Enhanced UD dependencies with Neutralized Diathesis Alternations

Marie Candito\textsuperscript{1}, Bruno Guillaume\textsuperscript{2},
Guy Perrier\textsuperscript{3} and Djamé Seddah\textsuperscript{4}

\textsuperscript{1}Univ Paris Diderot, \textsuperscript{2}Loria,
\textsuperscript{3}Univ de Lorraine, \textsuperscript{4}Univ Paris Sorbonne
Introduction

• UD scheme favors dependencies between content words
  • better cross-linguistic generalization
  • more semantic-oriented dependencies

• Yet, UD dependencies remain syntactic trees
  • Pb for well-known syntactic/semantic mismatches
Syntactic/Semantic mismatches

• Argument sharing
  • control verbs, Right-node raising, coordination…

• 1 syntactic argument = no semantic argument
  • e.g. impersonal construction

  FR: *il est arrivé* 3 personnes
  it is arrived 3 people
  « 3 people arrived »

• 2 syntactic arguments = 1 semantic argument
  • e.g. raising verbs, predicative complements

  FR: *Marie a trouvé Anna fatiguée*
  Marie has found Anna tired
  « Marie found that Anna was tired »
Beyond dependency trees

- Many proposals towards predicate-argument structures
  - Stanford dependencies (de Marneffe and Manning 08)
  - Graph banks
    - cf. in-depth analysis of 4 English graph-banks by Kuhlman & Oepen (CL, 2016)
    - the Semeval 2014 shared task on « broad coverage semantic dependency parsing » (Oepen et al. 14)
  - « Deep syntax »
    - Spanish: MTT deep trees (Ballesteros et al. 16)
    - French: Deep syntactic graphs (Candito et al. 14)
  - Tectogrammatical structures in Prague Dependency treebank …
More or less semantics

• In these proposals, e.g. labels are more or less semantic-oriented
  • syntactic labels
  • numbered arguments
    • arg0, arg1, arg2 …
    • MTT: deep syntactic arguments I, II, III …
  • semantic roles
    • patient, addressee, beneficiary …
    • as in tectogrammatical structures in Prague DT
Enhanced UD graphs

- « Enhanced dependencies »
  - Enhanced / enhanced+++ for English (Schuster & Manning, 16)
  - proposed as optional in UD v2.0
  - available for a few languages (Russian, Finnish)
Enhanced UD graphs

• 5 enhancements
  • subj. of infinitives in control/raising constructions
    \[ Paul \text{ seems to run: } run \rightarrow nsubj \rightarrow Paul \]
  • propagation of conjuncts
  • antecedent of relative pronouns
  • markers as suffixes in labels
    \[ went \rightarrowobl:into\rightarrow house \]
  • null nodes for elided predicates
    \[ Mary \text{ wants to buy a book and Jenny N1 N2 a CD} \]
This work

• Yet another proposal for enhanced UD: « Enhanced-diat »
  • that neutralizes syntactic alternations

• Implemented and evaluated on French
**Enhanced-diat**

- Enhanced-diat graphs remain mostly syntactic
  - in particular, we keep **UD syntactic labels**
  - as starting point for various kinds of semantic representations

```
Syntactic tree

Deep syntactic graph

PAS  AMR  MRS  ...
```
Enhanced-diat

• 2 enhancements over enhanced UD:
  • Add even more argumental edges, either
    • some fully determined by syntax:
      • control nouns, adj, some participles, gerunds
    • other cases not fully determined but most frequent
  • Neutralize syntactic alternations
    • recover canonical subcat frame
More argumental edges:
Example: noun-modifying participle

(a) ceux (étant) apparus en 2001

(b) ceux (ayant été) embauchés en 2007
More argumental edges: Example: infinitive adverbial clauses

- When main verb is active, with non expl subject
- subject of infinitive = subject of main verb
- in most cases (83% on Sequoia corpus)
  
  *Il mangera avant de jouer*
  
  *He will-eat before to play*
  
  « *He will eat before playing* »

- counter-example:

  *D’autres photos ont subi des retouches pour **accentuer** le drame*

  *Other photos have undergone modifications to **accentuate** the drama*
Neutralizing syntactic alternations

