Lexical Quantile-Based Text Complexity Measure

Maksim Eremeev
AITHEA
m.eremeev@aithea.com

Konstantin Vorontsov
National University of Science and Technology MISIS
voron@aithea.com

Abstract

This paper introduces a new approach to estimating the text document complexity. Common readability indices are based on average length of sentences and words. In contrast to these methods, we propose to count the number of rare words occurring abnormally often in the document. We use the reference corpus of texts and the quantile approach in order to determine what words are rare, and what frequencies are abnormal. We construct a general text complexity model, which can be adjusted for the specific task, and introduce two special models. The experimental design is based on a set of thematically similar pairs of Wikipedia articles, labeled using crowdsourcing. The experiments demonstrate the competitiveness of the proposed approach.

1 Introduction

Automated text complexity measurement tools have been proposed in order to help teachers to select textbooks that correspond to the students' comprehension level and publishers to explore whether their articles are readable. Thus, plenty of readability indexes were developed. Measures like *Automated Readability Index* (Senter and Smith, 1967), *Flesch-Kincaid readability tests* (Flesh, 1951), SMOG index (McLaughlin, 1969), Gunning fog (Gunning, 1952) and etc. use heuristics based on simple statistics such as total number of words, mean number of words per sentence, total number of sentences or even number of syllables to evaluate how complex given text is. By combining these statistics with different weighting factors, readability indexes assign the given document a *complexity score*, which is, in most cases, the approximate representation of the US grade level needed to comprehend the text. For instance, an Automated Readability Index (ARI) has the following form for the document d:

$$ARI(d) = 4.71 \times \frac{c}{w} + 0.5 \times \frac{w}{s} - 21.43 \quad (1)$$

where c refers to the total number of letters in the document d, w is the total number of words and s denotes the total number of sentences in d.

Since readability indexes rely on a few basic factors, precise assessment requires aggregation of many scores. Thus, Coh-Metrix-PORT tool (Aluisio et al., 2010) includes more than 50 different indexes for Portuguese language. The tool is based on Coh-Metrix (Graesser et al., 2004) principles to estimate complexity and cohesion not only for explicit text, but for the mental representation of the document.

Readability indexes are interpretable and easy to implement. However, the great number of constants tuned specifically for the English language texts, lack of the semantics consideration and tailoring to the US grade level system restrains the number of possible applications.

As for the non-English languages, several lexical and morphological features for Italian to solve text simplification problem were presented (Brunato et al., 2015), supervised approach in readability estimations was introduced (vor der Brck et al., 2008) and the complexity estimations for legal documents in Russian were explored (Dzmitryieva, 2017).

In this paper we introduce a new approach to gauge the complexity of the documents based on their lexical features. Our research is motivated by information retrieval applications such as exploratory search for learning or editorial purposes (Marchionini, 2006; White and Roth, 2009; Palagi et al., 2017). 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 content consumption that departs from the typical ranked lists of documents

Proceedings of Recent Advances in Natural Language Processing, pages 270–275,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_031

based on their relevance (Koutrika et al., 2015). The more specific terms document contains, and the more rare they are, the more complex the document is. To formalize this consideration, we estimate the complexity of each term in the document and then aggregate them to get the complete document complexity score. We use Wikipedia as a *reference collection* of moderately complex texts in order to determine what term frequencies are abnormal.

In section 2 we describe quantile approach to estimate the single term complexity. We present highly flexible general model in section 3 and models in subsections 3.1 and 3.2. The way of evaluating the proposed methods is introduced in section 4 and the experiments result are provided in section 5.

2 Single Term Complexity Estimation

Reference collection: Let D denote a reference collection. Let document $d \in D$ consist of terms $t_1, t_2, \dots t_{n_d}$, where n_d refers to the length of document d. Each term can be either a single word or a key phrase.

