Прикладные проблемы оптимизации

УДК 004.4 DOI 10.25559/SITITO.2017.3.486

Калегин С.Н.

Московский научно-исследовательский телевизионный институт (ЗАО «МНИТИ»), г. Москва, Россия

ОПТИМИЗАЦИЯ РЕСУРСОЁМКОСТИ ЯЗЫКОВЫХ ОПРЕДЕЛИТЕЛЕЙ ЗА СЧЁТ СПОСОБОВ И АЛГОРИТМОВ ИДЕНТИФИКАЦИИ

Аннотация

Данная статья является продолжением серии публикаций автора по тематике языковой идентификации текстов. Рассматривается возможность оптимизации ресурсоёмкости программ и систем для определения языковой принадлежности информационных блоков за счёт модификации идентификационных и результатирующих алгоритмов, а также подбора способов языковой идентификации. Это позволяет существенно повысить их эффективность и произвести расчёт требований к ресурсам на стадии проектирования подобных программных решений, что значительно сокращает время их разработки и отладки. Также приводятся: классификация способов языковой идентификации, их сравнительная таблица, блок-схемы и градации соответствующих алгоритмов.

Работа будет интересна специалистам в области компьютерной лингвистики и разработчикам автоматизированных комплексов обработки неструктурированных данных, как например: системы глобального мониторинга, информационно-поисковые системы, каталогизаторы литературы, системы автоматического реферирования документов, автоматизированные переводчики текстов и т.д.

Ключевые слова

Языковая идентификация текста; определение языковой принадлежности текста; система языковой идентификации; языковой определитель.

Kalegin S.N.

Moscow research institute of television (CJSC MNITI), Moscow, Russia

RESOURCE VOLUME OPTIMIZATION OF LANGUAGE IDENTIFIERS FOR THE ACCOUNT OF METHODS AND IDENTIFICATION ALGORITHMS

Abstract

This article is a continuation of the author's series of publications on the topic of text's language identification. Here is considered the possibility of optimizing the resource capacity of programs and systems for language identification of information blocks by modifying the identification and resultant algorithms, as well as the selection of language identification methods. This allows significantly increase their efficiency and to calculate the requirements for resources at the design stage of such software solutions, which significantly reduces the time of their development and debugging. Also given are: classification of language identification methods, their comparative table, block diagrams and gradations of the corresponding algorithms.

The work will be of interest to specialists in the field of computer linguistics and developers of automated complexes for processing unstructured data, such as: global monitoring systems, information retrieval systems, literature cataloguer, automatic document abstracting systems, automated text translators, etc.

Keywords

Language identification of the text; language identification system; language identifier.

Введение

С развитием коммуникаций, компьютерных сетей и международного общения всё чаще стали использоваться автоматические системы, позволяющие обрабатывать информацию на различных языках, а также выдавать отчёты по результатам анализа множества разноязычных источников (Интернет-сайтов, форумов, социальных сетей, радиовещательных станций и т.д.), что повлекло за собой развитие соответствующего программного обеспечения и аппаратно-программных комплексов. Сегодня получили широкое распространение многоязычные информационно-поисковые системы, базы знаний, системы глобального тематического мониторинга, автоматизированные системы реферирования документов, аналитические системы искусственного интеллекта [1] и т.д. Вследствие совершенствования подобных систем возникла необходимость оптимизации каждого модуля, включая языковые определители, которые являются основой анализаторов входящего информационного потока, так как без определения языка, использованного при фиксации данной информации, её семантический анализ практически невозможен. С целью уменьшения ресурсоёмкости и повышения эффективности работы языковых определителей, а также сокращения времени на отладку подобных программ, автором проведено исследование по оптимизации их алгоритмов на стадии проектирования, результаты которого приводятся в этой статье. Под эффективностью здесь понимается оптимальное сочетание качества результатов языковой идентификации и ресурсоёмкости данного процесса. При этом основными ресурсами считаются: время и вычислительная мощность компьютера (или комплекса), а дополнительными: базы данных, словари, языковые модели и т.д.