• recover « canonical » grammatical functions
  • the function you would get in active personal voice

• cheap way to limit linking diversity
  • e.g. proved useful for FrameNet parsing (Michalon et al. 16)

• massive for passive

• other cases (see paper):
  • impersonal, causative, mediopassive
Neutralizing syntactic alternations

• Note:
  • nsubj:pass / csubj:pass not enough to recover all arguments of passive (obl / obl:agent)
  • UD choice to distinguish functions according to POS of dependent (nsubj/csubj, obj/xcomp…) augments linking diversity
Syntactic alternation normalization for English ditransitives

- Take canonical subcat:
  - They\textit{(nsubj)} gave him\textit{(iobj)} orders\textit{(obj)}

\begin{itemize}
  \item (a) He was given orders by them
  \item (b) Orders were given to him
  \item (c) They often give orders to him
\end{itemize}
Obtaining enhanced-diat graphs for French

- 2 teams, 2 graph-rewriting systems
  - GREW (Guillaume et al. 12) : 157 rules
  - OGRE (Ribeyre et al. 12) : 115 rules
  - building on rules written for producing deep-sequoia
    (Candito et al. 14; Perrier et al. 14)

- rules written supposing gold surface tree

- mix of
  - purely deterministic cases (e.g. control verbs)
  - cases previously analyzed as "almost deterministic"
    - cf. previous example of infinitive adverbial clauses
Gold corpus for evaluation

- We produced gold graphs for 200 sentences
  - 100 from UD_French
  - 100 from UD_French-Sequoia
  - bias: obtained through adjudication of the 2 rule-based systems outputs
Quantitative assessment of enhancements

- **4804** edges in the 200 sentence gold corpus
- **956** are argumental dependents of verbs
  - approximated using core argument labels (nsubj,csubj,obj,iobj,ccomp,xcomp) + obl label
- edges added (set N): **18.9 %**
- edges with neutralized label (set A): **13.9 %**
- N U A represent **26.7 %** of arguments of verbs
Evaluation in 2 modes

• **PA+**: with manual pre-annotation of certain phenomena
  • expletive « il »
  • reflexive clitic « se » status (for mediopassive)
  • canonical subjects in causative constructions
  • agents of passives (by-phrases : obl:agent)

• **PA-**: no pre-annotation, handling by rules known to be approximative
Evaluation in 2 modes

<table>
<thead>
<tr>
<th></th>
<th>PA−</th>
<th></th>
<th>PA+</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEQ_{test}</td>
<td>UD_{test}</td>
<td>SEQ_{test}</td>
<td>UD_{test}</td>
</tr>
<tr>
<td>All edges</td>
<td>OGRE</td>
<td>98.81</td>
<td>99.17</td>
<td>99.46</td>
</tr>
<tr>
<td></td>
<td>GREW</td>
<td>99.44</td>
<td>99.54</td>
<td>99.69</td>
</tr>
<tr>
<td>N ∪ A edges</td>
<td>OGRE</td>
<td>86.20</td>
<td>89.89</td>
<td>92.51</td>
</tr>
<tr>
<td></td>
<td>GREW</td>
<td>93.42</td>
<td>94.31</td>
<td>95.77</td>
</tr>
</tbody>
</table>

Table 1: Evaluation of rule-based systems producing enhanced graphs: F-measures computed on all edges (top) or only on edges in $N$ or $A$ (bottom);
Conclusion

- Production of high quality enhanced UD graphs proved feasible for French
  - a little better with pre-annotation of a few not-so-deterministic phenomena

- **Quality**: accurate enough to serve as pseudo-gold for data-driven methods

- **Impact**: when considering arguments of verbs:
  - 19% are enhanced edges
  - 14% have a label modified by neutralizing syntactic alternation
Conclusion (cont)

• Other languages?
  • Romance
  • English:
    • diathesis alternations used for some experiments for the EPE shared task
    • Paris / Stanford system (Schuster et al. 17)
Thank you!

Questions?

data / rules available at https://github.com/bguil/Depling2017