Kateřina Veselovská, Jan Hajič, jr., Jana Šindlerová

Subjectivity Lexicon for Czech: Implementation and Improvements

Abstract

The aim of this paper is to introduce the Czech subjectivity lexicon[1], a new lexical resource for sentiment analysis in Czech. We describe particular stages of the manual refinement of the lexicon and demonstrate its use in the state-of-the art polarity classifiers, namely the Maximum Entropy classifier. We test the success rate of the system enriched with the dictionary on different data sets, compare the results and suggest some further improvements of the lexicon-based classification system.

1 Introduction

Subjectivity lexicon generation is one of the tasks in sentiment analysis widely worked on both in the academic and in the commercial sphere. The estimation of positive or negative polarity is usually performed by detecting the polarity items, i.e. words or phrases inherently bearing a positive or negative value. There are many methods for compiling a subjectivity lexicon. One of the most straightforward ways is a translation (and further expansion) of an already existing lexicon (see Section 2). Also, the list of evaluative items for specific domains can be extracted directly from the evaluative data, either manually, or by use of probabilistic models. However, it seems profitable for the polarity classification to combine both manually annotated data and a set of the most frequent domain-independent polarity indicators. In this article, we describe the results of an implementation of a method combining classification trained on the reviews with polarity items from Czech subjectivity lexicon.

2 Related Work

The issue of building a subjectivity lexicon is generally described e.g. in (Taboada et al., 2011) or (Liu, 2009). One of the earliest papers that is related to the collection of words with polarity is (Hatzivassiloglou and McKeown, 1997). In their research they experimented with adjectives of the same orientation of polarity. They identify and validate conjunction constraints with respect to the polarity of the adjectives they conjoin. Finally, they collected and manually labelled 1,336 adjectives for their semantic orientation. The idea of words or phrases that inherently bear certain polarity is also exploited in (Turney, 2002).

(Banea, Mihalcea and Wiebe, 2008) use a small set of subjectivity words and apply a bootstrapping method of finding new candidates on the basis of a similarity measure. The authors get to the number of 4000 top frequent entries for the final lexicon. They also describe another method for gaining a subjectivity lexicon: translation of an existing foreign language subjectivity lexicon. Mostly, the authors employ subjectivity lexicons and sentiment analysis in general for machine translation purposes. They are interested e.g. in how the information about polarity should be transferred from one language to another, if the polarity could differ in the corresponding text spans and if it is possible to compile a subjectivity lexicon for the target language during the translation.

There are a number of papers dealing with the topic of building subjectivity lexicons for particular languages (see e.g. Baklival et al., 2012, De Smedt et al., 2012, Jijkoun and Hofmann, 2009 or Peres-Rosas et al., 2012). Also, there is an ongoing research on sentiment analysis in Czech, including the efforts to build a subjectivity lexicon (e.g. as part of a multilingual system, see Steinberger et al., 2011). Still, as far as we know, there is no Czech language subjectivity lexicon publicly available which would help to improve the task and reach the state-of-the-art results.

3 Czech Subjectivity Lexicon

The core of the Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon, also known as the Pittsburgh subjectivity clues, introduced in (Wilson et al., 2005)[2]. The original lexicon, containing more than 8000 polarity expressions, is a part of the OpinionFinder, the system for subjectivity detection in English. The clues in this lexicon were collected from a number of both manually and automatically identified sources (see Riloff and Wiebe, 2003). The patterns and words are expanded iteratively. Some scoring mechanisms were used to ensure the extracted words are in the same semantic category as the seed words.

For translating the data to Czech, we only used parallel corpus CzEng 1.0 (Bojar and Žabokrtský, 2006) containing 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation. By translation, we gained 7228 potentially evaluative expressions. However, some of the items or the assigned polarities appeared rather unreliable at first sight. For this reason, the lexicon has been manually surveyed by one annotator and all the obviously non-evaluative items were excluded. In the end we gained the first applicable version of the lexicon which contained 4947 evaluative expressions. The most frequent items in this set were nouns (e.g. "hulvát" – a boor, 1958) followed by verbs (e.g. "mít rád" – to like, 1699), adjectives (e.g. "špatný" – bad, 821) and adverbs (e.g. "dobře" – rightly/well/correctly, 469).

3.1 Refining the Lexicon

After excluding clearly non-evaluative items, the lexicon has been manually checked again for other incorrect entries. Below we mention the most significant types of inappropriate entries, revealed in the checking phase by an experienced annotator.

