УДК 004.05

Калегин С.Н.

ЗАО «МНИТИ», Москва, РФ

ВАЖНОСТЬ ВЫБОРА ОСНОВНОГО ИДЕНТИФИКАЦИОННОГО ПРИНЦИПА ПРИ ПРОЕКТИРОВАНИИ ЯЗЫКОВЫХ ОПРЕДЕЛИТЕЛЕЙ

АННОТАЦИЯ

В статье приводятся результаты сравнительного исследования разнотипных программ определения языковой принадлежности текста, в основу алгоритмов которых заложены различные способы идентификации, с целью выявления зависимости их характерных особенностей от выбранных идентификационных принципов. Приведённые материалы наглядно демонстрируют причины общности преимуществ и недостатков рассмотренных решений проблемы определения языковой принадлежности информации.

КЛЮЧЕВЫЕ СЛОВА

Идентификационный принцип, способ языковой идентификации; языковая идентификация; идентификация языка; определение языка; определение языковой принадлежности.

Kalegin S.N.

CJSC MNITI, Moscow, RF

THE IMPORTANCE OF CHOOSING THE MAIN IDENTIFICATION PRINCIPLE IN THE DESIGN OF LANGUAGE IDENTIFIER

ABSTRACT

The article presents the results of a comparative research of different types of texts language identification programs, based on different methods of identification, to identify the dependence of the characteristics of the selected identification principles. The given materials clearly demonstrate the reasons of community advantages and disadvantages of the examined solutions to the problem of language detection.

KEYWORDS

Identification principle; identification method; language identification; language detection.

В данной статье приводятся результаты сравнительного исследования разнотипных программ определения языковой принадлежности текста, в основу алгоритмов которых заложены различные способы идентификации, что позволило выявить существенные преимущества и недостатки как самих способов, так и их программной реализации [1]. Этот обзор позволяет выявить зависимость эффективности работы языковых определителей от идентификационных принципов и способов, заложенных в их основу. Результаты проведённого анализа помогут программистам улучшить существующие и сделать более совершенными разрабатываемые программы, а пользователям более осознанно подходить к их выбору в зависимости от целей применения.

На сегодняшний день существует множество языковых определителей, однако каждый из них обладает некоторыми недостатками, обусловленными характерными особенностями используемых способов идентификации, показанных на рисунке 1.

Как видно из приведённой блок-схемы [2], существуют различные способы и подходы к языковой идентификации неструктурированного текста, а в качестве определяющих язык элементов могут быть выбраны различные компоненты: символы, их сочетания или наборы n-грамм, служебные слова и частицы, значимые слова и т.д., от выбора и состава которых зависит эффективность и результативность процесса определения языка. Таким образом, выбор базового принципа, способа и типа идентифицирующих элементов обуславливает характерные особенности программ, в основу которых они положены, и существенно влияет на результаты их применения, что будет показано ниже.

