Automatically generating annotator rationales to improve sentiment classification

Ainur Yessenalina **Yejin Choi** **Claire Cardie**
Department of Computer Science, Cornell University, Ithaca NY, 14853 USA
{ainur, ychoi, cardie}@cs.cornell.edu

Abstract

One of the central challenges in sentiment-based text categorization is that not every portion of a document is equally informative for inferring the overall sentiment of the document. Previous research has shown that enriching the sentiment labels with human annotators' "rationales" can produce substantial improvements in categorization performance (Zaidan et al., 2007). We explore methods to *automatically* generate annotator rationales for document-level sentiment classification. Rather unexpectedly, we find the automatically generated rationales just as helpful as human rationales.

1 Introduction

One of the central challenges in sentiment-based text categorization is that not every portion of a given document is equally informative for inferring its overall sentiment (e.g., Pang and Lee (2004)). Zaidan et al. (2007) address this problem by asking human annotators to mark (at least some of) the relevant text spans that *support each document-level sentiment decision*. The text spans of these "rationales" are then used to construct additional training examples that can guide the learning algorithm toward better categorization models.

But could we perhaps enjoy the performance gains of rationale-enhanced learning models without any additional human effort whatsoever (beyond the document-level sentiment label)? We hypothesize that in the area of sentiment analysis, where there has been a great deal of recent research attention given to various aspects of the task (Pang and Lee, 2008), this might be possible: using existing resources for sentiment analysis, we might be able to construct annotator rationales automatically.

In this paper, we explore a number of methods to automatically generate rationales for document-level sentiment classification. In particular, we investigate the use of off-the-shelf sentiment analysis components and lexicons for this purpose. Our approaches for generating annotator rationales can be viewed as *mostly unsupervised* in that we do not require manually annotated rationales for training.

Rather unexpectedly, our empirical results show that automatically generated rationales (91.78%) are just as good as human rationales (91.61%) for document-level sentiment classification of movie reviews. In addition, complementing the human annotator rationales with automatic rationales boosts the performance even further for this domain, achieving 92.5% accuracy. We further evaluate our rationale-generation approaches on product review data for which human rationales are not available: here we find that even randomly generated rationales can improve the classification accuracy although rationales generated from sentiment resources are not as effective as for movie reviews.

The rest of the paper is organized as follows. We first briefly summarize the SVM-based learning approach of Zaidan et al. (2007) that allows the incorporation of rationales (Section 2). We next introduce three methods for the automatic generation of rationales (Section 3). The experimental results are presented in Section 4, followed by related work (Section 5) and conclusions (Section 6).

2 Contrastive Learning with SVMs

Zaidan et al. (2007) first introduced the notion of *annotator rationales* — text spans highlighted by human annotators as support or evidence for each document-level sentiment decision. These rationales, of course, are only useful if the sentiment categorization algorithm can be extended to exploit the rationales effectively. With this in mind, Zaidan et al. (2007) propose the following *con-*

Proceedings of the ACL 2010 Conference Short Papers, pages 336–341,
Uppsala, Sweden, 11-16 July 2010.

trastive learning extension to the standard SVM learning algorithm.

Let $\vec{x}_i$ be movie review i, and let $\{\vec{r}_{ij}\}$ be the set of *annotator rationales* that support the positive or negative sentiment decision for $\vec{x}_i$. For each such rationale $\vec{r}_{ij}$ in the set, construct a *contrastive training example* $\vec{v}_{ij}$, by removing the text span associated with the rationale $\vec{r}_{ij}$ from the original review $\vec{x}_i$. Intuitively, the contrastive example $\vec{v}_{ij}$ should not be as informative to the learning algorithm as the original review $\vec{x}_i$, since one of the supporting regions identified by the human annotator has been deleted. That is, the *correct* learned model should be *less confident* of its classification of a contrastive example vs. the corresponding original example, and the classification boundary of the model should be modified accordingly.

