Индексы удобочитаемости как мера оценки сложности текста

О. Н. Ляшевская

Доклад на семинаре НУГ

Интуитивное понятие сложности / легкости текста для чтения и связанной с этим скорости чтения и понимания текста в лингвистике XX века было формализовано в виде индексов удобочитаемости (readability). В их основе лежит ряд презумпций — например, о том, что

1. короткие предложения читать легче, чем длинные;
2. длинные слова затрудняют чтение;
3. читатель замедляется или «спотыкается», встречая низкочастотные и / или незнакомые ему слова и т. п.

В наиболее общепринятом варианте оценка сложности текста не вытекает из незнания читателем темы, сложности материала как такового (информационная сложность), не зависит от способа шрифтового оформления, верстки блоков текста и других аспектов графического дизайна (визуальная сложность), но связана с особенностями языка текста (подбором лексики, выбором грамматических форм, строением предложений и т. п.), логической организацией текста, выстроенностью аргументации, использованием ожидаемых риторических приемов и другими аспектами организации дискурса[1].

Представляется, что оценка сложности также связана с субъективными факторами, такими как: языковой опыт (по-разному воспринимают текст носитель языка (L1) / изучающий язык как неродной (L2) / «несовершенный» носитель, например, носитель эритажного языка (LH) или студент, осваивающий плохо знакомый ему жанр академического письма);
возраст носителя (например, если читающий – ребенок или пожилой человек);
мотивированность читателя;
индивидуальные когнитивные, психологические, неврологические особенности читателя.

В связи с разнообразием ситуаций, в которых встречаются Текст и его Читатель, проводятся исследования в отдельных областях:

1. оценка удобочитаемости упражнений и учебных текстов для иностранцев, изучающих язык как неродной (L2);
2. экспертиза школьных учебников, экзаменационных тестов и других материалов (для носителей L1);

1 Обратим внимание, что наличие таблиц и иллюстраций может также существенно облегчать понимание текста, однако взаимодействие текстовых и графических стимулов выходит за рамки рассматриваемой нами проблемы.

3. оценка читабельности деловой документации; рекламных материалов; медицинской документации;
4. оценка текстов веб-сайтов с точки зрения привлекательности для целевой аудитории (например, для детей-подростков);
5. создание рекомендательных систем для библиотек и ряд других.

Доклад представляет краткий обзор методов измерения сложности в свете разработки онлайн-ресурсов, позволяющих пользователю оценить любой выбранный им русскоязычный текст.

Метрики сложности

Простые метрики удобочитаемости, такие как:
Flesh-Kincaid [Kincaid et al., 1975];
Koleman-Liau [Coleman, Liau, 1975];
SMOG [McLaughlin, 1969],
строятся на характеристиках, которые легко получить из текста без привлечения дополнительных лингвистических ресурсов или разметки. К таким характеристикам относятся, например, средняя длина слова в словах или слогах, количество слов длиной более 5 символов, средняя длина предложения в словах или слогах, количество знаков препинания и т. д.

В большинстве случаев метрики удобочитаемости представляют собой формулу линейной регрессии вида:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ...,$$

где X_1, X_2, X_3 и т. д. – отдельные характеристики текста, а Y – индекс удобочитаемости, который соответствует либо возрасту читателя («текст для детей возраста N лет»), либо уровню образования, либо уровню владения языком (L2), см. пример на Рис. 1. В некоторых случаях индекс не имеет мнемонической интерпретации, но задает известный диапазон значений, который позволяет сравнивать между собой текстовые материалы как в пределах одного языка, так и между языками. Коэффициенты β_0, β_1 и т. д. подбираются таким образом, чтобы на заданной выборке текстов оценка Y наилучшим образом соответствовала оценкам, поставленным экспертами.

Разработчики анонсировали следующий Doom ещё в 2008 году. В конце прошлого года из компании ушёл «отец» серии Джон Кармак, чтобы заниматься созданием очков виртуальной реальности, так что завершают проект уже без него.

