Исследования и разработки в области новых информационных технологий и их приложений

УДК 004.051

Калегин С.Н.

Московский научно-исследовательский телевизионный институт, г. Москва, Россия
Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, Россия

МЕТОДИКА ОЦЕНКИ РЕЗУЛЬТАТОВ ОПРЕДЕЛЕНИЯ ЯЗЫКОВОЙ ПРИНАДЛЕЖНОСТИ ИНФОРМАЦИИ

Аннотация

В статье приводится авторская методика оценки результатов определения языковой принадлежности информации, разработанная в ходе экспериментальных исследований и показывающая эффективность соответствующих способов, технологий, алгоритмов и программного обеспечения, а также недостатки существующих подходов к решению данной проблемы. Это позволяет оценивать результативность программ и систем языковой идентификации на этапе проектирования, что существенно сокращает ресурсозатраты на их разработку.

Ключевые слова

Языковая идентификация; определение языковой принадлежности; определение языка; методика оценки результатов языковой идентификации; результат языковой идентификации.

Kalegin S.N.

Moscow Research TV Institute Joint Stock Company, Moscow, Russia
Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia

ESTIMATION METHODOLOGY OF THE LANGUAGE IDENTIFICATION RESULTS

Abstract

The article presents the author's methodology for evaluating the language identification results, developed in the course of experimental research and showing the effectiveness of appropriate methods, technologies, algorithms and software, as well as the shortcomings of existing approaches to solving this problem. This allows to evaluate the effectiveness of language identification programs and systems at the design stage, which significantly reduces the resource costs for their development.

Keywords

Language identification; language detection; estimation (evaluating) methodology of the language identification results; result of language identification.

Введение

В связи с развитием электронно-вычислительных машин, компьютерных сетей и коммуникационных систем проблема языковой идентификации информации становится всё более актуальной. Процесс глобализации создаёт необходимость международного общения во всех сферах жизнедеятельности, что ведёт к возникновению многоязычной информационной среды, где информация существует на множестве языков в неструктурированном виде. Последнее обстоятельство требует разработки средств и способов автоматизации поиска, упорядочивания и градации данных во всём информационном пространстве в зависимости от тематики и запросов общества или отдельных пользователей,

на каком бы языке они не вводились. С целью упрощения международной и межъязыковой коммуникации созданы специальные программы и устройства-переводчики, которые призваны помочь человеку в работе с информацией на различных языках. Кроме того, на сегодняшний день создано множество различных электронных каталогов, библиотек, сетевых баз данных, поисковых и других подобных информационных систем, которые содержат и обрабатывают информацию на множестве различных языков, что приводит к необходимости их автоматического определения на этапе получения информации той или иной системой. Без этого практически невозможно распознать речь или текст, сделать тематическую выборку и распределить информацию в базе данных, определить, какие модули потребуются для её обработки или какому специалисту она должна быть направлена. Также необходимость определения языка информационных блоков возникает у студентов, аспирантов и сотрудников различных организаций, например, при работе с корреспонденцией или поиске информации по нужной тематике, у программистов, которым требуется автоматизировать процессы ввода данных, обработки документов и т.д. Например, при сканировании и распознавании документов или проверке орфографии и грамматики вводимого в компьютер текста обязательно нужно знать, на каком языке он написан, чтобы выбрать для него подходящий словарь или справочник, то же касается и машинного распознавания речи.

Постановка задачи

На сегодняшний день существует множество технологических и программных решений упомянутой проблемы, но каждое из них обладает некоторыми недостатками и ограничениями, что позволяет использовать его только при определённых условиях. Этот факт вынуждает разработчиков искать альтернативные пути и подходы к решению данной проблемы или комбинировать известные, что подтверждается количеством и разнообразием существующих способов и реализующих их алгоритмов, а также программ и систем языковой идентификации неструктурированной информации – языковых определителей. Кроме того, возникает вопрос адекватной оценки подобных решений с целью выбора наиболее подходящего из них для конкретной работы. Существующая методика такой оценки будет рассмотрена далее с учётом её особенностей и условий использования, что позволит выявить преимущества и недостатки, а также показать ее практическую эффективность. Однако её применимость в области языковой идентификации вызывает множество сомнений, обоснованных конкретными доводами, в связи с чем автор предлагает альтернативную систему оценки результатов языковой идентификации.

