IJCLA VOL. 3, NO. 1, JAN-JUN 2012, PP. 11–24
RECEIVED 01/11/11 ACCEPTED 09/12/11 FINAL 16/06/12

Extracting Emotive Patterns for Languages with Rich Morphology

ALEKSANDER WAWER

Institute of Computer Science, Poland

ABSTRACT

This paper describes a method of acquiring emotive patterns for morphosyntactically rich languages. The goal is to maximize the recall of automatically generated sentiment lexicons in a resource lean fashion. The algorithm requires a small corpus with morphosyntactic annotations to acquire candidates for emotive patterns and evalute the vocabulary, and web as corpus for lexical expansion. The approach, which involves rule mining and contrast sets discovery, is demonstrated and evaluated for Polish.

1 INTRODUCTION AND EXISTING WORK

The research on automated sentiment lexicon acquisition typically falls into one of several categories.

One of the popular related paradigms is focused on using extraction patterns to acquire subjective resources [17, 1]. Subjectivity is a wide term which includes not only evaluations, but also opinions, emotions, and speculations. While its recognition is clearly useful in multiple tasks including opinion mining, the goal of this paper is to demonstrate that sentiment resource acquisition can be succesfully applied using resource-lean methods that do not directly depend on subjectivity[1].

Another approach is to extend WordNet lexicon with sentiment [7]. Despite obvious benefits, evaluative experiments [2, 4] reveal rather moderate successes of applications of this resource. Perhaps an even bigger

[1] We prefer to begin with sentiment acquisition and then extend it with subjectivity, which is both more general and resource-demanding.

problem is that it can not be applied to languages without a WordNet or a poorly developed WordNet – which is an issue we try to avoid with our algorithm.

For the reasons outlined above, the method described in this paper expands and continues efforts to automatically obtain lexicons of evaluatively connotatated words described by [8]. In this approach a set of "emotive patterns"[2] is submitted to a search engine to find candidates for evaluatively charged words. In the second step, semantic orientations (polarity and evaluative strengths) of candidate words are computed using pointwise mutual information – $SO - PMI$ measure as described by [16]. This is done by submitting them to a search engine to find their distributions in neighborhoods of paradigm positive and negative words.

The purpose of using emotive patterns is to select likely candidates for evaluative words. While it is theoretically possible to try all lexemes in a language using any sentiment assessment algorithm such as the $SO - PMI$, in practice it is extremely resource intensive and simply not feasible. The purpose of extraction using emotive patterns is to increase the probability of acquiring evaluatively charged words.

It is notable that any method relying on a fixed, predefined set of emotive patterns is constrained to extracing words that appear in a limited number of syntactic configurations. Consequently, acquired lexemes are often limited to certain part of speech types only, as is the case with patterns described in [8] extracting just adjectives and adverbs.

Put generally, the technique proposed in this paper aims at improving the recall of automated sentiment dictionary generation. Another closely related goal is to extract not only adjectives and adverbs, but also other part of speech types.

While the founding research [6, 12] and multiple subsequent sentiment studies such as [9, 18, 10] focused exclusively on adjectives, it is hardly disputable that nominal word forms and nouns can also carry negative connotations. This fact is at least partially confirmed by research on automated acquisition of subjective resources [15] and WordNet's sentiment extensions [7], but none of the methods can be applied in a resource-lean manner to acquire sentiment vocabulary.

This paper is organized as follows: in Section 2 we start by explaining the reasons of using two types of corpora and separation of pattern generation from lexical acquisition. Section 4 describes how we gener-

[2] We follow the original terminology here as proposed in [8], although one could use the notion of "sentiment extraction patterns" interchangeably.

ate initial pattern candidates, then in Section 5 we give an overview of the algorithm and cover in details contrastive sets attribute selection and greedy tag mining. Sections 6 and 7 present results and discussion of future work.

2 Morphosyntactic and Lexical Corpora

We start by hypothesizing that emotive patterns share certain *morphosyntactic properties*, relevant for acquiring new patterns. The goal is to learn properties which lead to discovering the best patterns given quality criteria discussed in section 3. The approach is similar to that in [15], where a bootstrapping process looks for words that appear in the same extraction patterns as the seeds and assumes that they belong to the same semantic class.

