ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

УДК 002.53/55:001.814

А. В. Жарковский, А. А. Лямкин, С. А. Тревгода

АЛГОРИТМИЗАЦИЯ ПРОЦЕССА ОБРАБОТКИ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

Рассматриваются алгоритмы автоматизации процесса обработки научно-технических текстов. Предложен критерий корректности структуры текста, формализованы характеристики и ограничения на корректные структуры. Приведена структура системы автоматизации реферирования научно-технического текста.

Ключевые слова: *алгоритм, формализация, предикат, система автоматизации, реферирование, обработка информации.*

Развитие информационных технологий и значительный рост оборота электронной документации на предприятиях различного уровня выдвигает задачи создания методов обработки большого потока текстовой информации, в частности, методов сжатия текстовых документов — рефератов или аннотаций — с целью минимизации времени на их анализ.

Попытки решения проблемы автоматического реферирования текста предпринимались многими исследователями как за рубежом, так и в России. Многочисленные подходы к ее решению достаточно четко подразделяются на два направления [1]:

1) извлечение из исходного текста всех релевантных предложений (экстракция);

2) генерация реферата с использованием методов искусственного интеллекта (абстракция).

Второе направление связано с развитием методов искусственного интеллекта и в настоящее время представлено экспериментальными исследованиями, но до его широкой реализации еще далеко. Исследования в рамках первого направления в области автоматического реферирования базируются, главным образом, на статистических методах, которые относятся к „поверхностным“, они рассматривают текст как набор линейно упорядоченных слов, словосочетаний и предложений и не учитывают особенностей структуры текста. Как показала практика, различные статистические методы анализа текста недостаточно эффективны. Дальнейшее усложнение их математического аппарата без привлечения параметров структуры текста не позволит заметно повысить качество подобных систем.

В данном случае рассматривается подход к автоматизации обработки научно-технических текстов на русском языке, основанный на учете их структуры. Этот подход базируется на использовании теории риторической структуры текста (ТРС), которая рассматривает текст в виде древовидной структуры, узлы которой связаны между собой функциональными отношениями [2]. Единицами (узлами) структуры являются элементарные текстовые элементы (ЭТЭ), представляющие собой части предложения.

Анализ основных положений ТРС показал, что ее непосредственное применение для построения модели автоматического реферирования текста невозможно из-за нечеткого

формализованного описания структуры текста, критериев отличия корректной структуры от некорректной и алгоритмов построения такой структуры [1].

Научно-техническим текстам присущи логичность, аргументированность, наличие опорных слов и словосочетаний (ключевых фраз), которые являются единицами структурно-смысловой организации фрагментов текста. Это позволяет выделить подмножество функциональных отношений, а также ключевых фраз, которые позволяют провести формализацию описания структуры текста. Такая формализация включает в себя:

— определение критерия корректности структуры текста;

— описание характеристик структуры текста;

— описание ограничений на корректные структуры.

Критерий „корректность структуры текста“ формулируется следующим образом: если функциональное отношение R лежит между двумя элементами текстовой структуры, то оно лежит, по крайней мере, между двумя ключевыми ЭТЭ-потомками этих элементов.

На следующем этапе формализации определяются характеристики структуры текста: статус, тип и множество ключевых потомков, которые связаны с каждым узлом. Статус отражает роль данного узла в функциональном отношении, тип содержит функциональное отношение, соединяющее прямых потомков, множество ключевых узлов содержит потомков, играющих роль „ядра“ в функциональном отношении. Данные характеристики дают достаточное количество информации для описания текстовой структуры и использования этой информации в алгоритмах. Формальное описание характеристик структуры текста для текстового фрагмента $[l, h]$ представлено ниже.

Параметр $S(l, h, status)$ показывает статус фрагмента $[l, h]$, где $status$ — роль элементов функционального отношения — может иметь значения $NUCLEUS(ЯДРО)$, $SATELLITE(САТЕЛЛИТ)$ или $NONE(НЕ\ ОПРЕДЕЛЕН)$, l — левый индекс ЭТЭ, h — правый индекс ЭТЭ.