Существующие системы оценки

В настоящее время для оценки систем идентификации информации широко используется несколько простых векторных и вероятностных моделей, что связано с легкостью их реализации [1]. Основным направление их применения являются системы поиска информации, спрос на которые постоянно растет, поэтому большинство реализаций написаны непрофессионалами в данной сфере, для которых фактор простоты реализации является определяющим. Но даже если такую систему создает разработчик, осведомленный о достижениях в этой области, вероятнее всего он изберет один из проверенных и зарекомендовавших себя методов с целью экономии ресурсов. Это приводит к ситуации, когда технологии, дающие хорошие показатели в одних условиях, оказываются практически неэффективными в других, в результате чего более 90% всех поисковых систем используют различные вариации векторных моделей, эффективность которых оставляет желать лучшего, что вызывает заслуженное недоверие пользователей к соответствующим системам. По тем же причинам подобные подходы распространяются и на системы языковой идентификации речи или текста, что не всегда оказывается полезным, а в некоторых случаях приводит к неприемлемым результатам, примеры которых будут показаны ниже.

В современной практике все системы идентификации (дикторов, языковой принадлежности и т.д.) принято оценивать с использованием ошибок 1-го и 2-го рода, или же показателей точности, полноты и f-меры, по формулам, созданным на основе теоремы Байеса из области математической статистики [2]. Ошибки 1-го рода (*англ. type I errors, α errors, false positives*) показывают, сколько объектов других классов были приняты за искомый, а ошибки 2-го рода (*англ. type II errors, β errors, false negatives*) – сколько объектов искомого класса были пропущены. Такая методика оценки позволяет комплексно оценить алгоритм выборки и даёт хорошие результаты при определении эффективности поисковых систем за счёт простоты задачи [3], где не требуется точный бинарный ответ: «да» или «нет», то есть, присутствует ли заданный текст или документ в данной выборке, а достаточно процентной вероятности его нахождения. Обычно эта вероятность определяется по следующей формуле [4]:

$$P(c|d) = P(d|c)P(c) / P(d),$$

где: P(c|d) – вероятность принадлежности документа **d** классу **c**; P(d|c) – вероятность встретить документ **d** среди всех документов

класса **c**; P(c) – безусловная вероятность встретить документ класса **c** в корпусе документов (выборке); P(d) – безусловная вероятность нахождения документа **d** в корпусе документов (выборке).

Результаты тестирования существующих технологий

Расчеты по данной формуле позволяют наглядно представить эффективность поиска документов в определённом объеме информации и в большинстве случаев оказываются приемлемыми, так как в результате работы поисковой системы выдаётся множество более или менее подходящих документов и особая точность не требуется. Однако данный подход не всегда применим к другим системам идентификации, в том числе к системам определения языковой принадлежности информации, потому что не учитывает их особенности. Прежде всего это касается вероятности верного определения – ассоциативной вероятности (*термин предложен автором*) – которая показывает соответствие полученных результатов действительности, а не вероятность результата вообще, характерную для систем поиска информации. То есть при выяснении эффективности способа, технологии, алгоритма, программы или системы языковой идентификации важен не сам факт идентификации, а верность результата. Более того, требуется определить языковую принадлежность точно, а не приблизительно, и выборка ведётся среди всех языков мира, а не каких-то конкретных классов (или групп), что нивелирует параметры P(c) и P(d), делая использование указанной формулы практически бессмысленным. Это показывает следующий пример языковой идентификации кастильской фразы «yo no quiero hacerlo, hombre» (*я не хочу этого делать, друг*) с помощью системы **Gusser** [5], результат работы которой выведен в таблицу 1 [6].

