Большие данные и их приложения

УДК 004.4

Калегин С.Н.

ЗАО «МНИТИ», г. Москва, Россия
Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, Россия

РАСЧЁТ РЕСУРСОЁМКОСТИ ПРОГРАММ ЯЗЫКОВОЙ ИДЕНТИФИКАЦИИ ТЕКСТА

Аннотация

В статье приводится расчёт ресурсоёмкости программных решений для определения языковой принадлежности текста в зависимости от их способов идентификации. Рассмотрены наиболее используемые сегодня способы с указанием их преимуществ и недостатков, но без учёта вариантов их технической реализации. Это позволит разработчикам более осмысленно подходить к проектированию подобного программного обеспечения, а потребителям более осознанно его выбирать.

Ключевые слова

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

Kalegin S.N.

CJSC MNITI, Moscow, Russia
Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia

RESOURCE INTENSITY CALCULATION FOR TEXTS LANGUAGE IDENTIFICATION PROGRAMS

Abstract

The article deals with the resource intensity calculation of texts language identification programs depending on their identification methods. The most commonly used methods are described, together with an indication of their advantages and disadvantages, but without taking into account the technical implementation variants. This gives developers more sensible approach to the design of such software, and consumers will be able to make more informed choice.

Keywords

Text language identification; determining linguistic identity of text; language identification program; detector of text language affiliation.

Введение

В связи с развитием международных отношений и глобализацией информационно-коммуникационных систем, таких как телевидение и компьютерные сети, получили широкое распространение технологии обработки и перевода иноязычных текстов. Для примера можно привести многочисленные системы автоматического и автоматизированного перевода, которые сегодня имеются в различном исполнении для большинства компьютеров, и универсальные translate-системы от таких компаний как Google или Promt, доступные через Интернет любому пользователю, а также различные автоматизированные системы обработки информации, внедрённые сегодня практически на всех уровнях сбора данных и управления. Количество подобных систем и программ машинной обработки информации свидетельствует об их популярности среди пользователей и востребованности на рынке информационных технологий. Такие системы создаются коллективами специалистов в течение нескольких лет и постоянно совершенствуются. Качество их работы зависит от тех принципов, которые были заложены в функциональную основу конкретной системы, алгоритмов взаимодействия составных частей и их

программной реализации. При этом имеет значение всё, начиная с интерфейса взаимодействия с пользователем и заканчивая дополнительными программными модулями, выполняющими вспомогательные функции. Но основные преимущества и недостатки любой программы характеризуются логическими и технологическими принципами, на базе которых были построены её алгоритмы, что наглядно показано ниже.

В данной статье приводится обзор наиболее используемых способов определения языковой принадлежности неструктурированного текста [4, 5] при их программной реализации с целью выявления зависимости ресурсоёмкости программ от особенностей способов идентификации, положенных в их основу. На рисунке 1 приведена блок-схема существующих решений [3].

Ниже приводится краткое изложение сути основных современных способов языковой идентификации неструктурированного текста с указанием их качественных характеристик, а также существенных преимуществ и недостатков, определяющих особенности реализующих их программ. Для проведения тестирования способов идентификации в их программной реализации автором разработана соответствующая методика, в которой определены существенные характеристики, выявляющие ресурсоёмкость и целесообразность применения каждого конкретного способа, а также его результативность. Следуя данной методике, основными отличительными особенностями программы языковой идентификации рационально считать:

- основополагающий идентификационный принцип;
- объём и структура идентификационного набора;
- длина (объём) идентификационного маркера;
- количество операций сравнения;
- требуемый объём оперативной памяти;
- требуемая вычислительная мощность компьютера;
- необходимость использования дополнительных систем, моделей, баз данных, словарей и т.п.;
- вероятность верного результата.

Как видно из приведённого списка, на результат машинного определения языковой принадлежности текста тем или иным способом существенное влияние оказывают как способы идентификации, на базе которых строится логика алгоритма данной программы, так и технические показатели конкретной реализации. Для сравнения были выделены наиболее известные в современном мире способы языковой идентификации на основе:

- использования словарей;
- использования уникальных знаков;
- использования статистики комбинаций символов (n-грамм);
- грамматического анализа;
- поиска коротких слов.

Изложение расчёта их ресурсоёмкости, приведённое ниже, структурировано по основополагающим идентификационным принципам для удобства сравнения их эффективности и технической реализации с учётом рассматриваемых способов и подходов.

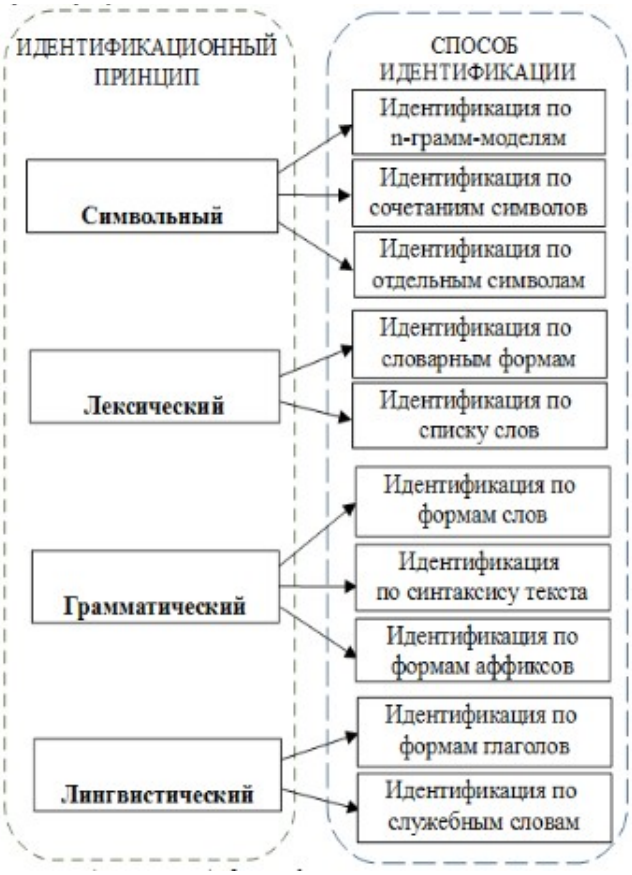

Рис. 1. Классификация способов языковой идентификации текста

Ресурсоёмкость способов языковой идентификации текста

Способы языковой идентификации неструктурированного текста могут быть классифицированы по различным признакам, например, по сложности, эффективности и т.д. В рамках данной статьи наиболее рационально классифицировать их по типу идентифицирующих элементов – идентификационных маркеров, так как это является основным отличием и определяющим фактором при разработке, который задаёт ограничения по использованию и требования к ресурсам. Для наглядности названия способов даны полужирным шрифтом, а их описание приводится в порядке усложнения алгоритмов реализации. Для каждого способа

указаны его расчётная ресурсоёмкость, относительная вероятность верной идентификации и условия использования, а также преимущества и недостатки, позволяющие оценить рациональность его применения в различных ситуациях. В представленном ниже обзоре эти способы приводятся в порядке увеличения сложности их реализации и с учётом распространённости. При этом для каждого способа приводятся расчёты максимального количества операций сравнения (N_{opr}) и объёма оперативной памяти (V_{pam}), требуемого для хранения массива идентифицируемых элементов, идентификационной матрицы (языкового фильтра), служебной информации и т.д. Сопутствующая ресурсоёмкость вспомогательных процессов не рассчитывается, так как не является атрибутом рассматриваемого способа и может меняться в широких пределах по различным причинам или оставаться практически неизменной. Например, ресурсоёмкость считывания в память компьютера анализируемого текста практически не изменяется, но преобразование считанных данных в массив анализируемых элементов в различных способах идентификации и их конкретных реализациях проводится по-разному, что не позволяет вывести общую формулу и произвести расчёты. При существенных дополнительных затратах ресурсов при использовании описываемого способа этот факт выносится в перечень его недостатков.

Общие формулы расчёта ресурсоёмкости для всех способов языковой идентификации имеют следующий вид:

$$N_{opr} = N_t * N_{ids} + N_{dop}, \quad (1)$$

где N_{opr} – максимальное количество операций сравнения;

N_t – количество идентифицируемых элементов анализируемого текста;

N_{ids} – общее количество идентификационных маркеров;

N_{dop} – количество дополнительных (сопутствующих) операций.

$$V_{pam} = (N_{slt} * L_{slt} + N_{idelm} * L_{idelm} + N_{idm} * L_{idm}) * L_{sim} + V_{sd} + V_{mtd}, \quad (2)$$

где V_{pam} – требуемый объём оперативной памяти;

N_{slt} – количество слов в тексте;

L_{slt} – средняя длина слова в тексте;

N_{idelm} – количество идентифицируемых элементов;

L_{idelm} – средняя длина идентифицируемого элемента;

N_{idm} – количество идентификационных маркеров в матрице;

L_{idm} – средняя длина идентификационного маркера;

L_{sim} – средняя длина символов в байтах;

V_{sd} – объём служебных данных;

V_{mtd} – объём метаданных.

