональных отношений между фрагментами текста в виде ЭТЭ или совокупности ЭТЭ (блоки 1–2); процедура построения оптимальной структуры текста (блоки 3–4); ранжирование по важности листьев (ЭТЭ) построенного структурного дерева для всего текста (блок 5) и составление аннотации в соответствии с требуемым объемом (блок 6).

Основной задачей при построении структуры текста является определение набора функциональных отношений между элементарными текстовыми элементами или частями предложений. Известные подходы к решению этой задачи основаны на использованиb глубокого семантического анализа текста, требующего полных баз знаний и соответствующих словарей русского языка и до практической реализации не доведены. В настоящей работе предлагается другой подход.

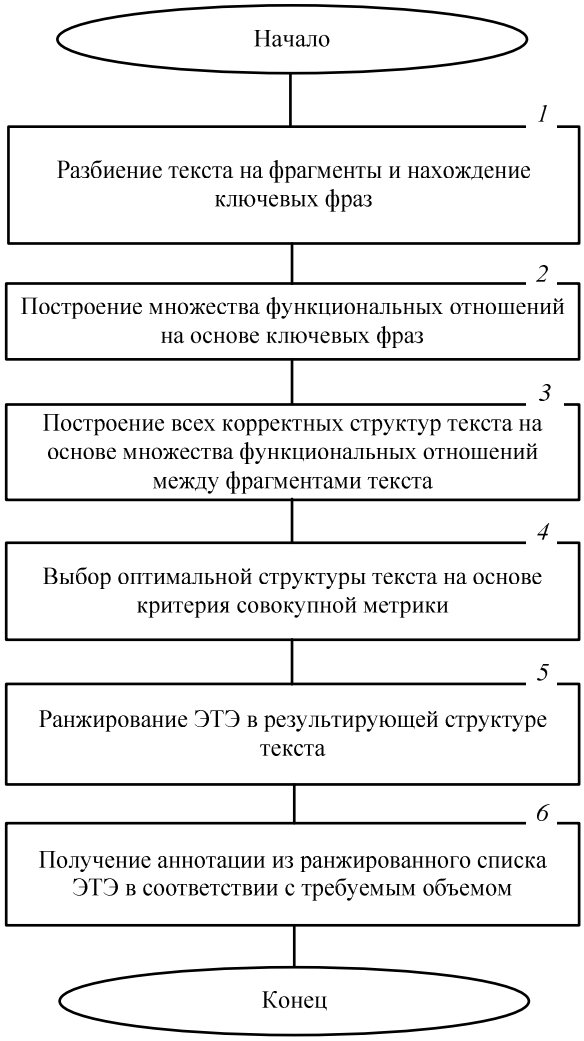

Рис. 1

На основе анализа корпуса научно-технических текстов на русском языке (информационно-справочная система, основанная на собрании текстов в электронной форме) был разработан узкоспециализированный словарь ключевых фраз русского языка. Он имеет сложную структуру – с каждой фразой связан список возможных функциональных отношений, позиция этой фразы в предложении (в начале, в середине или в конце), типы фрагментов текста, которые она обычно связывает (часть предложения, предложение или параграф), расстояние между связанными элементами, измеряемое в ЭТЭ, статусы связанных элементов (ядро или сателлит).

Примерами ключевых фраз служат: «если», «вследствие», «но», «наоборот». С ключевой фразой «вследствие» могут быть связаны такие функциональные отношения, как «предпосылка» и «условие», значениями статусов связанных элементов могут быть как ядро, так и сателлит.

