A Language-independent Metric for Measuring Text Simplification that does not Require a Parallel Corpus

Lucas Mucida, Alcione de P. Oliveira, Maurilio de A. Possi
Federal University of Vicosa
Vicosa - Minas Gerais - Brazil
{lucas.mucida,alcione,maurilio}@ufv.br

Abstract

Natural language processing encompasses several tasks, one of which is the automatic text simplification. Telling whether one text is simpler than another involves not only knowledge about the language being analyzed, but also a cultural knowledge of the target audience to which the text is being directed. Most of the current metrics used to measure text simplification are based on the use of parallel corpora, prepared by humans, which makes it difficult to apply the metrics in automatic text simplification in real time. In this paper, we present **ISiM** (**I**ndependent **Si**mplification **M**etric), a metric that dismiss a parallel corpus, is simple, fast, language and human annotation independent, capable of quantifying the simplicity/complexity of a sentence, thus contributing improve automating text simplification. The results of the tests performed indicate that the proposed metric has the potential to be used to evaluate automatic methods of simplification.

Introduction

Text simplification is the process of reducing linguistic complexity while ensuring that the text's original information and meaning are preserved (Siddharthan 2014). Simplifying texts can be beneficial in some situations, like reaching a broader audience by making the text more readable and understandable for particular groups, such as: children (Kauchak 2013), foreigners (Paetzold and Specia 2016) , people with educational limitations (Gasperin et al. 2009), people with language impairments like aphasia, dyslexia and autism (Devlin and Unthank 2006; Rello et al. 2013; Evans, Orasan, and Dornescu 2014).

In Brazil, according to a survey made by the Brazilian Institute of Geography and Statistics (IBGE), (IBGE 2019) about 6.6% of the population is illiterate. The most worrying information comes from a survey carried out by the Functional Illiteracy Indicator (*INAF*) in 2018, showing that about 30% of the Brazilian population is functionally illiterate (INAF 2018). Furthermore, 64% of the population has a rudimentary or lower degree of literacy, which shows how important it would be to use simpler text to efficiently reach this portion of the population.

Considering the importance of text simplification, automatic simplification performed by artificial intelligence turns out to be a relevant tool. Several studies have been done trying to create an efficient automation model for this purpose (Kincaid et al. 1975; Papineni et al. 2002; Sun and Zhou 2012; Xu et al. 2016; Zhang and Lapata 2017; Kriz, Apidianaki, and Callison-Burch 2020; Omelianchuk, Raheja, and Skurzhanskyi 2021). However, some of the proposed ideas are imported from other areas such as text translation (Štajner, Béchara, and Saggion 2015), where one basically try to "translate" a complex text into a simpler text, and also from the area of text summarization (Kriz, Apidianaki, and Callison-Burch 2020) where the goal of reduce sentence size intertwines with the goals of text simplification, although in (Davison and Kantor 1982) we see that not always reducing the length of a sentence makes it simpler. The two most widely used metrics to evaluate the results of automated text generation models are BLUE (Papineni et al. 2002) and SARI (Xu et al. 2016), which also have their ideas borrowed from the areas of translation and summarization of texts. BLUE tends to focus on the meaning and grammar, while SARI focus on the lexical side. Both need reference phrases to carry out the measurement.

Our metrics follows the *Lexile* approach cited in (Smith and others 1989), where the author shows that the symbols used for human communication are divided into two categories that will be described below: semantic difficulty and syntactic complexity. (Scarton and Aluísio 2010) classifies these two properties as follows:

- **Semantic difficulty**: the frequency of daily use of each word contained in the sentence. According to (Bormuth 1966), inferring the meaning of a text includes the probability of the person identifying the word in a context, that is, the low frequency of a word can interfere in the logical process of inferring the text and make the sentence incomprehensible.
- **Syntactic Complexity**: the sentence length is an influence on the syntactic complexity. The tendency is that the longer the sentence, the more syntactic structures are added to the sentence and the more difficult it is to interpret.