The most common problem was including items that are evaluative only in a rare or infrequent meaning or in a specific semantic context whereas mostly they represent non-evaluative expressions (e.g. "bouda" is in most cases used as a word for a "shed", though it can as well mean "dirty trick"). This concerns also the cases where the word is part of a multi-word expression. The main criterion for marking the given item as evaluative was its universal usability in a broader context. Thus we excluded most of the domain-dependent items. The non-evaluativeness of the item was sometimes caused by wrong translation of the original English expression. In case they had not been present in the lexicon yet, the correct translations were added manually.

On the other hand, we found a lot of items with twofold polarity. These were mostly intensifiers like "neuvěřitelně" ('incredibly'), quantifiers like "moc" ('a lot'), general modifiers or words which are frequently connected both with positive and negative meaning (e.g. "[dobré/špatné] svědomí" – [clear/guilty] conscience). The different polarities should be distinguished later on by recording such words in the lexicon together with their prototypical collocations. There are also other instances falling under this category of dual polarity, such as ambiguous words which can be used both in positive and negative meaning – e.g. "využít někoho", meaning to abuse somebody (negative), and "využít příležitosti", to take the opportunity (positive). We put these expressions aside for further research of their semantic features and corpus analysis of their collocations, since they seem to be crucial for more fine-grained sentiment analysis (see also Benamara et al., 2007).

Another problem concerns words assigned an incorrect polarity value. These could be divided into several categories. One of them are e.g. diminutives marked with positive polarity although they are very often used in negative (mostly ironic) sense – e.g. "svatoušek" – goody-goody. Another large group consists of incorrect translations of negated words like "nečestný" – not honest, "nemilosrdný" – not forgiving etc. In this case, the system did not take into account the negative particle preceding the given word and assigned a positive polarity.

After the manual refinement, we got 4,625 evaluative items altogether, of which 1,672 are positive, 2,863 are negative and 90 have both polarities assigned.

4 Evaluating the Lexicon

There are two basic ways to evaluate the quality of a subjectivity lexicon: looking directly at the statistical properties of the lexicon, and plugging the lexicon into classification experiments and measuring potential improvement it brings. We use datasets from various sources and domains, with varying degree of annotation quality, to evaluate its usefulness in various scenarios.

The lexicon can tell us whether a word encountered in the data has (or can have, or usually has) some polarity. We wish to evaluate how exact its estimate is and how useful it is for polarity classification. This evaluation is twofold: while evaluating how accurate the lexicon is, we are also evaluating how well human judgment on prior, context-less polarity of words agrees with their usage and how much of evaluative language is actually expressed through prototypical usage of words that humans judge by themselves evaluative.

Polarity (or, in a wider sense, subjectivity) disambiguation – deciding whether the given token is polar – is a different topic; for the purposes of testing the lexicon, we assume that for each lexicon entry, all its occurrences in the data are polar. By omitting a disambiguation stage, we are estimating the upper bound on lexicon coverage of polar items (lexicon "recall"); disambiguating polar and neutral usage could, on the other hand, increase lexicon "precision".

4.1 Data Sets

For testing the credibility of the lexicon, we used four datasets on which we had previously performed sentiment classification experiments. First, we worked with the data obtained from the Home section of the Czech news website Aktualne.cz – sentences from articles manually identified as evaluative. We identified 175 articles (89,932 words) bearing some subjective information and randomly picked 12 of them for annotation. The annotators annotated 428 segments (i.e. mostly sentences, but also headlines and subtitles) of texts (6,944 words, 1,919 unique lemmas). Second, we used the data from Czech-Slovak Movie Database, CSFD.cz. The data contained 531 segments (14,657 words, 2,556 unique lemmas) and was annotated similarly to the Aktualne dataset (see Veselovská, Hajič and Šindlerová, 2012). In spite of the proportion of the data being rather small, annotating those datasets made clear the challenges to determining the polarity of segments in both domains (see Veselovská, Hajič and Šindlerová, 2012). Third, we used domestic appliance reviews from the Mall.cz retail server. We have worked with 10,177 domestic appliance reviews (158,955 words, 13,370 distinct lemmas) from the Mall.cz retail server. These reviews had been divided into positive (6,365) and negative (3,812) by their authors. We also used the Czech Facebook dataset compiled at the University of Western Bohemia (see Habernal, Ptáček and Steinberger, 2013). This dataset contains 10,000 items, of which 2,587 are positive, 5,174 neutral, 1,991 negative, and 248 "bipolar" posts (posts containing both polarities); the set comprises of 139,222 words and 15,206 distinct lemmas.

