Multiword Expression Identification with Recurring Tree Fragments and Association Measures

Proceedings of the 11th Workshop on Multiword Expressions

WG3

Parseme 5th GM
Iași, 23-24 September 2015

Federico Sangati
federico.sangati@gmail.com

Andreas van Cranenburgh
andreas.van.cranenburgh@huygens.knaw.nl

FONDAZIONE BRUNO KESSLER

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

huygens ing

GOAL

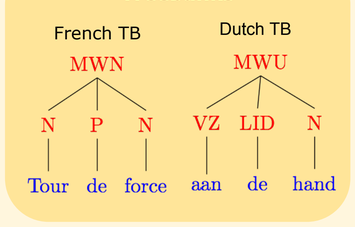
Investigate ways of **automatically discovering MWEs** in treebanks by searching for **recurring patterns**.

Treebank	Trees	Total Frags	Selected Frags
French	13K	274K	86K
Dutch	52K	536K	193K
English	500K	4.3M	2.8M

SUPERVISED

Input: Small MANUAL TREEBANKS with ANNOTATED MWE

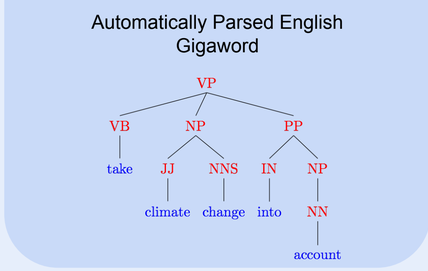
- Manually annotated MWEs.
- Only **contiguous** and **flat** MWEs



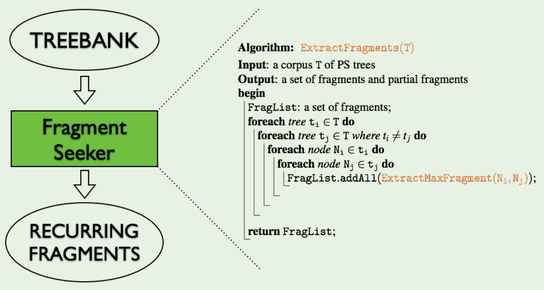
UNSUPERVISED

LARGE AUTOMATIC TREEBANK **without** ANNOTATED MWE

- Also **discontiguous** and **hierarchical** MWEs
- Not manually annotate MWEs



RECURRING FRAGMENTS SEEKER



Parsing

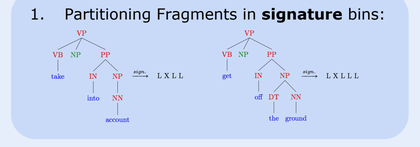
Using Disco-DOP: Tree Substitution Grammar
(implicit identification of MWEs as in Green et. al 2011)

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**

	#gold	DP-TSG	Stanford	This work
MWN	457	65.7	64.8	**68.9**
MWADV	220	**77.2**	75.0	70.0
MWP	162	79.5	81.2	**81.9**
MWC	47	85.8	**86.3**	80.7
MWV	26	56.2	**57.1**	55.9
MWPRO	17	75.3	72.2	**78.1**
MWD	15	65.1	**68.4**	66.7
MWA	8	36.0	26.1	**37.5**
Total	955	71.3	70.5	**71.9**

MWEs Discovery

1. Partitioning Fragments in **signature** bins:



2. **Ranking fragments** based on

a. **Association Measures**

Log-Likelihood Ratio

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

Pointwise Mutual Information

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

PMI	Freq.	Sequence Pattern
10.0	6	VB.take NP IN.into NN.account
13.0	6	VB.take DT NN.place IN.in NN.effect
13.6	7	VB.take DT NN.look TO.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 sequance pattern VB.take X L L, sorted by PMI.

PMI	Freq.	Sequence Pattern
13.3	13	VB.take IN.into NN.account
9.8	6	VB.take NN.advantage IN.for
9.7	5	VB.take NN.credit IN.for
9.3	12	VB.take DT.a NN.look
8.4	88	VB.take NN.advantage IN.of
8.4	7	VB.take NN.place IN.on
8.3	6	VB.take NN.effect IN.in
8.1	14	VB.take NNS.steps TO.to
..	..	..
4.6	6	VB.take DT.the NN.money

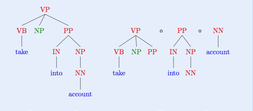
A sample of English fragments conforming to the sequance pattern VB.take L L, sorted by PMI.

b. **Syntax Compositionality Index**

New

$$\mathrm{SCI}(\mathrm{frag}) = \log \frac{p(\mathrm{frag})}{\mathit{inside}(\mathrm{frag})}$$

High when a fragment is often seen as a single block, low when it is typically generated by smaller units



Fragment generated as a single block

Fragment generated by smaller units

RESULTS

Computing single-figure **F1 evaluation** of the 3 metrics, obtained by aggregating the top 1/5 candidates of each bin. For this evaluation, recall and precision are computed, with the gold set consisting of all the extracted lexicalized fragments with MWE gold tags.

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

This **evaluation** is **not ideal**, as our method aims to go beyond the small, contiguous MWE strings annotated in the treebanks.

Manual inspection of the selected candidates reveals that many of them are MWEs, while not part of the gold standard.