Generating a Context-Aware Sentiment Lexicon for Aspect-Based Product Review Mining

Jürgen Broß
Institute of Computer Science
FU Berlin
Berlin, Germany
Email: juergen.bross@fu-berlin.de

Heiko Ehrig
Neofonie Technologieentwicklung
und Informationsmanagement GmbH
Berlin, Germany
Email: heiko.ehrig@neofonie.de

Abstract—A great share of current sentiment analysis techniques is based on special purpose lexicons providing information about the semantic orientation (e.g. positive, negative, neutral) of its entries. Due to the high labor costs of manually assembling such resources, recent work has focused on automatically inducing the polarity of given terms. We follow this line of work while focusing on the domain of user-generated product reviews, a main field of application for sentiment analysis. In this domain, a major observation is that the semantic orientation of terms is often context-dependent which poses an additional challenge to the automatic construction of such lexicons (e.g. positive: “long battery life” vs. negative: “long shutter lag time”). We propose a novel unsupervised method to induce a context-aware sentiment lexicon by utilizing the semi-structuredness of user-generated product reviews.

I. INTRODUCTION

Research in sentiment analysis examines the problem of discovering and evaluating opinion or sentiment expressions in a large number of textual resources. Fields of application are manifold, especially against the background of a continuously growing stream of user-generated content produced in the so-called Web 2.0. Undoubtedly, an increasing share of public discourse and popular opinion is taking place in there (e.g. by writing blogs, using dedicated product review sites, etc.). The sheer amount of these textual resources provides new opportunities for automatic mining and analysis. In this context, an often cited application is the mining of user-generated product reviews [1], [2], [3]. From a consumer’s perspective, summarizing the plethora of available content allows for more informed purchase decisions. From a producer’s perspective, product reviews represent a rich and useful source for market monitoring or quality control.

The major share of current sentiment analysis techniques depends on the existence of special purpose lexicons, so called sentiment lexicons. Such resources provide information about the semantic orientation of single words or whole phrases, i.e. an entry is associated with categories like positive, negative or neutral appraisal. For instance, an occurrence of the verb “like” generally indicates a positive or affirmative sentiment. Due to the huge effort required to manually assemble such lexicons, several techniques have been proposed to automatically induce the semantic category of given words (e.g. [4], [5]).

We follow this line of work while explicitly focusing on the domain of product reviews. We point out two important observations which characterize this application domain. First, the major part of evaluative expressions becomes manifest in the use of adjectives as part of speech. Secondly, the semantic orientation of a term is often context-dependent [2]. As a consequence, terms that are marked with a priori neutral orientation in standard lexicons such as SentiWordNet [4] may in fact express positive or negative sentiment. We address this problem by augmenting an existing sentiment lexicon with context-aware entries. Our approach is based on corpus statistics and exploits the semi-structuredness of online reviews. In particular, we suggest to utilize the inherent semantic of the pros and cons lists (see Figure 1) available in many product reviews.

| Pros: | intuitive menu, nice LCD, good picture quality, extremely portable, good ISO 400 images, fashionable |
| Cons: | blury edges, minimal flash control, difficult to press buttons |

Freetext: This is my first purchase of a camera in the Canon Elph series ...

Fig. 1. A review of a digital camera on opinions.com

The remainder of this article is organized as follows: Section II gives a more detailed insight into previous work on sentiment lexicon induction. In Section III, we define and formalize the basic terms used in subsequent parts. Our approach to a context-aware lexicon induction process is detailed in Section IV. Section V explains our evaluation methodology and reports the obtained results, while Section VI concludes the article.

II. RELATED WORK

The general modus operandi of lexicon construction is either corpus-based or knowledge-based:

The majority of corpus-based approaches is based on co-occurrence statistics. Turney and Littman [5] hypothesize that the “semantic orientation of a word tends to correspond to the semantic orientation of its neighbours”. They apply standard methods such as Pointwise Mutual Information (PMI) to derive a correlation statistic of an unseen word with a set of positive and negative seed words. Using the World Wide Web as a corpus, frequency counts are estimated by the hits count of a
web search engine. The convincing advantage of this technique is its general applicability (domain and language independent). Kaji and Kitsuregawa [7] also use a massive collection of HTML documents as a corpus. Besides using language clues, their main idea is to exploit HTML layout structure to decide on the semantic orientation of a sentence. Chi-square and PMI statistics are used to derive the predominant polarity of a word. However, both approaches fail to address the problem of context dependent polarity words.

