УДК 004.021+004.421

ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ ПОНЯТИЙ В ТЕКСТОВОМ СОДЕРЖИМОМ С ИСПОЛЬЗОВАНИЕМ СТАТИСТИЧЕСКОЙ ОЦЕНКИ

Белая Т.И., Пасечник П.А.

[1]*Санкт-Петербургский государственный университет технологии и дизайна, Северо-западный институт печати, Санкт-Петербург, Россия, (191186, Санкт-Петербург, ул. Большая Морская, 18), e-mail: studentszip@yandex.ru*

Проведен анализ проблемы компьютерной обработки русскоязычного текста, нацеленной на выделение ключевых понятий в текстовом содержимом. В качестве объекта рассмотрения выбраны термины, вводимые в текст впервые, а также сопровождающие их определения. Рассмотрены исключительно статистические средства выделения понятий, выделены преимущества над словарными методами. Имеется направленность работы на автоматическое реферирование. Выделены четыре ключевых этапа для решения проблемы, в которых использованы шаблонные конструкции, анализ слов и комбинаций, статистика встречаемости слов в тексте. Выделены формулы для получения вероятностных характеристик терминов и предложений, их определяющих. Сформирован алгоритм проведения анализа текста, приведены рекомендации по использованию данного алгоритма в разработке программных средств.

Ключевые слова: автоматизированная обработка текста, выделение понятий, реферирование

SEPARATION OF KEY CONCEPTS OF TEXT CONTENTS WITH USE OF THE STATISTICAL ASSESSMENT

Belaya T.I.[1], Pasechnik P.A.[1]

[1]*St. Petersburg State University of technology and design, Northwestern University Press, Saint Petersburg, Russia, (St. Petersburg, 191186, St. Bolshaya Morskaya, 18), e-mail: studentszip@yandex.ru*

We have done the analysis of text processing using statistical estimation of clauses or particular terms. Main purpose of this article is describing terms evaluation method without using thesaurus methods. As the object of consideration selected terms introduced in the text for the first time , as well as their accompanying definitions. Considered an exclusively statistical tools allocation concepts highlighted advantages over dictionary methods. There is a focus of the work on automatic summarization . Identified four key steps to solve the problem , which are used in the template design , analysis of words and combinations of words in the statistics of occurrence of the text. Select the formula for the probability characteristics of terms and defining their proposals . Formed algorithm analyzes the text provides guidance on the use of this algorithm in the development of software tools. Evaluated data can be used in automation of educational test formation process, science material coverage estimation, translation of Russian texts, grammatical correcting automation and purposes of artificial intelligence theory.

Keywords: automated text processing, selection of concepts, referencing

Введение

В данной работе рассмотрена проблема компьютерной обработки текста на русском языке. Основной задачей данной работы является автоматизированное формирование массива терминов, опираясь на статистические характеристики текстового содержимого и его ключевых единиц, без использования словарных методов обработки текста, за исключением использования словаря шаблонных конструкций и комбинаций слов, сопровождающих определения.

Полученные данные позволяет решить широкий круг проблем, связанных с анализом текстового содержимого, среди них могут быть выделены:

- формирование автоматизированных систем тестирования;

- оценка научного или образовательного материала, на предмет охвата существующих понятий;
- перевод русскоязычного текста;
- автоматизированная коррекция пунктуационных и смысловых ошибок в русскоязычном тексте;
- оптимизация алгоритмов поисковых систем [3];
- проблема распознавания смысла естественного языка компьютерным оборудованием, как подраздел теории искусственного интеллекта.

Предоставленные данные ориентированы на обработку с помощью императивного процедурного языка программирования, но допускает использования логической и функциональной парадигм программирования.

Актуальность проблемы и существующие методы

Актуальной проблемой анализа текстового содержимого является выделение ключевых понятий. Под ключевыми понятиями понимается наиболее значимые термины рассматриваемого текста, которые отражают его основной смысл [2]. Они формируют общее смысловое содержание, позволяя проанализировать глубину рассмотрения предметной области, а также в автоматическом режиме отнести рассматриваемый текст к определенной предметной области.

Существующие методы обработки текста [7] ориентированы в большей мере на выявление закономерностей между отдельными токенами [6] (словами или словосочетаниями составляющими одну смысловую единицу предложения), а также опираются исключительно на словарный анализ. Словарный анализ требует наличия максимально полных баз данных, содержащих слова, а также их взаимосвязи и свойства. Также метод является высоко требовательным к ресурсам, как хранения, так и пополнения и обработки. Метод ограничен представленными базами данных и имеет свойства, которые напрямую зависит от ее полноты. В данном случае появление нового термина, может быть обработано неверно, поскольку он не присутствует в словаре.