Как видно из приведённых формул, при анализе одного текста различными способами некоторые показатели изменяться практически не будут, поэтому далее по тексту они не приводятся. Например, это относится к размерам самого текста ($N_{slt}*L_{slt}$) и средней длине символов в байтах (L_{sim}), а также к количеству сопутствующих операций (N_{dop}). Кроме того, при выполнении любой программы в оперативную память загружается её код, а также используются временные, служебные и дополнительные переменные, обозначенные в формуле (2) как V_{sd}, объём которых зависит от множества различных факторов, абстрактное вычисление которых невозможно. Поэтому в отдельный параметр ресурсоёмкости в данной работе они не выделяются и в дальнейших расчётах не приводятся.

Здесь требуется отметить, что каждый символ обычно соответствует 1-2 байтам в зависимости от кодировки и самих символов, но в разных текстах и идентификационных матрицах могут использоваться различные символы и кодировки, что не позволяет произвести точный расчёт абстрактно и требует введения соответствующего коэффициента. Более того, во многих случаях используются дополнительные служебные символы и записи (метаданные), размер которых зависит от конкретной реализации программы и её рабочей среды. Но, как правило, совокупный размер метаданных относительно мал и в данном обзоре может игнорироваться. Поэтому для упрощения восприятия и чистоты сравнения при расчётах ресурсоёмкости указывается только максимальное количество возможных операций сравнения (сопоставления) идентифицируемых элементов анализируемого текста с идентификационными маркерами, объём требуемой оперативной памяти рассчитывается исходя из общего объёма идентификаторов матрицы, а длина одного символа по умолчанию подразумевается равной 1-2 байтам.

Использование словарей. Этот способ заключается в поиске совпадений слов текста со словами из словарей идентифицируемых языков, в которые обычно включаются наиболее употребительные лексемы, покрывающие более 80 % речи и современной литературы, что по разным оценкам составляет приблизительно (далее используется знак «~») 1500-3000 слов [1, 13, 18]. То есть, идентификационными маркерами являются словарные формы наиболее употребительных слов. Способ относительно простой и легко реализуемый как программными средствами, так и без них. Однако энергетические и временные затраты на поиск каждого слова текста в каждом словаре множества языков делают данный способ практически нецелесообразным из-за большого количества операций сравнения, которое упрощённо может

быть выражено формулой:

$$N_{opr} = N_t * N_{sl} * N_{svr}, \quad (3)$$

где N_{opr} – количество операций;

N_t – количество слов в тексте;

N_{sl} – среднее количество слов в одном словаре;

N_{svr} – количество задействованных словарей.

Таким образом, для автоматического определения языковой принадлежности неструктурированного текста, к примеру, из 100 слов при использовании одного словаря на язык, среднем объёме каждого из них в 2500 слов и 20 идентифицируемых языках получаем следующее выражение:

$$N_{opr} = 100*2500*20=5000000.$$

То есть, при заданных условиях для определения одного из 20 идентифицируемых языков потребуется до 5 миллионов операций сравнения слов и удерживания всех активных словарей в памяти компьютера, иначе потребуется дополнительное время на их считывание с накопителя информации, что существенно замедлит процесс идентификации. При увеличении количества слов текста, языков или словарей число операций будет увеличиваться кратно. Например, при 40 определяемых языках максимальное количество операций сравнения достигнет 10 миллионов, а при увеличении количества слов анализируемого текста до 400 – 20 миллионов.

Кроме того, потребуется определённый объём оперативной памяти, соответствующий общему объёму массива идентифицируемых элементов текста и идентификационной матрицы, что может быть выражено следующей формулой:

$$V_{pam} = (N_{svr} * N_{sl} * L_{sl}) * L_{sim}, \quad (4)$$

где V_{pam} – требуемый объём оперативной памяти;

N_{svr} – количество задействованных словарей;

N_{sl} – среднее количество слов в одном словаре;

L_{sl} – средняя длина идентифицирующего слова в словарях;

L_{sim} – длина символа в байтах.

Перед проведением расчётов по этой формуле требуется уточнить понятие «средней длины слова», которая обычно вычисляется по словарю и для различных языков может составлять от 5 до 12 символов [7, 9, 20]. Но результаты специального исследования, проведённого в рамках диссертационной работы автора, показывают, что средняя длина слова в текстах на европейских языках составляет 5-6 символов. С учётом этих данных средняя длина слова при дальнейших расчётах берётся равной 7 символам. Также требуется учитывать объём символов в 1-2 байта или вводить усреднённый коэффициент для анализируемых текстов. То есть, при тех же входных данных по формуле (4) получаются следующие выражения:

V_{pam} = (20*2500*7)*1 = 350000 (для символов в 1 байт)

или

V_{pam} = (20*2500*7)*2 = 700000 (для символов в 2 байта).

По результатам расчётов видно, что для хранения словарей и массива анализируемых элементов текста при данных условиях может быть задействовано более 700000 байт или ~ 700 килобайт оперативной памяти, причём большую часть занимают словари. Для современных вычислительных систем этот объём не является критичным. Однако увеличение количества идентифицируемых языков или объёма словарей на порядок потребует до ~7 мегабайт оперативной памяти, что может существенно повлиять на быстродействие некоторых программ, например, запускаемых удалённо на web-сервере, где для работы каждого скрипта выделяется строго определённое количество памяти (обычно несколько мегабайт). А с учётом ресурсоёмкости других сопутствующих операций и особенностей алгоритма определителя языка, при использовании web-технологий приведённые требования к количеству оперативной памяти могут стать критичными для его работоспособности.

К тому же при простом сравнении слов текста со словарными формами их грамматические вариации не учитываются, а значит, вероятность обнаружения совпадений слов из словарей со словами одного языка резко понижается (особенно в небольших текстах), что приводит к ошибкам идентификации. Более того, при использовании данного способа обычно не учитываются короткие слова длиной в 2-4 символа (что требует отдельного рассмотрения), которые могут полностью совпадать в разных языках, в том числе неродственных. Это касается, например, артиклей, предлогов и т.п. А вычленение этих слов из текста требует дополнительных ресурсозатрат и в несколько раз уменьшает количество идентифицирующих слов (идентификационных маркеров), что отрицательно сказывается на результате. Таким образом, уточнённая формула вычисления количества операций сравнения (3) приобретает следующий вид:

$$N_{opr} = N_t * (N_{sl} * N_{svr} - N_{ks}), \quad (5)$$

где N_{ks} – количество коротких слов в тексте.

Кроме того, количество слов в словарной форме существенно меньше их количества в тексте по причине морфологических изменений (склонений и спряжений), вызванных грамматическими требованиями согласования. Ниже приведён пример анализа неструктурированных текстов данным способом (словарные формы выделены полужирным шрифтом, короткие слова не учитываются).

Представитель *Минобороны ДНР Эдуард Басурин проинформировал, что* ***сегодня*** *была совершена очередная* ***попытка*** *прорыва обороны*

ополчения республики на линии соприкосновения в пригороде Донецка. По словам представителя Минобороны, силовики ВСУ пытаются ***реализовать*** *свои предыдущие планы по разъединению* ***стратегически*** *значимых населенных пунктов ДНР. (Из новостной ленты)*

Нова была красивой планетой, первой успешной земной колонией. ***Сейчас*** *это* ***пустыня****. Целые города исчезли с её лица, уничтоженные взрывами нейтронных бомб. Нечего* ***опознать****. Нечего* ***похоронить****. Некого* ***оплакать****.* ***Вторжение*** *началось* ***внезапно****. Объединённые силы Земли нанесли коварные удары по всей территории планеты.* ***После*** *них остались растерзанные тела. Крики женщин и детей о помощи захлебнулись в плазменном огне, прожегшем их* ***плоть****. Мы слышали* ***много*** *проповедей о добродетелях прогресса и науки. Что хорошего от них, если целые цивилизации разрушаются в* ***мгновение*** *ока? (Вступление из компьютерной игры "Power DOLLS")*

Как видно из примеров, в первом тексте из 40 слов только 5 стоят в словарной форме и имеют длину более 4 символов, что составляет 12,5 % от общего числа, а во втором содержится около 14 % словарных форм из 78 слов, то есть в среднем каждое седьмое слово, остальные формы при таком подходе в идентификации не участвуют, что отрицательно сказывается на качестве результатов применения данного способа, так как пропускаются целые предложения. Для повышения процентного содержания идентификационных маркеров путём приведения всех слов текста к словарным формам потребуется дополнительный морфологический анализ каждого слова, что на порядки увеличит общее количество операций и сделает данный процесс ещё более ресурсоёмким и, как следствие, не всегда целесообразным. Например, в русском языке имеется более 350 аффиксов и флексий, а также возможность изменения корня слова [6]. Это увеличивает указанные выше объём оперативной памяти и количество операций на несколько порядков, так как для приведения слов текста к словарной форме потребуется сделать предположение о языковой принадлежности текста, подобрать соответствующую грамматическую или морфологическую модель, разобрать каждое слово по составу, выделить корень и сформировать на его основе новое слово в нужной форме. К тому же для проведения подобного анализа требуется иметь отдельный грамматический анализатор или модель для каждого определяемого языка, а для большинства языков их на данный момент не существует.

