Extraction of the Multiword Lexical Units in the Perspective of the Wordnet Expansion

Maciej Piasecki Wrocław Univ. of Technology maciej.piasecki@pwr.edu.pl

Michał Wendelberger Wrocław Univ. of Technology michal.wendel@gmail.com

Marek Maziarz Wrocław Univ. of Technology mawroc@gmail.com

Abstract

The paper focuses on selecting an optimal set of the Multiword Expressions Extraction methods used as a tool during wordnet expansion. Wordnet multiword lexical units are a broad class and it is difficult to find a single extraction method fulfilling the task. Many extraction association measures were tested on very large corpora and a very large wordnet, namely plWordNet. Several new measures are proposed and compared with selected methods in the literature. Two ways of combining measures into ensembles were analysed too. We showed that method selection and the tuning of their parameters can be transferred between two large corpora. The comparison of the extracted collocations with the huge set of plWordNet multiword lexical units revealed that the performance of the methods is much below the optimistic levels reported in the literature. However, the carefully selected set and combination of the methods can be a valuable tool for lexicographers.

1 Introduction

A large number of different methods for the extraction Multiword Expressions (henceforth, MWE) have been proposed in literature. Most of them are focused on particular properties of MWEs, e.g. non-compositionality. Before applying selected methods as the support for the construction of a large lexicon we have to answer several questions.

- What method should we select if we need to extract MWEs of different subtypes?
- How effective are different methods in helping lexicographers who use a complex but well specified definition of Multiword Lexical Units (MLUs)[1] (Maziarz et al., 2015a)?
- What is the performance of the known extraction methods when they are applied to big corpora (e.g. >1 billion words) and next evaluated against very large lexicons of MWEs?

We aim at the development of a method for the extraction of MLUs from large corpora for the needs of wordnet expansion. MLUs encompass a broad spectrum of MWEs: from non-compositional MWEs to specialist terminology. The starting point for this work were the seminal papers of Pečina, e.g. (Pečina, 2010), including tests of a very large number of MWE extraction methods. However, those tests were done on relatively limited data set. The corpus used by Pečina in his experiments consisted of about 1.5 million words. In our work we utilised different corpora including the testing *Merged Corpus* of 1.6 billion words covering rich variety of topics and genres[2]. So, it was more than 1 000 times bigger than that used in (Pečina, 2010).

We wanted to perform large scale evaluation done on big corpora, utilising a very large lexicon of MWEs and focused on Polish, a language which is significantly different from English. The first experiments (completed) inspired us also to the development of a couple of additional association measures focused on selected MWE subtypes and meant to enrich the variety of MWE types covered by the combined measure.

[1] MLUs are, shortly speaking, MWEs that are elements of a lexical system, see Sec. 2

[2] The Merged Corpus combines Polish Wikipedia (`http://pl.wikipedia.org/`) the version from 28th Apr. 2012 and the corpus of electronic edition of the Rzeczpospolita newspaper (Rze, 2008). It was completed with texts collected from the internet. All texts from the internet were filtered: only larger texts with a relatively small number of words not recognised by the morphological analyser Morfeusz (Woliński, 2006) have been included in the corpus.

Proceedings of Recent Advances in Natural Language Processing, pages 512–520,
Hissar, Bulgaria, Sep 7–9 2015.

2 Background

plWordNet is a wordnet of Polish. Every wordnet is a lexico-semantic network describing lexical meanings in terms of lexico-semantic relations (Fellbaum, 1998). There are about 40 types of relations with more than 90 subtypes in total in plWordnet (Maziarz et al., 2013). After the series of projects starting 2005, plWordNet has now become the largest wordnet worldwide. The version 2.3 published in the year 2015 includes more than 171 000 lemmas, 244 000 lexical units (LUs)[3] and 184 000 synsets[4]. It has achieved very comprehensive coverage of Polish LUs that is comparable to the largest Polish dictionaries.

