A Machine Learning Approach to Multiword Expression Extraction

MWE 2008 Shared Task Evaluation

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Charles University, Prague

June 1, 2008

Outline

Introduction

MWE 2008 Shared Task:

- ranking multiword expression candidates
- best candidates to be concentrated on the top of the list

Evaluation on three data sets:

- German Adj-N collocations
- German PP-Verb collocations
- Czech PDT collocations

  (those provided with corpus frequency information)

System:

- based on machine learning combination of multiple association measures
- described in (Pecina, 2005), (Pecina and Schlesinger, 2006), (Pecina, 2008)

System overview

Association measures

- comprehensive inventory of 55 association measures
- implementation in Perl and R

Data

- a set of MWE candidates x^i, $i = 1 \ldots N$ split for training and testing
- each x^i described by the feature vector $\mathbf{x}^i = (x_1^i, \ldots, x_{55}^i)^T$ consisting of 55 association scores computed from joint and marginal frequencies
- each x^i provided with a label $y^i \in \{0, 1\}$ indicating true positives ($y = 1$) and true negatives ($y = 0$).

Combination

- statistical classification models (supervised machine learning)
- trained for $0 - 1$ classification but used to produce scores for ranking

Methods

- Linear Logistic Regression (GLM)
- Linear Discriminant Analysis (LDA),
- Neural Networks with 1 and 5 units in the hidden layer (NNet.1, NNet.5)

Evaluation scheme

Crossvalidation

- data randomly split into 7 folds of the same size
- each fold contained the same ration of TP/N
- models trained on 6 folds, tested on one fold - total of 7 runs
- each run produced a ranked list of MWE candidates from the test fold

Evaluation means

- Precision-Recall curves for each run (crossvalidation data fold)
- Average Precision (AP) for each run - expected value of precision for all possible values of recall (assuming uniform distribution of recall) (AUC)
- Mean Average Precision (MAP) for each crossvalidated experiment - - mean of average precision computed for each data fold.
- Significance testing by nonparametric paired Wilcoxon test.

Baseline scores

A. an expected MAP of a system ranking MWE candidates randomly (TP/N)
B. MAP of the best association measure used individually (no combination)

Experiment design

1. choose the evaluation data set
2. specify true positives (where applicable)
3. set the baseline score (TP/N)
4. split the data for crossvalidation (7 stratified folds of equal size)
5. compute association scores for all candidates in all folds and estimate their MAP
6. select the best individual AM and set the second baseline score
7. train and test the classification models (crossvalidation) and estimate their MAP
8. present the results

German Adj-N collocations: Data

Data

- a random sample of 1252 German collocation candidates selected from 8546 Adjective-Noun pairs occurring more then 20 times in FR corpus.

Annotation categories

1. true lexical collocations, other multiword expressions
2. customary and frequent combination, often part of collocational pattern
3. common expression, but no idiomatic properties
4. unclear / boundary cases
5. not collocational, free combinations
6. lemmatization errors corpus-specific combinations

Statistics

Category	*1*	*2*	*3*	*4*	*5*	*6*	*total*
Items	367	153	117	45	537	33	1252
%	29.3	12.2	9.3	3.6	42.9	2.6	100.0

- frequency information provided for 1 213 candidates

German Adj-N collocations: Data

Data

- a random sample of 1252 German collocation candidates selected from 8546 Adjective-Noun pairs occurring more then 20 times in FR corpus.

Annotation categories

1. true lexical collocations, other multiword expressions
2. customary and frequent combination, often part of collocational pattern
3. common expression, but no idiomatic properties
4. unclear / boundary cases
5. not collocational, free combinations
6. lemmatization errors corpus-specific combinations

Statistics

	TP		TN				
Category	*1*	*2*	*3*	*4*	*5*	*6*	*total*
Items	367	153	117	45	537	33	1252
%	29.3	12.2	9.3	3.6	42.9	2.6	100.0

- frequency information provided for 1 213 candidates

German Adj-N collocations: Data

Data

- a random sample of 1252 German collocation candidates selected from 8546 Adjective-Noun pairs occurring more then 20 times in FR corpus.

Annotation categories

1. true lexical collocations, other multiword expressions
2. customary and frequent combination, often part of collocational pattern
3. common expression, but no idiomatic properties
4. unclear / boundary cases
5. not collocational, free combinations
6. lemmatization errors corpus-specific combinations

Statistics

	TP			TN			
Category	*1*	*2*	*3*	*4*	*5*	*6*	*total*
Items	367	153	117	45	537	33	1252
%	29.3	12.2	9.3	3.6	42.9	2.6	100.0

- frequency information provided for 1 213 candidates

German Adj-N collocations: Results

	1–2	*1–2–3*
Baseline	42.12	51.78
Best AM	**62.88**	69.14
GLM	60.88	70.62
LDA	61.30	**70.77**
NNet.1	60.52	70.38
NNet.5	59.87	70.16

- Best AM: Piatersky-Shapiro coefficient $P(xy) - P(x*)P(*y)$
- TP *1-2*: no performance gain from combination methods
- TP *1-2-3*: improvement not significant
- possible explanation: small data
  (1213/7=173 candidates in one fold, 72 and 88 TPs, resp.)

