Исследование алгоритмов анализа и генерации технической документации

А.А. Андрюкеева, E-mail: andryukeeva@yandex.ru

Московский авиационный институт

Аннотация. *Работа посвящена исследованию алгоритмов обработки технической документации. В ходе работы были проанализированы существующие инструменты работы с документацией, разработан алгоритм валидации терминологии.*

Ключевые слова: *техническая документация, анализ текста, валидация документации, генерация документов, информационная система.*

Введение

В настоящее время работа с технической документацией стала неотъемлемой частью практически любого процесса создания товаров и предоставления услуг. От того, насколько качественно она составлена, зависит эффективность создания продукта и его использование. Техническая документация необходима на производственных предприятиях, при разработке программного обеспечения, при исполнении государственного задания и т. д.

Среди множества существующих проблем подготовки документации наиболее распространенными являются проверка соответствия структуры документации принятым стандартам, требованиям описываемого процесса или предмета, составление многочисленных однотипных документов, выявление сути документа, оперирование большим количеством сокращений и др.

Более узкой является проблема нарушения терминологии в тексте. Часто, в силу невнимательности автора или по незнанию, терминологическая целостность документа может быть нарушена. Например, в дипломной работе студент в введении может заявить, что создает информационную систему, а в заключении говорить о созданном интерфейсе.

Многие из указанных проблем могут решаться автоматически, тем самым существенно сократив трудозатраты разработки и обеспечив возможность больше времени уделить содержательной составляющей документов.

Для решения этих проблем была создана информационная система, которая позволяет анализировать документ на наличие типичных ошибок и осуществлять валидацию по заданным шаблонам. Система может применяться для создания шаблонов различных видов документов, анализа содержания документа, валидации структуры документа, валидации документа на соответствие шаблону.

Была поставлена задача расширить возможности созданной системы, за счет валидации стилей и шрифтов на соответствие заданным в шаблоне, детализации ошибок в отчете, обработку подразделов, генерации документов.

Для создания более гибкой системы так же необходимо реализовать редактор шаблонов, позволяющий пользователю задавать собственные настройки документа для валидации.

Не менее важной задачей является исследование алгоритма валидации терминологии, который решает задачу устранения противоречия, применяемых в тексте документа, терминов.

1. Проверка шрифтов документа

В документации важно, чтобы текст был оформлен в определенном стиле, единообразие должно распространятся на весь документ, тогда его чтение и верстка значительно упростится. Для этого применяются настройки шрифта, такие как выравнивание, отступы, межстрочный интервал, размер шрифта, его цвет и т. д. Кроме того, часто эти настройки жестко задаются в требованиях к документу.

Для решения задачи проверки параметров шрифтов был применен API Spire.Doc для Java от компании E-ICEBLUE [1]. Spire.Doc предназначен для программной обработки и создания документов Word без использования Microsoft Office.

В системе с использованием инструмента Spire.Doc реализован класс FontValidator, предназначенный для проверки шрифтов. Методы класса можно разделить на две группы: проверяющие стили абзаца и шрифт текста.

Метод validateParagraphStyle осуществляет проверку у абзаца межстрочного интервала, отступа первой строки и выравнивания. Метод возвращает результат проверки в виде списка элементов “правило, true/false”.

Метод validateTextStyle проверяет шрифт той части текста, которая имеет одинаковый стиль. Проверяется жирность, курсив, имя шрифта, его размер и цвет, а также регистр. Как и в методе проверяющем абзац возвращается отчет о проверке в виде списка.

Отчеты, возвращающиеся в методах проверки, дают возможность полноценно описать допущенные в документе ошибки. Для удобства

отслеживания полученных результатов проверок предложено отражать несоответствия в документе Word при помощи примечаний, выделяя текст, с допущенной ошибкой, и приписывая к нему расшифровку. Таким образом ошибки хорошо видно, и при этом текст документа не засоряется.

2. Редактор шаблонов и генератор документов

Шаблон документа представляет собой файл в формате XSD [2]. В системе с помощью специальных методов из входящего документа формируется шаблон, который сопоставляется с имеющимся. Например, по заготовленному шаблону «Технического задания» [3], пользователь может проверить свой документ, из которого генерируется файл XML, описывающий структуру: с заголовками, подразделами, содержимым разделов.

В готовых шаблонах описывается структура документа и правила проверки. Так как невозможно описать заранее все типы документов, которые пригодились бы пользователям для проверки, предлагается реализовать редактор шаблонов, где можно было бы задавать пользовательские параметры документа.

Для решения проблемы составления многочисленных однотипных документов предлагается реализация генератора документов, при помощи которого на основе стандартного или пользовательского шаблона будет создаваться документ Word. Работа с таким документом сильно упрощается, так как он уже имеет необходимую структуру и настройки шрифтов, и позволяет сконцентрироваться только над содержимым.

3. Алгоритм валидации терминологии

Так как единообразие текста документа должно распространяться не только на его оформление, но и на содержимое, необходимо проверять последовательность его изложения. Если автор неправильно использовал терминологию в своей работе, его необходимо предупредить об этом.

Для решения этой задачи предлагается следующий алгоритм:

1. Поиск ключевых слов.
2. Выделение окрестностей ключевых слов.
3. Проверка того, что в центре одинаковых окрестностей находятся одинаковые слова, являющиеся ключевыми.

Например, для работы, посвященной информационной системе из примера выше, во введении будет найдено следующее ключевое слово: «система».

Выделение окрестности ключевого слова требует нормализации, то есть необходимо исключить предлоги, союзы, частицы. После нормализации окрестность состоит из двух слов до ключевого и двух после: "разработка информационной системы (для) обучения автомобилистов".

Найденные пары слов необходимо найти в тексте и проверить на то, что между ними одно слово и его значение – "система", а не "интерфейс" или что-то еще. При отрицательном результате проверки употребления термина выводится найденная ошибка и отражается в примечаниях.

Автоматизация этого процесса позволит значительно сэкономить время на проверке документа и исключить человеческий фактор при проверке текста документа.

Заключение

Для решения более широкого спектра проблем обработки документации расширяются возможности созданной ранее системы для работы с технической документацией.

Реализованы методы проверки стилей абзацев и шрифтов текста с использованием инструментов Spire.Doc. Предложено решение проблемы детализации ошибок в документе Word при помощи примечаний, выделяя текст, с допущенной ошибкой, и приписывая к нему расшифровку.

Разработан проект редактора шаблонов для настроек пользовательских параметров и генератора документов, решающий проблему составления множества однотипных документов. Предложен алгоритм валидации терминологии, который решает задачу устранения противоречий применяемой терминологии.

Новая функциональность в разрабатываемой и развиваемой системе автоматического анализа технической документации позволит сократить время на подготовку документов техническими писателями и другими специалистами, а также повысить качество разрабатываемой документации.

Литература

1. Spire.Doc [Электронный ресурс] : API. – Дата обращения: 09.10.2020 –Режим доступа: https://www.e-iceblue.com/Introduce/doc-for-java.html#.X9dYJ9gzbIU

2. XML schema [Электронный ресурс] : учебник. – Дата обращения: 15.10.2020 – Режим доступа: https://www.w3.org/TR/xmlschema-0

3. ГОСТ 3.1105-2011. Единая система технологической документации. Формы и правила оформления документов общего назначения [Электронный ресурс] : – Введ. 2012-01-01. – Дата обращения: 20.10.2020 – Режим доступа: http://docs.cntd.ru/document/1200086391