Also according to (Scarton and Aluísio 2010), semantic units vary in familiarity, while syntactic structures vary in

complexity. The comprehensibility of a message is largely governed by the familiarity of the semantic units and the complexity of the syntactic structures used in the construction of the message, which in this work are addressed in measuring the frequency that the word appears in a corpus (affects comprehensibility) and sentence length (affects syntactic complexity).

Another problem that researchers face when working with text simplification is the particularity of each language. Depending on the metric we are using, it cannot be simply reused for a corpus in another language as this metric may be taking into account linguistic aspects characteristic of a specific language, and this will bring distorted results in another language. Our method does not have this problem because it is language independent.

This paper is organized as follows. The next section gives an overview of the works that have a relation with this research. The section after the related works presents the the proposed metric. After that, the results obtained using the metric are shown and finally the conclusions are presented.

Related Works

BLEU (Papineni et al. 2002) is a widely used metric for measuring the quality of machine translation and it is also applied in the text simplification task. It requires an input, a reference and the output generated by the model,in order to calculate how close the output emitted by the model is to the reference sentence. It is a conservative metric, the more there are changes from the reference sentence, the more the output is penalized. This idea is clearly coming from the text translation, where the closer the translated sentence gets to the reference, the better the translation. With this conservative characteristic, BLEU tends to preserve the grammar and the meaning of the sentence, penalizing strong alterations. However, this idea is the source of the biggest criticism of BLEU, since an output that differs greatly from the reference sentence can still be a good translation, or as in our case, a good simplification.

The SARI (**S**ystem Output **A**gainst **R**eferences and **I**nput Sentence) (Xu et al. 2016), measures how good were the words added, deleted or kept in relation to the entry and the reference phrase. As in the previous case, it needs reference sentences to carry out the evaluation.

Simple-QE (Kriz, Apidianaki, and Callison-Burch 2020) is a metric based on BERT (Devlin et al. 2019) derived from *Sum-QE* (Xenouleas et al. 2019), which is a metric to measure the quality of text summaries. The acronym *QE* comes from the English *quality estimation*. An important aspect of this metric is that it does not depend on reference phrases, reducing the need for human work to make the final estimate of the quality of the generated simplification. In their metric, the authors focused on three important aspects of the sentences: ***Fluency***: says how well formed the generated text is; ***Adequacy***: says whether how much of the original text's meaning has been preserved; ***Complexity***: says how simple the generated sentence is The main difference from our approach is the need to fine-tune the BERT model, which is therefore a more difficult metric to implement. The paper is still in pre-print mode and the authors have not made the data and code available. Therefore, it was not possible to perform comparative tests. According to the results presented in the paper, *Simple-QE* obtained better results in terms of simplification compared to SARI and BLEU.

The Lexile framework[1] (Lennon and Burdick 2004) is paid software used extensively to assess the relevance of books to schools, mainly in the USA. According to in (Smith and others 1989), Lexile is composed of two components: the Lexile measure and the Lexile scale. The first measures the reader's skill or the complexity of a text. The second is a reading mastery scale ranging from 200L (beginner readers) to 1700L (advanced readers). The authors believe that, by applying these measures, they would come up with ideal books for each type of reader measured by their scale. That is, the framework would be able to score a reader and, through that score, choose books (also scored by Lexile) that would be better understood by the reader with regard to the complexity of the text contained therein.

Lexile measures involving text complexity are based on two factors: word frequency and sentence length. However, as it is a paid framework, the authors didn't show the final equation. The only information about what might be present in the new framework is in (Smith and others 1989) from 1989, but this paper wasn't even cited by the latest version (Lennon and Burdick 2004) of Lexile. In that paper, the author explains how he formulated the equations using linear regression to calculate the constants. However, the author based himself only on a test carried out in three thousand American students, where a text with some white space was given and the student had to choose a word from the available options that best completed the text. From the results of these tests, the linear regression used by the author resulted in the following Lexile scale equation:

$$tl = (9.82247 \times LMSL) - (2.14634 \times MLWF) - c \quad (1)$$

Where tl stands for *Theoretical Logit*; $LMSL$ stands for *Log of the Mean Sentence Length*; $MLWF$ stands for *Mean of the Log Word Frequency*; and c is a constant.