Рассматриваемая задача сходна с задачей реферирования [1], которое подразумевает автоматическое формирование аннотации или реферата к представленному текстовому содержимому. Задача имеет более узкую направленность, не требует формирования связного текста из полученных данных, что является основным отличием от реферирования. Также реферирование является избыточным по отношению к представленной задаче.

Ключевые понятия текста имеют широкую область применения, как в области анализа текста, так и в области его автоматической обработки, перевода, а также автоматической проверки на смысловые ошибки [3]. Понятия, предоставленные в тексте, могут быть как

общеиспользуемые, так и новые, вводимые в рамках рассматриваемого текста. По этой причине процесс анализа наиболее целесообразно разделить на две отдельно выполняемые задачи: анализ текста на предмет общеиспользуемых [5] терминов и анализ текста на предмет терминов, вводимых и определяемых в рассматриваемом тексте, которые, как правило, предоставляют собой ключевые понятия. Обе задачи кардинально различаются по степени сложности. Поиск вводимых терминов также во многих случаях может быть не реализован словарными методами, поскольку термин в тексте может вводиться впервые или может быть новым и отсутствовать в словаре.

Анализ текстового содержимого на предмет вводимых терминов

Рассмотрим анализ текста на предмет терминов, вводимых в рамках рассматриваемого текста. Для выделения терминов может быть использована следующая последовательность:

- анализ пунктограмм, используемых в рассматриваемом тексте, а также использование шаблонных конструкций, сопровождающих определения нового термина;
- обработка текста на предмет слов и комбинаций, сопровождающих определения нового термина;
- сбор статистики встречаемости слов в тексте с отсеиванием заведомо не являющихся терминами, по полученным статистическим данным.

Анализ пунктограмм [8], а также комбинаций пунктограмм и слов, называемых шаблонными конструкциями, позволяют выделить термины, явно определяемые в тексте, а также является вспомогательным средством на этапе анализа частоты встречаемости, позволяя выявлять сложные предложения и анализировать их как отдельную единицу. При обработке языком программирования, данный этап не требует использование статистических методов, он построен исключительно на использовании теоретических сведений и позволяет достичь высокой степени точности. Этап требует наличия базы шаблонных конструкций, полнота которой напрямую влияет на точность полученных данных. Начальные данные базы формируются вручную, а впоследствии пополняется автоматизировано при взаимодействии с пользователем.

Для решения проблем, не затронутых на предыдущем этапе, производится словарная обработка текста, которая также требует использования теории, но дополняется использованием статистики расположения слов и их комбинаций, отсеиваемых на основе грамматических правил. На данном этапе собирается максимально полная база слов и комбинаций слов, сопровождающих определение новых терминов. Из полученной базы выбирается набор слов и комбинаций, имеющих наибольшую вероятность наличия определения при использовании. Затем производится поиск элементов набора в тексте, что позволяет сузить круг поиска. Таким образом, для каждого элемента набора формируется

массив предложений, которые могут содержать определения терминов с определенной вероятностью $P_у$, которая является вероятностью события, согласно которому рассматриваемый элемент набора слов или комбинаций указывает на наличие определения в данном предложении.

Поскольку вводимые определения, согласно существующим требованиям к оформлению научного текста, как правило, присутствуют в начале текстового содержимого, порядковый номер предложения в тексте также играет весомую роль. По этой причине вводится порядковый коэффициент K, который рассчитывается как отношение порядкового номера предложения к числу всех предложений в тексте согласно формуле 1, где i – номер рассматриваемого предложения, а N – это число всех предложений в тексте.

$$K_i = \frac{i}{N} \tag{1}$$

Используя формулу 2, рассчитаем вероятность наличия определения нового термина в рассматриваемом предложении.

$$P_н = P_у \cdot K \tag{2}$$

По полученной вероятности производится сортировка предложений.

Производится сбор статистики встречаемости слов в тексте, то есть производится занесение всех слов текста в один двумерный массив, который содержит анализируемое слово, а также $N_{появл.}$ – количество его появлений. Полученный массив обрабатывается на предмет союзов, предлогов и местоимений, которые затем исключаются. Следующей задачей является поиск элементов массива слов в элементах массива предложений в порядке убывания встречаемости. К каждому слову формируется массив предложений, в которых может быть определено данное слово. Количество появлений слова, а также содержащих его предложений, являются параметрами, определяющими вероятность того, что рассматриваемое слово является термином, поэтому примем ее согласно формуле 3.

