УДК 004.62

Калегин С.Н.

ЗАО «МНИТИ», Москва, РФ

ЛОГИЧЕСКАЯ СТРУКТУРА ИДЕНТИФИКАЦИОННЫХ НАБОРОВ В СИСТЕМАХ ЯЗЫКОВОЙ ИДЕНТИФИКАЦИИ

АННОТАЦИЯ

В данной статье рассматриваются варианты логической структуры идентификационных наборов для программ и систем языковой идентификации с целью оптимизации существующих решений, а также даются примеры их упрощения для повышения эффективности.

КЛЮЧЕВЫЕ СЛОВА

Идентификационный набор; идентификационная структура; языковая идентификация; идентификация языка; определение языка; определение языковой принадлежности.

Kalegin S.N.

CJSC MNITI, Moscow, RF

LOGICAL STRUCTURE OF IDENTIFICATION KITS IN LANGUAGE IDENTIFICATION SYSTEMS

ABSTRACT

This article discusses the options for the logical structure of identification kits for language identification software and systems in order to identify the advantages and disadvantages of existing solutions, and provides examples of their simplification and optimization to increase efficiency.

KEYWORDS

Identification kit; identification structure; language identification; language detection.

Системы определения языковой принадлежности неструктурированного текста в современном мире востребованы гораздо больше, чем это может показаться на первый взгляд, так как перед началом обработки любого текста требуется определить язык, на котором он написан, иначе невозможно будет с ним работать [1, 2]. Особое значение эта проблема приобретает при автоматической обработке текстовых данных в многоязычной информационной среде, такой как Интернет, базы знаний и т.п. Для этой цели создаются программные средства определения языковой принадлежности неструктурированного текста. Они применяются при машинной идентификации языка документов в подразделениях организаций и системах по обработке многоязычной информации, например, в бюро переводов, отделах по работе с корреспонденцией, почтовых отделениях или программах по работе с электронной почтой, библиотеках, программных комплексах по сбору и обработке неструктурированной информации, системах автоматического перевода, текстовых процессорах, классификаторах различных текстовых документов, программах распознавания и конвертации изображений или речи и т.д., а также при определении языка транслитерированных, транскрибированных и записанных специальными (например, шрифтом Брайля) или нетипичными письменными системами. В задачу такой программы входит точная идентификация преобладающего языка информационного блока (например, при сортировке корреспонденции) или всех использованных в нём языков (при распознавании текстовых документов). Работа автоматического определителя языка является сложным и многоуровневым процессом, который упрощённо можно представить следующей схемой (Рисунок 1).

Таким образом, процесс идентификации представляет собой комплекс нескольких основных операций, перечисленных ниже.

- Получение анализируемых данных.
- Преобразование полученных данных в массив анализируемых элементов.

- Считывание идентификационного набора.
- Формирование идентификационной матрицы.
- Сопоставление анализируемых элементов с идентификационными маркерами.
- Ассоциация анализируемых элементов с конкретным языком или группой по результатам сопоставления.
- Вывод результатов языковой идентификации.

