Mining Co-Occurrence Matrices for SO-PMI Paradigm Word Candidates

Aleksander Wawer
Institute of Computer Science, Polish Academy of Science
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
axw@ipipan.waw.pl

Abstract

This paper is focused on one aspect of SO-PMI, an unsupervised approach to sentiment vocabulary acquisition proposed by Turney (Turney and Littman, 2003). The method, originally applied and evaluated for English, is often used in bootstrapping sentiment lexicons for European languages where no such resources typically exist. In general, SO-PMI values are computed from word co-occurrence frequencies in the neighbourhoods of two small sets of paradigm words. The goal of this work is to investigate how lexeme selection affects the quality of obtained sentiment estimations. This has been achieved by comparing ad hoc random lexeme selection with two alternative heuristics, based on clustering and SVD decomposition of a word co-occurrence matrix, demonstrating superiority of the latter methods. The work can be also interpreted as sensitivity analysis on SO-PMI with regard to paradigm word selection. The experiments were carried out for Polish.

1 Introduction

This paper seeks to improve one of the main methods of unsupervised lexeme sentiment polarity assignment. The method, introduced by (Turney and Littman, 2003), is described in more detail in Section 2. It relies on two sets of paradigm words, positive and negative, which determine the polarity of unseen words.

The method is resource lean and therefore often used in languages other than English. Recent examples include Japanese (Wang and Araki, 2007) and German (Remus et al., 2006).

Unfortunately, the selection of paradigm words rarely receives sufficient attention and is typically done in an *ad hoc* manner. One notable example of manual paradigm word selection method was presented in (Read and Carroll, 2009).

In this context, an interesting variation of the semantic orientation–pointwise mutual information (SO-PMI) algorithm for Japanese was suggested by (Wang and Araki, 2007). Authors, motivated by excessive leaning toward positive opinions, proposed to modify the algorithm by introducing balancing factor and detecting neutral expressions. As will be demonstrated, this problem can be addressed by proper selection of paradigm pairs.

One not entirely realistic, but nevertheless interesting theoretical possibility is to pick pairs of opposing adjectives with the highest loadings identified in Osgood's experiments on semantic differential (Osgood et al., 1967). In the experiments, respondents were presented with a noun and asked to choose its appropriate position on a scale between two bipolar adjectives (for example: *adequate-inadequate*, *valuable-worthless*, *hot-cold*). Factor analysis of the results revealed three distinctive factors, called Osgood dimensions. The first of the dimensions, often considered synonymous with the notion of *sentiment*, was called Evaluative because its foundational adjective pair (one with the highest loading) is *good-bad*.

The first problem with using adjective pairs as exemplary for word co-occurrence distributions on the basis of their loadings, is the fact that factor loadings as measured by Osgood et al. are not necessarily reflected in word frequency phenomena.

Proceedings of the EACL 2012 Student Research Workshop, pages 74–80,
Avignon, France, 26 April 2012.

The second problem is translation: an adjective pair, central in English, may not be as strongly associated with a dimension (here: Evaluative) in other languages and cultures.

The approach we suggest in this paper assumes a latent structure behind word co-occurrence frequencies. The structure may be seen as a mixture of latent variables of unknown distributions that drives word selection. Some of the variables are more likely to produce certain types of highly evaluative words (words with high sentiment scores). We do not attempt to model the structure in a generative way as in for example probabilistic latent semantic analysis (PLSA) or latent Dirichlet allocation (LDA). A generative approximation is not feasible when using corpora such as the balanced, 300-million version of the National Corpus of Polish (NKJP) (Przepiórkowski et al., 2008; Przepiórkowski et al., 2012) [1] applied in the experiments described in the next sections, which does not enable creating a word-document matrix and organizing word occurrences by documents or narrowly specified topics.

Therefore, we propose different techniques. We begin with a symmetric matrix of word co-occurences and attempt to discover as much of its structure as possible using two well established techniques: Singular Value Decomposition and clustering. The discovered structures are then used to optimize the selection of words for paradigm sets used in SO-PMI.

The paper is organized as follows. In Section 2 we define the SO-PMI measure and briefly formulate the problem. Section 3 describes obtaining the set of sentiment word candidates, which are then used to generate a symmetric co-occurence matrix as outlined in Section 4. Section 5 delineates the details of human word scoring, which serves as a basis for evaluations in 9. Sections 6, 7 and 8 describe three distinct approaches to paradigm sets generation.