Резюмируя результаты вычислений и учитывая специфику данного способа, его можно использовать только при небольшом количестве заранее известных неродственных языков и достаточном количестве анализируемых элементов. При автоматическом определении языка неструктурированного текста данным способом в условиях жёсткого ограничения потребляемых ресурсов (например, с помощью скриптов на web-сервере) потребуется увеличение лимита ресурсоёмкости, а процесс идентификации может занять много времени, что в большинстве случаев сделает её нецелесообразной.

Использование уникальных знаков. Данный способ заключается в отличии языков друг от друга по особенным буквам или знакам традиционной письменной системы в частности, по буквам с диакритическими знаками (диакритиками), которые обычно используются при записи текстов на данном языке. Диакритика бывает над гласной (например, в буквах «й» или «ё»), над согласной (буква «č») или может как-то иначе сопровождать букву или буквосочетание. Кроме того, во многих алфавитах дополнительно используются специальные символы для обозначения фонем данного языка. Например, в польском алфавите есть знак, напоминающий перечёркнутую букву «L» (Ł), а в украинском – буква, похожая на русское «э», но повёрнутая в другую сторону (є). Более того, некоторые языки имеют собственную письменную систему, как например, японский или корейский. Это наталкивает многих специалистов на ассоциацию конкретной письменности с определённым языком. Так, в «Определителе языков мира по письменностям» Гиляревского и Гривнина [2] собрано около 200 языков в ассоциации с их письменными системами, по особенностям которых авторы предлагают определять язык нужного текста. Однако это не всегда приводит к ожидаемому результату по разным причинам. Например, если определять русский язык по наличию в тексте буквы «ё», то многие технические тексты, где данная буква практически не используется, не будут идентифицированы, а тексты на других языках (германских или тюркских), записанных кириллицей и имеющих подобную фонему, как раз будут отнесены к русскому. Также язык может быть определён по наличию в тексте специальных непроизносимых букв или знаков. В русском алфавите это твёрдый (ъ) и мягкий (ь) знаки. То есть для идентификации языка текста данным способом достаточно выделить из традиционной для конкретного языка письменности несколько характерных символов (обычно 4-6 штук), например, для русского могут быть взяты **ё**, **й**, **ъ** и **ь**. Это позволит существенно сократить затраты ресурсов.

Как следует из описания, идентификационными маркерами при этом способе определения языка являются символы, а значит, количество сравнений будет равно количеству символов текста кратному количеству идентификационных маркеров, что может быть

выражено следующей формулой:

$$N_{opr} = N_{sim} * N_{ids}, \quad (6)$$

где N_{opr} – количество операций сравнения;

N_{sim} – количество символов в тексте;

N_{ids} – общее количество идентификационных маркеров.

Следовательно, исходя из описания данного способа и количества идентифицирующих язык символов в 4-6 штук, то есть в среднем ~ 5, получается следующая модификация формулы (6):

$$N_{opr} = N_{sim} * 5 * N_{iaz}, \quad (7)$$

где N_{iaz} – количество определяемых языков.

Таким образом, если в анализируемом тексте содержится 100 слов, средняя длина которых ~ 7 символов, то при 20 определяемых языках получаем:

$$N_{opr} = (100*7)*5*20 = 70000.$$

Из приведённых результатов видно, что при проведении идентификации языка текста данным способом требуется примерно в 71 раз меньше операций сравнения, чем при использовании словарей. Кроме того, сами идентификационные маркеры занимают существенно меньше места в памяти компьютера, что также позволяет экономить ресурсы. Так, при среднем количестве идентифицирующих символов по 5 на каждый язык и объёме символа в 1-2 байта, получаем следующие модификации приведённого выше выражения:

$$V_{pam} = 20*5*1 = 100$$

или

$$V_{pam} = 20*5*2 = 200.$$

Как следует из расчётов, в данном случае вся идентификационная матрица займёт от 100 до 200 байт в оперативной памяти компьютера, что практически не повлияет на быстродействие и не является критичным для современных ЭВМ.

Эти факторы указывают на явное техническое преимущество данного способа по сравнению с предыдущим, однако его эффективность характеризуется не только количественным выражением затрачиваемых ресурсов, но и результативностью идентификации, о чём будет сказано ниже.

Далее приведён пример определения языка текста с помощью поиска характерных символов национальных алфавитов или оригинальных письменных систем при идентификации текстов на русском (славянская группа) и кумыкском (тюркская группа) языках по 4 маркерам: **ё**, **й**, **ъ** и **ь**, характерным для русской письменности (идентификационные маркеры выделены полужирным шрифтом, слова подчёркнуты).

Русский язык:

*В Курске возбудили дело против депутата областно**й** думы Ольги Ли, которая разместила в интернете видеоролик с обращением к президенту Владимиру Путину. В этих выступлениях следователи нашли возбуждение ненависти к власти….*

*Представитель Минобороны ДНР Эдуард Басурин проинформировал, что сегодня была совершена очередная попытка прорыва обороны ополчения республики на линии соприкосновения в пригороде Донецка. По словам представителя Минобороны, силовики ВСУ пытаются реализовать**ь** свои предыдущие планы по разъединению стратегически значимых населенных пунктов ДНР. (Отрывки из новостной ленты)*

*Перед началом работы с данным устро**й**ством внимательно изучите следующие меры безопасности во избежание получения травм и повреждения изделия Nikon. Сохраните инструкции по технике безопасности в месте, доступном всем поль**з**ователям данного устройства для дальнейшего ознакомления (Из инструкции по пользованию фотоаппаратом Nikon D3200).*

*…Наша способност**ь** усваивать необычное не беспредельна, а когда путешествуешь на другие планеты, пределы оказываются очен**ь** узкими. Слишком много новых впечатлений; их приток становится невыносимым, и мозг ищет отдыха в буферном процессе аналогизирования. Этот процесс как бы созда**ё**т мост между воспринятым известным и неприемлемым неизвестным, облекает невыносимое неизвестное в желанную мантию привычного…. (Роберт Шекли Обмен разумов)*

Кумыкский язык:

*…Сюув юрек яллатг**ъ**ансан*
*Мени о**й**лягъа салг**ъ**ансан*
*Г**ё**злеримни пашмалык**ъ**,*
*Г**ё**з яшларым ак**ъ**тырг**ъ**ансан*
*Сюювг**ъ**е инандырдип ,*
*Сюювну алма**й**дип,*
Сенсиз яшап болмай эдим,
Сюедим сени, сюедим

*А**й**тдим мен сагъа, сю**й**геним,*
*Унутма мени, тиле**й**мен,*
Унутдунгму сен, аявлум,
*Башхагъа г**ё**зюнг г**ъ**ара**й**му?*
Англамадынг сен мени,
Мени сен англамадынг.
*Мен а**й**тагъан с**ё**злеге бурулуп сен г**ъ**арамадынг ….*

(Отрывок из песни «Сюув юрек яллатгъансан, сюйгеним»)

Как видно из данных отрывков, текст на кумыкском языке содержит больше идентификационных маркеров, чем первые четыре, так как из 1170 символов в текстах на русском языке всего 7 являются определяющими, что составляет ~ 0,6 % от общего количества знаков. В противоположность этому, в тексте на кумыкском языке из 320 символов 22 являются идентификационными маркерами, что составляет более 6 % от общего числа знаков, а это на порядок

больше, чем в текстах на русском. То есть статистически, по результатам сравнения приведённых текстов, с русским языком должен ассоциироваться последний (на кумыкском), что абсолютно неприемлемо.

Подобное сравнение с аналогичным результатом может быть проведено для текстов на других языках, например: немецком, турецком и финском, характерным признаком которых являются буквы «**ä**» «**ö**» и «**ü**».