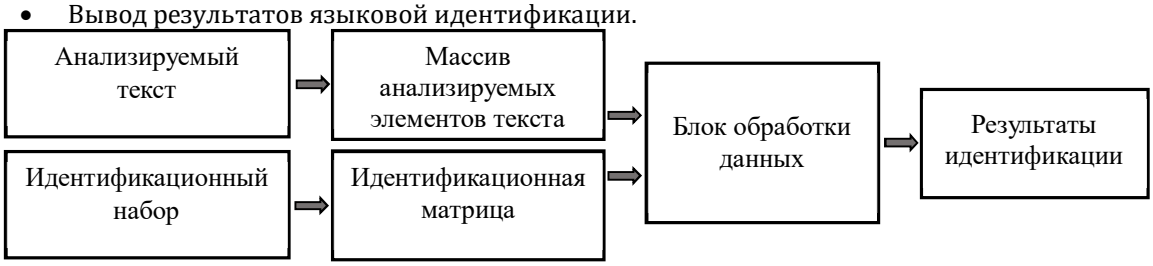

Рис. 1. Упрощённая блок-схема работы определителя языка текста

При этом программа должна учитывать письменную систему, лексические и грамматические особенности языков, различия в диалектах, форматы ввода и вывода информации, особенности способов идентификации, недостатки языковых моделей и множество других факторов. Качество идентификации и эффективность языкового определителя зависят от выбора способа языковой идентификации и его технической реализации. Основой всего процесса идентификации является набор идентифицирующих элементов – идентификационных маркеров, на базе которого генерируется идентификационная матрица, содержащая ассоциативные связи маркеров с конкретными языковыми группами, вариантами, языками и диалектами, а также другие характерные параметры. В этой статье подробно рассмотрены возможные модификации и логическая структура такого набора, а также его техническая реализация с указанием особенностей каждого технологического решения. В качестве основы языковой идентификации выбран один из лингвистических способов [3], позволяющий определять языковую принадлежность по формам предикатов [4], где в качестве идентификационных маркеров обычно используются формы глаголов. Это позволит наиболее наглядно продемонстрировать вариации идентификационного набора и возможности его оптимизации.

Выбор конкретных форм идентифицирующих элементов, включаемых в идентификационный набор, зависит от языка, цели и требуемой детализации. Например, для определения только языковой группы и для определения конкретного варианта или диалекта будут использоваться различные наборы идентификационных маркеров. А для некоторых языков, например, английского или французского, характерным является частое использование вспомогательных глаголов, формы которых требуется обязательно учитывать. С помощью комбинаций выбранных групп и форм глаголов или аналогичных предикативных образований идентифицируемых языков, и при условии исключения из составляемых наборов совпадающих форм в разных языках и/или языковых группах, может быть достигнута высокая точность определения языковой принадлежности текста. От качества составления таких наборов зависит эффективность и область применения данного способа, количество идентифицируемых языков и вероятность возникновения ошибок.

Разработка подобного идентификационного набора и его применение могут быть кратко описаны двумя наглядными последовательностями.

1. Разработка:
   a. выбор идентификационных маркеров;
   b. разработка идентификационной структуры;
   c. формирование идентификационного набора.
2. Применение:
   a. считывание идентификационного набора;
   b. формирование идентификационных маркеров;
   c. создание идентификационной матрицы.

Из приведённого описания выделяются три ключевых компонента, необходимых для использования таких наборов на практике:

- идентификационные маркеры;
- идентификационная структура;
- идентификационная матрица.

Эти составляющие являются основой всего процесса идентификации данным способом, так как от них зависят эффективность и условия применения его программной реализации. Ниже

детально рассматривается разработка указанных компонентов, что позволяет показать возможные варианты и модификации как самого набора, так и идентификационной матрицы.

Идентификационные маркеры формируются программой на основании списка идентифицирующих элементов, которые сопоставляются с определённым языком или группой в идентификационном наборе. Для простоты восприятия в следующих примерах будут использоваться только формы наиболее употребительных и вспомогательных глаголов.

Идентификационный набор может быть представлен в нескольких формах и различных вариациях. Например, в виде списка с определённой структурой, таблицы или многомерного массива, где имеется одна или несколько групп идентифицирующих элементов каждого определяемого языка, указана связь этих групп с конкретным языком и языковой группой (и/или подгруппой), а также языковой ветвью, семьёй или макросемьёй по мере необходимости. Конкретная схема деления зависит от выбранной классификации языков и требуемых ассоциаций. В данной статье за основу берётся генеалогическая классификация [5], но в идентификационном наборе могут использоваться и другие, например: реляционная, морфологическая (типологическая) или структурная [6]. При этом структура набора может быть линейной, в которой просто перечисляются языки или группы с их идентификаторами (Рисунки 2-3), или иерархической (Рисунок 4), где градация может быть разветвлённая и многоуровневая (например, элементы близкородственных или однотипных языков находятся на одном уровне отдельной ветви иерархии), а для каждой языковой группы и отдельных языков могут даваться уточнения или более подробная языковая классификация, например, деление на подгруппы, варианты и/или диалекты. К примеру, английский язык относится к западногерманской языковой подгруппе и для него существуют британский, американский, австралийский и другие варианты со множеством диалектов внутри каждого из них, что может быть отражено в структуре набора, позволяя устанавливать степень идентификации на уровне генеалогической группы, языка, варианта, диалектной группы или диалекта. Ниже приводятся примеры обеих структур.