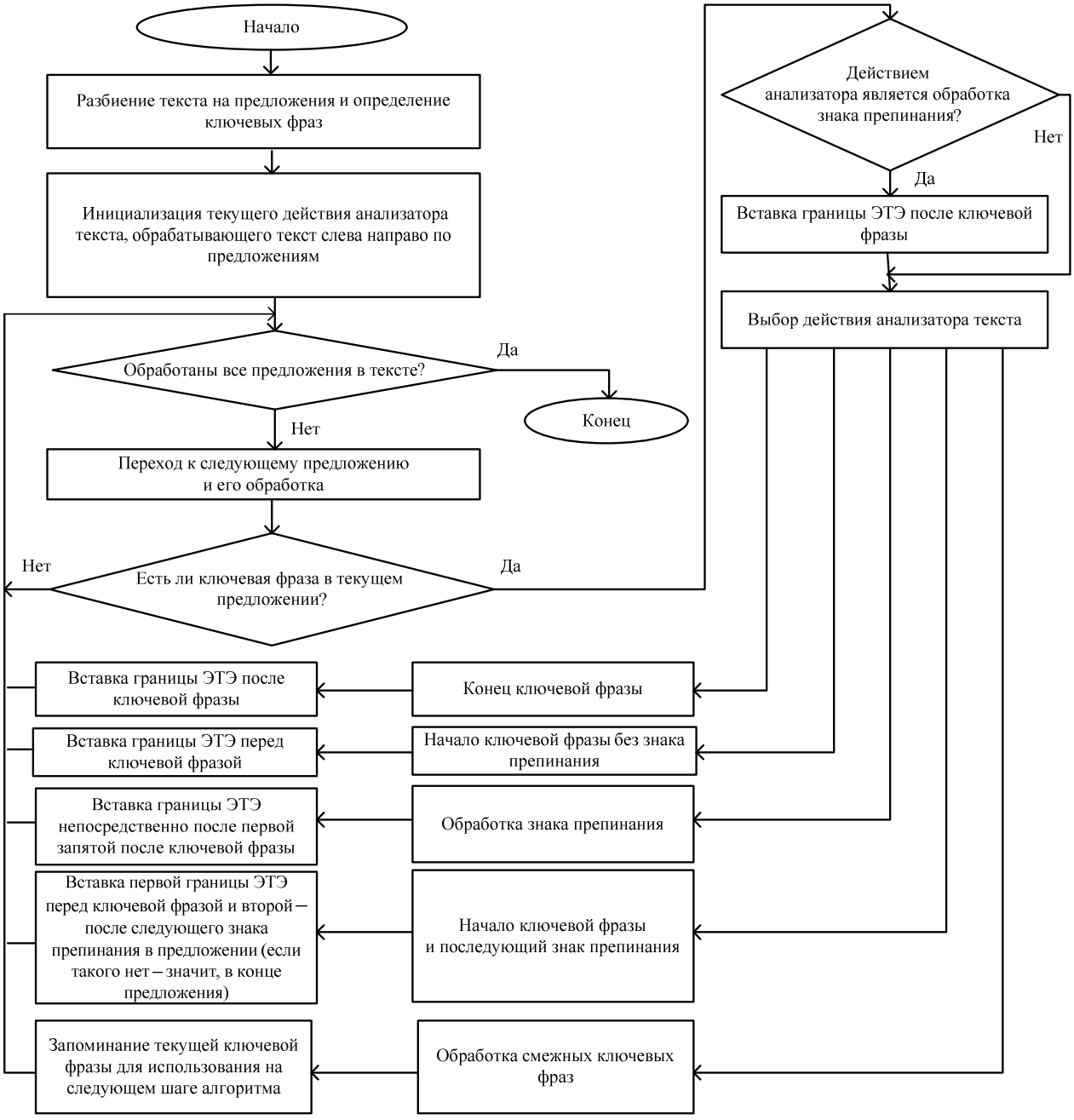

Рис. 2

С помощью словаря ключевых фраз текст разбивается на фрагменты, определяются границы ЭТЭ и построение функциональных отношений между ними.

Алгоритм определения границ ЭТЭ с помощью указанного словаря представлен на рис. 2. Входными данными для алгоритма служат предложение и связанный с ним набор потенциальных ключевых фраз, а выходными – то же предложение, разбитое на ЭТЭ, с установленными реальными ключевыми фразами, которые определяют функциональное отношение.

Алгоритм построения множества функциональных отношений на основе списка ЭТЭ представлен на рис. 3.

