ИНФОРМАТИКА
2011 июль-сентябрь № 3

*УДК 004.912*

**Ю.Б. Крапивин**

# АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ЯЗЫКА ТЕКСТОВОГО ДОКУМЕНТА ДЛЯ ОСНОВНЫХ ЕВРОПЕЙСКИХ ЯЗЫКОВ

*Проводится анализ основных методов решения задачи автоматического определения языка текстового документа и предлагается алгоритм, основанный на комбинировании алфавитного метода, метода грамматических слов и алфавитно-триграммного метода, сочетающий в себе возможности минимального статистического и лингвистического анализа языковых данных и обеспечивающий эффективное решение указанной задачи.*

## Введение

В работе [1] была определена базовая функциональность, а также представлена и описана принципиальная схема системы автоматического распознавания воспроизведенных фрагментов текстового документа, в соответствии с которой в первую очередь должна быть решена задача определения языка его представления.

Как показало проведенное исследование, наиболее известные в настоящее время методы решения указанной задачи строятся по одинаковому принципу, описать который с помощью терминологии информационного поиска можно следующим образом. Во-первых, для каждого естественного языка (ЕЯ) из конечного набора языков, участвующих в процессе идентификации, строится поисковый образ языка (ПОЯ), во-вторых, для каждого анализируемого (входного) документа в соответствии с правилами построения ПОЯ создается поисковый образ документа (ПОД) и, в-третьих, выбирается стратегия сравнения ПОД и ПОЯ, на основании которой и выносится решение о принадлежности входного документа определенному языку.

## 1. Методы решения задачи автоматического определения языка текстового документа

В зависимости от применяемых правил построения поисковых образов и стратегий их сравнения различают следующие основные методы определения языка текстовых документов: коротких слов [2, 3], частотных слов [4, 5], N-грамм [2, 6–8], статистический [9–11], строковых ядер [12], алфавитный [13, 14], грамматических слов [13], неграмматических слов [13].

Метод коротких слов при построении ПОЯ использует слова определенной длины, не превышающей заданный порог. Например, в работе [2] на основании корпуса текстов, содержащего документы общим объемом один миллион символов для каждого идентифицируемого языка, извлекались лексемы длиной до пяти символов, встретившиеся в тексте более трех раз. Вероятность появления в тексте *i*-й лексемы рассчитывалась как отношение ее частоты к общей сумме частот всех лексем из полученного набора. Предложение входного документа разбивалось на лексемы, и лексемам, присутствующим в ПОЯ, назначались их частоты, а отсутствующим – некоторая минимальная частота. Вероятность принадлежности предложения языку рассчитывалась как произведение вероятностей его лексем.

Согласно методу частотных слов ПОЯ представляет собой набор слов, обладающих наибольшей частотой встречаемости в сравнении с остальными словами документов из обучающего корпуса. Количество слов в ПОЯ, необходимых для определения языка, у разных авторов отличается на порядок; например, в [4] использовалась тысяча слов, а в работе [5] – всего сто. Вероятность принадлежности входного документа языку определяется на основании пословного сравнения с ПОЯ с учетом рассчитанных вероятностей его элементов.

Как и в методах, описанных выше, ПОЯ по методу N-грамм [2, 6–8] строится на основании тренировочного корпуса текстов. Составляющими ПОЯ являются N-граммы – последовательности символов строки текста переменной [6] или строго фиксированной длины [2]. Например, в [2] в ПОЯ включались триграммы, встречавшиеся в текстах тренировочного корпуса более ста раз, и их количество изменялось от 2550 до 3560 в зависимости от языка. Вероятность отнесения документа к какому-либо языку рассчитывалась подобно методу коротких слов.

Работы [9–11] реализуют статистический метод решения указанной задачи. Например, в работе [9] на основании тренировочного корпуса текстов для каждого языка определялось распределение вероятностей униграмм и биграмм, образующих поисковые образы языков, или триграмм в случае экспериментов с отсканированными документами. Далее для входного документа также строилось распределение вероятностей составляющих его N-грамм. Используя метрику Кульбака – Лейблера [15], вычислялась относительная энтропия между вероятностными распределениями входного документа и всеми ПОЯ. Тексту назначался язык с минимальной относительной энтропией.

Вероятность, с которой входной документ принадлежит некоторому языку, в [10] определялась с помощью метода Монте-Карло.

В работе [11] предложен способ автоматического определения числа задействованных языков и принадлежащих им текстовых фрагментов документа на основании статистических данных о совместной встречаемости слов языка.