На рисунке 2 демонстрируется упрощённый вариант линейной структуры идентификационного набора для языков различных генеалогических групп. Деление идентификаторов для вариантов и диалектов не производится, идентифицирующие элементы заменены отточием.

Язык: список идентифицирующих элементов

русский: ……..……..
польский: ……..……..
латинский: ……..……..……
кастильский: ……..……..……
гаэльский: ……..……..……
валлийский: ……..….
литовский: ……
эстонский: ……..……..
венгерский: ……..…….
албанский: ……
английский: ……..……
немецкий: ……..……..…
шведский: ……..
волапюк: ……..….
эсперанто: ……..……

Рис.2. Пример линейной структуры идентификационного набора для определения отдельных языков

Языковая группа: список идентифицирующих элементов

славянская: ……..……..…
романская: ……..……..…
кельтская: ……..……
балтийская: ……..…….
финно-угорская: ……..….
палеобалканская: ……..….
германская: ……..…….

Рис. 3. Пример линейной структуры идентификационного набора для определения языковых групп

Преимущество такой структуры в простоте реализации и обработки. Однако она не

позволяет определять языковые группы и подгруппы в случае, если язык определить не удалось. Поэтому требуется использовать дополнительный идентификационный набор, в котором элементы идентифицируют только группы и/или подгруппы (Рисунок 2), или усложнять предыдущий набор, что изменит его структуру и алгоритм обработки.