Because plWordNet 2.1 contained mainly one-word lemmas[5] contrary to most dictionaries, we have decided to add also many MLUs to plWordNet 3.0. We have estimated that the future plWordNet 3.0 should be expanded with about 60 000 multi-word LUs in comparison to 2.1.

plWordNet has been developed on the basis of the corpus-based semi-automatic method with all editing decisions having been made by linguists. In order to follow this development model, word combinations that seem to be good candidates for MLUs should be extracted from the large corpora, verified by lexicographers and added to plWordNet in this semi-automated way. In this paper we concentrate on the first phase: extraction MLU candidates from large corpora in a way facilitating their manual verification.

The crucial point in the evaluation procedure of extraction algorithms (Sec. 6) was the utilization of plWordNet as a gold-standard. We sought for such algorithms which gave us good precision in recognising MLUs from plWordNet. It must be emphasised that in previous versions of plWordNet MLUs were added on the basis of linguists' intuition of what is and what is not lexical. This intuition was supported by lexicographic resources, mainly general, phraseological and specialist dictionaries, lexicons and encyclopaedias.

The newest version of plWordNet now contains more than 30k MLUs added with the usage of detailed guidelines. The procedure of assessing lexicality of a given MLU candidate is presented in (Maziarz et al., 2015b) and (Maziarz et al., 2015a) in this volume. Summarising, it is based on a decision tree guiding linguists. Every extracted collocation is analysed in a sequence of tests before it is rejected or accepted as a plWordNet MLU. The application of with the guidelines tree improved consistency of lexicographers' decisions.

In order to check trustworhtiness of plWordNet as a gold-standard lexical resource we asked 5 linguists to intuitively assess lexicality of 200 MLUs randomly taken from plWordNet 2.1. They were given a definition pointing to the notion of LU being the part of our mental lexicon[6] and non-reproducibility of a word combination (whether it is set or free). Having averaged their answers we found that the confidence interval for the proportion of genuine MLUs is 90-98% (α=0.05). For instance, linguists rejected such word combinations as *koszt zakupu* 'the cost of buying something' or *kolor włosów* 'hair coloring', while accepted *płaca minimalna* 'minimum wage' or *ośrodek zdrowia* 'health centre'. [7] Thus, finally, we obtained a good argument for basing the estimation procedure on plWordNet.

3 Starting Point: Association Measures for MWEs

MWE elements occur together in text more frequently than it would be caused by chance. This idea has been expressed in more than hundred association measures based on statistical association measures, information theory or just heuristics, e.g. cf a rich overview in (Pečina, 2010). It would be difficult to repeat such an overview in a short paper, so our starting point were the results reported in (Pečina, 2010) and the set of the best performing measures according to those tests, e.g. Unigram Subtuples, Frequency Biased Mutual Dependency, Mutual Expectation or Pearsońs Chi$\hat{2}$, see the complete list in Sec. 5.2. Next, we extended this basic set with several more measures reported in the literature as having good performance: e.g. Contonni T1 (Paradowski, 2015) or Specific Exponential Correlation (Buczyński, 2004), see Sec. 5.2.

[3] A lexical unit is understood here technically as a triple: a lemma plus sense number plus a part of speech; MLUs have multi-word lemmas.

[4] Synsets are traditionally sets of near synonyms (Fellbaum, 1998), in plWordNet they group lexical units sharing the same lexico-semantic relations (Maziarz et al., 2013).

[5] In version 2.1 of plWordNet 1/5 of all LUs were MLUs.

[6] «The basic prerequisite for according lemma status to a multi-word items is that it has undergone some kind of lexicalisation, i.e., that it has been stored in our mental lexicon as a unit.» (Svensén, 2009, pp. 102-3).

[7] (Maziarz et al., 2015a) provide arguments for taking averaged decision of 5 linguists as a fair sign of lexicalicity.