Таблица 1. Результат языковой идентификации фразы в системе Guesser

№ 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 из примера, идентификация произведена и результаты получены, но практически они бесполезны, так как языковая принадлежность информации не определена, следовательно, цель идентификации не достигнута. Это показывает несостоятельность выбранной разработчиками **Guesser**-а технологии для определения языковой принадлежности текста, а также ставит под сомнение эффективность этой системы, что никак не выявляется с помощью принятой методики.

Результаты определения языковой принадлежности той же фразы в других программах и системах языковой идентификации приведены в таблице 2 (оформление результатов соответствует оригинальному выводу программы).

Таблица 2. Результаты языковой идентификации одной фразы различными программами и системами

№ п/п	Программа или система языковой идентификации	Результат
1	**Automatic language identifier** (T-Labs) [7]	**Spanish** (The text is too short!)
2	**Полиглот 3000** [8]	Испанский (Вероятность 71 %)
3	**TextCat** [9]	Spanish
4	**Language Identifier by Henrik Falck** [10]	**Spanish** or **French**
5	**SILC** [11]	Spanish
6	**Talenknobbel** [12]	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%

Как видно из данных примеров, языковая принадлежность текстов не всегда верно определяется из-за недостатка материала для анализа (текст слишком короткий) или программа даёт бесполезные результаты. Например, **Talenknobbel** почти с одинаковой вероятностью относит текст к норвежскому, испанскому или польскому языку, которые принадлежат разным генеалогическим группам. Также показательно, что вероятность ассоциации текста с португальским языком более 62 %, а с испанским (вероятно кастильским) только ~ 39 %, при том, что эти языки являются близкородственными. Кроме того, программа **Полиглот 3000** показала вероятность определения с точностью в 71 %, что

не позволяет сделать окончательный вывод о языковой принадлежности информации.

Ещё более показательны результаты следующего эксперимента (Таблица 3), при котором тем же программам и системам была передана кастильская фраза «**es mi casa**» (*это мой дом*).

*Таблица 3. Результаты языковой идентификации фразы «**es mi casa**»*

№ п/п	Программа или система языковой идентификации	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 группы) и финский (финно-угорской группы), что свидетельствует о неадекватности самой программы. Эти особенности результатов обязательно должны учитываться методикой оценки, иначе определить объективную эффективность подобной системы языковой идентификации будет невозможно.

Приведённые доводы и примеры показывают, что методика оценки результатов языковой идентификации информации должна существенно отличаться от методик оценки эффективности поисковых систем, основанных на формулах математической статистики.

Суть разработанной методики

По итогам экспериментального исследования можно выделить 3 возможных результата языковой идентификации:

1) языковая принадлежность информации не определена;

2) языковая принадлежность информации определена (верность не гарантирована);

3) языковая принадлежность информации определена верно.

На практике для большинства современных языковых определителей наиболее вероятен 2-й вариант, но что значит «вероятность определения языковой принадлежности, например, 87%»? Язык этого текста определен или нет? Если языковой определитель выдаёт вероятностный результат, значит он автоматически должен расцениваться как ненадежный. Кроме того, сам вероятностный результат может существенно варьироваться, и в зависимости от его стабильности и точности в различных условиях языковой определитель, способ или технология могут считаться более точными, менее точными или неточными. Например, если почти все переданные тексты идентифицированы, ассоциативная вероятность всегда близка к 100% или языковая группа всегда определяется верно, но не всегда идентифицируется язык, то такой результат может считаться хорошим и высокоточным. Если же результатирующие показатели выдаются с неопределенной вероятностью, группы языков не различаются, а логика определения не прослеживается, то такой результат должен считаться неудовлетворительным и неточным.

Таким образом, с учётом указанных особенностей можно составить следующую условную градацию способов, алгоритмов, технологий, программ и систем определения языковой принадлежности информации.

1. **Сверхточные** – качество результатов отличное. Языковая принадлежность информации всегда определяется верно.
2. **Высокоточные** – качество результатов хорошее. Не допускаются вероятностные результаты, верно идентифицируется более 97% полученной информации при любых условиях.
3. **Точные** – качество результатов хорошее. Имеют высокую ассоциативную вероятность (более 95%) и незначительное количество неидентифицированных информационных блоков (1-3%).
4. **Нормальные** – качество результатов удовлетворительное. Верно идентифицируется более 80 % полученной информации.
5. **Низкоточные** – качество результатов низкое. Дают стабильные вероятностные результаты, верно идентифицируется более 60% полученной информации,

логика идентификации прослеживается.

