Lexical Quantile-Based Text Complexity Measure

(In proceedings of Conference on Recent Advances in Natural Language Processing 2019)

Maksim Eremeev and Konstantin Vorontsov (Machine Intelligence MIPT Lab)

OpenTalks.AI conference

Motivation

- Build a simple **Reading Order** technique to rank search results
- In the **exploratory search**, the user needs a hint which of the found documents to read first, gradually moving from simple to more complex documents.
- Reading Order optimization is an alternative way to consume content. It departs from the ranking-by-relevance way, which is typical and spread.

Main Idea

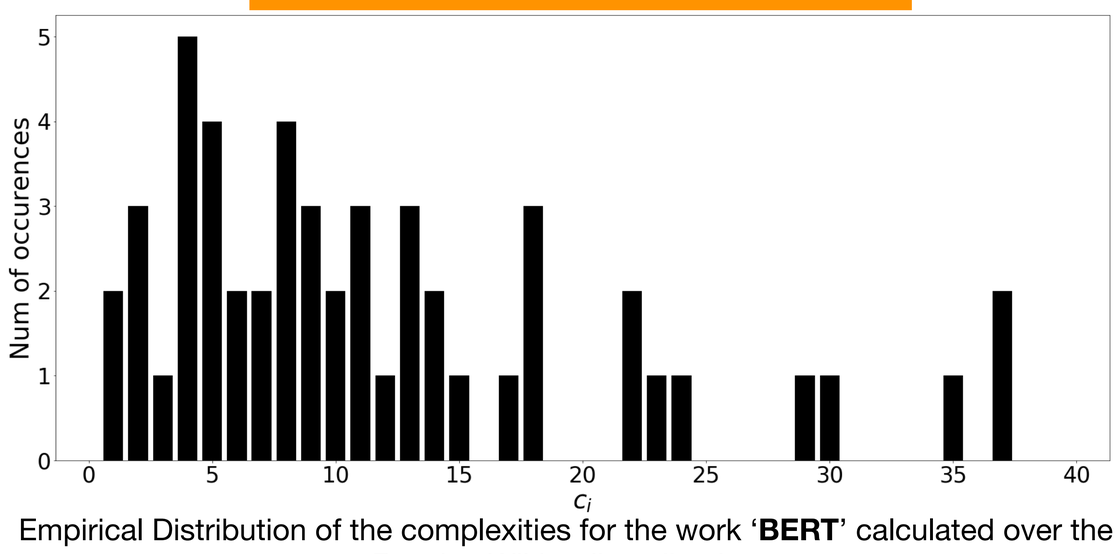
- The more specific terms document contains, and the more rare they are, the more complex the document is.
- We estimate the complexity of each term in the document and then aggregate them to get the complete document complexity score.
- We used the **Russian Wikipedia** as a ***reference collection*** of moderately complex texts in order to determine what term frequencies are abnormal.

Evaluation

- By using crowdsourcing, we labeled 10K pairs of the Russian Wikipedia documents. Assessors were asked to read both articles and to choose which was more difficult to comprehend. Yandex.Toloka was used as a platform.

Left Document	Right Document	Which document is more complex
Matrix	Tensor	RIGHT
Neural Network	Linear Regression	LEFT
Electric Charge	Molecule	EQUAL
Mac OS X	Convex Hull	INVALID

Sample Distribution



Empirical Distribution of the complexities for the work '**BERT**' calculated over the Russian Wikipedia collection

General Model

Document length

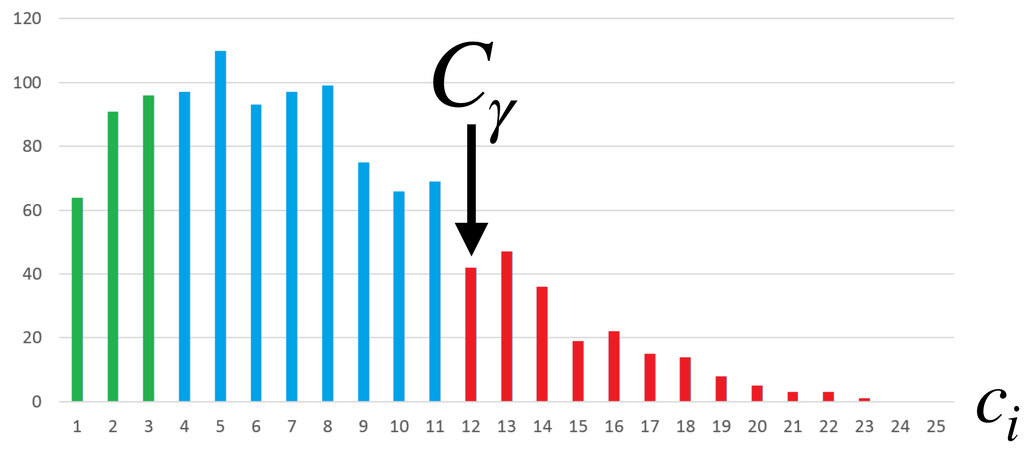
$$W(d) = \sum_{i=1}^{n_d} w_i [c_i > C_\gamma(x_i)]$$

Term complexity

Document complexity

Term weight

γ-quantile for empirical term complexity distribution in the reference collection



Distance-Based Model

Assumption: The frequency of each word correlates with the brain load when reading the text *(Birkin, 2007)*.

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

$$c_i = \bar{r}(x_i) - r(x_i)$$

We count complexity as a difference between mean distance over the reference collection and current distance value

Weight examples

$w_i = c_i$ – Total complexity | $w_i = \frac{c(t_i)}{n_d}$ – Mean complexity

Experiment Results

We tested our model on 10K labeled Russian Wikipedia dataset. We used **accuracy** score and compared the model with SOTA readability indexes - ARI and Flesh-Kincaid test.

Flesh-Kincaid test - **57 %**
ARI - **63 %**
Total Complexity Distance-Based model - **74 %**
Mean Complexity Distance-Based model - **81 %**