Quantile-base approach to measuring cognitive complexity of text

Maksim Eremeev[1]⋆ maks5507@yandex.ru
Konstantin Vorontsov[1,2] vokov@forecsys.ru
[1]Moscow State University
[2]Moscow Institute of Physics and Technology

The indexes of readability or cognitive complexity of the text are used to compare educational texts, websites, business and promotional materials. It seems promising to use them also in information retrieval and text recommendation systems for ranking search results in the order "from simple to complex", "from popular, educational, and surveys to highly specialized". This principle of ranking can be used in educational platforms, personalized exploratory search and recommendation systems aimed at automating the process of studying new subject areas by the user.

The well-known readability indices use simple quantitative features of the text, such as the average word length, the frequency of long words, the average sentence length, the average length of subordinates or composed clauses, etc. Discursive features are less commonly used: the number of anaphoric connections, the complexity of the rhetorical structure, etc.

All these estimations have two main disadvantages.

Firstly, they do not take into account the relative nature of the complexity concept itself. The complexity of the text should depend on which texts can be considered as simple, and for which readership, including language experience, age, education, and profession factors.

Secondly, they do not unify all levels of the language: phonetic, morphological, syntactic, and discursive.

We propose a quantile-based approach to the cognitive complexity of a text, free of these shortcomings.

Firstly, the complexity is determined with respect to a representative reference corpus of texts that we consider as simple for the implied readership. Depending on the objectives of the study, the reference corpus can be an electronic library of fiction, Wikipedia, educational literature for a given specialty, a topic or a subset of topics from the topic model of a multidisciplinary text collection.

Secondly, for each level of the language, its own alphabet of tokens is determined. These are: phonemes or letters for the phonetic level; morphemes or syllables for the morphological one; words or terms for the lexical one; types and lengths of syntactic links, rhetorical structures, or sentences for the discursive level.

Our mathematical formalism is based on the following ideas from neurophysiology and psychophysiology. The perception of text or speech goes through several stages of decoding, approximately corresponding to the levels of the language. At each stage, the tokens of the corresponding level are recognized and analyzed. Decoding processe takes place in various areas of the nervous system, from the visual and auditory analyzers to the cerebral cortex. Each zone is specialized in decoding a specific code. Having completed the decoding, the zone goes into a refractoriness state and is restored for some time. In a refractory state, the zone is not able to decode the same code. If it occurs again in the input sequence, then another zone will be used for decoding. If the frequency of the code in the text significantly exceeds the comfortable (evolutionarily determined) frequency, then the recipient's brain is subjected to additional stress being unable to process it in the usual mode.

Thus, the text is complicated if it contains abnormally many rare tokens, which are unfamiliar, incomprehensible or unusual for the recipient. This intuitive definition of complexity allows for a natural statistical formalization.

Having a representation of the reference corpus of texts in a given alphabet, we calculate the empirical frequency distribution for each token over the reference corpus. Let us have to measure the complexity of a text, not necessarily from the reference corpus. A token is considered complex in the text if it appears abnormally often in the text, if compared to the reference corpus. The abnormally high frequency of the token is determined through the quantile of the empirical distribution of its frequency. We take the percentage of complex tokens in the text as a complexity measure of this text at a given language level with respect to a given reference corpus. Thus, we use a unified quantile-based approach at all levels of the language, and all level-wise complexities are measured as a percentage of the text length. This makes it easy to build aggregate complexity measures for all levels.

For the empirical comparison of complexity measures of the text, we prepared a set of pairs of Wikipedia articles and marked it up on the crowdsourcing platform Yandex.Toloka. First, we built a topic model of Wikipedia, from which we selected 10 thousand pairs of articles of similar topic distributions and lengths (an example of a suitable pair is the articles "Lead" and "Tin"). Then annotators labeled these pairs by four marks: "the first article is simpler than the second one", "the second article is simpler than the first one", "the articles are approximately the same in complexity", "the articles cannot be compared because they relate to different topics". After discarding incomparable pairs, 8 thousand of pairs of articles remained.

Complexity measures were calculated from the reference corpus of 1.5 million Russian-language Wikipedia articles. In a series of experiments, our complexity measures were agreed with assessors in 81–84% pairs, whereas the well-known readability indices ARI and Flash-Kincaid yielded only from 41% to 58% matches.

The research is funded by RFBR, grant 17-07-01536.

[1] *M.Eremeev, K.Vorontsov.* Semantic-Based Text Complexity Measure. Recent Advances inNatural Language Processing, RANLP-2019.

Квантильный подход к оцениванию когнитивной сложности текста

Еремеев Максим Алексеевич[1]⋆ maks5507@yandex.ru
Воронцов Константин Вячеславович[1,2] voron@forecsys.ru
[1]Московский государственный университет им. М. В. Ломоносова
[2]Московский физико-технический институт (НИУ)

Индексы удобочитаемости или когнитивной сложности текста используются для сравнения учебных текстов, веб-сайтов, деловых и рекламных материалов. Представляется перспективным их применение также в системах разведочного информационного поиска и текстовых рекомендательных системах для ранжирования поисковой выдачи в порядке «от простого к сложному», «от популярного, учебного и обзорного к специализированному и узко профессиональному». Такой принцип ранжирования может быть использован в образовательных платформах и поисково-рекомендательных системах, нацеленных на автоматизацию процесса изучения новых предметных областей пользователем.