Zaidan et al. (2007) formulate exactly this intuition as SVM constraints as follows:

$$(\forall i, j):\ y_i\ (\vec{w}\vec{x}_i - \vec{w}\vec{v}_{ij}) \geq \mu(1 - \xi_{ij})$$

where $y_i \in \{-1, +1\}$ is the negative/positive sentiment label of document i, $\vec{w}$ is the weight vector, $\mu \geq 0$ controls the size of the margin between the original examples and the contrastive examples, and ξ_{ij} are the associated slack variables. After some re-writing of the equations, the resulting objective function and constraints for the SVM are as follows:

$$\frac{1}{2}||\vec{w}||^2 + C\sum_i \xi_i + C_{contrast}\sum_{ij}\xi_{ij} \quad (1)$$

subject to constraints:

$$(\forall i):\quad y_i\ \vec{w}\cdot\vec{x}_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$(\forall i, j):\quad y_i\ \vec{w}\cdot\vec{x}_{ij} \geq 1 - \xi_{ij} \quad \xi_{ij} \geq 0$$

where ξ_i and ξ_{ij} are the slack variables for $\vec{x}_i$ (the original examples) and $\vec{x}_{ij}$ ($\vec{x}_{ij}$ are named as *pseudo examples* and defined as $\vec{x}_{ij} = \frac{\vec{x}_i - \vec{v}_{ij}}{\mu}$), respectively. Intuitively, the pseudo examples ($\vec{x}_{ij}$) represent the difference between the original examples ($\vec{x}_i$) and the contrastive examples ($\vec{v}_{ij}$), weighted by a parameter μ. C and $C_{contrast}$ are parameters to control the trade-offs between training errors and margins for the original examples $\vec{x}_i$ and pseudo examples $\vec{x}_{ij}$ respectively. As noted in Zaidan et al. (2007), $C_{contrast}$ values are generally smaller than C for noisy rationales.

In the work described below, we similarly employ Zaidan et al.'s (2007) contrastive learning method to incorporate rationales for document-level sentiment categorization.

3 Automatically Generating Rationales

Our goal in the current work, is to generate annotator rationales automatically. For this, we rely on the following two assumptions:

(1) Regions marked as annotator rationales are more subjective than unmarked regions.

(2) The sentiment of each annotator rationale coincides with the document-level sentiment.

Note that assumption 1 was not observed in the Zaidan et al. (2007) work: annotators were asked only to mark a few rationales, leaving other (also subjective) rationale sections unmarked.

And at first glance, assumption (2) might seem too obvious. But it is important to include as there can be subjective regions with seemingly conflicting sentiment in the same document (Pang et al., 2002). For instance, an author for a movie review might express a positive sentiment toward the movie, while also discussing a negative sentiment toward one of the fictional characters appearing in the movie. This implies that not all subjective regions will be relevant for the document-level sentiment classification — rather only those regions whose polarity matches that of the document should be considered.

In order to extract regions that satisfy the above assumptions, we first look for subjective regions in each document, then filter out those regions that exhibit a sentiment value (i.e., polarity) that conflicts with polarity of the document. Assumption 2 is important as there can be subjective regions with seemingly conflicting sentiment in the same document (Pang et al., 2002).

Because our ultimate goal is to reduce human annotation effort as much as possible, we do not employ supervised learning methods to directly learn to identify good rationales from human-annotated rationales. Instead, we opt for methods that make use of only the document-level sentiment and off-the-shelf utilities that were trained for slightly different sentiment classification tasks using a corpus from a different domain and of a different genre. Although such utilities might not be optimal for our task, we hoped that these basic resources from the research community would constitute an adequate source of sentiment information for our purposes.

We next describe three methods for the automatic acquisition of rationales.

3.1 Contextual Polarity Classification

The first approach employs OpinionFinder (Wilson et al., 2005a), an off-the-shelf opinion analysis utility.[1] In particular, OpinionFinder identifies phrases expressing positive or negative opinions. Because OpinionFinder models the task as a word-based classification problem rather than a sequence tagging task, most of the identified opinion phrases consist of a single word. In general, such short text spans cannot fully incorporate the contextual information relevant to the detection of subjective language (Wilson et al., 2005a). Therefore, we conjecture that good rationales should extend beyond short phrases.[2] For simplicity, we choose to extend OpinionFinder phrases to sentence boundaries.

In addition, to be consistent with our second operating assumption, we keep only those sentences whose polarity coincides with the document-level polarity. In sentences where OpinionFinder marks multiple opinion words with opposite polarities we perform a simple voting — if words with positive (or negative) polarity dominate, then we consider the entire sentence as positive (or negative). We ignore sentences with a tie. Each selected sentence is considered as a separate rationale.