German PP-Verb collocations

Data

- 21 796 German combinations of a prepositional phrase and a governing verb extracted from the FR corpus

Annotation categories

1. collocational: support-verb constructions (*FVG*)+figurative expressions (*figur*)
2. non-collocational

Statistics

	items	*%*
total	21796	100.0
TPs	1149	5.3
FVG	549	2.5
figur	600	2.8
in.fr30	5102	23.4
light.v	6892	31.6

- frequencies provided for 18 649 candidates (4 098 *in.fr30*, 6272 *light.v*)

German PP-Verb collocations: Support-verb constructions

	all	*in.fr30*	*light.v*
Baseline	2.91	5.75	7.25
Best AM	18.26	28.48	43.97
GLM	28.40	26.59	41.25
LDA	28.38	40.44	**45.08**
NNet.1	**30.77**	42.42	44.98
NNet.5	30.49	**43.40**	44.23

- Best AM (*all*, *in.fr30*): Confidence measure $\max[P(y|x), P(x|y)]$
- Best AM (*light.v*): Poisson significance m. $\frac{\hat{f}(xy) - f(xy)\log\hat{f}(xy) + \log f(xy)!}{\log N}$
- substantial improvement of MAP for *all* and *in.fr30*
- slight improvement for *light.v*

German PP-Verb collocations: Figurative expressions

	all	*in.fr30*	*light.v*
Baseline	3.16	5.70	4.56
Best AM	14.98	21.04	23.65
GLM	**19.22**	15.28	10.46
LDA	18.34	**23.32**	24.88
NNet.1	19.05	22.01	24.30
NNet.5	18.26	22.73	**25.86**

- Best AM (*all*): Confidence measure $\max[P(y|x), P(x|y)]$
- Best AM (*in.fr30*): Piatersky-Shapiro $P(xy) - P(x*)P(*y)$
- Best AM (*light.v*): t test $\frac{f(xy) - \hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
- moderate improvement in all subtasks

German PP-Verb collocations: Support-verb constructions and Figurative expr

	all	*in.fr30*	*light.v*
Baseline	6.07	11.45	11.81
Best AM	31.17	43.85	63.59
GLM	44.66	47.81	65.37
LDA	41.20	57.77	65.54
NNet.1	44.71	**60.59**	65.10
NNet.5	**44.77**	59.59	**66.06**

- Best AM (*all*, *in.fr30*): Confidence measure $\max[P(y|x), P(x|y)]$
- Best AM (*light.v*): Poisson significance m. $\frac{\hat{f}(xy) - f(xy)\log \hat{f}(xy) + \log f(xy)!}{\log N}$
- substantial improvement of MAP for *all* and *in.fr30*
- slight improvement for *light.v*

Czech PDT collocations: Data

Data

- 12 233 normalized dependency bigrams occurring in PDT more than five times, with part-of-speech patterns that can possibly form a collocation
- three parallel annotations

Annotation categories

0. non-collocations
1. stock phrases, frequent unpredictable usages
2. names of persons, organizations, geographical locations, and other entities
3. support verb constructions
4. technical terms
5. idiomatic expressions

Statistics

Category	*0*	*1-5*	*total*
Items	9661	2572	12233
%	79.99	21.01	100.0

Czech PDT collocations: Results

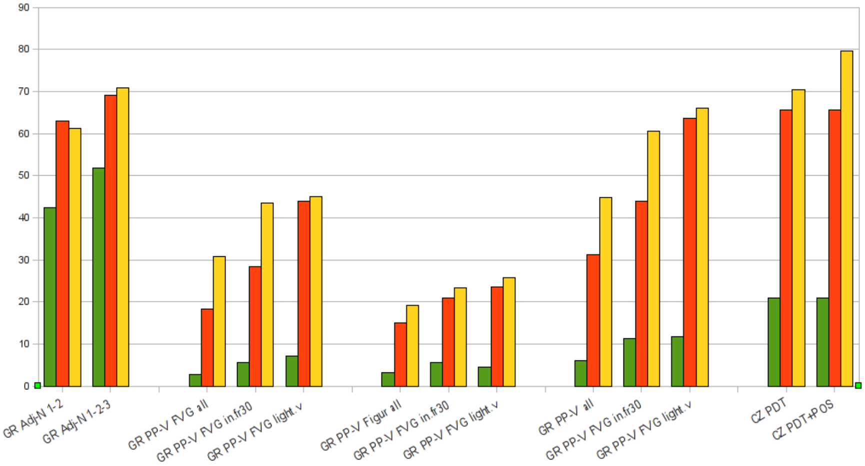
	AMs	*AMs+POS*
Baseline	21.01	
Best AM	65.63	
GLM	67.21	77.27
LDA	67.23	75.83
NNet.1	67.34	77.76
NNet.5	**70.31**	**79.51**

