A Machine Learning Approach to Multiword Expression Extraction

Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz

Abstract

This paper describes our participation in the MWE 2008 evaluation campaign focused on ranking MWE candidates. Our ranking system employed 55 association measures combined by standard statistical-classification methods modified to provide scores for ranking. Our results were crossvalidated and compared by Mean Average Precision. In most of the experiments we observed significant performance improvement achieved by methods combining multiple association measures.

1. Introduction

Four gold standard data sets were provided for the MWE 2008 shared task. The goal was to re-rank each list such that the "best" candidates are concentrated at the top of the list[1]. Our experiments were carried out only on three data sets – those provided with corpus frequency data by the shared task organizers: German Adj-N collocation candidates, German PP-Verb collocation candidates, and Czech dependency bigrams from the Prague Dependency Treebank. For each set of experiments we present the best performing association measure (AM) and results of our own system based on combination of multiple association measures (AMs).

2. System Overview

In our system which was already described in (Pecina and Schlesinger, 2006) and (Pecina, 2005), each collocation candidate x^i is described by the *feature vector* $\mathbf{x}^i = (x^i_1, \ldots, x^i_{55})^T$ consisting of 55 association scores from Table 1 computed from the corpus frequency data (provided by the shared task organizers) and assigned a label $y^i \in \{0,1\}$ which indicates whether the bigram is considered as true positive ($y = 1$) or not ($y = 0$). A part of the data is then used to train standard statistical-classification models to predict the labels. These methods are modified so they do not produce 0–1 classification but rather a score that can be used (similarly as for association measures) for ranking the collocation candidates (Pecina and Schlesinger, 2006). The following statistical-classification methods were used in experiments described in this paper: Linear Logistic Regression (GLM), Linear Discriminant Analysis (LDA), Neural Networks with 1 and 5 units in the hidden layer (NNet.1, NNet.5).

For evaluation we followed a similar procedure as in our previous work (Pecina and Schlesinger, 2006). Before each set of experiments every data set was split into seven stratified folds each containing the same ratio of true positives. Average precision (corresponding to the area under the precision-recall curve) was estimated for each data fold and its mean was used as the main evaluation measure (Mean Average Precision - MAP). The methods combining multiple association measures used 6 data folds for training and one for testing (7-fold crossvalidation).

[1] http://multiword.sf.net/mwe2008/

3. German Adj-N Collocation Candidates

3.1. Data Description

This data set consits of 1 252 German collocation candidates randomly sampled from the 8 546 different adjective-noun pairs (attributive prenominal adjectives only) occurring at least 20 times in the Frankfurter Rundschau corpus (FR, 1994). The collocation candidates were lemmatized with the IMSLex morphology (Lezius et al., 2000), preprocessed with the partial parser YAC (Kermes, 2003) for data extraction, and annotated by professional lexicographers with the following 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

3.2. Experiments and Results

Frequency counts were provided for 1 213 collocation candidates from this data set. We performed two sets of experiments on them. First, only the categories 1–2 were considered true positives. There was a total of 511 such cases and thus the baseline precision was quite high (42.12%). The highest MAP of 62.88% achieved by *Piatersky–Shapiro coefficient* (51) was not outperformed by any of the combination method.

In the second set of experiments, the true positives comprised categories 1–2–3 (total of 628 items). The baseline precision was as high as 51.78%. The best association measure was again *Piatersky–Shapiro coefficient* (51) but it was slightly outperformed by most of the combination methods. The best one was based on LDA and achieved MAP of 70.77%. See detailed results in Table 2.

	1–2		*1–2–3*	
Baseline	42.12		51.78	
Best AM	**62.88**	(51)	69.14	(51)
GLM	60.88		70.62	
LDA	61.30		**70.77**	
NNet.1	60.52		70.38	
NNet.5	59.87		70.16	

