COST Action IC1207

PARSing and Multi-word Expressions

Towards linguistic precision and computational efficiency
in natural language processing

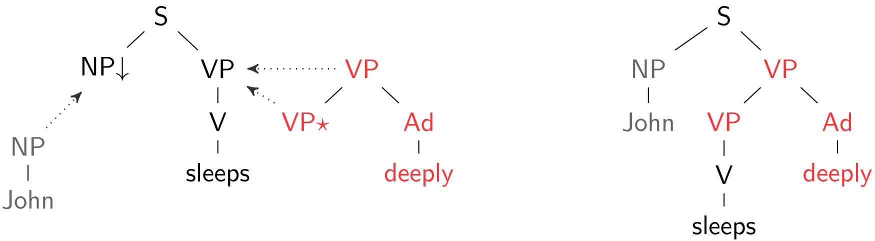
Working Group 2: PARSING TECHNIQUES FOR MWEs

Pre-processing MWEs in TAG Parsing

PARSEME General Meeting
Frankfurt-am-Main, 09 September 2014

Introduction (recall Tree-Adjoining Grammar, TAG)

- **Tree-rewriting** system [Joshi and Schabes, 1997]
- Tree-rewriting operations: Substitution / Adjunction

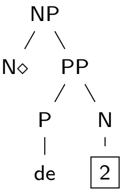


- Elementary trees built on **linguistic well-formedness constraints** (lexicalization, predicate/arguments coocurrency, semantic minimality) [Abeillé, 1993]

Introduction (continued, recall MWEs in TAG)

- **TAG's extended domain of locality** makes it possible to express long-distance dependencies within single elementary rules (trees)
- Following [Abeillé, 1995], MWEs can be represented via **dedicated TAG tree families** (often made of single trees)
- Example : tour de {magie | force}, pomme de terre

(α_{tree})



(α_{lex})

HEAD	tour
CAT	N
MORPH	[gen masc, num sg]
CO-ANCHOR	[2 magie]
FAMILY	α_{tree}
SEM	*trick*(t)

- Problem : **high redundancy** → computational cost at parsing

Plan

1. Lexical selection for TAG Parsing
2. Lexical selection and Multi-Word Expressions
3. Conclusion

TAG Parsing

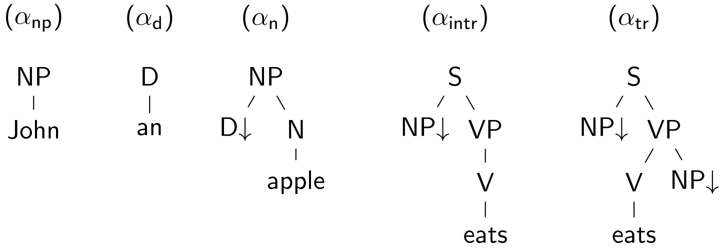
- TAG is mainly used as a **lexical formalism** → each rule is associated with at least one lexical item (≈ word)
- Parsing process :
  1. **Segmentation / POS tagging**
  2. **Subgrammar extraction** (also known as *supertagging* or *lexical selection*)
  3. **Core TAG parsing** (tree rewriting, using either top/down or bottom/up algorithms)
  4. **Feature unification** (on a factorised structure called parse forest)
- Proposal : enhancing step 2 by performing a *better filtering* to reduce the search space at (core) parsing

Supertagging

- Original idea from [Bangalore and Joshi, 1999] : **learning** which TAG tree is most likely to be associated with a lexical item *in a given context*
- Idea from [Boullier, 2003] : **finding out** which TAG tree**s** are relevant *in a given context*
- Technique used : **approximating** the input TAG with a CFG, and use the latter for parsing
- Drawback : on-line **computation cost** is still high with real-size grammars
- Idea from [Gardent et al., 2014] : approximating the input TAG with a **polarity-based automaton** encapsulating information about *left context*

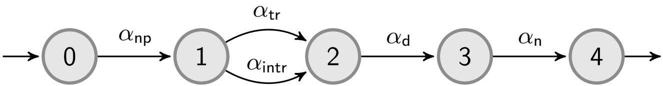
Toy example

- Input grammar



- Sentence to parse : *John eats an apple*

- Initial automaton-based grammar selection :



About polarities

- Automaton's paths contain **all** possible tree selections
  $\rightarrow$ surgeneration
- Polarities' role : keeping track of missing constituents (unsolved TAG substitutions) to remove *unsatisfiable* trees during selection
- Technique : enriching the automaton's states with couples of the form (CAT, INT) where INT is:
  - a positive integer when a constituent is given (root node)
  - a negative integer when it is needed (substitution node)

Toy example (continued)

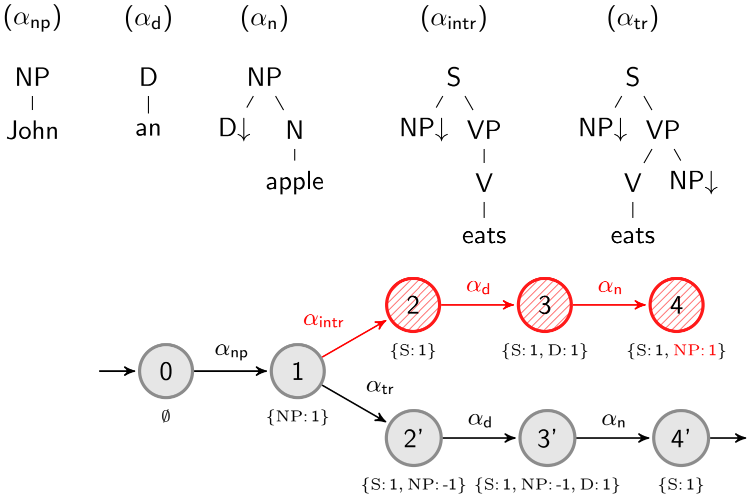


Figure : A polarity automaton for the sentence ‘John eats an apple.’