Работа [12] является примером решения задачи определения языка документа методом строковых ядер, принципы работы которого близки методу N-грамм. Рассматривались два способа: первый заключается в ускорении вычислений строковых ядер с помощью суффиксных деревьев, создаваемых по алгоритму Укконена [16], и применении метода поддерживающих векторов (SVM) для классификации к одному из языков; второй – в построении ядерного центроидного вектора ПОЯ и дальнейшем определении принадлежности документа языку путем вычисления квадратичного евклидова расстояния, а не величины несоответствия, как в [6].

В [13] приводится способ, основанный на комбинировании двух лингвистических методов: алфавитного и грамматических слов – с методом неграмматических слов – аналогом метода *N*-грамм, согласно которому в данном случае элементами ПОЯ являются наиболее частые окончания слов.

Алфавитный метод [13, 14] заключается в определении языка на основании обнаруженных в анализируемом тексте характерных диакритических знаков – специальных значков, добавляемых к буквам того или иного алфавита с целью обозначить изменение их стандартного чтения или же указать на какую-либо особую роль, которую звук, обозначенный буквой с диакритикой, играет в слове.

Метод грамматических слов состоит в поиске слов, обладающих малым лексическим значением, но необходимых для выражения грамматических и других отношений в предложении, которые характерны для рассматриваемого языка. Этими словами являются предлоги, союзы, артикли и т. д. [17].

Как показал проведенный анализ, указанные методы достаточно эффективны только при определенных ограничениях на размер входного текста и сравнительные характеристики тех естественных языков, на множестве которых решается задача, и, таким образом, не могут в отдельности обеспечить построение эффективного соответствующего алгоритма в общем случае, который как раз и имеется в виду в данной работе.

### 2. Алгоритм автоматического определения языка текстового документа

Любой текст $T$ языка $L$ можно рассматривать как конечную цепочку $T = a_1 a_2 ... a_n$, где $a_i \in A$, $i = \overline{1,n}$, $A$ – алфавит языка $L$, образованную в соответствии с множеством его лексико-грамматических, синтаксических и семантических правил. Исходя из этого задачу распознавания языка текста можно решать, начиная уже с уровня алфавита, наращивая при необходимости объем привлекаемых знаний о ЕЯ из других уровней его глубины и, таким образом, имея возможность продвигаться от использования простейшего алфавитного метода решения задачи до *N*-граммного; метода грамматических слов; метода, основанного на использовании полных эталонных словарей ЕЯ, и далее – вплоть до методов, затрагивающих синтаксический и семантический уровни ЕЯ. Очевидно, что все перечисленные ранее методы ориентированы на использование знаний о ЕЯ в пределах от уровня алфавита до лексико-грамматического уровня глубины ЕЯ, что вполне приемлемо по трудоемкости для их использования в промышленных

системах автоматической обработки текста. Проблема заключается только в том, чтобы, оставаясь на этих уровнях, добиться большей эффективности (точности) данных методов при решении задачи, в данном случае на множестве белорусского, русского и ряда европейских языков (английского, французского, немецкого).

В проведенных экспериментах были задействованы триграммный метод, метод грамматических слов, алфавитный и алфавитно-триграммный методы. При исследовании первого метода в качестве обучающей выборки для европейских языков использовался известный лингвистический корпус текстов Leipzig Corpora Collection (LCC) [18]. Данный корпус построен на основании выбранных случайным образом предложений из газетных статей и веб-документов. Что касается русского и белорусского языков, то для них по принципу LCC был построен собственный текстовый корпус. Для каждого из пяти языков использовались текстовые наборы по 100 тыс. предложений. Для второго метода впервые лингвистически строго и достаточно полно был определен список грамматических слов (ПОЯ). В него вошли: предопределители (для английского языка это, например, quite, rather, such и др.), предквантификаторы (all, half), детерминативы (few, last, many, only), предлоги (about, above), союзы (and, or, but), артикли (a, an, the), модальные и вспомогательные глаголы (can, could, has), местоимения (I, we, it, him, mine), определители (as, more, very, enough), наречия (below, near, now, there) и частицы (to, not). Всего из больших, например порядка 200 тыс. словоформ для английского языка и порядка 1 млн 200 тыс. словоформ для белорусского языка, тэгированных лексико-грамматическими классами слов эталонных словарей автоматически было получено: 1532 грамматических слова для белорусского языка (BE), 1435 – для русского (RU), 311 – для английского (EN), 559 – для французского (FR), 968 – для немецкого (DE). Что касается последнего из указанных выше методов, то его идея заключается в том, чтобы, в отличие от классического триграммного метода, в качестве ПОЯ использовать списки наиболее частотных триграмм ЕЯ, в нашем случае – по 400 триграмм для каждого языка. При этом, что важно, сохраняется тот же, что для второго и третьего методов, характер алгоритма (сравнение двух списков): для всех трех методов язык входного текста определяется наличием в последнем уникальных для данного языка элементов алфавита, грамматических слов, триграмм.