On the basis of the first experiments and analysis of the measure similarity in (Paradowski, 2015), we formulated our own unique measures: W Specific Correlation, W Order, W Term Frequency Order and W Specific Exponential Correlation that are presented in Sec. 4. The last two measures have parameters and were tested for their different values.

We have also adopted from (Pečina, 2010) the method of combining by means of Machine Learning many association measures into one complex of better performance. Every MWE candidate is described by a vector of the measure values. Candidates that are know to be MWE define positive examples, the rest of candidates is used as negative examples. Several learning methods were used in (Pečina, 2010), namely: Linear Logistic Regression, Linear discriminant analysis, SVM (Support Vector Machines) and Multi-layer Perceptron (a neural network). A complex measure based on the Multi-layer Perceptron expressed the best performance, but the other complex measures were on the similar level.

Pečina tried to combine almost all single measures. However, (Paradowski, 2015) showed that many of them are correlated and even can be obtained from the same basic equation by changing its parameters. Such correlated measures are redundant attributes from the Machine Learning point of view and should not be used together.

Our approach differs significantly from the previous ones by the scale of the evaluation tests in terms of the size of: corpora used for the extraction and the MWE lexicon used for the comparison. Concerning the former we used the Merged Corpus of Polish described in Sec. 1, concerning the latter we used MLUs for plWordNet 2.2 as the gold set that includes almost 50 000 MWEs.

In (Pečina, 2010) the evaluation was performed on the basis of the *Prague Dependency Treebank* 2.0 that consists of 1 504 847 words from which 635 952 different word bi-grams were extracted. After Part of Speech based filtering 26 450 bi-grams were left. Next, all bi-grams occurring less than 6 times were removed and the bi-gram set was reduced to only 12 232. Those MWE candidates were evaluated manually by linguists according to the 5 MWE categories defined. The same definitions were used by all evaluators, but the inter-annotator agreement was moderate, cf. (Pečina, 2010). 2 557 bi-grams, i.e. 20.9% of the evaluated set, were found to be MWEs.

Summing up, hundreds of association measures were proposed in the literature, cf. (Pečina, 2010). On the basis of the evaluation results presented in the literature, especially for Polish data (Buczyński, 2004), and the possible generalisation of some measure to one equation with parameters (Paradowski, 2015), this set can be reduced to a much smaller number of the most promising ones. As a baseline we used the raw frequency of lemma bi-grams assuming that the more frequent bi-grams are more likely to be MWEs.

4 Extension: Additional Measures and the Complex Measure

4.1 W Specific Exponential Correlation

Pointwise Mutual Information, shortly mentioned in Eq. 1, is often used and expresses relatively good performance. In Eq. 1, x and y are words, $p(x)$, $p(y)$ and $p(x, y)$ are Maximum Likelihood Estimations of the probabilities, respectively, of single (marginal) and joint occurrences. However, PMI is known to overestimate the importance of infrequent events.

$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

PMI was modified in different ways to cope with this problem, e.g. one possibility is to refer to the 'full' Mutual Information in which the logarithm is multiplied by $p(x, y)$ probability. Applying this analogy to PMI, we obtain W Specific Correlation in Eq. 2 proposed in (Hoang et al., 2009a).

$$W\_SC(x, y) = p(x, y) log_2 \frac{p(x, y)}{p(x)p(y)} \quad (2)$$

Mutual Dependency in Eq. 3 is another modification of PMI in which the x and y joint frequency is emphasised inside the logarithm:

$$MD(x, y) = log_2 \frac{p(x, y)^2}{p(x)p(y)} \quad (3)$$

Buczyński (Buczyński, 2004) increased the power of the nominator to 3 and called this measure *Frequency Biased Mutual Dependency.* It produced good results in two evaluations on large Polish corpora (Buczyński, 2004) and (Broda et al., 2008). Later, Buczyński generalised his measure exchanging "3" to "$2 + \alpha$".

