Building Subjectivity Lexicon(s) from Scratch for Essay Data

Beata Beigman Klebanov¹, Jill Burstein¹, Nitin Madnani¹, Adam Faulkner², and Joel Tetreault¹

1 Educational Testing Service
{bbeigmanklebanov,jburstein,nmadnani,jtetreault}@ets.org
2 Graduate Center, The City University of New York
adamflkr@gmail.com

Abstract. While there are a number of subjectivity lexicons available for research purposes, none can be used commercially. We describe the process of constructing subjectivity lexicon(s) for recognizing sentiment polarity in essays written by test-takers, to be used within a commercial essay-scoring system. We discuss ways of expanding a manually-built seed lexicon using dictionary-based, distributional in-domain and out-of-domain information, as well as using Amazon Mechanical Turk to help “clean up” the expansions. We show the feasibility of constructing a family of subjectivity lexicons from scratch using a combination of methods to attain competitive performance with state-of-art research-only lexicons. Furthermore, this is the first use, to our knowledge, of a paraphrase generation system for expanding a subjectivity lexicon.

Keywords: essay writing, sentiment analysis, sentiment polarity, subjectivity lexicon, C5.0, lexicon expansion, paraphrase generation, thesaurus resources.

1 Introduction

For commercial applications of sentiment analysis, an in-house subjectivity lexicon needs to be constructed, since existing lexicons, such as MPQA [1] and GI [2], are available either for research and education only¹ or under GNU GPL license that disallows the incorporation of the resource into proprietary materials.² In this article, we describe a methodology for creating a family of subjectivity lexicons from scratch through the following phases: (1) a lexicon of about 400 words was manually constructed based on materials in our domain of interest (test-taker essays), (2) a small-scale annotation was conducted to augment the lexicon to 750 words, and (3) a variety of expansion methods with subsequent human and automated clean-up were implemented. We show that this process results in subjectivity lexicons that are comparable to state-of-art lexicons in terms of sentiment classification performance.

¹ “This version of the General Inquirer is made available exclusively for educational and research purposes.” From http://www.wjh.harvard.edu/~inquirer/j1_1/manual
² “The GNU General Public License does not permit incorporating your program into proprietary programs.” From http://www.gnu.org/copyleft/gpl.html
on our data as well as in terms of effective coverage (the number of words in a lexicon that appear in our data).

The article is organized as follows. Section 2 details the process of lexicon construction, starting from the 750-word seed lexicon (section 2.1), then discussing the automatic lexicon expansions (section 2.2), proceeding to the manual clean-up using Amazon Mechanical Turk (AMT) (section 2.3) and automatic clean-up through lexicon combination (section 2.4). Section 3 details the evaluation of the lexicons; the setup for evaluation is described in sections 3.1 and 3.2, section 3.3 compares the lexicons in terms of effective coverage of our data, section 3.4 provides the comparative evaluation of the lexicons on the sentence-level sentiment classification task. Table 4 in section 3.4 presents our main results. Section 4 surveys related work. We discuss our results and conclude in section 5.

2 Building Subjectivity Lexicons

2.1 Seed Lexicon

First, we randomly sample 5,000 essays from a corpus of about 100,000 essays containing writing samples across many topics. Essays were responses to several different writing assignments, including graduate school entrance exams, non-native English speaker proficiency exams, and accounting exams. We manually selected positive and negative sentiment words from the full list of word types in these data; these constitute our Lexicon 0, which contains 407 words.

We sampled 878 sentences containing at least one word from Lexicon 0, thus biasing the sample towards sentiment-bearing sentences. The motivation for the bias was increasing the incidence of sentiment-bearing – positive (POS) and negative (NEG) – sentences, under the assumption that sentiment-bearing sentences had more positive and negative words, and hence, were more effective for lexicon development. Using these sentences, we proceeded with an annotation task as follows. Two research assistants annotated 878 sentences with sentence-level sentiment polarity; 248 of these were also annotated for all words that contribute to the sentiment of the sentence or go against it. We refer to the 248 sentence set as L-1, and to the 630 sentence set L-2. For example, the following sentence was labeled as positive; words contributing to the positive sentiment are bold-faced and words (and phrases) going against it are underlined.

Some may even be impressed that we are confident enough to risk showing a lower net income.

In addition, positive and negative sentences from the T-1 dataset (to be described in section 3.2) were annotated using AMT for words that most contribute to the overall sentiment of the sentence (marking words that go against the dominant sentiment was omitted to simplify the protocol). Each sentence was assigned to 5 AMT annotators; all words marked by at least 3 AMT annotators were selected.

Finally, the Seed Lexicon was created by adding to Lexicon 0 all words marked in L-1 annotations and all words selected from the AMT annotations; the authors then

---

3 When annotators could not attribute sentiment to single words, they marked phrases. Our current lexicons make no use of multi-word annotations.