Both the datasets and the lexicon were lemmatized and morphologically tagged using the Morče tagger (Ptáček et al., 2005); from the morphological tags, we retained part of speech and negation values and combined them with the raw lemma. These combined tokens form the new "words" of the data sets and the lexicon entries. The dataset sizes are reported for the lemmatized version, since all experiments were run on lemmatized data (since Czech has a very rich morphology).

4.2 Statistical Properties of the Lexicon

There are several questions we can ask about the lexicon quality: What is the *coverage* of the lexicon. Do lexicon entries appear in the data at all? How often does a lexicon entry occur in the data and how many distinct lexicon entries appear in the data? This gives us a very loose upper bound on lexicon "density" in the given data: even if every negative/positive hit came from a text span of the given orientation, the proportion of lexicon items in the evaluative text would be the number of hits divided by the size of the data with the given orientation. Table 1 summarizes how many times a lexicon word occurred in the various data sets (we refer to the occurrence of a lexicon entry in the data as a lexicon hit). "Neg. words" is the total word count over all items tagged as negative in the dataset, "neg. hits" is the total count of words in the data that were found in the lexicon with the negative orientation (negative hits) and "dist. neg. hits" is the amount of distinct negative lexicon entries found in the data set. (Analogously for positive items and lexicon entries.)

Dataset	Neg. words	Pos. words	Neg. hits	Dist. neg. hits	Pos. hits	Dist. pos. hits
Aktualne	1003	358	119	53	102	59
CSFD	4739	6231	254	68	301	65
Reviews	60652	98303	1676	154	4174	146
Facebook	33091	30361	1166	186	2661	182

Tab. 1: Lexicon coverage

However, since many lexicon hits are not in the text span of the corresponding polarity, we need to proceed to testing how good the lexicon is as a predictor. To this end, we used a series of primitive, "raw" binary classifiers. Note that these classifiers are just helper constructs for measuring the relationship between lexicon hits and data item orientations.

We define *lexicon features*: the counts of positive and the count of negative items from the lexicon in the text span. We will call the features POS and NEG. If a lexicon item permits both polarities, it contributes both to POS and NEG counts. If the text span contained no lexicon item, it was given a technical NTR feature with count 1.

We then derive *lexicon indicator variables* from lexicon features: if a lexicon feature is greater or equal to some threshold frequency (denoted $threshold_{LI}$, by default 1) for a data item, the indicator variable value for the given data item is 1; otherwise it is 0. We will denote these features as LI_{POS}, LI_{NEG} *and* LI_{NTR} (*LI = Lexicon Indicator*).

The raw negative classifier then labels all items with negative hits – those with a LI_{NEG} value of 1 – as negative and all the others as non-negative. These binary "predictions" then are evaluated against the binarized "true classes" – all negative data items receive a 1, all non-negative a 0. Analogously for positive items. (Note that under this scheme, one data item may receive a 1 for multiple lexicon indicator features – if it contains both a negative and a positive lexicon hit; this would be a concern if we were building a classifier for all classes at once. However, it only has one true orientation, so it can only contribute once to a correct classification.)

The raw neutral classifier labels as neutral items without more than $threshold_{LI}$ lexicon hits. The "both" class is not predicted.

For each raw classifier on each dataset, we report its precision, recall and support (the true number of data items with the given polarity label) for the label of interest (NEG for the raw negative classifiers, etc.). Recall is the ratio of text spans of the given polarity "found" by the lexicon to the total amount of data items labelled with this polarity, precision is the proportion of correctly identified data items in the set. A recall of 0.5 for the label NEG and negative polarity data items means that in half of the negative data items, a negative lexicon entry appeared. A precision 0.5 means that half the data items in which a negative lexicon entry appeared are actually items labelled as negative in the data.

Given that we are building a separate raw classifier for each class, the baseline performance is also computed for each class separately. The baseline classification assigns a 1 to the LI feature for each data item. This simulates the situation of a lexicon which tags at least one word in every item with the given orientation. Baseline recall is thus 1.0 and so recall ceases to be of interest; our focus is precision, which will tell us how well the lexicon hits

are able to signal that an item actually has the orientation they indicate. At the same time, we watch recall to see a more detailed overview of lexicon coverage.

Recall and precision the raw classifiers achieved are captured in Table 2.

