Generating Focused Topic-specific Sentiment Lexicons

Valentin Jijkoun **Maarten de Rijke** **Wouter Weerkamp**
ISLA, University of Amsterdam, The Netherlands
jijkoun,derijke,w.weerkamp@uva.nl

Abstract

We present a method for automatically generating focused and accurate topic-specific subjectivity lexicons from a general purpose polarity lexicon that allow users to pin-point subjective on-topic information in a set of relevant documents. We motivate the need for such lexicons in the field of media analysis, describe a bootstrapping method for generating a topic-specific lexicon from a general purpose polarity lexicon, and evaluate the quality of the generated lexicons both manually and using a TREC Blog track test set for opinionated blog post retrieval. Although the generated lexicons can be an order of magnitude more selective than the general purpose lexicon, they maintain, or even improve, the performance of an opinion retrieval system.

1 Introduction

In the area of *media analysis*, one of the key tasks is collecting detailed information about opinions and attitudes toward specific topics from various sources, both offline (traditional newspapers, archives) and online (news sites, blogs, forums). Specifically, media analysis concerns the following system task: given a topic and list of documents (discussing the topic), find all instances of attitudes toward the topic (e.g., positive/negative sentiments, or, if the topic is an organization or person, support/criticism of this entity). For every such instance, one should identify the source of the sentiment, the polarity and, possibly, subtopics that this attitude relates to (e.g., specific targets of criticism or support). Subsequently, a (human) media analyst must be able to aggregate the extracted information by source, polarity or subtopics, allowing him to build support/criticism networks etc. (Altheide, 1996). Recent advances in language technology, especially in *sentiment analysis*, promise to (partially) automate this task.

Sentiment analysis is often considered in the context of the following two tasks:

- *sentiment extraction*: given a set of textual documents, identify phrases, clauses, sentences or entire documents that express attitudes, and determine the polarity of these attitudes (Kim and Hovy, 2004); and
- *sentiment retrieval*: given a topic (and possibly, a list of documents relevant to the topic), identify documents that express attitudes *toward this topic* (Ounis et al., 2007).

How can technology developed for sentiment analysis be applied to media analysis? In order to use a *sentiment extraction* system for a media analysis problem, a system would have to be able to determine which of the extracted sentiments are actually relevant, i.e., it would not only have to identify specific targets of all extracted sentiments, but also decide which of the targets are relevant for the topic at hand. This is a difficult task, as the relation between a *topic* (e.g., a movie) and specific targets of sentiments (e.g., acting or special effects in the movie) is not always straightforward, in the face of ubiquitous complex linguistic phenomena such as referential expressions ("... this beautifully shot *documentary*") or bridging anaphora ("the *director* did an excellent jobs").

In *sentiment retrieval*, on the other hand, the topic is initially present in the task definition, but it is left to the user to identify sources and targets of sentiments, as systems typically return a list of documents ranked by relevance and opinionatedness. To use a traditional sentiment retrieval system in media analysis, one would still have to manually go through ranked lists of documents returned by the system.

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 585–594,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

To be able to support media analysis, we need to combine the specificity of (phrase- or word-level) sentiment analysis with the topicality provided by sentiment retrieval. Moreover, we should be able to identify sources and specific targets of opinions.

Another important issue in the media analysis context is *evidence* for a system's decision. If the output of a system is to be used to inform actions, the system should present evidence, e.g., highlighting words or phrases that indicate a specific attitude. Most modern approaches to sentiment analysis, however, use various flavors of classification, where decisions (typically) come with confidence scores, but without explicit support.

In order to move towards the requirements of media analysis, in this paper we focus on two of the problems identified above: (1) pinpointing evidence for a system's decisions about the presence of sentiment in text, and (2) identifying specific targets of sentiment.

We address these problems by introducing a special type of lexical resource: a topic-specific subjectivity lexicon that indicates specific relevant targets for which sentiments may be expressed; for a given topic, such a lexicon consists of pairs (*syntactic clue*, *target*). We present a method for automatically generating a topic-specific lexicon for a given topic and query-biased set of documents. We evaluate the quality of the lexicon both manually and in the setting of an opinionated blog post retrieval task. We demonstrate that such a lexicon is highly *focused*, allowing one to effectively pinpoint evidence for sentiment, while being competetive with traditional subjectivity lexicons consisting of (a large number of) clue words.