Известные индексы удобочитаемости используют простые количественные признаки текста, такие, как средняя длина слов, доля длинных слов, средняя длина предложений, средняя длина подчинённых и сочинённых клауз, и т. д. Реже используются дискурсивные признаки: количество анафорических связей, сложность риторической структуры, и т. д.

Эти методы имеют два основных недостатка.

Во-первых, они не учитывают относительную природу самого понятия сложности. Оценка сложности текста должна зависеть от того, какие тексты мы согласны считать простыми, и для какой читательской аудитории, включая факторы языкового опыта, возраста, образования, профессии.

Во-вторых, они не позволяют учитывать одновременно и единообразно все уровни языка: фонетический, морфологический, синтаксический, дискурсивный.

Мы предлагаем квантильный подход к оцениванию когнитивной сложности текста, свободный от указанных недостатков.

Во-первых, сложность определяется относительно представительного референтного корпуса текстов, которые считаются

простыми для выбранной читательской аудитории. В зависимости от целей исследования референтным корпусом может быть электронная библиотека художественной литературы, Википедия, корпус учебной литературы по заданной специальности, тема или подмножество тем из тематической модели мультидисциплинарной текстовой коллекции.

Во-вторых, для каждого уровня языка определяется свой алфавит токенов: для фонетического — фонемы или буквы; для морфологического — морфемы или слоги; для лексического — слова или термины; для синтаксического — типы и длины синтаксических связей; для дискурсивного — типы и длины риторических структур или предложений.

Предлагаемая математическая формализация понятия сложности основана на следующих представлениях нейрофизиологии и психофизиологии. Восприятие текста или речи проходит через несколько этапов декодирования, приблизительно соответствующих уровням языка. На каждом этапе происходит распознавание и анализ токенов соответствующего уровня. Процессы декодирования происходят в различных зонах нервной системы, от зрительного и слухового анализаторов до коры головного мозга. Каждая зона специализирована на декодировании определённого кода. Завершив декодирование, зона переходит в состояние рефрактерности и некоторое время восстанавливается. Находясь в состоянии рефрактерности, зона не способна декодировать тот же код. Если он снова встретится во входном сигнале, то для декодирования будет задействована другая зона. Перераспределение ресурса может приводить к снижению эффективности анализа сигнала на последующих этапах, реализующих более сложные и эволюционно более молодые функции сознания и мышления. Если частота кода в тексте существенно превышает комфортную (эволюционно обусловленную) частоту, то мозг реципиента не успевает обрабатывать его в обычном режиме и испытывает дополнительную нагрузку.

Таким образом, текст является сложным для восприятия, нагруженным, если он содержит аномально много редких токенов — незнакомых, непонятных или непривычных для реципиента. Это интуитивное определение сложности допускает естественную статистическую формализацию.

Имея представление референтного корпуса текстов в заданном алфавите, мы вычисляем эмпирические распределения частот для каждого токена по референтному корпусу. Теперь допустим, что требуется оценить сложность некоторого текста, не обязательно из референтного корпуса. Токен считается сложным в данном тексте, если он встречается в нём аномально часто по сравнению с референтным корпусом. Аномально высокая частота токена определяется через квантиль эмпирического распределения его частоты. Доля сложных токенов (в процентах) в данном тексте принимается за оценку его сложности на заданном уровне языка относительно заданного референтного корпуса. Таким образом, на всех уровнях языка используется единый квантильный подход, и все оценки сложности измеряются в процентах от длины текста. Это облегчает построение агрегированных оценок сложности по всем уровням.

Для эмпирического сравнения различных оценок когнитивной сложности текста мы подготовили набор пар статей Википедии и разметили его на краудсорсинговой платформе Яндекс.Толока. Сначала была построена тематическая модель Википедии, с помощью которой мы отобрали 10 тысяч пар статей схожей тематики и длины (пример подходящей пары — статьи «Свинец» и «Олово»). Затем эти пары статей предъявлялись асессорам, которых просили поставить одну из четырёх отметок: «первая статья проще второй», «первая статья сложнее второй», «статьи примерно одинаковы по сложности», «статьи невозможно сравнить, так как они относятся к разным темам». После отбрасывания несравнимых пар осталось 8 тысяч пар статей.

Квантильные оценки сложности вычислялись по референтному корпусу из 1,5 миллионов русскоязычных статей Википедии. В серии экспериментов предложенные оценки сложности совпадали с асессорскими на 81–84% пар, тогда как известные индексы удобочитаемости ARI и Флеша–Кинкейда давали лишь от 41% до 58% совпадений.

Работа поддержана грантом РФФИ № 17-07-01536.

[1] *M.Eremeev, K.Vorontsov.* Semantic-Based Text Complexity Measure. Recent Advances inNatural Language Processing, RANLP-2019.