Dataset	Target label	Recall	Precision	Baseline p.	Support
Aktualne	POS	0.294	0.054	0.040	17
	NEG	0.324	0.230	0.166	71
	NTR	0.598	0.792	0.792	338
CSFD	POS	0.454	0.451	0.345	183
	NEG	0.377	0.333	0.284	151
	NTR	0.579	0.467	0.371	197
Reviews	POS	0.354	0.744	0.639	6500
	NEG	0.204	0.551	0.361	3677
	NTR	0.000	0.000	0.000	0
Facebook	POS	0.278	0.320	0.259	2587
	NEG	0.162	0.298	0.199	1991
	NTR	0.741	0.554	0.517	5174

Tab. 2: Lexicon feature "raw" performance

The most important finding from Table 2 is that raw classifier precision tends to follow the baseline for the given label (the proportion of text spans of that class in the data)[3]. This means that the presence or absence of lexicon words per se gives us no additional information: if a lexicon word were present in every data item, we would have the same precision.

Setting $threshold_{LI}$ to 2 very predictably slightly improves precision (at most on the order of 0.1) while drastically reducing recall (to between 0.03 and 0.1). Setting the threshold to 3 showed that no neutral item contained 3 or more lexicon hits and very few non-neutral items did.

While precision can be improved by using more sophisticated classification methods, recall is more limiting – if only 65 % of positive items contain a positive lexicon item, unless we are able to generalize from the lexicon to unseen words, we simply cannot improve recall over 0.65 unless we expand the lexicon.

Again, note that feature performance as measured above is not the performance of "real" classifiers using the lexicon features. The raw classifiers are among the most unsophisticated classification methods based on the lexicon; however, they set a *lower* bound on what should definitely be achievable with the lexicon, based on how lexicon words occur in or outside items with corresponding orientations.

4.3 Evaluation against annotated polar expressions

Since the Aktualne and CSFD data sets are annotated at the expression level[4] including explicitly tagged polar expressions (parts of data items that make the annotator believe the item contains an evaluation, see (Veselovská, Hajič jr. and Šindlerová 2012) for details), we can measure how much the lexicon hits correlate with these expressions. In this polar expression data, there are naturally only positive and negative data items, since only in them

the polar expressions were annotated. We again measure precision, which in this case is the proportion of hits that occur inside polar expressions to the total amount of hits, and recall, which is the proportion of polar expressions with lexicon hits to the number of all polar expressions. The results are reported in Table 3. In this case, support is the number of polar expressions annotated with the given orientation by the given annotator. Since the polar expressions were tagged by two annotators with both significant overlap and significant differences, we report precision and recall for annotators separately (annotator 1/annotator 2).

Dataset	Orientation	Recall	Precision	Support
Aktualne	POS	0.15/0.24	0.50/0.67	13/17
	NEG	0.26/0.26	1.00/0.94	58/66
CSFD	POS	0.09/0.14	0.72/0.87	194/143
	NEG	0.09/0.10	0.78/0.82	152/138

Tab. 3: Precision and Recall against annotated polar expressions

While recall is still low, if the lexicon identifies something, it does tend to lie in expressions of the corresponding orientation. This again suggests that a disambiguation stage is in order; once we know the lexicon hit lies in an evaluative statement, the hit orientation can be relied upon

4.4 Evaluation within Classification Experiments

A further way of testing the lexicon is using lexicon features directly in a classification task, comparing them to automatically extracted features (word and n-gram counts) and evaluating also the combination of automatic and lexicon features. Contrary to the precision/recall scores reported above, the results reported here are for "real" classifiers that classify items by orientation, so that the NEG, NTR, POS and BOTH labels are generated at once. (In section 4.2, each raw classifier was a separate entity.)

Automatic features used in classification were simply word counts. The value of feature f in a text span represents how many times the lemma corresponding to feature f was present.

All classification experiments report 5-fold cross-validation averages. We used the MaxEnt classifier (implemented as Logistic Regression in the scikit-learn Python library[5]). The regularization parameter was set to 1.0 with the exception of the Aktualne dataset, where setting it to values of several thousand significantly improves the performance on the positive text spans.

We report results for the individual classes. It is more informative, especially for datasets with large imbalances of classes, than to report the averaged performance. (Since the classifier performance was never significantly changed by including the lexicon features, the results are reported for classification with automatic and combined lexicon/automatic features in the same table.)

Table 4 shows the results on the Aktualne dataset (note that given the small size and heavily imbalanced nature of the dataset, the results for the negative and positive classes

were very unstable; the positives F-score varying by as much as 0.2 in consecutive cross-validation runs).