Unlike other methods for topic-specific sentiment analysis, we do not expand a seed lexicon. Instead, we make an existing lexicon more focused, so that it can be used to actually pin-point subjectivity in documents relevant to a given topic.

2 Related Work

Much work has been done in sentiment analysis. We discuss related work in four parts: sentiment analysis in general, domain- and target-specific sentiment analysis, product review mining and sentiment retrieval.

2.1 Sentiment analysis

Sentiment analysis is often seen as two separate steps for determining subjectivity and polarity. Most approaches first try to identify subjective units (documents, sentences), and for each of these determine whether it is positive or negative. Kim and Hovy (2004) select candidate sentiment sentences and use word-based sentiment classifiers to classify unseen words into a negative or positive class. First, the lexicon is constructed from WordNet: from several seed words, the structure of WordNet is used to expand this seed to a full lexicon. Next, this lexicon is used to measure the distance between unseen words and words in the positive and negative classes. Based on word sentiments, a decision is made at the sentence level.

A similar approach is taken by Wilson et al. (2005): a classifier is learnt that distinguishes between polar and neutral sentences, based on a prior polarity lexicon and an annotated corpus. Among the features used are syntactic features. After this initial step, the sentiment sentences are classified as negative or positive; again, a prior polarity lexicon and syntactic features are used. The authors later explored the difference between prior and contextual polarity (Wilson et al., 2009): words that lose polarity in context, or whose polarity is reversed because of context.

Riloff and Wiebe (2003) describe a bootstrapping method to learn subjective extraction patterns that match specific syntactic templates, using a high-precision sentence-level subjectivity classifier and a large unannotated corpus. In our method, we bootstrap from a subjectivity lexicon rather than a classifier, and perform a topic-specific analysis, learning indicators of subjectivity toward a specific topic.

2.2 Domain- and target-specific sentiment

The way authors express their attitudes varies with the domain: An unpredictable movie can be positive, but unpredictable politicians are usually something negative. Since it is unrealistic to construct sentiment lexicons, or manually annotate text for learning, for every imaginable domain or topic, automatic methods have been developed.

Godbole et al. (2007) aim at measuring overall subjectivity or polarity towards a certain entity; they identify sentiments using domain-specific lexicons. The lexicons are generated from manually selected seeds for a broad domain such as *Health* or *Business*, following an approach similar to (Kim and Hovy, 2004). All named entites in a sentence containing a clue from a lexicon are considered targets of sentiment for counting. Because of the data volume, no expensive linguistic processing is performed.

Choi et al. (2009) advocate a joint topic-sentiment analysis. They identify "sentiment topics," noun phrases assumed to be linked to a sentiment clue in the same expression. They address two tasks: identifying sentiment clues, and classifying sentences into positive, negative, or neutral. They start by selecting initial clues from SentiWordNet, based on sentences with known polarity. Next, the sentiment topics are identified, and based on these sentiment topics and the current list of clues, new potential clues are extracted. The clues can be used to classifiy sentences.

Fahrni and Klenner (2008) identify potential targets in a given domain, and create a target-specific polarity adjective lexicon. To this end, they find targets using Wikipedia, and associated adjectives. Next, the target-specific polarity of adjectives is detemined using Hearst-like patterns.

Kanayama and Nasukawa (2006) introduce polar atoms: minimal human-understandable syntactic structures that specify polarity of clauses. The goal is to learn new domain-specific polar atoms, but these are not target-specific. They use manually-created syntactic patterns to identify atoms and coherency to determine polarity.

In contrast to much of the work in the literature, we need to specialize subjectivity lexicons not for a domain and target, but for "topics."

2.3 Product features and opinions

Much work has been carried out for the task of mining product reviews, where the goal is to identify features of specific products (such as *picture*, *zoom*, *size*, *weight* for digital cameras) and opinions about these specific features in user reviews. Liu et al. (2005) describe a system that identifies such features via rules learned from a manually annotated corpus of reviews; opinions on features are extracted from the structure of reviews (which explicitly separate positive and negative opinions).

Popescu and Etzioni (2005) present a method that identifies product features for using corpus statistics, WordNet relations and morphological cues. Opinions about the features are extracted using a hand-crafted set of syntactic rules.

Targets extracted in our method for a topic are similar to features extracted in review mining for products. However, topics in our setting go beyond concrete products, and the diversity and generality of possible topics makes it difficult to apply such supervised or thesaurus-based methods to identify opinion targets. Moreover, in our method we directly use associations between targets and opinions to extract both.

