Квантильный подход к оцениванию когнитивной сложности текста

Еремеев Максим Алексеевич,
Воронцов Константин Вячеславович
(ВМК МГУ • МФТИ • ФИЦ ИУ РАН)

Москва • 26-29 ноября 2019

Содержание

Индексы удобочитаемости текста

Индексы удобочитаемости — меры сложности восприятия текста

- Наиболее распространенные:
  — *индекс Флеша*,
  — *Automated Readability index (ARI)*,
  — *SMOG-index*
- Вычисляются на основе простых параметров:
  — длины предложений и слов,
  — среднего количества букв в слове
  — среднего количества слов в предложении
- содержат большое количество констант, которые подбираются отдельно для каждого языка

Flesh, R. How To Test Readability. 1951.
Senter, R.J. and Smith, E.A. Automated Readability Index. 1967.

Индексы удобочитаемости текста

Автоматический индекс удобочитаемости

$$ARI(d) = 4.71 \times \frac{c}{w} + 0.5 \times \frac{w}{s} - 21.43,$$

c — общее количество букв в документе d,
w — общее количество слов,
s — общее количество предложений в d.

Индекс Флеша

$$Flesh(d) = 206.835 - 1.015 \times \frac{w}{s} - 84.6 \times \frac{syl}{w},$$

syl — общее количество слогов.

Ранжирование документов в порядке чтения

Reading Order — порядок документов от общих и простых к сложным и узкоспециализированным.

Применения:

- Сортировка выдачи разведочного поиска
- Построение персонализированных образовательных траекторий

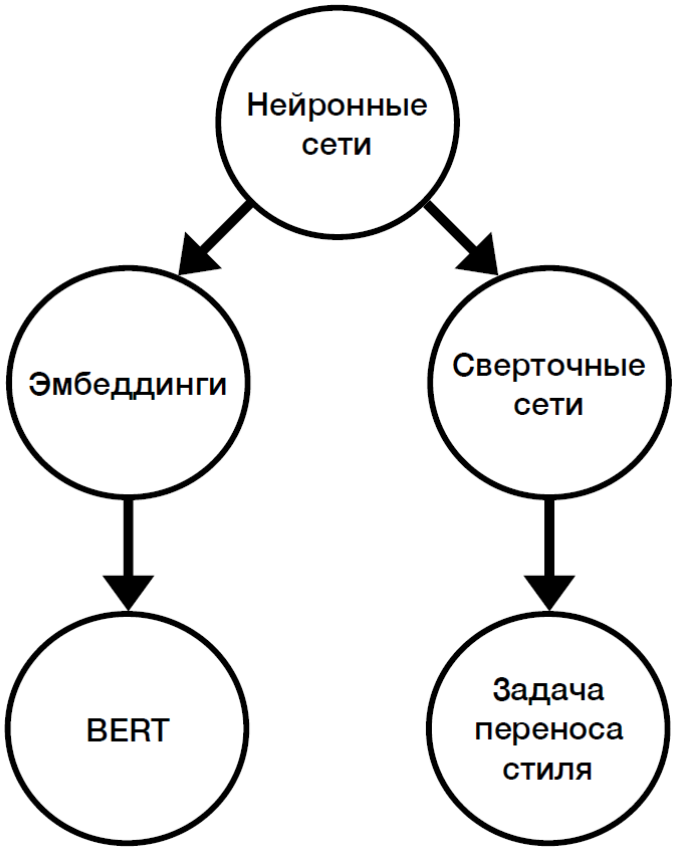
Подход, основанный на тематическом моделировании:

- Оценка общности текста — энтропия тематического вектора
- Оценка близости документов — косинусное расстояние между тематическими эмбеддингами
- В цепочку попадают близкие документы в порядке убывания общности

Georgia Koutrika, Lei Liu, Steven Simske. Generating Reading Orders over Document Collections. 2015.

Ранжирование документов в порядке чтения

Порядок может быть не только списком, но и деревом.



Общности и близости недостаточно для того, чтобы строить полные целочки. Нужны дополнительные параметры.