Most knowledge-based approaches make use of the lexical relations defined in WordNet [6]. Hu and Liu [8] suggest to measure the relative distance between unclassified words and seed words by using the synonym and antonym relations available in WordNet. Another relevant work is SentiWordNet due to Esuli and Sebastiani [4]. Using a seed word list, they train an ensemble of ternary (neutral, positive, negative) classifiers over associated glosses which are subsequently applied to gloss texts of all previously unseen entries. SentiWordNet is context-aware in the sense that it resembles the syn-sets available in WordNet thus capturing the polysemy of words. However, in a concrete application, the major problem still is to disambiguate the actual meaning of word in a given context. Choi and Cardie [9] focus on the domain adaption of an existing sentiment lexicon. They reformulate the adaptation to a given corpus as a linear optimization problem and apply integer linear programming to solve it. Agreeing to Choi and Cardie, we point out that a major drawback of the above cited research efforts is their tacit assumption that it is feasible to construct a general-purpose lexical resource which is usable in all kinds of relevant sentiment analysis applications.

Approaches similar to ours in the sense that they use meta information available in product reviews are due to Kim and Hovy [10] as well as Branavan et al. [11]. While Kim and Hovy propose to use the pros/cons lists in order to generate a labeled dataset for training a maximum entropy classifier, [11] suggest an approach based on topic models.

III. MODEL AND DEFINITIONS

The term aspect-based review mining refers to the fine-grained sentiment analysis of product reviews. Its goal is to detect individual opinion expressions, determine their semantic orientation and relate them to specific product aspects mentioned in the review. Referring to [2], we regard a product aspect \( P \) as being a component of a product (e.g. lens, camera body) or as an attribute of the product itself (e.g. weight) or of a component (e.g. lens resolution, lens cap size). Further, we define an opinion expression as a tuple \((O, P)\) with \(O\) being a phrase – the opinion indicator – that expresses positive or negative opinion on a product aspect \(P\).

An opinion indicator is denoted as context-dependent if its semantic orientation depends on the product aspect. Phrases having the same polarity independent of any related product aspect are termed as context-independent. Take note that the semantic orientation may differ even when considering only a single product category (e.g. digital cameras: “low price” vs. “low battery drain”).

A sentiment lexicon \(L\) defines a mapping between lexical entries as keys and polarity scores \(S\) as values. We define the key to be an opinion expression, that is a tuple \((O, P)\) and the value \(S\) to be a real number in the interval \([-1; 1]\). The sign of \(S\) determines the polarity of the opinion expression (negative or positive), whereas the absolute value relates to the intensity.

In contrast to earlier proposals such as SentiWordNet, the above definition accommodates the problem of context-dependent polarity. This is simply achieved by defining a dictionary key to be a tuple \((O, P)\). However, in order to be able to incorporate standard, term-based lexicons we allow the second value of the tuple to be \(NULL\). Such an entry then represents a context-independent opinion indicator (e.g. “great”, “horrible”, “like”, “hate”, etc.).

IV. PROPOSED METHOD

A. General Idea

The proposed method takes as input a list of product aspects and a corpus of product reviews (each accompanied with a descriptions of pros and cons). The output is a context-aware sentiment lexicon \(L\) containing entries of type \((O, P) \rightarrow S\) where the opinion indicator \(O\) is an adjective. It is reasonable to assume that an author chooses positive expressions when describing a product aspect in the pros, while she uses negative ones for aspects mentioned in the cons. In that sense, we regard the labeled pros or cons as implicit annotations provided by the review authors.

The idea is to detect all occurrences of product aspects \(P\) and potential opinion indicators \(O\) in such a labeled text. A high precision heuristic is applied to relate observed aspects to opinion indicators. Then co-occurrence statistics are employed to decide whether a tuple \((O, P)\) obtained by this process predominantly originates from a pros or cons document. A tuple that significantly more often occurs in pros than in cons is considered as having positive semantic orientation and vice versa for tuples originating from the cons.

