УДК 004 DOI 10.25559/SITITO.2017.4.492

Калегин С.Н.

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

ЯЗЫКОВАЯ ИДЕНТИФИКАЦИЯ ИНФОРМАЦИОННЫХ БЛОКОВ НА ОСНОВЕ ЛЕКСИКО-ГРАММАТИЧЕСКИХ МАРКЕРОВ

Аннотация

Данная статья является продолжением серии публикаций автора по тематике языковой идентификации текстов. Рассматривается создание технологической основы для систем языковой идентификации неструктурированных информационных блоков на основе лексико-грамматических маркеров, в качестве которых используются формы глаголов, отглагольных образований или функционально аналогичных конструкций, даётся описание предлагаемого способа и алгоритма его программной реализации. Данные разработки позволят существенно уменьшить ресурсоёмкость и улучшить качество подобных систем, что даст значительный экономический эффект и возможность создания принципиально новых технологий для определения языковой принадлежности информации в многоязычной среде. Следовательно, работа представляет интерес для компьютерных лингвистов и разработчиков систем автоматической обработки текстов, таких как: системы глобального мониторинга, многоязычные базы знаний, автоматические системы перевода, информационно-поисковые комплексы, системы реферирования документов, каталогизаторы литературы и т.д.

Ключевые слова

Языковая идентификация текста; определение языковой принадлежности текста; технология языковой идентификации; языковой определитель; лексико-грамматический маркер.

Kalegin S.N.

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

LANGUAGE IDENTIFICATION OF INFORMATION BLOCKS BASED ON LEXICO-GRAMMATIC MARKERS

Abstract

This article is a continuation of the author's series of publications on the subjects of language identification of texts. In the article is being considered the creation of a technological basis for language identification systems of unstructured information blocks based on lexico-grammatical markers, in which are used the forms of verbs, verbal formations or functionally analogous constructions, are described method and algorithm for its software implementation. These developments will significantly reduce the resource intensity and improve the quality of such systems, which will give a significant economic effect and the possibility of creating fundamentally new technologies for determining the linguistic affiliation of information in a multilingual environment. Consequently, the study is of interest for computer linguists and developers of automatic word processing systems, such as: global monitoring systems, multilingual knowledge bases, automatic translation systems, information retrieval systems, document summarizing systems, literature catalogers, etc.

Keywords

Language identification of the text; language identification technology; language identifier; lexico-grammatic marker.

Введение

В последние годы резко возросли потоки информации во всех сферах человеческой деятельности и появилось множество технических средств для её передачи, фиксации, хранения и обработки. С ростом экономических, политических и культурных связей между государствами, а также развитием коммуникационных систем, объёмы этой информации быстро увеличиваются, а доступ к ней упрощается, что позволяет создавать многоязычные системы глобального поиска данных. Для хранения и обмена данными чаще всего используется неструктурированный текст, который является основным средством фиксации информации в современном мире. Такая информация может быть представлена в любом стиле, на любом языке и с помощью любой письменности (письменной системы), что создаёт существенные проблемы при попытке её обработки, особенно если это делается компьютерами в автоматическом режиме. Поэтому сегодня актуальна разработка интеллектуальных систем обработки неструктурированного текста, которые используются в программных комплексах машинного перевода, электронных базах знаний, различных поисковых и мониторинговых системах, комплексах сбора и обработки данных, каталогизаторах, системах обработки публикаций и корреспонденции, и т.д. Одной из основных проблем в подобных системах является определение языка, который был использован при фиксации информации в конкретном тексте, так как без этого его дальнейшая обработка практически невозможна (не считая статистических операций).

В данной статье представлены разработки автора, призванные решить проблему языковой идентификации неструктурированных текстовых блоков в автоматическом режиме, которые могут использоваться при создании технологии определения языковой принадлежности неструктурированного текста для соответствующих систем. Особенностью этих разработок является дуалистический подход к решению задач, представляющий собой комбинацию лингвистических и технических аспектов, что делает рассматриваемую работу интересной как техническим специалистам, так и языковедам, работающим в области автоматизации обработки неструктурированной информации.