Психофизиологическая нагрузка декодирования речи

Что такое «тяжёлый текст» и чем он отличается от лёгкого?

Пример. Частота буквы р в русском языке 0.04, но здесь 0.17:

Гипотезы из нейрофизиологии и психофизиологии речи:

- декодирование каждого элемента языка вызывает нагрузку определенной зоны коры головного мозга
- зоне требуется *время рефрактерности* для восстановления
- в ходе эволюции языка устанавливаются существенно неравномерные распределения частот (закон Ципфа)
- в ходе освоения языка для высокочастотных элементов устанавливаются более короткие периоды рефрактерности

А.А.Биркин. Код речи. СПб.: Гиппократ, 2007.

Психофизиология кода речи

«Код речи» — программа для диагностики нагрузок декодирования и оптимизации восприятия текстов (Биркин)

Возможности:

- оценивание динамической нагрузки восприятия текста
- пересчет нагрузок и подсветка сложных элементов в тексте при его модификации с целью упрощения

Ограничения:

- фиксированный референтный корпус
- используются только частоты букв, длины слов, количество логических связок в предложении
- используются гауссовские приближения распределений частот и дисперсионные оценки

А.А.Биркин. Природа речи. М.: Ликбез, 2009.

Идея комплексного оценивания сложности текста

Основные положения предлагаемого подхода:

- *уровни языка*: фонетический, морфологический, лексический, синтаксический, дискурсивный
- на уровне h текст представляется в виде последовательности *токенов* алфавита A_h
- *сложность текста* на уровне h — это доля токенов, имеющих аномально высокую нагрузку
- нагрузка токена *аномально высокая*, если она превышает 95%-ю квантиль его нагрузки в референтном корпусе
- *референтный корпус* — тексты, которые можно считать простыми для выбранной читательской аудитории

M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

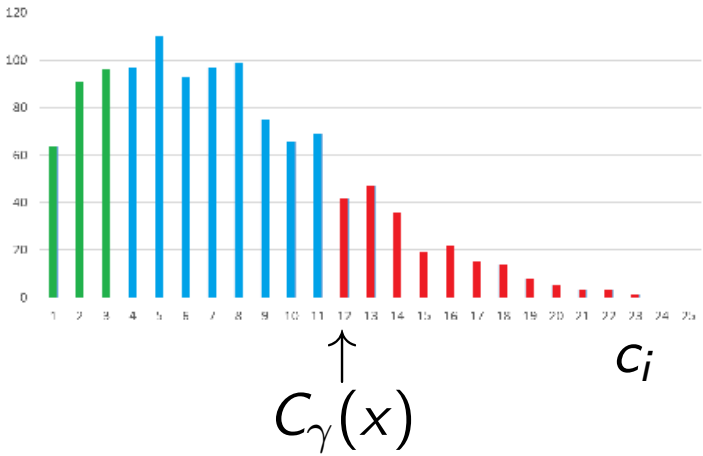
Квантильный подход к оцениванию сложности текста

$x_1, \ldots, x_n$ — последовательность токенов из A_h в тексте d;
c_i — *нагрузка* (оценка сложности) токена x_i
w_i — *вес нагрузки* токена x_i

Оценка сложности текста — суммарный вес сложных токенов:

$$W(d) = \sum_{i=1}^{n} w_i \left[c_i > C_\gamma(x_i)\right]$$

$C_\gamma(x)$ — γ-квантиль распределения сложности токена x в *референтном корпусе* несложных текстов



M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

Два типа оценок токенов — частотные и сложностные

Частотная оценка нагрузки токена

- для каждого токена $a \in A_h$ строится эмпирическое распределение значений его нагрузок $\{c_i : x_i = a\}$ по референтному корпусу несложных текстов
- $c_i = f(r_i)$ — убывающая функция расстояния r_i от токена x_i до его предыдущего вхождения
- примеры: частоты букв, n-грамм, слов

