Quality Check and Expansion of Small Treebanks

Akshay Aggarwal[⋆], **Chiara Alzetta**[◇]

[⋆]Twilio Czechia s.r.o, Prague, Czechia

[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa, ItaliaNLP Lab – *www.italianlp.it*

aaggarwal@twilio.com, chiara.alzetta@edu.unige.it

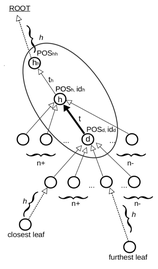
Linguistically annotated language resources (e.g., treebanks) are fundamental for training and testing tools and to acquire linguistic evidence from corpora. Ideally, they should be large (showing as many different examples of language use as possible) and coherent (having similar constructions sharing the same annotation representation). This poster presents a methodology to perform treebank **quality check** and **expansion** in a single workflow. The methodology is specifically designed to be applied to **small treebanks** (word count $<$ 100,000).

Approach

We propose a methodology which allows to perform both treebank quality check and expansion. Our methodology relies on *LISCA* (*LInguiStically–driven Selection of Correct Arcs*) [Dell'Orletta et al.], an unsupervised linguistically–driven algorithm which assigns a score quantifying the plausibility of individual arcs (deprels) within dependency–based representations. Traditionally, plausibility is computed based on a large set of examples seen during a preliminary linguistic model creation phase. We adapted the traditional LISCA workflow as proposed by Aggarwal [2020] in order to apply the methodology to small treebanks.

Method Workflow

Step 1) Split the treebank into 4 equally-sized portions (1/4 of the sentences each).
Step 2) Use LISCA to collect statistics about linguistically–motivated features from the examples reported within 3 portions of the treebank and obtain a statistical linguistic model (SLM).
Step 3) Calculate a plausibility score (as a product of individual feature weights) for each deprel of the 4th treebank portion based on the LISCA SLM.
Repeat Steps 2–3 until all portions are analysed.
Step 4) Merge all portions and re-order the relations based on their obtained plausibility score in order to have a single ranking containing all relations of the treebank.
Step 5a) To perform **quality check**: inspect relations obtaining lowest scores, which have more chances to be errors.
Step 5b) To perform **treebank expansion**: use the obtained scores to collect test suites containing sentences with similar LISCA scores. These could be added to the treebank to expand it with novel unseen examples.



Global and local features characterising a deprel, defined as a triple (d=dependent, h=head, t=dependency)

Data and Languages

The 1,000 sentences of Parallel Universal Dependencies (PUD) treebanks, covering *Newswire* and *Wikipedia* texts, for the following languages (language family between parenthesis): **Arabic** (Afro-Asiatic, Semitic) **Czech** (IE, Slavic), **English** (IE, Germanic), **Hindi** (IE, Indic), **Italian** (IE, Romance), **Indonesian** (Austronesian), **Finnish** (Finno-Uralic), and **Thai** (Tai-Kadai).

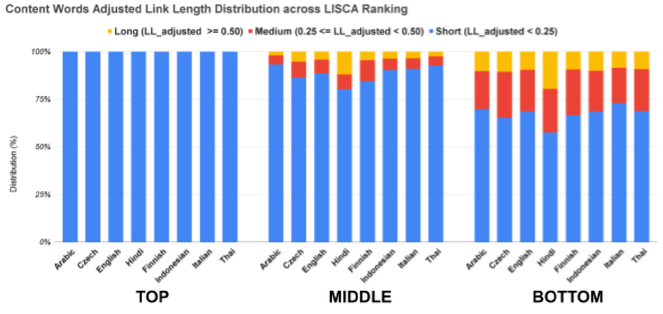
Application 1: Quality Check

Goal: Guarantee that similar constructions share the same annotation. As the LISCA score is computed on the basis of contextual linguistic information, deprels obtaining low LISCA scores have the higher chances of showing annotation errors (or at least anomalous constructions). We split the LISCA–based ranking and inspected the characteristics of deprels located in the last positions (having lower scores), as in [Alzetta et al., 2017].