Finally, inspired by (Petrovica et al., 2010), we combined all the above modifications in a measure that is called *W Specific Exponential Correlation* and presented in Eq. 4.

$$W\_SEC(x,y) = p(x,y)log_2\frac{p(x,y)^e}{p(x)p(y)} \quad (4)$$

In *W_SEC* the frequency of the pair increases both components of the measure: one inside the logarithm and the one outside of it. The *W_SEC* behaviour is controlled by the parameter e and can be adapted to the task to some extent.

Following (Petrović et al., 2010) and (Van de Cruys, 2011, p. 2), *W_SEC* can be also easily modified for the extraction of candidates with n constituents, see Eq. 5.

4.2 W Order

Several criteria can be used in the MWE recognition. One of them is how the word order of the candidate is fixed. The more constrained possible linear word orders of the MWE candidate constituents are, the more likely it is an MWE. In order to test this, we need to calculate the number of possible word orders for the given candidate. This assumption is the basis for the *W Order* measure proposed in Eq. 6 where t is a sequence of the candidate constituents (words), S is a set of all possible orders of the same constituents, $S(t)_i$ – ith tuple from the set $S(t)$ and $f(\ldots)$ is the frequency.

$$W\_Ord(t) = \frac{1}{\prod_{i=1}^{n}(1+\frac{f(S(t)_i))}{max(f(S(t)))+1})} \quad (6)$$

In *W Order* the components are multiplied and the result of this multiplication is the biggest if all of them are equal. The smallest result is 0 if one of them is 0 – in the extreme situation only one is non-zero and equals the whole sum. Thus, the larger the multiplication result is, the more distributed the word orders are. It means that we need to reverse the fraction in order to get the needed behaviour of the measure.

W Order abstracts from the interpretation of the word order, i.e. it does not give any preference to any word order. The measure tests the number of the orders and their relative frequencies. The value of the measure does not depend on the exact frequencies of the candidate tuples, but it tests their mutual ratio.

By adding 1 to every frequency value in the denominators in Eq. 6, we wanted to secure against possible zero values, e.g. caused by one zero-frequency tuple. Secondly, we avoid assigning the same value of the measure and the position in the ranking to those candidates that have at least one zero frequency tuple. As a result, candidates with the greater number of zero frequency tuples obtain higher values of the measure, which promotes more fixed word order candidates. Thirdly, adding 1 causes that the amount of statistical information collected about a given candidate is taken into account in the measure. If the product of the tuple frequencies for different order variants of a given candidate is equal, a candidate with more statistical information, i.e. such that occurred more times, will be promoted. Finally, adding 1 modifies the range of possible values of the measure, eliminates problems with dividing and practically increases the number of possible values produced by the measure.

W Order does not require generalisation, as it is defined already for n-element tuples.

4.3 W Term Frequency Order

The frequency of a candidate is a very simple feature that is not sufficient on its own. However, it is correlated with the acceptance of candidates as MWE. Thus, we proposed also a *W Order* version, Eq. 7, in which the raw candidate frequency increases the measure value. It is already defined for n-element candidates.

$$W\_TFO(t) = f(t)W\_Ord(t) \quad (7)$$

4.4 Combined Measure

Complex measures are built as follows:

1. for each measure a ranking of the candidates is created;
2. each ranking is multiplied by the weights assigned to the measures;
3. weighted rankings are combined into the resulting ranking of the complex measures.

Weights for the individual measures were optimised by the genetic algorithm using a system described in (Kłyk et al., 2012). Genotypes consisted of measure weights. The precision of the extracted candidates in comparison to plWordNet MWEs

$$W\_SEC(x_1, \ldots, x_n) = p(x_1, x_2, ..., x_n) log_2 \frac{p(x_1, x_2, ..., x_n)^e}{\prod_{i=1}^{n} p(x_i)} \quad (5)$$

was used as the fitness function value. By mapping all candidate extraction results on the rankings we remove different ranges of different measures. The linear combination of the rankings has clear interpretation. The applied genetic algorithm is very flexible and does not need any assumptions concerning the combined measures. The algorithm was run on the tuning corpus only. For the test corpus, we used weights optimised on the test corpus. Henceforth, the complex measure will be called *VAM* (*Vector Attribute Measure*).