Для тестирования указанных четырех методов было выбрано по 100 текстов длиной по 14 и столько же – по 7 слов каждый для всех пяти языков. Такие значения длины были взяты исходя из того, что если построенный в итоге алгоритм будет достаточно точно решать рассматриваемую задачу для текстов длиной в 14 слов, значит, он тем более будет ее эффективно решать для среднестатистического по длине предложения (например, для известного корпуса текстов LOB Corpus [19] значение этого параметра оказалось равным примерно 19 слов). Если он окажется таковым и для текстов вдвое короче (7 слов), значит, его можно будет с успехом использовать и для распознавания во входном тексте вставок на других языках в виде отдельных предложений и даже коротких фраз. Критерием для признания языка $L_i$ в качестве решения задачи являлась максимальная оценка для $|A(T) \cap A_i|$, где $A(T)$ – алфавит входного текста $T$ и $A_i$ – алфавит языка $L_i$ , их соответственно множество попарно различных слов и множество уникальных грамматических слов, множество попарно различных триграмм и множество наиболее частотных уникальных триграмм соответственно для алфавитного метода, метода грамматических слов и алфавитно-триграммного метода. Если несколько языков получали одинаковую максимальную оценку, то это хотя бы позволяло определить языковую группу для текста $T$.

Результаты экспериментов показали следующее. Алфавитный метод, используя минимальные ресурсы, позволяет точно отнести входной текст к языкам одного или близких алфавитов, в нашем случае – к языковой группе BE-RU или EN-FR-DE. Значит, его можно, например, использовать как одно из средств оптимизации строящегося алгоритма. Кроме того, этот метод позволил, например, определить язык 82 белорусских и 97 русских текстов длиной в 14 слов (для оставшихся соответственно 18 и 3 документов не определился из указанной пары язык-победитель, поскольку в этих текстах не оказалось уникальных элементов алфавита ни одного из данных языков). На текстах длиной в 14 слов каждый из оставшихся трех методов дает близкий к абсолютному результат (исключение составил триграммный метод для русских текстов – 90 %), который становится на 1–3 % ниже при переходе к текстам длиной в 7 слов. При этом самым эффективным остается метод грамматических слов, который, учитывая, что в

некоторых коротких текстах их может просто не оказаться, очевидно, допускает свое последующее усиление ресурсами уникальных триграмм.

Таким образом, эффективный алгоритм решения задачи, сочетающий в себе возможности минимального статистического и лингвистического анализов языковых данных, может быть основан на комбинировании алфавитного метода, метода грамматических слов и алфавитно-триграммного метода. В целом, в силу изложенного имеет место следующая принципиальная схема алгоритма:

1. Начало.

2. Обработать текст $T$ алфавитным методом для всех ЕЯ из их заданного множества. Если на выходе получен единственный язык $L_i$ , то обозначить его как решение задачи и перейти к п. 5.

3. Зафиксировать для $T$ языковую группу и обработать его методом грамматических слов для тех ЕЯ, которые входят в эту группу. Если на выходе получен единственный язык $L_i$ , то обозначить его как решение задачи и перейти к п. 5.

4. Зафиксировать для $T$ языковую группу и обработать его алфавитно-триграммным методом для тех ЕЯ, которые входят в эту группу. Обозначить $L_i$ в качестве решения задачи.

5. Конец.

Понятно, что если $L_i$ на шаге 4 окажется не единственным, то необходим некоторый дополнительный ресурс снятия многозначности, хотя бы, например, диалог с пользователем.

