What to do with a parser? Learn!

Éric de la Clergerie

INRIA Paris & University Paris-Diderot
http://alpage.inria.fr

NLP Meetup
Paris, November 23rd 2016
My main research topics:

- parsing technologies (symbolic, statistics, hybrid)
- FRMG a large coverage French (meta)grammar \( \sim \) parser

Several output annotation schemas: richer native DepXML, but also PASSAGE, FTB/CONLL, Universal Dependencies, ...
What can be done with parsing?

Since 2004, FRMG has become an efficient, accurate, & large coverage parser (on journalistic French TreeBank [FTB]: \( \text{LAS} \sim 88\%, \text{coverage} > 97\% \))

but 2 main questions:

- **What to do with a parser?**
  - Information Extraction ([http://passage.inria.fr/SAPIENS](http://passage.inria.fr/SAPIENS))
    Citation extraction from AFP news about Presidential Campaign 2007
  - Question-Answer
  - ...
  - **Knowledge Acquisition** (knowledge bottleneck)

- **How to continue to improve parsing?**
  - \( \sim \) **knowledge injection** for syntactic disambiguation
    *tremblement de terre de forte magnitude* (earthquake with high magnitude)

  \( \sim \) **virtuous circle** between language and knowledge
Knowledge Acquisition experiments

Two main directions explored during FUI SCRIBO (circa 2010)

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Events</th>
</tr>
</thead>
</table>
| Terminology extraction  
*garde à vue*  
implant chirurgical non actif  
[implant/nc] [chirurgical/adj] [non/adv] [actif/adj]  
| event-based verb clustering  
▷ /transfer/ *donner, offrir, céder*  
▷ /communication act/ *annoncer, indiquer, affirmer*  
| Semantic networks  
| Word clustering *(synset)*  
| Ontological relations (eg. hyperonymy)  
*warship: destroyer, aviso*  
| verb-noun pairs  
▷ *déclarer/déclaration* ;  
▷ *identifier/identification* ;  
▷ *commencer/commencement/début*  
| relations between named entities  
appartenance(PERS,ORG)  

Step 0 – Knowledge sources: parsed corpora

A large heterogeneous “general” corpus

<table>
<thead>
<tr>
<th>Corpora</th>
<th>#Msent</th>
<th>#Mwords</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia (fr)</td>
<td>18.0</td>
<td>178.9</td>
<td>504K encyclopedic pages</td>
</tr>
<tr>
<td>Wikisource (fr)</td>
<td>4.4</td>
<td>64.0</td>
<td>12.8K literacy texts</td>
</tr>
<tr>
<td>EstRepublicain</td>
<td>10.5</td>
<td>144.9</td>
<td>journalistic</td>
</tr>
<tr>
<td>JRC</td>
<td>3.5</td>
<td>66.5</td>
<td>European directives</td>
</tr>
<tr>
<td>EuroParl</td>
<td>1.6</td>
<td>41.5</td>
<td>parliamentary debates</td>
</tr>
<tr>
<td><strong>AFP</strong></td>
<td>14.0</td>
<td>248.3</td>
<td>400K news</td>
</tr>
<tr>
<td><strong>Total ALL</strong></td>
<td>52.0</td>
<td>744.2</td>
<td></td>
</tr>
</tbody>
</table>

But also smaller specialized corpora (some from a law editor)

<table>
<thead>
<tr>
<th>Corpora</th>
<th>#Msent</th>
<th>#Mwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>fiscal</td>
<td>7.2</td>
<td>145.2</td>
</tr>
<tr>
<td>social</td>
<td>6.8</td>
<td>127.5</td>
</tr>
<tr>
<td>civil</td>
<td>2.6</td>
<td>40.9</td>
</tr>
<tr>
<td>business</td>
<td>7.2</td>
<td>133.8</td>
</tr>
</tbody>
</table>

And several others: botanical corpus, medical, automobile, travel stories, ...
“Beware the Jabberwock, my son! 
The jaws that bite, the claws that catch! 
Beware the Jubjub bird, and shun 
The frumious Bandersnatch!”

Il était grilheure; les slictueux toves 
Gyraient sur l’alloinde et vriblaient: 
Tout flivoreux allaient les borogoves; 
Les verchons fourgus bourniflaient.

Paul s’est cassé la binti. 
Sa fracture à la binti a été correctement réduite. 
Il a des douleurs dans la binti.
Harris distributional hypothesis

Meanings of words are (largely) determined by their distributional patterns (Harris 1968)

You shall know a word by the company it keeps (Firth 1957)

1. attach to each word a (weighted) vector of contexts, dependency-based ones in our case

2. exploit these vectors to measure the similarity of pairs of words

3. exploit word similarity to organize/group words

Many variants on these 3 points (Lin, Pantel, Pedersen, Bourrigault, ...)