About left-context

- Idea: reducing the automaton as soon as possible, that is, once a left context is not satisfied

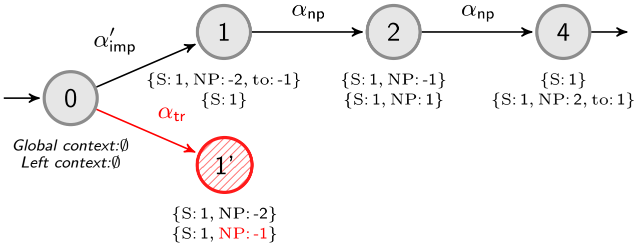


Figure : Lexical selection using left context for 'Say it to John.'

Plan

1. Lexical selection for TAG Parsing
2. Lexical selection and Multi-Word Expressions
3. Conclusion

Selecting MWEs in a TAG

- Polarity-based lexical selection can be used to process MWEs, so that:
  - **both** the trees of the literal meaning and those of the idiomatic meaning **are selected**
  - the idiomatic meaning is **prioritised**

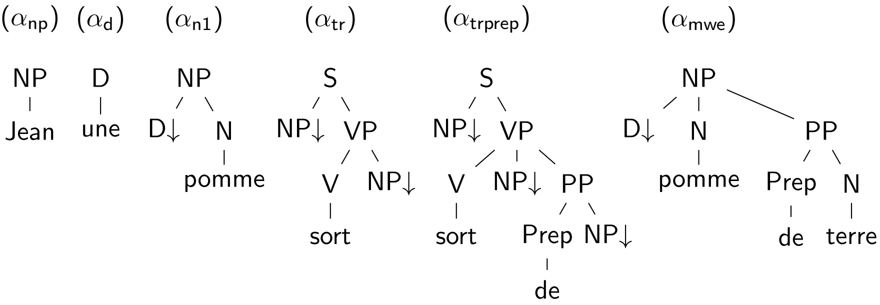
- Prioritisation is achieved by **comparing the length of the automaton paths**
  (recall that TAG trees for MWEs do not have substitution nodes)

Representing MWEs in TAG (continued)

- Example :

(1)

Jean	sort	une	pomme	de	terre
John	plucks	an	apple	from	earth
John	plucks	a	*potato*		



Representing MWEs in TAG (continued)

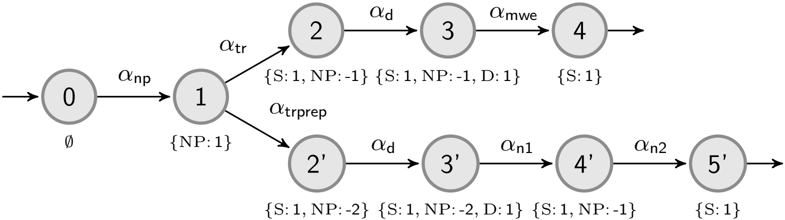


Figure : Polarity automaton for the sentence 'Jean sort une pomme de terre.'

Plan

Conclusion

- Lexicalized TAG can encode various MWEs at the price of structural redundancy
- Resulting parsing cost can be reduced by lexical selection
- Polarity-based lexical selection offers a way to characterize MWEs (useful for parsing ranking)

References I

Abeillé, A. (1993).
Les Nouvelles Syntaxes.
A. Colin – Paris.

Abeillé, A. (1995).
The Flexibility of French Idioms: A Representation with Lexicalised Tree Adjoining Grammar.
In Everaert, M., van der Linden, E.-J., Schenk, A., and Schreuder, R., editors, Idioms: Structural and Psychological Perspectives, chapter 1. Lawrence Erlbaum Associates.

Bangalore, S. and Joshi, A. K. (1999).
Supertagging: An approach to almost parsing.
Computational Linguistics, 25(2):237–262.

References II

Boullier, P. (2003).
Supertagging: A non-statistical parsing-based approach.
In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03), pages 55–65, Nancy, France.

Gardent, C., Parmentier, Y., Perrier, G., and Schmitz, S. (2014).
Lexical Disambiguation in LTAG using Left Context.
In Human Language Technology. Challenges for Computer Science and Linguistics. 5th Language and Technology Conference, LTC 2011, Poznan, Poland, November 25-27, 2011, Revised Selected Papers, volume 8387 of LNCS/LNAI, pages 67–79. Springer.

Joshi, A. K. and Schabes, Y. (1997).
Tree adjoining grammars.
In Rozenberg, G. and Salomaa, A., editors, Handbook of Formal Languages. Springer Verlag, Berlin.