Технологическую основу могут составить следующие разработки:

- перечень лексико-грамматических маркеров, составляющих синтаксическую основу текста, а также их представления в текстах различных тематик (на примере европейских языков) и выявленные статистические закономерности их употребления;
- способ определения языка или языковой группы текста по указанным маркерам, позволяющий не только верно определить язык, но и языковую группу в случае, если однозначно идентифицировать язык анализируемого текста не удаётся;
- принципы построения набора идентифицирующих элементов – идентификационного набора – и идентификационной матрицы (фильтра), созданной на его основе, позволяющие существенно сократить количество ресурсов, затрачиваемых на языковую идентификацию текста, а также воспроизвести иерархию матрицы в соответствии с нужной градацией языков (генеалогической, типологической и т.д.), что даёт возможность определять принадлежность языка анализируемого текста к конкретной группе или типу, причём на любом уровне иерархии;
- алгоритмы реализации указанных разработок в программном обеспечении.

Синтаксическая основа текста

Выделение ключевых элементов, обычно, составляет первый этап обработки текста [1]. В их качестве могут выступать слова, словосочетания и конструкции, несущие базовую смысловую нагрузку, составляющие информационную основу и представляющие особую значимость для понимания сути данного текста. Выбор ключевых элементов – это основной процесс смыслового свёртывания материала при анализе и компрессии первичного документа, например, при реферировании и других аналогичных операциях. Такие элементы считаются опорными для семантики, так как составляют логическую основу текста, их нельзя заменить без искажения сути, а их устранение приводит к её исчезновению [2], однако проводить языковую идентификацию на основе таких элементов не всегда рационально вследствие их статичности и совпадения в разных языках.

На основе проведённых исследований автором выявлены тематиконезависимые синтаксические маркеры, позволяющие не только выделить грамматическую основу текста,

но и идентифицировать язык, который был использован для фиксации информации в нём. Для их выявления проведён глубокий историко-лингвистический анализ, результаты которого показали универсальность выбранного подхода и верность направления работы, обусловленную особенностями развития человека как в антропологическом, так и биологическом плане.

По мнению языковедов, ещё в протоязыке выделились такие категории, как «существительное» и «глагол» или «подлежащее», «сказуемое» и «дополнение», которые имеют чётко выделяемое универсальное «ядро» (основу) и более размытую «периферию» [3]. Существительные и именные группы прототипически обозначают дискретные, стабильные во времени объекты, в то время как глаголы обозначают положения вещей, подверженные изменению в пространстве и времени. Следовательно, конституирующим ядром предложения является не обозначение действия или указание на него, а та составляющая, которая содержит информацию о времени, модальности и т. п. категориях. Чаще всего это глагольные и отглагольные формы, но в некоторых языках такая информация может содержаться в именной группе или специальных конструкциях.

К похожим выводам приходят и специалисты по детской речи, так как на первом этапе формирования представлений о языке ребенку важно усвоить слова-глаголы (слова-действия), которые играют ключевую роль в процессе развития коммуникативной функции. Если не акцентировать внимание на появлении этих простых слов в активном словаре ребёнка, то он ещё долго будет употреблять в общении ограниченное количество слов, обозначающих некоторые обиходные предметы из его окружения, и обращения к близким людям [4, 5], а его речь будет бессвязной.

Ещё одним доказательством в пользу основополагающей функции глагола (или семантически аналогичных образований) в тексте является наличие во многих языках однословных самодостаточных глагольных предложений, не требующих логического подлежащего, например:

Дождит. Снежит. Светает.