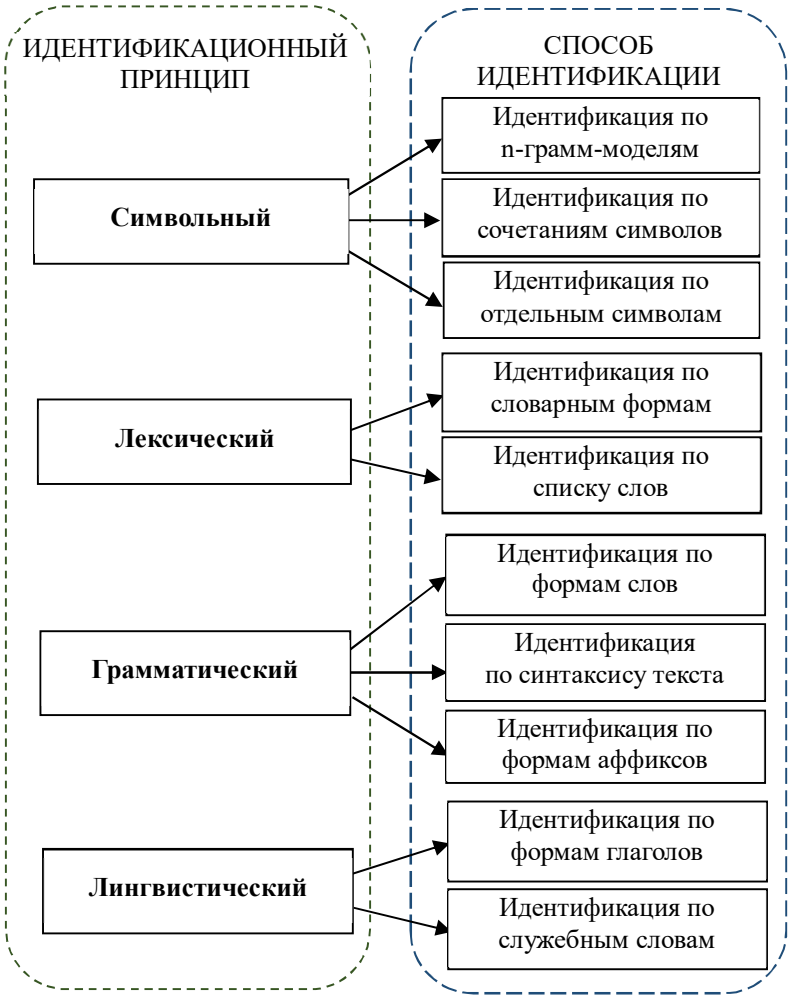

Рис. 1. Классификация способов языковой идентификации текста

При описании и сравнении языковых определителей рассматриваются существенные преимущества, недостатки, условия и особенности применения. Приводятся наиболее существенные результаты тестирования (с примерами), которые наглядно демонстрируют особенности работы каждой программы в зависимости от её способа идентификации.

Guesser [3]. Автоматический определитель языка текста российской компании «Фларус», позволяющий по нескольким введенным словам определить язык, на котором они написаны. Технически, идентификация языка в этой программе реализуется при помощи словарей всех определяемых языков. Введённый текст разбивается на слова, которые сравниваются со словами из базы определителя. При этом, в процессе идентификации участвуют не все слова, например, исключаются предлоги, союзы и слова длиной менее четырёх символов, так как они часто совпадают в различных языках. В качестве дополнительных идентификаторов языка могут использоваться характерные сочетания букв или символы национальных письменных систем. Определение языка текста с помощью данной программы происходит с некоторой ошибкой, которая зависит от особенностей языка и письменности, а также уникальности и количества слов загруженного текста. Поэтому результатом работы Guesser-а может являться список наиболее подходящих языков с большей или меньшей вероятностью. По заявлению разработчиков, для нормальной идентификации текст должен содержать не менее семидесяти символов или десяти слов.

Ниже даны примеры тестирования данной программы по фразам на кастильском (испанском) языке (Таблица 1).

Результаты определения языка фразы "**no lo creo**" (*я в это не верю*):

Таблица 1

№ п/п	Предполагаемый язык текста	Ассоциативная вероятность
1	испанский	8%
2	турецкий	8%
3	португальский	8%
4	норвежский	8%
5	датский	8%
6	финский	8%
7	немецкий	8%
8	нидерландский	8%
9	эстонский	8%
10	словацкий	8%
11	венгерский	8%
12	шведский	8%

Как видно из приведённого примера, язык практически не определён, так как вероятность ассоциации данного текста со всеми перечисленными языками одинакова. Видимо, причиной такого результата стали короткие слова, которые не были учтены при идентификации. Однако в таком случае, об этом программа должна была сообщить во избежание недоразумений. Кроме того, при таком подходе невозможно идентифицировать язык даже имея целые предложения. Например, фраза на кастильском **¿Que tal la vida?** (*Как жизнь?*) не будет анализироваться, так как полностью состоит из коротких слов. Это подтверждается приведённым ниже тестом (Таблица 2), где определителю передаётся фраза «**es mi casa**» (*это мой дом*), состоящая из коротких словарных форм.

Таблица 2

№ п/п	Предполагаемый язык текста	Ассоциативная вероятность
1	португальский	8%
2	норвежский	8%
3	датский	8%
4	финский	8%
5	немецкий	8%
6	нидерландский	8%
7	эстонский	8%
8	словацкий	8%
9	венгерский	8%
10	шведский	8%
11	испанский	8%
12	турецкий	8%

В третьем примере (Таблица 3) **Guesser**-у была передана фраза с более длинными словами на том же языке, содержащая словарную форму «**hombre**».

