An Orthographic Similarity Measure for Graph-Based Text Representations

Maxime Deforche [1], Ilse De Vos [2], Antoon Bronselaer [1], and Guy De Tré [1]

[1] Department of Telecommunications and Information Processing, Ghent University, Belgium
{maxime.deforche,antoon.bronselaer,guy.detre}@ugent.be
[2] Flanders AI Academy, VAIA, Belgium
ilse.devos@kuleuven.be

With the rapid growth of data, and textual data in particular, the need for adequate techniques to analyse and extract information from huge data volumes has grown substantially. Similarity measures between (parts of) textual documents play a pivotal role in these automatic techniques. Conventional similarity approaches, treating texts as paradigmatic examples of unstructured data, tend to overlook their structural nuances by only taking characters and/or tokens into account. As a consequence, valuable information related to the underlying structure of texts or connections between them is lost.

In our work, we propose a novel orthographic similarity measure tailored for the semi-structured analysis of texts. We explore a graph-based textual representation, where the graph's structure is shaped by a hierarchical decomposition of textual discourse units (e.g., tokens, words, sentences, documents...). Employing the concept of edit distances, our orthographic similarity measure is computed hierarchically across all components in this textual graph, integrating precomputed similarity values among lower-level nodes. In essence, this method can be construed as a generalised soft measure [1] over entire texts, transcending beyond the idea of only combining character- and token-based similarity measures

The relevance and applicability of the presented approach are illustrated by using a corpus of Byzantine book epigrams [2], featuring texts that exhibit numerous and intricate interconnections among their components.

The resulting similarity scores, between all different structural levels of the graph, allow for a deeper understanding of the (structural) interconnections among texts and enhances the explainability of similarity measures as well as the tools using them. Moreover, the resulting graph alongside the computed similarity scores can be implemented in a graph database system, enabling flexible and nuanced queries for textual analysis as well as the computation of graph-based statistics on texts.

References

[1] Cohen, W., Ravikumar, P., Fienberg, S. (2003) A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, vol 3, 73–78.

[2] Ricceri, R., et al. (2023): The Database of Byzantine Book Epigrams Project: Principles, Challenges, Opportunities. *Journal of Data Mining & Digital Humanities*.