Class	Recall	Precision	F-score	Support	Class	Recall	Precision	F-score	Support
NEG	0.12	0.5	0.2	71	NEG	0.01	0.2	0.03	71
NTR	0.94	0.82	0.87	338	NTR	1	0.79	0.88	338
POS	0.47	1	0.62	17	POS	0	0	0	17
BOTH	0	0	0	2	BOTH	0	0	0	2

Tab. 4: Aktualne dataset, classification with/without lexicon features and using only LFs

Table 5 shows the CSFD dataset (while as small, the dataset proved much more stable, varying within 0.05 in consecutive runs). Note that using only the lexicon features improves recall on positive items.

Class	Recall	Precision	F-score	Support	Class	Recall	Precision	F-score	Support
NEG	0.6	0.71	0.6	151	NEG	0.32	0.54	0.4	151
NTR	0.88	0.68	0.76	197	NTR	0.75	0.57	0.65	197
POS	0.53	0.71	0.6	183	POS	0.64	0.63	0.63	183

Tab. 5: CSFD dataset, classification with/without lexicon features and using only LFs

In Table 6 we present the results for the Reviews dataset:

Class	Recall	Precision	F-score	Support	Class	Recall	Precision	F-score	Support
NEG	0.94	0.94	0.94	3677	NEG	0.4	0.73	0.52	3677
POS	0.89	0.89	0.89	6500	POS	0.91	0.73	0.81	6500

Tab. 6: Reviews dataset, classification with/without lexicon features and using only LFs

Table 7 gives the Facebook dataset results:

Class	Recall	Precision	F-score	Support	Class	Recall	Precision	F-score	Support
NEG	0.43	0.61	0.51	1991	NEG	0.06	0.46	0.1	1991
NTR	0.85	0.71	0.77	5174	NTR	0.88	0.56	0.68	5174
POS	0.7	0.77	0.73	2587	POS	0.3	0.48	0.37	2587
BOTH	0.05	0.36	0.08	248	BOTH	0	0	0	248

Tab. 7: Facebook dataset, classification with/without lexicon features and using only LFs

4.5 Identifying problematic lexicon entries

By looking at the lexicon entries which appear in items of opposite or neutral polarity, we can try to detect problematic patterns – those left over from the translation phase that have slipped through the refining process, or problems connected to the usage of lexicon entries in Czech. We report the top ten "mischief" words for each problem category, the English lexicon entries they were translated from, their frequencies in the opposite data and in their "home" data and notes on the prevailing nature of the error after manually inspecting error

sites. Tables 8 and 9 show problems with orientations, Tables 10 and 11 with detecting evaluations vs. neutrality.

Negative hits, positive data	pos.freq	neg.freq	note
manipulace (manipulation, tamper)	178	27	domain-specific (household apps.)
chyba (error, mistake, flaw, etc.)	65	56	negation mismatch ("no flaw at all")
nastavit (plot)	32	35	mistranslated: nastavit=set
vypnout (disable)	24	41	mistrans./lost in trans.: vypnout=turn off
manipulovat (manipulate, manipulation)	18	3	see (1)
komedie (comedy, farce)	18	1	domain mismatch (film reviews)
hluk (din, clamor)	17	28	domain+negation mismatch ("little noise")
odpad (waste, drain)	13	20	domain mismatch (household apps.)
zkusit (try)	9	12	homonymy: try the car vs. a trying test
skvrna (stain, blemish)	9	7	domain+neg. mismatch (household apps.)

Tab. 8: Positive entries occurring most often in negative segments

Positive hits, negative data	neg.freq	pos.freq	Note
dost (pretty, plenty)	135	58	lost in trans.: positive->neutral intensifier
smlouva (agreement, covenant)	30	1	domain mismatch (phone operator trouble)
informace (intelligence)	28	28	mistranslation (intelligence as in CIA)
cena-2 (worth)	24	12	lemmatization disambiguation error
dodat (embolden)	22	16	split phraseme: embolden=dodat+courage
lehce (easily)	20	56	lost in trans.: positive->neutral modifier
vypadat (minister)	19	35	mistranslation: rare Eng. to common Cz.
energie (energize)	19	158	lost in trans.+mistrans.: wrong POS
super (super)	17	127	irony/sarcasm + adversative constructions
snadno (easily, ease, attractively)	16	69	analogous to (6)

Tab. 9: Negative entries occurring most often in positive segments

We see that the most frequent causes of misclassification are domain mismatches, where a word that is a priori – or in the source domain – oriented one way is oriented differently (manipulation, comedy) in another domain. Other frequent problems arise from translation: either a "lost in translation" phenomenon, where what is an originally subjective and evaluative word becomes a more or less neutral word, or a word that is evaluative only weakly or in a very specific context (and thus escaped manual cleansing), or a straight mistranslation. The statistical MT system can also translate rare words as more frequent ones due to the target-side language model. Some other problems suggested by our inspection are the use of words frequently negated in a domain ("hasn't got a single error"), words that are translated as colloquial phrases with only one part of the phrase included in the lexicon, and the occasional use of frequent and strong evaluative words ironically ("super").