Сложностная оценка нагрузки токена

- одноэлементный алфавит $A_h = \{a\}$, строится эмпирическое распределение всех нагрузок $\{c_i\}$ по референтному корпусу несложных текстов
- c_i — числовая характеристика сложности элемента x_i
- примеры: длина слова, предложения, синтаксической связи

Фонетический уровень: токены — это буквы

r_i — расстояние от токена x_i до его предыдущего вхождения:

$$\ldots \quad \boxed{x_{i-r_i} = a} \quad \underbrace{x_{i-r_i+1} \quad x_{i-r_i+2} \quad \ldots \quad x_{i-2} \quad x_{i-1} \quad \boxed{x_i = a}}_{r_i} \quad \ldots$$

Если предыдущего вхождения нет, доопределяем r_i «через хвост» (тогда сумма r_i по всем вхождениям $x_i = a$ равна длине текста n):

токен	г	е	р	о	й	н	а	ш	е	г	о	в	р	е	м	е	н	и
r_i исходное	–	–	–	–	–	–	–	–	7	9	7	–	10	5	–	2	11	–
r_i доопред.	9	4	8	11	18	7	18	18	7	9	7	18	10	5	18	2	11	18

Варианты определения c_i:

$$c_i = \bar{r}(x_i) - r_i, \qquad c_i = 1/r_i$$

где $\bar{r}(x)$ — среднее r_i токена x в *референтном корпусе*.

А.А.Биркин. Природа речи. М.: Ликбез, 2009.

Морфологический уровень: токены — это буквенные n-граммы

Варианты определения токенов:

- слоги,
- буквенные n-граммы,
- буквенные n-граммы, не сохраняющие порядок букв,
- морфемы (приставки, корни, суффиксы, окончания)

Пример.

1-граммы	г	е	р	о	й	н	а	ш	е	г	о	в	р	е	м	е	н	и

n-граммы, сохраняющие порядок букв

2-граммы	ге	ер	ро	ой		на	аш	ше	ег	го		вр	ре	ем	ме	ен	ни	
3-граммы	гер	ero	рой			наш	аше	шег	его			вре	рем	еме	мен	ени		

n-граммы, не сохраняющие порядок букв

2-граммы	ге	ер	ор	йо		ан	аш	еш	ге	го		вр	ер	ем	ем	ен	ин	
3-граммы	гер	еор	йор			анш	аеш	геш	гео			вер	емр	еем	емн	еин		

Какой смысл брать n-граммы без сохранения порядка букв

1. Меньше словарь, надёжнее эмпирические распределения
2. Мозг легко обрабатывает локальные перестановки букв:

По рзелульаттам илссеовадний одонго анлигйского унвиертсиета, не иеемт занчения, в кокам пряокде рсапожолены бкувы в солве. Галвоне, чотбы преавя и псолендяя бквуы блыи на мсете. Осатьлыне бкувы мгоут селдовтаь в плоонм бсепордяке, все рвано ткест чтаитсея без порбелм. Пичрионй эгото ялвятеся то, что мы не чиатем кдаужю бкуву по отдльенотси, а все солво цликеом.
По результатам исследований одного английского университета, не имеет значения, в каком порядке расположены буквы в слове. Главное, чтобы первая и последняя буквы были на месте. Остальные буквы могут следовать в полном беспорядке, все равно текст читается без проблем. Причиной этого является то, что мы не читаем каждую букву по отдельности, а все слово целиком.

Лексический уровень: токены — это слова

Варианты определения токенов для частотных оценок:

- слова до лемматизации,
- слова после лемматизации,
- словные n-граммы,
- словные n-граммы, не сохраняющие порядок слов,
- термины из тезауруса или автоматически выделенные

Варианты сложностных оценок нагрузки:

- c_i — длина слова
- $c_i = \frac{1}{\text{count}(x_i)}$ — редкость слова в референтном корпусе, где $\text{count}(x_i)$ — частота слова

Синтаксический уровень

Используется синтаксический парсер UDPipe или SyntaxNet

Варианты определения токенов для частотных оценок:

- грамматические структуры — *синтагмы*,
  в которых отброшены слова и оставлены теги частей речи и/или членов предложения

Варианты сложностных оценок нагрузки:

- c_i — длина синтаксической связи с родительским словом
- c_i — количество связей с подчинёнными словами
- c_i — суммарная длина связей с подчинёнными словами

Дискурсивный уровень

Варианты определения токенов для частотных оценок:

- типы риторических структур

Варианты сложностных оценок нагрузки:

- c_i — число слов в предложении
- c_i — число логических связок в предложении (и, или, значит, который, чтобы, ... около 150 выражений)
- c_i — суммарная длина кореферентных связей со словами в данном предложении и соседних предложениях

Обучаемая линейная модель когнитивной сложности текста

Пусть $W_k(d), k = 1, \ldots, K$ — различные оценки сложности.

Линейная агрегированная оценка сложности с параметрами α_k:

$$W(d, \alpha) = \sum_{k=1}^{K} \alpha_k W_k(d), \quad \alpha_k \geqslant 0.$$

Данные экспертного сравнения пар документов:
$d \prec d'$ — документ d' сложнее документа d.

Критерий обучения агрегированной оценки:

$$\sum_{d \prec d'} \mathscr{L}\Big(\underbrace{W(d', \alpha) - W(d, \alpha)}_{\text{pair-wise margin}}\Big) \to \min_{\alpha},$$

где $\mathscr{L}(M)$ — гладкая невозрастающая функция отступа M.

Сбор ассессорских оценок

Для оценивания качества была сгенерирована выборка пар статей русскоязычной Википедии из категорий математики, физики, химии, информатики.

Ассессорам предлагалось выбрать из двух статей ту, которая потребовала больше усилий для её понимания и содержала больше незнакомых терминов, либо указать, что статьи примерно равны по сложности, либо что они совершенно из разных областей.

Какая из статей сложнее?

Эксперимент 1. Референтный корпус — Википедия

Accuracy — доля пар, на которых и модель, и ассессоры выбрали одну и ту же статью как сложную.

Референтный корпус — вся Википедия.
Сравнение наших оценок сложности с ARI и индексом Флеша:

Модель	**Тип токенов**	**Accuracy**
ARI	–	46%
Индекс Флеша	–	57%
Частотные оценки	Слова	63%
Частотные оценки	Слова+Биграммы	71%
Сложностные оценки	Слова+Биграммы	74%
Сложностные оценки	Биграммы	**81%**

- квантильные оценки лучше коррелируют с экспертными
- сложностные оценки лучше частотных
- оценки по биграммам лучше, чем по отдельным словам

Эксперимент 2. Референтный корпус — тема из Википедии

Референтный корпус — статьи Википедии из фиксированной темы, выделенной с помощью тематической модели ARTM.

Тема: математика

Модель	w_i	**Accuracy**
ARI	-	41%
Индекс Флеша	-	49%
Частотная	c_i	55%
Частотная	c_i/n	61%
Сложностная	c_i	79%
Сложностная	c_i/n	**84%**

Тема: физика

Модель	w_i	**Accuracy**
ARI	-	52%
Индекс Флеша	-	58%
Частотная	c_i	65%
Частотная	c_i/n	63%
Сложностная	c_i	**82%**
Сложностная	c_i/n	81%

- квантильные оценки лучше коррелируют с экспертными
- сложностные оценки лучше частотных
- качество оценок слабо зависит от темы

K.Vorontsov, A.Potapenko. Additive regularization of topic models. 2014.

Выводы

- *Предложен* статистический (квантильный) подход к оцениванию сложности текстов естественного языка,
- использующий референтный корпус несложных текстов,
- и применимый единообразно на всех уровнях языка.
- *Предложена* методика верификации оценок сложности.
- *В экспериментах показано*, что квантильные оценки сложности лучше коррелируют с экспертными оценками, чем известные индексы удобочитаемости.

Открытые проблемы:

- построить квантильные оценки для всех уровней языка;
- построить агрегированные оценки сложности;
- использовать их в системе разведочного поиска.