Table 2: MAP results of ranking German Adj-N collocation candidates

#	Name	Formula
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***	$\log \frac{P(xy)^2}{P(x*)P(*y)} \quad \log P(xy)$
7.	**Normalized expectation**	$\frac{2f(xy)}{f(x*) \quad f(*y)}$
8.	**Mutual expectation**	$\frac{2f(xy)}{f(x*) \quad f(*y)} \cdot P(xy)$
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.	**Poisson significance measure**	$\frac{\hat{f}(xy)-f(xy)\log \hat{f}(xy) \quad \log f(xy)!}{\log N}$
15.	**Log likelihood ratio**	$-2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$
16.	**Squared log likelihood ratio**	$-2\sum_{i,j} \frac{\log f_{ij}^2}{\hat{f}_{ij}}$
17.	**Russel-Rao**	$\frac{a}{a \quad b \quad c \quad d}$
18.	**Sokal-Michiner**	$\frac{a \quad d}{a \quad b \quad c \quad d}$
19.	**Rogers-Tanimoto**	$\frac{a \quad d}{a \quad 2b \quad 2c \quad d}$
20.	**Hamann**	$\frac{(a \quad d)-(b \quad c)}{a \quad b \quad c \quad d}$
21.	**Third Sokal-Sneath**	$\frac{b \quad c}{a \quad d}$
22.	**Jaccard**	$\frac{a}{a \quad b \quad c}$
23.	**First Kulczynsky**	$\frac{a}{b \quad c}$
24.	**Second Sokal-Sneath**	$\frac{a}{a \quad 2(b \quad c)}$
25.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a \quad b} \quad \frac{a}{a \quad c})$
26.	**Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a \quad b} \quad \frac{a}{a \quad c} \quad \frac{d}{d \quad b} \quad \frac{d}{d \quad c})$
27.	**Odds ratio**	$\frac{ad}{bc}$
28.	**Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad} \quad \sqrt{bc}}$
29.	**Yulle's Q**	$\frac{ad-bc}{ad \quad bc}$
30.	**Driver-Kroeber**	$\frac{a}{\sqrt{(a \quad b)(a \quad c)}}$
31.	**Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a \quad b)(a \quad c)(d \quad b)(d \quad c)}}$
32.	**Pearson**	$\frac{ad-bc}{\sqrt{(a \quad b)(a \quad c)(d \quad b)(d \quad c)}}$
33.	**Baroni-Urbani**	$\frac{a \quad \sqrt{ad}}{a \quad b \quad c \quad \sqrt{ad}}$
34.	**Braun-Blanquet**	$\frac{a}{\max(a \quad b, a \quad c)}$
35.	**Simpson**	$\frac{a}{\min(a \quad b, a \quad c)}$
36.	**Michael**	$\frac{4(ad-bc)}{(a \quad d)^2 \ (\ b \quad c)^2}$
37.	**Mountford**	$\frac{2a}{2bc \quad ab \quad ac}$
38.	**Fager**	$\frac{a}{\sqrt{(a \quad b)(a \quad c)}} - \frac{1}{2}\max(b,c)$
39.	**Unigram subtuples**	$\log \frac{ad}{bc} - 3.29\sqrt{\frac{1}{a} \quad \frac{1}{b} \quad \frac{1}{c} \quad \frac{1}{d}}$
40.	***U* cost**	$\log(1 \quad \frac{\min(b,c)}{\max(b,c)} \quad \frac{a}{a})$
41.	***S* cost**	$\log(1 \quad \frac{\min(b,c)}{a} \quad 1)^{-\frac{1}{2}}$
42.	***R* cost**	$\log(1 \quad \frac{a}{a \quad b}) \cdot \log(1 \quad \frac{a}{a \quad c})$
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**	$\frac{P(xy) \quad P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
46.	***J* measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)} \quad P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)} \quad P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
47.	**Gini index**	$\max[P(x*)(P(y\|x)^2 \quad P(\bar{y}\|x)^2) - P(*y)^2$ $P(\bar{x}*)(P(y\|\bar{x})^2 \quad P(\bar{y}\|\bar{x})^2) - P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2 \quad P(\bar{x}\|y)^2) - P(x*)^2$ $P(*\bar{y})(P(x\|\bar{y})^2 \quad P(\bar{x}\|\bar{y})^2) - P(\bar{x}*)^2]$
48.	**Confidence**	$\max[P(y\|x), P(x\|y)]$
49.	**Laplace**	$\max[\frac{NP(xy) \quad 1}{NP(x*) \quad 2}, \frac{NP(xy) \quad 1}{NP(*y) \quad 2}]$
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**	$\frac{P(xy) \quad P(\bar{x}\bar{y})}{P(x*)P(y) \quad P(\bar{x}*)P(*y)} \cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
55.	**Klosgen**	$\sqrt{P(xy)} \cdot AV$