### 3. Тестирование алгоритма

Описанный алгоритм был реализован на множестве указанных ранее пяти ЕЯ. В таблице приводятся результаты его тестирования на 500 текстах, по 100 текстов для каждого языка, из них по 25 текстов для каждой группы (тексты длиной 4 КБ, 5 предложений, 14 слов, 7 слов). Тестовый набор использует тексты из указанных ранее корпусов текстов для исследования триграммного метода решения задачи. Здесь $T$ – это язык входного текста; $M_1$, $M_2$, $M_3$ – соответственно алфавитный метод, метод грамматических слов и алфавитно-триграммный метод; 4 КБ, 5 предложений, 14 слов, 7 слов – длины входных текстов. В столбцах, соответствующих $M_1$, показано количество текстов из 25 поступивших на вход, для которых правильно определен их язык, в последующих столбцах – аналогичное количество, но только из числа тех текстов, для которых на предыдущем этапе алгоритма язык не был определен.

Результаты тестирования алгоритма

| Язык $T$ | BE | | | RU | | | DE | | | EN | | | FR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Метод | $M_1$ | $M_2$ | $M_3$ | $M_1$ | $M_2$ | $M_3$ | $M_1$ | $M_2$ | $M_3$ | $M_1$ | $M_2$ | $M_3$ | $M_1$ | $M_2$ | $M_3$ |
| 4 КБ | 25 | – | – | 25 | – | – | 0 | 25 | – | 0 | 25 | – | 0 | 25 | – |
| 5 предл. | 25 | – | – | 24 | 1 | – | 0 | 25 | – | 0 | 25 | – | 0 | 25 | – |
| 14 слов | 24 | 1 | – | 23 | 2 | – | 0 | 25 | – | 0 | 25 | – | 0 | 25 | – |
| 7 слов | 21 | 3 | 1 | 23 | 2 | – | 0 | 24 | 1 | 0 | 24 | 0 | 0 | 24 | 1 |

Применение алфавитного метода на первом этапе алгоритма позволило идентифицировать группы языков одного алфавита: BE-RU и DE-EN-FR, а также точно определить языки в группе BE-RU: BE – для текстов длиной 4 Кб и 5 предложений, RU – для текстов длиной 4 КБ. В этом случае точность метода составила 100 %, т. е. 25 из 25 текстов. В группе BE-RU для RU из 25 текстов длиной 5 предложений были верно идентифицированы 24 (96 %). В случае текстов длиной 14 слов метод верно определил язык: BE – 24 из 25 (96 %), RU – 23 из 25 (92 %), для текстов длиной по 7 слов: BE – 21 из 25 (84 %), RU – 23 из 25 (92 %).

Применение метода грамматических слов на втором этапе алгоритма позволило точно (100 %) доопределить языки в группе BE-RU: BE – для текстов длиной 14 слов – 1 из 1, RU – для текстов длиной 5 предложений – 1 из 1, длиной 14 слов и 7 слов – 2 из 2 соответственно. В группе DE-EN-FR точно идентифицирован язык для текстов длиной 4 КБ, 5 предложений, 14 слов для языков DE, EN и FR. Для текстов длиной 7 слов метод точно доопределил язык BE – 3 из 4 (75 %). В случае текстов длиной 7 слов метод верно определил языки DE, EN, FR в 96 % случаев (24 из 25).

Применение алфавитно-триграммного метода на третьем этапе алгоритма позволило точно доопределить языки в группе BE-RU: BE – для текстов длиной 7 слов – 1 из 1. В группе DE-EN-FR точно идентифицированы языки FR и DE для текстов длиной 7 слов – 1 из 1 соответственно.

Из 500 входных текстов установить язык не удалось для одного текста английского языка длиной 7 слов следующего содержания: «In Hong Kong, gains in Hong Kong». Алфавитный метод на первом этапе позволил определить для него группу языков одного алфавита DE-EN-FR, метод грамматических слов на втором этапе не позволил снять многозначность, так как грамматическое слово «in» является общим для DE и EN, т. е. не является уникальным, а алфавитно-триграммный метод на третьем этапе дал одинаковые результаты. Таким образом, точность предложенного алгоритма в общем составила 99,8 %. При этом общее время решения задачи для всех 500 текстов на ЭВМ Athlon 1800, ОЗУ 1 Гб под управлением операционной системы Windows XP составило 920,526 с, а среднее время, затрачиваемое на определение языка одного текста, – 1,841 с.

**Заключение**

Полученные результаты позволяют считать, что предложенный алгоритм комбинации методов – алфавитного, грамматических слов и алфавитно-триграммного – может рассматриваться в качестве эффективного инструмента идентификации языка текста, используемого, в частности, в системе автоматического распознавания воспроизведенных фрагментов текстового документа.