2.4 Sentiment retrieval

At TREC, the Text REtrieval Conference, there has been interest in a specific type of sentiment analysis: opinion retrieval. This interest materialized in 2006 (Ounis et al., 2007), with the opinionated blog post retrieval task. Finding blog posts that are not just about a topic, but also contain an opinion on the topic, proves to be a difficult task. Performance on the opinion-finding task is dominated by performance on the underlying document retrieval task (the topical baseline).

Opinion finding is often approached as a two-stage problem: (1) identify documents relevant to the query, (2) identify opinions. In stage (2) one commonly uses either a binary classifier to distinguish between opinionated and non-opinionated documents or applies reranking of the initial result list using some opinion score. Opinion add-ons show only slight improvements over relevance-only baselines.

The best performing opinion finding system at TREC 2008 is a two-stage approach using reranking in stage (2) (Lee et al., 2008). The authors use SentiWordNet and a corpus-derived lexicon to construct an opinion score for each post in an initial ranking of blog posts. This opinion score is combined with the relevance score, and posts are reranked according to this new score. We detail this approach in Section 6. Later, the authors use domain-specific opinion indicators (Na et al., 2009), like "interesting story" (movie review), and "light" (notebook review). This domain-specific lexicon is constructed using feedback-style learning: retrieve an initial list of documents and use the top documents as training data to learn an opinion lexicon. Opinion scores per document are then computed as an average of opinion scores over all its words. Results show slight improvements (+3%) on mean average precision.

3 Generating Topic-Specific Lexicons

In this section we describe how we generate a lexicon of subjectivity clues and targets for a given *topic* and a list of *relevant documents* (e.g., re-

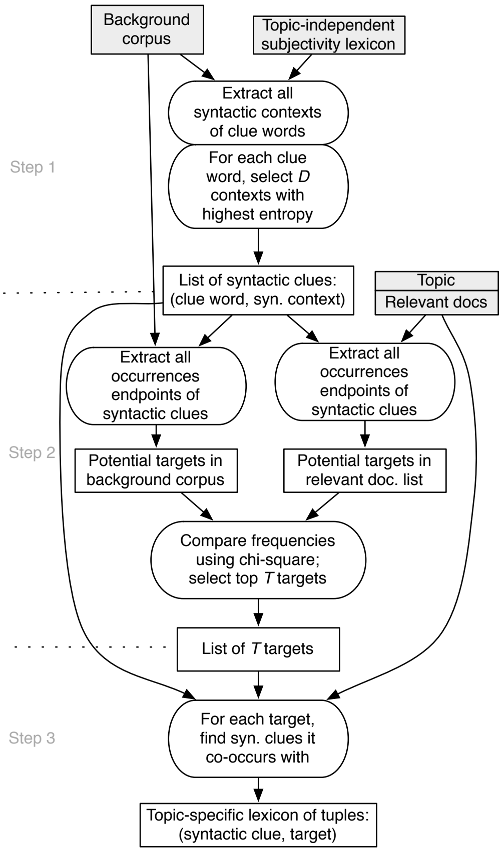
Background corpus
Topic-independent subjectivity lexicon
Extract all syntactic contexts of clue words
For each clue word, select *D* contexts with highest entropy
Step 1
List of syntactic clues: (clue word, syn. context)
Topic
Relevant docs
Extract all occurrences endpoints of syntactic clues
Extract all occurrences endpoints of syntactic clues
Step 2
Potential targets in background corpus
Potential targets in relevant doc. list
Compare frequencies using chi-square; select top *T* targets
List of *T* targets
For each target, find syn. clues it co-occurs with
Step 3
Topic-specific lexicon of tuples: (syntactic clue, target)

Figure 1: Our method for learning a topic-dependent subjectivity lexicon.

trieved by a search engine for the topic). As an additional resource, we use a large background corpus of text documents of a similar style but with diverse subjects; we assume that the relevant documents are part of this corpus as well. As the background corpus, we used the set of documents from the assessment pools of TREC 2006–2008 opinion retrieval tasks (described in detail in section 4). We use the Stanford lexicalized parser[1] to extract labeled dependency triples (*head, label, modifier*). In the extracted triples, all words indicate their category (*noun, adjective, verb, adverb,* etc.) and are normalized to lemmas.

Figure 1 provides an overview of our method; below we describe it in more detail.