For expanding the set of emotive patterns, a corpus with morphosyntactical tags is required: queries to this corpus are based on combinations of morphosyntactic attributes, selected as potentially improving the quality of patterns which extract evaluatively connotated words. The key property of this approach is that size of this morphosyntactically tagged corpus is of secondary importance because it is not used for lexical acquisition in the sense of extracting sentiment candidate words.

The corpus we used was The National Corpus of Polish [3] with 3 millions segments[4] [13] and Poliqarp[5] query formalism. While sufficient to acquire candidates for emotive patterns, it is not extensive enough to acquire evaluative vocabulary by pattern continuations, because emotive patterns occur in the small, morphosyntactic corpus no more than tens of times and typically exactly once. For the same patterns, search engines like Google or Bing return hundreds or even thousands of results, depending on pattern. Then, on one hand it is straightforward that acquisition of candidates for evaluative words has to take advantage of lexical magnitude and proceed along the web as corpus approach. On the other, candidates for emotive patterns could potentially be extracted from smaller corpora with benefits from morphosyntactic information.

[3] `http://www.nkjp.pl`

[4] The notion of a segment roughly corresponds to a word. The distinction was introduced due to non-standard behaviour of certain rare morphological forms.

[5] `http://poliqarp.sourceforge.net`

3 Pattern Productivity

Algorithms presented in the latter parts of this paper may be evaluated in multiple ways. Quantifiable comparisons of the methods demand formulations of emotive pattern productivity metrics. Evaluation criteria for emotive patterns should involve two main factors:

- Number of distinctive lexemes set L returned by a pattern p.
- Aggregated, absolute $SO - PMI$ of all lexemes in L.

Thus, productivity pr of a pattern p is computed according to the following formula:

$$pr_p = \frac{\sum_{i=0}^{n} |SO - PMI(L_n)|}{\parallel L \parallel}$$

It promotes patterns which return many unique lexemes with high evaluative loading, either positive or negative.

4 Pattern Generation

This section discusses in a step by step fashion how we generate emotive pattern candidates, explaining the rationale and design choices.

4.1 *Part of Speech Sequences*

Probably the most simple conceivable approach to generating sequences which may be suspected of being emotive patterns is by focusing on parts of speech types that constitute known emotive patterns and searching the morphosyntactic corpus for sequences of lexemes which consist of selected part of speech types only. Let emotive patterns be created only by lexemes belonging to part of speech types listed in POS_{emot}:

$$POS_{emot} = \left\{ \substack{conj, ppron*, prep, qub, \\ praet, inf, fin, ger} \right\}$$

The above list is too unrestrictive because it admits a large class of nearly meaningless, but very frequent patterns, as for example sequences of particles (qub) and prepositions (prep). In fact, sequences composed of conjunctions, pronouns, prepositions and particles *only* are very unlikely to be emotive patterns because their role is mostly syntactic. Therefore, we introduced the second list of part of speech types POS_{na}: at least one part of speech type from this list is required to appear in any emotive pattern candidate, but also no emotive pattern may consist of lexemes belonging to these part of speech types only.

$$POS_{na} = \{conj, ppron*, prep, qub\}$$

Firstly, the morphosyntactic corpus is queried for sequences of at least two tokens consisting of lexemes that belong to POS_{emot}. Secondly, all sequences which contain only POS_{na} tokens are removed. However, the number of sequences is still too large and they mostly lack emotive pattern characteristics[6]. This is why subsequent filterings are still needed to approach a plausible set of emotive pattern candidates.

4.2 *Frequency Filtering*

The benefit of using word frequencies in pattern discovery was demonstrated recently by [5]. The key idea, applicable also in the context of emotive pattern discovery, is dividing patterns into high frequency syntactic word slots and slots for less frequent, content words which are actually the subject of interest in concept discovery. While the Davidov and Rappoport method ignored morphosyntactic information, we propose to mix both types of data, frequency and morphosyntactic tags, which seems reasonable for Slavic or German languages.

We introduce frequency as three valued rank variable, dividing word rankings at 200 and 5000. Contrary to Davidov and Rappoport, we include also frequency information about the medium frequency range.

4.3 *Sequence Contexts Filtering*

An intuitive method of such filtering is by checking left and right contexts of a sequence (candidate for an emotive pattern) for occurrences of known, highly evaluative words – the value was operationalized as $SO - PMI$ threshold. Number of rules after filtering with a range of tresholds values, for the set of rules $R1$, is illustrated in Figure 1.

