

# Semantics Based MWE Extraction

Automatically acquiring multi-word expressions  
by exploiting non-compositionality

Tim Van de Cruys    Begoña Villada Moirón

A Broader Perspective on Multiword Expressions  
ACL 2007 Workshop  
June 28, 2007



# Outline

- 1 Introduction
- 2 Methodology
  - Overview
  - Measures
- 3 Results & Evaluation
  - Examples
  - Evaluation
- 4 Conclusion & Future Work



# Introduction

- Task: automatic extraction of multi-word expressions (MWEs) from large corpora
- Starting point: many MWEs are non-compositional, i.e. the meaning of the MWE is not the sum of the meaning of the individual words
- Intuition: a noun within a MWE cannot easily be replaced by a semantically similar noun
- → Use of semantic clusters to determine whether a MWE-candidate is compositional or not



# Intuition

break the ice  $\longleftrightarrow$  break the vase  
                  \**snow*                                  cup  
                  \**hail*                                  dish

- In the first expression, it is not possible to replace *ice* with semantically related nouns such as *snow* or *hail*;
- in the second expression, it is possible to replace *vase* with semantically related words such as *cup* or *dish*.



# Overview of the Method

- verb + prepositional complement instances are extracted from 500M word corpus (focus on MWE with PP)
- matrix of 5K verb-preposition combinations  $\times$  10K nouns is created
- 10K nouns are automatically clustered using distributional similarity measures
- number of statistical measures is applied to determine 'unique associations' given the cluster in which a noun appears



## Measures 1/3

- inspired by selectional preferences (Resnik 1993), entropy-based
- KL divergence between  $P(n)$  and  $P(n | v)$

$$S_v = \sum_n p(n | v) \log \frac{p(n | v)}{p(n)} \quad (1)$$



## Measures 2/3

- The preference of the verb for the noun  $\rightarrow [0, 1]$

$$A_{v \rightarrow n} = \frac{p(n | v) \log \frac{p(n|v)}{p(n)}}{S_v} \quad (2)$$

- Ratio of strength of verb preference for a particular noun, compared to other nouns in the cluster  $\rightarrow [0, 1]$

$$R_{v \rightarrow n} = \frac{A_{v \rightarrow n}}{\sum_{n' \in C} A_{v \rightarrow n'}} \quad (3)$$



## Measures 3/3

- The preference of the noun for the verb  $\rightarrow [0, 1]$

$$A_{n \rightarrow v} = \frac{p(v | n) \log \frac{p(v|n)}{p(v)}}{S_n} \quad (4)$$

- Ratio of strength of noun preference for a particular verb, compared to other nouns in the cluster  $\rightarrow [0, 1]$

$$R_{n \rightarrow v} = \frac{A_{n \rightarrow v}}{\sum_{n' \in C} A_{n' \rightarrow v}} \quad (5)$$





## An elaborated example 1/3

in de smaak vallen  $\longleftrightarrow$  in de put vallen  
\*geur kuil  
\*voorkeur krater  
\*stijl greppel

*in the taste* *fall* *in the well* *fall*  
'to be appreciated' 'to fall down the well'  
\*smell hole  
\*preference crater  
\*style trench



# An elaborated example 2/3

- **smaak**: idioom, karakter, persoonlijkheid, stijl, temperament, thematiek, uiterlijk, uitstraling, voorkomen

MWE candidate	(1)	(2)	(3)	(4)	MWE?
val#in smaak	.12	1.00	.07	1.0	yes
val#in karakter	.00	.00	.00	.00	no
val#in stijl	.00	.00	.00	.00	no



## An elaborated example 3/3

- **put**: gaatje, gat, kloof, krater, kuil, lek, scheur, valkuil

MWE candidate	(1)	(2)	(3)	(4)	MWE?
val#in put	.00	.04	.10	.05	no
val#in kuil	.00	.11	.10	.38	no
val#in kloof	.00	.01	.10	.04	no
val#in gat	.04	.72	.10	.25	both



## More examples

MWE candidate	(1)	(2)	(3)	(4)	MWE?
ben#op moment	0.24	0.88	0.01	0.42	no
ga#met pensioen	0.28	0.99	0.03	0.99	yes
ben#aan kant	0.32	0.930	0.01	0.61	no
schrijf#in brief	0.27	0.77	0.03	0.37	no
voldoe#aan eis	0.32	0.46	0.09	0.14	no
haal#uit kast	0.23	0.97	0.06	0.94	yes
neem_op#in ziekenhuis	0.35	0.91	0.03	0.39	no
stel#in vooruitzicht	0.24	1.00	0.23	1.00	yes
houd#op peil	0.24	0.97	0.13	0.95	yes
houd#op vlakte	0.24	0.99	0.21	0.99	yes
houd#aan regel	0.24	0.43	0.03	0.11	no



# Quantitative Evaluation

- Fully automated, compared to Referentie Bestand Nederlands
- Upper bound consists of all RBN MWE's *present in the data*

Parameters					Prec.	Rec.	F-Measure
(1)	(2)	(3)	(4)	n	(%)	(%)	(%)
.10	.80	–	–	2916	24.66	16.64	19.87
.10	.90	–	–	2425	26.72	14.99	19.55
.10	.80	–	.80	2014	28.95	13.49	18.40
.10	.90	–	.90	1699	30.14	11.85	17.01
.10	.80	.01	.80	1694	30.99	12.15	17.45
.20	.99	.05	.99	387	50.39	4.51	8.28
Random baseline				4332	1.02	1.02	1.02



# Qualitative Evaluation

- Algorithm is able to filter out expressions that cause problems with other MWE extraction algorithms, e.g.
  - *benoemen tot* { *minister, secretaris-generaal, ...* }
  - *voldoen aan* { *eisen, voorwaarden, ...* }
- Overall results seem better when only using (1) and (2)
- **But:** in many cases, (3) and (4) are able to filter out bad MWE's, e.g.
  - *verschijnen op toneel* ↔ *zingen op toneel*
  - *lig in ziekenhuis* ↔ *lig in geheugen*
  - *naar school willen, naar huis willen*



# Conclusion

- Non-compositionality based algorithm is able to rule out expression that are coined as MWE's by traditional algorithms
- Using measures (1) and (2) gives best results; using (3) and (4) increases precision but degrades recall.



## Future Work

- Compare the method of semantic uniqueness with other methods used to find MWE's:
  - salience/log-likelihood
  - Decision Tree Classifier
  - Maximum Entropy Model
- Perform an evaluation of the algorithm by using human judges





# Clustering Method

- frequency matrix of 10K nouns \* 100K syntactic relations is extracted (smoothed with mutual information)
- dimensionality reduction to 1800 dimensions with random indexing
- simple and computationally very efficient dimensionality reduction technique
- clustered into 1K clusters with k-means clustering



## Automatic Evaluation with Mutual Information

Parameters					Prec	Rec	F-Measure
(1)	(2)	(3)	(4)	N	(%)	(%)	(%)
.10	.80	–	–	2916	24.66	16.64	19.87
.10	.90	–	–	2425	26.72	14.99	19.55
.10	.80	–	.80	2014	28.95	13.49	18.40
.10	.90	–	.90	1699	30.14	11.85	17.01
.10	.80	.01	.80	1694	30.99	12.15	17.45
.20	.99	.05	.99	387	50.39	4.51	8.28
MI	3.0	.90		5406	12.80	16.01	14.23
MI	4.0	.80		3470	16.89	13.56	15.04
Random baseline				4332	1.02	1.02	1.02