3.1 Step 1: Extracting syntactic contexts

We start with a general domain-independent prior polarity lexicon of 8,821 clue words (Wilson et al., 2005). First, we identify *syntactic contexts* in which specific clue words can be used to express attitude: we try to find how a clue word can be syntactically linked to targets of sentiments. We take a simple definition of the syntactic context: a single labeled directed dependency relation. For every clue word, we extract all syntactic contexts, i.e., all dependencies, in which the word is involved (as head or as modifier) in the background corpus, along with their endpoints. Table 1 shows examples of clue words and contexts that indicate sentiments. For every clue, we only select those contexts that exhibit a high entropy among the lemmas at the other endpoint of the dependencies. E.g., in our background corpus, the verb *to like* occurs 97,179 times with a nominal subject and 52,904 times with a direct object; however, the entropy of lemmas of the subjects is 4.33, compared to 9.56 for the direct objects. In other words, subjects of *like* are more "predictable." Indeed, the pronoun *I* accounts for 50% of subjects, followed by *you* (14%), *they* (4%), *we* (4%) and *people* (2%). The most frequent objects of *like* are *it* (12%), *what* (4%), *idea* (2%), *they* (2%). Thus, objects of *to like* will be preferred by the method.

Our entropy-driven selection of syntactic contexts of a clue word is based on the following assumption:

> *Assumption 1:* In text, targets of sentiments are more diverse than sources of sentiments or other accompanying attributes such as location, time, manner, etc. Therefore targets exhibit higher entropy than other attributes.

For every clue word, we select the top D syntactic contexts whose entropy is at least half of the maximum entropy for this clue.

To summarize, at the end of Step 1 of our method, we have extracted a list of pairs (*clue word, syntactic context*) such that for occurrences of the clue word, the words at the endpoint of the syntactic dependency are likely to be targets of sentiments. We call such a pair a *syntactic clue*.

3.2 Step 2: Selecting potential targets

Here, we use the extracted syntantic clues to identify words that are likely to serve as specific targets for opinions about the topic in the relevant documents. In this work we only consider individual words as potential targets and leave exploring other options (e.g., NPs and VPs as targets) for future work. In extracting targets, we rely on the following assumption:

[1] `http://nlp.stanford.edu/software/lex-parser.shtml`

Clue word	Syntactic context	Target	Example
to like	has direct object	*u2*	*I do still like U2 very much*
to like	has clausal complement	*criticize*	*I don't like to criticize our intelligence services*
to like	has *about*-modifier	*olympics*	*That's what I like about Winter Olympics*
terrible	is adjectival modifier of	*idea*	*it's a terrible idea to recall judges for...*
terrible	has nominal subject	*shirt*	*And Neil, that shirt is terrible!*
terrible	has clausal complement	*can*	*It is terrible that a small group of extremists can . . .*

Table 1: Examples of subjective syntactic contexts of clue words (based on Stanford dependencies).

> *Assumption 2*: The list of relevant documents contains a substantial number of documents on the topic which, moreover, contain sentiments about the topic.

We extract all endpoints of all occurrences of the syntactic clues in the relevant documents, as well as in the background corpus. To identify potential attitude targets in the relevant documents, we compare their frequency in the relevant documents to the frequency in the background corpus using the standard χ^2 statistics. This technique is based on the following assumption:

> *Assumption 3:* Sentiment targets related to the topic occur more often in subjective context in the set of relevant documents, than in the background corpus. In other words, while the background corpus contains sentiments towards very diverse subjects, the relevant documents tend to express attitudes related to the topic.

For every potential target, we compute the χ^2-score and select the top T highest scoring targets.

As the result of Steps 1 and 2, as candidate targets for a given topic, we only select words that occur in subjective contexts, and that do so more often than we would normally expect. Table 2 shows examples of extracted targets for three TREC topics (see below for a description of our experimental data).

3.3 Step 3: Generating topic-specific lexicons

In the last step of the method, we combine clues and targets. For each target identified in Step 2, we take all syntactic clues extracted in Step 1 that co-occur with the target in the relevant documents. The resulting list of triples (*clue word, syntactic context, target*) constitute the lexicon. We conjecture that an occurrence of a lexicon entry in a text indicates, with reasonable confidence, a subjective attitude towards the target.