Результаты определения языка фразы "**yo no quiero hacerlo, hombre**" (*я не хочу это делать, друг*):

Таблица 3

№ п/п	Предполагаемый язык текста	Ассоциативная вероятность
1	венгерский	8%
2	шведский	8%
3	испанский	8%
4	турецкий	8%
5	португальский	8%
6	норвежский	8%
7	Датский	8%
8	Финский	8%
9	немецкий	8%
10	нидерландский	8%

11	эстонский	8%
12	словацкий	8%

Как видно, несмотря на длину слов, результат идентификации практически не изменился. Возможно, данная программа имеет нерационально ограниченный лексикон кастильского языка, что не позволило идентифицировать словарную форму «**hombre**».

В четвёртом тесте (Таблица 4) **Guesser**-у передаётся фраза «**Han elegido unos colores muy alegres para decorar la sala.**» (*Они выбрали несколько очень весёлых цветов для украшения зала*.):

Таблица 4

№ п/п	Предполагаемый язык текста	Ассоциативная вероятность
1	вьетнамский **uy**	34%
2	финский	5%
3	немецкий	5%
4	нидерландский	5%
5	эстонский	5%
6	словацкий	5%
7	венгерский	5%
8	шведский	5%
9	испанский	5%
10	турецкий	5%
11	португальский	5%
12	норвежский	5%
13	датский	5%

Здесь, несмотря на наличие во фразе длинных слов и словарных форм (**muy, para, sala, decorar**), наиболее вероятным языком текста был выбран вьетнамский, что абсолютно не соответствует действительности. Ошибка произошла из-за наличия в тексте сочетания «**uy**», которое было привлечено в качестве дополнительного идентифицирующего элемента. Это показывает несостоятельность выбранного разработчиками Guesser-а подхода для определения языковой принадлежности текста, а также ставит под сомнение рациональность применения способов идентификации языка по отдельным символам или их сочетаниям.

Как следует из результатов тестирования, программа подсчитывает количество совпадений слов разных языков со словами заданного текста, и выводит отчёт в виде названия одного или нескольких языков, которые были признаны наиболее подходящими. Такой результат в большинстве случаев оказывается практически бесполезным или приводит к критическим ошибкам в работе программных комплексов, где требуется точное определение языка (например, при автоматическом переводе или проверке орфографии текста). Причиной всех недостатков Guesser-а, которые выявляются при тестировании, является примитивный способ определения языка (использование словарей) и применённый подход в данной его реализации (без учёта коротких слов). Например, очевидно, что близкородственные языки различаться практически не будут, а транскрипция, транслитерация и направление письма в этой программе не учитываются в принципе, так же как и грамматические изменения слов или их взаимосвязи, что приводит к многочисленным ошибкам и вероятностным результатам. Кроме того, использованный способ не технологичен, так как в процессе анализа текста в память компьютера должно быть загружено и обработано множество словарей, что накладывает определённые технические ограничения. По этим причинам данный способ языковой идентификации без модификаций используется крайне редко и в чистом виде практически не встречается, а программы на его основе пишутся, чаще всего, учебные или тестовые.

Automatic language identifier [4]. Программа итальянского исследовательского центра *Translated Labs (T-Labs)* с web-интерфейсом, доступным удалённо любому пользователю Интернета.

Полиглот 3000 (П3000) [5]. Программа кипрской компании Likasoft, разработанная для применения в среде операционных систем семейства Windows.

TextCat [6]. Эта программа известна с середины 90-х годов прошлого века, а на её основе создано множество модификаций.

Language Identifier by Henrik Falck [7]. Программа Генриха Фалка, использующая Википедию (сетевую энциклопедию) как источник своего «обучения» и пополнения базы данных.

SILC (Système d'Identification de la Langue et du Codage) [8]. Система определения языка и кодировки текста от канадской лаборатории RALI.

Talenknobbel [9]. Тестовая версия голландского определителя языка.