$a=f(xy)$	$b=f(x\bar{y})$	$f(x*)$
$c=f(\bar{x}y)$	$d=f(\bar{x}\bar{y})$	$f(\bar{x}*)$
$f(*y)$	$f(*\bar{y})$	N

A contingency table contains observed frequencies and marginal frequencies for a bigram xy; $\bar{w}$ stands for any word except w; $*$ stands for any word; N is a total number of bigrams. The table cells are sometimes referred to as f_{ij}. Statistical tests of independence work with contingency tables of expected frequencies $\hat{f}(xy) = f(x*)f(*y)/N$.

Table 1: Lexical association measures used for ranking MWE candidates.

4. German PP-Verb Collocation Candidates

4.1. Data Description

This data set comprises 21 796 German combinations of a prepositional phrase (PP) and a governing verb extracted from the Frankfurter Rundschau corpus (FR, 1994) and used in a number of experiments, e.g. (Krenn, 2000). PPs are represented by combination of a preposition and a nominal head. Both the nominal head and the verb were lemmatized using the IMSLex morphology (Lezius et al., 2000) and processed by the partial parser YAC (Kermes, 2003). See (Evert, 2004) for details of the extraction procedure. The data were manually annotated as lexical collocations or non-collocational by Brigitte Krenn (Krenn, 2000). In addition, distinction was made between two subtypes of lexical collocations: support-verb constructions (FVG), and figurative expressions (Figur).

4.2. Experiments and Results

On this data we carried out several series of experiments. First, we focused on the support-verb constructions and figurative expressions separately, then we attempted to extract all of them without making this distinction. Frequency data were provided for a total of 18 649 collocation candidates. The main experiments were performed on all of them. Further, as suggested by the shared task organizers, we restricted ourselves to a subset of 4 908 candidate pairs that occur at least 30 times in the Frankfurter Rundschau corpus (*in.fr30*). Similarly, additional experiments were restricted to candidate pairs containing one of 16 typical *light verbs*. This was motivated by assumption that filtering based on this condition should significantly improve the performance of association measures. After applying this filter the resulting set contained 6 272 collocation candidates.

Support-Verb Constructions

The baseline precision for ranking only the support-verb constructions in all the data is as low as 2.91%, the best MAP (18.26%) was achieved by *Confidence* measure. Additional substantial improvement was achieved by all combination methods. The best score (30.77%) was obtained by Neural Network (1 unit). When focused on the candidates occurring at least 30 times (baseline precision 5.75%), the best individual association measure appeared to be again *Confidence* measure with MAP 28.48%. The best combination method was then Neural Network with 5 units: MAP 43.40%. The best performing individual association measure on light verb data was *Poisson significance measure* (14) with MAP as high as 43.97% (baseline 7.25%). The performance gain achieved by the best combination method was not, however, so significant (45.08%, LDA). Details are shown in Table 3.

	all		*in.fr30*		*light.v*	
Baseline	2.91		5.75		7.25	
Best AM	18.26	(48)	28.48	(48)	43.97	(14)
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	

Table 3: MAP results of ranking German PP-Verb support-verb construction candidates.

Figurative Expressions

Ranking figurative expressions seems more difficult. The best individual association measure on all data is again *Confidence* measure with MAP of only 14.98%, although the baseline precision is a little bit higher then in the case of support-verb constructions (3.16%). The best combination of multiple AMs is obtained by Logistic Regression (GLM) with MAP equal to 19.22%. Results for the candidates occurring at least 30 times (baseline precision 5.70%) are higher: the best AM (*Piatersky-Shapiro coefficient*) with MAP 21.04% and LDA with MAP 23.32%. In case of PP combinations with light verbs, the winning individual AM is *t test* (12) with MAP of 23.65% and the best combination method is Neural Network (5 units) with 25.86%. Details are depicted in Table 4.

	all		*in.fr30*		*light.v*	
Baseline	3.16		5.70		4.56	
Best AM	14.98	(48)	21.04	(51)	23.65	(12)
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**	

Table 4: MAP results of ranking German PP-Verb figurative expression candidates.