5 Experimental Setting

5.1 Data Sets

We used two corpora: the first one for tuning the measure parameters and the second for testing. The tuning corpus was also utilised for training different versions of the complex VAM.

As a tuning corpus we used *The Corpus of IPI PAN of Polish* (IPIC) (Przepiórkowski, 2004) – the first large corpus of Polish, still the only bigger Polish corpus available and used in many different experiments. IPIC consists of 255 516 328 tokens (of the word level) from which we extracted 19 752 289 possible word bi-grams.

All tests were performed on the Merged Corpus, cf Sec. 1. It consists of 1 610 753 950 tokens. 77 770 719 word bi-grams of different types were extracted from it. There is no overlapping between the tuning corpus (i.e. IPIC) and the test corpus. We checked for duplicated texts and removed them from the test corpus.

MWEs from the plWordNet version 17th April2015 were used as a gold standard set. The set contains 48 735 multi-word lemmas that represent a larger number of MLUs but all corpora were processed on the level of words not word senses.

5.2 Association Measures

On the basis of the results reported in the literature, we selected a number of association measures for the tests plus our own measures: Contonni T1 (Paradowski, 2015), Contonni T2 (Paradowski, 2015), Sorgenfrei (Paradowski, 2015), Dice (Pečina, 2010), Jaccard (Pečina, 2010), Unigram Subtuples (Pečina, 2010), Frequency Biased Mutual Dependency (Pečina, 2010), Mutual Expectation (Pečina, 2010), W Specific Correlation (Hoang et al., 2009b), T-Score (Pečina, 2010), Z-Score (Pečina, 2010), Pearson's Chi$\hat{2}$ (Pečina, 2010), Loglikelihood (Pečina, 2010), Specific Exponential Correlation (Buczyński, 2004), W Specific Exponential Correlation, W Order, and W Term Frequency Order.

5.3 Candidate Extraction Process

In the case of inflectional languages like Polish, a direct application of the statistical measures to word forms would not be feasible for the extraction of MWE candidates. There are too many word forms and each candidate has several inflectional forms on average. Thus, both corpora were first preprocessed by the morphosyntactic tagger WCRFT2 (Radziszewski, 2013) that maps words on their lemmas[8]. Next, the extraction process was performed on the level of lemmas annotated with morphosyntactic information.

MLUs in plWordNet are described with complex information including: multi-word lemmas, partial description of the syntactic structure and syntactic heads, cf (Kurc et al., 2012). The partial description of a MLU is expressed in the WCCL language of morpho-syntactic constraints (Radziszewski et al., 2011). Each MLU is assigned a minimal set of constraints that refer to its lemma and enable recognition of its occurrences in text, e.g. the constraints define the order of constituents (if it is fixed) and morphosyntactic agreements between them. plWordNet editors tried to use the same single constraint set for the description of many MLUs. As a result a limited set of structural classes of MWEs was defined. About 100 MWE structural classes are used in plWordNet, but most of them represent Proper Names and specific idioms. Due to the large size of plWordNet we can assume that the set of MWE classes is representative for Polish.

Annotated lemma bi-grams extracted from the tagged corpus were filtered with morpho-syntactic patterns, cf (Seretan, 2011), written in WCCL language[9] (Radziszewski et al., 2011) and acquired

[8] A lemma is a basic morphological form representing a set of word forms that differ only in the values of the morphosyntactic categories.

[9] See also: `http://nlp.pwr.wroc.pl/redmine/`

from the MWE representation in plWordNet. Only 38 more frequent MWE classes were used, e.g. classes describing Proper Names were excluded.

The extracted statistical data concerning all extracted candidates were stored in a contingency table to be available for the computation of different association measures.