В основе этих и им подобных программ языковой идентификации лежит статистический способ, суть которого заключается в создании n-мерного представления текста (модели векторного пространства) при помощи статистики встречаемости байтовых последовательностей (комбинаций символов или n-грамм), найденных в тексте как координаты или опорные точки. После их подсчёта программа составляет соответствующую модель и сравнивает её с результатами подобного моделирования справочных текстов, определяя язык заданного текста по максимальному соответствию одной из известных программе моделей, которая была заранее ассоциирована с конкретным языком.

Существует множество вариаций и реализаций данного способа идентификации и алгоритм работы в этом случае гораздо более сложный, нежели применённый в Guesser-е или подобных программах, так как основан на математических вычислениях и обработке большого объёма информации. Но несмотря на это, данный способ часто берётся за основу программистами из-за простоты его машинной реализации. Как и в предыдущем случае, такие программы не лишены недостатков в силу особенностей используемой методики идентификации. Символьный принцип, который положен в основу функционирования этих программ, опирается на характерные комбинации кодов символов (байтовых последовательностей), то есть на их написание, что порождает грубые ошибки при определении языка. Например, слово «быти (byti)», записанное кириллицей, программа относит к македонскому языку славянской группы, а это же слово, написанное латиницей, относится к корсиканскому романской группы языков. Примеры тестирования упомянутых программ приведены ниже с использованием тех же фраз (Таблицы 5-8).

Результаты определения языка фразы "**no lo creo**":

Таблица 5

№ п/п	Программа идентификации языка	Результат
1	**Automatic language identifier** (T-Labs)	**Spanish** (The text is too short!)
2	**Полиглот 3000**	Не распознан
3	**TextCat**	**italian**
4	**Language Identifier by Henrik Falck**	Text too short
5	**SILC**	**Italian**
6	**Talenknobbel**	Portuguese 51.47% Polish 48.33% Hawaiian 44.73% Tongonian 37.88% Italian 22.94% Spanish 3.57% English 3.48%

Результаты определения языка фразы "**yo no quiero hacerlo, hombre**":

Таблица 6

№ п/п	Программа идентификации языка	Результат
1	**Automatic language identifier** (T-Labs)	**Spanish** (The text is too short!)
2	**Полиглот 3000**	**Испанский** (Вероятность 71 %)
3	**TextCat**	**Spanish**
4	**Language Identifier by Henrik Falck**	**Spanish** or **French**
5	**SILC**	**Spanish**
6	**Talenknobbel**	Portuguese 62.1% English 44.31% Norwegian 40.32% Spanish 39.59% Polish 39.55% Dutch 38.73% Suid Afrikaans 33.08% Italian 32.09%

		French	29.02%
		Fries	28.3%
		Swedish	27.81%
		German	24.86%
		Latin	18.92%
		Hungarian	17.67%
		Tongonian	14.97%
		Hawaiian	6.44%
		Latvian	0.54%

Результаты определения языка фразы "**Han elegido unos colores muy alegres para decorar la sala.**":

Таблица 7

№ п/п	**Программа идентификации языка**	**Результат**	
1	**Automatic language identifier** (T-Labs)	**Spanish**	
2	**Полиглот 3000**	**Испанский** (Вероятность 14 %)	
3	**TextCat**	**Spanish**	
4	**Language Identifier by Henrik Falck**	**Spanish** or **Tagalog (Filipino)**	
5	**SILC**	**Spanish**	
6	**Talenknobbel**	Spanish	70.19%
		Portuguese	70.07%
		Swedish	67.78%
		French	67.47%
		Hungarian	66.46%
		Latin	64.45%
		Norwegian	59.56%
		English	59.05%
		Italian	57.26%
		Polish	56.39%
		Suid Afrikaans	55.76%
		Tongonian	54.3%
		Latvian	53.97%
		Finnish	51.37%
		Dutch	51.33%
		Indonesian	51.09%
		Fries	50.02%
		Hawaiian	46.76%
		German	43.05%