Эффективность современных технологий языковой идентификации

Качество и ресурсоёмкость современных технологий, программ и систем языковой идентификации зависят от двух основных факторов: способов языковой идентификации и алгоритмов их применения [2, 3]. Если эти факторы не учитывать на этапе проектирования языкового определителя, то результат окажется практически непредсказуемым, так как техническая реализация такой программы или системы несущественно влияет на качество конечного продукта. Например, выбор языка программирования или среды разработки может определить набор инструментов, скорость разработки и состав программистов, а также повлиять на объём (вес) и структуру конечной программы или системы, но её эффективность определяется только способами и алгоритмами идентификации [4], что наглядно показано далее.

Проблема выбора способов идентификации

В качестве идентификационной базы разработчики, обычно, выбирают наиболее известные способы языковой идентификации на основе символьного или лексического принципа (Рисунок 1), что сразу ограничивает сферу применения конечного продукта и определяет его особенности, так как все способы имеют свои недостатки.

Как видно из представленной блок-схемы, каждый способ требует использования конкретных идентификационных маркеров, с помощью которых производится определение языка информационного блока.

Таблица 1. Обзорная таблица способов языковой идентификации и результатов их сравнения

№ п/п	Способ	Идентификационные маркеры	Преимущества	Недостатки
1	Словарный	слова в исходной форме	простота реализации	низкая эффективность и повышенная ресурсоёмкость
2	Символьный	характерные символы традиционной письменности	простота реализации, низкая ресурсоёмкость	низкая эффективность и высокая вероятность ошибок
3	Сравнение n-грамм-моделей	характерные сочетания символов или фонем (n-граммы)	высокая эффективность при определённых условиях	сложность реализации, повышенная ресурсоёмкость и вероятностный результат
4	Поиск коротких слов	характерные короткие слова (до 4-5 символов)	простота реализации, низкая ресурсоёмкость	низкая эффективность и высокая вероятность ошибок
5	Грамматический анализ	характерные аффиксы, словоформы и грамматические особенности	высокая эффективность	требуется подготовительный этап и грамматические анализаторы для каждого языка, сложность реализации и высокая ресурсоёмкость

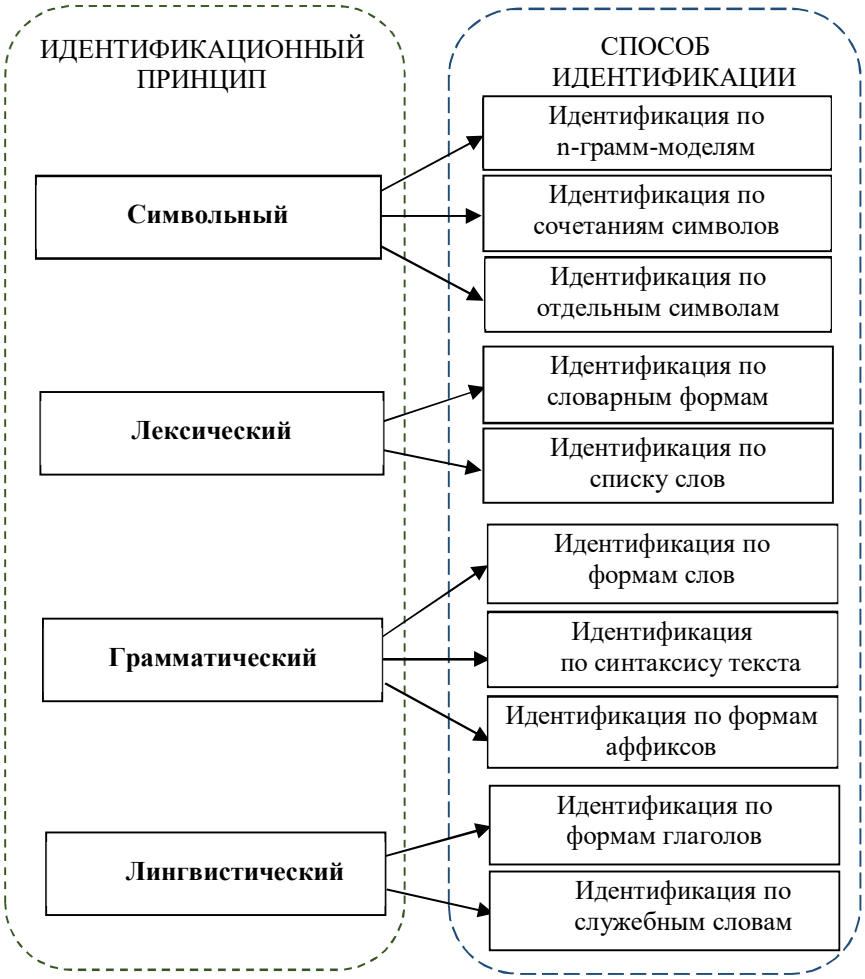


Рис. 1. Классификация способов языковой идентификации текста

В качестве таких маркеров могут выступать символы, слова или их сочетания, которые определитель пытается обнаружить в анализируемом тексте и по их наличию, количеству или положению ассоциирует его с определённым языком.