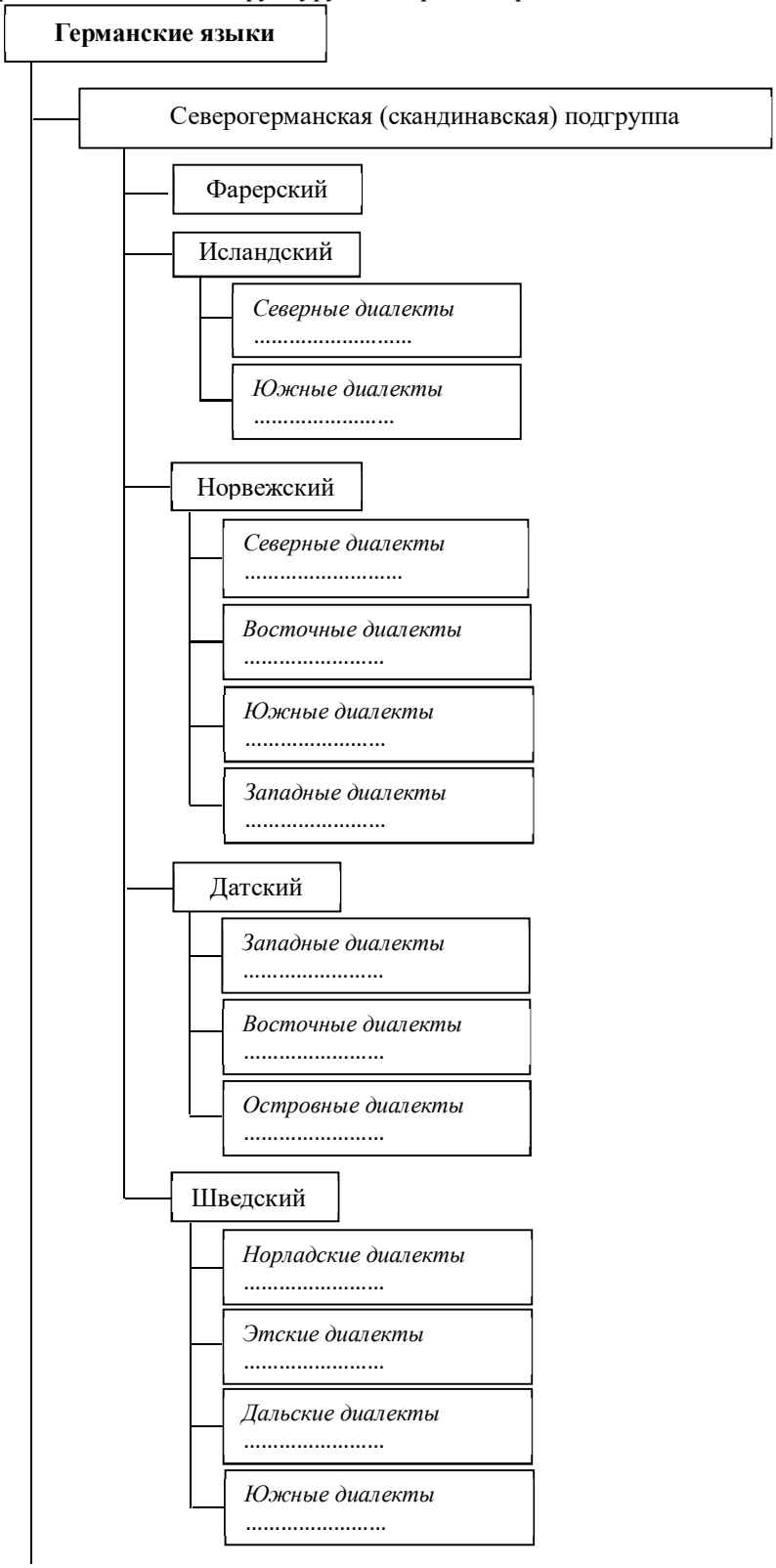

Рис. 4. Часть иерархической структуры идентификационного набора

На рисунке 4 схематически изображена возможная структура иерархического идентификационного набора на примере северной (скандинавской) подгруппы германских языков,

которые объединены по генеалогическим признакам. Языки, варианты и диалектные группы здесь различаются условно, так как их общепринятой градации не существует. Для краткости отдельные диалекты не указываются, а списки идентифицирующих элементов для каждой группы диалектов заменены отточием.

Такая структура идентификационного набора позволяет определять языковые ветви, группы или подгруппы без определения самого языка анализируемого текста, так как позволяет объединять идентифицирующие элементы, а также может включать совпадающие формы идентификаторов на разных уровнях иерархии, которые обычно не используются в линейных наборах и, следовательно, не участвуют в идентификации отдельных языков.

Ниже приводится несколько примеров рассмотренных идентификационных наборов с линейной структурой, созданных на основе простого текстового файла (приводятся в сокращённом и упрощённом виде с комментариями), где идентификаторы записываются в одну строку после названия языка и определённого разделителя, а первые 2 строки служат для оформления контента и программой не анализируются. В качестве идентифицирующих элементов используются формы наиболее употребительных глаголов.

Язык → Идентифицирующие элементы

русский → быть, есть, был, была, было, были, иметь, имею, имеешь и т.д.
болгарский → съм, си, сме, сте, имам, имаш, има, имаме и т.д.
сербский → бити, је, имати, имамо, имају, немам, немаш и т.д.
польский → jestem, jest, mam, masz, mamy, macie и т.д.
немецкий →sein, bin, ist, sind, war, gewesen, haben, habe и т.д.
латинский →esse, sum, sunt, fui, habere, habeo и т.д.
гаэльский → bi, ta, taim, taimid, bhfuil, nil, tha и т.д.
латышский → but, esmu, esam, esat, darit и т.д.
эстонский → olla, olen, oled, on, oleme, olete, evima, omama и т.д.
венгерский → lenni, vagyok, vagy, vagyunk, vagytok, vannak и т.д.
эсперанто → esti, estas, estos, estis, estus, havi, havas, havos и т.д.