Как видно из данных примеров, языковая принадлежность коротких текстов (Таблица 5) при таком принципе идентификации не всегда определяется из-за недостатка материала для анализа (текст слишком короткий) или программа даёт бесполезные результаты. Например, **Talenknobbel** примерно с одинаковой вероятностью относит текст к португальскому, польскому или гавайскому языку, которые принадлежат разным генеалогическим группам. Также показательно, что вероятность ассоциации текста с португальским языком более 51 %, а с испанским только ~ 3,5 %, при том, что эти языки являются близкородственными. Тоже самое прослеживается в результатах следующего теста (Таблица 6), где вероятность ассоциации анализируемого текста с испанским, польским или датским языком практически одинакова. Результаты идентификации языка длинной фразы (Таблица 7) ещё более странные, так как вероятность определения программы **Полиглот 3000** упала до 14 %, а определитель языка Генриха Фалка с одинаковой вероятностью ассоциировал данный текст с испанским романской группы и тагальским языком, который относится к филиппинской группе австронезийской семьи, не имеющей прямого генеалогического родства с языками Европы. Всё это свидетельствует об

абсолютной абстракции результатов идентификации от лингвистических реалий, вследствие чего надёжность выбранного принципа определения языка текста вызывает сомнения.

Для полноты тестирования всем вышеназванным программам была передана фраза "**es mi casa**" (Таблица 8), состоящая из коротких слов, но результаты не очень отличаются от выданных программой **Guesser**.

Результаты определения языка по фразе "**es mi casa**":

Таблица 8

№ п/п	Программа идентификации языка	Результат
1	**Automatic language identifier** (T-Labs)	**Spanish** (The text is too short!)
2	**Полиглот 3000**	Не распознан
3	**TextCat**	**romanian**
4	**Language Identifier by Henrik Falck**	Text too short
5	**SILC**	**Spanish**
6	**Talenknobbel**	Latvian 57.9% Indonesian 43.99% Finnish 43.28% Hungarian 38.75% Swedish 34.63% Portuguese 30.88% Hawaiian 27.86% Latin 25.88% Tongonian 25.07% Spanish 20.25% Polish 20.01% French 12.66% English 11.16% Norwegian 4.97% Afrikaans 2.95% Fries 1.06%

В этом примере верно идентифицировать язык удалось только двум определителям из шести, что, вероятно, объясняется недостаточным объёмом анализируемого текста. Также показателен результат программы **Talenknobbel**, которая с вероятностью в 57 % сопоставила текст с латышским (балтийской группы), а следующими наиболее вероятными языками оказались индонезийский (австронезийской семьи) и финский (финно-угорской группы), что свидетельствует о неадекватности самой программы.

Таким образом, результаты работы программ, основанных на n-граммном принципе идентификации, во многих случаях будут практически бесполезны, так как не всегда определяется даже языковая группа и непонятно специалист какого направления (или какой программный модуль) потребуется для перевода заданного текста или к какой категории его отнести при автоматической сортировке корреспонденции, статей, публикаций или другой литературы, так как языки с близкими показателями вероятности часто даже неродственные. Более того, как следует из примеров, часто подобные программы выдают список наиболее подходящих языков с некоторой вероятностью. Такой результат, как сказано выше, в большинстве случаев, абсолютно неприемлем, например, при определении языка вводимого текста в текстовом процессоре, так как это вызовет тотальные ошибки в работе системы проверки орфографии.

Из всего выше сказанного следует, что подобные идентификаторы языка могут использоваться в строго ограниченных условиях и при небольшом количестве определяемых языков, так как требуют предварительного обучения или создания базы данных статистики использования n-грамм во всех идентифицируемых языках, а также опираются на коды (изображения) символов, что приводит к грубым ошибкам и даёт вероятностный результат. К тому же, подобные системы трудно назвать технологичными, так как их ресурсоёмкость далеко не всегда целесообразна.

PHPLangautodetect [10]. Эта программа представляет собой библиотеку (модуль) на языке PHP, в основу которой положен подсчёт статистики использования символов «национального» алфавита или традиционной письменности. Программа использует разновидности символьного способа языковой идентификации текста, суть которого состоит в различении языков по специфическим символам или их сочетаниям, характерным для литературы на каждом конкретном

языке. Также в процессе идентификации может использоваться набор заранее сформированных правил, которые устанавливают идентичность при помощи уникальных или типичных для текстов на конкретных языках букв (например, «**ъ**» и «**ё**» для русского или "**є**" для украинского языка).