Topic "Relationship between Abramoff and Bush" *abramoff lobbyist scandal fundraiser bush fund-raiser republican prosecutor tribe swirl corrupt corruption norquist democrat lobbying investigation scanlon reid lawmaker dealings president*
Topic "MacBook Pro" *macbook laptop powerbook connector mac processor notebook fw800 spec firewire imac pro machine apple powerbooks ibook ghz g4 ata binary keynote drive modem*
Topic: "Super Bowl ads" *ad bowl commercial fridge caveman xl endorsement advertising spot advertiser game super essential celebrity payoff marketing publicity brand advertise watch viewer tv football venue*

Table 2: Examples of targets extracted at Step 2.

4 Data and Experimental Setup

We consider two types of evaluation. In the next section, we examine the quality of the lexicons we generate. In the section after that we evaluate lexicons quantitatively using the TREC Blog track benchmark.

For extrinsic evaluation we apply our lexicon generation method to a collection of documents containing opinionated utterances: blog posts. The Blogs06 collection (Macdonald and Ounis, 2006) is a crawl of blog posts from 100,649 blogs over a period of 11 weeks (06/12/2005–21/02/2006), with 3,215,171 posts in total. Before indexing the collection, we perform two preprocessing steps: (i) when extracting plain text from HTML, we only keep block-level elements longer than 15 words (to remove boilerplate material), and (ii) we remove non-English posts using TextCat[2] for language detection. This leaves us with 2,574,356 posts with 506 words per post on average. We index the collection using Indri,[3] version 2.10.

TREC 2006–2008 came with the task of *opinionated blog post retrieval* (Ounis et al., 2007). For each year a set of 50 topics was created, giv-

[2] http://odur.let.rug.nl/~vannoord/TextCat/

[3] http://www.lemurproject.org/indri/

ing us 150 topics in total. Every topic comes with a set of relevance judgments: Given a topic, a blog post can be either (i) nonrelevant, (ii) relevant, but not opinionated, or (iii) relevant and opinionated. TREC topics consist of three fields (*title*, *description*, and *narrative*), of which we only use the *title* field: a query of 1–3 keywords.

We use standard TREC evaluation measures for opinion retrieval: MAP (mean average precision), R-precision (precision within the top R retrieved documents, where R is the number of known relevant documents in the collection), MRR (mean reciprocal rank), P@10 and P@100 (precision within the top 10 and 100 retrieved documents). In the context of media analysis, recall-oriented measures such as MAP and R-precision are more meaningful than the other, early precision-oriented measures. Note that for the opinion retrieval task a document is considered relevant if it is on topic and contains opinions or sentiments towards the topic.

Throughout Section 6 below, we test for significant differences using a two-tailed paired t-test, and report on significant differences for $\alpha = 0.01$ ($^{\blacktriangle}$ and $^{\blacktriangledown}$), and $\alpha = 0.05$ ($^{\vartriangle}$ and $^{\triangledown}$).

For the quantative experiments in Section 6 we need a topical baseline: a set of blog posts potentially relevant to each topic. For this, we use the Indri retrieval engine, and apply the Markov Random Fields to model term dependencies in the query (Metzler and Croft, 2005) to improve topical retrieval. We retrieve the top 1,000 posts for each query.

5 Qualitative Analysis of Lexicons

Lexicon size (the number of entries) and selectivity (how often entries match in text) of the generated lexicons vary depending on the parameters D and T introduced above. The two rightmost columns of Table 4 show the lexicon size and the average number of matches per topic. Because our topic-specific lexicons consist of triples (*clue word, syntactic context, target*), they actually contain more words than topic-independent lexicons of the same size, but topic-specific entries are more selective, which makes the lexicon more focused. Table 3 compares the application of topic-independent and topic-specific lexicons to on-topic blog text.

We manually performed an explorative error analysis on a small number of documents, annotated using the smallest lexicon in Table 4 for the topic "March of the Pinguins." We assigned 186 matches of lexicon entries in 30 documents into four classes:

There are some *tragic* moments like eggs freezing , and predators *snatching* the females and *little* ones-you know the whole *NATURE* thing ... but this movie is *awesome*	There are some tragic moments l ike eggs freezing , and predators snatching the females and little ones-you know the whole NATURE thing ... but this **movie** is *awesome*
Saturday was more errands, then spent the evening with Dad and Stepmum, and *finally* was *able* to see March of the Penguins, which was *wonderful*. Christmas Day was *lovely*, surrounded by family, *good* food and drink, and *little* L to play with.	Saturday was more errands, then spent the evening with Dad and Stepmum, and finally was able to see March of the **Penguins**, which was *wonderful*. Christmas Day was lovely, surrounded by family, good food and drink, and little L to play with.