Quantile approach: In general case each term can occur in different complexity states, which may depend on a position in text or context surrounding the term. Each complexity state of the term t_i standing in position i is described with a *term complexity score* $c(t_i)$. Consider a complexity scores empirical distribution for each term over the reference collection. Assume that term t_i is *in complex state* if its complexity $c(t_i)$ in current text position i is greater than γ-quantile $C_\gamma(t_i)$ of the distribution over $c(t_i)$, where γ is a hyperparameter, responsible for the complexity level. Therefore, when estimating complexity score of the document, we count $c(t_i)$ only for terms t_i which are in the complex state, defined by the γ parameter.

For instance, $c(t_i)$ can be a constant, which means all terms have identical complexity, or can be set equal to 0 if it occurs in the reference collection and 1 otherwise. In this case, we count new terms (for the reference collection) as complex and all other terms as simple.

3 General Document Complexity Model

Document d complexity $W(d)$ can be calculated by aggregating complexity scores of terms that form d. In this paper we propose a weighed sum over the complex terms to be the aggregate function.

$$W(d) = \sum_{i=1}^{n_d} w(t_i)[c(t_i) > C_\gamma(t_i)] \qquad (2)$$

where $[\,]$ refers to the Iverson notation (i.e. $[true] = 1, [false] = 0$).

By defining weights $w(t_i)$ and complexity scores $c(t_i)$ for all terms t_i specialize the complexity model.

Some examples of interpretable weights $w(t_i)$ are presented in Table 1.

$w(t_i)$	**Meaning of** $w(t_i)$
1	number of complex terms
$1/n_d \times 100\%$	complex terms percentage
$c(t_i)$	total complexity
$c(t_i)/n_d$	mean complexity
$c(t_i) - C_\gamma(t_i)$	excessive complexity
$(c(t_i) - C_\gamma(t_i))/n_d$	mean excessive complexity

Table 1: Weights $w(t_i)$ examples.

3.1 Distance-Based Complexity Model

The following model relies on the assumption, proposed in (Birkin, 2007). Consider an arbitrary document d which is the sequence of terms $t_1, t_2, \dots t_{n_d}$. Let $r(t_i)$ be a distance in terms to the previous occurrence of the same term t_i in document d. Formally,

$$r(t_i) = \min_{1 \le j < i} \{i - j \mid t_i = t_j\}. \qquad (3)$$

If i is the first occurrence of term t_i in document d, it means that $r(t_i)$ is undefined. In such cases we take $r(t_i)$ equal to n_d. Hence, for terms with the only occurrence in d complexity scores are the greatest.

If term t does not appear in the reference collection, we set C_γ equal to $-\infty$, therefore counting it as a constantly complex term.

Assume that term t in the position i is more complex than the same term in the position j if $r(t_i) > r(t_j)$. Consider there are no separators between documents in the reference collection, so it becomes a single document d_{all}. Thus, it is possible to count distributions of $r(t)$ of each unique term t in d_{all} and corresponding γ-quantiles $C_\gamma(t)$ of these distributions.

For the document d, which complexity we try to estimate, we calculate $r_d(t_i)$ values for all terms $t_i \in d$.

We define mean distance $r_{d,i}(t_i)$ for term t_i in i-th position in the document d as

$$\bar{r}_{d,i}(t_i) = \frac{\sum_{j=1}^{i} r_d(t_i)[t_i = t_j]}{\sum_{j=1}^{i}[t_i = t_j]} \quad (4)$$

which aggregates all occurrences of the term t_i from the document start.

Finally $c(t_i)$ has the form:

$$c(t_i) = \bar{r}(t_i) - \bar{r}_{d,i}(t_i) \quad (5)$$

where $\bar{r}(t_i)$ is the mean distance of the reference collection scores $r(t_i)$ for the term t_i.

Intuitively, this means, that term is more complex if it occurs less in reference collection and occurs more in document d.

Figures 1 and 2 show distributions of distances $r(t)$ for the simple term 'algebra' and the complex term 'nlp', calculated over the reference collection containing 1.5M documents of the Russian Wikipedia. For the 'algebra' term most occurrences are relatively close to each other, whether 'nlp' occurrences have fairly greater distance scores.