2 Problem Statement. SO-PMI

When creating a sentiment lexicon, the strength of association between candidate words and each of the two polar classes (positive and negative, for instance) can be calculated using several measures. Perhaps most popular of them, employed in this experiment after (Turney and Littman, 2003) and (Grefenstette et al., 2006) is Pointwise Mutual Information (PMI). The Pointwise Mutual Information (PMI) between two words, *w1* and *w2*, is defined as:

$$\text{PMI}(w1, w2) = \log_2 \left(\frac{p(w1\&w2)}{p(w1)p(w2)} \right)$$

where p(w1 & w2) is the probability of co-occurrence of (w1) and (w2). For the task of assigning evaluative polarity, it is computed as number of co-occurrences of candidate words with each of the paradigm positive and negative words, denoted as pw and nw. Optimal selection of these two sets of words is the subject of this paper.

Once the words are known, the semantic orientation PMI (SO-PMI) of each candidate word c can be computed as:

$$\text{SO-PMI}(c) =$$
$$= \sum_{pw \in PW} \text{PMI}(c, pw) - \sum_{nw \in NW} \text{PMI}(c, nw)$$

The equation above demonstrates that optimization of both word lists, pw and nw, is of crucial importance for the performance of SO-PMI.

3 Generating Sentiment Word Candidates

This section describes the acquisition of sentiment word candidates. The method we followed could be substituted by any other technique which results in a set of highly sentimental lexemes, possibly of varying unknown polarity and strength. A similar experiment for English has been described by (Grefenstette et al., 2006).

The procedure can be described as follows. In the first step, a set of semi-manually defined lexical patterns is submitted to a search engine to find candidates for evaluatively charged terms. Then, the downloaded corpus is analyzed for pattern continuations – lexemes immediately following pattern matches, which are likely to be candidates for sentiment words. In the last step, candidate terms selected this way are tested for their sentiment strength and polarity (in other words, how positive or negative are the conotations). In original experiment described in the cited paper, words were evaluated using the SO-PMI technique.

[1] http://www.nkjp.pl/index.php?page=0&lang=1

The purpose of using extraction patterns is to select candidates for evaluative words. In this experiment, 112 patterns have been created by generating all combinations of elements from two manually prepared sets[2], **A** and **B**:

- **A:** *[0] wydawać się, [1] wydawał się, [2] wydawała się, [3] czuć się, [4] czułem się, [5] czułam się, [6] czułem, [7] być* [3]
- **B:** *[0] nie dość, [1] niewystarczająco, [2] niedostatecznie, [3] za mało, [4] prawie, [5] niemal, [6] tak, [7] taki, [8] zbyt, [9] zbytnio, [10] za bardzo, [11] przesadnie, [12] nadmiernie, [13] szczególnie* [4]

Each pattern (a combination of A and B) has been wrapped with double quotes ("**A B**") and submitted to Google to narrow the results to texts with exact phrases. The Web crawl yielded 17657 web pages, stripped from HTML and other web tags to filter out non-textual content. Two patterns are grammatically incorrect due to gender disagreement, namely *wydawała się taki* and *czułam się taki* [5], thus did not generate any results.

The corpus of 17657 web pages has been analyzed using Spejd[6], originally a tool for partial parsing and rule-based morphosyntactic disambiguation, adapted in the context of this work for the purpose of finding pattern continuations. Again, 112 patterns were constructed by generating all combinations of elements from the two sets, **A** and **B** above. Spejd rules were written as "**A B** *" where the wildcard can be either an adjective or an adverb.

Parsing the web pages using the 112 patterns resulted in acquiring 1325 distinct base word forms (lexemes) recognized by the morphologic analyser and related dictionaries. This list is subsequently used for generating the co-occurrence matrix as delineated in the next Section and for selecting paradigm words.

4 Word Co-Occurrence Matrix

Each word (base form) from the list was sought in the balanced, 300 million segments[7] version of the National Corpus of Polish (NKJP). For each row i and column j of the co-occurrence matrix m, its value was computed as follows:

$$m_{ij} = \frac{f_{ij}}{f_i f_j}$$

where f_{ij} denotes the number of co-occurences of word i within the window of 20 segments left and right with word j, f_i and f_j denote the total numbers of occurrences of each word. The selection of a window of 20 follows the choice in (Turney and Littman, 2003).

This design has been found optimal after a number of experiments with the singular value decomposition (SVD) technique described further. Without the denominator part, decompositions are heavily biased by word frequency. In this defnition, the matrix resembles the PMI form in (Turney and Pantel, 2010), however we found that the logarithm transformation flattens the eigenvalue distribution and is not really necessary.