Однако применение любого из подобных способов идентификации будет давать заведомо вероятностный результат с различной точностью и требовать конкретных условий работы (Таблица 1). К примеру, словарный способ требует обязательной загрузки словарей всех определяемых языков в оперативную память компьютера, относительно высокой производительности центрального процессора и наличия словарных форм в тексте.

Из таблицы следует, что универсальных способов языковой идентификации не существует, а значит при их выборе требуется учитывать множество факторов, включая условия использования конечного продукта.

Для компенсации указанных недостатков часто создаются гибридные системы, в основу которых входит сразу несколько способов идентификации, что позволяет частично решить данную проблему [5], однако при этом количество ошибок увеличивается, вся система многократно усложняется, а её ресурсоёмкость повышается, снижая общую эффективность. В таблице 2 показана градация различных языковых определителей, где тип определителя указывает на количество используемых способов идентификации и их приоритетность. Монометодные программы и системы основаны только на одном способе идентификации, в монометодных+ используются вспомогательные способы, позволяющие уточнить результаты работы основного, а в гибридных системах используемые способы равнозначны.

Для упрощения восприятия и наглядности в данную таблицу не были включены уточнения гибридизации программ, разновидностей идентифицирующих элементов и т.д., которые могут быть получены из описания каждого способа идентификации и его конкретной реализации.

Таблица 2. Градация программ и систем языковой идентификации

№ п/п	Название	Тип	Идент. принципы	Идент. элементы	Ресурсо-ёмкость	Доп. ресурсы
1	Guesser	монометодный +	лексический символьный	словарные формы, символы	высокая	словари, доп. наборы
2	PHPLang-autodetect	монометодный	символьный	символы	низкая	идент. набор
3	МОЯТ	монометодный	лингвисти-ческий	глаголы и отглагольные образования	низкая	идент. набор
4	Eureka Eng.	гибридный	символьный, словарный, грамматич.	символы, слова, морфемы	высокая	словари, идент. набор, морфолог. анализаторы
5	Xerox Lang. Ident.	гибридный	лексический символьный	короткие слова, n-граммы	высокая	идент. набор, языковые модели
6	Auto.Lang. Identifier (T-Labs)	монометодный	n-граммный	n-граммы	средняя	языковые модели
7	Полиглот 3000	монометодный	n-граммный	n-граммы	средняя	языковые модели
8	Lang. Ident. of H.Falck	монометодный	n-граммный	n-граммы	средняя	языковые модели
	SILC	монометодный	n-граммный	n-граммы	средняя	языковые модели

Но и представленных данных достаточно для иллюстрации критериев, влияющих на эффективность конечного продукта, которые должны учитываться на стадии его проектирования в зависимости от условий предполагаемого применения, так как универсального решения не существует.

Проблема выбора алгоритмов идентификации

На ряду со рассмотренными способами на эффективность языкового определителя не меньше влияют идентификационные и результатирующие алгоритмы, которыми определяются не только количество и последовательность применяемых способов, но и качество результата идентификации. Так, выбор между полученными промежуточными данными при использовании нескольких способов является решающим, поскольку результатирующий алгоритм будет определять верность или неверность сделанных ассоциаций с различными языками или же вычислять вероятность их соответствия действительности, а от этого зависит качество работы всей программы, системы или комплекса.

Идентификационные алгоритмы можно разделить на линейные и циклические (Рисунок 2), монометодные и полиметодные (Рисунок 3а), а также на последовательные и параллельные (Рисунок 3б). Линейным называется алгоритм, осуществляющий идентификацию за один проход по набору идентифицируемых элементов до первого совпадения с идентифицирующим маркером, а циклическим называют процесс многократных проверок таких элементов; монометодный от полиметодного отличается количеством используемых способов идентификации – от одного до нескольких; последовательные и параллельные алгоритмы задают очерёдность процессов обработки данных в зависимости от задач и возможностей аппаратного обеспечения.