But often: black box, no explanations, hard classes (no polysemy), ...  
⇒ looking for a more flexible approach
## Step 1 – Collecting and counting dependencies

<table>
<thead>
<tr>
<th>&lt;governor&gt;</th>
<th>&lt;rel&gt;</th>
<th>&lt;governee&gt;</th>
<th>&lt;freq&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>chaise_nc</td>
<td>et</td>
<td>table_nc</td>
<td>235</td>
</tr>
<tr>
<td>asseoir_v</td>
<td>sur</td>
<td>chaise_nc</td>
<td>227</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>modifier</td>
<td>long_adj</td>
<td>168</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>de=</td>
<td>poste_nc</td>
<td>115</td>
</tr>
<tr>
<td>tomber_v</td>
<td>sur</td>
<td>chaise_nc</td>
<td>103</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>modifier</td>
<td>musical_adj</td>
<td>102</td>
</tr>
<tr>
<td>se_asseoir_v</td>
<td>sur</td>
<td>chaise_nc</td>
<td>93</td>
</tr>
<tr>
<td>prendre_v</td>
<td>cod</td>
<td>chaise_nc</td>
<td>87</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>modifier</td>
<td>electrique_adj</td>
<td>82</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>modifier</td>
<td>vide_adj</td>
<td>80</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>à=</td>
<td>porteur_nc</td>
<td>80</td>
</tr>
<tr>
<td>dossier_nc</td>
<td>de</td>
<td>chaise_nc</td>
<td>78</td>
</tr>
<tr>
<td>avoir_v</td>
<td>cod</td>
<td>chaise_nc</td>
<td>71</td>
</tr>
<tr>
<td>table_nc</td>
<td>et</td>
<td>chaise_nc</td>
<td>62</td>
</tr>
<tr>
<td>chaise_nc</td>
<td>de=</td>
<td>paille_nc</td>
<td>56</td>
</tr>
</tbody>
</table>
Abstracting and completing PASSAGE dependencies (at collect time):

- rectification of passives (surface subject $\rightsquigarrow$ deep object)
- addition of se for pronominal verbs
- direct relation between an attribute and a subject $(apple, att, red)$ in *the apple is red*
- abstraction of verbs in sentential arguments  
  $(can, object, eat) \rightsquigarrow (can, object, *sentence*)$
- distribution over coordinated elements
  *he takes an apple and a beer* $\rightsquigarrow (take, object, apple) \& (take, object, beer)$
- addition of potential (ambiguous) PP attachments

  terre_nc de= magnitude_nc 344  
  tremblement_nc de=* magnitude_nc 357

- injection of candidate terms

  qualité_nc de= président_du_conseil 189  
  tremblement_de_terre de=* magnitude_nc
A dependency \((to\_sit, on, chair)\) provides

- a syntactic context \(<to\_sit \ on \ •>\) for word \(chair\)

- and, symmetrically, \(<• \ on \ chair>\) for \(to\_sit\)

<table>
<thead>
<tr>
<th>Corpora</th>
<th>#dep (millions)</th>
<th>#(distinct forms) (thousands)</th>
<th>#(distinct contexts) (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPL</td>
<td>170</td>
<td>1149</td>
<td>4</td>
</tr>
<tr>
<td>AFP</td>
<td>93</td>
<td>378</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>263</td>
<td>1366</td>
<td>5</td>
</tr>
</tbody>
</table>
Step 2 – Clustering algorithm

Inspired from Markov clustering [MCL, van Dongen] in a weighted graph connecting words to contexts, we try

- to reinforce high density of short paths
- to weaken long paths

\[
WW_{i,j} = \frac{1}{Z_i} \left( \sum_{a,b} WC_{i,a} CC_{a,b} WC_{j,b} \right)^\alpha
\]

\[
CC_{a,b} = \frac{1}{Z_a} \left( \sum_{i,j} CW_{a,i} WW_{i,j} CW_{b,j} \right)^\alpha
\]

with inflation \( \alpha > 1 \) (default: 2) et normalization \( \frac{1}{Z} \)

⇒ strengthen high coefficients, lower weak ones!
Compact matrix formulation:

\[
\begin{align*}
W &= \Gamma_\alpha (F^t C F) \\
C &= \Gamma_\alpha (G^t W G)
\end{align*}
\]

with the inflation and normalization operator \( \Gamma_\alpha \)

where:
- \( W = (ww_{i,j}) \) and \( C = (cc_{a,b}) \) are the similarity matrices to be computed
- \( F = (wc_{i,a}) \) and \( G = (cw_{a,i}) \) parameter matrices
  - \( wc_{i,a} : \) weight of context \( c_a \) for word \( w_i \)
  - \( cw_{a,i} : \) weight of word \( w_i \) for context \( c_a \)

Recursive formulation \( \leadsto \) iterative fix-point algorithm
starting from initial matrix \( W^{(0)} \)

Many extensions: bonus/malus,
transfer words \( \leftrightarrow \) contexts (\textit{chair} \( \leadsto \) \textit{stool} \( \leadsto \) \textquote{\textbullet on chair} \( \leadsto \) \textquote{\textbullet on stool}),
\ldots
What’s the usage of chairs?