We used the same approach to see which negative and positive words most often appear in neutral segments (Tables 10 and 11). Aside from legitimate language use reasons (regular

non-evaluative usage), the discovery of which is again a task for disambiguating whether an entry is *used* as an evaluative word, the most frequent problems stemmed from translation.

Negative hits, neutral data	ntr.freq	neg.freq	note
zkusit (try, difficult)	48	12	homonymy: try the car vs. a trying test
chyba (error, mistake, failure, flaw...)	46	56	regular non-evaluative usage of "chyba"
situace (crisis, predicament, plight...)	17	7	lost in translation: crisis->situation
nastavit (plot)	17	2	mistranslated: nastavit = set
chybit (miss)	16	2	see (2)
ztratit (lose, vanish, doom, dishearten)	12	1	regular non-evaluative usage of "lose"
smrt (death, martyrdom, dying)	11	2	regular non-evaluative usage of "death"
zmizet (vanish, abscond, swagger)	9	5	lost in translation: "zmizet" is neutral
vypnout (disable)	9	41	lost in translation: "vypnout" = "turn off"
sranda (fun, goof)	9	7	orientation error in lexicon refinement

Tab. 10: Negative entries occurring most often in neutral segments

Positive hits, neutral data	ntr.freq	pos.freq	note
cena (worth)	40	12	lemmatization disambiguation error
doufat (hope, hopefully, hopefulness)	36	32	lost in translation: neutral colloquial usage
vypadat (minister)	30	35	mistranslation: rare Eng. to common Cz.
informace (intelligence)	29	28	mistranslation: rare Eng. to common Cz.
dost (pretty, plenty)	28	56	lost in trans.: positive->neutral modifier
dobro (good)	27	42	phrase "dobrý den" (greeting phrase)
souhlasit (agree, consent, concur...)	21	15	regular non-evaluative usage of "agree"
smlouva (agreement, covenant)	20	1	domain mismatch (cell phone operators)
radost (joy, pleasure, delight, happines...)	15	33	non-eval. usage, misannotated items
chystat (solace)	14	6	mistranslation: "chystat" = "to prepare"

Tab. 11: Positive entries occurring most often in neutral segments

4.6 Automated lexicon pruning

Since the number of incorrect hits drops off roughly exponentially, we hypothesised that we could significantly improve lexicon indicator precision by pruning. To see how much we could gain by removing misleading lexicon entries, we combined half of the Facebook and Reviews data to find lexicon entries that impede classification. We then computed the recall and precision statistics of lexicon indicator features and coverage statistics on the second halves of the data (see Fig. 1).

An entry was classified as misleading if we couldn't reject the hypothesis that its occurrences are evenly distributed across items of its class vs. items of all other classes combined, or if we could reject this hypothesis *and* it occurred less frequently in items of its class than in other items. We used the binomial exact test since lexicon hits are often low-frequency words and we thus cannot accurately use the chi-square test.