3.2 Polarity Lexicons

Unfortunately, domain shift as well as task mismatch could be a problem with any opinion utility based on supervised learning.[3] Therefore, we next consider an approach that does not rely on supervised learning techniques but instead explores the use of a manually constructed polarity lexicon. In particular, we use the lexicon constructed for Wilson et al. (2005b), which contains about 8000 words. Each entry is assigned one of three polarity values: positive, negative, neutral. We construct rationales from the polarity lexicon for every instance of positive and negative words in the lexicon that appear in the training corpus.

As in the OpinionFinder rationales, we extend the words found by the PolarityLexicon approach to sentence boundaries to incorporate potentially relevant contextual information. We retain as rationales only those sentences whose polarity coincides with the document-level polarity as determined via the voting scheme of Section 3.1.

3.3 Random Selection

Finally, we generate annotator rationales randomly, selecting 25% of the sentences from each document[4] and treating each as a separate rationale.

3.4 Comparison of Automatic vs. Human-annotated Rationales

Before evaluating the performance of the automatically generated rationales, we summarize in Table 1 the differences between automatic vs. human-generated rationales. All computations were performed on the same movie review dataset of Pang and Lee (2004) used in Zaidan et al. (2007). Note, that the Zaidan et al. (2007) annotation guidelines did not insist that annotators mark **all** rationales, only that some were marked for each document. Nevertheless, we report precision, recall, and F-score based on overlap with the human-annotated rationales of Zaidan et al. (2007), so as to demonstrate the degree to which the proposed approaches align with human intuition. Overlap measures were also employed by Zaidan et al. (2007).

As shown in Table 1, the annotator rationales found by OpinionFinder (F-score 49.5%) and the PolarityLexicon approach (F-score 52.6%) match the human rationales much better than those found by random selection (F-score 27.3%).

As expected, OpinionFinder's positive rationales match the human rationales at a significantly lower level (F-score 31.9%) than negative rationales (59.5%). This is due to the fact that OpinionFinder is trained on a dataset biased toward negative sentiment (see Section 3.1 - 3.2). In contrast, all other approaches show a balanced performance for positive and negative rationales vs. human rationales.

4 Experiments

For our contrastive learning experiments we use SVM^{light} (Joachims, 1999). We evaluate the usefulness of automatically generated rationales on

[1] Available at www.cs.pitt.edu/mpqa/opinionfinderrelease/.

[2] This conjecture is indirectly confirmed by the fact that human-annotated rationales are rarely a single word.

[3] It is worthwhile to note that OpinionFinder is trained on a newswire corpus whose prevailing sentiment is known to be negative (Wiebe et al., 2005). Furthermore, OpinionFinder is trained for a task (word-level sentiment classification) that is different from marking annotator rationales (sequence tagging or text segmentation).

[4] We chose the value of 25% to match the percentage of sentences per document, on average, that contain human-annotated rationales in our dataset (24.7%).

Method	% of sentences selected	Precision			Recall			F-Score		
		ALL	POS	NEG	ALL	POS	NEG	ALL	POS	NEG
OPINIONFINDER	22.8%	54.9	56.1	54.6	45.1	22.3	65.3	49.5	31.9	59.5
POLARITYLEXICON	38.7%	45.2	42.7	48.5	63.0	71.8	55.0	52.6	53.5	51.6
RANDOM	25.0%	28.9	26.0	31.8	25.9	24.9	26.7	27.3	25.5	29.0

Table 1: Comparison of Automatic vs. Human-annotated Rationales.

five different datasets. The first is the movie review data of Pang and Lee (2004), which was manually annotated with rationales by Zaidan et al. (2007)[5]; the remaining are four product review datasets from Blitzer et al. (2007).[6] Only the movie review dataset contains human annotator rationales. We replicate the same feature set and experimental set-up as in Zaidan et al. (2007) to facilitate comparison with their work.[7]

The contrastive learning method introduced in Zaidan et al. (2007) requires three parameters: (C, μ, $C_{contrast}$). To set the parameters, we use a grid search with step 0.1 for the range of values of each parameter around the point (1,1,1). In total, we try around 3000 different parameter triplets for each type of rationales.

4.1 Experiments with the Movie Review Data

We follow Zaidan et al. (2007) for the training/test data splits. The top half of Table 2 shows the performance of a system trained with **no annotator rationales** vs. two variations of human annotator rationales. HUMANR treats each rationale in the same way as Zaidan et al. (2007). HUMANR@SENTENCE extends the human annotator rationales to sentence boundaries, and then treats each such sentence as a separate rationale. As shown in Table 2, we get almost the same performance from these two variations (91.33% and 91.61%).[8] This result demonstrates that locking rationales to sentence boundaries was a reasonable choice.