Таким образом, основным недостатком данного способа является смешение понятий языка и письменности, что приводит к принципиальным ошибкам идентификации. Как следует из определений понятий «письменность» и «язык», практически все письменности подходят для фиксации информации, выражаемой в словах и синтагмах, а значит, их можно использовать для записи текста на любом языке. Например, по-русски можно писать кириллицей, латиницей, глаголицей, руницей, еврейскими, арабскими или греческими буквами, а также слоговым письмом, идеограммами или иероглифами, и в любом направлении, что не приведёт к каким-либо искажениям передаваемой информации. Это подтверждается историческим развитием письменности на Руси [14] и многочисленными примерами народов, которые переходили с одной письменной системы на другую за короткое время, что никак не отражалось на их языке. Для примера можно взять языки Средней Азии (казахский, узбекский, киргизский и т.п.) и Кавказа (адыгейский, аварский, азербайджанский и т.п.), на которых писали разными письменностями в различные исторические периоды [8], а сейчас они используют модифицированную кириллицу или латиницу. Более того, тексты на некоторых языках могут быть записаны несколькими письменными системами без особой разницы. Например, на языках бывшей Югославии (хорватском, словенском и т.п.) и сейчас пишут либо латиницей, либо кириллицей, а на вьетнамском языке чаще всего пишут латинскими буквами с диакритиками, хотя существует собственная вьетнамская письменность, созданная на основе китайских иероглифов, которые когда-то также использовались для записи текстов на вьетнамском языке. Похожая ситуация наблюдается в Японии – для записи текстов на японском языке могут использоваться иероглифы или слоговое письмо типа кана (хирагана или катакана) [19]. Однако иероглифы могут служить для записи текстов на абсолютно любом языке, что подтверждается наличием подобных знаков в древности у различных народов и их заимствованием друг у друга. Так, например, китайские иероглифы заимствовались для записи текстов на японском, корейском, вьетнамском и других языках Юго-Восточной Азии, а значит, определить язык текста по этим знакам без дополнительного анализа практически невозможно, то есть вероятность верного определения языковой принадлежности текста описанным способом невелика вследствие использования одной письменной системы или алфавита для фиксации информации на множестве различных языков, что продемонстрировано примерами выше и подтверждается историческими фактами распространения одной письменности для фиксации информации на языках различных неродственных народов, например, латиницы в Европе или арабского письма в Азии и Африке.

Несмотря на неоспоримые недостатки, подобным способом язык часто определяется в системах распознавания текста. Например, в описании «Способа автоматического определения языка распознаваемого текста при многоязычном распознавании» [10] приводится следующее:

… Известные способы распознавания текста с бумажного или другого жесткого носителя, например микрофиш, перевод изображения в графический файл, разбивку графического файла на области (блоки), предположительно содержащие признаки изображения символов текста, с последующим сопоставлением изображения в блоках с эталонным изображением, в нескольких специальных признаковых (или растровых) классификаторах, содержащих символы одного определенного языка…

и далее в том же документе:

… Вместо нескольких отдельных классификаторов иногда используют единственный, содержащий признаки символов всех языков, предположительно присутствующих в документе.

Такой способ представлен, например, в патенте США 6370269 April 9, 2002. [12]

Из приведённых цитат следует, что авторы подобных способов неразрывно связывают распознаваемые символы (буквы, слоговые знаки или иероглифы) с конкретными языками, что, с точки зрения лингвистики, в корне ошибочно по определению и в большинстве случаев не соответствует действительности. На это всё чаще обращают внимание различные специалисты [16], предлагая использовать лингвистические методы при анализе текста и речи. Однако подобные способы языковой идентификации широко используются в современных системах обработки неструктурированного текста и применяются при проектировании новых программных решений, что вынуждает рассматривать их в качестве возможных вариантов при проектировании программ и систем языковой идентификации текста, а также рассчитывать их ресурсоёмкость.

Использование статистики комбинаций символов (байтовых последовательностей [11]

или n-грамм). Данный способ обычно заключается в определении языка по количеству типичных для него комбинаций символов или фонем – n-грамм, характерных для каждого языка в соответствии со специальной моделью. Чаще всего подсчитываются комбинации двух (диграммы) или трёх (триграммы) символов или фонем, хотя могут встречаться и другие варианты n-грамм. Таким образом, анализируемый текст ассоциируется с тем языком, n-грамм-модели которого с наибольшей вероятностью соответствует большинство найденных в тексте комбинаций (или байтовых последовательностей). Однако при такой идентификации не учитывается возможность использования нетрадиционной письменности или направления письма и не производится транскрипция или транслитерация, а это значит, что уже на стадии формирования n-грамм может быть выбран неверный метод деления текста, что отрицательно отражается на результате.

Данный способ определения языковой принадлежности текста имеет множество вариаций и подходов, но их суть сводится к идентификации языка и кодировки документа, основываясь на статистических показателях текстов, для которых язык и кодировка известны заранее. В них подсчитываются частоты n-грамм длиной не более **n** и предполагается, что примерно 300 самых часто используемых n-грамм зависят от языка [23]. На основе выделенных комбинаций символов создаются языковые идентификационные модели, с которыми впоследствии будет производиться сравнение. Один из вариантов выделенных n-грамм для некоторых европейских языков показан в таблице 1 (приводится в сокращении) [25].

Таблица 1

Шведский	Английский	Немецкий	Французский	Итальянский
en_	_th	en_	_de	_di
_._	he_	er_	es_	to_
er_	the	_de	de_	_de
et_	_,_	der	ent	di_
tt_	nd_	ie_	nt_	_co
_de	ed_	ich	_le	la_
ar_	_an	sch	e_d	re_
-,-	and	ein	le_	ion
fr	_._	che	ion	ent
om_	_to	die	s_d	e_d
_oc	ing	ch_	e_l	le_
ch_	to_	den	_la	o_d
de_	ng_	nd_	la_	ne_
och	er_	_di	re_	no_
an_	_of	ung	on_	_in

Алгоритм определения языка данным способом заключается в сравнении частот n-грамм для наборов тестовых документов (корпусов текстов), язык которых известен, и каждого документа, язык которого требуется определить. После этого, среди всех тестовых документов находится тот, для которого отличие его n-граммной статистики и статистики тестируемого документа минимально. Следовательно, языком анализируемого документа считается язык соответствующего тестового документа или заранее созданной идентификационной модели (графа). Вариант данного алгоритма реализован, например, в поисковой машине MnogoSearch [30].

Ресурсоёмкость данного процесса идентификации для современных компьютеров невелика в силу относительно небольшого количества элементов сравнения и их длины (или объёма). Основными недостатками описанного алгоритма являются необходимость предварительного набора статистики встречаемости n-грамм в большом количестве текстов на каждом языке и непредсказуемая вероятность идентификации малых текстовых фрагментов – отдельные предложения в текстах на похожих языках распознаются неуверенно и часто с ошибками. Кроме того, требуется дополнительный ресурс в виде базы языковых моделей.

Иногда дополнительно применяется более простой подход. Например, в ABBYY RME [21], до четвертой версии, морфологическая машина при словарной лемматизации перебирала все загруженные языковые словари, что сближало её метод идентификации языка со словарным, который описан выше. Улучшающий результаты процесс морфологической обработки текста является сравнительно ресурсоёмким, полные сборники словоформ могут занимать в памяти от единиц до десятков мегабайт на один язык, поэтому подбор идентификационных моделей даже для 20 определяемых языков требует существенных ресурсозатрат.

Кроме того, определить язык документа при данном подходе за один шаг очень сложно по причине вероятности ошибки идентификации родственных языков. Для решения этой проблемы обычно рекомендуется производить определение языка для каждого предложения отдельно. Поскольку предложения представляют собой некие законченные мысли, то, за исключением некоторых малораспространённых случаев, их можно считать самыми крупными моноязычными единицами текста. Следовательно, когда в предложении на каком-то языке встречаются отдельные слова на другом, их можно считать инвариантами для основного языка.

Несмотря на относительно результативную работу данного алгоритма, его оригинальных характеристик часто бывает недостаточно в случае, когда обрабатывается текст, содержащий части на близкородственных языках, например, на русском и украинском [17]. Для решения этой

проблемы алгоритм можно сделать двухпроходным, что позволит проводить процедуру идентификации в 2 этапа (Рисунок 2).

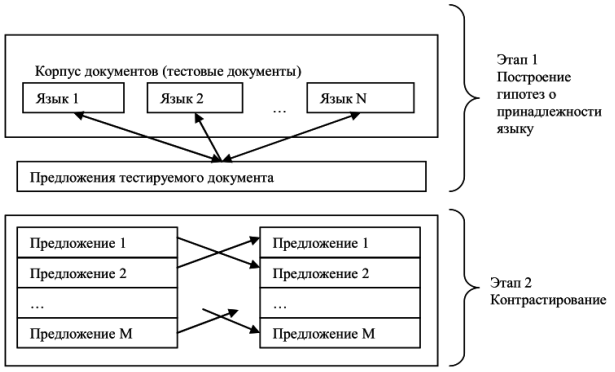

Рис. 2. Двухпроходной алгоритм идентификации языка

При первом проходе выдвигаются гипотезы об ассоциации частей текста с каким-либо языком (или языками) в случае уверенного определения, либо определяется только один язык с весовым коэффициентом не ниже определённого порога, либо самый вероятный язык опережает следующего претендента не менее, чем в 2 раза. Оставшиеся предложения подвергаются обработке при втором проходе – так называемой процедуре контрастирования. Она заключается в том, что языком предложения становится неуверенно распознанный язык, если окружающие его предложения уже сопоставлены с тем же языком [34]. По результатам обоих проходов весь текст ассоциируется с самым вероятным языком. Однако такое усложнение алгоритма не гарантирует верной языковой идентификации, а только повышает её вероятность и существенно увеличивает ресурсоёмкость языковой идентификации за счёт дополнительных этапов и обработки предположений.