The desired treshold value of $SO - PMI$ should be high enough to avoid patterns without highly evaluative words in their continuations.

Because most of sequences of interest appear exactly once in the morphosyntactic corpus, no further statistical elaboration and systematic

[6] It has been determined by random sampling and comparing average quality to known, effective emotive patterns.

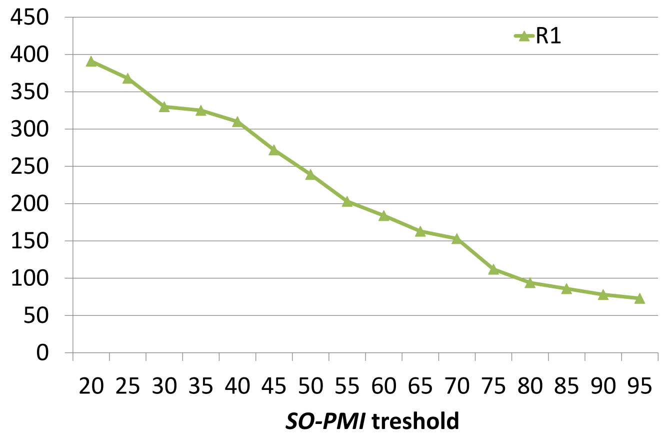

Fig. 1. Number of rules generated for a range tresholds values

analysis of sequence properties is possible, at least not on results obtained from the morphosyntactic corpus.

The proposed method of sequence generation and subsequent multistep filtering is designed to maximize the probability that obtained sequences are indeed emotive patterns. One obvious drawback of imposing presence of known evaluative words around pattern candidates is that it renders impossible detection of emotive patterns surrounded by evaluative words that are not yet recognized. On one hand, it is less of a problem after multiple iterations – once the list of emotive words is expanded. However, it still does not guarantee obtaining all evaluative words, because certain such words may (at least in theory) be obtainable only by a set of emotive patterns which never co-occur with any known evaluative word. The issue is at least partially taken into account by promoting patterns which lead to wide range of emotive words as in section 3.

5 Algorithm Overview

Figure 2 illustrates the processing scheme and core ideas of our approach.

The prerequisite is to prepare an initial seed of emotive patterns. Next steps of the method are iterative and can be summarized as follows:

1. Using any search engine and web as corpus method, issue queries for emotive patterns, get results, extract candidate words and compute their $SO-PMI$. Apply morphosyntactic analysis and disam-

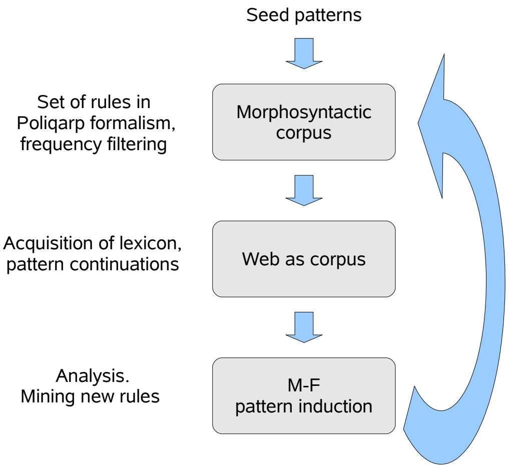
Seed patterns
Set of rules in Poliqarp formalism, frequency filtering
Morphosyntactic corpus
Acquisition of lexicon, pattern continuations
Web as corpus
Analysis. Mining new rules
M-F pattern induction

Fig. 2. Processing phases of the iterative emotive pattern acquisition

biguation so that each recognized token carries a list of associated tags.

2. Create a matrix M such that each row corresponds to one pattern match occurrence with contexts of k tokens each side of the sought candidate for evaluative word, along with morphosyntactic tags for each token and $SO - PMI$. Each emotive pattern is associated to corresponding M rows.
3. Select two sets of rows from M of approximately equal frequency by highest and lowest productivity of patterns that generated them. Call rows generated by high quality patterns M_{Hi-PR} and rows by low quality patterns M_{Low-PR}[7].
4. Using contrast sets mining [3] and association rule discovery as in Algorithm 1, learn a set of morphosyntactic and frequency rules $M - F$, describing the rows in M_{Hi-PR} according to optimization criteria defined by $SO - PMI$ as described below in more detail.
5. Apply $M - F$ rules on morphosyntactically annotated corpus to obtain a new set of emotive patterns.