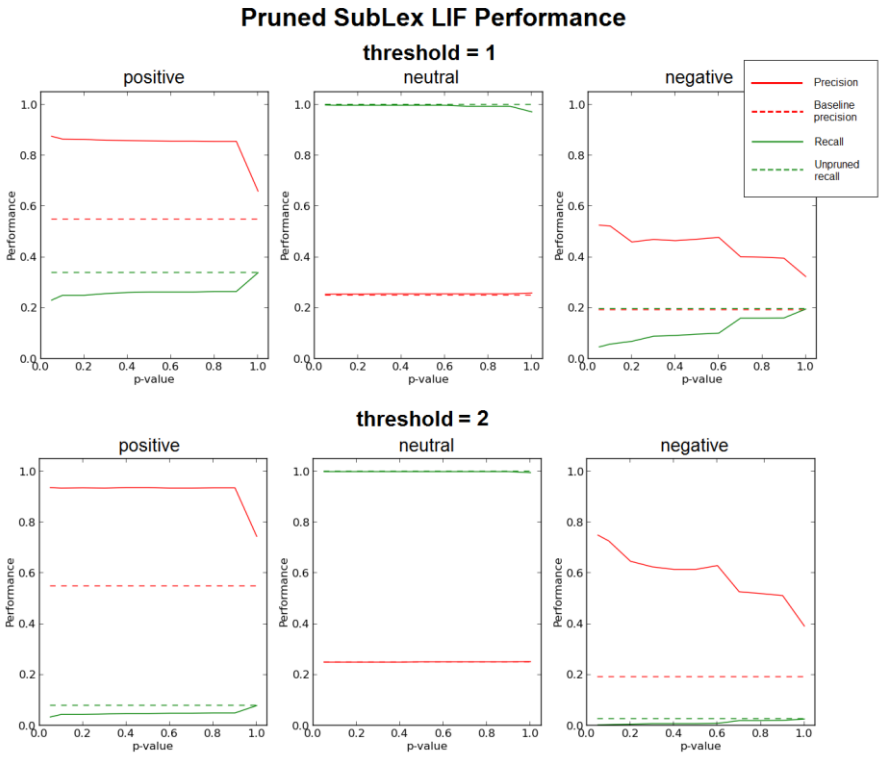

Fig. 1: Pruned lexicon performance. Red lines are precision, green lines recall; dotted lines are baseline precision and pre-pruning recall. From left to right in one sub-figure, pruning is less strict.

We tried pruning at various levels of the test, to find a good tradeoff between gaining precision and not losing too much recall, so that the pruning isn't too severe. The results are reported in Fig. 1. The rightmost data point ($p = 1.0$, $\alpha = 0.0$) is for the lexicon before pruning, so the large skip between $p = 0.9$ and 1.0 is caused by removing words which appear more frequently in items of other orientations than their own orientation. We also used both $threshold_{LI} = 1$ and 2 (setting the indicator threshold to 3 is mostly useless, since very few items contain 3 lexicon hits; see 4.2).

The very low recall for some classes meant that less than 10 items actually contained a lexicon hit of their polarity. However, after such automated pruning, the lexicon may be suitable for building a high-precision classifier such as in (Riloff and Wiebe, 2003).

On the Aktualne dataset, the pruned lexicon never achieved higher precision than the unpruned version. However, on the CSFD data set, for $p = 0.05$ and, $threshold_{LI} = 2$, the precision for LI_{POS} defeated the unpruned (0.793 vs. 0.543) with precision for the other indicators not significantly different from the unpruned lexicon scores.

5 Conclusions and Future Work

From the experiment with lexicon feature recall and precision, we believe that a disambiguation stage, where the occurrence of a lexicon item is assigned some confidence that the occurrence actually is polar, could be highly beneficial – words from the lexicon frequently appear in text spans of opposite polarities or neutral text spans.

Adding the lexicon features to sentiment classifiers did not significantly improve the results in any experiment we have run so far, with the exception of positive text spans in the CSFD dataset. Using the lexicon features alone, which is an option in a scenario where manually annotated data is not available, might work decently on the datasets with preeminently evaluative user-generated content: Aktualne and CSFD. However, to confirm this claim it would be useful to repeat the experiments using other classifiers.

As for the general usefulness of the lexicon, it is apparent that the lexicon by itself – at least by using lexicon features in the manner described above – cannot compete with statistical methods on a representative in-domain annotated dataset such as Reviews, and even when the automatic features are combined with the lexicon features, classifier performance does not improve. However, the lexicon does not hurt classification either, and it remains to be seen whether it can help in classifying previously unseen domains (the Aktualne and CSFD datasets are not large enough for conclusive testing), although the prevalence of domain mismatch among frequent causes of entry/data item orientation mismatch suggests that this will at least require a more sophisticated method.

In order to improve the automatic polarity classification, it could also be advantageous to enhance the subjectivity lexicon by several methods. Firstly, we could use the dictionary-based approach as described by Hu and Liu (2004) or Kim and Hovy (2004) and grow the basic set of words by searching for their synonyms in Czech WordNet (Pala and Ševeček, 1999).

Secondly, we could employ the corpus-based approach based on syntactic or co-occurrence patterns as described in (Hatzivassiloglou & McKeown, 1997). Also, we can extend the lexicon manually by Czech evaluative idioms and other common evaluative phrases. Moreover, it would be useful to add back some special domain-dependent modules for the different areas of evaluation.