Источник: https://vc.ru/p/doom-4

Индекс: 7.27, 7-9 класс, 12-14 лет

Рис. 1. Пример оценки удобочитаемости веб-сайта.

Другие характеристики, которые привлекаются при оценке сложности текстов, можно разделить на лексические, морфологические, синтаксические и дискурсивные. Например, индекс Dale-Chall [Chall, Dale, 1995] анализирует, входят ли слова, встречающиеся в тексте, в 3000 наиболее частотных слов. Аналогично, может оцениваться доля слов за пределами лексических минимумов для иностранцев (A1, A2 и т. п.), средняя частота лемм (вычисляемая на основе частотных словарей), доля абстрактных слов, аббревиатур, служебных слов, лексическое разнообразие в конкретном тексте (чем чаще повторяются отдельные словоформы или леммы, тем, предположительно, легче текст).

Морфологические факторы включают долю разных частеречных классов в тексте, а также присутствие слов с определенной словообразовательной структурой (например, существительных с тем или иным суффиксом).

Синтаксические факторы оценивают сложность синтаксической структуры предложений, в частности, среднюю долю подчиненных, сочиненных и т. п. клауз, причастных и деепричастных оборотов.

Дискурсивные характеристики, которые используются реже всего из-за сложности их автоматического распознавания, могут учитывать среднюю долю диалогических единиц на предложение, количество анафорических местоимений и других слов, требующих для понимания разрешения кореференции, сложность риторической структуры и т.д.

Помимо линейно-регрессионных формул, предлагаются и более сложные модели (ср. индекс SMOG, в котором характеристики перемножаются друг на друга, и из этого значения извлекается квадратный корень). В пользу того, что функция $Y=f(\beta_i, X_i)$ должна быть устроена нелинейно, говорят и те соображения, что языковая компетенция развивается неравномерно для разных возрастных периодов и даже деградирует у пожилых и больных носителей, а важность таких факторов, как длина слова, может меняться – например, становиться малорелевантной для взрослых образованных носителей языка.

Очевидно и то, что многие перечисленные факторы не-независимы друг от друга. В этой связи неудивительно, что в последнее время появляется все больше моделей-классификаторов, которые обучаются на размеченных коллекциях текстов на большом наборе предлагаемых исследователем факторов, которых может быть несколько сотен.

Так, в работе [Reynolds, 2016] модель RandomForest, работающая на 179 факторах, смогла аккуратно предсказать сложность текстов РКИ согласно принятым уровням (A1, A2, B1, B2, C1, C2) в 66% случаев, причем только в 8% она ошибалась более чем на один уровень, см. Рис. 2.

	A1	A2	B1	B2	C1	C2
A1	234	120	48	0	0	0
A2	41	553	192	17	0	0
B1	16	76	1130	90	5	5
B2	1	57	311	478	83	4
C1	1	20	66	98	394	6
C2	0	3	40	58	9	78

Рис. 2. Кросс-валидация предсказания уровня текстов РКИ в [Reynolds, 2016: 135]: предсказанные системой ответы сопоставляются с классом, приписанным тексту в тестовой коллекции.

Вместе с тем, в той же работе отмечается, что результаты машинного обучения сильно зависят от коллекции, на которой проводится оценка, и требуется разрабатывать модели, более устойчивые к ошибкам и шумам в обучающей выборке (например, если текст атрибутирован экспертом неверно или если в тексте представлено одно, но очень длинное слово). Вклад каждого из предлагаемых исследователем факторов в работу классификатора также оценивается по-разному, в зависимости от выбранного способа ранжирования. Например, в одних случаях системы ранжирования факторов выше рейтингуют лексические факторы, а в других – морфологические.