6. **Неточные** – качество результатов неудовлетворительное. В большинстве случаев языковая принадлежность информации не определяется или определяется неверно, верно идентифицируется менее 60% полученной информации, логика идентификации не прослеживается.

В соответствии с данной градацией результативность существующих решений в области определения языковой принадлежности информации просто определяется опытным путём или вычисляется математически, а разрабатываемые способы, алгоритмы, программы и системы языковой идентификации могут изначально проектироваться с расчётом на конкретную точность результатов, так как их итоговые показатели будут зависеть от базовых технологий, эффективность которых рассчитывается по основным критериям [13]. Например, если в основе технологии языковой идентификации лежит символьный способ различения языков, основанный на выявлении специальных знаков или их сочетаний в тексте, то такое решение заведомо будет неэффективным, так как недостатки подобных способов давно известны [14].

Резюмируя всё сказанное выше, для оценки результатов определения языковой принадлежности информации предлагается использовать следующую методику.

Описание методики. Данная методика предназначена для оценки результатов определения языковой принадлежности информации посредством различных способов, подходов и алгоритмов, а также для выяснения идентификационной эффективности технологий и языковых определителей. Ее разработка произведена автором на основании результатов теоретических и экспериментальных исследований, проведенных в ходе диссертационной работы по тематике языковой идентификации текстов. Основными методами являются: анализ, сопоставление, синтез и методы математической статистики.

Область применения. Может применяться для определения идентификационной эффективности существующих способов, алгоритмов, технологий, программ и систем языковой идентификации, а также при проектировании подобных разработок.

Цель. Оценка качества результатов языковой идентификации способов, алгоритмов, технологий, программ и систем в области определения языковой принадлежности информации.

Необходимое оснащение (материалы). Полное описание полученных результатов, условий проведения экспериментов (тестирования) и оцениваемых способов, алгоритмов, технологий, программ и систем, а также реализующее их программное обеспечение (если имеется) и идентифицируемые материалы требуемых тематик.

Порядок работы. При наличии программного обеспечения производится его тестирование следующим образом. Тестируемой программе или системе передаётся не менее 1000 случайных фраз заданной длины (зависит от способов, целей и условий идентификации) или 1000 неструктурированных и неадаптированных случайных текстов объёмом от 30 до 60 слов различной тематики на каждом определяемом языке (что в среднем соответствует одному абзацу в большинстве документов). Тестирование проводится с помощью фраз и текстов на всех определяемых языках. В итоге проведённого анализа каждого информационного блока программой выдаются результаты идентификации или сообщается о ее невозможности. Все выданные результаты и сообщения должны фиксироваться вкупе с анализируемыми материалами.

При отсутствии программного обеспечения производится расчёт эффективности способов и алгоритмов их взаимодействия по описанию разработчиков, тестовым таблицам и с учётом критики специалистов (анализа преимуществ и недостатков).

Обработка данных. По окончании эксперимента все результаты сопоставляются с действительной языковой принадлежностью каждого текста, на основе чего вычисляется количество допущенных ошибок, выраженное в процентном соотношении. Следовательно, оставшееся число процентов будет являться показателем вероятности верной идентификации (ассоциативной вероятности) тестируемой программы. Например, если из 1000 переданных программе текстов языковая принадлежность 8-ми определена неверно, то расчёт в процентах производится следующим образом:

$$(8/1000)*100 = 0{,}8\ \%;$$

$$P_{id} = 100 - 0{,}8 = 99{,}2\ \%,$$

где P_{id} – показатель вероятности верной идентификации программы.

Данный расчёт может быть произведён и в обратном направлении. То есть, процентное соотношение текстов с верно определённой языковой принадлежностью, которое является показателем идентификационной эффективности программы, подсчитывается сразу. Однако такой подход более ресурсоёмкий и менее рациональный при небольшом количестве ошибок идентификации.