**Список литературы**

1. Крапивин, Ю.Б. К задаче автоматического распознавания воспроизведенных фрагментов текстовых документов / Ю.Б. Крапивин // Вестник БрГТУ : Физика, математика, информатика. – 2009. – № 5 (59). – С. 120–123.

2. Grefenstette, G. Comparing two language identification schemes / G. Grefenstette // The Third Intern. Conf. on Statistical Analysis of Textual Data. – Rome, 1995.

3. Sibun, P. Language Determination: Natural Language Processing from Scanned Document Images / P. Sibun, A.L. Spitz // Proc. of the 4th ACL Conf. on Applied Natural Language Proceeding (ANLP). – Stuttgart, Germany, 1994.

4. Cowie, J. Language recognition for mono- and multilingual documents / J. Cowie, Y. Ludovic, R. Zacharski // Proc. of the Vextal Conference. – Venice, 1999.

5. Natural Language Identification using Corpus-based Models / C. Souter [et al.] // Hermes Journal of Linguistics. – 1994. – № 13. – P. 183–203.

6. Cavnar, W.B. N-Gram-Based Text Categorization / W.B. Cavnar, J.M. Trenkle // Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR). – Las Vegas, 1994. – P. 161–175.

7. Prager, J.M. Linguini: Language identification for multilingual documents / J.M. Prager // Proc. of the 32nd Hawaii Intern. Conf. on System Sciences. – Maui, Hawaii, USA, 1999.

8. Dunning, T. Statistical Identification of Language / T. Dunning // Computing Research Laboratory. Technical report MCCS. – New Mexico State University, 1994. – P. 94–273.

9. Sibun, P. Language identification: Examing the issues / P. Sibun, J.C. Reynar // Proc. of the 5th Symposium on Document Analysis and Information Retrieval (SDAIR). – Las Vegas, 1996. – P. 125–135.

10. Poutsma, A. Applying MonteCarlo Techniques to Language Identification / A. Poutsma // Proc. of Computational Linguistics in the Netherlands. – Amsterdam, Netherlands, 2001.

11. Biemann, C. Disentangling from Babylonian Confusion – Unsupervised Language Identification / C. Biemann, S. Teresniak // Proc. of the CICLing-2005. – Mexico City, 2005.

12. Kruengkrai, C. Language Identification Based on String Kernels / C. Kruengkrai // Proc. of the 5th Intern. Symposium on Communications and Information Technologies (ISCIT-2005). – Beijing, China, 2005.

13. Giguet, E. Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning / E. Giguet // 4th Intern. Workshop of Parsing Technologies. – Prague, Karlovy Vary, Czech Republic, 1995.

14. Newman, P. Foreign language identification: First step in the translation process / P. Newman // Proc. of the 28th Annual Conf. of the American Translators Accociation. – Albuquerque NM, USA, 1987. – P. 509–516.

15. Kullback – Leibler_divergence [Electronic resource] // Wikipedia. – Mode of access : http://en.wikipedia.org/wiki/ Kullback-Leibler_divergence. – Date of access : 15.12.2010.

16. Ukkonen, E. On-line construction of suffix trees / E. Ukkonen // Algorithmica. – 1995. – № 14 (3). – P. 249–260.

17. Function word [Electronic resource] // Wikipedia. – Mode of access : http://en.wikipedia.org/wiki/Function_word. – Date of access : 14.12.2010.

18. Quasthoff, U. Corpus Portal for Search in Monolingual Corpora / U. Quasthoff, M. Richter, C. Biemann // Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation, LREC 2006. – Genoa, 2006. – P. 1799–1802.

19. Lancaster-Oslo-Bergen Corpus [Electronic resource] // Wikipedia. – Mode of access : http://en.wikipedia.org/wiki/LOB_Corpus. – Date of access : 15.12.2010.

**Поступила 24.02.11**

*Брестский государственный*
*технический университет,*
*Брест, ул. Московская, 267*
*e-mail: ybox@list.ru*

**Y. Krapivin**

## AUTOMATIC IDENTIFICATION OF THE LANGUAGE OF TEXT DOCUMENT FOR BASIC EUROPEAN LANGUAGES

An analysis of the existing methods of solving the problem of the automatic identification of the text document has been proposed. The efficient algorithm ensuring the solution of the stated problem has been defined. It is based on the combination of the alphabetic method, the function word method and the alphabetic-trigram method, that unites possibilities of minimal statistical and linguistic analysis of lingual data.