На сегодня эта программа способна определять три языка: русский, украинский или английский, а её работоспособность гарантируется только при анализе орфографически правильно написанных длинных текстов с использованием традиционных письменных систем. По этим причинам она представляет больше академический интерес, чем практический, а её тестирование не имеет смысла в рамках данного исследования.

Гибридные системы

Гибридными называются те системы, в которых комбинируются известные принципы определения языковой принадлежности текста с целью улучшения эффективности, причём их роли в идентификации языка примерно равнозначные. Ниже приводятся примеры гибридных систем с указанием их преимуществ и недостатков, а также результаты тестирования.

К таким системам относится определитель языка от американской компании Microsoft, подключённый к программному комплексу автоматического перевода текстов **BingTranslator**, которая доступна удалённо через web-интерфейс [11]. Этот определитель использует комбинации различных принципов и способов идентификации, включая символьные и словарные.

Примеры тестирования данной программы по тем же фразам:

Таблица 9

№ п/п	**Анализируемая фраза**	**Результат**
1	no lo creo	Испанский
2	yo no quiero hacerlo, hombre	Испанский
3	Han elegido unos colores muy alegres para decorar la sala.	Испанский
4	es mi casa	Испанский

Как следует из результатов, язык всех тестируемых фраз определён верно, что свидетельствует о существенном превосходстве данного типа определителей по сравнению с предыдущими. Однако при попытке идентификации коротких текстов и отдельных слов программа начинает давать сбои. Например, при вводе буквы «**ё**» сразу происходит ассоциация с русским языком, хотя это может быть текст на кумыкском или другом подобном языке, а такие слова как «**in**» или «**of**» автоматически ассоциируются с английским, что не всегда соответствует действительности.

Кроме того, ресурсоёмкость таких программ, обусловленная комбинированием идентификационных процессов, требует использования мощных аппаратных комплексов и не позволяет автономно подключать их к бытовым системам, что существенно ограничивает сферу применения.

Xerox Language Identifier [12]. Определитель языка текста от американской компании Xerox, доступен для тестирования через сайт проекта “Open Xerox”. Определяет язык заданного текста с помощью коротких слов и n-грамм, что позволяет отнести его к гибридным системам.

Примеры определения языка данной программой приведённых выше фраз на кастильском языке показаны ниже (Таблица 10).

Таблица 10

№ п/п	**Анализируемая фраза**	**Результат**
1	no lo creo	**Español** (Spanish; Castilian)
2	yo no quiero hacerlo, hombre	**Español** (Spanish; Castilian)
3	Han elegido unos colores muy alegres para decorar la sala.	**Español** (Spanish; Castilian)
4	es mi casa	**Català** (Catalan; Valencian)

Тестирование показывает, что комбинация нескольких принципов идентификации и качество алгоритма работы программы существенно улучшают результаты. В данном случае способ определения языка с помощью коротких слов является удачным дополнением к идентификации по наборам комбинаций символов (n-грамм), что дало высокую вероятность и несравнимо лучший результат по отношению к применению скомбинированных принципов в отдельности. Однако это же обстоятельство увеличивает количество недостатков программы, присущих всем использованным способам, таким как: вероятностный результат, необходимость

набора статистики встречаемости n-грамм и создания языковых моделей, зависимость от письменной системы и т.д. Это наглядно показывает попытка определения языка кастильской фразы «**es mi casa**» (*это мой дом*), которую данный определитель ассоциирует с каталонским языком Валенсии.

Более того, к программе добавляются и все атрибуты, необходимые для идентификации каждым скомбинированным способом, то есть словари, списки коротких слов, таблицы характерных символов и языковые модели, что усложняет весь процесс идентификации и увеличивает время её проведения.

Eureka Engine [13]. Это система лингвистического анализа текстов, в составе которой имеется модуль автоматического определения языка LanguageDetector, объединяющий несколько принципов идентификации, в том числе символьный, словарный и морфологический.

Примеры тестирования данного модуля приведены в таблице 11.

