УДК 004.4:62-52

Автоматизация процесса языковой идентификации текста на основе существующих решений

С.Н. Калегин – *аспирант, начальник сектора НТО, Московский научно-исследовательский телевизионный институт; соискатель ИПУ РАН*
E-mail: skalegin@inbox.ru

Приведен анализ эффективности существующих программ языковой идентификации текста с целью выяснения возможности их применения в системах автоматической обработки многоязычной информации. Протестировано несколько разнотипных программных решений, использующих различные способы идентификации; выявлены общие недостатки всех применяемых решений и оценены перспективы их автоматизации в различных условиях, что может помочь разработчикам соответствующих программных комплексов и систем по сбору и обработке данных сделать обоснованный выбор специального программного обеспечения и рассчитать риски при полной автоматизации процесса языковой идентификации.

Ключевые слова: *языковая идентификация текста, определение языковой принадлежности, языковой определитель, автоматизация определения языка, автоматизация языковой идентификации.*

The article provides an analysis of the effectiveness of existing language identification programs in order to clarify the possibility of their use in automatic multilingual information processing systems. By this analysis were tested several different types of program solutions which use different identification methods, that allowed to reveal common weaknesses of applied solutions and to measure the prospects of automation in various conditions. This will help developers to relevant software systems and data processing systems to make informed choices, and special software to calculate the risks of full automation of the language identification process.

Keywords: *language identification of the text, determining the language, language identifier, automating language identification.*

В настоящее время возрастает потребность в автоматической обработке неструктурированных данных в связи с большими объемами электронных документов, более 90 % которых представлены в текстовом виде. Особенно это заметно в области информационных технологий и средств массовой информации. Также в число активно развивающихся направлений обработки неструктурированных данных входят: поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документов, поиск похожих документов и дубликатов, сегментирование документов, обработка коротких сообщений (на форумах, в чатах, социальных сетях, электронной почте и т.п.) и многие другие.

Наиболее востребованными в разных сферах деятельности являются информационно-аналитические, поисковые и мониторинговые системы, применение которых стало фактическим стандартом. Об этом свидетельствует множество работ и проектов по созданию систем обработки информации на разных уровнях управления и для решения различных задач [1–6]. Например, информационно-аналитическая система «Лавина» [7], которая предназначена для сбора, обработки и консолидации разнородной неструктурированной информации (текстовой и аудиовизуальной) из внутренних и внешних источников (базы данных, Интернет, файловые системы, корпоративные информационные системы, телевизионный и радио эфир и др.) и ее аналитической обработки в режиме, близком к реальному времени. Чаще всего анализируемая информация в таких системах приводится к текстовому виду, а сам анализ проводится на основе статистических методов учета частотной встречаемости определенных комбинаций символов или слов в синтагме, предложении, абзаце, документе и т.п. [8].

Однако из-за большого объема и разнородности накопленной информации требуется не только применение современных аналитических систем, но и полная автоматизация многих рутинных процессов, что, возможно, существенно сократит ресурсозатраты при обработке данных. Одним из таких процессов является процесс определения языковой принадлежности текстов в многоязычной инфор-