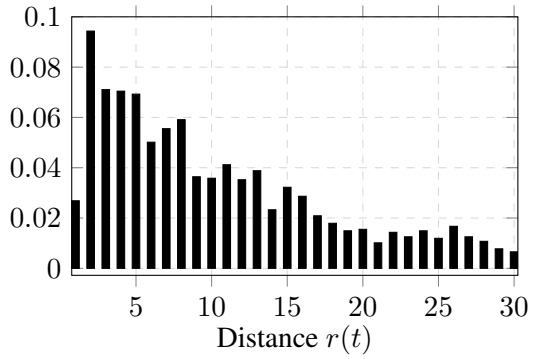

Figure 1: Distribution of distances $r(t)$, calculated over the complete Wikipedia dataset for the word 'algebra'.

So, using the formula for $c(t_i)$ as above and choosing weights $w(t_i)$ we get the distance-based complexity model.

3.2 Counter-Based Complexity Model

The second model presented in this paper is based on the assumption that each term has an independent fixed complexity in the whole language. Thus, in this section we consider not the complexity distribution of a single term, but the general complexity distribution over all terms in the language. Hence, each term t is assigned the only

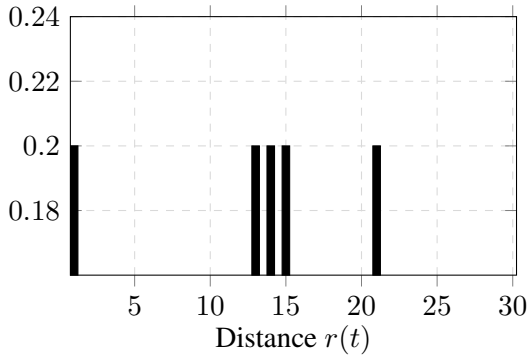

Figure 2: Distribution of distances $r(t)$, calculated over the complete Wikipedia dataset for the word 'nlp'.

complexity score $c(t)$ and the γ-quantile we count is now a constant C_γ.

Hence, the model has the following form:

$$W(d) = \sum_{i=1}^{n_d} w(t_i) \left[\frac{1}{count(t_i)} > C_\gamma\right] \quad (6)$$

where $w(t_i)$ corresponds to the term weights introduced before.

Assume the term t_1 is more complex than the term t_2 if number of occurrences in the reference collection of the term t_1 is lesser than the number of occurrences of the term t_2.

Let $count(t)$ denote number of occurrences of the term t in the reference collection. Thus, the complexity score function can be defined as

$$c(t) = \frac{1}{count(t)} \quad (7)$$

so the assumption above is satisfied.

For each term t we calculate counters $count(t)$ and complexity scores $c(t)$ over the reference collection. Having the distribution of $c(t)$, we obtain γ-quantiles C_γ. The described distribution for the Russian Wikipedia reference collection is shown on Figure 3.

Thus, we have defined $c(t)$ for all terms possible and the distribution necessary to count the C_γ. By varying weights $w(t_i)$ described in section 3, we obtain the counter-based model for the complexity estimation.

4 Quality Metric

To measure the quality of proposed algorithms, we asked assessors to label 10K pairs of Russian Wikipedia articles. Assessors were asked to carefully read both articles and to choose which was

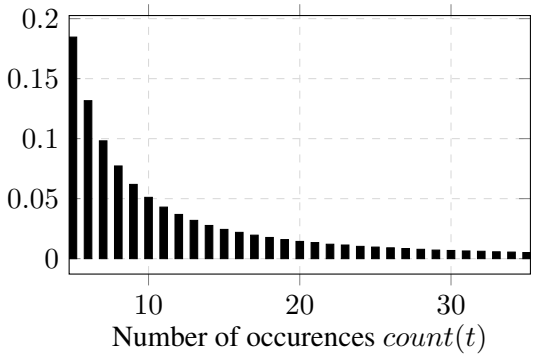

Figure 3: Distribution of $count(t)$, calculated over complete Wikipedia articles dataset.

more difficult to comprehend. If person cannot determine which document is more complex, then he was asked to choose 'documents are equal' option. If documents in the given pair are from different scientific domains, then we ask assessor to choose 'invalid pair' option.