На каждом уровне разбиения текста (часть предложения, параграф, предложение) алгоритм определения функциональных отношений пробегает по всем текстовым элементам этого уровня и по всем ключевым фразам, которые им принадлежат.

Для каждой ключевой фразы алгоритм строит множество взаимно исключающих функциональных отношений. Рассматриваются два случая. Первый случай, ключевая фраза «*m*» *i*-го тексто-

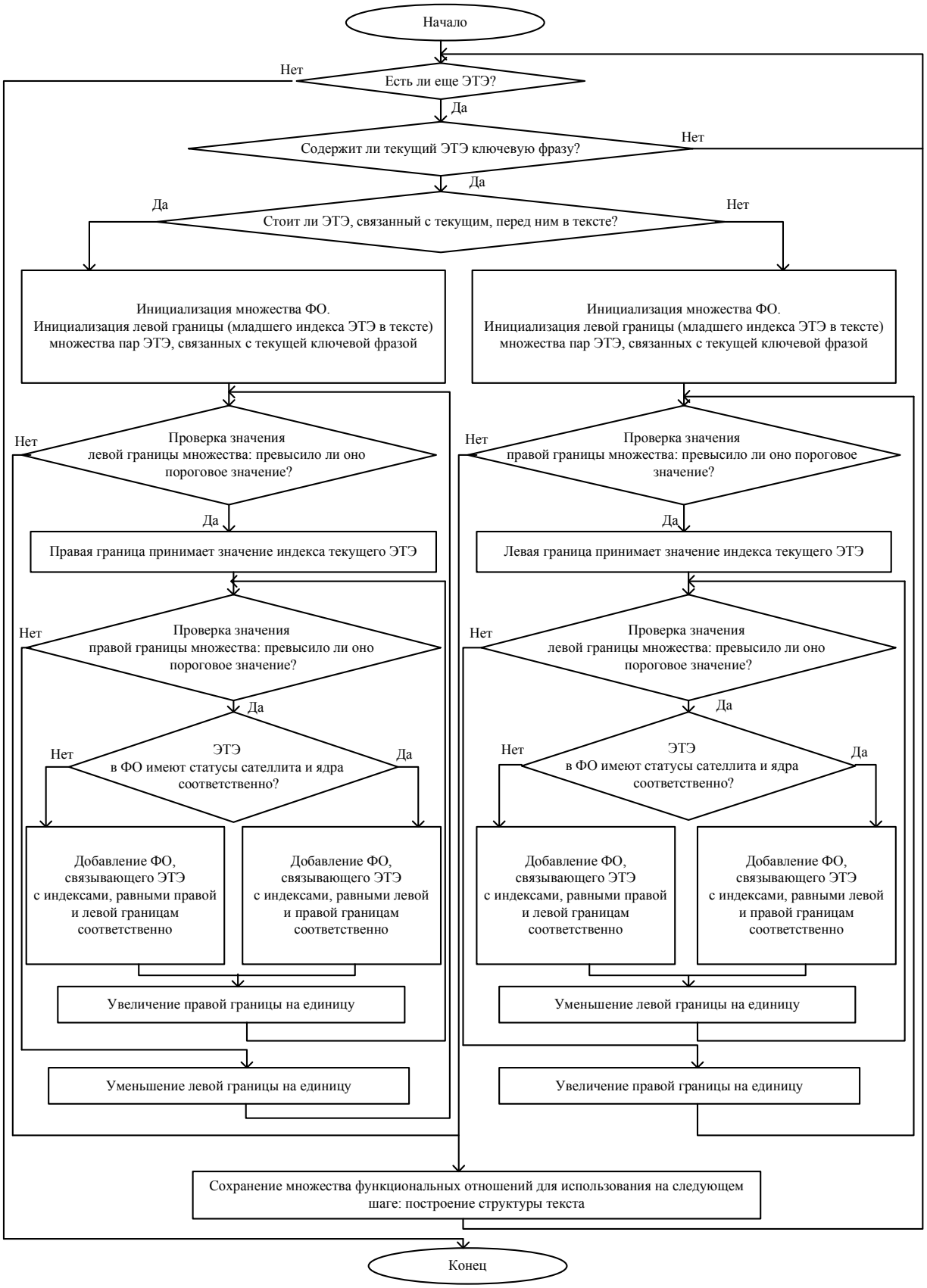
Начало
Нет
Есть ли еще ЭТЭ?
Да
Содержит ли текущий ЭТЭ ключевую фразу?
Нет
Да
Да
Стоит ли ЭТЭ, связанный с текущим, перед ним в тексте?
Нет
Инициализация множества ФО.
Инициализация левой границы (младшего индекса ЭТЭ в тексте) множества пар ЭТЭ, связанных с текущей ключевой фразой
Инициализация множества ФО.
Инициализация левой границы (младшего индекса ЭТЭ в тексте) множества пар ЭТЭ, связанных с текущей ключевой фразой
Нет
Проверка значения левой границы множества: превысило ли оно пороговое значение?
Да
Правая граница принимает значение индекса текущего ЭТЭ
Нет
Проверка значения правой границы множества: превысило ли оно пороговое значение?
Да
Левая граница принимает значение индекса текущего ЭТЭ
Нет
Проверка значения правой границы множества: превысило ли оно пороговое значение?
Да
Нет
ЭТЭ в ФО имеют статусы сателлита и ядра соответственно?
Да
Добавление ФО, связывающего ЭТЭ с индексами, равными правой и левой границам соответственно
Добавление ФО, связывающего ЭТЭ с индексами, равными левой и правой границам соответственно
Увеличение правой границы на единицу
Уменьшение левой границы на единицу
Нет
Проверка значения левой границы множества: превысило ли оно пороговое значение?
Да
Нет
ЭТЭ в ФО имеют статусы сателлита и ядра соответственно?
Да
Добавление ФО, связывающего ЭТЭ с индексами, равными правой и левой границам соответственно
Добавление ФО, связывающего ЭТЭ с индексами, равными левой и правой границам соответственно
Уменьшение левой границы на единицу
Увеличение правой границы на единицу
Сохранение множества функциональных отношений для использования на следующем шаге: построение структуры текста
Конец

