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

This paper is a compressed version of Jijkoun et al. (2010).

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. 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* (identify subjective phrases/ sentence in a document) and *sentiment retrieval* (identify/rank documents with subjective attitude on a topic).

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.

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.

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. 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., retrieved 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 3). We use the Stanford lexicalized parser 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.

3. DATA AND EXPERIMENTAL SETUP

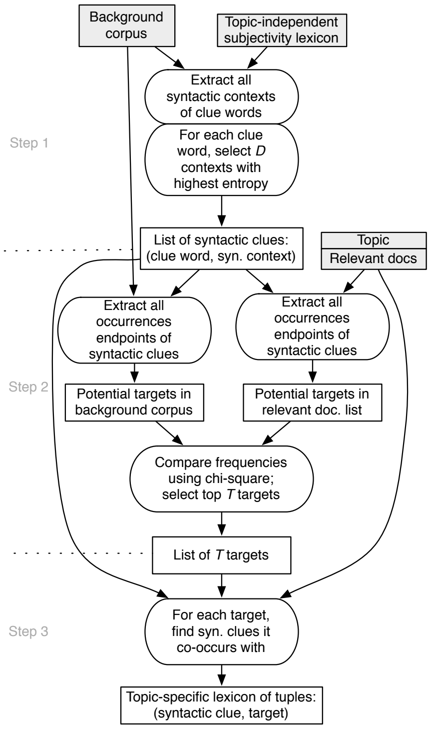

Figure 1: Our method for learning a topic-dependent subjectivity lexicon.

For extrinsic evaluation we apply our lexicon generation method to a collection of documents containing opinionated utterances: the TREC Blog06 collection.

TREC 2006–2008 came with the task of *opinionated blog post retrieval*. For each year a set of 50 topics was created, giving 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.

4. 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. 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 and is fairly straightforward to implement. For all experiments we use the collection described in Section 3.

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.

4.1 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. In addition to using Okapi BM25 for opinion scoring, we also consider a simpler method:a simple count of lexicon matches in a document.

4.1.1 Results and observations

There are several parameters that we can vary when generating a topic-specific lexicon and when using it for reranking: the number of syntactic contexts per clue, the number of extracted targets, the opinion scoring function, the weight of the opinion score in the linear combination with the relevance score.

First, we note that reranking using all lexicons 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.

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 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.

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

Jijkoun, V., de Rijke, M., and Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden. ACL, ACL.

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*.