Результаты исследований специалистов, изучавших формирование языка на ранних стадиях развития, показывают, что глаголы, отглагольные образования или их функциональные аналоги, являются основой любого предложения, а, следовательно, и ключевыми элементами текста. Таким образом, их можно использовать в качестве идентификационных или опорных маркеров, например, при синтаксическом разборе, грамматическом или стилистическом анализе, определении языковой принадлежности текста и т.д., так как без таких элементов связных текстов практически не встречается. Надёжность таких маркеров обусловлена их разнообразием и характерными особенностями, связанными с необходимостью выражения модально-темпоральных отношений, присущих каждому языку и диалекту, что делает их универсальными идентификаторами языка и связующими звеньями любого текста. Без глагольных форм и отглагольных образований (или аналогичных конструкций) тексты утрачивают синтаксические связи и перестают существовать, а на их месте образуется набор слов и сочетаний, относящийся к тематике текста.

Следует отметить, что наиболее употребительными являются формы инфинитивов, причастий и деепричастий, а также личные формы третьего лица настоящего и прошедшего времён. Это доказывает определённую универсальность таких маркеров и закономерность их использования в разных языках и диалектах. А их количество практически постоянно и мало зависят от языка или тематики текста, что иллюстрируется следующими подсчётами.

Русский язык

Среднее количество глагольных и отглагольных форм в текстах ~ 16,9 %. Из них ~ 3 % относятся к 10 наиболее частотным глаголам, что составляет ~ 17,5 % глагольных и отглагольных форм текста.

Кастильский язык

Среднее количество глагольных и отглагольных форм в текстах ~ 16,6 %. Из них ~ 5,8 % относятся к 10 наиболее частотным глаголам, что составляет ~ 34,6 % глагольных и отглагольных форм текста.

Немецкий язык

Среднее количество глагольных и отглагольных форм в текстах ~ 19,8 %. Из них ~ 5,8 % относятся к 10 наиболее частотным глаголам, что составляет ~ 29,3 % глагольных и отглагольных форм текста.

Подсчёты показывают, что среднее количество глагольных и отглагольных форм в текстах на разных языках является общей статистической закономерностью, а использование наиболее употребительных глаголов (и производных от них) практически постоянно для конкретного языка, что позволяет уменьшить количество синтаксических маркеров при обработке текстов

достаточного объёма.

По итогам проведённых исследований можно заключить, что в среднем содержание глагольных и отглагольных форм в общем количестве слов составляет ~ 18 % и несущественно зависит от тематики или языка текста, а значит, такие формы могут использоваться как универсальные лексико-грамматические маркеры, составляющие синтаксическую основу текста. Кроме того, число подобных форм в тексте очевидно превышает количество других маркеров, что существенно понижает сложность и ресурсоёмкость анализа, а также требует меньше анализируемого материала.

Языковая идентификация на основе лексико-грамматических маркеров

Разработанный автором способ языковой идентификации [6] позволяет по-новому посмотреть на проблему определения языковой принадлежности неструктурированного текста и открывает два направления дальнейших изысканий: 1) выявление ключевых маркеров (и их грамматических форм) текста на каждом языке, однозначно его идентифицирующих, и 2) сужение спектра подходящих языков при идентификации до группы или подгруппы, а также выявление определяющих элементов для каждой из них. Решение этих задач позволит значительно продвинуться в совершенствовании технологий языковой идентификации вообще и неструктурированного текста в частности.

В качестве основных идентификационных маркеров выбраны формы глаголов отглагольных образований и предикатов, которые выражают модально-темпоральные отношения и составляют семантико-грамматическую основу текста на большинстве языков мира. Основное отличие данного способа от аналогов заключается в использовании статистически наиболее употребительных форм глаголов, отглагольных образований, предикативов или изменяемых форм (частей) предикатов в качестве ключевых элементов идентификационной матрицы, которая применяется для фильтрации анализируемых текстов. А значит, таким образом, может быть идентифицирован любой язык, основной частью речи или основным связующим элементом предложений которого является глагол или функционально аналогичная составляющая предиката, так как подобные формы присутствуют в большинстве синтагм и высказываний. Наглядное представление программной реализации данного способа на примере определения славянских языков по формам 3-х групп глаголов, отглагольных образований и предикативов показано на рисунке 1 (приводится в сокращении).

