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)

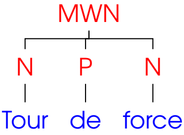
$\langle$ JJ_mountain, NN_bike $\rangle$

MWE representations

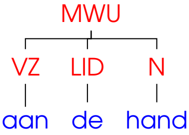
Word (POS) *n*-grams (e.g., Ramisch et al 2010)

⟨ JJ_mountain, NN_bike ⟩

French Treebank (Green et al. 2011)



Dutch Lassy treebank



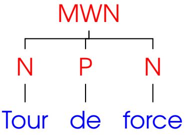
lit.: *on the hand*, “going on.”

MWE representations

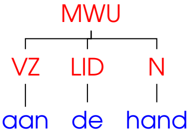
Word (POS) *n*-grams (e.g., Ramisch et al 2010)

⟨ JJ_mountain, NN_bike ⟩

French Treebank (Green et al. 2011)

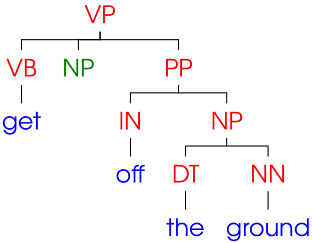


Dutch Lassy treebank



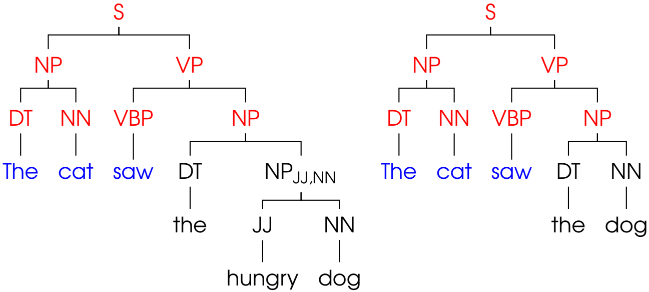
lit.: *on the hand*, “going on.”

Annotated English Gigaword



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

Data

Treebank	Trees	Total Frags	Selected Frags
French (FTB)	13K	274K	86K
Dutch (Lassy)	52K	536K	193K
English (Gigaword subset)	500K	4.3M	2.8M

Selected fragments: at least 1 content word,
1 other non-punctuation token.

Overview

Main Idea MWEs from recurring syntactic tree fragments

Data Treebanks (French, Dutch, English)

Experiments

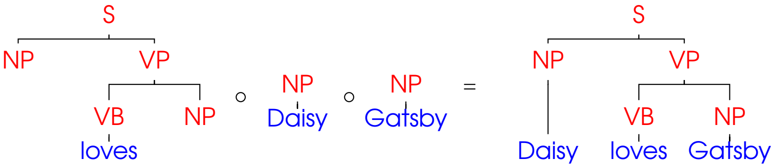
- MWEs by parsing with tree fragments (supervised)
- MWEs by ranking tree fragments (unsupervised)

Parsing

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:

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

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

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

Parser	F1	EX	MWE-F1
FRENCH			
Green et al. (2013): DP-TSG	76.9	16.0	71.3
Green et al. (2013): Stanford	79.0	17.6	70.5
disco-dop, 2DOP	**79.3**	**19.9**	**71.9**
DUTCH			
disco-dop, PCFG baseline	63.9	21.8	50.4
disco-dop, 2DOP	**77.0**	**35.2**	**75.3**

Ranking: flat

Association Measures
generalized to *n*-ary sequences.

- Pointwise Mutual Information (PMI):

$$\text{PMI}(S) = \log \frac{p(S_1, S_2, \ldots, S_n)}{\prod_{i=1}^{n} p(S_i)}$$

Ranking: flat

Association Measures

generalized to *n*-ary sequences.

- Pointwise Mutual Information (PMI):

$$\mathrm{PMI}(S) = \log \frac{p(S_1, S_2, \ldots, S_n)}{\prod_{i=1}^{n} p(S_i)}$$

- Log-Likelihood Ratio (LLR):

$$\mathrm{LLR}(S) = \log \frac{p(S_1, \ldots, S_n)}{\sum_{\sigma \in \mathrm{CSP}(S_1, \ldots, S_n)} \prod_{s \in \sigma} p(s)}$$

CSP = *Contiguous Sequence Partition*

Ranking: hierarchical

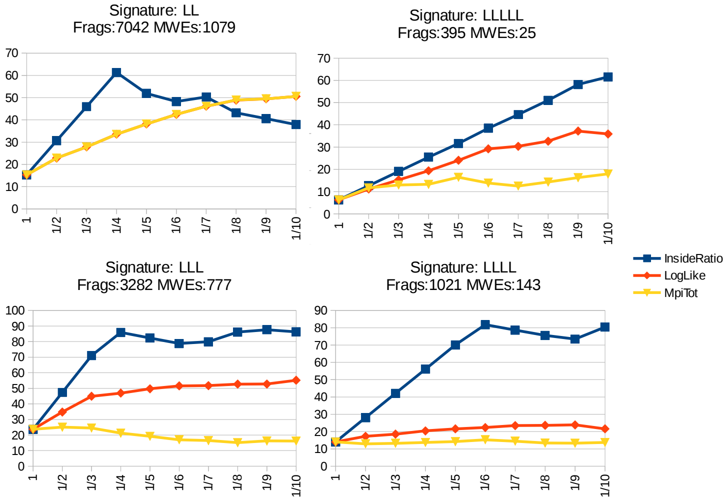
Definition

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.

- i.e., a `compositionality index'
- $$\mathrm{LIR}(S) = \log \frac{p(\mathrm{frag})}{inside(\mathrm{frag})}$$

Ranking results

FRENCH TREEBANK RESULTS



Signature: LL
Frags:7042 MWEs:1079

Signature: LLLLL
Frags:395 MWEs:25

Signature: LLL
Frags:3282 MWEs:777

Signature: LLLL
Frags:1021 MWEs:143

Ranking results

Treebank	PMI	LLR	LIR
French	33.0	32.3	45.8
Dutch	49.4	46.6	50.5

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

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

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