Таблица 11

№ п/п	Анализируемая фраза	Результат
1	no lo creo	Испанский
2	yo no quiero hacerlo, hombre	Испанский
3	Han elegido unos colores muy alegres para decorar la sala.	Испанский
4	es mi casa	Сербский

Как следует из тестов, идентификация даёт очень высокую вероятность определения. Однако фразу на кастильском «**es mi casa»** данная программа ассоциирует с сербским языком, что выявляет несовершенство анализирующего алгоритма. Промежуточные результаты нормализации словоформ (Таблица 12) и морфологического анализа (Таблица 13) этой фразы приведены ниже.

Нормализация словоформ

Таблица 12

Слово	Нормальная форма	Часть речи
Es	Es	Существительное
mi	Mi	Существительное
casa	Casa	Существительное

Морфологический анализ

Таблица 13

Слово	Морфологические атрибуты и словоформы	
Es	Основа слова:	-
	Падеж:	-
	Число:	-
	Род:	-
	Лицо:	-
	Залог:	-
	Время:	-
	Переходность:	-
	Часть речи:	Существительное
	Возможные словоформы:	
mi	Основа слова:	mi
	Падеж:	-
	Число:	единственное
	Род:	-
	Лицо:	-
	Залог:	-
	Время:	-
	Переходность:	-
	Часть речи:	Существительное
	Возможные словоформы:	Mi, Mis, mi, mis
casa	Основа слова:	Casa
	Падеж:	-

	Число:	единственное
	Род:	-
	Лицо:	-
	Залог:	-
	Время:	-
	Переходность:	-
	Часть речи:	Существительное
	Возможные словоформы:	Casa, Casas

Из приведённых таблиц видно, что ошибки были допущены на промежуточных этапах морфологического разбора текста, что привело к неприемлемым результатам. Этот факт показывает, что избыточное усложнение процесса идентификации может отрицательно повлиять на результаты, а значит, при улучшении алгоритма определителя должен поддерживаться баланс между сложностью программы, качеством результатов и целесообразной ресурсоёмкостью.

По итогам тестирования рассмотренных программ языковой идентификации следует отметить, что все они так или иначе зависят от письменной системы, применяемой в анализируемом тексте, а также правил грамматики и орфографии каждого анализируемого языка, что обуславливает невозможность их применения при записи текста нетрадиционным или несовременным способом. Кроме того, для верной идентификации определителю часто требуется существенно больше анализируемого материала, чем декларируют разработчики, что ставит под сомнение заявляемые ими преимущества рассмотренных способов, программ или систем. Следовательно, в настоящее время не существует универсальных программ определения языковой принадлежности текста. Каждая конкретная программа имеет свои преимущества и недостатки, обусловленные особенностями принципов и способов идентификации, положенных в основу её алгоритма. Это демонстрирует прямую зависимость основных характеристик языкового определителя от выбора идентифицирующей технологии, что обязательно должно учитываться при разработке подобной программы или системы для конкретного применения, поскольку неверное решение непременно отразится на результате. Например, для использования в мобильных компьютерах и веб-приложениях нерационально выбирать ресурсоёмкие способы и системы, которые существенно замедляют процесс языковой идентификации текста при ограниченных ресурсах или даже делают его нецелесообразным, а к системам каталогизации или автоматической сортировки текстов не имеет смысла подключать внешние комплексы идентификации, доступные только через сеть, поскольку это существенно снижает надёжность работы и гибкость настройки, что в свою очередь отражается на общей функциональности системы.

Приведённые материалы позволяют выявить причины общности недостатков рассмотренных программ. В большинстве случаев эти причины, обусловлены несовершенством алгоритма программы и не его технической реализацией, а выбором способов языковой идентификации. Поэтому при разработке программ важно стремиться не только к совершенству алгоритмов и технологических решений, но и обращать внимание на их логическую основу, от которой зависят результаты работы и условия применения конечного программного продукта.

Приведённые результаты сравнительного исследования наглядно иллюстрируют современное состояние проблемы определения языка неструктурированного текста и демонстрируют несовершенство существующих методов идентификации, а также реализующих их программ и систем. Отсюда следует необходимость дальнейшего изучения рассмотренной проблемы как с лингвистикой, так и с технической точек зрения, для чего требуются дополнительные исследования на данную тему и поиск новых решений.