Table 3: Posts with highlighted targets (bold) and subjectivity clues (blue) using topic-independent (left) and topic-specific (right) lexicons.

- REL: sentiment towards a relevant target;
- CONTEXT: sentiment towards a target that is irrelevant to the topic due to context (e.g., opinion about a target "film", but refering to a film different from the topic);
- IRREL: sentiment towards irrelevant target (e.g., "game" for a topic about a movie);
- NOSENT: no sentiment at all

In total only 8% of matches were manually classified as REL, with 62% classified as NOSENT, 23% as CONTEXT, and 6% as IRREL. On the other hand, among documents assessed as opionionated by TREC assessors, only 13% did not contain matches of the lexicon entries, compared to 27% of non-opinionated documents, which does indicate that our lexicon does attempt to separate non-opinionated documents from opinionated.

6 Quantitative Evaluation of Lexicons

In this section we assess the quality of the generated topic-specific lexicons numerically and extrinsically. To this end we deploy our lexicons to the task of opinionated blog post retrieval (Ounis et al., 2007). A commonly used approach to this task works in two stages: (1) identify topically relevant blog posts, and (2) classify these posts as being opinionated or not. In stage 2 the standard approach is to rerank the results from stage 1, instead of doing actual binary classification. We take this approach, as it has shown good performance in the past TREC editions (Ounis et al., 2007) and is fairly straightforward to implement. We also explore another way of using the lexicon: as a source for query expansion (i.e., adding new terms to the original query) in Section 6.2. For all experiments we use the collection described in Section 4.

Our experiments have two goals: to compare the use of topic-independent and topic-specific lexicons for the opinionated post retrieval task, and to examine how different settings for the parameters of the lexicon generation affect the empirical quality.

6.1 Reranking using a lexicon

To rerank a list of posts retrieved for a given topic, we opt to use the method that showed best performance at TREC 2008. The approach taken by Lee et al. (2008) linearly combines a (topical) relevance score with an opinion score for each post. For the opinion score, terms from a (topic-independent) lexicon are matched against the post content, and weighted with the probability of term's subjectivity. Finally, the sum is normalized using the Okapi BM25 framework. The final opinion score S_{op} is computed as in Eq. 1:

$$S_{op}(D) = \frac{Opinion(D) \cdot (k_1 + 1)}{Opinion(D) + k_1 \cdot (1 - b + \frac{b \cdot |D|}{avgdl})}, \quad (1)$$

where k_1, and b are Okapi parameters (set to their default values $k_1 = 2.0$, and $b = 0.75$), $|D|$ is the length of document D, and $avgdl$ is the average document length in the collection. The opinion score $Opinion(D)$ is calculated using Eq. 2:

$$Opinion(D) = \sum_{w \in O} P(sub|w) \cdot n(w, D), \quad (2)$$

where O is the set of terms in the sentiment lexicon, $P(sub|w)$ indicates the probability of term w being subjective, and $n(w, D)$ is the number of times term w occurs in document D. The opinion scoring can weigh lexicon terms differently, using $P(sub|w)$; it normalizes scores to cancel out the effect of varying document sizes.

In our experiments we use the method described above, and plug in the MPQA polarity lexicon.[4] We compare the results of using this topic-independent lexicon to the topic-dependent lexicons our method generates, which are also plugged into the reranking of Lee et al. (2008).

In addition to using Okapi BM25 for opinion scoring, we also consider a simpler method. As we observed in Section 5, our topic-specific lexicons are more selective than the topic-independent lexicon, and a simple number of lexicon matches can give a good indication of opinionatedness of a document:

$$S_{op}(D) = \min(n(O, D), 10)/10, \quad (3)$$

where $n(O, D)$ is the number of matches of the term of sentiment lexicon O in document D.

6.1.1 Results and observations

There are several parameters that we can vary when generating a topic-specific lexicon and when using it for reranking:

D**:** the number of syntactic contexts per clue
T**:** the number of extracted targets
$S_{op}(D)$**:** the opinion scoring function.
α**:** the weight of the opinion score in the linear combination with the relevance score.

Note that α does not affect the lexicon creation, but only how the lexicon is used in reranking. Since we want to assess the quality of lexicons, not in the opinionated retrieval performance as such, we factor out α by selecting the best setting for each lexicon (including the topic-independent) and each evaluation measure.

In Table 4 we present the results of evaluation of several lexicons in the context of opinionated blog post retrieval.