[7] It seems reasonable to split the rows into groups based on the pattern that generated them and assign those groups into either M_{Hi-PR} or M_{Low-PR} – rather selecting individual rows. This is why frequencies might be different.

5.1 *Seed Patterns*

In the beginning of our experiment, 112 gramatically correct patterns were created by generating cartesian pairs from two sets of words, **A** and **B**. We disclose neither the exact words of **A** and **B** nor the two lists of paradigm positive and negative words, used in $SO - PMI$ computation. This is because both lists are language specific and do not contribute much to this paper. For a reference list, applied successfully in English, see [8].

What is important is that the initial (seed) patterns, submitted to a search engine, resulted in over 11 thousands web pages and 1381 unique lexemes obtained as pattern continuations.

5.2 *M-F Rules*

This section describes rules used to generate lexical patterns from the morphosyntactic corpus.

Rules, in case of Polish, follow tagging scheme defined in [14]. It is notable that this scheme also defines the feature space, which in our case is constituted by 17 variables spanned over 6 word positions (assuming 3 tokens left and 3 right from the extracted sentiment word placeholder). For Slavic and German languages the numbers and consequently tagging schemes will be not far from the one used here:

- Part of speech (in Polish, specific tags defined for over 35 token types).
- Morphosyntactic attributes (example: case, number, person for nouns and adjectives).
- Frequency information (as 3-valued rank).

Not all tag combinations are always present on all positions, therefore the actual feature space is typically more narrow. Rules are defined as sets of type:

$$position_{attribute} = value$$

For example:

$$k - 2_{case} = nom$$

denotes attribute $case$ positioned two tokens left ($k - 2$) from the placeholder for extracted evaluative word, whose value is nom (nominative). Rules define values (tag or frequency ranks) for corresponding attributes and thus every rule potentially selects a number of rows from M – where the values match.

A real example of a rule is presented below:

$$k{-}1_{number}{=}sg,k{+}3_{rank}{=}2,k{-}2_{aspect}{=}imperf \\ k{-}3_{rank}{=}1,k{-}3_{gender}{=}n$$

The rule consists of five attributes, two of them are frequency ranks at different positions. Extracted rules were very rarely longer than 5 or 6 attributes. This is caused by the greedy rule generation algorithm described in listing 1.

5.3 *Contrast Sets Attribute Selection*

For attribute (feature) selection, our method relies on contrast sets. In essence, rules that are later applied to the morphosyntactic corpus are formed only from those attributes (morphosyntactics, frequency and position), for which corresponding values (tags) differ meaningfully in their distributions across more and less productive rules.

For each attribute, we compare the corresponding distributions in M_{Hi-PR} and M_{Low-PR} filtering out attributes that do not differ significantly between the two matrices in terms of χ^2. As an example, let us consider attribute $k-2_{case}$ (morpholosyntactic attribute "case" positioned two tokens to the left from where candidates for evaluative words occurred). Distributions of five cases, which occurred in M_{Hi-PR} and M_{Low-PR}, are as follows:

$$[121, 10, 197, 4, 3], [2, 0, 64, 0, 0]$$

We compute χ^2 and significance (0.0), thus $k-2_{case}$ is considered in further analysis. For the first (seed) data matrix obtained, the contrastive sets method leaves only 54 attributes out of 102 possible for the initial matrix.

5.4 *Rule Generation*

In typical descriptive rule mining, algorithms typically seek surprising deviations or identify significant relationships using support, confidence, information theory measures or various combinations thereof.[8]

[8] An extensive coverage of various approaches, proposing unified terminology, is presented in [11].

Identification of morphosyntactic pattern configurations relevant for acquiring productive emotive patterns should, in our assesment, follow a different objective: the mechanism of rule induction should maximize absolute $SO - PMI$. This formulation seems the most relevant for our purpose and at the same time the most intuitive.