Рис. 3

вого элемента определяет функциональное отношение, которое лежит между *i*-м текстовым элементом и текстовым элементом, предшествующим ему. Второй случай, в котором ключевая фраза «*m*» *i*-го текстового элемента определяет функциональное отношение, лежащее между *i*-м текстовым элементом и текстовым элементом, следующим за *i*-м.

Не всегда алгоритм определения функциональных отношений с помощью ключевых фраз выдает на выходе отношения для всех частей предложений. Чтобы определить отношения несвязанных элементов, необходимо выполнить дополнительные действия. Предполагается, что если два предложения «говорят» об одном и том же, то вероятно, что последующее предложение детализирует предыдущее; иначе они относятся к разным темам. Решение по поводу «схожести» вычисляется через количество похожих слов в двух предложениях. Если оно больше некоторого порога, то используется функциональное отношение *ELABORATION* (уточнение), иначе – *JOINT* (соединение).

На следующем этапе сформированный набор функциональных отношений дает возможность перейти к построению структуры текста. Задача построения формулируется следующим образом: дано множество ЭТЭ $U = \{u_1 u_2 ... u_n\}$ и множество функциональных отношений RR, которые лежат между ЭТЭ из U; найти все корректные структуры текста, исходя из множества U.

Параметрами алгоритма построения структуры текста являются:

– множество ЭТЭ $U = \{u_1 u_2 ... u_n\}$;

– множество констант *NUCLEUS*, *SATELLITE*, *LEAF*, *NULL*;

– имена всех функциональных отношений;

– объекты типа $tree(S, T, P, left, right)$.

Объекты, имеющие форму $tree(S, T, P, left, right)$, обеспечивают функциональное представление корректных деревьев. Переменная S может иметь значения *NUCLEUS* или *SATELLITE*; T содержит имя функционального отношения; P представляет собой подмножество элементов из множества U; *left* и *right* могут быть либо *NULL*, либо рекурсивным определением через объект *tree*.