First, we note that reranking using all lexicons in Table 4 significantly improves over the relevance-only baseline for all evaluation measures. When comparing topic-specific lexicons to the topic-independent one, most of the differences are not statistically significant, which is surprising given the fact that most topic-specific lexicons we evaluated are substantially smaller (see the two rightmost columns in the table). The smallest lexicon in Table 4 is seven times more selective than the general one, in terms of the number of lexicon matches per document.

The only evaluation measure where the topic-independent lexicon consistently outperforms topic-specific ones, is Mean Reciprocal Rank that depends on a single relevant opinionated document high in a ranking. A possible explanation

[4] `http://www.cs.pitt.edu/mpqa/`

Lexicon			MAP	R-prec	MRR	P@10	P@100	|lexicon|	hits per doc
no reranking			0.2966	0.3556	0.6750	0.4820	0.3666	—	—
topic-independent			0.3182	0.3776	**0.7714**	**0.5607**	0.3980	8,221	36.17
D	T	S_{op}							
3	50	count	0.3191	0.3769	0.7276$^{\triangledown}$	0.5547	0.3963	2,327	5.02
3	100	count	0.3191	0.3777	0.7416	0.5573	0.3971	3,977	8.58
5	50	count	0.3178	0.3775	0.7246$^{\triangledown}$	0.5560	0.3931	2,784	5.73
5	100	count	0.3178	0.3784	0.7316$^{\triangledown}$	0.5513	0.3961	4,910	10.06
all	50	count	0.3167	0.3753	0.7264$^{\triangledown}$	0.5520	0.3957	4,505	9.34
all	100	count	0.3146	0.3761	0.7283$^{\triangledown}$	0.5347$^{\triangledown}$	0.3955	8,217	16.72
all	50	okapi	0.3129	0.3713	0.7247$^{\blacktriangledown}$	0.5333$^{\triangledown}$	0.3833$^{\triangledown}$	4,505	9.34
all	100	okapi	0.3189	0.3755	0.7162$^{\blacktriangledown}$	0.5473	0.3921	8,217	16.72
all	200	okapi	**0.3229**$^{\blacktriangle}$	**0.3803**	0.7389	0.5547	**0.3987**	14,581	29.14

Table 4: Evaluation of topic-specific lexicons applied to the opinion retrieval task, compared to the topic-independent lexicon. The two rightmost columns show the number of lexicon entries (average per topic) and the number of matches of lexicon entries in blog posts (average for top 1,000 posts).

is that the large general lexicon easily finds a few "obviously subjective" posts (those with heavily used subjective words), but is not better at detecting less obvious ones, as indicated by the recall-oriented MAP and R-precision.

Interestingly, increasing the number of syntactic contexts considered for a clue word (parameter D) and the number of selected targets (parameter T) leads to substantially larger lexicons, but only gives marginal improvements when lexicons are used for opinion retrieval. This shows that our bootstrapping method is effective at filtering out non-relevant sentiment targets and syntactic clues.

The evaluation results also show that the choice of opinion scoring function (Okapi or raw counts) depends on the lexicon size: for smaller, more focused lexicons unnormalized counts are more effective. This also confirms our intuition that for small, focused lexicons simple presence of a sentiment clue in text is a good indication of subjectivity, while for larger lexicons an overall subjectivity scoring of texts has to be used, which can be hard to interpret for (media analysis) users.

6.2 Query expansion with lexicons

In this section we evaluate the quality of targets extracted as part of the lexicons by using them for query expansion. Query expansion is a commonly used technique in information retrieval, aimed at getting a better representation of the user's information need by adding terms to the original retrieval query; for user-generated content, selective query expansion has proved very beneficial (Weerkamp et al., 2009). We hypothesize that if our method manages to identify targets that correspond to issues, subtopics or features associated with the topic, the extracted targets should be good candidates for query expansion. The experiments described below test this hypothesis.

Run	MAP	P@10	MRR
Topical blog post retrieval			
Baseline	0.4086	0.7053	0.7984
Rel. models	0.4017$^{\triangledown}$	0.6867	0.7383$^{\blacktriangledown}$
Subj. targets	0.4190$^{\vartriangle}$	0.7373$^{\vartriangle}$	0.8470$^{\vartriangle}$
Opinion retrieval			
Baseline	0.2966	0.4820	0.6750
Rel. models	0.2841$^{\blacktriangledown}$	0.4467$^{\blacktriangledown}$	0.5479$^{\blacktriangledown}$
Subj. targets	0.3075	0.5227$^{\blacktriangle}$	0.7196

Table 5: Query expansion using relevance models and topic-specific subjectivity targets. Significance tested against the baseline.