В данном примере для определения языка используются грамматические формы нескольких наиболее употребительных глаголов каждого идентифицируемого языка различных генеалогических групп Европы (без указания на эти группы). В качестве разделителя выбран знак «**→**» (стрелка) для наглядности, хотя может использоваться любой другой символ или сочетание, например: знаки препинания, цифры и т.д.. С помощью такого набора идентифицирующих элементов можно определить язык текста, если формы глаголов в анализируемом тексте не имеют совпадений со словами других языков. В противном случае язык не будет идентифицирован или результат будет ошибочным. Хотя эта проблема может быть решена усложнением алгоритма обработки данного набора, но применение такого решения не всегда оправдано и рационально.

При необходимости определения принадлежности языка текста к одной из языковых групп приведённый выше набор может быть изменён следующим образом (показан в сокращённом и упрощённом виде):

Группа → (Язык) → Идентифицирующие элементы

славянская → (русский) → быть, есть, был, была, иметь, имею, имеешь и т.д.
славянская → (болгарский) → съм, си, сме, сте, имам, имаш, има, имаме и т.д.
славянская → (сербский) → бити, је, имати, имамо, имају, немам, немаш и т.д.
славянская → (польский) → jestem, jest, mam, masz, ma, mamy, macie и т.д.
германская → (немецкий) → sein, bin, ist, sind, war, gewesen, haben, habe и т.д.
искусственные → (эсперанто) → esti, estas, estos, estis, havi, havas, havos и т.д.

В приведённых примерах сознательно исключены совпадающие формы глаголов в близкородственных языках (например, южнославянских: сербском и болгарском), чтобы избежать возникновения ошибок в программе. Для краткости и наглядности здесь деление на подгруппы не производится, хотя оно может быть добавлено по мере необходимости, а названия языков указаны в скобках для уточнения принадлежности идентификаторов. Как явствует из приведённого набора, наличие совпадений форм глаголов в языках одной группы (например, славянской) не повлияет на результат, если целью анализа текста является только определение языковой группы. В таком случае указания на конкретный язык могут быть исключены из набора идентифицирующих элементов, а совпадающие формы глаголов в близкородственных языках в него добавлены, что улучшит результат определения языковой группы.

Теоретически, линейная структура может усложняться и дальше, однако количество повторяющихся названий групп и необходимость их обработки делают применение подобной

структуры нетехнологичным, поэтому в подобных случаях её предлагается заменять на иерархическую. Образец возможной иерархии такого набора в текстовом виде на примере северогерманской генеалогической подгруппы приведён ниже (идентификационные маркеры заменены отточием).

Германские языки
–> Северогерманские
–> –> Фарерский
–> –> –> : ……
–> –> Исландский
–> –> –> южные диалекты: ……
–> –> –> северные диалекты: ……
–> –> Норвежский
–> –> –> южные диалекты: ……
–> –> –> западные диалекты: ……
–> –> –> северные диалекты: ……
–> –> –> восточные диалекты: ……
–> –> Датский
–> –> –> островные диалекты: ……
–> –> –> западные диалекты: ……
–> –> –> восточные диалекты: ……
–> –> Шведский
–> –> –> южные диалекты: ……
–> –> –> этские диалекты: ……
–> –> –> дальские диалекты: ……
–> –> –> норландсике диалекты: ……

В данном примере уровни иерархии определяются количеством сочетаний знаков «–>», например, «–>–>» указывает на 2-й уровень. Однако в рабочем идентификационном наборе использование подобных сочетаний представляется нетехнологичным, поэтому рекомендуется заменять их более простыми символами (тире, точками, запятыми, цифрами и т.д.).

Рабочий вариант такой двухуровневой реализации на основе форм вспомогательных глаголов приведён ниже (показан в сокращении). В качестве разделителя идентифицирующих элементов здесь используется двоеточие, сокращённые элементы заменены отточием.