Как следует из описания, основными недостатками данного способа являются его абстрактность и вероятностный результат, так как никакого лексического или грамматического анализа текста не производится, варианты письменности игнорируются, а в завершении процесса определения выдаётся список различных языков, к которым можно было бы отнести данный текст с некоторой вероятностью. Более того, для приемлемой работоспособности данного способа требуется набрать определённую статистику встречаемости комбинаций символов или фонем (n-грамм) в различных языках, а для этого нужно проанализировать десятки или сотни тысяч текстов и создать большую базу данных, что само по себе достаточно ресурсоёмко. В силу указанных причин этот способ редко используется лингвистами, однако часто применяется математиками и программистами, так как не требует специальных знаний в области языковедения.

Эффективность данного способа при идентификации языка неструктурированного текста варьируется в широких пределах – обычно от 50 до 100 % [15] – и зависит от многих факторов, среди которых можно выделить следующие:

- количество определяемых языков;
- генеалогическое родство определяемых языков;
- используемая письменная система анализируемого текста;
- используемая кодировка текста;
- количество слов в анализируемом тексте;
- применяемые вариант и алгоритм идентификации;
- качество языковых моделей в базе определителя;
- размер и качество статистической базы данных.

Например, при 2 неродственных определяемых языках и относительно длинном тексте, записанном традиционной для данного языка письменностью в кодировке Unicode, вероятность определения может достигать 100 %. Однако при увеличении количества языков и смешении письменных систем результативность данного способа резко понижается и может доходить до бесполезных результатов, когда вероятность ассоциации текста с различными языками становится практически одинаковой, но ресурсоёмкость при этом многократно увеличивается.

Резюмируя сказанное выше, можно сделать вывод, что данный способ языковой идентификации является одним из самых сложных и ресурсоёмких.

Грамматический анализ текста

Суть этого способа заключается в морфологическом разборе слов и синтаксическом анализе предложений [33]. Его идея основана на тактике анализа текста человеком при попытке идентификации языка (вкупе с лексическим сопоставлением), что придаёт данному способу определённую логичность. Однако автоматически определить язык текста данным способом сразу невозможно, так как неизвестно какая модель должна быть выбрана для идентификации и какие словари или наборы аффиксов загружены. Поэтому требуется отдельный подготовительный этап, на котором выдвигаются гипотезы об ассоциации анализируемого текста с определённым языком или языковой группой по характерным признакам (письменности, лексике и т.п.), после чего делается попытка применить соответствующие языковые модели для окончательного выбора языка. То есть идентификационный процесс должен быть как минимум двухпроходным, что естественно его усложнит и многократно увеличит ресурсоёмкость. Для проведения подобного

анализа требуются специальные лингвистические модели и морфологические словари (или анализаторы) для каждого определяемого языка, а для большинства языков, как было сказано выше, их не существует. Также требуется совершить множество действий с каждым словом анализируемого текста, что в сумме составит миллионы операций даже для небольшого абзаца, на реализацию которых потребуется затратить нецелесообразное количество ресурсов. Об этом свидетельствуют логика процесса и элементарные расчёты, которые приведены ниже.

Самый простой вид морфологического словаря представляет собой таблицу, в первой колонке которой записывается словоформа, во второй – нормальная форма, в третьей – набор параметров [28]. Для морфологической идентификации слов текста необходимо в данном словаре найти все соответствующие анализируемым словам формы и выдать их в словарном виде. Если предположить, что средняя длина слова составляет 7 символов (что обычно равно 7-14 байтам), среднее количество параметров – 4 и для обозначения параметров использовать 8 байт, то с указанием части речи одна запись потребует минимум 69 байт, что демонстрируется следующими выражениями:

$$V_{рам} = 7*7+4+8+8 = 69$$

или

$$V_{рам} = 7*14+4+8+8 = 118.$$

Если предположить, что для каждой нормальной (словарной) формы имеется в среднем ~8 словоформ, то для морфологического словаря объёмом 2500 слов потребует до 2,3 мегабайта оперативной памяти на каждый определяемый язык, что для 20 языков составит 46 мегабайт. Однако эти оптимистические предположения не всегда достижимы на практике. Так, в русском языке существительные обычно имеют 12 форм, прилагательные – 24, наречия и предлоги – 1-2 формы, но их количество существенно превосходит число существительных, также имеются числительные и местоимения. Следовательно, максимальное количество словоформ может превышать 300, и среднее количество в этом случае составит около 25 форм на парадигму. На практике потребуется также место для индексов и дополнительных параметров, в случае реляционной таблицы потребуется резервировать место не под среднюю длину слова, а под максимальную и т.д. В итоге словарь в 2500 слов (которого не всегда достаточно для анализа текстов широкой тематики) будет содержать в среднем более 62000 форм одного языка, что при заданных условиях может занять ~ 7,3 мегабайта оперативной памяти компьютера, что для тех же 20 языков потребует объёма ~ 147 мегабайт. В результате, с учётом количества определяемых языков, возможных ошибок гипотетических предположений о языковой принадлежности текста и особенностей реализующего алгоритма, общее число операций при данном способе идентификации может составить миллионы, потребление оперативной памяти – десятки и сотни мегабайт, а системные требования к данному процессу оказываются нецелесообразно высокими.

Таким образом, данный способ представляет собой существенно усложнённый вариант словарной идентификации, о которой сказано выше, что неизбежно сопровождается теми же недостатками. А следовательно, несмотря на разумный подход, этот способ может использоваться только в ограниченных условиях для текстов, грамотно записанных традиционной письменной системой, при небольшом количестве определяемых языков и наличии для них всех необходимых материалов (морфологических словарей, грамматических анализаторов и т.д.), а также соответствующих вычислительных ресурсов.

Поиск коротких слов [22]**.** Этот способ редко используется в чистом виде, но регулярно предлагается программистами. Его основная идея заключается в идентификации языка текста по формам характерных служебных слов и частиц, таких как союзы, предлоги, артикли и т.п. или коротких слов длиной до 4-5 символов. Например, при определении английского языка предлагается искать артикль «the» и другие подобные маркеры. Этот способ имеет низкую ресурсоёмкость и эффективен при анализе длинных и грамотно написанных текстов с большим количеством слов, однако практически не применим к коротким и упрощённым текстам, в которых такие слова могут отсутствовать [31] (например, к сообщениям с форумов, электронным письмам и т.д.). Более того, он не учитывает множества совпадений служебных слов в родственных языках и похожих коротких слов, междометий и грамматических форм в других языках. Для примера можно привести романские языки (итальянский, французский, португальский, кастильский, румынский и т.п.), в большинстве из которых встречаются артикль «**la**», союз «**e**», предлог «**en**» и форма «**ma**» или «**mas**», а в большинстве германских языков – «**a**», «**na**», «**ten**», «**do**», «**to**» и т.д. Кроме того, такие формы встречаются в других языках, например, в славянских и кельтских, а также в греческом, эсперанто, идо и т.д.

Ниже приведены наглядные примеры текстов на разных языках, которые позволяют провести визуальное сравнение похожих служебных и коротких слов (совпадающие идентификационные маркеры выделены полужирным шрифтом).

Чешский язык.

*Byli jeden dědeček **a** babička **a** měli malou vnučku. Každý **den** vnučka Evička pomáhala svým prarodičům.*

Evička krmila zvířátka, babička pracovala v kuchyni, děda měl ***na*** *starost pole. Vypěstoval řepu takovou, že* ***se na ni*** *chodili dívat sousedi. Byla veliká, zabírala polovinu pole* ***a*** *stále rostla* ***a*** *rostla. Co dělat? Děda ji musí vykopat. Motyka* ***se*** *mu zlomila. Co teď? Jak* ***tu*** *řepu dostat?*

Zavolal ***na*** *pomoc babičku. Ta vzala dědu v pase,* ***ten*** *chytil* ***ze*** *všech sil silné listy* ***a*** *společnými silami táhli, táhli, ale řepu nevytáhli.(Отрывок из сказки про репку)*

Немецкий язык.