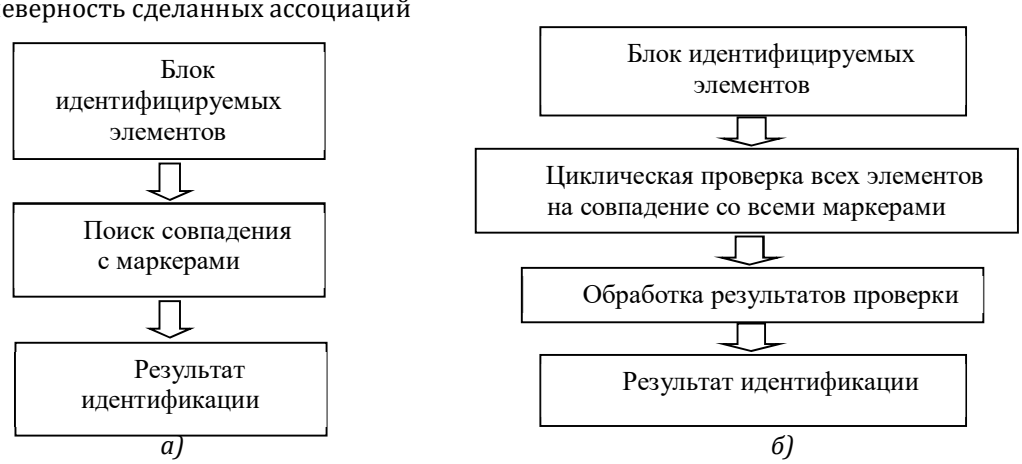
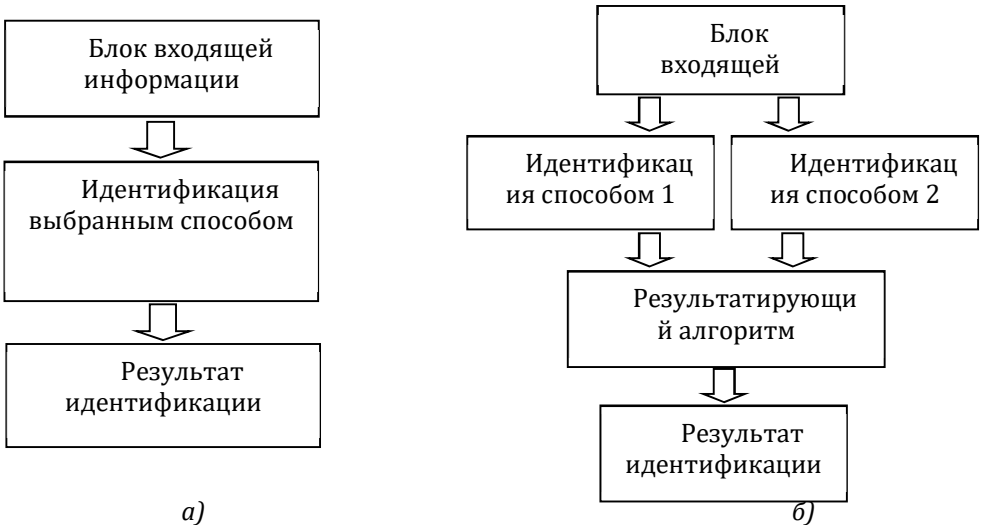


Рис. 2. Блок-схема линейного (а) и циклического (б) идентификационных алгоритмов



а) *б)*

Рис. 3. Блок-схема монометодного (а) и полиметодного (б) идентификационных алгоритмов

Однако на практике часто совместно используются различные типы алгоритмов для повышения эффективности определителя, что позволяет определять такие алгоритмы как гибридные.

Тип алгоритма выбирается в зависимости от способов идентификации, степени гибридизации языкового определителя и с учётом условий применения конечного продукта. Например, для тестовой или учебной программы достаточно линейного алгоритма и одного способа идентификации, наиболее соответствующего заданным условиям, а для создания системы глобального мониторинга потребуется тщательный выбор способов идентификации и полиметодный алгоритм.

