Towards a MWE-driven A* parsing with LTAGs [WG2, WG3]

Jakub Waszczuk and Agata Savary
Université François Rabelais Tours, France

Introduction

Motivation: Syntactic parsing is known to potentially produce a high number of interpretations for a sentence. Reaching MWE-based derivations faster than the compositional alternatives can be advantageous for parsing efficiency, for instance in:

(1) ***Acid rains*** *in Ghana are equally grim.*

Objective: Propose a TAG parsing architecture that:

- ◇ will allow to systematically promote MWE-oriented interpretations within a kind of online ranking,
- ◇ can be informed about corpus-based probabilities of lexical units (MWEs, in particular), which will improve (or replace) the ad-hoc behavior described above.

Promoting MWEs in A* parsing

A*-based chart parsing allows to find the lower-weight derivations *before* the less probable alternatives.
Starting point: We assign the same positive weight w_0 to all elementary trees. As a result, the lowest-weight derivations are naturally those which contain MWEs.
Heuristic: For a given chart item, we compute the lower-bound estimate on the weight of the elementary tree and the part of the sentence remaining to be parsed.

- ◇ For each terminal t, we estimate the minimal weight $w(t)$ of scanning it based on the underlying grammar.
- ◇ For a given span s, we define the estimated weight of parsing the remaining part of the sentence as:

$$h(s) = \sum_{t \in out(s)} \nu_{out(s)}(t)\, w(t)$$

where $out(s)$ is a multiset of terminals present in the input sentence outside of span s and $\nu_m(t)$ is the multiplicity of element t in multiset m.

To estimate the remaining weight for a chart item q and the corresponding span s, which together represent a partial matching of one or more elementary trees over s:

- ◇ We consider all grammar subtrees t and the corresponding elementary trees r that q can potentially represent,
- ◇ For each (t, r) pair, we compute the estimated weight (augmented by r's weight) of parsing $out(s)$ after removing from it the terminals present in r but not in t.

Example

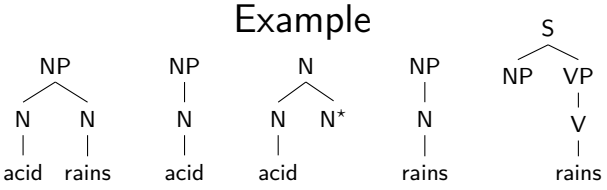

Figure: A toy LTAG grammar, each tree is assigned the weight of 1

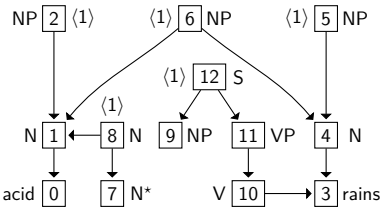

Figure: Weighted DAG representation of the TAG

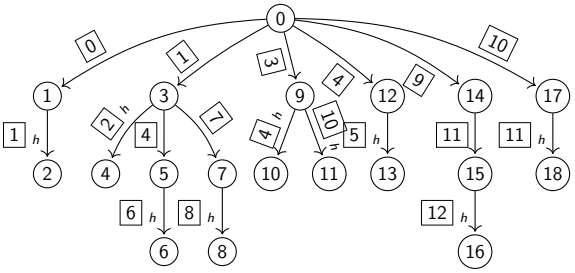

Figure: TAG compression as a prefix tree

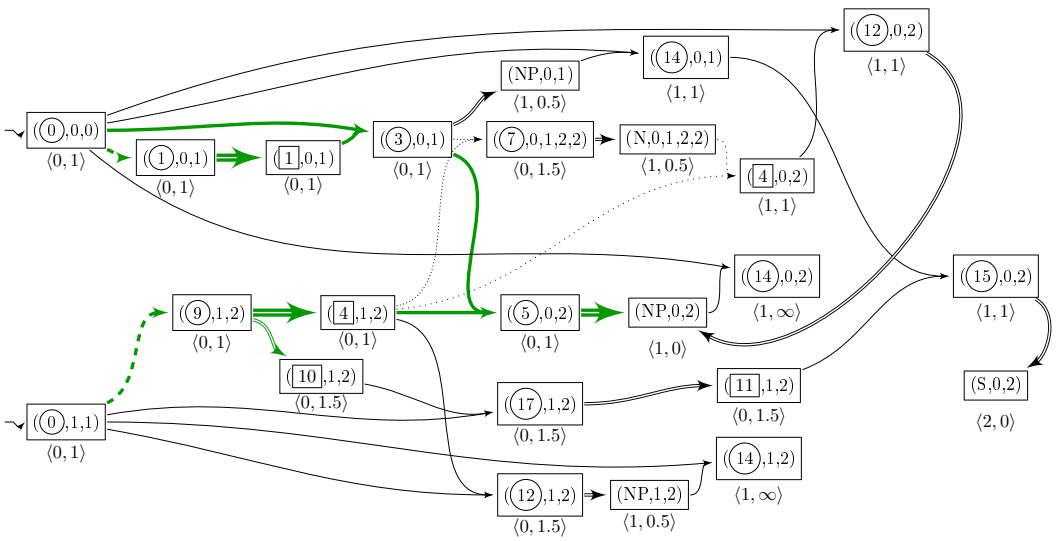

Figure: Hypergraph representing the chart parsing of the substring *acid rains* with the TAG grammar given above. The snake, plain, double, dashed, densely dotted and loosely dotted hyperarcs represent axioms, pseudo substitution, deactivate, scan, foot adjoin and root adjoin inference rules, respectively. The lowest-cost path representing the idiomatic interpretation is highlighted in bold.