- Best AM: Unigram subtuples measure $\log \frac{ad}{bc} - 3.29\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
- considerable performance improvement by combination methods
- additional improvement after using POS pattern as an additional feature

Summary results

Data Set	*Var*	*Baseline*	*Best AM*	*Best CM*	*+%*
GR Adj-N	1-2	42.40	62.88	61.30	**-2.51**
	1-2-3	51.74	69.14	70.77	2.36
GR PP-V FVG	all	2.89	18.26	30.77	**68.51**
	in.fr30	5.71	28.48	43.40	52.39
	light.v	7.26	43.97	45.08	2.52
GR PP-V Figur	all	3.15	14.98	19.22	**28.30**
	in.fr30	5.71	21.04	23.32	10.84
	light.v	4.47	23.65	25.86	9.34
GR PP-V all	all	6.05	31.17	44.77	**43.63**
	in.fr30	11.43	43.85	60.59	**38.18**
	light.v	11.73	63.59	66.06	3.88
CZ PDT		21.01	65.63	70.31	7.13
	+POS	21.01	65.63	79.51	**21.15**

Summary graph



Conclusions

- MAP seems a reasonable evaluation metrics
- different association measures give different results for different tasks (data)
- it is not possible to recommend "the best general association measure"
- instead, let the machine learning methods do the job: to select the right measures and give them the right weights in the combination model
- many AMs in the models are redundant and should be removed so the models can be trained properly (Pecina, 2008)

Thank you!

Association Measures I

1.	**Joint probability**	$P(xy)$
2.	**Conditional probability**	$P(y\|x)$
3.	**Reverse conditional prob.**	$P(x\|y)$
4.	**Pointwise mutual inform.**	$\log \frac{P(xy)}{P(x*)P(*y)}$
5.	**Mutual dependency (MD)**	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
6.	**Log frequency biased MD**	$\log \frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$
7.	**Normalized expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}$
8.	**Mutual expectation**	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$
9.	**Salience**	$\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
10.	**Pearson's χ^2 test**	$\sum_{i,j} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
11.	**Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
12.	**t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
13.	**z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
14.	**Poison significance measure**	$\frac{\hat{f}(xy)-f(xy)\log\hat{f}(xy)+\log f(xy)!}{\log N}$
15.	**Log likelihood ratio**	$-2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$

Association Measures II

16.	**Squared log likelihood ratio**	$-2\sum_{ij} \frac{\log {f_{ij}}^2}{\hat{f}_{ij}}$
17.	**Russel-Rao**	$\frac{a}{a+b+c+d}$
18.	**Sokal-Michiner**	$\frac{a+d}{a+b+c+d}$
19.	**Rogers-Tanimoto**	$\frac{a+d}{a+2b+2c+d}$
20.	**Hamann**	$\frac{(a+d)-(b+c)}{a+b+c+d}$
21.	**Third Sokal-Sneath**	$\frac{b+c}{a+d}$
22.	**Jaccard**	$\frac{a}{a+b+c}$
23.	**First Kulczynsky**	$\frac{a}{b+c}$
24.	**Second Sokal-Sneath**	$\frac{a}{a+2(b+c)}$
25.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
26.	**Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$
27.	**Odds ratio**	$\frac{ad}{bc}$
28.	**Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
29.	**Yulle's Q**	$\frac{ad-bc}{ad+bc}$
30.	**Driver-Kroeber**	$\frac{a}{\sqrt{(a+b)(a+c)}}$

Association Measures III

31.	**Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
32.	**Pearson**	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
33.	**Baroni-Urbani**	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
34.	**Braun-Blanquet**	$\frac{a}{\max(a+b,a+c)}$
35.	**Simpson**	$\frac{a}{\min(a+b,a+c)}$
36.	**Michael**	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
37.	**Mountford**	$\frac{2a}{2bc+ab+ac}$
38.	**Fager**	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2}\max(b,c)$
39.	**Unigram subtuples**	$\log\frac{ad}{bc} - 3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
40.	**U cost**	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
41.	**S cost**	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
42.	**R cost**	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$
43.	**T combined cost**	$\sqrt{U\times S\times R}$
44.	**Phi**	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
45.	**Kappa**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$

Association Measures IV

46.	**J measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})}, P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
47.	**Gini index**	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2 +P(\bar{x}*)(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2, P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2 +P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
48.	**Confidence**	$\max[P(y\|x), P(x\|y)]$
49.	**Laplace**	$\max[\frac{NP(xy)+1}{NP(x*)+2}, \frac{NP(xy)+1}{NP(*y)+2}]$
50.	**Conviction**	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})}, \frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
51.	**Piatersky-Shapiro**	$P(xy)-P(x*)P(*y)$
52.	**Certainity factor**	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)}, \frac{P(x\|y)-P(x*)}{1-P(x*)}]$
53.	**Added value (AV)**	$\max[P(y\|x)-P(*y), P(x\|y)-P(x*)]$
54.	**Collective strength**	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)} \cdot \frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
55.	**Klosgen**	$\sqrt{P(xy)}\cdot AV$