„Wir müssen Briefkastenfirmen und Stiftungen, deren wirtschaftlich Berechtigte anonym bleiben, weltweit verbieten", sagte ***der*** *SPD-Politiker* ***der*** *„Süddeutschen Zeitung". Er sprach von „Geldgier* ***der*** *Superreichen",* ***die*** *sich verbinde „****mit der*** *Gewissenlosigkeit* ***im*** *Banken- und Finanzsektor". Beides zerstöre* ***das*** *Vertrauen* ***in den*** *Rechtsstaat.* ***Es*** *gehe* ***um*** *„organisierte Kriminalität von Banken und Finanzjongleuren",* ***die mit*** *allen Mitteln zu bekämpfen* ***sei****. (Из новостной ленты)*

Голландский язык.

De *taal* ***en het*** *literaire leven* ***in de*** *steden* ***en*** *dorpen van Nederland* ***en*** *Vlaanderen. Wie werd waar geboren? Welke romans spelen* ***in*** *Antwerpen, hoe klonk het Katwijks* ***in de*** *19de eeuw?* ***De*** *Atlas* ***is*** *een snel groeiende informatiebron voor plaatsgebonden verschijnselen* ***in de*** *Nederlandstalige cultuur. Alfabetische overzichten* ***en*** *aanklikbare kaarten wijzen* ***de*** *weg.* ***En wie*** *alleen* ***in de*** *Middeleeuwen* ***of in de*** *Achttiende Eeuw wil rondkijken,* ***kan*** *dat doen aan* ***de*** *hand van deeloverzichten* ***en*** *historisch kaartmateriaal. Daarnaast* ***is er*** *een afdeling Buitengaats, waar alle mogelijke literatuur* ***is te*** *vinden* ***over de*** *relaties van Nederlandstaligen* ***met de*** *rest van* ***de*** *wereld. (Из новостной ленты)*

Кастильский (испанский) язык.

En ***su*** *intervención ante* ***el*** *pleno* ***del*** *Congreso, Iglesias* ***se ha*** *dirigido* ***a*** *Sánchez para subrayar* ***que en*** *España hace falta* ***un*** *gobierno comprometido* ***con el*** *derecho* ***de*** *asilo y* ***con los*** *derechos humanos.*

"Quizás ***eso no sea*** *posible cuando* ***se*** *llega* ***a*** *acuerdos* ***con los que*** *querían negar* ***la*** *tarjeta sanitaria* ***a*** *personas* ***que sólo*** *estaban huyendo* ***de la*** *pobreza, señor Sánchez",* ***ha*** *señalado,* ***en*** *alusión* ***a*** *Ciudadanos y* ***al*** *pacto* ***al que el*** *PSOE* ***ha*** *llegado* ***con este*** *partido.*

Ha *añadido* ***que en ese*** *acuerdo* ***de más de*** *200 propuestas entre socialistas y* ***el*** *partido* ***de*** *Albert Rivera, podían haber incluido alguna referencia* ***la*** *derecho* ***de*** *asilo. (Из новостной ленты)*

Гаэльский (ирландский) язык.

Cuireadh ***tús*** *inniu* ***le*** *héisteacht Ard-Chúirte* ***i*** *dtaca* ***le*** *cúiteamh* ***do*** *bhuachaill ceithre bliana d'aois* ***as*** *Baile Átha Cliath ar bhain drochghortú* ***dá*** *inchinn agus* ***é*** *ag dul faoin scian* ***in*** *Ospidéal* ***na*** *Leanaí* ***i*** *gCromghlinn* ***in*** *2012.*

Cuireadh obráid ar Jude Miley ***as*** *Dún Droma agus* ***é*** *ina naíonán féachaint* ***le*** *faoiseamh* ***a*** *thabhairt* ***dó*** *faoi bhrú* ***a*** *bhí ar* ***a*** *scairt.*

Ní raibh greim ***a*** *cuireadh ina chliabhrach gearrtha mar ba cheart trí dhearmad agus tolladh* ***a*** *chroí* ***dá*** *bharr. Stop* ***a*** *chroí* ***an lá*** *tar éis* ***na*** *hobráide agus rinneadh damáiste mór* ***dá*** *inchinn. (Из новостной ленты)*

Эсперанто.

Bonokaze ***mi*** *esperantistighis* ***en*** *Odessa, nigramara urbo, kie loghis unu* ***el la*** *grandaj sovetiaj esperanto-tradukistoj, Sergeo Rublov. Iufoje* ***en*** *1961* ***mi*** *pro ne-memoras-kio trafis* ***al*** *lia loghejo.* ***La*** *mastro regalis min* ***per*** *ege interesa konversacio pri diversaj esperanto-problemoj kaj pri* ***la*** *problemoj* ***de*** *tradukarto. En* ***la*** *fino li proponis legi kelkajn tradukojn* ***al*** *esperanto* ***el*** *Pushkin-poezio.* ***Mi****, sidante* ***en la*** *mola fotelo, fermis* ***la*** *okulojn por pli bone percepti* ***la*** *shatatan poeton* ***en*** *alia (nerusa) lingvo. Tiam* ***mi*** *jam perfekte posedis* ***la*** *lingvon. Ekfluis* ***la*** *konataj strofoj. Jes, tio estis Pushkin, ruse intima Pushkin. Pushkin, centfoje Pushkin! Mi scipovis abstraktighi* ***de la*** *rusa lingvo,* ***sed post la*** *lego* ***al mi*** *shajnis, ke* ***mi*** *audis chion* ***en la*** *rusa lingvo. Fidela, brila traduko! Ech pli,* ***en*** *kelkaj lokoj* ***al mi*** *ekshajnis, ke* ***la*** *recitanto diris ech pli bellingve, pli esprimive ol Pushkin mem. Chu? Chu eblas diri pri* ***la*** *samo pli bone* ***ol*** *Pushkin? (Отрывок из статьи Superi originalon: chu eble?)*

По результатам сравнения наиболее употребительных коротких слов из текстов на европейских языках ниже представлена сводная таблица совпадений таких маркеров в разных языковых группах (Таблица 2).

Таблица 2

№ п/п	Идентифицирующее слово	В каких языках встречается
1	a	польский, чешский, английский, кастильский (испанский), португальский, румынский, французский, гаэльский, идо
2	na	польский, чешский, гаэльский
3	do	польский, английский, португальский, гаэльский
4	to	польский, английский
5	i	польский, шведский, гаэльский, итальянский
6	ten	польский, чешский, английский
7	o	польский, португальский, румынский
8	se	чешский, кастильский, португальский, французский
9	den	чешский, немецкий, шведский
10	**of, is**	английский, голландский
11	as	английский, португальский, гаэльский
12	in	английский, немецкий, голландский, итальянский, гаэльский
13	das	немецкий, португальский
14	er	немецкий, голландский
15	es	немецкий, кастильский
16	an	немецкий, гаэльский, идо
17	sei	немецкий, итальянский

18	kan	голландский, шведский
19	over	голландский, румынский
20	en	голландский, кастильский, французский, эсперанто, идо
21	de	голландский, шведский, итальянский, кастильский, португальский, румынский, французский, эсперанто
22	e	итальянский, португальский, идо
23	mi	итальянский, эсперанто
24	ma	польский, итальянский, французский
25	la	итальянский, кастильский, французский, румынский, эсперанто, идо
26	le	итальянский, французский, гаэльский
27	un	французский, кастильский

Как видно из приведённых данных, определить язык по служебным и коротким словам практически невозможно, так как они могут полностью совпадать в языках разных групп. Например, формы «**do**» и «**as**» имеются в польском (славянская группа), португальском (романская группа), английском (германская группа) и гаэльском (кельтская группа), что не позволяет использовать их в качестве надёжных идентификаторов, а формы «**da**», «**la**», «**i**», «**le**», «**in**» и т.п. позволяют ассоциировать текст на гаэльском с итальянским языком. Кроме того, такие маркеры как «**a**», «**en**» или «**de**» встречаются во многих языках и должны исключаться из наборов идентификационных элементов как ненадёжные.

Объём приведённых текстов и количество взятых языков не позволяет выявить все совпадения, которые исчисляются десятками даже в языках разных генеалогических групп, но и поверхностного обзора достаточно, чтобы показать практическую несостоятельность данного способа. Например, слово «**den**» (*день*) в чешском совпадает с формой артикля немецкого языка «**den**», а форма глагола «**ma**» (*имеет*) в польском – с союзом «**ma**» (*но*) в итальянском и местоимением «**ma**» (*моя*) во французском языках. А если учесть современную тенденцию к упрощению письменности, при котором не используются диакритические знаки, то такие формы как **si**, **şi** и **sí** будут полностью совпадать, что уменьшит вероятность верной идентификации на порядки. Следовательно, в результате применения этого способа определения языковой принадлежности текста при большом количестве возможных языков либо будет выдан список различных языков, в которых встречаются заданные слова, либо язык будет определён неверно в соответствии с ассоциативной вероятностью. Такой результат на практике оказывается бесполезным.