The total number of candidates extracted from the tuning corpus and filtered by 38 WCCL-based patterns was 13 384 814. Most candidates, i.e. 8 249 314, 61,63% of all, were covered by only two patterns that require a noun or a word not recognised by the morphological analyser used in the WCRFT tagger.

All extracted and pre-filtered candidates were used during the extraction process. However the final ranking was created by post-filtering based on a narrow subgroup of only 6 WCCL pattern related to nouns and adjectives. This subgroup was selected on the basis of the frequency of MWEs represented in plWordNet[10]. Only patterns covering the largest number of plWordNet MWEs that were found among the candidates extracted from the corpus were preserved. The selected patterns decreased the number of candidates to 878 096, but the precision was increased very much, as only infrequent classes were removed.

In the case of the test corpus, the initial non-filtered set of 77 770 719 was reduced by the selected 6 patterns to 3 867 835 candidates. However, we could observe that most of them are very infrequent, i.e. below 5 occurrences (for more than 1.6 billion tokens). As such infrequent MWEs would not be interesting for extending plWordNet, we decided to add post-filtering based on the candidate frequency. The threshold was set to at least 6 occurrences. This threshold reduced the number of candidates to 524 760 that is still large number beyond the possibility of the manual verification before adding to plWordNet.

6 Results

Results of experiments are presented in Table 1. First, we tried to optimise the parameter values for different measures on IPIC – the tuning corpus, cf Sec. 5. IPIC was also used for learning weights for the individual measures in the complex VAM measure. In Table 1 we present also the results obtained on the large test set – the Merged Corpus, cf Sec. 5. Parameter values established on the tuning corpus were used during the tests. As both corpora do not have any overlap, we can notice how stable the applied measures are when moved between corpora. It s was especially important for our intended application to the plWordNet expansion, since with the advancement of the work we are interested in new MWEs not yet covered and we use bigger and bigger corpora. The process of collecting texts for the merged corpora is ongoing.

The weights established for the single measures in VAM on the running corpus are as follows: Mutual Expectation: -0.21, T-Score: 0.97, Loglikelihood: 0.68, Jaccard: -0.57, Sorgenfrei: 0.39, Unigram Subtuples: 0.46, SEC($E = 2.8$): 0.77, WSEC($E = 1.1$): -0.65, W Order: 0.04, W Term Frequency Order: 0.52, Contonni T1: 0.63, Contonni T2: -0.58.

In order to evaluate the results we applied two different evaluation measures. The first measure, called *Average Precision* in Table 1 was taken from (Pečina, 2010) and it is based on calculating cut-off precisions for every ranking position on which a true MWE (from plWordNet) was found. Next, in a similar way to (Pečina, 2010), values lower than 0.1 and greater than 0.9 were filtered out. From the rest, the average was computed and used as an evaluation result for the given measure.

As the second evaluation measure we used a simple cut-off precision, called *Cut-off* in Table 1. In this case, the same cut-off ranking position was used for all measures. As the tuning and test corpus have very different size we set the cut-off ranking position on 7 685 for IPIC (tunning) and on 19 687 for the Merged Corpus (test). These values were defined as the minimal number of candidates after filtering across all measures tested, i.e. no measure produced less candidates after filtering, but many extracted more. With the help of the cut-off precision we analyse what is the percentage of extracted candidates on the ranking up to this position that are included in plWordNet. The cut-off precision is a simple measure and does not show the distribution of MWEs across different ranking positions. In the worst case they can be all grouped at the end of the ranking. However, the cut-off value signals what is the estimated percentage of good hints for new MLUs (the real value should be higher, as many MWEs are not included in the