При использовании нескольких способов языковой идентификации следует тщательно разрабатывать результатирующий алгоритм, от которого будет зависеть качество работы определителя и результаты всей системы в целом. На сегодняшний день не существует совершенных результатирующих алгоритмов для решения задач языковой идентификации, поэтому каждый разработчик вынужден создавать свой вариант и многократно его проверять. Для примера ниже приведён вероятностный результат языковой идентификации текста на латышском языке системой Gesser, где даётся несколько вариантов соответствия, выбор из которых должен сделать результатирующий алгоритм.

Tāpēc viņš no jauna izņēma no kastes tējas katliņu, kuram bija pazudusi āpša galva un aste, un kurš bija atguvis savu agrāko izskatu, un parādīja to skārdniekam. Apskatījis tējas katliņu, skārdnieks par to piedāvāja divdesmit vara monētas.

Язык текста: латышский

Языковая группа: балтийская

Guesser: венгерский (zs) 32 %, эстонский 8 %, нидерландский 7 % и т.д.

Выданный результат практически бесполезен без доработки, так как язык не определён.

Заключение

По результатам проведенного исследования можно заключить, что сегодня не существует совершенных способов языковой идентификации и алгоритмов их реализации, что не позволяет разработать соответствующую технологию для создания универсальных языковых определителей. Следовательно, при проектировании подобных программ и систем требуется исследовать каждый способ идентификации отдельно, не взирая на его популярность, и определить алгоритмы их реализации с учётом выявленных недостатков и условий применения конечного продукта.

Благодарности

Статья написана по материалам диссертационной работы автора в области языковой идентификации текста.

Научный руководитель: Зубарев Юрий Борисович, доктор технических наук, профессор, член-корреспондент РАН.

Научный консультант: Фархадов Маис Паша оглы, доктор технических наук, старший научный сотрудник.

Литература

1. Калегин С.Н. Распознавание языка неструктурированной информации в системах глобального мониторинга // Известия ЮЗГУ. Серия Управление, вычислительная техника, информатика. Медицинское приборостроение. Том 7, № 2 (23). – Курск, 2017. – С. 20-27.
2. Калегин С.Н. Способы определения языка текста // Филологические науки. Вопросы теории и практики. № 12 (54): в 4-х ч.

Ч. II. Тамбов: Грамота, 2015. – С. 84-89.
3. Калегин С.Н. Важность выбора основного идентификационного принципа при проектировании языковых определителей // Современные информационные технологии и ИТ-образование. Том 12, № 2. – Москва, 2016. – С. 194-204.
4. Калегин С.Н. Расчёт ресурсоёмкости программ языковой идентификации текста // Современные информационные технологии и ИТ-образование. Том 13, № 1. – Москва, 2017. – С. 59-73.
5. Калегин С.Н. Классификация программ языковой идентификации текста // Динамика сложных систем — XXI век. № 4. – Москва: Радиотехника, 2016. – С. 27-33.

References

1. Kalegin S.N. Raspoznavanie yazyka nestrukturirovannoj informatsii v sistemakh global'nogo monitoringa // Izvestiya YUZGU. Seriya Upravlenie, vychislitel'naya tekhnika, informatika. Meditsinskoe priborostroenie. Tom 7, № 2 (23). – Kursk, 2017. – S. 20-27.
2. Kalegin S.N. Sposoby opredeleniya yazyka teksta // Filologicheskie nauki. Voprosy teorii i praktiki. № 12 (54): v 4-kh ch. CH. II. Tambov: Gramota, 2015. – S. 84-89.
3. Kalegin S.N. Vazhnost' vybora osnovnogo identifikatsionnogo printsipa pri proektirovanii yazykovykh opredelitelej // Sovremennye informatsionnye tekhnologii i IT-obrazovanie. Tom 12, № 2. – Moskva, 2016. – S. 194-204.
4. Kalegin S.N. Raschyot resursoyomkosti programm yazykovoj identifikatsii teksta // Sovremennye informatsionnye tekhnologii i IT-obrazovanie. Tom 13, № 1. – Moskva, 2017. – S. 59-73.
5. Kalegin S.N. Klassifikatsiya programm yazykovoj identifikatsii teksta // Dinamika slozhnykh sistem — XXI vek. № 4. – Moskva: Radiotekhnika, 2016. – S. 27-33.

Поступила: 1.10.2017

Об авторе:
Калегин Сергей Николаевич, аспирант, начальник сектора НТО ЗАО «МНИТИ», ksn@mniti.ru

About author:
Sergey Kalegin N., aspirant (applicant), chief of section of CJSC MNITI, ksn@mniti.ru