Расчёт ресурсоёмкости процесса идентификации данным способом аналогичен таковому для словарного способа идентификации по формулам (3) и (4). Следовательно, с учётом относительно малой длины слов-идентификаторов и их небольшого количества для каждого языка, ресурсоёмкость данного способа будет низкой, что позволяет использовать его для тестовых целей и предварительной экспресс-идентификации в многоэтапных алгоритмах.

Отдельно следует сказать о тех случаях, когда в тексте смешиваются слова на разных языках, записанные национальными алфавитами или традиционными письменными системами. Например, имена или названия компаний и товаров могут быть написаны на оригинальном языке, а всё предложение сформулировано по-русски, или в тексте могут встретиться цитаты на других языках. К примеру, в художественных произведениях европейских писателей 18-19 веков часто используются фразы и «крылатые» выражения на латинском или французском языке, что превращает данный текст в многоязычный и требует специального подхода при анализе, а ресурсоёмкость такого процесса многократно увеличится.

Также в связи с развитием компьютерных сетей и сетевого общения, стоит упомянуть о современной тенденции писать текст не традиционной письменностью, которая обычно применяется для данного языка, а использовать наиболее доступные пишущему или наиболее понятные целевой аудитории символы (например, на Интернет-форумах часто пишут по-арабски кириллицей, изменяя, при этом, направление письма), что никак не предусмотрено упомянутыми способами определения языка. То есть, при транскрипции, транслитерации или изменении направления письма, для упомянутых способов текст становится неопределяемым в силу специфики их подходов к идентификации, что потребует дополнительных ресурсов для «нормализации» текста перед его анализом и многократно увеличит ресурсоёмкость всего процесса.

Для иллюстрации характерных особенностей рассмотренных способов языковой идентификации ниже приведена обзорная таблица результатов их сравнения (Таблица 3), в которой наглядно представлены особенности данных способов, а также их основные преимущества, недостатки и ресурсоёмкость.

По итогам проведённого сравнения используемых сегодня способов определения языковой принадлежности текста видно, что совершенных решений не существует и каждый способ имеет ограниченное применение. Это подтверждается исследованиями и других учёных, проводивших похожие сравнения и тестирования технологий языковой идентификации на основе

различных наборов текстов [24, 26, 27, 29, 32, 35].

Таблица 3

№ п/п	Способ	Идентификационные маркеры	Преимущества	Недостат-ки
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

Несмотря на разницу подходов, методов и анализируемых языков, результаты этих исследований очень схожи между собой и с показанными выше, что свидетельствует об их объективности и достоверности. Кроме того, при анализе результатов часто указывается на эффективность комбинирования различных методов и подходов, что позволяет улучшить многие показатели. Но это существенно усложняет как саму систему, так и процесс идентификации, а также повышает ресурсоёмкость и вероятность программного сбоя.

Общим недостатком большинства рассмотренных способов и подходов является отсутствие лингвистического анализа, которое не позволяет отличать языки различных генеалогических групп, что может вызвать критические ошибки при автоматической обработке текста. Это вынуждает разработчиков комбинировать известные решения в одной гибридной системе, что многократно её усложняет и повышает ресурсоёмкость процесса идентификации, но существенно улучшает результаты.

Заключение

Представленный в статье перечень способов и подходов к определению языковой принадлежности неструктурированного текста не является исчерпывающим, поскольку он не охватывает многочисленные вариации и комбинации современных технологий. Но даже краткое и поверхностное изложение исследованной темы показывает множество проблем в данной области, которые остаются нерешёнными. Это является определённым стимулом для разработчиков и причиной поиска новых оригинальных решений при создании соответствующих технологий и программного обеспечения.

Как следует из проведённого анализа, основные отличительные признаки технологий, программ и систем языковой идентификации зависят от применяемых способов, алгоритмов и их технических реализаций. Это позволяет не только оценивать эффективность и ресурсоёмкость существующих разработок, но и осмысленно подходить к проектированию новых с учётом недостатков каждого известного решения.

Литература

1. 1500 необходимых испанских слов / Под. Ред. Пудковой А.В. – Москва: Астрель, 2007. – 63 с.
2. ГИЛЯРЕВСКИЙ Р.С. Определитель языков мира по письменностям / Гривнин В.С. – Москва, 1961. – 303 с.
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. Способы определения языка текста / Филологические науки. Вопросы теории и практики. № 12 (54): в 4-х ч. Ч. II. Тамбов: Грамота, 2015. – С. 84-89.
5. КАЛЕГИН С.Н. Способы определения языковой принадлежности неструктурированного текста в мультиязычной информационной среде. Конференция «CONCORT-2016», Нижний Новгород, 2016.
6. КАМЫНИНА А.А. Современный русский язык. Морфология: Учебное пособие для студентов филологических факультетов государственных университетов. М.: Изд-во МГУ, 1999. 240 с.
7. КУЯНОВ Ю.В. Количественный анализ Большого русского словаря-справочника синонимов / Тришин В.Н. Журнал «Научное обозрение: гуманитарные исследования», №9, 2015 г. – С. 105-111
8. Лингвистический энциклопедический словарь / Гл. ред. В.Н. Ярцева. – Москва: Советская энциклопедия, 1990. – 685 с.
9. Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М.: Азбуковник, 2009. – 1087 с.
10. Патент № 2251737 РФ, G06K9/68. Способ автоматического определения языка распознаваемого текста при многоязычном распознавании / Анисимович К.В., Терещенко В.В., Рыбкин В.Ю.; Аби Софтвер Лтд. (CY); Опубликован 10.05.2005.
11. Патент № 2500024 РФ, G06F17/00. Способ автоматизированного определения языка и (или) кодировки текстового

документа / Лапшин В.А., Пшехотская Е.А., Перов Д.В.; «Центр Инноваций Натальи Касперской» (RU); Опубликован 27.11.2013
12. Патент № 6370269 США. Optical character recognition of handwritten or cursive text in multiple languages. / Al-Karmi, Abdel Naser, Shamsher S., Baldev Singh; International Business Machines Corporation (USA); Опубликован 9.04.2002.
13. ПЕТРОЧЕНКОВ А.В. 2000 самых употребительных английских слов / 2000 Most Frequent English Words. – Москва: Айрис-Пресс, 2011. – 304 с.
14. РУСИНОВ Н.Д. Древнерусский язык / М: Высшая школа, 1997. – 207 с.
15. СЕЛЕЗНЕВ К. Обработка текстов на естественном языке. «Открытые системы», № 12, 2003 [Электронный ресурс]. URL: http://www.osp.ru/os/2003/12/183694/
16. СМИРНОВ В. А. Автоматизированная система анализа неструктурированной речевой информации с применением лингвистических знаний : диссертация … кандидата технических наук : 05.13.11
17. СОТНИК С.Л. Идентификация языка UNICODE-текста по N-граммам длиной до 4-х включительно (квадрограммам) / журнал «Математичне моделювання» № 1,2(15). – Днепродзержинск: издательство ДГТУ, 2006. – С. 111-114.
18. СУХАНОВ В.Ф. Китайско-русский тематический словарь 3000 наиболее употребительных слов. – Москва: Экономика, 2001. – 307 с.
19. Учим японские иероглифы / Под ред. Ф. Сакаде и др. – СПб: Лань, 1997. – 320 с.
20. Частотный словарь русского языка под ред. Засориной Л. Н. – М.: Русский язык, 1977. – 936 с.
21. ABBYY [Электронный ресурс]. URL: http://www.abbyy.ru/products/ (дата обращения: 20.10.2016).
22. ALMEIDA-CRUZ, Y., ESTÉVEZ-VELARDE, S., PIAD-MORFFIS, A. Detección de Idioma en Twitter / Revista Internacional de Gestión del Conocimiento y la Tecnología. Vol. 2 (3), 2014.
23. CAVNAR W. B., TRENKLE J. M.. 1994. N-gram-based text categorization. In Proceedings of SDAIR'94, pages 161–175.
24. GOTTRON T., LIPKA N. A Comparison of Language Identification Approaches on Short, Query-Style Texts. / 32nd European Conference on Information Retrieval (ECIR 10), Springer, 2010.
25. GREFENSTETTE G. Comparing two language identification schemes. / 3rd International Conference JADT 1995 (Statistical Analysis of Textual Data). Rome, Italy.
26. GROTHE L., E. DE LUCA W., A. NURNBERGER A Comparative Study on Language Identification Methods / Conference LREC 2008, Marrakech, Morocco.
27. INDHUJA K, INDU M, SREEJITH C, P. C. REGHU RAJ Text Based Language Identification System for Indian Languages Following Devanagiri Script / International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 4, 2014.
28. JOHNSON S. Solving the problem of language recognition [Report] / Technical report, School of Computer Studies, University of Leeds, 1993.
29. KRANIG S. Evaluation of Language Identification Methods / University of Tubingen, Bachelor of Arts Thesis.
30. mnoGoSearch [Электронный ресурс]. URL: http://www.mnogosearch.org/ (дата обращения: 20.10.2016).
31. ÖLVECKÝ T. N-Gram Based Statistics Aimed at Language Identification / Konferencia IIT.SRC 2005, Bratislava.
32. PADRO M., PADRO Ll. Comparing methods for language identification. URL: http://www.sepln.org/revistaSEPLN/revista/33/33-Pag155.pdf (дата обращения: 20.10.2016).
33. ROMSDORFER, H., PFISTER, B. (2007). Text analysis and language identification for polyglot text-to-speech synthesis. Speech Communication 49, 697-724.
34. TIMOTHY J. Hazen Automatic Language Identification Using a Segmen t-Based Approach / S.B., Massachusetts Institute of Technology, 1991.
35. VOJTEK P., BIELIKOVA M. Comparing Natural Language Identification Methods based on Markov Processes.