Documents were chosen from math, physics, chemistry and programming areas. Clustering was performed using the topic modeling technique (Hofmann, 1999). BigARTM open-source library was used to perform the clustering (Vorontsov et al., 2015). Pairs were formed so that both documents belong to a single topic and their lengths are almost identical. Examples of document pairs to assess are introduced in Table 2.

Document 1	**Document 2**	**Result**
Matrix	Tensor	RIGHT
Neural network	Linear regression	LEFT
Electric charge	Molecule	EQUAL
Mac OS X	Convex Hull	INVALID

Table 2: Examples of labeled document pairs.

Each pair was labeled twice in order to avoid human factor mistakes. We assume that the pair was labeled correctly if labels were not controversial, i.e. first assessor labeled the first document as more complex, while second assessor chose the second document. If one or both grades were 'documents are equal' then we assume the pair to be correctly labeled.

8K pairs out of 10K were labeled correctly and were used to compare for the different versions of algorithms. For each we calculated the accuracy score, which is the rate of correctly chosen document in the pair.

5 Experiments

Two types of experiments were done. In first case we used full Russian Wikipedia articles dataset (1.5M documents) as a reference collection. In second type we used only Wikipedia articles from the math domain. To do that, we built a topic model using ARTM (Additive Regularization of Topic Models) technique (Vorontsov and Potapenko, 2015), which clusters documents into monothematic groups.

5.1 Complete Wikipedia Dataset

Preprocessing: All Wikipedia articles were lemmatized (i.e. reduced to normal form). In this experiment we assume term to be either a single word or a bigram (i.e. two words combination). To extract them, RAKE algorithm (Rose et al., 2010) was used. Hence, each document in the collection was turned into the sequence of such terms.

Reference collection: Preprocessed Wikipedia articles were used as a reference collection. $r(t)$ for every term position and $count(t)$ for every unique term were counted.

Documents to estimate complexity on: We used the labeled pairs described in Section 4 to evaluate the models. Accuracy was used as a quality metric.

Models to evaluate: Models introduced in 3.1 and 3.2 with different $w(t_i)$ parameters were tested. We took ARI and Flesch-Kincaid readability test as benchmarks.

The results of the experiments are introduced in Table 3. Also we tested how the bigrams extraction affects final quality with fixed weight function $w(t) = c(t)/n_d$. The results are given in Table 4.

Model	$w(t)$	**Accuracy**
ARI	-	**46%**
Flesch-Kincaid	-	**57%**
Distance-based	$c(t)$	**68%**
Distance-based	$c(t)/n_d$	**71%**
Counter-based	$c(t)$	**77%**
Counter-based	$c(t)/n_d$	**81%**

Table 3: Results of experiment 1 with different weight function.

Results show that both distance- and counter-based approaches work twice as well as readability indexes. Counter-based model with $w(t) = c(t)/n_d$ weights show the best results.

Model	Terms	Accuracy
Distance-based	Words	**63%**
Distance-based	Words+Bigrams	**71%**
Counter-based	Words+Bigrams	**74%**
Counter-based	Bigrams	**81%**

Table 4: Results of experiments 1 with terms differently defined.

5.2 Single Topic Wikipedia Dataset

In experiment 2 we shortened the reference collection to include only documents from specific topic.

ARTM model: To divide documents into single-topic clusters, topic modeling is used. Topic Models are unsupervised machine learning models and perform soft clustering (i.e. assign each document a distribution over topics). The set of such vectors for all documents form a matrix, which is usually denoted by Θ. *ARTM model* was trained on the preprocessed Wikipedia dataset. ARTM features dozens of various types of regularizers and allows to treat modalities (i.e. types of terms) differently.