If the distributions of words i and j are statistically independent, then by the defnition of independence $f_i f_j = f_{ij}$. The product $f_i f_j$ is what we would expect for f_{ij}, if i occurs in the contexts of j by the matter of a random chance. The opposing situation happens when there exists a relationship between i and j, for instance when both words are generated by the same latent topic variable, and we expect f_{ij} to be larger than in the case of independency.

5 Evaluating Word Candidates

In order to evaluate combinations of paradigm words, one needs to compare the computed SO-PMI scores against a human made scoring. Ideally, such a scoring should not only inform about polarity (indication whether a word is positive or negative), but also about association strength (the degree of positivity or negativity). Reliable and

[2] Terms are translations of words listed in (Grefenstette et al., 2006). Many of the expressions denote either excess or deficiency, as for example *not enough* or *too much*.

[3] English translations (morphosyntactic tags in parentheses): [0] *seem to* (inf), [1] *seemed to* (sg,pri,perf,m), [2] *seemed to* (sg,pri,perf,f), [3] *feel* (inf), [4] *felt* (sg,pri,perf,m), [5] *felt* (sg,pri,perf,f), [7] *to be* (inf)

[4] items [0-3] are various ways of expressing *not enough*, items [4-5] *almost*, items [6-7] *such*, items [8-12] *too much*, item [13] *especially*

[5] *seemed(f) so(m)* and *felt(f) so(m)*

[6] `http://nlp.ipipan.waw.pl/Spejd/` (Przepiórkowski and Buczyński, 2007)

[7] A segment usually corresponds to a word. Segments are not longer than orthographic words, but sometimes shorter. See `http://nkjp.pl/poliqarp/help/ense1.html#x2-10001` for a detailed discussion

valid measurement of word associations on a multipoint scale is not easy: the inter rater agreement is likely to decrease with the growing complexity of the scale.

Therefore, we decided that each lexeme was independently scored by two humans using a five point scale. Extreme values denoted **very** negative or positive words, the central value denoted neutral words and remaining intermediate values were interpreted as **somehow** positive or negative. Discrepancies between raters were solved by arithmetic means of conflicting scores rather than introducing the third human (often called the Golden Annotator) to select one value of the two. Consequently, the 5-point scale extended to 10 points.

Human word scores were used in evaluations of methods described in forthcoming sections.

6 Random Selection

The baseline method to compare against is to select lexemes in a random fashion. In order to ensure highest possible performance of the method, lexemes were selected only from those with at least one extreme human score (very positive or very negative) and at least 500 occurrences in the corpus. The last condition renders this method slightly favourable because in the case of SVD, in many eigenvectors the highly loaded terms were not as frequent and had to be selected despite relative rarity.

7 SVD

The word co-occurrence matrix m (1325x1325) was the subject of singular value decomposition (SVD), a well-known matrix factorization technique which decomposes a matrix A into three matrices:

$$A = U\Sigma V^T$$

where Σ is a matrix whose diagonals are the singular values of A, U and V are left and right eigenvectors matrices.

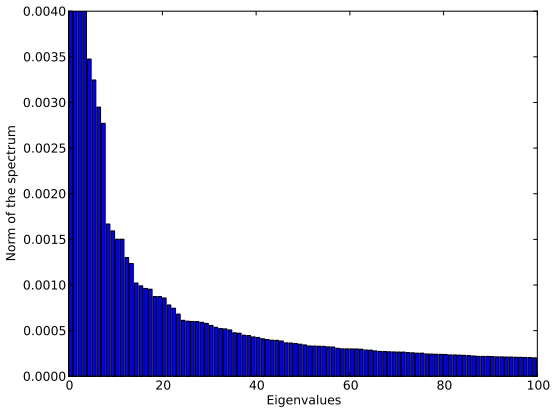
The usage of SVD decompositions has a long and successful history of applications in extracting meaning from word frequencies in word-document matrices, as for example the well established algorithm of latent semantic indexing (LSI). More recently, the usability of analyzing the structure of language via spectral analysis of co-occurrence matrices was demonstrated by studies such as (Mukherjee et al., 2009). The focus was on phonology with the intention to discover principles governing consonant inventories and quantify their importance. Our work, as we believe, is the first to apply SVD in the context of co-occurrence matrices and SO-PMI.