References

1. 1500 neobhodimyh ispanskih slov / Pod. Red. Pudkovoj A.V. – Moskva: Astrel', 2007. – 63 s.
2. GILYAREVSKIJ R.S. Opredelitel' yazykov mira po pis'mennostyam / Grivnin V.S. – Moskva, 1961. – 303 s.
3. KALEGIN S.N. Vazhnost' vybora osnovnogo identifikacionnogo principa pri proektirovanii yazykovyh opredelitelej. / Sovremennye informacionnye tekhnologii i IT-obrazovanie. Tom 12, № 2. – Moskva 2016. – S. 194-204.
4. KALEGIN S.N. Sposoby opredeleniya yazyka teksta / Filologicheskie nauki. Voprosy teorii i praktiki. № 12 (54): v 4-h ch. CH. II. Tambov: Gramota, 2015. – S. 84-89.
5. KALEGIN S.N. Sposoby opredeleniya yazykovoj prinadlezhnosti nestrukturirovannogo teksta v mul'tiyazychnoj informacionnoj srede. Konferenciya «CONCORT-2016», Nizhnij Novgorod, 2016.
6. KAMYNINA A.A. Sovremennyj russkij yazyk. Morfologiya: Uchebnoe posobie dlya studentov filologicheskih fakul'tetov gosudarstvennyh universitetov. M.: Izd-vo MGU, 1999. 240 s.
7. KUYANOV YU.V. Kolichestvennyj analiz Bol'shogo russkogo slovarya-spravochnika sinonimov / Trishin V.N. ZHurnal «Nauchnoe obozrenie: gumanitarnye issledovaniya», №9, 2015 g. – S. 105-111
8. Lingvisticheskij ehnciklopedicheskij slovar' / Gl. red. V.N. YArceva. – Moskva: Sovetskaya ehnciklopediya, 1990. – 685 s.
9. Lyashevskaya O. N., SHarov S. A. CHastotnyj slovar' sovremennogo russkogo yazyka (na materialah Nacional'nogo korpusa russkogo yazyka). – M.: Azbukovnik, 2009. – 1087 s.
10. Patent № 2251737 RF, G06K9/68. Sposob avtomaticheskogo opredeleniya yazyka raspoznavaemogo teksta pri mnogoyazychnom raspoznavanii / Anisimovich K.V., Tereshchenko V.V., Rybkin V.YU.; Abi Softver Ltd. (CY); Opublikovan 10.05.2005.
11. Patent № 2500024 RF, G06F17/00. Sposob avtomatizirovannogo opredeleniya yazyka i (ili) kodirovki tekstovogo dokumenta / Lapshin V.A., Pshekhotskaya E.A., Perov D.V.; «Centr Innovacij Natal'i Kasperskoj» (RU); Opublikovan 27.11.2013
12. Patent № 6370269 SSHA. Optical character recognition of handwritten or cursive text in multiple languages. / Al-Karmi, Abdel Naser, Shamsher S., Baldev Singh; International Business Machines Corporation (USA); Opublikovan 9.04.2002.
13. PETROCHENKOV A.V. 2000 samyh upotrebitel'nyh anglijskih slov / 2000 Most Frequent English Words. – Moskva: Ajris-Press, 2011. – 304 s.
14. RUSINOV N.D. Drevnerusskij yazyk / M: Vysshaya shkola, 1997. – 207 s.
15. SELEZNEV K. Obrabotka tekstov na estestvennom yazyke. «Otkrytye sistemy», № 12, 2003 [EHlektronnyj resurs]. URL: http://www.osp.ru/os/2003/12/183694/
16. SMIRNOV V. A. Avtomatizirovannaya sistema analiza nestrukturirovannoj rechevoj informacii s primeneniem lingvisticheskih znanij : dissertaciya … kandidata tekhnicheskih nauk : 05.13.11
17. SOTNIK S.L. Identifikaciya yazyka UNICODE-teksta po N-grammam dlinoj do 4-h vklyuchitel'no (kvadrogrammam) / zhurnal «Matematichne modelyuvannya» № 1,2(15). – Dneprodzerzhinsk: izdatel'stvo DGTU, 2006. – S. 111-114.

18. SUHANOV V.F. Kitajsko-russkij tematicheskij slovar' 3000 naibolee upotrebitel'nyh slov. – Moskva: EHkonomika, 2001. – 307 s.
19. Uchim yaponskie ieroglify / Pod red. F. Sakade i dr. – SPb: Lan', 1997. – 320 s.
20. CHastotnyj slovar' russkogo yazyka pod red. Zasorinoj L. N. – M.: Russkij yazyk, 1977. – 936 s.
21. ABBYY [EHlektronnyj resurs]. URL: http://www.abbyy.ru/products/ (data obrashcheniya: 20.10.2016).
22. ALMEIDA-CRUZ, Y., ESTÉVEZ-VELARDE, S., PIAD-MORFFIS, A. Detección de Idioma en Twitter / Revista Internacional de Gestión del Conocimiento y la Tecnología. Vol. 2 (3), 2014.
23. CAVNAR W. B., TRENKLE J. M.. 1994. N-gram-based text categorization. In Proceedings of SDAIR'94, pages 161–175.
24. GOTTRON T., LIPKA N. A Comparison of Language Identification Approaches on Short, Query-Style Texts. / 32nd European Conference on Information Retrieval (ECIR 10), Springer, 2010.
25. GREFENSTETTE G. Comparing two language identification schemes. / 3rd International Conference JADT 1995 (Statistical Analysis of Textual Data). Rome, Italy.
26. GROTHE L., E. DE LUCA W., A. NURNBERGER A Comparative Study on Language Identification Methods / Conference LREC 2008, Marrakech, Morocco.
27. INDHUJA K, INDU M, SREEJITH C, P. C. REGHU RAJ Text Based Language Identification System for Indian Languages Following Devanagiri Script / International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 4, 2014.
28. JOHNSON S. Solving the problem of language recognition [Report] / Technical report, School of Computer Studies, University of Leeds, 1993.
29. KRANIG S. Evaluation of Language Identification Methods / University of Tubingen, Bachelor of Arts Thesis.
30. mnoGoSearch [EHlektronnyj resurs]. URL: http://www.mnogosearch.org/ (data obrashcheniya: 20.10.2016).
31. ÖLVECKÝ T. N-Gram Based Statistics Aimed at Language Identification / Konferencia IIT.SRC 2005, Bratislava.
32. PADRO M., PADRO Ll. Comparing methods for language identification. URL: http://www.sepln.org/revistaSEPLN/revista/33/33-Pag155.pdf (data obrashcheniya: 20.10.2016).
33. ROMSDORFER, H., PFISTER, B. (2007). Text analysis and language identification for polyglot text-to-speech synthesis. Speech Communication 49, 697-724.
34. TIMOTHY J. Hazen Automatic Language Identification Using a Segmen t-Based Approach / S.B., Massachusetts Institute of Technology, 1991.
35. VOJTEK P., BIELIKOVA M. Comparing Natural Language Identification Methods based on Markov Processes.

---

*Статья написана по материалам диссертационной работы автора на тему «**Автоматическая идентификация языковой принадлежности текста по формам предикатов**». Научный руководитель: Зубарев Юрий Борисович, доктор технических наук, профессор, член-корреспондент РАН. Научный консультант: Фархадов Маис Паша оглы, доктор технических наук, старший научный сотрудник.*

Поступила: 21.03.2017

Об авторе:

Калегин Сергей Николаевич, начальник сектора НТО ЗАО «МНИТИ»; аспирант, ЗАО «МНИТИ», Институт проблем управления им. В.А. Трапезникова РАН, ksn@mniti.ru

Note on the author:

Kalegin Sergey, Head of the NTO sector, CJSC MNITI; aspirant of CJSC MNITI and applicant of Institute of Control Sciences of the Russian Academy of Sciences, ksn@mniti.ru