As said before, this framework is similar to our approach. However, their final work was not disclosed and not published in a scientific journal. In addition, it was built based on the English language and American students, which makes it subject to modifications for its application in other countries.

The proposed metric

Our metric is based on two linguistic elements: the length of the sentence and the frequency of the appearance of words in a corpus. This was based on (Stenner 1996) where it is stated that, regarding to the semantic component, the probability of a person finding a word in the context is what weighs most when inferring its meaning (Bormuth 1966). This is the basis for the so-called "exposure theory", which explains how receptive or auditory vocabulary develops (Miller and Gildea 1987; Stenner, Smith III, and Burdick 1983). Two other important studies related on familiarity and scarcity of words (Klare and others 1963;

[1] https://lexile.com/

Carroll and Davies 1971), reinforce this idea. Also according to (Stenner 1996), knowing how often words are used in written and oral communication provides the best way to infer the likelihood that a word will be found and thus become part of an individual's receptive vocabulary.

Based on this idea, we used the corpus *wordfreq* in the Python library, using the function *zipf frequency* that brings the word frequency to a user-friendly logarithmic scale. The proposed metric is based solely on the frequency and length of the words in the sentence. The base equation is as follows:

$$\frac{1}{n^{1.19}} \left(\sum_{i=1}^{n} \log_{10} \left(max \left(freq \left(w_i \right), 1 \right) \right) \right) \quad (2)$$

Where: ***n*** is the number of words contained in the sentence. It is inversely proportional to the sum of the log and also raised to 1.19 to penalize longer sentences. w_i is the i-th word of the sentence. $freq(w_i)$ is the frequency with which the i-th word appears in the wordfreq corpus. ***max*** is a function to ensure that words not found in the corpus are not counted, always guaranteeing a value of 1. log_{10} to prevent the function from growing too large for words with a very high frequency in the corpus.

The weight in the power of the variable *n* was empirically tested with values ranging from 1.1 to 1.5, and the value 1.19 was the one that obtained the best results. In this metric, the higher the sentence punctuation, the simpler it is. If a simplified sentence generation algorithm generates a sentence whose punctuation exceeds punctuation of the original sentence, it means that the text has been simplified. Since the function *zipf frequency* has already brought the value in a log base, the function can be rewritten as follows:

$$\frac{1}{n^{1.19}} \left(\sum_{i=1}^{n} freq \left(w_i \right) \right) \quad (3)$$

Results

To assess the results, we tested the metric on two separate corpora, the SimPA (Scarton, Paetzold, and Specia 2018) with 2200 pairs of complex/simple sentences and the TurkCorpus(Xu et al. 2016) with 352 pairs of complex/simple sentences. TurkCorpus also has 8 reference sentences for each case.

To measure the effectiveness of our metric, we calculated the scores for each of the sentences in the two *corpora* mentioned above. The expected result is a higher score for the sentence indicated as the simplest in the *corpus*. After applying the algorithm, we measured how many sentences were classified by the metric as being simpler than the original. For that, several different values were tested for the power of *n* that represents the length of the sentence, as well as tests with and without *stop words*. In the table 1 we can see the results for the SimPA *corpus*.

Table 2 displays the metric results for TurkCorpus. Note that in both *corpora* the best results presented were for the same power of *n*. We also compared the time spent calculating the score, where we used the SARI measure as the baseline. The time taken to calculate the score for each of the

	With stop words	Without stop words
1.4	68.45%	71.64%
1.3	70.09%	74.36%
1.25	70.36%	78.45%
1.19	71.55%	**80.27%**
1.1	71.27%	79.36%

Table 1: Percentage of phrases correctly classified by our metric for the SimPA corpus varying the power of *n*.

718 TurkCorpus sentences by SARI averaged 32.7 seconds, while our metric took just 0.17 seconds. As for the SimPA corpus that has 5,000 sentences, the SARI took an average of 65 seconds while our metric took an average of 572 ms.

