Semantics Based MWE Extraction

Automatically acquiring multi-word expressions by exploiting non-compositionality

Tim Van de Cruys Begoña Villada Moirón

CLIN 17
January 12, 2007

RuG

Introduction

- Task: Automatic extraction of multi-word expressions (MWE) from large corpora
- Starting point: many MWE's 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
- $\longrightarrow$ Use of semantic clusters to determine whether a MWE-candidate is compositional or not

RuG

Intuition

in	de	pen	klimmen	$\longleftrightarrow$	in	de	eik	klimmen
		*potlood					boom	
		*penseel					beuk	
		*viltstift					den	

- In the first expression (*'to climb in the pen'*, i.e. to write an article), it is not possible to replace *pen* with other kinds of writing instruments;
- in the second expression (*'to climb up the oak'*), it is possible to replace *eik* with other kinds of trees.

RuG

Overview of the Method

- verb + prepositional complement instances are extracted from Twente News Corpus (focus on MWE with PP)
- matrix of 5K verb-preposition combinations * 10K nouns is created
- 10K nouns are automatically clustered using distributional similarity measures and random indexing
- number of statistical measures is applied to determine 'unique associations' given the cluster in which a noun appears

RuG

Measures 1/2

- inspired by selectional preferences (Resnik 1993), entropy-based
- The preference of the verb for the noun $\rightarrow [0, 1]$

$$S_{v,n} = P(n \mid v)\ log \frac{P(n \mid v)}{P(n)} \quad (1)$$

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

$$R_{v,n,C} = \frac{S_{v,n}}{\sum_{n' \epsilon C} S_{v,n'}} \quad (2)$$

RuG

Measures 2/2

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

$$S_{n,v} = P(v \mid n)\ log \frac{P(v \mid n)}{P(v)} \tag{3}$$

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

$$R_{n,v,C} = \frac{S_{n,v}}{\sum_{n' \epsilon C} S_{n',v}} \tag{4}$$

RuG

An elaborated example 1/3

in	de	smaak		vallen	⟷	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	

RuG

An elaborated example 2/3

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

MWE candidate	$S_{v,n}$	$R_{v,n,C}$	$S_{n,v}$	$R_{n,v,C}$	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

RuG

An elaborated example 3/3

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

MWE candidate	$S_{v,n}$	$R_{v,n,C}$	$S_{n,v}$	$R_{n,v,C}$	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

RuG

More examples

MWE candidate	$S_{v,n}$	$R_{v,n,C}$	$S_{n,v}$	$R_{n,v,C}$	MWE?
ga#met pensioen	0.28	0.99	0.03	0.99	yes
haal#uit kast	0.23	0.97	0.06	0.94	yes
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
schrijf#in brief	0.27	0.77	0.03	0.37	no
voldoe#aan eis	0.32	0.46	0.09	0.14	no
houd#aan regel	0.24	0.43	0.03	0.11	no
ben#aan kant	0.32	0.93	0.01	0.61	no
ben#op moment	0.24	0.88	0.01	0.42	no
neem_op#in ziekenhuis	0.35	0.91	0.03	0.39	no

RuG

Quantitative Evaluation

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

Parameters					Prec.	Rec.	F-Measure
$S_{v,n}$	$R_{v,n,C}$	$S_{n,v}$	$R_{n,v,C}$	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

RuG

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 $S_{v,n}$ and $R_{v,n,C}$
- **But:** in many cases, $S_{n,v}$ and $R_{n,v,C}$ are able to filter out non MWE's, e.g.
  - *verschijnen op toneel* $\leftrightarrow$ *zingen op toneel*
  - *lig in geheugen* $\leftrightarrow$ *lig in ziekenhuis*
  - *naar school willen*, *naar huis willen*

RuG

Conclusion

- Non-compositionality based algorithm is able to rule out expression that are coined as MWE's by traditional algorithms
- Using measures $S_{v,n}$ and $R_{v,n,C}$ gives best results; using $S_{n,v}$ and $R_{n,v,C}$ increases precision but degrades recall.

RuG

Future Work

- Compare and combine the method of semantic uniqueness with other methods used to find MWE's:
  - salience/log-likelihood
  - Decision Tree Classifier
  - Maximum Entropy Model
- Perform a manual evaluation of the algorithm (human judges)

RuG

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

RuG

Automatic Evaluation with Mutual Information

Parameters				N	Prec (%)	Rec (%)	F-Measure (%)
$S_{v,n}$	$R_{v,n,C}$	$S_{n,v}$	$R_{n,v,C}$				
.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

RuG