The algo provides (weighted) explaining contexts for close words

<table>
<thead>
<tr>
<th>Action</th>
<th>chaise</th>
<th>divan</th>
<th>chaise</th>
<th>tabouret</th>
<th>banquette</th>
<th>divan</th>
<th>canapé</th>
<th>chaise</th>
</tr>
</thead>
</table>
Visualisation: so many bones!

Graph with about 40K edges
Visualization with TULIP (http://tulip.labri.fr/), layout BubbleTree

Others on http://alpage.inria.fr/~clerger/wnet/wnet.html
Step 3 – Validation with LIBELLEX interface

Need for local views, browsing, and validation ⇒ collaborative WEB interface
http://alpage.inria.fr/Lbx (guest/guest)

Note: collaboration with startup Lingua & Machina
Coarse-grained view already useful to detect some topological structures:

- **strongly connected bushes**: very close from semantic classes

- **threads**: progressive sense shifts

- **star-like structures**: a center with many satellites
  sometimes pertinent, often not!

- some polysemic words at the junctions between bushes

  - char (*carriage*) and chariot
    \(<\bullet \text{ modifieur atteler}>, <\text{promenade en } \bullet>\)

  - char (*tank*) and tank
    \(<\bullet \text{ de combat}>, <\text{régiment de } \bullet>\)
Some topological classes

The bushes may be used to extract classes ⇒ 4000 classes (ALL)

<79> (a cluster of various kinds of dogs) sulky malinois fox-terrier setter cocker colley chiot fox labrador ratier griffon caniche teckel épagneul

<80> (a cluster of various kinds of soldiers and military groups) arrière-garde canonnier cavalerie carabinier tirailleur hussard panzer voltigeur blindé grenadier cuirassier avant-garde zouave lancier

<83> (a cluster of various kinds of diseases) pneumonie paludisme diphtérie pneumopathie variole dysenterie malaria botulisme poliomyélite septicémie varicelle polio rougeole ménningite
Step 4 – Injecting and “reasoning”

Injecting knowledge in FRMG (similarity + contexts)

*il mange une tartelette maison à la quetsche.*

- tartelette close to tarte
- quetsche a kind of fruit
- aux_fruits frequent context for tarte

⇒ tartelette à la quetsche
Recent buzz on word embeddings ("low-dimension" dense word vectors)

- word2vec [Mikolov] and Glove [Pennington, Socher, Manning]
- ≡ distributional-based approaches [Goldberg]

**DepGlove**: minimization of objective function $J \sim$ vectors $w_i$

extended to syntactic dependencies $r$ in subject, object, ... with matrices $M_r$

$$J = \sum_{i,r,j} f(X_{irj})(w_i^T M_r w_j + b_i + b_j + b_r - \log X_{irj})$$

with $f(X_{irj}) = \begin{cases} 
\left( \frac{X_{irj}}{x_{\text{max}}} \right)^\alpha & \text{if } X_{irj} < x_{\text{max}} \\
1 & \text{otherwise}
\end{cases}$

$r$ extended to longer syntactic paths between words

- 2 brothers of a same governor: cat subject+eat+object mouse
- a grand-father with a grand-son: cat subject/of eat
  in *a large majority of cats does not eat mouses*

**Note**: $w_i^T M_r$ similar to a vector for a syntactic context (word + relation)
Random generation of TOEFL-like tests from French WordNet (FWN) synsets

toutefois  néanmoins  complètement  progressivement  sensiblement
exploit  prouesse  offset  plie  bit

MCL and DepGlove not evaluated exactly the same way
(graph shortest paths for MCL, minimal cosine for DepGlove)

Many parameters to explore

<table>
<thead>
<tr>
<th>categories</th>
<th>passage/MCL</th>
<th>passage/depGlove</th>
<th>depxml/depGlove</th>
<th>d+path/depGlove</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>51.00</td>
<td>67.68</td>
<td>71.52</td>
<td>76.34</td>
</tr>
<tr>
<td>nouns</td>
<td>94.00</td>
<td>68.36</td>
<td>72.01</td>
<td>77.00</td>
</tr>
</tbody>
</table>

\[ d = 200, \text{wmin} = 20 \]

- Influence of the algorithm: MCL << DepGlove on recall, mais interest of MCL for precision
- Influence of annotation schema: Passage < DepXML
- Influence of collected data: dependancies < paths
Playing with FRMG

Long-standing effort to ease installation and use of Alpage’s tools

FRMGWIKI a linguistic wiki with extended functionalities

http://alpage.inria.fr/frmgwiki

- exploring and using FRMG
- discussing syntactic phenomena with sample sentences
- access to a corpus processing service
- links to LIBELLEX and Word Vectors

http://alpage.inria.fr/Lbx
http://alpage.inria.fr/depglove/process.pl

large parsed corpora available (Wikipédia, Wikisource, ...)
corpus indexing and querying
Merci

Questions bienvenues