The rule generation algorithm is given in Algorithm 1. Let A denote the set of attributes. Rule r is then generated as follows:

```
$R_r$ = $\{a_r = t\}$ , $t : \arg\max |SO - PMI|$
foreach iteration $i$ do
    foreach attribute $a$ in $A$ do
        foreach tag $t$ of $a$ do
            $|SO - PMI|$ for $\{R_i, a = t\}$
        end
    end
    $a, t : \arg\max \Delta |SO - PMI|$
    $R_{i+1}$ = { $R_i$ , $a = t$ }
end
```

Algorithm 1: Rule generation

The algorithm begins by initializing empty rules R, one for every attribute r in A, and finds a tag which – for this specific attribute – maximizes its absolute $SO - PMI$ in M. Then, in iterative fashion, all other attributes and values are tested, and at the end of each iteration, an attribute and its tag value are appended to a rule, which maximizes absolute $SO - PMI$ of newly created rule.

The algorithm is greedy, because every tag and attribute selection is based on the principle of $SO - PMI$ maximization. This guarantees that every part of created rule contributes to absolute $SO - PMI$, but does not ensure finding globally optimal combinations which maximize $SO - PMI$. Such algorithm would have to compare every tag value of an attribute with all tag combinations of all other attributes, which gives complexity of $O(n!)$.

6 Results

Figure 3 presents $SO - PMI$ distributions for newly acquired lexemes – not seen in any of the previous iterations (unique).

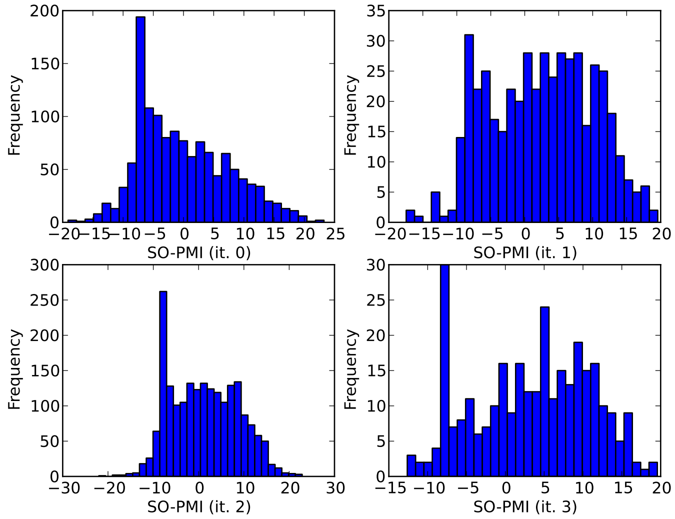

Fig. 3. Histograms of unique SO-PMI words acquired in 4 iterations

The histograms reveal that patterns generated automatically are at least as productive as the human made ones in terms of acquiring highly positive and highly negative words. However, the distribution of new vocabulary in terms of polarity appears to be different in each iteration: the first set of manually crafted patterns (iteration 0) leans towards negative words, iterations 1 and 3 are slightly skewed towards positive, while iteration 2 seems the most balanced, perhaps due to frequency.

Arguably, patterns created automatically have the advantage of generating more balanced distributions. Such results are easy to explain given the randomized pattern generation technique described in Section 4 which does not favour any specific polarity.

Interestingly, negative words tend to be much more densely distributed with a clear peak between -9 and -7. Distributions of positive words are more uniform. In each iteration certain amount of discovered words falls in between -5 and +5; it is likely that these should be considered neutral. The ratio of such words remains similar in each iteration and must be considered not avoidable.

Table 1 presents the results from a different perspective: organized by part of speech[9].

Table 1. Standard Deviation of $SO - PMI$ (std) and percentages of unique vocabulary acquired in each iteration, by part of speech types

	it.0		it.1		it.2		it.3	
POS	std	%	std	%	std	%	std	%
adv	8.04	17.49	6.91	3.97	5.72	3.6	4.68	2.29
pact	7.86	0.99	6.32	1.05	1.69	0.15	0.0	0.33
ger	0.0	0.0	8.27	1.05	7.48	0.64	5.79	0.98
pcon	0.0	0.0	0.0	0.0	0.0	0.05	0.0	0.0
fin	0.0	0.0	0.0	0.21	0.0	0.1	0.0	0.0
subst	6.68	27.06	7.55	66.11	7.71	55.46	7.54	65.69
ppas	0.0	0.33	8.8	0.84	6.03	0.69	1.04	0.65
inf	6.97	15.51	7.45	15.69	7.22	24.59	6.74	16.99
adj	7.69	38.61	7.57	9.62	6.81	13.73	7.18	11.11

The first iteration acquired many adjectives (38%). In subsequent iterations, percents of adjectives among unseen lexemes remained relatively smaller, with much larger relative amounts of nouns (subst). The quality of extracted lexemes, measured by the standard deviation of $SO - PMI$, did not vary a lot between iterations: further iterations were on average as good as the first one, with two exceptions: adverbs, which tended to be more centered around mean, and nouns, which tended to be more spread.

