The Journal of Open Source Software

TRUNAJOD: A text complexity library to enhance natural language processing

Diego A. Palma[1], **Christian Soto**[1], **Mónica Veliz**[1], **Bruno Karelovic**[1], **and Bernardo Riffo**[1]

1 Universidad de Concepción

DOI: 10.21105/joss.03153

Software

- Review
- Repository
- Archive

Editor: Daniel S. Katz

Reviewers:

- @mbdemoraes
- @apiad

Submitted: 16 March 2021
Published: 21 April 2021

Summary

We present TRUNAJOD, a text complexity analysis tool that includes a wide variety of linguistics measurements that can be extracted from texts as an approximation for readability, coherence, and cohesion. The features that TRUNAJOD can extract from the text are based on the literature and can be separated into the following categories: discourse markers, emotions, entity grid-based measurements, givenness, lexical-semantic norms, semantic measures, surface proxies, etc. In this first version of TRUNAJOD, we mainly support the Spanish language, but several features support any language that has proper natural language processing POS tagging and dependency parsing capabilities. Finally, we show how TRUNAJOD could be used in applied research.

Statement of need

TRUNAJOD aims to address three challenges:

1. A standardized API for text complexity measurements
2. An open-source code, so any researcher in the linguistics field could contribute to it
3. Easy-to-build applications and tools that rely upon text complexity assessment

Other tools aim to make it easy for the public to get coherence and cohesion metrics. One such tool is TAACO (Crossley et al., 2019), which is written in Python and can be freely downloaded. A problem with TAACO is that it is a desktop application, which encloses the code. This makes it impossible to contribute modifications or new features, as it is a closed system. Moreover, it does not implement other relevant features to assess cohesion and coherence of discourse, for example, entity grid-based features. One open-source project with this purpose is the `Cohere` framework (Smith et al., 2016), which is written in a mixture of Java and Python. However, it does not seem to be actively maintained and it does not implement other measurements that could be used by other researchers. On the other hand, most of the tools only support English languages and do not provide support for a plethora of metrics available in a comprehensible API. TRUNAJOD aims to be different, in the sense that we do not present a closed system, but rather, an open-source project, trying to follow the best Python development patterns. Furthermore, we rely on spaCy, enabling us to support not only one language but multiple languages for coherence and cohesion tasks, which enables TRUNAJOD to improve its performance when spaCy does, promoting collaboration.

Moreover, TRUNAJOD not only implements state-of-the-art measurements for text complexity assessment, but also bundles new sets of predictors for this task. In this sense, TRUNAJOD´s contributions are:

- Fixing periphrasis of texts, because many NLP tools have issues dealing with periphrasis. In this release, this only applies to Spanish.
- Adding heuristics for measurements based on clause count. TRUNAJOD provides a new algorithm for clause segmentation.
- TRUNAJOD provides several approximations to narrativity in these new clause segmentation-dependent indices.

Text complexity assessment is a natural language processing task that can be applied to multiple problems, such as automatic summarization, automatic essay scoring, automatic summary evaluation, intelligent tutoring systems, and so on. Text complexity is usually related to the readability of a text, which is dependent on several of its intrinsic properties, mainly cohesion and coherence.

Automatic coherence evaluation is an open problem, and there have been several studies addressing it. On one side, text coherence assessment has been related to how sentences connect either semantically, or by co-referencing noun phrases. In the semantic view of coherence, Latent Semantic Analysis (LSA) (Foltz et al., 1998) has been widely used because of its simplicity. In essence, sentences are represented as vectors, and the coherence of the text is computed as the average sentence similarity, using similarity vector measurements (such as cosine distance). This approach has drawbacks, such as that sentence ordering does not matter. To solve this issue, other methodologies have been proposed based on discourse theory, in particular the centering theory. One such approach is entity grids (Barzilay & Lapata, 2008) and entity graphs (Guinaudeau & Strube, 2013) that treat coherence as to how are entities take different roles between sentences and how are they connected in the text. `TRUNAJOD` implements all these models, and thus `TRUNAJOD` can compute coherence based on sentence similarities using word vectors. `TRUNAJOD` also provides an API for dealing with entity grids and entity graphs, to extract such measurements.

