Lab 3. Crossings in syntactic dependency structures.

Introduction to Quantitative Linguistics

2022-2023 (2nd edition)

Ramon Ferrer-i-Cancho

May 27, 2023

1 Introduction

In this lab session, we are going to investigate the hypothesis the scarcity of crossing dependencies in languages is one of the manifestations of the syntactic dependency distance minimization principle. You will have to reproduce previous results on $\mathbb{E}_2[C]$, a predictor of the number of crossings (C) based on the probability that two dependencies cross assuming that their length is as in the original sentence but the linear placement of vertices is at random [Gómez-Rodríguez and Ferrer-i-Cancho, 2017, Ferrer-i-Cancho, 2014]. The relative error of a predictor x is defined as

$$\Delta_x = \frac{\mathbb{E}_x[C] - C}{q}, \tag{1}$$

where C is the true number of crossings and q is the number of independent pairs of edges. Hence, the relative error of $\mathbb{E}_2[C]$ is

$$\Delta_2 = \frac{\mathbb{E}_2[C] - C}{q}. \tag{2}$$

The first novelty of this lab with respect to previous research is to test the hypothesis that chunk structure constraints even further the number of crossings. Put differently, that the scarcity of crossings dependencies results from a combination of dependency distance minimization and chunking. To that aim, the relative error is adapted as

$$\Delta_2^{chunk} = \frac{\mathbb{E}_2^{chunk}[C] - C}{q}, \tag{3}$$

where

- $\mathbb{E}_2^{chunk}[C]$ if the prediction of $\mathbb{E}_2[C]$ on the new sentence that results from collapsing all vertices forming a chunk into a single node, procedure inspired by renormalization in physics.
- C is the number of edge crossings in the original sentence (before renormalization).
- q is the number of independent edges in the original sentence (before renormalization).

The second novelty of this lab is the use of a new error score Γ where normalization by q is replaced by normalization by $\mathbb{E}_0[C]$. Γ is defined as

$$\Gamma_x = \frac{\mathbb{E}_x[C] - C}{\mathbb{E}_0[C]}, \tag{4}$$

Accordingly,

$$\Gamma_2^{chunk} = \frac{\mathbb{E}_2^{chunk}[C] - C}{\mathbb{E}_0[C]}. \tag{5}$$

2 Data preparation

The starting point is the Parallel Universal Dependencies (PUD) treebank collection. Here we consider two annotation styles, the original Universal Dependencies style (UD) and the Surface-Syntactic Universal Dependencies style (SUD), which leads to two versions of the PUD collection respectively, namely the PUD collection and the PSUD collection [Ferrer-i-Cancho et al., 2022]. These versions are already available from `https://cqllab.upc.edu/lal/universal-dependencies/`. Within the treebank of a collection, each sentence is coded using the head vector format. See the documentation of LAL to understand this format.

For each version, you have to produce a variant where chunks have been collapsed into a single node. You have to consider two definitions of chunk: Anderson's chunks [Anderson, 2021, Section 3.3.1] and Macutek et al segments [Mačutek et al., 2021]. This leads to four collections.

Chunks must be computed from the syntactic dependency structure as head vector (do not use more sophisticated methods based on the original texts or treebanks). For Anderson's chunks, the relevant definition is [Anderson, 2021, p. 53] *"Here we loosen the definition of a chunk and consider any base-level subtree a possible chunk defined by the following criteria: (i) the components of a chunk are syntactically linked; (ii) there is only one level of dependency (one head and its dependents); (iii) the components are continuous; and (iv) no dependent within a chunk has a dependent outside the chunk."*

3 Data analysis

You have to use the Linear Arrangement Library (LAL) [Alemany-Puig et al., 2021] to check that the head vectors in each of collections are correct and then compute

- n, the number of words in the sentence.
- C, the number of crossings.
- q, the number of independent pairs of edges.
- $\mathbb{E}_0[C]$, the expected number of crossings in random shuffling of the word of a sentence.
- The prediction on the number of crossings by $\mathbb{E}_2[C]$.

The verification step is critical for the new collections that you have to produce renormalizing chunks. Check also that the number of sentences per language within each version of the collection is the same.

Documentation and releases are available from `https://cqllab.upc.edu/lal/`

There are two options available to apply the different chunking methods to the treebanks. The first, and also the simplest, is to use the treebank-parser application that is available here `https://github.com/LAL-project/treebank-parser/`. It is an easy-to-use command-line application to parse a file in a given format and to apply filtering and transformation operations on the trees as a whole and on particular words. The documentation of this tool is available online in its github repository. The second option is to use the treebank reader class available in LAL to iterate over the trees and apply whatever filtering and transformation operations you wish. Documentation and usage examples of this class are available in the Quick Guide of the latest version of LAL.

To calculate the relevant metrics, there are again two options. The first one, and also the simplest, is to use the treebank collection processor class available in LAL. This class calculates whatever "features" you want on the treebank collection of your choice. Documentation and usage examples of this class are available in the Quick Guide of the latest version of LAL. The second option, which requires more coding than the first, is to use the treebank reader class available in LAL to iterate over the trees and calculate the metrics you want.