Для идентификации языка или языковой группы данным способом производятся следующие основные операции.

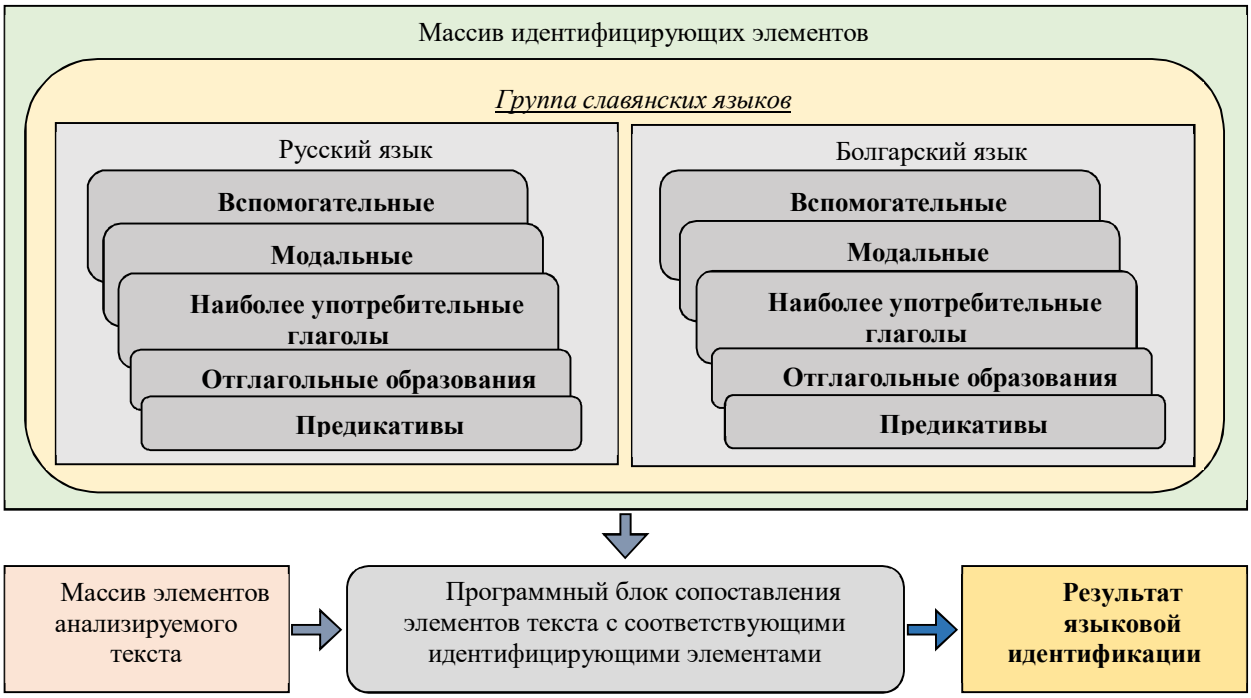

Рис. 1. Блочная схема программной реализации способа языковой идентификации текста по формам глаголов, отглагольных образований и основ предикатов

- создаётся массив идентифицирующих элементов – идентификационная матрица – на основе имеющегося идентификационного набора;
- формируется массив слов или других элементов анализируемого текста;
- сопоставляются элементы анализируемого текста с элементами идентификационной матрицы, в которой хранится набор идентификационных маркеров и их соответствия определённым языкам и/или группам;
- результат идентификации фиксируется и выдаётся на устройство вывода.