	With stop words	Without stop words
1.4	93.31%	95,82%
1.3	94.15%	96.10%
1.25	95,54%	96.38%
1.19	96.10%	**96.94%**
1.1	96.38%	96.38%

Table 2: Percentage of phrases correctly classified by our metric for the TurkCorpus varying the power of *n*.

Conclusion

As we can see in the previous session, we obtained excellent results considering that our metric is simple, free of human intervention, independent of language and extremely fast. In the SimPA corpus, the highest percentage of correct answers was 80.27%, while in the TurkCorpus we obtained 96.94%, which shows its effectiveness for text simplification analysis. The time spent for the calculation in these corpora is also a differential of ISiM, being executed in less than 1 second in both corpora. In addition, the metric was performed independently of human references and annotations, being also language-independent, simply changing the word frequency corpus to the target language. A limitation of our metric is the fact that it does not guarantee semantic equivalence between the analysed sentences. Regarding the next steps, we intend to add a measure based on dense vectorization of the sentences to indicate the degree of semantic equivalence between the sentences. Furthermore, a research goal is the development of a model that generates simplified phrases.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and FAPEMIG.

References

Bormuth, J. R. 1966. Readability: A new approach. *Reading research quarterly* 79–132.

Carroll, J. B., and Davies, P. 1971. Barry rich man. *Word Frequency Book. Boston: Hough ton Mifflin*.

Davison, A., and Kantor, R. N. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly* 187–209.

Devlin, S., and Unthank, G. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, 225–226.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Evans, R.; Orasan, C.; and Dornescu, I. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Association for Computational Linguistics.

Gasperin, C.; Maziero, E.; Specia, L.; Pardo, T.; and Aluisio, S. M. 2009. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *The proceedings of SEMISH-XXXVI seminário integrado de software e hardware* 387–401.

IBGE. 2019. Taxa de analfabetismo no brasil.

INAF. 2018. Indicador de analfabetismo funcional. *Instituto Paulo Montenegro, Ação Educativa. São Paulo*.

Kauchak, D. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 1537–1546.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Klare, G. R., et al. 1963. *Measurement of readability*. Iowa State University Press.

Kriz, R.; Apidianaki, M.; and Callison-Burch, C. 2020. Simple-qe: Better automatic quality estimation for text simplification. *arXiv preprint arXiv:2012.12382*.

Lennon, C., and Burdick, H. 2004. The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*.

Miller, G. A., and Gildea, P. M. 1987. How children learn words. *Scientific American* 257(3):94–99.

Omelianchuk, K.; Raheja, V.; and Skurzhanskyi, O. 2021. Text simplification by tagging. *arXiv preprint arXiv:2103.05070*.

Paetzold, G., and Specia, L. 2016. Understanding the lexical simplification needs of non-native speakers of english. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 717–727.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Rello, L.; Baeza-Yates, R.; Bott, S.; and Saggion, H. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, 1–10.

Scarton, C. E., and Aluísio, S. M. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática* 2(1):45–61.

Scarton, C.; Paetzold, G.; and Specia, L. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Siddharthan, A. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165(2):259–298.

Smith, D. R., et al. 1989. *The Lexile Scale in Theory and Practice. Final Report.* MetaMetrics, Inc.

Štajner, S.; Béchara, H.; and Saggion, H. 2015. A deeper exploration of the standard pb-smt approach to text simplification and its evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 823–828.

Stenner, A. J.; Smith III, M.; and Burdick, D. S. 1983. Toward a theory of construct definition. *Journal of educational measurement* 305–316.

Stenner, A. J. 1996. Measuring reading comprehension with the lexile framework. In *Proceedings of the North American Conference on AdolescentAdult Literacy*.

Sun, H., and Zhou, M. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 38–42.

Xenouleas, S.; Malakasiotis, P.; Apidianaki, M.; and Androutsopoulos, I. 2019. SUM-QE: a BERT-based summary quality estimation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6005–6011. Hong Kong, China: Association for Computational Linguistics.

Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.

Zhang, X., and Lapata, M. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.