For every test topic, we select the 20 top-scoring targets as expansion terms, and use Indri to return 1,000 most relevant documents for the expanded query. We evaluate the resulting ranking using both topical retrieval and opinionated retrieval measures. For the sake of comparison, we also implemented a well-known query expansion method based on Relevance Models (Lavrenko and Croft, 2001): this method has been shown to work well in many settings. Table 5 shows evaluation results for these two query expansion methods, compared to the baseline retrieval run.

The results show that on topical retrieval query expansion using targets significantly improves retrieval performance, while using relevance models actually hurts all evaluation measures. The failure of the latter expansion method can be attributed to the relatively large amount of noise in user-generated content, such as boilerplate material, timestamps of blog posts, comments etc. (Weerkamp and de Rijke, 2008). Our method uses full syntactic parsing of the retrieved documents, which might substantially reduce the amount of noise since only (relatively) well-formed English sentences are used in lexicon generation.

For opinionated retrieval, target-based expansion also improves over the baseline, although the differences are only significant for P@10. The consistent improvement for topical retrieval suggests that a topic-specific lexicon can be used both for query expansion (as described in this section) and for opinion reranking (as described in Section 6.1). We leave this combination for future work.

7 Conclusions and Future Work

We have described a bootstrapping method for deriving a topic-specific lexicon from a general purpose polarity lexicon. We have evaluated the quality of generated lexicons both manually and using a TREC Blog track test set for opinionated blog post retrieval. Although the generated lexicons can be an order of magnitude more selective, they maintain, or even improve, the performance of an opinion retrieval system.

As to future work, we intend to combine our method with known methods for topic-specific lexicon *expansion* (our method is rather concerned with lexicon "restriction"). Existing sentence- or phrase-level (trained) sentiment classifiers can also be used easily: when collecting/counting targets we can weigh them by "prior" score provided by such classifiers. We also want to look at more complex syntactic patterns: Choi et al. (2009) report that many errors are due to exclusive use of unigrams. We would also like to extend potential opinion targets to include multi-word phrases (NPs and VPs), in addition to individual words. Finally, we do not identify polarity yet: this can be partially inherited from the initial lexicon and refined automatically via bootstrapping.

Acknowledgements

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.-815, 640.004.802.

References

Altheide, D. (1996). *Qualitative Media Analysis*. Sage.

Choi, Y., Kim, Y., and Myaeng, S.-H. (2009). Domain-specific sentiment analysis using contextual feature generation. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 37–44, New York, NY, USA. ACM.

Fahrni, A. and Klenner, M. (2008). Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In *Proc.of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention, 1st-2nd April 2008. University of Aberdeen, Aberdeen, Scotland*, pages 60 – 63.

Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Kanayama, H. and Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363, Morristown, NJ, USA. Association for Computational Linguistics.

Kim, S. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of COLING 2004*.

Lavrenko, V. and Croft, B. (2001). Relevance-based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*.

Lee, Y., Na, S.-H., Kim, J., Nam, S.-H., Jung, H.-Y., and Lee, J.-H. (2008). KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval. In *Proceedings of TREC 2008*.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*.

Macdonald, C. and Ounis, I. (2006). The TREC Blogs06 collection: Creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow.

Metzler, D. and Croft, W. B. (2005). A markov random feld model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pages 472–479, New York, NY, USA. ACM Press.

Na, S.-H., Lee, Y., Nam, S.-H., and Lee, J.-H. (2009). Improving opinion retrieval based on query-specific sentiment lexicon. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 734–738, Berlin, Heidelberg. Springer-Verlag.

Ounis, I., Macdonald, C., de Rijke, M., Mishne, G., and Soboroff, I. (2007). Overview of the TREC 2006 blog track. In *The Fifteenth Text REtrieval Conference (TREC 2006)*. NIST.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical methods in Natural Language Processing (EMNLP)*.

Weerkamp, W., Balog, K., and de Rijke, M. (2009). A generative blog post retrieval model that uses query expansion based on external collections. In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-ICNLP 2009)*, Singapore.

Weerkamp, W. and de Rijke, M. (2008). Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*, page 923931, Columbus, Ohio. Association for Computational Linguistics, Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Morristown, NJ, USA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.