Для применения данного способа определения языковой принадлежности текста составление идентификационных наборов с указанием на соответствие конкретному языку или языковой группе (а также с другими нужными индикаторами) является необходимым и достаточным условием [7]. Эти наборы могут быть составлены как вручную, так и с помощью компьютера в автоматическом или автоматизированном режиме, и храниться в любой форме, например, в виде простых текстовых файлов, таблиц или массивов, что позволяет оперативно вносить изменения в такие наборы.

Идентификационный набор может быть представлен в нескольких формах и различных вариациях. Например, в виде списка с определённой структурой или графа (схемы), где имеется одна или несколько групп идентифицирующих элементов каждого определяемого языка, указана связь этих групп с конкретным языком, типом и языковой группой (и/или подгруппой), а также языковой ветвью, семьёй и т.д. по мере необходимости. Конкретная схема деления зависит от выбранной классификации языков и требуемых ассоциаций. При этом структура набора может быть линейной, в которой просто перечисляются языки или группы с их идентификаторами, или иерархической, где градация может быть разветвлённая и многоуровневая, а для каждой языковой группы и отдельных языков могут даваться уточнения или более подробная языковая классификация, например, деление на подгруппы, варианты и/или диалекты.

Представленный способ позволяет идентифицировать как естественные языки, такие как: русский, немецкий, английский, кастильский, латинский и т.д., так и созданные искусственно, как например: волапюк, эсперанто, идо, интерлингва и т.д. Более того, в тех случаях, когда идентифицировать язык не удаётся, этот способ позволяет определять генеалогическую, типологическую или другую группу, в которую он входит согласно заданной классификации. При необходимости идентификационная матрица может быть дополнена формами вспомогательных и модальных глаголов, а также других ключевых синтаксических элементов и компонентов предиката, не вошедших в первоначальный список идентификационных маркеров.

Техническая реализация данного способа языковой идентификации в программном обеспечении может осуществляется в простых линейных алгоритмах последовательного сопоставления анализируемых элементов текста с языковыми маркерами, хранящимися в идентификационной матрице. Ресурсоёмкость таких программ минимальная вследствие простоты алгоритма и краткости машинного кода, а также количества требуемых операций и языковых маркеров, а эффективность высокая за счёт применения лингвистического подхода, положенного в основу идентификации.

Заключение

В результате проведённых исследований и анализа итоговых выводов предложены оригинальные решения в сфере языковой идентификации, имеющие как практическую, так и теоретическую значимость. На основе сделанных разработок может быть создана новая технология языковой идентификации неструктурированного текста.

Согласно проведенным расчётам и тестам предложенный способ является более универсальным, эффективным и технологичным по сравнению с применяемыми сегодня и позволяет существенно улучшить результаты по ряду показателей. Среди преимуществ этого способа можно выделить следующие.

Возможность работы с многоязычными текстами и верного определения всех языков, используемых в анализируемом тексте, при наличии в нём форм из идентификационной матрицы.

Возможность верного определения типа или генеалогической семьи, ветви или группы языков, к которой относится язык анализируемого текста (например, славянская, германская, романская, кельтская и т.д.).

Возможность идентифицировать язык по грамматическим формам и/или их семантически значимым частям (основам или корням) небольшой группы глаголов, например,

вспомогательных, модальных и т.д., или комбинации таких групп, а также отглагольным образованиям или их аналогам.

Независимость от системы письма или представления информации в анализируемом тексте.

Существенное повышение точности идентификации языка при небольших объёмах текста и исключение вероятностных результатов.

При использовании компьютера имеется возможность обойтись без сложных алгоритмов и мощных вычислительных средств.

Возможность регулирования функциональности, точности определения и скорости работы с помощью расширения и уточнения или сокращения и упрощения идентификационных наборов.

Не требуется использования словарей определяемых языков и баз данных, а также грамматического анализа, создания модели языка или дерева грамматических зависимостей, сбора статистики по использованию комбинаций символов и т.д., что позволяет существенно сократить количество выделяемых на обработку ресурсов, а также ускорить и упростить её.