$$P_т \approx \frac{1}{N_{появл.} \; N_{предл.}} \tag{3}$$

Производится группировка синонимичных понятий, в результате чего вероятности $P_т$, группируемых понятий, пересчитывается, а также производится исключение понятий, несущих вспомогательный характер. Затем слова и предложения рассматриваются попарно, в случае если слову соответствует более одного предложения, оно рассматривается с каждым

по отдельности, иначе выносится в отдельный массив, а также считается потенциальным определением. Каждой паре формируется вероятность потенциального определения, согласно формуле 4.

$$P_{п.о.} = P_{н} \cdot P_{т} \tag{4}$$

Используя вероятность потенциального определения, выделяются наиболее вероятные пары, которые далее обрабатываются человеком. На данном этапе целесообразно введение автоматизированного средства, обработки результатов. Предложения, одобренные и не принятые пользователями, заносятся в банк знаний, который в дальнейшем обрабатывает их с целью выявления шаблона, который может быть использован на первом этапе. В случае если одно из предложений, соответствующее полученному шаблону, является не принятым, шаблон в данном случае должен быть переработан или исключен из рассмотрения.

При использовании данного алгоритма, также целесообразным является повтор аналогичных действий как внутри подразделов, так и во всем тексте в целом, а затем сравнение полученных результатов, что позволяет повысить их достоверность.

Визуальное представление данного алгоритма в виде блок-схемы предоставлено на рисунке 1. Его использование в большей степени ориентировано на автоматизацию формирования исходного материала для тестирования знаний обучаемого по существующему материалу преподавателя. При использовании методов автоматизации процесса реферирования, с целью автоматизации формирования тестов, возникают общепринятые понятия рассматриваемой области знаний, что является избыточным.

Использование словарных методов также является нецелесообразным, поскольку они не являются достаточными ввиду разнородности используемых терминов, которые зависят от преподаваемой дисциплины, а также имеют низкое быстродействие. В рассматриваемом методе решаются данные недостатки за счет вероятностного характера алгоритма и исключения общепринятых понятий из рассмотрения.