Link Length Distribution across LISCA–based Rankings

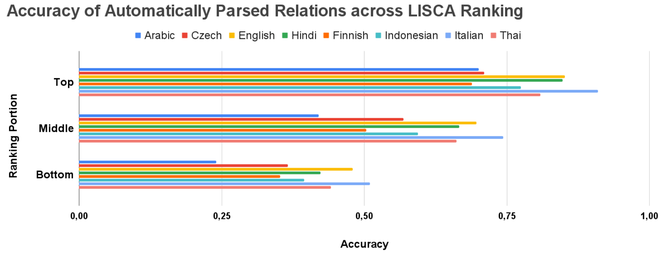
To allow multilingual comparison, we use *(a)* only content words; *(b)* a normalised adjusted link-length instead of the raw link-length, factoring in a Brevity Penalty for small sentences. Results show that for all languages, LISCA assigns lower scores to longer links, possibly owing to their higher complexity.

$$\text{Norm. } LL_{adjusted} = \begin{cases} \frac{LL_{raw} \cdot \exp\left(min\left(1-\frac{TrbAvgSentLen}{SentLength}, 0\right)\right)}{SentLength} & \text{if } \frac{SentLength}{TrbAvgSentLen} < 0.5 \\ \frac{LL_{raw}}{SentLength} & \text{if } 0.5 \leq \frac{SentLength}{TrbAvgSentLen} \leq 1.25 \ \& \\ & LL_{raw} < \lfloor TrbAvgSentLen \rfloor \\ min\left(1, \frac{LL_{raw}}{TrbAvgSentLen}\right) & \text{otherwise} \end{cases}$$



LISCA Ranking Portions Accuracy

Computed for automatically parsed sentences, it allows us to verify whether low-scores deprels are also more difficult to parse. Results show that wrongly parsed relations mostly concentrate in the bottom part of the ranking for all languages.



Application 2: Treebank Expansion

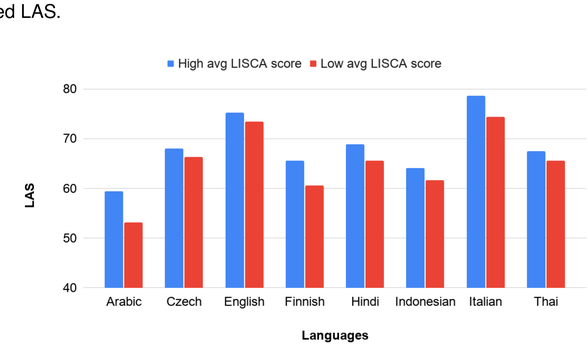
Goal: Treebank expansion is extremely valuable for low resourced languages as it allows the addition of new unseen examples to treebanks. We verify if LISCA can support a faster and efficient way to expand training sets or to create test suites homogeneous in their complexity.

Hypothesis

Sentences with low average LISCA scores (containing deprels ranked mostly towards the bottom) should be also more difficult to parse than those with higher average LISCA score.

Experiment

We collected 2 sets of 100 sentences from 1/4 of PUD: one containing sentences with highest average LISCA scores, the other containing sentences showing lowest average LISCA scores. We parsed each set with UDPipe (trained with remaining 3/4 of PUD) and computed LAS.



Note: for language comparability, avg LISCA score is computed considering only content words.

Result

The set with higher average LISCA score is more accurately parsed than the lower-score set for all languages.
LISCA can be used for automatic collection of test suites that are homogeneous in parsing complexity. These are particularly valuable for small treebanks as the suites with low average LISCA scores contain information not yet present in the treebank. Furthermore, they can also be used for not-low-resourced languages to test parsers on specific linguistic phenomena, such as those of language complexity, or for corpus analyses.

References

A. Aggarwal. Consistency of Linguistic Annotation. Master's thesis, Univerzita Karlova (ÚFAL), Prague, Czechia, 2020.

C. Alzetta, F. Dell'Orletta, S. Montemagni, and G. Venturi. Dangerous Relations in Dependency Treebanks. In *Proceedings of the TLT16*, 2017.

F. Dell'Orletta, G. Venturi, and S. Montemagni. Linguistically-driven Selection of Correct Arcs for Dependency Parsing. *Computación y Sistemas*.