Текст может быть представлен в любой воспринимаемой компьютером или человеком форме (например, в виде изображений символов, комбинаций точек шрифта Брайля и т.д. с применением одной из известных письменных систем, а также передан в виде блока (набора) сигналов, например, звуковых волн, азбуки Морзе и т.п.), что делает предлагаемый способ более универсальным.

Количество идентифицирующих элементов и операций сравнения при реализации данного способа в разы меньше, чем при использовании словарей, лингвистических моделей или последовательностей символов (байтовых последовательностей) в наиболее употребительных сегодня технологиях.

Технический результат применения разработанного способа языковой идентификации текста позволяет повысить эффективность процесса языковой идентификации и получить существенный экономический эффект, а его применение как технологической основы для языковой идентификации позволит создать принципиально отличные от существующих системы и программные модули, которые позволят вывести процесс определения языковой принадлежности текста на новый качественный уровень.

Благодарности

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

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

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

Литература

1 Черновалюк И.В. Реферирование научного текста // Методические рекомендации по русскому языку для иностранных студентов и аспирантов Часть II. – Одесса, 2012. – 47 с.
2 Пузырёв А.В. О разграничении понятий «опорные и ключевые элементы художественного текста» // Актуальные проблемы филологии и педагогической лингвистики, № 16. Владикавказ, 2014. – с. 308-317.
3 Бурлак С. А. Происхождение языка: Факты, исследования, гипотезы. М: Астрель, 2011. – 464 с.
4 Громова О.Е. Алло! Ляля? Речь. Первые глаголы. Для детей до 2 лет / М.: Карапуз, 2003. – 19 с.
5 Янушко Е.А. Развитие речи. Первые глаголы. Для детей от 1 года / М.: Эксмо, 2011. – 16 с.
6 Патент РФ № 2607989, МПК G06F 17/27. Способ автоматизированного определения языка или языковой группы текста / Калегин С.Н.; ЗАО «МНИТИ»; заяв. 08.07.2015; опубл. 11.01.2017, бюл. № 2.
7 Калегин С.Н. Логическая структура идентификационных наборов в системах языковой идентификации. / Современные информационные технологии и ИТ-образование. Том 12, № 1. – Москва, 2016. – С. 173-181.

References

1 Chernovalyuk I.V. Referirovanie nauchnogo teksta // Metodicheskie rekomendatsii po russkomu yazyku dlya inostrannykh studentov i aspirantov CHast' II. – Odessa, 2012. – 47 s.
2 Puzyryov A.V. O razgranichenii ponyatij «opornye i klyuchevye ehlementy khudozhestvennogo teksta» // Aktual'nye problemy filologii i pedagogicheskoj lingvistiki, № 16. Vladikavkaz, 2014. – s. 308-317.
3 Burlak S. A. Proiskhozhdenie yazyka: Fakty, issledovaniya, gipotezy. M: Astrel', 2011. – 464 s.
4 Gromova O.E. Allo! Lyalya? Rech'. Pervye glagoly. Dlya detej do 2 let / M.: Karapuz, 2003.
5 YAnushko E.A. Razvitie rechi. Pervye glagoly. Dlya detej ot 1 goda / M.: Eksmo, 2011.
6 Patent RF № 2607989, MPK G06F 17/27. Sposob avtomatizirovannogo opredeleniya yazyka ili yazykovoj gruppy teksta / Kalegin S.N.; ZAO «MNITI»; zayav. 08.07.2015; opubl. 11.01.2017, byul. № 2.
7 Kalegin S.N. Logicheskaya struktura identifikatsionnykh naborov v sistemakh yazykovoj identifikatsii. / Sovremennye informatsionnye tekhnologii i IT-obrazovanie. Tom 12, № 1. – Moskva, 2016. – S. 173-181.

Поступила: 1.10.2017

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

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