B. Extraction Process

To extract tuples \((O, P)\) we preprocess each pros/cons document with a part-of-speech tagger (we use the Stanford POS Tagger) and additionally annotate all identified product aspects. Due to the simple structure of pros and cons lists (nearly 90% are simple enumerations of rated product aspects), we are able to use very simple, but high precision heuristics to correlate adjectives and aspects: An adjective is considered to be a proper opinion indicator if it either directly precedes an aspect (e.g. “long battery life”), directly succeeds an aspect (e.g. “pictures blurry”) or is connected via a simple predicative construction (e.g. “focus is loud”).

Then for each unique tuple \((O, P)\) obtained via this process, two counters \(C^+\) and \(C^-\) are established, which de facto represent the document frequency of a tuple \((O, P)\) in the pros and cons corpus. Negation (e.g. ”pictures not very sharp”) is

\(^{1}\text{Note that our approach does not aim at determining an intensity score.}\)
detected using indicator terms such as “not” and “n’t”. As it
swaps the semantic orientation of an opinion, the incremen-
tional operation is performed on the opposite counter.

C. Statistical Means

Basically, we need to examine if a given tuple \((O, P)\) ex-
bhibits a strong correlation with the pros or with the cons
document set. Following Manning and Schütze [12], we pro-
pose to apply hypothesis testing to examine the strength of
such a correlation. In particular our approach is based on the
likelihood-ratio test due to Dunning [13].

The likelihood-ratio test (LRT) is a parametric hypotesis
which is used to compare two different hypotheses \(H_0\) and
\(H_1\) with regard to the fit to observed data. The LRT computes
the ratio \(\lambda = \frac{L(H_1)}{L(H_0)}\) of the maximum likelihood \(L(H_0)\) for
the null hypothesis and the maximum likelihood \(L(H_1)\) for
the alternative hypothesis. The value \(\lambda\) expresses how much more
likely it is to make a certain observation under the assumption
\(H_0\) than under the assumption \(H_1\). Furthermore, it is known
that the value \(-2\ln(\lambda)\) is asymptotically \(\chi^2\)-distributed [13].

To decide whether a co-occurrence \((O, P)\) is significantly
more likely to originate from the pros or from the cons corpus,
we apply the test as follows: Let \(D^+\) be the set of all pros and
\(D^-\) be the set of all cons documents contained in a corpus of
reviews. Then \(p = \text{prob}(d \in D^+)\) denotes the probability that
a randomly chosen document \(d\) stems from the pros. Obvi-
ously \(p = 0.5\) as the number of documents in the pros and cons
corpora is the same. With \(p_1 = \text{prob}(d \in D^+ \mid (O, P) \in d)\)
we denote the probability that \(d\) is a document from the pros,
conditioned on the occurrence of the tuple \((O, P)\) in \(d\). The
probability \(p_2 = \text{prob}(d \in D^+ \mid (O, P) \notin d)\) is defined
analogously.

To reason about the origin of a specific tuple \((O, P)\), we
formulate the null hypothesis \(H_0\), that independent of \((O, P)\)
occuring in a document \(d\) or not, the probability of \(d\) stem-
ing from the pros is the same. Formally: \(H_0 : p_1 = p = p_2\).
The alternative hypothesis \(H_1\) assumes that it is not just by
coincidence that a document \(d\) originates from the pros if the
tuple \((O, P)\) is contained. Formally: \(H_1 : p_1 \neq p_2\). Then
maximum likelihood estimates are used to determine \(p, p_1\)
and \(p_2\). The estimates are based on the following frequencies:

<table>
<thead>
<tr>
<th>((O, P)) in (d)</th>
<th>(d \in D^+)</th>
<th>(d \in D^-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>((O, P)) \notin (d)</td>
<td>(c_{11})</td>
<td>(c_{12})</td>
</tr>
<tr>
<td>(c_{21})</td>
<td>(c_{22})</td>
<td></td>
</tr>
</tbody>
</table>

And are computed as:

\[
p = \frac{c_{11} + c_{21}}{c_{11} + c_{21} + c_{12} + c_{22}} = \frac{1}{2}
\]

\[
p_1 = \frac{c_{11}}{c_{11} + c_{21} + c_{12} + c_{22}}
\]

\[
p_2 = \frac{c_{21}}{c_{11} + c_{21} + c_{12} + c_{22}}
\]

