Multiword Expression Identification with Recurring Tree Fragments and Association Measures

Federico Sangati, Andreas van Cranenburgh

Fondazione Bruno Kessler (FBK) Trento, Italy
Huygens ING, Royal Netherlands Academy of Arts & Sciences
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam

June 4, 2015

Denver, NAACL 2015 MWE workshop
Overview

Main Idea  MWEs from recurring syntactic tree fragments
Data  Treebanks (French, Dutch, English)
Experiments
MWE representations

Word (POS) n-grams (e.g., Ramisch et al 2010)

\( \langle \text{JJ\_mountain}, \text{NN\_bike} \rangle \)
MWE representations

Word (POS) n-grams (e.g., Ramisch et al 2010)

\[
\langle \text{JJ\_mountain, NN\_bike} \rangle
\]

French Treebank (Green et al. 2011)

- MWN
  - N
  - P
  - N
  - Tour
  - de
  - force

Dutch Lassy treebank

- MWU
  - VZ
  - LID
  - N
  - aan
  - de
  - hand

lit.: on the hand, “going on.”
MWE representations

Word (POS) $n$-grams (e.g., Ramisch et al. 2010)

\[ \langle \text{JJ\_mountain, NN\_bike} \rangle \]

French Treebank (Green et al. 2011)

MWN

MWN

\[ \text{N \ P \ N} \]

\[ \text{Tour \ de \ force} \]

Dutch Lassy treebank

MWU

\[ \text{VZ \ LID \ N} \]

\[ \text{aan \ de \ hand} \]

lit.: on the hand, “going on.”

Annotated English Gigaword

VP

VP

\[ \text{VB \ NP \ PP} \]

\[ \text{get \ IN \ NP} \]

\[ \text{off \ DT \ NN} \]

the ground
Recurring fragments

- Extract only **recurring** tree fragments from treebank
- For every pair of trees, extract maximal overlapping fragments
- Using a **linear average time** tree kernel
- Number of fragments is small enough to parse with directly

Sangati & Zuidema (2011). Accurate parsing w/compact TSGs: Double-DOP
van Cranenburgh (2014). Extraction of […] fragments w/linear average time tree kernel
## Data

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Trees</th>
<th>Total Frags</th>
<th>Selected Frags</th>
</tr>
</thead>
<tbody>
<tr>
<td>French (FTB)</td>
<td>13K</td>
<td>274K</td>
<td>86K</td>
</tr>
<tr>
<td>Dutch (Lassy)</td>
<td>52K</td>
<td>536K</td>
<td>193K</td>
</tr>
<tr>
<td>English (Gigaword subset)</td>
<td>500K</td>
<td>4.3M</td>
<td>2.8M</td>
</tr>
</tbody>
</table>

**Selected fragments**: at least 1 content word, 1 other non-punctuation token.
Main Idea  MWEs from recurring syntactic tree fragments

Data  Treebanks (French, Dutch, English)

Experiments  
- MWEs by parsing with tree fragments  
  (supervised)
- MWEs by ranking tree fragments  
  (unsupervised)
Data-Oriented Parsing (Scha 1990; Bod 1992)

- A language user exploits arbitrary parts of previous language experience in the analysis/construction of new sentences.
- “idiomaticity is the rule rather than the exception” (Scha, 1990)
- Implementation: Tree-Substitution Grammar
Tree-Substitution Grammar

fragment:

\[ P(f) = \frac{\text{count}(f)}{\sum_{f' \in F} \text{count}(f')} \]

where \( F = \{ f' \mid \text{root}(f') = \text{root}(f) \} \)

derivation:

\[ P(d) = P(f_1 \circ \cdots \circ f_n) = \prod_{f \in d} p(f) \]

parse tree:

\[ P(t) = P(d_1) + \cdots + P(d_n) = \sum_{d \in D(t)} \prod_{f \in d} p(f) \]
## Parsing results

<table>
<thead>
<tr>
<th>Parser</th>
<th>F1</th>
<th>EX</th>
<th>MWE-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F R E N C H</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green et al. (2013): DP-TSG</td>
<td>76.9</td>
<td>16.0</td>
<td>71.3</td>
</tr>
<tr>
<td>Green et al. (2013): Stanford</td>
<td>79.0</td>
<td>17.6</td>
<td>70.5</td>
</tr>
<tr>
<td>disco-dop, 2DOP</td>
<td><strong>79.3</strong></td>
<td><strong>19.9</strong></td>
<td><strong>71.9</strong></td>
</tr>
<tr>
<td><strong>D U T C H</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>disco-dop, PCFG baseline</td>
<td>63.9</td>
<td>21.8</td>
<td>50.4</td>
</tr>
<tr>
<td>disco-dop, 2DOP</td>
<td><strong>77.0</strong></td>
<td><strong>35.2</strong></td>
<td><strong>75.3</strong></td>
</tr>
</tbody>
</table>
Association Measures generalized to $n$-ary sequences.