We suspect that the SVD technique can be helpful by selecting lexemes that represent certain amounts of latent co-occurrence structure. Furthermore, the fact that 20 eigenvalues constitutes approximately half of the norm of the spectrum (Horn and Johnson, 1990), as on Table 1, suggests that there may exist a small number of organizing principles which could be potentially helpful to improve the selection of lexemes into paradigm sets.

	c	m
10	0.728	0.410
20	0.797	0.498
100	0.924	0.720

Table 1: Frobenius norm of the spectrum for 10, 20 and 100 first eigenvalues.

Table 1 depicts also the problem of frequency bias, stronger in case of 10 and 20 eigenvalues than for 100. The values were computed for two matrices: c contains only co-occurrence frequencies and m is the matrix described in section 4. Figure 1 plots the eigenvalue spectrum restricted to the first 100 values.



""

Figure 1: Eigenvalue distribution (limited to the first 100).

In order to "discover" the principles behind the co-occurrences, we examine eigenvectors associ-

ated with the largest eigenvalues. Some of the vectors indeed appear to have their interpretations or at least one could name common properties of involved words. The meaning of vectors becomes usually apparent after examination of the first few top component weights.

The list below consits of four eigenvectors, top three and the eighth one (as ordered according to their eigenvalues), along with five terms with highest absolute weights and interpretations of each vector.

1 *sztuczny* (artificial), *liryczny* (lyrical), *upiorny* (ghastly), *zrzędliwy* (grouchy), *przejrzysty* (lucid).
⇒ abstract properties one could attribute to an actor or a play.

2 *instynktowny* (instinctive), *odlotowo* (super/cool), *ostrożny* (careful), *bolesny* (painful), *przesadnie* (excessively)
⇒ physical and sensual experiences

3 *wyemancypować* (emancipate), *opuszczony* (abandoned), *przeszywać* (pierce), *wścibski* (inquisitive), *jednakowo* (alike)
⇒ unpleasant states and behaviours

8 *gładki* (smooth), *kochany* (beloved), *starać się* (make efforts), *niedołężny* (infirm), *intymnie* (intimately)
⇒ intimacy, caring, emotions

As it has been noted before, the eigenvectors of pure co-occurrence matrix c did not deliver anything close in terms of conceivable interpretations. It is also fairly clear that some of the eigenvectors, as for example the third one, are more related to sentiment than the others. This is also evident by examination of average lexeme sentiment of top loaded terms of each vector, not disclosed in the paper.

The heuristic of SVD backed selection of paradigm words maximizes three factors:

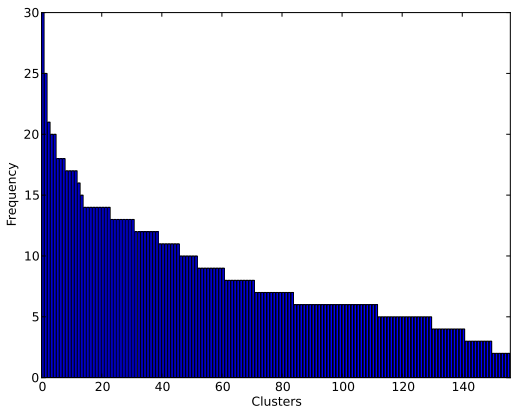
- corpus frequency: avoid rare words where possible;
- eigenvector component weights: select words that contribute the most to a given eigenvector;
- sentiment polarity: select words with the highest absolute human scores.

8 Affinity Propagation

Affinity Propagation (Frey and Dueck, 2007) method was selected because of two distinct advantages for our task. First is the fact that it clusters data by diffusion in the similarity matrix, therefore does not require finding representations in Euclidean space. Second advantage, especially over cluster analysis algorithms such as k-means, is that the algorithm automatically sets its number of clusters and does not depend on initialization.

Affinity Propagation clusters data by exchanging real-valued messages between data points until a high-quality set of exemplars (representative examples, lexemes in our case) and corresponding clusters gradually emerges.

Interestingly, in each parameter setting the algorithm found exactly 156 clusters. It hints at the fact that the number of "latent" variables behind the co-occurrences could indeed be over 100. This is further confirmed by the percentage of norm of the spectrum covered by top 100 eigenvalues.



"""

Figure 2: Histogram of cluster counts.

The five most frequent clusters cover only 116 words. We restrict the selection of paradigm words to the same frequency and polarity conditions as in the case of random method. We pick one paradigm word from each most frequent cluster because we assume that it is sufficient to approximate the principle which organizes that cluster. The heuristic is very similar to the one used in case of SVD.

9 Evaluation

Using continous SO-PMI and multi point scales for human scoring facilitates formulating the problem as a regression one, where goodness of fit of the estimations can be computed using different measures than in the case of classification.