Assuming that each occurrence of a tuple is a Bernoulli event,
the counts in the above table follow a binomial distribution.
With \(b(p, k, n)\) being the standard notation for the probability
mass function of the binomial distribution\(^2\), the log-likelihood
ratio computes as:

\[-2 \ln(\lambda) = \ln \frac{L(H_1)}{L(H_0)}
\]

\[
= \ln b(p; c_{11}, c_{11} + c_{12}) - b(p; c_{21}, c_{21} + c_{22})
\]

\[
= \ln b(p; c_{11}, c_{11} + c_{12}) - b(p; c_{21}, c_{21} + c_{22})
\]

\[
= \frac{c_{11} \ln p_1 - c_{21} \ln p_2 - c_{12} \ln (1 - p_1) - c_{22} \ln (1 - p_2)}
\]

Then, if the test rejects \(H_0\) and if \(p_1 > p_2\), \((O, P)\) is
statistically significantly more often part of pros than of cons. In
that case the tuple is assigned a positive semantic orientation.
Analogously, hypotheses are formulated to decide whether a
tuple exhibits negative orientation. In the case that neither
positive nor negative orientation can be assigned, a tuple is
classified as being neutral.

V. EXPERIMENTS AND RESULTS

We conducted experiments to find answers to the following
questions.

Importance of Context-Dependency: What exactly is the
proportion of opinion expressions that are based on an ad-
jectival construction? And in particular, how many of these
expressions exhibit a context dependency on a product aspect?

Intrinsic Evaluation: What is the accuracy of the proposed
approach on the examined datasets? In particular, when exam-
ing the extracted lexicons, how many entries are correct and
what is the fraction of false entries?

Extrinsic Evaluation: When applying an automatically
constructed lexicon, how well does it perform in the task of
classifying extracted opinion expressions? In particular, what
is the effect of augmenting a standard sentiment lexicon
with entries extracted by our approach?

In the following we describe in detail the experimental
setups and datasets and as well discuss the obtained results:

A. Experiment - Importance of Context-Dependency

The aim of this corpus study is to show the importance of
adjectives and to quantify the percentage of adjectives being
context-dependent. To do so we examined an excerpt of the
product review corpus provided by Liu et al. [2]. Basically, Liu
marked each sentence (of a review) that contains at least one
opinion expression with the product aspect mentioned and the
semantic orientation of the opinion indicator. A single sentence
may contain more than one of these annotations.

Results: Our excerpt comprises 685 opinion expressions
from customer reviews of digital cameras and cell phones.
For each opinion expression we manually check whether the
opinion becomes manifest by an adjective and whether it is
context-dependent. The study reveals that in total 59% of
opinion indicators are due to adjectival constructs, whereas
31% of these are context-dependent. That is, at least 18% of
all opinion expressions are context-dependent. We regard this

\(^2\)which computes the probability of exactly \(k\) successes in \(n\) trials given a
success probability of \(p\)
as a strong motivation for our proposal to augment existing lexicons.

B. Experiment - Intrinsic Evaluation

The goal of this experiment is to quantify the accuracy of the lexicon induction process. We report ROC curves for the evaluation. The threshold value varied in the computation of the ROC curves is the minimum confidence required for a classification decision of the likelihood-ratio test. To compute the required evaluation statistics we apply our method to the below-mentioned corpora and manually annotate each extracted tuple with its true class (positive, negative, neutral).

We compare the accuracy of the induction process on two different corpora. We therefore crawled a corpus \(A\) of 13903 product reviews on digital cameras and a corpus \(B\) of 6893 reviews on cell phones, both from the website epinions.com.

Results: Applying the proposed method on corpus \(A\) we were able to extract a total of 3324 unique tuples from which 936 occurred at least two times. We apply the statistical tests to these 936 tuples. Initially, i.e. without considering the computed significance score as a threshold value, 559 tuples are classified as positive orientation, 317 being negatively connoted and 60 are classified as neutral. Tables I and II exemplify the results of the classification process. The 5 topmost context-dependent opinion expressions (extracted from the digital camera corpus) are depicted for positive and the negative class.