In this specific experiment we used regularizers to sparse Θ matrix and make each topic distribution over terms more different. Words and bigrams (i.e. pairs of words) modalities were used with weights 1 and 5 respectively. Using this model, we detect the most likely topic for each document.

Experiment setup: In the following experiment we chose math and physics documents to be the reference collection. Documents were preprocessed in the same way as they were in the previous experiment. We also divided labeled pairs into same single-topic groups to test models configured with different reference collections on various single-topic groups of labeled pairs.

Math collection included 200K documents in reference collection and 3.5K labeled pairs, while for the physics collection it was 250K documents in reference collection and 1.5K labels. The results are shown in Table 5 and Table 6.

As it can be seen from results, using tailored reference collection improves the score. Indeed, that solves terms ambiguity problem and eliminates terms unrelated to the topic from the reference collection, so they are treated complex in the estimating document, which is fairly logical.

Model	$w(t)$	Accuracy
ARI	-	**41%**
Flesch-Kincaid	-	**49%**
Distance-based	$c(t)$	**55%**
Distance-based	$c(t)/n_d$	**61%**
Counter-based	$c(t)$	**79%**
Counter-based	$c(t)/n_d$	**84%**

Table 5: Results of experiment 2 on math collection of Wikipedia articles with different weights.

Model	$w(t)$	Accuracy
ARI	-	**52%**
Flesch-Kincaid	-	**58%**
Distance-based	$c(t)$	**65%**
Distance-based	$c(t)/n_d$	**63%**
Counter-based	$c(t)$	**82%**
Counter-based	$c(t)/n_d$	**81%**

Table 6: Results of experiment 2 on physics collection of Wikipedia articles with different weights.

6 Conclusions

We have presented an approach to estimating text complexity based on lexical features. Document complexity is an aggregation of terms' complexities. Introduced general model is highly flexible, it can be adjusted by tuning weights $w(t)$ and choosing proper reference collection.

Complexity score can only be count with respect to the reference collection. Reference collection can be a large set of documents on different topics or just contain single-topic texts.

The proposed complexity measures are used in AITHEA exploratory search system (http://aithea.com/exploratory-search) for ranking search results in complexity-based reading order.

Acknowledgements

Application of topic modeling in this research was supported by the Russian Research Foundation grant no. 19-11-00281. The work of K.Vorontsov was partially supported by the Government of the Russian Federation (agreement 05.Y09.21.0018) and the Russian Foundation for Basic Research (grants 17-07-01536).

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for

text simplification.

A.A. Birkin. 2007. *Speech Codes*. Hippocrat, Saint-Peterburg.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. pages 31–41.

Tim vor der Brck, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators.

Aryna Dzmitryieva. 2017. The art of legal writing: A quantitative analysis of russian constitutional court rulings. *Sravnitel'noe konstitucionnoe obozrenie*, 3:125–133.

R. Flesh. 1951. How to test readability. *New York, Harper and Brothers*.

Arthur Graesser, Danielle McNamara, Max Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, computers : a journal of the Psychonomic Society, Inc*, 36:193–202.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Georgia Koutrika, Lei Liu, and Steven Simske. 2015. Generating reading orders over document collections. In *2015 IEEE 31st International Conference on Data Engineering*, pages 507–518.

Gary Marchionini. 2006. Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41–46.

G. H. McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Emilie Palagi, Fabien Gandon, Alain Giboin, and Raphaël Troncy. 2017. A survey of definitions and models of exploratory search. In *ESIDA17 - ACM Workshop on Exploratory Search and Interactive Data Analytics, Mar 2017, Limassol, Cyprus*, pages 3–8.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*.

R.J. Senter and E.A. Smith. 1967. Automated readability index. *AMRL-TR*, 66(22).

K. V. Vorontsov and A. A. Potapenko. 2015. Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1):303–323.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Petr Romov, and Marina Suvorova. 2015. Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *AIST'2015, Analysis of Images, Social networks and Texts*, pages 370–384. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS).

Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan and Claypool Publishers.