Группа / Язык | Идентифицирующие элементы

славянская
русский : быть : есть : был : была : были : будет : будут : иметь : имею : имеют ……..
болгарский : съм : си : сме : са : съ : имам : имаш : има : имаме : имате ……
Сербский : бити : сам : смо : су : јесам : јеси : јесте : јесмо : јесу : нисам ……
польский : jestem : jest : mam : masz : ma : mamy : macie : delae : delas ……..
романская
латинский : esse : sum : est : sunt : fui : habere : habeo : habes : habet ……..
итальянский : essere : sono : siamo : siete : ho : hai : habbiamo : avete : hanno ……..
каталанский : soc : ets : som : sou : estat : fer : faig : fas : fem : feu : fan ……
кастильский : ser : soy : eres : somos : son : haber : has : hemos : han : hecho ……
португальский : sou : haver : hei : havemos : fazer : faz : fazemos : fazem ……..
французский : etre : suis : sommes : sont : avoir : ai : avons : avez : ont : faire ……
кельтская
гаэльский : bi : ta : taim : taimid : bhfuil : nil : tha ……..
Валлийский : bod : dw : wyt : mae : dyn : dych : maen : dydy : dydyn ……..
балтийская
литовский : buti : esu : yra : esame : esate : nesu : nesi : nesame : nesate ……..
латышский : but : esmu : ir : esam : esat : darit ……..
финно-угорская
эстонский : olla : olen : oled : on : oleme : olete : evima : omama : teha : teen ……..
венгерский : lenni : vagyok : vagy : vagyunk : vagytok : vannak : nincs : nincsenek …….
Палеобалканская
албанский : eshte : jemi : jeni : ke : ka : kemi : keni : kini : kane : bej ……
германская
английский : be : are : were : have : has : had : does : did ……..

немецкий : sein : bin : ist : sind : war : gewesen : haben : habe : hatte : gehabt ……
голландский : zijn : ben : bent : geweest : hebben : heb : hebt : gehad ……
Шведский : vara : var : voro : varit : hava : har : haver ……
искуственные языки
волапюк : binob : binol : binor : binof : binos : binom : binon : binobs : binols ……
эсперанто : esti : estas : estos : estis : estus : havi : havas : havos : havis : fari ……
Идо : esar : esor : esir : esas : esos : esis : havar : havir : havor : facar ……

Такой идентификационный набор может быть размещён в простом текстовом файле, который при считывании программой будет преобразован в многомерный массив данных, представляющий собой двухуровневую идентификационную матрицу, где на первом уровне находятся названия языковых групп (выделены полужирным шрифтом), а на втором – язык, определяемый первым местом в списке, и его идентификационные маркеры. В данном случае маркеры являются уникальными, поэтому язык будет определяться однозначно за один проход программы по элементам. Для определения только языковой группы при недостатке идентификаторов в данный набор должны быть включены все совпадающие формы в родственных языках, а названия самих языков могут исключаться. Таким образом, идентификационный набор приобретает следующий вид (приводится в сокращении).