Идея алгоритма построения структуры текста состоит в следующем. Изначально каждый i-й ЭТЭ ассоциирован с элементарным деревом (деревом, состоящим из одного элемента), которое имеет статус либо *NUCLEUS*, либо *SATELLITE*, тип *LEAF* и множество ключевых ЭТЭ-потомков $\{i\}$. Вначале любое отношение из множества функциональных отношений *RR* может быть использовано для связи двух элементов в более сложные деревья. После построения всех элементарных деревьев структура текста формируется соединением смежных деревьев в бóльшие, при условии, что на каждом шаге получается корректная древовидная структура. С каждым таким шагом связано множество функциональных отношений, которые могут быть использованы на следующих шагах. Но как только одно из отношений было использовано, оно становится недоступным для дальнейших преобразований. Этот процесс повторяется рекурсивно до тех пор, пока не будет получена результирующая структура, покрывающая весь текст. При реализации данного алгоритма используется система правил вывода корректных структур текста [4], которые определяют условия объединения двух смежных фрагментов текста в более сложные структуры в различных ситуациях.

Отличительной особенностью алгоритма построения структуры текста является учет неоднозначности функциональных отношений путем генерации альтернативных корректных структур текста. Выбор предпочтительной альтернативы осуществляется по критерию совокупной метрики, вычисляемому на основе линейной комбинации шести индикаторов, используемых на практике для определения информационной значимости структур дерева.

Следующим этапом обобщенного алгоритма является ранжирование по важности ЭТЭ построенного структурного дерева для всего текста. Наиболее простой способ определения важности ЭТЭ – это подсчет веса для каждого ЭТЭ на основе анализа высоты дерева относительно того узла, где впервые встретился данный ЭТЭ во множестве ключевых потомков. Чем больше этот вес, тем важнее ЭТЭ.

Далее из ранжированного списка ЭТЭ выбирается их необходимое количество в соответствии с заданным объемом аннотации.

Для реализации предложенного алгоритма автоматического реферирования научно-технических текстов была разработана программная система на основе объектно-ориентированного подхода в системе программирования Java.

Эффективность разработанного метода и алгоритма автоматического реферирования оценивалась по качеству получаемых аннотаций. Оценка качества аннотаций, получаемых с помощью разработанного алгоритма, проводилась на основе различных процедур с помощью метода экспертных оценок.

Проведенные исследования показали, что качество аннотаций, полученных с помощью разработанного алгоритма, в среднем на 20 % выше по параметрам полноты и точности по сравнению с аннотациями, полученными с помощью традиционных статистических методов для научно-технических текстов на русском языке.

Разработанный алгоритм автоматического реферирования текста использует процедуру автоматического построения структуры текста на основе множества функциональных отношений, что позволяет получать качественные рефераты без использования обширных словарей и баз знаний общего назначения.

СПИСОК ЛИТЕРАТУРЫ

1. Заболеева-Зотова А. В., Камаев В. А. Лингвистическое обеспечение автоматизированных систем. М.: Высш. шк., 2008. 245 с.

2. Mann W., Matthiessen C., Thompson S. A. Rhetorical structure theory and text analysis // Discourse Description. Amsterdam: Benjamins, 1992. С. 39–78.

3. Жарковский А. В., Лямкин А. А., Тревгода С. А. Структурный подход к автоматизации реферирования научно-технических текстов // Программная инженерия. 2013. № 3. С. 33–36.

4. Тревгода С. А., Сабинин О. Ю. Технология автоматического реферирования технического текста // Изв. СПбГЭТУ «ЛЭТИ». 2008. № 7. С. 25–34.

A. V. Zharkovskiy, A. A. Lyamkin, S. A. Trevgoda
Saint-Petersburg state electrotechnical university «LETI»

Algorithm for automatic abstracting scientific and technical texts basing on a structural approach

The algorithm of automatic abstracting scientific and technical texts basing on taking account the text structure has been proposed. The algorithms have been presented for implementing the processes of analysis and construction of the text structure at various stages of abstract preparation.

Automation, algorithm, formal characterization, automatic abstracting, functional relations, text structure