- Pointwise Mutual Information (PMI):

$$\text{PMI}(S) = \log \frac{p(S_1, S_2, \ldots, S_n)}{\prod_{i=1}^{n} p(S_i)}$$
Association Measures generalized to $n$-ary sequences.

- **Pointwise Mutual Information (PMI):**
  \[
  \text{PMI}(S) = \log \frac{p(S_1, S_2, \ldots, S_n)}{\prod_{i=1}^{n} p(S_i)}
  \]

- **Log-Likelihood Ratio (LLR):**
  \[
  \text{LLR}(S) = \log \frac{p(S_1, \ldots, S_n)}{\sum_{\sigma \in \text{CSP}(S_1, \ldots, S_n)} \prod_{s \in \sigma} p(s)}
  \]

  \[\text{CSP} = \text{Contiguous Sequence Partition}\]
Definition

Log Inside Ratio (LIR): The probability of generating a given fragment in a single step with respect to the total probability of generating it in any possible way.

\[\text{LIR}(S) = \log \frac{p(\text{frag})}{\text{inside}(\text{frag})}\]

- i.e., a ‘compositionality index’
FRENCH TREEBANK RESULTS

Signature: LL
Frags:7042 MWEs:1079

Signature: LLLL
Frags:1021 MWEs:143

Signature: LLL
Frags:3282 MWEs:777

Signature: LLLLL
Frags:395 MWEs:25
Ranking results

<table>
<thead>
<tr>
<th>Treebank</th>
<th>PMI</th>
<th>LLR</th>
<th>LIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>33.0</td>
<td>32.3</td>
<td>45.8</td>
</tr>
<tr>
<td>Dutch</td>
<td>49.4</td>
<td>46.6</td>
<td>50.5</td>
</tr>
</tbody>
</table>

F1 scores for the top 1/5 candidates wrt. extracted recurring fragments.

Gold standard from treebank annotations.
Dutch examples not in gold standard

zo nu en dan
now and then
naar aanleiding van
prompted by
in vergelijking met
in comparison with
Europese Unie
European Union
Sociale Zaken
Socioeconomic Affairs
Tweede Kamerfractie
parliamentary caucus
English examples

<table>
<thead>
<tr>
<th>PMI</th>
<th>Freq.</th>
<th>Sequence Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.0</td>
<td>6</td>
<td>VB_take NP IN_into NN_account</td>
</tr>
<tr>
<td>14.6</td>
<td>6</td>
<td>VB_take NP IN_for VBN_granted</td>
</tr>
<tr>
<td>13.6</td>
<td>7</td>
<td>VB_take DT NN_look IN_at</td>
</tr>
<tr>
<td>12.9</td>
<td>6</td>
<td>VB_take NP TO_to NN_court</td>
</tr>
<tr>
<td>12.5</td>
<td>6</td>
<td>VB_take NN RB_away IN_from</td>
</tr>
<tr>
<td>12.4</td>
<td>17</td>
<td>VB_take NP RB_away IN_from</td>
</tr>
<tr>
<td>12.0</td>
<td>6</td>
<td>VB_take JJ NN_action TO_to</td>
</tr>
<tr>
<td>11.2</td>
<td>5</td>
<td>VB_take NP RB_away IN_from</td>
</tr>
<tr>
<td>10.5</td>
<td>6</td>
<td>VB_take QP NNS_years TO_to</td>
</tr>
<tr>
<td>8.3</td>
<td>10</td>
<td>VB_take DT NN_time TO_to</td>
</tr>
</tbody>
</table>

List of English fragments conforming to the sequence pattern VB\_take X L L, sorted by PMI.
Conclusion

- MWEs from **recurring syntactic tree fragments**
- MWEs with gaps, hierarchical structure
- Improved results with Probabilistic Tree-Substitution Grammar (PTSG)
- Ranking with Association Measures
  - Log Inside Ratio (LIR) based on PTSG