This, however, demands a mapping such that ranges of the continuous SO-PMI scale correspond to discrete human scores. We propose to base such a mapping on dividing the SO-PMI range into 10 segments $\{s_0, ..., s_{10}\}$ of various length, each of which corresponds to one discrete human value.

The choice of values (locations) of specific points is a subject of minimization where the error function E over a set of words W is as follows:

$$E = \sum_{w \in W} dist(s_c, s_e)$$

For each word w, the distance function $dist$ returns the number of segments between the correct segment s_c and the estimated segment s_e using the SO-PMI. We minimize E and find optimum locations for points separating each segment using Powell's conjugate direction method, determined the most effective for this task. Powell's algorithm is a non-gradient numerical optimization technique, applicable to a real valued function which does not need not be differentiable (Powell, 1964).

10 Results

Table 2 presents E errors and extreme (min and max) SO-PMI values computed over two independent samples of 500 lexemes. Error columns indicated as E denote errors computed either on non-optimized default (def) or optimized segments (min). Each combination of paradigm words and each sample required re-computing optimum values of points dividing the SO-PMI scale into segments.

Generally, the randomized selection method performs surprisingly well – most likely due to the fact that the frequency and polarity conditions are the key factors. In either case, the best result was obtained using the selection of paradigm words using the heuristic based on svd, closely followed by aff. In one case, random selection performed better than the aff.

		SO-PMI		E	
sample		min	max	def	min
S1	r_1	-14	29	1226	908
	r_2	-15	23	1131	765
	r_3	-18	8.6	844	710
	aff	-9	25	1057	716
	svd	-13	26	1002	701
S2	r_1	-18	19	983	812
	r_2	-17	15	910	756
	r_3	-11	20	1016	789
	aff	-13	28	1033	732
	svd	-13	35	1028	724

Table 2: SO-PMI ranges and error (E) values on two independent random samples of N=500. 3 randomized selections ($r_1 - r_3$), Affinity Propagation (aff) and SVD (svd).

The small margin of a victory could be explained by the fact that the size of each set of paradigm SO-PMI words is limited to five lexemes. Consequently, it is very difficult to represent a space of over one hundred latent variables – because such appears to be the number indicated by the distribution of eigenvalues in SVD and the number of clusters.

The ranges of SO-PMI values (in the columns min and max) were often non symmetric and leaned towards positive. This shift did not necessarily translate to higher error rates, especially after optimizations.

11 Discussion and Future Work

The methods presented in this article, based on the assumption of latent word co-occurrence structures, performed moderately better than the baseline of random selections. The result is ambiguous because it still requires a more in-depth understanding of underlying mechanims.

The work will be continued in several aspects. One is to pre-determine lexeme type before it is actually evaluated against particular members of paradigm word sets. This could be acheved using a two-step model consisting of lexeme type classification (with regard to over one hundred latent variables) followed by SO-PMI computation, where the selection of paradigm words is not fixed, as in this paper, but dependens on previously selected latent variables. Another promising direction is to focus on explanations and word features: how adding or removing particu-

lar words change the SO-PMI, and more importantly, why (in terms of features involved)? What are the features that change SO-PMI in specific directions? How to extract them?

Acknowledgment

This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund

References

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315:972–976.

Gregory Grefenstette, Yan Qu, David A. Evans, and James G. Shanahan, 2006. *Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes*. Springer. Netherlands.

Roger A. Horn and Charles R. Johnson. 1990. *Matrix Analysis*. Cambridge University Press.

Animesh Mukherjee, Monojit Choudhury, and Ravi Kannan. 2009. Discovering global patterns in linguistic networks through spectral analysis: a case study of the consonant inventories. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 585–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1967. *The Measurement of Meaning*. University of Illinois Press.

M. J. D. Powell. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, January.

Adam Przepiórkowski and Aleksander Buczyński. 2007.
spade: Shallow parsing and disambiguation engine. In *Proceedings of the 3rd Language & Technology Conference*, Poznań.

Adam Przepiórkowski, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, and Marek Łaziński. 2008. Towards the national corpus of polish. In *The proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco*.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw. Forthcoming.

J. Read and J. Carroll. 2009. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 45–52. ACM.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2006. Sentiws: a publicly available german-language resource for sentiment analysis. In *Proceedings of LREC*.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37:141–188, January.

Guangwei Wang and Kenji Araki. 2007. Modifying so-pmi for japanese weblog opinion mining by using a balancing factor and detecting neutral expressions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, NAACL-Short '07, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.