Группа | Идентифицирующие элементы

славянская
быть : есть : был : была : было : были : буду : будешь : будет : будем : будут : иметь : имею : имеешь : имеет : имеем : имеют : имел : имела : имели : делать : сделать : делал : делала : сделал : сделала : делает : сделает : делают : сделают : делаем : сделаем
съм : си : сме : сте : са : съ : имам : имаш : има : имаме : имате : имат : правя : правиш : прави : правим : правите : правят
бити : сам : смо : су : јесам : јеси : јесте : јесмо : јесте : јесу : нисам : ниси : није : нисмо : нисте : нису:
имати : имамо : имају : немам : немаш : нема : немамо : немате : немају
jestem : jest : mam : masz : mamy : macie : delae : delas : dela : delame
романская
esse : sum : est : sunt : fui : habere : habeo : habes : habet : habemus : habetis : habent : habui : facere : facio : facis : facit : facimus : facitis
essere : sono : siamo : siete : hai : habbiamo : avete : hanno : fare : faccio : fai : facciamo : fate :
soc : ets : som : sou : estat : fer : faig : fas : fem : feu : fan
ser : soy : eres : somos : son : haber : has : hemos : han : hecho : hay : hacer : hago : haces : hace
sou : haver : hei : havemos : fazer : faz : fazemos : fazem : fazendo
etre : suis : sommes : sont : avoir : ai : avons : avez : ont : faire : fais : fait : faisons : faites
кельтская
bi : ta : taim : taimid : bhfuil : nil : tha
bod : dw : wyt : mae : dyn : dych : maen : dydy : dydyn : gwneud : gwnaf : gwnei : gwna : gwnawn : gwnewch : gwnant
балтийская
buti : esu : yra : esame : esate : nesu : nesi : nesame : nesate
but : esmu : esi : esam : esat : darit
финно-угорская
olla : olen : oled : oleme : olete : evima : omama : teha : teen : teed : teeb : teeme : teete : teevad
lenni : vagyok : vagy : vagyunk : vagytok : vannak : nincs : nincsenek
германская
be : are : is : were : have : has : had : does : did : I
sein : bin : ist : sind : war : gewesen : haben : habe : hatte : gehabt : 179d war : mache : macht : gemacht : tun : tue : tut : tat : getan
zijn : ben : bent : geweest : hebben : heb : hebt : gehad : maken : maak : maakt : maakten :
vara : var : voro : varit : hava : har : haver
искусственные языки
binob : binol : binor : binof : binos : binom : binon : binobs : binols : binofs : binoms : binons : binors : labob : labol : labom : labof : labon : labobs : labols : laboms : labons : labofs
esti : estas : estos : estis : estus : havi : havas : havos : havis : fari : faras : faros : faris
esar : esor : esir : esas : esos : esis : havar : havir : havor : facar : facas : facos : facor : facir
не определена

Как видно из примера, в наборе разделителем идентифицирующих элементов является двоеточие, а каждая строка содержит глагольные формы одного из языков указанной выше группы.

Названия групп выделены в отдельные строки для простоты обработки, последняя запись выводится в том случае, если языковую группу определить не удалось.

Здесь следует отметить, что в рабочем варианте идентифицирующие элементы могут записываться наиболее рациональным образом для данной программы или её конкретной реализации. Например, если идентификация проводится по формам таких глаголов как «делать», «иметь» и т.д., то они могут быть записаны в следующих вариациях:

1. … делать, делаю, делаешь, делает, делаем, делаете, делают …….
2. … дела|ть|ю|ешь|ет|ем|ете|ют …….

То есть, основа глагола может выделяться и больше не повторяться в наборе, что существенно сократит его длину и время считывания, но потребует дополнительных ресурсов для генерации идентификационных маркеров. Также основы могут храниться отдельно от аффиксов, например, в разных строках, файлах или таблицах. Построчно это может быть реализовано следующим образом:

- основы: дела, име ……;
- аффиксы: ть, ю, ешь, ет, ем, ете, ют ……

После считывания и объединения основ со всеми аффиксами получится два набора идентификационных маркеров:

1. … делать, делаю, делаешь, делает, делаем, делаете, делают …….;
2. … иметь, имею, имеешь, имеет, имеем, имеете, имеют ……..

В рабочем варианте идентификационного набора это может выглядеть примерно следующим образом (приводится фрагмент).

Группа / Язык | Идентифицирующие элементы (основы : аффиксы)

искусственные языки

волапюк : bin, lab : ob|ol|or|of|os|om|on|obs|ols|ofs|oms|ons|ors

эсперанто : est, hav, far : i|as|os|is|us

идо : es, hav, far : ar|or|ir|as|os|is

Для примера взята группа искусственных языков, так как в них отсутствуют изменения корня при спряжении глаголов, но для естественных языков потребуется более сложная система морфологических сокращений. В приведённом фрагменте видно, что при таком подходе структура каждой строки идентифицирующих элементов существенно меняется и требует более сложной обработки, так как список теперь включает комплексные разнородные составляющие. Но при этом размер записи идентификаторов сократился с 299 знаков (при полной записи) до 127, то есть более чем в 2 раза, что существенно экономит ресурсы.