Support-Verb Constructions and Figurative Expressions

The last set of experiments performed on the German PP-Verb data aimed at ranking both support-verb constructions and figurative expressions without making any distinction between these two types of collocations. The results are shown in Table 5 and are not very surprising. The best individual AM on all the candidates as well as on the subset of the frequent ones was *Piatersky-Shapiro coefficient* with MAP 31.17% and 43.85%, respectively. *Poisson significance measure* (14) performed best on the candidates containing light verbs (63.59%). The best combination method were Neural Networks with 1 or 5 units. The most substantial performance improvement obtained by combining multiple AMs was observed on the set of all candidates (no filtering applied).

	all		*in.fr30*		*light.v*	
Baseline	6.07		11.45		11.81	
Best AM	31.17	(48)	43.85	(48)	63.59	(14)
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**	

Table 5: MAP results of ranking German PP-Verb candidates of both support-verb constructions and figurative expressions.

5. Czech PDT Bigrams

5.1. Data Description

The PDT data consist of notated list of 12 233 normalized dependency bigrams occurring in the manually annotated Prague Dependency Treebank (2.0, 2006) more than five times and having part-of-speech patterns that can possibly

form a collocation. Every bigram is assigned to one of the six categories described below by three annotators. Only the bigrams that all annotators agreed to be collocations (of any type, categories 1–5) are considered true positives. The entire set contains 2 572 such items. See (Pecina and Schlesinger, 2006) for details.

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

5.2. Experiments and Results

The baseline precision on this data is 21.02%. In our experiments, the best performing individual association measure was *Unigram subtuple measure* (39) with MAP of 65.63%. The best method combining all AMs was Neural Network (5 units) with MAP equal to 70.31%. After introducing a new (categorial) variable indicating POS patterns of the collocation candidates and adding it to the combination methods, the performance increased up to 79.51% (in case of the best method – Neural Network with 5 units) .

	AMs	*AMs+POS*
Baseline	21.01	
Best AM	65.63 (39)	
GLM	67.21	77.27
LDA	67.23	75.83
NNet.1	67.34	77.76
NNet.5	**70.31**	**79.51**

Table 6: MAP results of ranking Czech PDT collocation candidates. The second column refers to experiments using combination of association measures and information about POS patterns.

6. Conclusions

The overview of the best results achieved by individual AMs and by combination methods on all the data sets (and their variants) is shown in Table 7. With only one exception the combination methods significantly improved ranking of collocation candidates on all data sets. Our results showed that different measures give different results for different tasks (data). It is not possible to recommend "the best general association measure" for ranking collocation candidates. Instead, we suggest to use the proposed machine learning approach and let the classification methods do the job. Although it seems that Neural Network is probably the most suitable method for this task, we treat all the combination methods as equally good. We only recommend to use models that are fitted properly. Further, we also suggest to reduce the number of AMs employed in the combination methods by removing those that are redundant or do not help the prediction (see Pecina and Schlesinger (2006) for details.

Acknowledgments

This work has been supported by the Ministry of Education of the Czech Republic, projects MSM 0021620838.

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	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 Bigram		21.01	65.63	70.31	7.13
	+POS	21.01	65.63	79.51	21.15

Table 7: Summary of the results obtained on all data sets and their variants. The last two columns refer to the best method combining multiple association measures and the corresponding relative improvement compared to the best individual association measure. The last row refers to the experiment using combination of association measures and information about POS patterns.

7. References

PDT 2.0. 2006. http://ufal.mff.cuni.cz/pdt2.0/.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Univ. of Stuttgart.

1994. The FR corpus is part of the ECI Multilingual Corpus I distributed by ELSNET. See http://www.elsnet.org/eci.html for more information and licensing conditions.

Hannah Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. Ph.D. thesis, Saarland University.

Wolfgang Lezius, Stefanie Dipper, and Arne Fitschen. 2000. IMSLex - representing morphological and syntactical information in a relational database. In *U. Heid, S. Evert, E. Lehmann, and C. Rohrer (eds.), Proceedings of the 9th EURALEX International Congress*, Stuttgart, Germany.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, Australia.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL 2005 Student Research Workshop*, Ann Arbor, USA.