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. Калегин С.Н. Экспериментальное исследование возможности автоматизации процесса языковой идентификации текста. / Международная конференция "CNews FORUM 2016: Информационные технологии завтра", Москва, 2016.
2. Калегин С.Н. Способы определения языковой принадлежности неструктурированного текста в мультиязычной информационной среде. / Международная конференция «CONCORT-2016», Нижний Новгород, 2016.
3. Автоматический определитель языка текста [Электронный ресурс]. URL: “Guesser.ru” http://guesser.ru/.
4. Automatic language identifier (Автоматический определитель языка) [Электронный ресурс]. URL: http://labs.translated.net/.

5. Автоматический определитель языка текста Полиглот 3000 (П3000) [Электронный ресурс]. URL: http://www.polyglot3000.com/.
6. Программа TextCat [Электронный ресурс]. URL: http://odur.let.rug.nl/~vannoord/TextCat/.
7. Language Identifier by Henrik Falck [Электронный ресурс]. URL: http://whatlanguageisthis.com/.
8. SILC RALI [Электронный ресурс]. URL: http://rali.iro.umontreal.ca/rali/.
9. Автоматический определитель языка Talenknobbel [Электронный ресурс]. URL: http://www.fuzzums.nl/~joost/talenknobbel/
10. Лозовюк А. PHPLangautodetect [Электронный ресурс]. [Электронный ресурс]. URL: http://abrdev.com/; сайт проекта: http://code.google.com/p/phplangautodetect/.
11. MS BingTranslator [Электронный ресурс]. URL: http://www.microsofttranslator.com/.
12. Open Xerox Language Identifier [Электронный ресурс]. URL: http://open.xerox.com/Services/LanguageIdentifier/
13. EUreka ENgine [Электронный ресурс]. URL: http://eurekaengine.ru/, http://palitrumlab.ru/products.

References

1. Kalegin S.N. Jeksperimental'noe issledovanie vozmozhnosti avtomatizacii processa jazykovoj identifikacii teksta. / Mezhdunarodnaja konferencija "CNews FORUM 2016: Informacionnye tehnologii zavtra", Moskva, 2016.
2. Kalegin S.N. Sposoby opredelenija jazykovoj prinadlezhnosti nestrukturirovannogo teksta v mul'tijazychnoj informacionnoj srede. / Mezhdunarodnaja konferencija «CONCORT-2016», Nizhnij Novgorod, 2016.
3. Avtomaticheskij opredelitel' jazyka teksta [Jelektronnyj resurs]. URL: “Guesser.ru” http://guesser.ru/.
4. Automatic language identifier (Avtomaticheskij opredelitel' jazyka) [Jelektronnyj resurs]. URL: http://labs.translated.net/.
5. Avtomaticheskij opredelitel' jazyka teksta Poliglot 3000 (P3000) [Jelektronnyj resurs]. URL: http://www.polyglot3000.com/.
6. Programma TextCat [Jelektronnyj resurs]. URL: http://odur.let.rug.nl/~vannoord/TextCat/.
7. Language Identifier by Henrik Falck [Jelektronnyj resurs]. URL: http://whatlanguageisthis.com/.
8. SILC RALI [Jelektronnyj resurs]. URL: http://rali.iro.umontreal.ca/rali/.
9. Avtomaticheskij opredelitel' jazyka Talenknobbel [Jelektronnyj resurs]. URL: http://www.fuzzums.nl/~joost/talenknobbel/.
10. Lozovjuk A. PHPLangautodetect [Jelektronnyj resurs]. [Jelektronnyj resurs]. URL: http://abrdev.com/; sajt proekta: http://code.google.com/p/phplangautodetect/ .
11. MS BingTranslator [Jelektronnyj resurs]. URL: http://www.microsofttranslator.com/.
12. Open Xerox Language Identifier [Jelektronnyj resurs]. URL: http://open.xerox.com/Services/LanguageIdentifier/.
13. EUreka ENgine [Jelektronnyj resurs]. URL: http://eurekaengine.ru/, http://palitrumlab.ru/products.

Поступила 12.10.2016

Об авторе:

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