Таким образом, варианты записи идентифицирующих элементов могут влиять на размеры идентификационных наборов и ресурсоёмкость их обработки. Поэтому вопрос формирования идентификационных маркеров должен решаться до или во время планирования структуры идентификационного набора и учитываться при расчёте ресурсоёмкости всей системы языковой идентификации.

Рассмотренные варианты идентификационного набора могут модифицироваться в зависимости от целей и задач, поставленных перед программой-определителем, а также от алгоритма и конкретной реализации процесса языковой идентификации. Но в итоге обработки такого набора в памяти компьютера должна быть сформирована идентификационная матрица (фильтр), позволяющая решить поставленные задачи. При этом её структура может не соответствовать структуре идентификационного набора в силу особенностей алгоритма обработки или выбранных программистом технологий, что не должно отражаться на результате.

Из приведённых примеров наборов идентифицирующих элементов следует, что предлагаемая технологическая реализация не зависит от используемой в анализируемом тексте системы письма или фиксации информации, так как в самом наборе могут быть использованы различные системы письма и фиксации информации (например, слоговые знаки, идеограммы или комбинации точек шрифта Брайля). Кроме того, сам идентификационный набор может быть многократно транслитерирован, что позволяет работать именно с языковыми маркерами, а не с символами письменной системы, как это делается в большинстве современных определителей языковой принадлежности текста. Кроме того, текстовый тип данных легко обрабатывается и редактируется, не требуя специализированных программ и систем, например, СУБД (систем управления базами данных), что существенно упрощает применение таких систем в различных областях и уменьшает ресурсозатраты.

Статья написана по материалам диссертации автора «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.

Литература

1. Beesley K. R. Language identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text / 29th Annual Conference of the American Translators Association. Seattle, Washington, USA, 1988.
2. Sibun P., Reynar J. C. Language Identification: Examining the Issues / 5th Symposium on Document Analysis and Information Retrieval (SDAIR-96). Las Vegas, Nevada, USA, 1996.
3. Калегин С.Н. Способы определения языковой принадлежности неструктурированного текста в мультиязычной информационной среде. Конференция «CONCORT-2016», Нижний Новгород, 2016.
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. Лингвистический энциклопедический словарь / Гл. ред. В.Н. Ярцева. – Москва: Советская энциклопедия, 1990. – 685с.
6. Даниленко В.П. Общая типология языков в концепции Э.Сепира [Электронный ресурс]. URL: http://old.islu.ru/danilenko/articles/sepirtipol.htm

References

1. Beesley K. R. Language identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text / 29th Annual Conference of the American Translators Association. Seattle, Washington, USA, 1988.
2. Sibun P., Reynar J. C. Language Identification: Examining the Issues / 5th Symposium on Document Analysis and Information Retrieval (SDAIR-96). Las Vegas, Nevada, USA, 1996.
3. Kalegin S.N. Sposoby opredelenija jazykovoj prinadlezhnosti nestrukturirovannogo teksta v mul'tijazychnoj informacionnoj srede. Konferencija «CONCORT-2016», Nizhnij Novgorod, 2016.
4. Kalegin S.N. Sposoby opredelenija jazyka teksta / Filologicheskie nauki. Voprosy teorii i praktiki. № 12 (54): v 4-h ch. Ch. II. Tambov: Gramota, 2015. – S. 84-89.
5. Lingvisticheskij jenciklopedicheskij slovar' / Gl. red. V.N. Jarceva. – Moskva: Sovetskaja jenciklopedija, 1990. – 685 s.
6. Danilenko V.P. Obshhaja tipologija jazykov v koncepcii Je.Sepira [Jelektronnyj resurs]. URL: http://old.islu.ru/danilenko/articles/sepirtipol.htm

Поступила: 13.10.2016

Об авторе:

Калегин Сергей Николаевич, аспирант, начальник сектора НТО ЗАО «МНИТИ», e-mail: ksn@mniti.ru.