Экспертная оценка удобочитаемости

Как уже было сказано, метрики удобочитаемости настраиваются, как правило, по коллекциям, размеченным экспертами. Например, в [Karpov, 2014; Reynolds, 2016] использовались коллекции текстов для чтения в курсах РКИ, такие как CIE (МГУ), Red Kalinka: «Russian books with audio», TORFL (тексты на понимание), «Златоуст» (тексты для чтения). Коллекция LingQ, собранная методом краудсорсинга, включает оценки пользователей — как правило, преподавателей РКИ. Однако из-за того, что оценку в этом случае ставит один эксперт (человек, выкладывающий текст на сайт) и критерии оценки четко не определены и могут варьировать от эксперта к эксперту, степень доверия к таким оценкам будет ниже.

В [Tanaka-Ishii et al., 2010] (для английского и японского языка) применен другой стандарт оценки, когда тексты в тесте попарно сравниваются друг с другом носителями

языка. В результате создается граф с попарно упорядоченными текстами, и классификатор обучается и тестируется на его оценках.

Наконец, имеет смысл понимать разброс оценок носителей языка и изучающих язык как неродной для отдельных текстов и однородных коллекций текстов. Например, в [Кошелева, 2015] (на материале польских текстов для студентов-иностранцев) проводится сопоставление между множеством индексов удобочитаемости, разработанных для польского языка, и множеством оценок носителей. Не исключено, что разные лингвистические факторы оказываются по-разному важны для разных носителей.

Интересны попытки включить более «объективные» критерии в сопоставление текстов по сложности. В пионерской работе [Микк, 1974] было предложено использовать для оценки сложности время реакции носителя при восприятии слов. В [Петрова, Окладникова, 2009], наряду с субъективными оценками сложности, получаемыми от испытуемых, используется мера времени чтения текста. В [Шпаковский, 2012] описано, что наряду с просьбой оценить сложность по шкале испытуемые получали контрольные задания: ответить на вопросы к тексту, заполнить пропуски, упорядочить абзацы по порядку.

6. Заключение

Завершая обзор, мы бы хотели обратить внимание на две тенденции в анализе удобочитаемости, в т. ч. с помощью онлайн-сервисов. Во-первых, это привлечение к анализу не только низкоуровневых единиц (длина слова, длина предложения), но и единиц более высокого уровня (например, морфологических и синтаксических характеристик). Во-вторых, видна тенденция к представлению отдельных критериев — в дополнение к единому индексу удобочитаемости. Последнее особенно полезно в случае, если онлайн-сервис используется самим автором текста для того, чтобы проанализировать текст и сделать его более доступным для чтения, улучшая отдельные параметры.

По мере настройки индексов удобочитаемости на разных коллекциях текстов и с помощью различных классификаторов становится ясно, что «вес» тех или иных критериев нельзя определить раз и навсегда (ср. коэффициенты в формуле Флеша-Кинкейда), и что предлагаемые модели пока слишком зависят от обучающей коллекции, состава текстов, схемы определения экспертных оценок и случайных «шумных» данных в самих текстах.

Источники

Chall J. S., Dale E. Readability Revisted: The New Dale-Chall Read- ability Formula. Brookline Book, 1995.

Coleman M., Liau T. L. A computer readability formula designed for machine scoring. Journal of Applied Psychology, Vol. 60, 1975. Pp. 283–284.

Kincaid J. P., Fishburne R. P. J., Rogers R. L., Chissom, B. S. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN, 1975.

McLaughlin H. SMOG grading – a new readability formula. Journal of Reading, № 22, 1969. Pp. 639–646.

Reynolds R. Russian natural language processing for computer-assisted language learning. PhD dissertation, UiT: The Arctic University of Norway, 2016.

Tanaka-Ishii K., Tezuka K., Terada H. Sorting texts by readability. Computational Linguistics, Vol. 36(2), 2010. Pp. 203-227.

Кошелева Д. Определение уровня языковой сложности текстов для изучающих польский язык как иностранный. Дипломная работа. М.: НИУ ВШЭ, 2015.

Микк Я. А. Методика разработки формул читабельности // Советская педагогика и школа, № 9, 1974. С. 78-163.

Оборнева И. В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров. Дисс... канд. пед. наук. М., 2006.