В итоге все результаты расчётов фиксируются и вычисляется суммарная ассоциативная вероятность, а также количество идентифицированных и не идентифицированных текстов.

При отсутствии программной реализации оцениваемого решения за основу берётся верность и адекватность результатов идентификации каждого способа, входящего в состав данной системы, а также алгоритм их итогового объединения и резюмирования выходных данных.

Выводы. При отсутствии вероятностных результатов и количестве верно идентифицированных фраз и текстов более 97% оцениваемые способ, алгоритм, технология, система или определитель условно могут считаться высокоточными; при наличии вероятностных результатов, ассоциативной вероятностью более 95% и верной идентификации более 90% переданной информации – точными; при наличии стабильных результатов и верной идентификации более 80% переданной информации – нормальными; при ассоциативной вероятности более 60%, стабильными результатами, чёткой логикой идентификации и незначительным количеством неидентифицированных информационных блоков – низкоточными; при нестабильных результатах и ассоциативной вероятности менее 60% – неточными. Соответственно качество результатов должно оцениваться как: хорошее, нормальное, удовлетворительное, низкое и неудовлетворительное. Решений языковой идентификации с отличным качеством результатов на данный момент не существует.

Достоинства и недостатки методики. Проста в использовании, не требует сложных математических расчётов и специальной подготовки, может быть адаптирована под любые условия эксперимента, однако имеет показатели точности (верности) идентификации всего по 6 уровням, что не позволяет производить детальные сравнения множества различных способов, алгоритмов, технологий, программ и систем в области определения языковой принадлежности информации. Однако на практике оценок предлагаемой градации достаточно для решения большинства задач.

Пример использования данной методики применительно к рассмотренным выше языковым определителям приводится в таблице 4.

Таблица 4. Оценка точности различных языковых определителей по результатам экспериментального исследования автора

№ п/п	Название определителя	Ассоциативная вероятность	Точность (верность) языковой идентификации
1	Guesser	50 %	неудовлетворительная
2	Automatic language identifier (T-Labs)	76 %	удовлетворительная
3	Полиглот 3000	95 %	нормальная
4	TextCat	81 %	нормальная
5	Language Identifier by Henrik Falck	71 %	удовлетворительная
6	SILC	72 %	удовлетворительная
7	Talenknobbel	60 %	неудовлетворительная

Заключение

Предложенная методика может быть полезна как разработчикам технологий в области языковой идентификации информации, так и потребителям таких технологий. Прежде всего это относится к программистам и создателям автоматизированных комплексов по обработке информации, таких как: системы глобального мониторинга, многоязычные базы знаний, автоматические системы межъязыковых переводов текстов и речи, поисковые системы, работающие в глобальной информационной среде и т.д.

Дополнительно стоит отметить, что при разработке технологий и систем определения языковой принадлежности информации требуется учитывать не только верность результатов, но и другие показатели, такие как: ресурсоёмкость, количество определяемых языков, возможность определения языковых групп без идентификации конкретного языка и т.д., которые существенно влияют на эффективность данной разработки. Например, при сверхвысокой точности (верности) результатов, но и сверхвысокой ресурсоёмкости, технология или система будет неэффективной вследствие её практической неприменимости.

ПОЯСНЕНИЕ ТЕРМИНА (предложен автором).

«Ассоциативная вероятность» – показатель вероятности верной ассоциации анализируемого материала с неким шаблоном, моделью или классом (например, текста с определённым языком или языковой группой).

Литература

1. Кураленок И.Е. Оценка систем текстового поиска : диссертация ... кандидата физико-математических наук : 05.13.01. Санкт-Петербург, 2004. – 112 с.
2. Гмурман В.Е. Теория вероятностей и математическая статистика. – М. : Высшая школа, 2003. – 479 с.
3. Rish I. An empirical study of the naive Bayes classifier / IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.
4. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, 2008.