7 Conclusions and Future Work

The contribution of this work affects at least two aspects of sentiment analysis research.

First, it demonstrates how to expand sentiment lexicons beyond the limits of emotive patterns method described by [8]. This is possible thanks to the new iterative technique of vocabulary expansion, based on splitting the task into emotive pattern generation, which may be done by mining small morphosyntactic corpora and frequency information, and lexical acquisition using web as corpus framework.

[9] A description of POS types along with their abbreviations as in the table, can be found here: `http://nkjp.pl/poliqarp/help/ense2.html#x3-30002.1`.

Second, it allows including other part of speech types than just adjectives and adverbs. While this achievement is not distinctive in the context of sentiment lexicons (especially [7] and [15]), it seems one of very few approaches that do not rely on additional resources such as WordNets or specifically annotated corpora.

The method has been implemented and tested in Polish, but it is equally applicable to any language with rich morphology and syntax and comparable resources.

Future work will focus on evaluations against other feasible techniques of emotive pattern acquisition. Certain efforts should also be dedicated to finetuning of various parameters of the algorithm, such as for example $SO - PMI$ thresholds, and examination of net effects of each step of the procedure.

Finally, one natural extension is to re-implement it in other languages, Slavic or German.

ACKNOWLEDGEMENTS The research presented in this paper was supported by the project POIG.01.01.02-14-013/09 co-financed by the European Union under the European Regional Development Fund.

REFERENCES

1. Learning Multilingual Subjective Language via Cross-Lingual Projection (2007)
2. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment analysis in the news. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
3. Bay, S.D., Pazzani, M.J.: Detecting change in categorical data: mining contrast sets. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 302–306. KDD '99, ACM, New York, NY, USA (1999), http://doi.acm.org/10.1145/312129.312263
4. Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.): Is Sentiment a Property of Synsets? Evaluating Resources for Sentiment Classification using Machine Learning. European Language Resources Association (ELRA), Valletta, Malta (may 2010)
5. Davidov, D., Rappoport, A.: Effcient unsupervised discovery of word categories using symmetric patterns and high frequency words. In: COLINGACL (2006)

6. Deese, J.: The associative structure of some common English adjectives. Journal of Verbal Learning and Verbal Behavior 3(5) (1964)
7. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC (2006)
8. Grefenstette, G., Qu, Y., Evans, D.A., Shanahan, J.G.: Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. Springer. Netherlands (2006)
9. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the eighth EACL conference. pp. 174–181 (1997)
10. Kamps, J., Marx, M., Mokken, R.J., Rijke, M.D.: Using wordnet to measure semantic orientations of adjectives. In: The proceedings of LREC. vol. IV, pp. 1115–1118 (2004)
11. Novak, P.K., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. J. Mach. Learn. Res. 10, 377–403 (June 2009)
12. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The Measurement of Meaning. University of Illinois Press (1967)
13. Przepiórkowski, A.: The IPI PAN Corpus: Preliminary version. IPI PAN (2004)
14. Przepiórkowski, A., Woliñski, M.: The unbearable lightness of tagging: A case study in morphosyntactic tagging of polish. In: The Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003 (2003)
15. Riloff, E., Wiebe, J., Wilson, T.: Learning subjective nouns using extraction pattern bootstrapping. In: Proceeings of the Seventh CoNLL conference at HLT-NAACL. pp. 25–32 (2003)
16. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21, 315–346 (2003)
17. Whitelaw, C., Navendu, G., Argamon, S.: Learning subjective language. Computational Linguistics (2005)
18. Wiebe, J.: Learning subjective adjectives from corpora. In: AAAI-00 Proceedings. pp. 735–740 (2000)

ALEKSANDER WAWER
INSTITUTE OF COMPUTER SCIENCE,
POLISH ACADEMY OF SCIENCE,
UL. JANA KAZIMIERZA 5, 01-238 WARSZAWA, POLAND
E-MAIL: <AXW@IPIPAN.WAW.PL>