Ph.D. *Deep Syntax and Optimisation*

Advisors: Joseph Le Roux & Adeline Nazarenko

This Ph.D. work aims at:

- developing efficient parsing methods for various formalisms used in natural language processing (NLP);
- extending these methods to systems processing syntactic and semantic information of natural languages.

In order to reach this goal, we plan to use methods from combinatorial optimisation, recently applied successfully to various tasks in NLP.

1 Administrative Context

This proposal is intended to master students with the following background:

- proficient in NLP and Computational Linguistics (CL) eager to discover optimisation, or
- proficient in optimisation with a **strong** motivation to learn NLP and CL.

The candidate will join RCLN team, specialized in NLP, at Université Paris 13 Computer Science Laboratory LIPN, and will collaborate with other teams, i.e. in optimisation and machine learning. This Ph.D is funded by the French *Laboratoire d'excellence* LABEX EFL, in particular by the strand dedicated to computational semantic analysis. As such the candidate will also interact with researchers from the various teams involved in the strand.

This work will be supervised by Joseph Le Roux and Adeline Nazarenko and funded from October 2014 to October 2017 (3 years). Applications must contain:

- a CV,
- a copy of Master's grades,
- a cover letter,
- reference letters.

To apply, and for any additional information, please contact leroux@univ-paris13.fr.

2 Scientific Context

The main focus of this work will be the syntactical analysis of sentences, parsing, and its relation to semantic processing.

Besides natural language ambiguity, several factors can impact the complexity of parsers:

1. Interactions between elementary parts associated to (groups of) words must be taken into account when evaluating global structures associated to sentences (i.e. interactions between edges or nodes when evaluating trees):
   - this is the case in formalisms commonly used in NLP such as higher-order dependency parsers, where the scoring function must consider tuple of edges, an rule-based formalisms such as PCFG-LAs where some interaction between nodes are implied by rewriting rules over latent variables.
   - this is also the case in formalisms which have been used for a long time in NLP/CL such as tree adjoining grammars (TAGs) and other closely related formalisms (MCTAGs, RCGs, LCFRSs, MCFGs...) with an expressive power greater than context-free grammars, at the expense of more complex composition rules.
2. Syntactic parsing is not an isolated task and must be thought as one step in a text comprehension architecture. As a result, it is often conjoined to tasks closer to text (ie, tokenization or morphological architecture) or tasks closer to meaning (ie, semantic analysis). Joint resolution of the associated tasks helps interactions between linguistic levels to be accounted for, but is a much more challenging problem.

In any case, it is interesting (linguistically, algorithmically, or from the point of view of software engineering) to divide the main problem into several subproblems. While this division is obvious when the problem is the simultaneous processing of linguistic levels, it may be more challenging when the problem is focused on a specific task. Once this division performed, it remains to define how to combine the resulting substructures. Traditionally two paradigms were used in NLP:

cascading subsystems are chained sequentially. The global consistence is guaranteed but this type of architecture leads to error accumulation in practice.

dynamic programming the set of states in the global system is constructed as the cartesian product of the possible states of each subsystems, where forbidden combinations are removed. As appealing as it looks, this type of architecture often leads to search space explosion, even when time and space complexity are polynomial. To cope with this issue, non admissible heuristics are used to prune the search space. As a consequence, this method is not optimal in practice.

Recently other types of resolutions and combinations have been proposed for NLP tasks relying on optimisation:

dual decomposition based on Lagrange relaxation this method can be used to derive iterative algorithms in which subsystems are encouraged or penalised according to their distance to a global consensus depending on the actual task [Rush et al., 2010]. The key advantages of this method are (1) its ability to reuse efficient specialised algorithms for the subproblems and (2) the formal guarantee to provide optimal solutions if they exist. However, in theory the optimal solution can be slow to obtain.

column/row generation this method, although not a proper type of combination strictly speaking, can be used to solve problems while ignoring a large portion of variables or constraints. It also leads to iterative algorithms where subproblems of increasing sizes are solved where variables/constraints are added only when they can improve the current solution. However, it is more difficult to incorporate a priori knowledge of the specifics a problem, for instance the projectivity of the resulting parse tree.

We propose to study in more depth these two latter types of combinations for mainly two applications:

- parsing with deep syntactic parsing formalisms such as tree adjoining grammars,
- semantic-augmented parsing.

3 Proposition

3.1 Towards a New Approach to Deep Syntax

For several formalisms (TAG, MC-TAG, LCFRS, SRCG...) which are known to offer a clean syntax/semantics interface, the parsing complexity, even when polynomial, makes it impossible to process large contents in practice. First, this work will begin with the careful study of how parsing using these formalisms, which may be seen as constrained phrase-structure grammars, can be cast as a constrained optimisation problem and thus solved with techniques such as lagrangian relaxation, using the work of [Le Roux et al., 2013] as a starting point. We hope to (1) design efficient algorithms and (2) be able to obtain a typology of these formalisms in terms of constraints over context-free grammars. The tree adjoining grammars are our main target, borrowing from [Carreras et al., 2008] and [Schmitz and Le Roux, 2008].

3.2 Optimal Syntax/Semantics Interface

Following work from [Rush et al., 2013] in machine translation, a promising application of lagrangian relaxation in NLP might be its capacity to filter the search space while

still providing optimal guarantees. Moreover one can see syntactic/semantic parsers as syntactic parsers with additional constraints induced by semantics. In this second part of the work we expect to evaluate this method in the case of syntactic-semantic parsing, for example in the case of synchronous frameworks derived from the formalisms studied in the previous part. Finally we would like to study column/row generation in this context. This technique has been successfully applied to higher-order dependency parsing in [Riedel et al., 2012] and we plan to carefully compare these approach with dual decomposition.

References

[Carreras et al., 2008] Carreras, X., Collins, M., and Koo, T. (2008). TAG, Dynamic Programming, and the Perceptron for Efficient, Feature-Rich Parsing. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 9–16, Manchester, England. Coling 2008 Organizing Committee.

[Le Roux et al., 2013] Le Roux, J., Rozenknop, A., and Foster, J. (2013). Combining PCFG-LA Models with Dual Decomposition: A Case Study with Function Labels and Binarization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Riedel et al., 2012] Riedel, S., Smith, D., and McCallum, A. (2012). Parse, Price and Cut – Delayed Column and Row Generation for Graph Based Parsers. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*.

[Rush et al., 2010] Rush, A., Sontag, D., Collins, M., and Jaakola, T. (2010). On dual decomposition and linear programming relaxations for natural language processing. In ACL, editor, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Rush et al., 2013] Rush, A. M., Chang, Y.-W., and Collins, M. (2013). Optimal Beam Search for Machine Translation. In ACL, editor, *Proceedings of EMNLP*.

[Schmitz and Le Roux, 2008] Schmitz, S. and Le Roux, J. (2008). Feature unification in TAG derivation trees. In Gardent, C. and Sarkar, A., editors, *TAG+9: Proceedings of the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms*, pages 141–148.