Тип $T(l, h, relation\_name)$ показывает имя функционального отношения, которое лежит между его прямыми потомками, где $relation\_name$ — функциональное отношение.

Параметр $P(l, h, unit\_name)$ показывает имя ключевого ЭТЭ среди своих прямых потомков, где $unit\_name$ — название или индекс ЭТЭ.

Совокупность ограничений на корректные структуры текста представляется в виде следующих предикатов.

1) Для каждого фрагмента текста $[l, h]$ параметр узла структуры S имеет домен значений $NUCLEUS, SATELLITE, NONE$:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \to \{[l = h \to (S(l, h, NUCLEUS) \vee S(l, h, SATELLITE))]$$

$$\wedge [l \ne h \to (S(l, h, NUCLEUS) \vee S(l, h, SATELLITE) \vee S(l, h, NONE))]\}.$$

2) Статус любого фрагмента текста уникален:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \to [(S(l, h, status_1) \wedge S(l, h, status_2)) \to status_1 = status_2.$$

3) По крайней мере, одно функциональное отношение лежит между двумя смежными фрагментами текста:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \to [(T(l, h, name_1) \wedge T(l, h, name_2)) \to name_1 = name_2].$$

4) Фрагменты текста не накладываются друг на друга:

$$[(1 \le h_1 \le N) \wedge (1 \le l_1 \le h_1) \wedge (1 \le h_2 \le N) \wedge (1 \le l_2 \le h_2) \wedge (l_1 \prec l_2) \wedge (h_1 \prec h_2) \wedge (l_2 \le h_1)]$$

$$\to [\neg S(l_1, h_1, NONE) \to S(l_2, h_2, NONE)].$$

5) Фрагмент текста со статусом $NONE$ не включается в результирующее дерево, описывающее структуру текста:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \rightarrow [(S(l,h,NONE) \wedge P(l,h,NONE) \wedge T(l,h,NONE))$$

$$\rightarrow (\neg S(l,h,NONE) \wedge \neg P(l,h,NONE) \rightarrow \neg T(l,h,NONE))].$$

6) Существует главный фрагмент текста, корень дерева, который покрывает весь текст:
$(\neg S(l,N,NONE) \wedge \neg P(l,N,NONE) \rightarrow \neg T(l,N,NONE)$.

Введенные ограничения необходимы для исключения некорректных структур из алгоритма автоматического построения структуры текста. В соответствии с приведенной формализацией общий алгоритм автоматического реферирования включает в себя четыре основных этапа:

1) определение функциональных отношений в тексте;

2) построение структуры текста на основе набора функциональных отношений;

3) ранжирование листьев (элементов) древовидной структуры текста по важности;

4) формирование аннотации по ранжированному списку ЭТЭ.

Рассмотрим подробнее каждый из этих этапов. Определение функциональных отношений в тексте включает в себя следующие действия:

— разбиение текста на параграфы и предложения и определение ключевых фраз в тексте;

— определение границ ЭТЭ;

— определение функциональных отношений между фрагментами текста (параграфами, предложениями и ЭТЭ);

— определение отношений между фрагментами текста для элементов, еще не связанных функциональным отношением.

Построение структуры текста состоит из следующих шагов:

— построение деревьев для каждого из уровней текста (параграф, предложение, часть предложения);

— нахождение оптимальной структуры дерева для каждого из уровней фрагментов текста;