projects/joskipi/wiki/

[10]MWEs have been added mostly to noun and adjective parts of plWordNet

	Average Precision		Cut-off Precision	
Measure	IPIC	Merged Corpus	IPIC	Merged Corpus
Frequency	0.2660	0.2116	0.2636	0.2292
Frequency Biased MD	**0.3585**	**0.2709**	**0.3256**	**0.2643**
Loglikelihood	0.3125	0.2202	0.2882	0.2286
Mutual Expectation	0.3150	0.2246	0.2990	0.2353
Pearsons Chi 2	0.3231	0.2523	0.2982	0.2598
Sorgenfrei	0.3239	0.2543	0.2986	0.2601
Specific Exp. Corr. E=2.8	**0.3592**	**0.2715**	**0.3266**	**0.2642**
Tscore	0.2895	0.2223	0.2766	0.2345
Unigram Subtuples	0.2375	0.1893	0.2373	0.2099
W Order	0.2476	0.1169	0.2393	0.1530
W Specific Correlation	0.3240	**0.2410**	0.2993	0.2434
W Specific Exp. Corr. E=1.1, E=0.9	**0.3339**	0.2394	**0.3049**	0.2442
W Term Frequency Order	0.2915	0.2027	0.2744	0.2263
Zscore	0.3234	0.2525	0.2982	**0.2597**
Jaccard	0.2799	0.2168	0.2743	0.2403
Dice	0.2799	0.2168	0.2743	0.2403
Consonni T1	0.1180	0.0962	0.1447	0.1331
Consonni T2	0.1180	0.0962	0.1447	0.1331
Vector Association Measure	**0.3929**	**0.3114**	**0.3521**	**0.2835**

Table 1: Average and cut off precision of MWEs extracted from tuning corpus (IPIC – the IPI PAN Corpus) and the test corpus (Merged Corpus) and plWordNet the version 17th Apr. 2015 as a source of MLUs to be used as a gold-standard.

applied version of plWordNet).

As we could expect, the results obtained on the test corpus are worse than those on tuning corpus. However, the test corpus is several times larger than the tuning corpus. This can negatively influence the average precision. For the cut-off precision we set much higher cut-off level for the test corpus. Surprisingly, not all measures performed better than the simple *Frequency* measure that can be treated as a baseline.

The complex measure VAM appeared to be the best in all tests. In the case of tuning corpus this was expected, as VAM was optimised on this corpus. However its improvement is even larger on the test corpus. It means that VAM improves moving the false candidates down to the more remote ranking positions. The next two best measures were well known Frequency Biased MD and SEC in the generalised version proposed by us. W Order produced results below the expected level. However, W Order is sensitive to the fixed word order of candidates while many MWEs in plWordNet have non-constrained word order. Other measures proposed by us were close to the top ones. It is worth to emphasise that VAM combines all single measures but with different weights.

Most measures showing good performance in tests in (Pečina, 2010) are also among higher results in our tests. The only difference is the poor performance o Unigram Subtuples – the best single measure in (Pečina, 2010) .

7 Conclusions

We have verified and confirmed the idea of Pečina of combining together many simple association measures. However, tests were done on much larger corpora and a lager set of manually described MWEs.

The obtained results show that a complex measure, even if it is so simple as a linear combination of individual association measures can produce results better than any single measure. What is more, the combined measure was trained on a different corpus and still it expresses better results on a different test corpus. During tests on two large corpora we revisited the evaluation performed by Pečina on much smaller scale and for a different language. In general, we conformed his findings, however, we added to the tests several additional measures including a couple of original measures proposed by us. Any single measure is not as good as their combination, but our results show that some measures, e.g. FBMD, SEC, are worth more attention than the others. Moreover, measures with better performance are interesting components for the complex combined measure. Following observations of (Paradowski, 2015), it is important to avoid combining together correlated measures that produce identical rankings.

Acknowledgments

Work financed by the Polish Ministry of Science and Higher Education, a program in support of scientific units involved in the development of a European research infrastructure for the humanities and social sciences in the scope of the consortia CLARIN ERIC (www.clarin-pl.eu) and ESS-ERIC, 2015-2016.

References

B. Broda, M. Derwojedowa, and M. Piasecki. 2008. Recognition of structured collocations in an inflective language. *Systems Science*, 34(4):27–36. The previous version was published in the Proceedings of AAIA'08, Wisła Poland.

