УДК 004.8

Метод оценки качества работы алгоритмов исправления опечаток

Плюхин Д.А.
(Национальный исследовательский университет ИТМО, г. Санкт-Петербург)
Научный руководитель - к.т.н. Муромцев Д.И.
(Национальный исследовательский университет ИТМО, г. Санкт-Петербург)

В работе рассматривается задача исправления опечаток в тексте на естественном языке, исследуются существующие методы оценки качества решения данной задачи алгоритмами исправления опечаток, выделяются достоинства и недостатки существующих методов, а также описываются результаты разработки метода оценки алгоритмов исправления опечаток на основе произвольного датасета, состоящего из текстов на естественном языке.

Введение. Задача исправления опечаток является одной из наиболее распространенных задач обработки текстов на естественном языке, для решения которой используются как детерминированные и интерпретируемые алгоритмы, так и алгоритмы и модели машинного обучения, процесс работы которых носит стохастический характер. Данная задача встречается во многих контекстах, таких как упрощение работы с мессенджерами, повышение качества текста, вводимого с использованием текстовых редакторов и текстовых процессоров, в редакторах исходного кода. Помимо всего прочего модули исправления опечаток могут использоваться в качестве фрагмента пайплайна систем машинного обучения для решения таких задач, как повышение качества существующих датасетов перед их использованием для обучения или тестирования моделей, для предобработки экземпляров обучающей выборки, поступающих в развернутую модель с целью ее инкрементного обучения. Более того, задача исправления опечаток может быть расширена и обобщена на случай преобразования абстрактных элементов некоторой совокупности в элементы заранее заданной выборки, частным случаем которого является обеспечение помехоустойчивости информационных систем.

На сегодняшний день существует ряд методов оценки качества работы систем автоматического исправления опечаток. В частности, для данной цели используются следующие подходы:

1. Подход, основанный на применении специализированных датасетов для обучения и тестирования алгоритмов и моделей с подсчетом доли токенов, не содержащих опечаток. Данный подход является общепринятым при работе с моделями машинного обучения. Преимуществом подобного подхода является высокая воспроизводимость результатов и возможность переиспользования результатов работы и инкрементального повышения его качества, однако, поскольку процесс внесения опечаток недетерминирован и носит случайный характер, решение задачи формирования датасета для оценки качества работы систем исправления опечаток вручную отличается высокой сложностью;
2. Ручная оценка результатов работы системы с привлечением экспертов. В частности, данный метод оценки применялся при разработке итеративного подхода с подсчетом условной вероятности и использованием статистических матриц замены, удаления и вставки символов. Данный подход к оценке отличается высокой простотой с точки зрения алгоритма подсчета оценки, но в то же время характеризуется сложностью с точки зрения реализации, что ограничивает его практическое применение. Также для

обеспечения объективности получаемых результатов требуется объединение результатов, полученных несколькими экспертами, что обуславливает необходимость адаптации существующего или разработки нового метода для решения подобной задачи. С одной стороны, подобный подход может обеспечить наиболее высокое качество результатов оценки, но в то же время данный подход ассоциирован со сложностью поиска экспертов и организации процесса их работы;

3. Подход, основанный на использовании журналов взаимодействия пользователей с программной системой и отдельным подсчетом точности по каждому классу ошибок[1]. Данный метод реализует возможность всесторонней оценки качества работы системы, однако характеризуется необходимостью привлечения экспертов для разметки датасетов и необходимостью доступа к журнальным записям системы, с которой взаимодействуют реальные пользователи. В связи с этим подобный подход может быть неудобен с точки зрения сложности обеспечения конфиденциальности персональных данных пользователей.

Основная часть. По результатам приведенных результатов анализа существующих методов оценки качества работы систем исправления опечаток был предложен и реализован новый метод, лишенный ряда недостатков существующих методов. Данный метод является компромиссом между подходами, используемыми на сегодняшний день, и описывается следующим алгоритмом:

1. Формирование словаря опечаток, в котором каждому фрагменту текста без опечаток соответствует один или более вариантов текста с опечатками. Для обеспечения возможности расширенной оценки допускается формирование дополнительных аннотаций, в которых фиксируются свойства соответствующих фрагментов текста;
2. Формирование произвольного датасета, который может содержать любые виды аннотаций либо не аннотированного вовсе. Допускается использование любого фрагмента текста, в частности - общедоступных текстов художественных произведений;
3. Осуществление стохастической замены фрагментов текста из выбранного датасета на фрагменты с опечатками из словаря, сформированного на первом шаге. При итерационном подходе к оценке на данном шаге может быть реализовано ограничение на выполняемые замены для уточнения формируемых результатов (например, удаление, вставка или замена только латинских символов на другие латинские символы). Сформированный датасет необходимо сохранить и использовать на следующих шагах;
4. Выполнение оценки точности на версии датасета, сформированной на предыдущем шаге. Для этого необходимо для каждого фрагмента текста до и после замены на фрагмент, содержащий опечатки. По результатам сравнения увеличивается счетчик общего количества фрагментов текста и, если две версии фрагмента текста совпадают, счетчик фрагментов текста без опечаток. Далее необходимо выполнить деление полученного количества фрагментов текста без опечаток на общее количество фрагментов текста - полученное значение является результатом подсчета метрики accuracy;
5. Исправление опечаток в версии датасета, сформированной на шаге 3, результат необходимо сохранить для использования на следующем шаге;
6. Выполнение подсчета значения метрики accuracy (см. шаг 4) с использованием исходной версии датасета и версии датасета, сформированной на предыдущем шаге;

[1] Под классами ошибок здесь подразумеваются свойства фрагмента текста и наблюдаемой ошибки, в которых произошла ошибка - например, отдельно выделяются ошибки в названии бренда и отдельно - фонетически обусловленные опечатки

7. Вычисление разности значений метрики accuracy, полученных на шаге 6 и на шаге 4 - результат является оценкой качества работы алгоритма исправления опечаток;
8. На основе оценки качества, полученной на предыдущем шаге, формирование взвешенного среднего с нормализованным показателем времени работы алгоритма.

Выводы. Предложенный метод был реализован и применен для сравнения ряда модификаций алгоритмов по исправлению опечаток. Был сделан вывод о том, что предложенная реализация характеризуется удобством использования, высокой универсальностью, высокой гибкостью и низкой сложностью процесса формирования нового датасета. По результатам практического применения метода были сформулированы следующие выводы:

1. Как и ожидалось, алгоритм генерирует более высокую оценку качества работы для метода Демарау-Левенштайна по сравнению с оценкой, генерируемой для классического метода Левенштайна;
2. Использование матриц замен символов, элементы которых пропорциональны расстояниям между клавишами на клавиатуре QWERTY, а также результатам подсчета статистики замены символов пользователями по датасету, включающему 4291 вариант написания токенов с опечатками, не оказывает существенного влияния на итоговый результат;
3. В случае формирования среднего значения на последнем шаге предложенного метода с одинаковыми весами для качества работы алгоритма и усредненного времени его выполнения оптимальным является вариант, предусматривающий предварительную фильтрацию по префиксу из двух символов в случае поиска по словарю;
4. При тестировании алгоритма, предусматривающего предварительную фильтрацию с префиксом длиной в 3 символа качество работы системы существенно падает по сравнению с фильтрацией по префиксу из одного символа (в среднем на 62.18%).

Плюхин Д.А. (автор) _____

Муромцев Д.И. (научный руководитель) _____