Секция «Вычислительная математика и кибернетика»

Автоматическое улучшение синтезированных правил коррекции документов в формате LaTeX

Чувилин Кирилл Владимирович
Аспирант
Московский физико-технический институт, Факультет управления и прикладной математики, Рыбинск, Россия
E-mail: kirill.chuvilin@gmail.com

Многие конференции и издательства принимают материалы от авторов в формате LaTeX. Обычно авторские тексты содержат значительное количество типографических ошибок, связанных с несоблюдением требований к оформлению, исправление которых производится корректорами вручную. Можно минимизировать рутинную работу с помощью правил для автоматической коррекции. Но ручное описание таких правил приведёт скорее к увеличению трудозатрат.

В работе рассматривается способ автоматического синтеза правил коррекции по обучающей выборке, составленной из пар документов «черновик-чистовик», и последующего их улучшения на основе статистики применимости к черновикам и чистовикам.

Файлы формата LaTeX обладают естественной древовидной структурой (синтаксическим деревом), исследуя которую, можно получить всю необходимую информацию для описания корректорской правки. Узлы этой структуры называются токенами. Правила замены удобно формулировать именно для деревьев.

Для выявления различий между синтаксическими деревьями документов используется алгоритм, основанный на алгоритме Zhang-Shasha [3]. Алгоритм позволяет вычислять редактирующее расстояние (минимальное количество операций) между двумя деревьями и, кроме того, определять, какую операцию нужно применить к каждой вершине для реализации такого расстояния.

После построения различия между черновиком и чистовиком синтезируется набор правил, которые преобразуют дерево первого в дерево второго [2]. Каждое построенное правило характеризуется шаблоном (последовательностью соседних токенов с общим родителем) и типом локализатора (токена, к потомкам которого применяется шаблон).

Оценка качества правил производится на основе статистики их применимости к черновикам и чистовикам [1]. Эксперименты показывают, что построенные правила позволяют обнаружить значительное число ошибок, но частота срабатываний на чистовиках составляет, в среднем, 40% от частоты срабатываний на черновиках, что довольно много.

Для улучшения качества синтезируемых правил используется следующее решение. Начальный набор генерируется с минимально возможными шаблонами. Для каждого правила собирается статистика применимости для чистовиков: каждый случай применимости считается отрицательным прецедентом для правила. Если количество отрицательных прецедентов превышает некоторый порог, правило заменяется новыми, которые получаются из исходного с помощью увеличения шаблонов. Эксперименты показывают, что после таких модификаций качество правил улучшается, в среднем, на 30%.

Литература

1. ЧувилинК.В. Автоматический синтез и статистическая оценка качества правил коррекции документов в формате LaTeX // Труды 54-й научной конференции МФТИ. — 2011.

2. ЧувилинК.В. Синтез правил коррекции документов в формате LaTeX с помощью сопоставления синтаксических деревьев // Математические методы распознавания образов. — 2011. — 15.

3. ZhangK., ShashaD. Simple fast algorithms for the editing distance between trees and related problems// SIAM Journal of Computing, 1989. — No.18. — Pp.1245-1262.