A. Buczyński. 2004. Pozyskiwanie z internetu tekstów do badań lingwistycznych. Master's thesis, Wydział Matematyki Informatyki i Mechaniki Uniwersytetu Warszawskiego, Warsaw.

Ch. Fellbaum, editor. 1998. *WordNet — An Electronic Lexical Database*. The MIT Press.

H. H. Hoang, S. N. Kim, and M.-Y. Kan. 2009a. A re-examination of lexical association measures. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 31–39. ACL.

H. H. Hoang, S. N. Kim, and M.-Y. Kan. 2009b. A re-examination of lexical association measures. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 31–39, Singapore. Suntec.

Ł. Kłyk, P. B. Myszkowski, B. Broda, M. Piasecki, and D. Urbansky. 2012. Metaheuristics for tuning model parameters in two natural language processing applications. In Allan Ramsay and Gennady Agre, editors, *Proceedings of the 15th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, volume 7557 of *Lecture Notes in Computer Science*, pages 32–37, Varna, Bulgaria. Springer.

R. Kurc, M. Piasecki, and B. Broda. 2012. Constraint based description of polish multiword expressions. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2408–2413, Istanbul, Turkey, may. European Language Resources Association (ELRA).

M. Maziarz, M. Piasecki, and S. Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: Synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

M. Maziarz, S. Szpakowicz, and M. Piasecki. 2015a. A procedural definition of multi-word lexical units. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2015*, page this volume, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

M. Maziarz, S. Szpakowicz, M. Piasecki, and A. Dziob. 2015b. Jednostki wielowyrazowe. Procedura sprawdzania leksykalności połączeń wyrazowych [Multi-word units. A procedure for testing the lexicality of collocations]. Technical Report PRE-11, Faculty of Computer Science and Management, Wrocław University of Technology. http://clarin-pl.eu/wp-content/uploads/2015/05/Jednostki-wielowyrazowe_Procecura-sprawdzania-leksykalno%C5%9Bci-po%C5%82%C4%85cze%C5%84-wielowyrazowych.pdf.

M. Paradowski. 2015. On order equivalence relation of binary association measures. *International Journal of Applied Mathematics and Computer Science*, 25(3):*to appear*.

S. Petrović, J. Šnajder, and B. D. Bašić. 2010. Extending lexical association measures for collocation extraction. *Computer Speech and Language*, 24(2):383–394.

S. Petrovica, J. Šnajder, and B. D. Bašic. 2010. Extending lexical association measures for collocation extraction. *Computer, Speech and Language*, 24:383–394.

P. Pečina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

A. Przepiórkowski. 2004. *The IPI PAN Corpus, Preliminary Version*. Institute of Computer Science PAS.

A. Radziszewski, A. Wardyński, and T. Śniatowski. 2011. WCCL: A morpho-syntactic feature toolkit. In I. Habernal and V. Matoušek, editors, *Text, Speech and Dialogue, Plzen 2011*, LNAI 6836, pages 434–441. Springer.

A. Radziszewski. 2013. A tiered CRF Tagger for polish. In *Intelligent Tools for Building a Scientific Information Platform. Studies in Computational Intelligence*, volume 467, pages 215–230. Springer Verlag.

2008. Korpus Rzeczpospolitej. [on-line] www.cs.put.poznan.pl/dweiss/rzeczpospolita. Corpus of text from the online edtion of Rzeczpospolita.

V. Seretan. 2011. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer Netherlands.

B. Svensén. 2009. *A Handbook of Lexicography: the Theory and Practice of Dictionary-making*. Cambridge University Press.

T. Van de Cruys. 2011. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 16–20, Portland.

M. Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining – Proceedings of the International IIS: IIPWM '06 Conference held in Wisła, Poland, June, 2006*, Advances in Soft Computing, pages 511–520, Berlin. Springer.