5. Автоматический определитель языка текста Gusser [Электронный ресурс]. URL: “Guesser.ru” http://guesser.ru/
6. Калегин С.Н. Оценка эффективности методов определения языковой принадлежности неструктурированного текста и варианты их программной реализации. Международная конференция «CONCORT-2016», Нижний Новгород, 2016.
7. Automatic language identifier (Автоматический определитель языка) [Электронный ресурс]. URL: http://labs.translated.net/.
8. Автоматический определитель языка текста Полиглот 3000 (П3000) [Электронный ресурс]. URL: http://www.polyglot3000.com/.
9. Программа TextCat [Электронный ресурс]. URL: http://odur.let.rug.nl/~vannoord/TextCat/.
10. Language Identifier by Henrik Falck [Электронный ресурс]. URL: http://whatlanguageisthis.com/.
11. SILC RALI [Электронный ресурс]. URL: http://rali.iro.umontreal.ca/rali/.
12. Автоматический определитель языка Talenknobbel [Электронный ресурс]. URL: http://www.fuzzums.nl/~joost/talenknobbel/.
13. Калегин С.Н. Автоматизация процесса языковой идентификации текста на основе существующих решений. / Нейрокомпьютеры: разработка, применение. № 1. – Москва : Радиотехника, 2017. – С. 56-65.
14. 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.

References

1. Kuralenok I.E. Ocenka sistem tekstovogo poiska : dissertacija ... kandidata fiziko-matematicheskih nauk : 05.13.01. Sankt-Peterburg, 2004. – S. 112.
2. Gmurman V.E. Teorija verojatnostej i matematicheskaja statistika. – M.: Vysshaja shkola, 2003. – 479 s.
3. Rish I. An empirical study of the naive Bayes classifier / IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.
4. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, 2008.
5. Avtomaticheskij opredelitel' jazyka teksta Gusser [Jelektronnyj resurs]. URL: “Guesser.ru” http://guesser.ru/
6. Kalegin S.N. Ocenka jeffektivnosti metodov opredelenija jazykovoj prinadlezhnosti nestrukturirovannogo teksta i varianty ih programmnoj realizacii. Mezhdunarodnaja konferencija «CONCORT-2016», Nizhnij Novgorod, 2016.
7. Automatic language identifier (Avtomaticheskij opredelitel' jazyka) [Jelektronnyj resurs]. URL: http://labs.translated.net/
8. Avtomaticheskij opredelitel' jazyka teksta Poliglot 3000 (P3000) [Jelektronnyj resurs]. URL: http://www.polyglot3000.com/
9. Programma TextCat [Jelektronnyj resurs]. URL: http://odur.let.rug.nl/~vannoord/TextCat/
10. Language Identifier by Henrik Falck [Jelektronnyj resurs]. URL: http://whatlanguageisthis.com/
11. SILC RALI [Jelektronnyj resurs]. URL: http://rali.iro.umontreal.ca/rali/
12. Avtomaticheskij opredelitel' jazyka Talenknobbel [Jelektronnyj resurs]. URL: http://www.fuzzums.nl/~joost/talenknobbel/
13. Kalegin S.N. Avtomatizacija processa jazykovoj identifikacii teksta na osnove sushhestvujushhih reshenij. / Nejrokomp'jutery: razrabotka, primenenie. № 1. – Moskva: Radiotehnika, 2017. – S. 56-65.
14. Kalegin S.N. Vazhnost' vybora osnovnogo identifikacionnogo principa pri proektirovanii jazykovyh opredelitelej. / Sovremennye informacionnye tehnologii i IT-obrazovanie. Tom 12, № 2. – Moskva, 2016. – S. 194-204.

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.

Поступила: 15.07.2017

Об авторе:

Калегин Сергей Николаевич, начальник сектора, аспирант, Московский научно-исследовательский телевизионный институт; соискатель, Институт проблем управления им. В.А. Трапезникова РАН, ksn@mniti.ru.

Note on the author:

Sergey Kalegin N., chief of section, aspirant (applicant), Moscow Research TV Institute Joint Stock Company; applicant, Institute of Control Sciences of the Russian Academy of Sciences, ksn@mniti.ru.