УДК 004.912

Юрий Борисович Крапивин, к. техн. н.
Минский государственный лингвистический университет, Минск, Беларусь
***э-почта*: ybox@list.ru**

Yury Borisovich Krapivin, Cand. of Sc. (Tech. Sciences)
Minsk State Linguistic University, Minsk, Belarus
***e-mail*: ybox@list.ru**

К ЗАДАЧЕ ОБ АВТОМАТИЧЕСКОМ ОПРЕДЕЛЕНИИ ЯЗЫКА ТЕКСТА

Рассматривается решение задачи автоматического определения языка текста, методом машинного обучения на материале построенного корпуса текстов на 49 языках.

Ключевые слова: естественный язык; искусственная нейронная сеть; автоматическое определение языка; корпус текстов.

TOWARDS THE PROBLEM OF AUTOMATIC LANGUAGE IDENTIFICATION OF TEXT

The solution to the problem of automatic language identification using the machine learning method based on the constructed corpus of texts in 49 languages is considered.

Key words: natural language; artificial neural network; automatic language identification; text corpus.

Обработка естественного языка (ЕЯ) – одно из наиболее динамично развивающихся направлений искусственного интеллекта в настоящее время, связанное с решением задач автоматической обработки информации, представленной на естественном языке. Существуя уже более полувека, оно в своем развитии опирается на результаты в области общей лингвистики, математики, информатики (computer

science), психологии, социологии и др., что, в свою очередь, означает междисциплинарный характер проводимых исследований. К актуальным прикладным задачам этого направления можно отнести задачи: автоматизации инженерии знаний; информационного поиска; машинного перевода (МП) текстов с одного ЕЯ на другой; автоматической классификации текстов; автоматического реферирования текстов; автоматического определения языка текста.

Последняя задача заключается в определении языка, на котором написан текст или его часть. Необходимость получения результатов ее точного и оперативного решения обусловлена широким кругом их применения при создании интеллектуальных систем обработки данных, ориентированных на работу в многоязычной языковой среде, и во многом определяет дальнейшее использование инструментальных и лингвистических ресурсов, необходимых для эффективного функционирования таких систем. Так, например, в контексте создания информационно-поисковых систем решение задачи автоматического определения языка требуется на предварительных этапах анализа текстов документов и запросов пользователей, связанных с процедурами индексирования. Учитывая ориентацию современных систем на многоязычную информационную среду, а значит, поддержку multilanguage- и cross-language- функциональности, актуальность рассматриваемой задачи возрастает ещё больше, поскольку корректное автоматическое распознавание языка текста документа позволяет системе выбрать те инструментальные и лингвистические ресурсы используемого лингвистического процессора, которые ей необходимы для автоматической обработки текстовых документов / запросов данного / данных языка / языков в каждый конкретный момент решения общей задачи информационного поиска. Функциональность лингвистического процессора в общем случае обеспечивает следующие основные уровни лингвистического анализа: 1) предварительное форматирование текста (преформатирование); 2) лексический анализ; 3) лексико-грамматический анализ; 4) синтаксический анализ; 5) семантический анализ.

Уровень глубины анализа текста ЕЯ, как и ЕЯ-запроса пользователя, в каждом конкретном случае зависит от особенностей целевой задачи, в том числе и от необходимых критериев оценки качества ее решения и их значений. Подобным образом результаты рассматриваемой задачи используются и при создании систем автоматического машинного перевода, автоматического реферирования, организации вопросно-ответного и диалогового взаимодействий с пользователем на ЕЯ и т. п.

Очевидно, что выбор подходящего метода из многообразия методов распознавания языка текста анализируемого документа (коротких слов, грамматических форм, статистический, алфавитный и т. д.) во многом связан с доступностью лингвистических ресурсов, которые могут быть ориентированы как на использование знаний о ЕЯ в пределах от уровня алфавита до лексико-грамматического уровня глубины ЕЯ, что вполне приемлемо по трудоёмкости для их использования в промышленных системах автоматической обработки текста, например системах информационного поиска, автоматического обнаружения заимствованных фрагментов текстовых документов, так и на знания, затрагивающие синтаксический и семантический уровни ЕЯ. Другим немаловажным фактором, влияющим на выбор метода, является доступность вычислительных ресурсов, в том числе ресурсов на базе технологии облачных вычислений, например, сервисы Google Cloud, Amazon Web Services, Microsoft Azure, Yandex.Cloud, которые, наряду с удачно подобран-

ными корпусами языковых данных, составляют основу получивших широкое распространение в настоящее время методов машинного обучения, в том числе на базе искусственных нейронных сетей (supervised-, unsupervised-, bootstrapping-, zero-shot-методы и т. п.). Однако, учитывая характерную особенность данной группы – достаточно затратную, в том числе по времени, фазу обучения модели машинного обучения (в особенности характерно для нейросетевых методов), при выборе целесообразно учитывать число языков, моделируемых параметров, архитектуру и т. д. Несмотря на возросшую популярность решений на базе нейросетевых моделей, построенных на архитектуре Transformer и обеспечивающих возможность обрабатывать всю последовательность текстовых элементов одновременно, что позволяет лучше учитывать контекст, особенно в длинных предложениях, модели на основе рекуррентных нейронных сетей (RNN), задействованные нами в результате экспериментов, часто используются для маркировки и классификации последовательностей, т. е. предсказания класса для каждого вектора признаков в последовательности и предсказания класса для всей последовательности соответственно способны обеспечить приемлемые результаты с меньшими затратами (временными и вычислительными). При этом сама последовательность состоит из предложений текстов как последовательности слов, символов и знаков препинания ЕЯ, представленной в виде матрицы, каждая строка которой соответствует вектору признаков, следующему в определенном порядке.

Для решения задачи автоматического определения языка текста методом машинного обучения исходный список языков [1] был расширен до 49 (чешский, польский, шведский и т. д.), что потребовало реорганизации исходного набора данных для обучения по принципу Wikipedia Language Identification Dataset (WLID) за счет его расширения недостающими сведениями для новых языков, в итоге обеспечив по 1000 предложений для каждого из 49 языков. Ресурсы сервиса Google Cloud использовались для обучения модели на основе двунаправленной сети с вентильными рекуррентными узлами (Gated Recurrent Unit). Ее тестирование на полученном текстовом корпусе, содержащем не вошедшие в обучающую выборку и не подвергавшиеся предредактированию, а значит, содержащие аббревиатуры, сокращения, грамматические и синтаксические ошибки, фрагменты на языках, отличающихся от основного языка документа и т. п. тексты, обеспечило точность 97,1%. Это позволяет сделать вывод о возможности получения приемлемых результатов решения задачи автоматического определения языка при относительной простоте реализации за счет применения специализированных программных библиотек (TensorFlow, Scikit-learn) и возможности использования полученной модели в условиях ограничений на доступные ресурсы или требования локального развертывания системы в рамках периметра безопасности организации. При этом наиболее трудоёмким и длительным процессом может являться построение самого корпуса текстов, определяющего как качество получаемого решения, так и его применимость с учетом возможных изменений требований к использованию создаваемой интеллектуальной системы обработки данных.

ЛИТЕРАТУРА

1. Крапивин Ю. Б. Автоматическое определение языка текстового документа для основных европейских языков // Информатика. 2011. № 3 (31). С. 112–117.