Analyzing and classifying language differences with graph pattern mining

Bachelor thesis proposal

Supervision: Francesco Bariatti & Matthijs van Leeuwen
✉ f.bariatti@liacs.leidenuniv.nl

June 2023

1 Context

Universal Dependencies. The Universal Dependencies (UD) project is "a framework for consistent annotation of grammar [...] across different human languages" [1]. The project annotates many corpora of texts available in many languages with the same set of annotations. Each sentence in a corpus is represented as a tree, with the tree nodes corresponding to words in the sentence and tree links corresponding to *dependency relationships* between words. Additionally, tree nodes are annotated with the lemma corresponding to the word and their *part-of-speech* (POS). For instance in the sentence *"The dog was chased by the cat"* the word "the" is annotated as a `determiner`, with a `determiner` relationship with the word "dog", which itself is a `noun` that is the `nominal subject` of the `verb` "chased". The goal of the UD project is to provide the same framework for annotating all languages [1] and all corpora, to allow comparisons between them.

The UD project provides a large quantity of data: in its current version it is composed of more than two hundred treebanks on more than a hundred languages, with each treebank containing around 10^2 to 10^4 sentences/trees. However, this is also a disadvantage for the analysis, as it can not be performed by humans alone.

GraphMDL. GraphMDL [2, 3] is a family of *graph pattern mining* algorithms which use the *Minimum Description Length*(MDL) principle [5] to select a small, human-sized set of graph patterns from a graph dataset. Their goal is to enable human analysis and exploration of large graph datasets[2] by extracting characteristic patterns (subgraphs). A classification method based on the extracted patterns and the MDL principle has been proposed, but has not been evaluated thoroughly on the UD dataset.

2 Goal and Challenges

For this thesis, we are interested in assessing the ability of tree/graph pattern mining approaches (in particular the GraphMDL algorithms) to surface differences between languages. We are both interested in the comparison of the *characteristic* structures (i.e. patterns) that can be found in different languages, and their usage to *classify* languages *automatically*.

This will imply (all or part of) the following steps and challenges:

- Exploration of the UD dataset and its online documentation, in order to assess which languages/corpora would be interesting to consider, and to get an idea on how the annotation method works.

[1] However in praqctice some differences may be present due to the differences between languages. These should be assessed during the thesis

[2] Note that trees —such as the ones in the UD data— are a subclass of graphs.

- Evaluate the difficulty/feasibility of the classification task(s) by devising baseline algorithms.
- Run the existing GraphMDL implementation on the chosen data to extract relevant patterns (probably developing some automation scripts for avoiding repetitive tasks).
- Qualitative analysis of the extracted patterns. Possibility of manual analysis of the extracted patterns w.r.t. linguistics knowledge (depending on student's expertise).
- Run the existing GraphMDL-based classification implementation, then *analyze and present* the results and compute metrics to assess the performances.
- Critical assessment of the usability of GraphMDL (and more generally patterns) as a way to compare and classify languages.

3 Relevant reading

We list here some relevant material that may be used as an entry point for the literature around this subject.

- The Universal Dependencies website: [1]
- For an introduction to graph and tree mining and their terminology, the first 15 pages of the following survey may be of interest (the survey is about mining *dynamic* graphs, which is out of scope here, but the section on static graph mining is well written): [4]
- The main GraphMDL paper: [2]
- Detecting language differences using the MDL principle on sentences represented as *sequences*: [6]

References

[1] Universal Dependencies. `http://universaldependencies.org`.

[2] Francesco Bariatti, Peggy Cellier, and Sébastien Ferré. GraphMDL: Graph Pattern Selection Based on Minimum Description Length. In *Advances in Intelligent Data Analysis XVIII*, Lecture Notes in Computer Science, pages 54–66, 2020. doi: 10.1007/978-3-030-44584-3_5.

[3] Francesco Bariatti, Peggy Cellier, and Sébastien Ferré. GraphMDL+: interleaving the generation and MDL-based selection of graph patterns. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, pages 355–363, March 2021. doi: 10.1145/3412841.3441917.

[4] Philippe Fournier-Viger, Ganghuan He, Chao Cheng, Jiaxuan Li, Min Zhou, Jerry Chun-Wei Lin, and Unil Yun. A survey of pattern mining in dynamic graphs. *WIREs Data Mining and Knowledge Discovery*, 10(6), 2020. doi: 10.1002/widm.1372.

[5] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[6] Martin Kroon, Sjef Barbiers, Jan Odijk, and Stéphanie van der Pas. Detecting syntactic differences automatically using the Minimum Description Length principle. *Computational Linguistics in the Netherlands Journal*, 10:109–127, December 2020. `https://www.clinjournal.org/clinj/article/view/109`.