**Преимущества и недостатки системы REALEC Inspector и пути её развития**

Иван Торубаров

06.12.17

**Введение. REALEC Inspector как новаторское средство. Преимущества продукта**

Разработка проекта по компьютерной оценке текстов REALEC Inspector в последние месяцы активно поддерживалась – был добавлен ещё один грамматический критерий, исправлены ошибки и отключены слабо предсказательные модели. На сегодняшний день REALEC Inspector представляет собой одно из немногих существующих средств для автоматического анализа учебных текстов. Он находится в открытом доступе, что делает его доступным и привлекательным для самых широких кругов целевой аудитории. При этом, однако, проект находится в стадии развития: текущий метод имеет ряд слабых мест и стратегии дальнейшего развития проекта ясно видны.

Для начала хотелось бы напомнить принцип работы REALEC Inspector. На текущий момент он автоматически определяет во введенном тексте на английском языке части речи, длину предложений и особые словосочетания, затем считает их относительное количество и сравнивает их со «средними по корпусу» величинами. Для показателей меньших, чем средние, система выводит сообщение о желательности их большего употребления в будущем, для остальных – напротив, о их достаточности.

В новом критерии с глагольными формами вместо «среднего по корпусу» уже немного другой подход: после расчёта относительных значений для худших и лучших работ мы предполагаем наличие линейной зависимости между количеством глагольных форм и оценкой, и далее вычисляем относительное количество форм в работе, оцененной в 60%. Это позволило лучше связать выдачу с реальным оцениванием, в отличие от сравнения со «средним» вне связи с оценкой.

Этот подход, однако, также не лишён недочётов, и на этом семинаре мы подробно рассмотрим все статистические проблемы, не позволяющие однозначно доверять REALEC Inspector как средству для предварительной оценки текста на текущий момент. Будут рассмотрены собственно проблемы, спектр возможных путей их решения и список текущих задач для начала следующего этапа разработки Inspector.

**Недостатки REALEC Inspector на текущий момент**

Итак, во-первых, зависимость в естественных процессах редко бывает линейной. Она может быть близка к линейной, но это необходимо выяснять и показывать. На текущий момент Inspector не учитывает, что распределение может быть и другим, мы сильно проигрываем по релевантности, допуская, что у нас все случаи приблизительно линейны – и это нельзя не учитывать при выдаче программы.

Кроме того, основывая подсчёт на сравнении между коллекциями «плохих» и «хороших» работ, мы не учитываем возможное распределение фактора между самими работами. Если папка «хороших работ» состоит из двух неравномерных частей с высоким и низким показателем фактора соответственно, это придаёт фактору качественно иную важность, чем в случае папки с относительно равномерным распределением признака.

Наконец, некой «общей проблемой» является то, что наши данные не выстраиваются в единую зависимость, а выглядят как множество точек, и линейная аппроксимация является самым грубым механизмом для работы с ними. Нам необходимо корректно применить более точные методы и сделать ключевым механизмом в программе предсказание по оценке текста, исходя из имеющихся факторов.

**Варианты решения проблем**

Одним из первых шагов для увеличения точности нашего инструмента, которые мы можем предпринять, является увеличение выборки. С добавлением в базу новых оценённых текстов возможно пересчитать текущие средние значения с учётом большего числа текстов, что повысит точность системы.

Вторым достаточно простым шагом может стать изменение текущего метода построения гистограмм в Inspector. Сейчас строятся гистограммы частот по данным всего корпуса, не только сокращая объём данных всего корпуса, представляя их в одном измерении, но и не связывая работы с оценками (а знание средней оценки по корпусу значительно повысило бы релевантность средства). Совмещение гистограмм для двух коллекций - плохих и хороших работ - повысит репрезентативность программы и добавит фактор оценки в представление, что сделает выдачу понятнее.

Далее необходимо назначить веса для признаков. Сейчас результаты по всем анализируемым критериям выдаются в рамках единой таблицы шрифтом одинакового размера. Пока модель не учитывает, что влияние факторов может быть различным (как и разница между их долями в плохих и хороших работах) и важность факторов также может варьироваться. Это необходимо исследовать при построении модели и, с учётом результатов, включить в итоговую выдачу программы.

Одним из главных предлагаемых улучшений является изменение направления вычисления результатов при разработке: сейчас метод основывается на подсчёте доли интересующих нас форм в коллекциях работ по оценкам и дальнейшем сравнении с ними автоматически рассчитанных долей. Для улучшения работы необходимо рассмотреть обратную связь: так как Inspector при анализе введённого текста работает с рассчитанными долями исследуемых форм в этом тексте, для создания корректной модели необходимо построить её на основе уже существующих данных таким образом, чтобы для имеющихся в корпусе оценённых текстов по имеющимся в них с удовлетворительной долей точности предсказывалась полученная оценка.

Оценку выгодно было бы построить, определив для совокупности факторов вероятность их “попадания” в коллекцию работ, оценённых определённым баллом - иным словом, кластеризовать. Необязательно представлять результат этой операции в выдаче, но применение её, как и прочих описанных методов, даст одновременно большую релевантность средству и позволит эффективнее отбирать тип представления результатов пользователю.

В качестве одного из дополнительных перспективных средств можно также рассмотреть возможность использования нейросетевых технологий при необходимости - для аппроксимации сложной функции или разработки семантических критериев оценки эссе.

**Текущие предложения**

Для успешного применения описанных методик улучшения REALEC Inspector я предлагаю ряд действий первой очерёдности, которые возможно выполнить в ближайшей перспективе. Для начала для применения подходов, связанных со сравнением данных с реальной оценкой, следует пересчитать относительные величины признаков для всех работ с известной оценкой. Это позволит проверить текущие контрольные величины и одновременно станет первым этапом сбора данных для разработки новых методов Inspector.

Полученные результаты было бы удобно занести в поддерживаемую базу данных. Это необходимо для более удобного обращения к данным впоследствии. Наконец, с той же целью хотелось бы предложить создать и поддерживать для каждого файла из корпуса файл разметки. Это позволит значительно экономить время при необходимости комплексного анализа корпуса в дальнейшем, когда REALEC Incpector, согласно плану, войдёт во вторую фазу развития, и для поддержания постоянной актуальности данных.