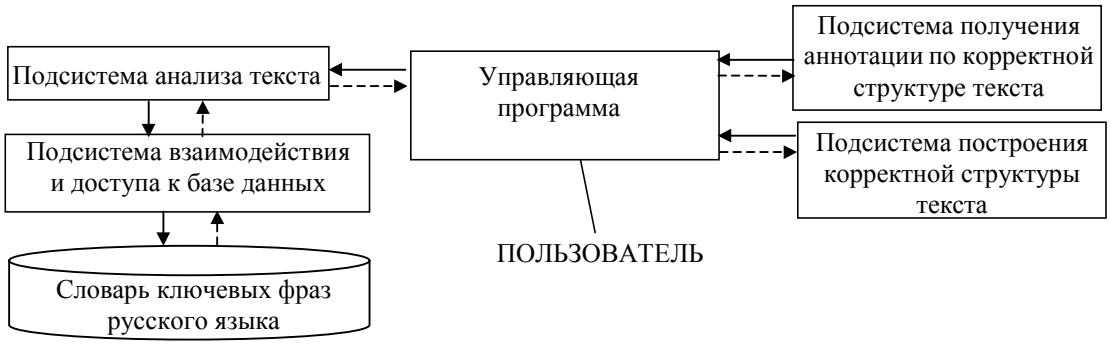
— объединение найденных деревьев в одно общее дерево.

Этап ранжирования единиц структуры текста по важности необходим для выбора самых значимых элементов текста при формировании аннотации требуемого объема.

Наиболее простой способ определения важности ЭТЭ — это подсчет весовых коэффициентов для каждого ЭТЭ на основе анализа высоты дерева относительно того узла, где впервые встретился данный ЭТЭ во множестве ключевых ЭТЭ-потомков. Чем больше значение коэффициента, тем важнее этот ЭТЭ.

На этапе формирования аннотации рассчитывается количество ЭТЭ в соответствии с заданным объемом аннотации и производится выборка этого количества ЭТЭ из начала отсортированного списка.

Структура системы автоматического реферирования, реализующей разработанные алгоритмы, представлена на рисунке.



При разработке программной системы, реализующей предложенный алгоритм, используются общие принципы системного проектирования и объектно-ориентированного программирования. Взаимосвязь подсистем обеспечивает управляющая программа. Процесс начинается с передачи управления подсистеме анализа текста, которая, используя словарь

ключевых фраз, полученный в результате анализа корпуса научно-технических текстов русского языка, разбивает текст на части (параграфы, предложения, части предложений) и определяет функциональные отношения между этими частями.

Далее управление передается подсистеме построения корректной структуры текста, которая по результатам разбиения текста на части строит результирующее дерево, покрывающее весь текст. На заключительном этапе работает подсистема формирования аннотации заданного размера.

Разработанная система автоматического реферирования реализована на языке Java. Словарь ключевых фраз хранится в базе данных MySQL.

Проведенные исследования показали, что качество аннотаций, полученных с помощью системы, реализующей предложенный алгоритм, значительно выше по сравнению с аннотациями, полученными с помощью традиционных статистических методов, при этом система имеет достаточно хорошее быстродействие, что служит основанием для ее практического использования.

СПИСОК ЛИТЕРАТУРЫ

1. *Marcu D.* The theory and practice of discourse parsing and summarization. Cambridge, Massachusetts: The MIT Press, 2000. 245 p.

2. *Inderjeet M.* Automatic summarization (Natural Language Processing). John Benjamins Publishing Company, 2001. 285 p.

Сведения об авторах

Аркадий Викторович Жарковский — канд. техн. наук; Санкт-Петербургский государственный электротехнический университет „ЛЭТИ“, заместитель проректора по научной работе; E-mail: av.jarkov@yandex.ru

Александр Анатольевич Лямкин — канд. техн. наук, доцент; Санкт-Петербургский государственный электротехнический университет „ЛЭТИ“, кафедра систем автоматического управления; E-mail: alex-ljamkin@yandex.ru

Сергей Александрович Тревгода — Санкт-Петербургский государственный электротехнический университет „ЛЭТИ“, кафедра систем автоматического управления, младший научный сотрудник; E-mail: troftu@mail.ru

Рекомендована кафедрой
систем автоматического управления

Поступила в редакцию
09.03.10 г.