A downside of the previous approaches for text complexity is that they only capture either CORPUS-based semantics or relationships between entities. Additionally, entity grids rely heavily on the dependency parser at hand and the co-reference resolution used because an entity might be mentioned in several ways across a text. The problem with this is that these measurements might be noisy depending on the use case, and simpler measurements would fit better in such cases. In `TRUNAJOD`, we compute several surface proxies that have been used by several state-of-the-art text assessment tools (McNamara et al., 2014) (Page, 1994). Such surface proxies try to approximate intrinsic properties of the text such as narrativity, connectiveness, givenness, cohesion, and coherence. `TRUNAJOD` includes classical measurements such as word count, sentence count, pronoun-noun ratio, type-token ratios, frequency index, etc. Moreover, `TRUNAJOD` comes bundled with heuristics to compute clause count-based metrics, such as subordination and clause length, among others. To achieve this, `TRUNAJOD` adds periphrasis tags to the text to heuristically segment clauses.

One drawback of using surface proxies (shallow measurements) is that they do not capture all the properties of the text, and just rely on approximations that are captured from raw text. In some use cases, this is not desirable (e.g., automatic essay scoring, intelligent tutoring systems), and these measurements should be complemented with other measurements that are desirable in those use cases, such as lexical-semantic norms and word emotions. Lexical semantic norms are norms for words that are related to the psychological degree of activation of a word in the reader (Guasch et al., 2016). Some examples of such variables are concreteness, imageability, familiarity, arousal, valence. These variables might be used for reading comprehension tools (e.g., in reading comprehension assessments, it is desirable that feedback is concrete). Moreover, emotions could be used in such cases, and even in opinion mining (Sidorov et al., 2012). `TRUNAJOD` comes bundled with both types of features, and thus, average lexical-semantic norms and emotions could be extracted from the text.

`TRUNAJOD` architectural design is shown in Figure 1.

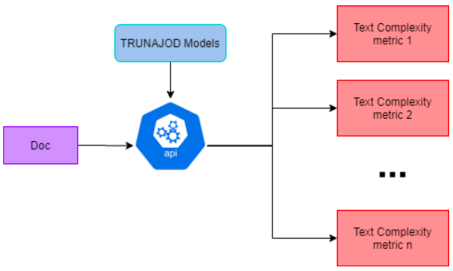

Figure 1: `TRUNAJOD` architectural design.

Basically, `TRUNAJOD` API takes as input a spaCy Doc and `TRUNAJOD` models (lemmatizer, synonym map, lexical-semantic norm map, etc.) and then it can compute supported text complexity metrics. It is worth noting that `TRUNAJOD` has available downloadable models from its GitHub repository, but currently only Spanish models are available. Nevertheless, it should be straightforward adding models for different languages.

Acknowledgements

This research was supported by FONDEF (Chile) under Grant IT17I0051 "Desarrollo de una herramienta computacional para evaluación automática de textos del Sistema escolar chileno" ("Development of a computational tool for automatic assessment of Chilean school texts")

References

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1–34.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2-3), 285–307. https://doi.org/10.1080/01638539809545029

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, *48*(4), 1358–1369. https://doi.org/10.3758/s13428-015-0684-y

Guinaudeau, C., & Strube, M. (2013). Graph-based local coherence modeling. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 93–103.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, *62*(2), 127–142. https://doi.org/10.1080/00220973.1994.9943835

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Trevino, A., & Gordon, J. (2012).

Empirical study of opinion mining in Spanish tweets. MICAI 2012. *Lect. Notes Comput. Sci.*, *7629*, 1–14.

Smith, K. S., Aziz, W., & Specia, L. (2016). Cohere: A toolkit for local coherence. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4111–4114.