To improve the lexicon itself by automatic means besides pruning by statistical significance, we can "ablate" the lexicon: try removing features and see how much the removal hurts (or helps) classification in various scenarios both already implemented and new.

6 Acknowledgments

The research described herein has been supported by the by SVV project number 260 140 and by the LINDAT/CLARIN project funded by the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2010013.

This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

[1] Available at *http://ufal.mff.cuni.cz/seance/data.*

[2] Available at *http://www.cs.pitt.edu/mpqa/subj_lexicon.html.*

[3] This is the same result we could get for evaluative text spans by tagging each with every feature. However, we avoid this degenerate case by also reporting statistics for neutral text spans, if available.

[4] We derived a segment-level polarity from the expression-level annotations.

[5] Available at *http://scikit-learn.org/stable.* For experiments with machine learning, the library has proven to be for us an excellent tool.

References

BAKLIWAL, A., ARORA, P., AND VARMA, V. (2012). "Hindi subjective lexicon: A lexical resource for hindi adjective polarity classification". In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Chair K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds., European Language Resources Association (ELRA).

BANEA, C., MIHALCEA, R., AND WIEBE, J. (2008). "A bootstrapping method for building subjectivity lexicons for languages with scarce resources". In Proceedings of LREC (2008).

BANEA, C., MIHALCEA, R., WIEBE, J. AND HASSAN, S. (2008). "Multilingual subjectivity analysis using machine translation". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 127-135). Association for Computational Linguistics.

BENAMARA, F., CESARANO, C. AND REFORGIATO, D. (2007). "Sentiment analysis: Adjectives and adverbs are better than adjectives alone". Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

BOJAR, O. AND ŽABOKRTSKÝ, Z. (2006). "CzEng: Czech-English Parallel Corpus, Release version 0.5". Prague Bulletin of Mathematical Linguistics, 86. Available from http://ufal.mff.cuni.cz/czeng/.

DE SMEDT, T. AND W. DAELEMANS (2012). "Vreselijk mooi! (terribly beautiful): A subjectivity lexicon for dutch adjectives". In Proceedings of the 8th Language Resources and Evaluation Conference (LREC'12).

HABERNAL, I., PTÁČEK, T. AND STEINBERGER, J. (2013). "Sentiment Analysis in Czech Social Media Using Supervised Machine Learning". WASSA 2013: 65.

HATZIVASSILOGLOU, V. AND MCKEOWN, K. R. (1997). "Predicting the semantic ori-entation of adjectives". Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguis-tics.

HU, M., AND LIU, B. (2004). "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

JIJKOUN, V. AND HOFMANN, K. (2009). "Generating a Non-English Subjectivity Lexicon: Relations That Matter". In proceeding of: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference.

KIM, S.-M., AND HOVY, E. (2004). "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics.

LIU, B. (2009). "Sentiment Analysis and Subjectivity". Invited Chapter for the Handbook of Natural Language Processing, Second Edition. Marcel Dekker, Inc: New York.

PALA, K. AND ŠEVEČEK, P. (1999). "The Czech WordNet, final report". Brno: Masarykova univerzita.

PEREZ-ROSAS, V., BANEA, C. AND MIHALCEA, R. (2012). "Learning Sentiment Lexicons in Spanish". In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC2012).

RILOFF, E. AND WIEBE, J. (2003). "Learning extraction patterns for subjective expressions". In Proceedings of EMNLP-2003.

STEINBERGER, J., LENKOVA, P., KABADJOV, M., STEINBERGER, R. AND VAN DER GOOT, E. (2011). "Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora". In Proceedings of the 8th International Conference Recent Advances in Natural Language Processing.

TABOADA, M., BROOKS, J., TOFILOSKI, M., VOLL, K. AND STEDE, M. (2011). "Lexicon-Based Methods for Sentiment Analysis". Computational Linguistics, 37(2), pp. 267-307.

TURNEY, P. D. (2002). "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews". In Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pp. 417–424.

VESELOVSKÁ, K., HAJIČ JR., J. AND ŠINDLEROVÁ, J. (2012). "Creating Annotated Resources for Polarity Classification in Czech". In Proceedings of the 11th Conference on Natural Language Processing, Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), Vienna, Austria, 2012

WILSON, T., WIEBE, J., AND HOFFMANN, P. (2005). "Recognizing contextual polarity in phrase-level sentiment analysis". In Proceedings of HLT/EMNLP 2005.