The application of our method on corpus \(B\) revealed 1517 unique tuples from which 322 tuples had a minimum support of 2. From these 210 are classified as positive, 98 as negative and 14 were deemed neutral.

We report the following accuracy statistics for the examined corpora: Setting the threshold value to 3.84 (95% confidence interval for the \(\chi^2\)-test), we achieve a precision of 97.06% for the positive class and of 92.45% for the negative class in corpus \(A\). In corpus \(B\) the numbers are 97.41 and 94.74 respectively. In order to measure the influence of the threshold value on the classification performance we vary the value from 0.0 to the maximum score achieved in the class under consideration and plot the results in a ROC curve. Figures 2 and 3 depict the characteristics for the positive and negative class in both corpora.

The main observation is that in both corpora the curves are quite steep, meaning that even at low false positive rates such as 0.1, the approach rejects only a small fraction of true positives. The good results for the intrinsic evaluation are mainly due to the following reasons. First, the results show that even for low log-likelihood scores precision is quite high. That is, even if we assume a rather low confidence level for the statistical tests we achieve a good accuracy. In essence this implies a high accuracy of the applied extraction heuristic and further that our assumption about the “semantic richness” of pros and cons lists is true. A second reason is that the statistic tests provide an adequate score that allows to effectively filter out false tuples which are initially found due to an imperfect extraction heuristic.
C. Experiment - Extrinsic Evaluation

Although revealing insightful numbers, we are aware that assessing the extracted lexicons in isolation (as in Section V-B) is only one part of a sound evaluation: While providing reasonably interpretable statistics for the accuracy of our method, the intrinsic evaluation fails to examine the true recall of the extracted lexicon. It is unclear to which degree our approach helps in the addressed application scenario.

In order to clarify this question we quantify the effect of augmenting a standard sentiment lexicon with entries extracted by our method. We choose the SentiWordNet lexicon as the baseline since it has already been used in other approaches for sentiment analysis.

Again, we make use of the corpus provided by Liu [2] to simulate a realistic setting. Utilizing the annotations conducted for the first experiment (Section V-A), we automatically extracted all sentences containing opinion expressions which are based on adjectival constructions. Since we have not implemented a complete sentiment analysis system, the process of relating the correct adjectives to observed product aspects in complex natural language sentences is simulated by manually extracting adjective/aspect tuples \((O, P)\). The set of tuples obtained by this procedure forms the corpus for the experiment.

We use our own extracted lexicon from Section V-B in a cascading fashion: if a tuple \((O, P)\) cannot be found in our lexicon, we determine its orientation by looking up \(O\) in SentiWordNet. Whereas the baseline use only SentiWordNet and classifies a lemma \(O\) according to the maximum score over all associated synonym sets of this lemma.

Results: We extracted and manually labeled a set of 184 tuples \((O, P)\) for the digital camera domain and 217 tuples for the cell phone domain. Theses are all possible occurring adjectival construstions in the corpus.

The results of the experiment are reported in Table III.

It can be observed that the F-measure for both classes (positive and negative) could be improved substantially. For the digital camera corpus we report improvements of 30.52% and 61.81% respectively. In the cell phone corpus the numbers are 5.30% and 5.32%. The improvement is mainly due to the higher recall value. A great share of context-dependent adjectives were classified as neutral by SentiWordNet and thus were not considered as part of an opinion expression. In case of the digital camera corpus 48.0% of all tuples could directly be classified by our extracted lexicon. For the cell phone corpus we report 27.2%.

VI. Conclusion

This paper set focus on the automatic induction of a context-aware sentiment lexicon to be applied in an aspect-based review mining scenario. We demonstrated that detecting the proper semantic orientation of context-dependent adjectives is a key to this task. Our suggested approach is based on exploiting the inherent semantic properties of user-generated pros and cons lists in product reviews. Statistical hypothesis testing is applied for classifying the semantic orientation of an adjective/product aspect tuple. Our experiments reveal a high accuracy of the proposed method. We also showed that our approach outperforms SentiWordNet as a baseline in the application scenario under consideration. We regard our approach as a simple and valuable component to a complete review mining system.

REFERENCES