3.1 Projectivity in languages

For each of the four original collections and the four ones that result from renormalizing chunks, you have to produce a table indicating, for each language, the average sentence length (n), the proportion of projective sentences, the proportion of planar sentences, the average number of crossings (C), the average expected number of crossings in random linear arrangements ($q/3$) and the average relative number of crossings $\frac{C}{q}$.

3.2 The hypothesis that chunks reduce the probability that two edges cross even further

3.2.1 Part 1

You have to generate the following materials

- For each combination annotation styles (UD versus SUD) and chunk definition (Anderson's and Macutek et al's), you have to produce a multipanel figure showing the average Δ_0, Δ_2 and Δ_2^{chunk} as a function of n, the number of vertices of the tree, for a selection of at least 6 languages, each language in a distinct panel. For these six languages, we suggest that you arrange the panels as a matrix with three rows and two columns. Plots must be generated using `ggplot2` with facets in R (one facet for every language) or another tool of equivalent power and visual quality. Use colors and dotted/dashed lines to ease readability.
- Tables summarizing distribution of Δ_0, Δ_2 and Δ_2^{chunk} for each language over sentences of any length. In these tables, a row corresponds to a specific language. You should use at most one table for each of the four collections. You may find convenient to display one table with Macutek et al's segments combining results for UD and SUD style and another table displaying the same information for Anderson's segments.

3.2.2 Part 2

Produce a new version of the materials in Part 1 where Δ_x and Δ_2^{chunk} are replaced by Γ_x and Γ_2^{chunk} respectively.

4 Report

The report must contain at least the following sections:

- Results. That section should contain the materials above and other results you may find relevant or useful.
- Discussion. A discussion of the results. Some possibilities are
  - Differences between languages.
  - The kind of error made by the predictor. According to previous research Δ_2 should indicate overestimation, i.e. $\Delta_2 > 0$ [Gómez-Rodríguez and Ferrer-i-Cancho, 2017, Ferrer-i-Cancho, 2014]. Is Δ_2^{chunk} still indicating overestimation?
  - The range of the error made by each predictor. In previous research, it has been found that Δ_2 indicates a relative error that does not exceed 5% in sufficiently large sentences [Gómez-Rodríguez and Ferrer-i-Cancho, 2017, Ferrer-i-Cancho, 2014]. What is Δ_2^{chunk} indicating?
  - Results with Δ versus results with Γ.

  - Reflections on the validity or power of the hypothesis that chunking reduces crossings. Is renormalization of chunks reducing error to realistic degree?
  - Differences in the results due to annotation style or chunk definition.
  - ...

- Methods section. That sections contains any relevant details about the methods that you have used (do not hide important ideas or decisions in your code).

4.1 Teaming

Work is done in pairs and two people cannot pair more than once for a lab session. If you think that you cannot satisfy these constraints ask the professor as soon as possible (before the deadline). Every team must work independently from other teams.

For this lab, pair in a way that there is at most one member of has a Mac computer. Installation in Mac is tricky. However, Lluís Alemany will help solve installation problems in Mac and other environments. You can contact him via lluis.alemany.puig@upc.edu (if you write, him please add ramon.ferrer@upc.edu as cc).

4.2 To deliver

One member of the team has to deliver the report and related materials as a single .zip file. The name of the .zip file has to indicate the names of the students (for instance, `name1_name2.zip`) and must contain

- A PDF with the report in the main directory.
- A folder called "text" with the source .tex file and the figures and tables used to produce the PDF of the report in the main folder.
- A folder called "code" with the code (R or Python scripts) that you have used to produce the materials.
- A folder called "data" with the original treebank collections, each of the four new treebank collections as head vectors and other data generated in this project. The structure of the subfolder data should be

```
data/treebanks/original/UD
data/treebanks/original/SUD
data/treebanks/Anderson/UD
data/treebanks/Anderson/SUD
data/treebanks/Macutek/UD
data/treebanks/Macutek/SUD
data/other
```

Within each treebank folders, each file contains head vectors and its name of the form `language.txt` where *language* is the ISO code of the language.

The .zip file has be sent to ramon.ferrer@upc.edu with the header "[IQL] Lab 3" before June 12, 2023.

References

[Alemany-Puig et al., 2021] Alemany-Puig, L., Esteban, J. L., and Ferrer-i-Cancho, R. (2021). The Linear Arrangement Library. A new tool for research on syntactic dependency structures. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 1–16, Sofia, Bulgaria. Association for Computational Linguistics.

[Anderson, 2021] Anderson, M. (2021). *An Unsolicited Soliloquy on Dependency Parsing*. PhD thesis, Departamento de Ciencias de la Computación y Tecnologías de la Información, Universidade da Coruña, Spain.

[Ferrer-i-Cancho, 2014] Ferrer-i-Cancho, R. (2014). A stronger null hypothesis for crossing dependencies. *Europhysics Letters*, 108(5):58003.

[Ferrer-i-Cancho et al., 2022] Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., and Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308.

[Gómez-Rodríguez and Ferrer-i-Cancho, 2017] Gómez-Rodríguez, C. and Ferrer-i-Cancho, R. (2017). Scarcity of crossing dependencies: a direct outcome of a specific constraint? *Physical Review E*, 96:062304.

[Mačutek et al., 2021] Mačutek, J., Čech, R., and Courtin, M. (2021). The menzerath-altmann law in syntactic structure revisited. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 65–73, Sofia, Bulgaria. Association for Computational Linguistics.