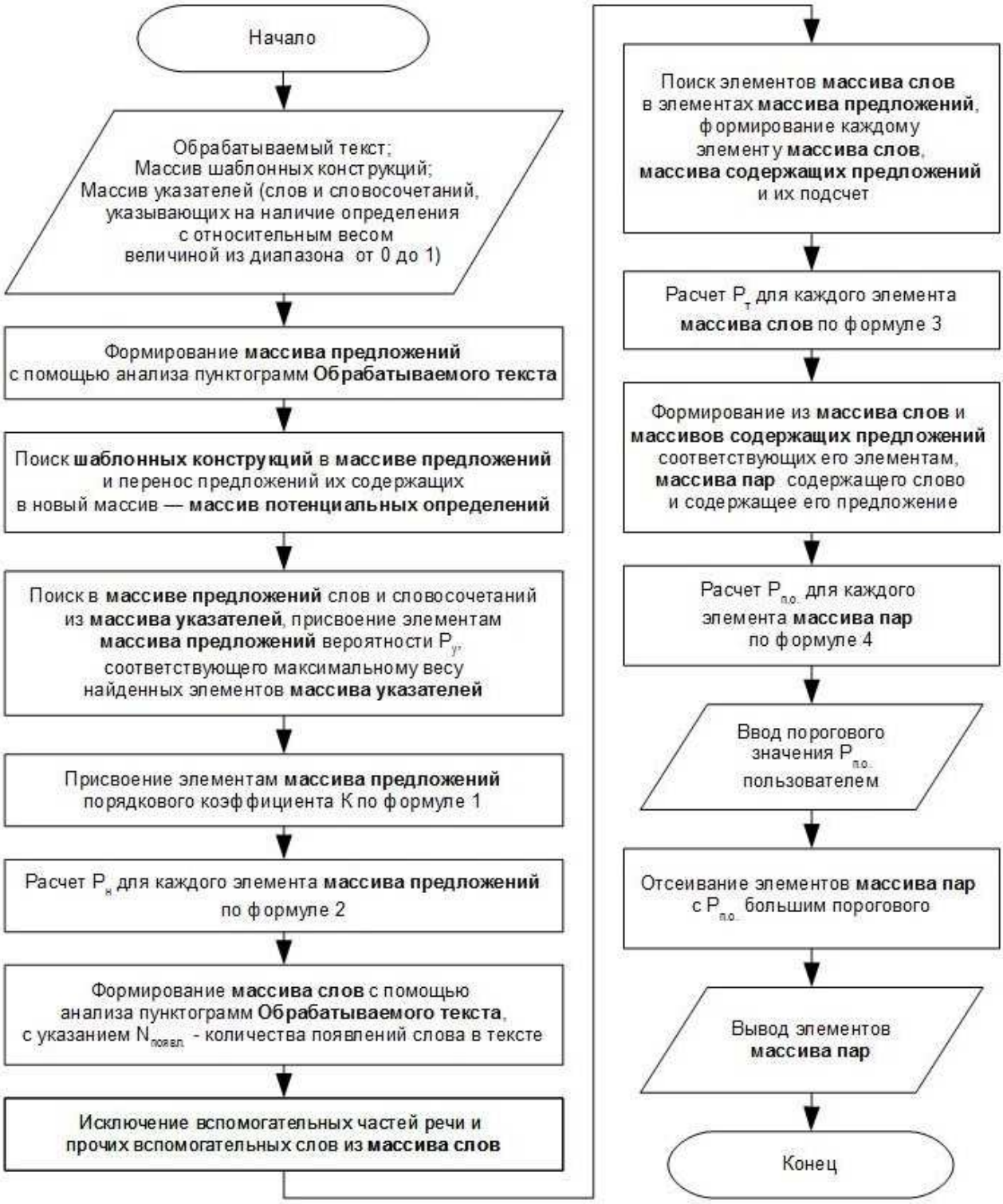

Рисунок 1 – Блок-схема алгоритма

В программном изделии для автоматизации формирования тестов взаимодействие с пользователем, при использовании данного алгоритма, является необходимым, поскольку в выходных данных алгоритма могут присутствовать ложные элементы, которые должны быть исключены. При одобрении элементов пользователем, они подвергаются обработке, с целью выявления шаблонных конструкций и слов-указателей, которые могут быть использованы

при повторном использовании приложения, что позволяет оптимизировать последующие результаты. Наличие данного взаимодействия позволяет решить следующие проблемы:

- недостаточность базы шаблонных конструкций, а также неточность некоторых шаблонных конструкций, которая приводит к появлению ложных элементов и отсутствию истинных;
- недостаточность базы слов и их комбинаций, которые сопровождают определения;
- выявление дополнительных вероятностных взаимосвязей, позволяющих повысить точность обработки.

Алгоритм может быть реализован с использованием языка высокого уровня, имеющего функции или библиотеки обработки текста. Также при его использовании достигается более высокая производительность по отношению к словарным методам.

Заключение

В данной работе ключевые понятия рассматриваются как отдельный класс распознаваемых элементов текста, написанного на естественном языке. Данная работа имеет большую направленность к области реферативной обработки текста, которая заключается в выявлении набора ключевых высказываний, описывающих содержимое текста. Большинство методов, используемых для реферирования, опираются на словарные методы, которые требуют наличия словарей, имеющих набор существующих терминов. Предложенный метод имеет исключительно вероятностный характер.

Список литературы

1. Абрамов В.Е. Автоматическое рубрицирование и реферирование текстовой информации (в том числе на иностранных языках) : автореф. дис. на соиск. учен. степ. канд. техн. наук. – М., 2008. – 27 с.
2. Горошкин А.Н., Обработка и распознавание рукописного текста в системах электронного документооборота : автореф. дис. на соиск. учен. степ. канд. техн. наук. – Красноярск, 2008. – 21 с.
3. Крищенко В.А., Программное обеспечение для метапоиска информации в гипертекстовой среде : автореф.дис. на соиск. учен. степ. . канд. техн. наук. – М., 2002. – 16 с.
4. Вишняков Р. Ю. Разработка и исследование формализованных представлений и семантических схем предложений текстов научно-технического стиля для повышения эффективности информационного поиска : автореф. дис. на соиск. учен. степ. канд. техн. наук. – Таганрог, 2012. – 18 с.

5. Суркова А.С. Разработка структурно-статистических методов и алгоритмов идентификации текста : автореф. дис. на соиск. учен. степ. канд. техн. наук спец. – Н. Новгород, 2004. – 19 с.
6. Кадомцев В.И. Распознавание коммуникативной функции составляющих текста (письменной речи) : автореф. дис. на соиск. учен. степени канд. психол. наук. – М., 1975. – 25 с.;
7. Файн В.С., Распознавание образов и машинное понимание естественного языка /Отв. ред. И.Т. Турбович; АН СССР, Ин-т пробл. передачи информ. – М.: Наука, 1987. – 172 с.
8. Шоломов Д.Л. Синтаксические методы контекстной обработки в задачах распознавания текста : дис. на соиск. учен. степ. канд. техн. наук. – М., 2007. – 24 с.

Рецензенты:

Колбанев М.О., д.т.н., профессор СПбГУСЭ, кафедра «Прикладные информационные технологии», г. Санкт-Петербург.

Татарникова Т.М., д.т.н., доцент, профессор, Институт информационных систем и защиты информации СПбГУАП, г. Санкт-Петербург.