Method	Accuracy
NORATIONALES	88.56
HUMANR	91.61$^{\bullet}$
HUMANR@SENTENCE	91.33$^{\bullet}$ †
OPINIONFINDER	91.78$^{\bullet}$ †
POLARITYLEXICON	91.39$^{\bullet}$ †
RANDOM	90.00*
OPINIONFINDER+HUMANR@SENTENCE	92.50$^{\bullet}$ $^{\triangle}$

Table 2: Experimental results for the movie review data.

– The numbers marked with $^{\bullet}$ (or *) are statistically significantly better than NORATIONALES according to a paired t-test with $p < 0.001$ (or $p < 0.01$).
– The numbers marked with $^{\triangle}$ are statistically significantly better than HUMANR according to a paired t-test with $p < 0.01$.
– The numbers marked with † are ***not*** statistically significantly worse than HUMANR according to a paired t-test with $p > 0.1$.

Among the approaches that make use of only automatic rationales (bottom half of Table 2), the best is OPINIONFINDER, reaching 91.78% accuracy. This result is slightly better than results exploiting human rationales (91.33-91.61%), although the difference is not statistically significant. This result demonstrates that automatically generated rationales are just as good as human rationales in improving document-level sentiment classification. Similarly strong results are obtained from the POLARITYLEXICON as well.

Rather unexpectedly, RANDOM also achieves statistically significant improvement over NORATIONALES (90.0% vs. 88.56%). However, notice that the performance of RANDOM is statistically significantly lower than those based on human rationales (91.33-91.61%).

In our experiments so far, we observed that some of the automatic rationales are just as good as human rationales in improving the document-level sentiment classification. Could we perhaps achieve an even better result if we combine the automatic rationales with human

[5] Available at http://www.cs.jhu.edu/∼ozaidan/rationales/.

[6] http://www.cs.jhu.edu/∼mdredze/datasets/sentiment/.

[7] We use binary unigram features corresponding to the unstemmed words or punctuation marks with count greater or equal to 4 in the full 2000 documents, then we normalize the examples to the unit length. When computing the pseudo examples $\vec{x}_{ij} = \frac{\vec{x}_i - \vec{v}_{ij}}{\mu}$ we first compute $(\vec{x_i} - \vec{v}_{ij})$ using the binary representation. As a result, features (unigrams) that appeared in both vectors will be zeroed out in the resulting vector. We then normalize the resulting vector to a unit vector.

[8] The performance of HUMANR reported by Zaidan et al. (2007) is 92.2% which lies between the performance we get (91.61%) and the oracle accuracy we get if we knew the best parameters for the test set (92.67%).

rationales? The answer is yes! The accuracy of OPINIONFINDER+HUMANR@SENTENCE reaches 92.50%, which is statistically significantly better than HUMANR (91.61%). In other words, not only can our automatically generated rationales replace human rationales, but they can also improve upon human rationales when they are available.

4.2 Experiments with the Product Reviews

We next evaluate our approaches on datasets for which human annotator rationales do not exist. For this, we use some of the product review data from Blitzer et al. (2007): reviews for Books, DVDs, Videos and Kitchen appliances. Each dataset contains 1000 positive and 1000 negative reviews. The reviews, however, are substantially shorter than those in the movie review dataset: the average number of sentences in each review is 9.20/9.13/8.12/6.37 respectively vs. 30.86 for the movie reviews. We perform 10-fold cross-validation, where 8 folds are used for training, 1 fold for tuning parameters, and 1 fold for testing.

Table 3 shows the results. Rationale-based methods perform statistically significantly better than NORATIONALES for all but the Kitchen dataset. An interesting trend in product review datasets is that RANDOM rationales are just as good as other more sophisticated rationales. We suspect that this is because product reviews are generally shorter and more focused than the movie reviews, thereby any randomly selected sentence is likely to be a good rationale. Quantitatively, subjective sentences in the product reviews amount to 78% (McDonald et al., 2007), while subjective sentences in the movie review dataset are only about 25% (Mao and Lebanon, 2006).

4.3 Examples of Annotator Rationales

In this section, we examine an example to compare the automatically generated rationales (using OPINIONFINDER) with human annotator rationales for the movie review data. In the following positive document snippet, automatic rationales are underlined, while **human-annotated rationales** are in bold face.

...But a little niceness goes a long way these days, and **there's no denying the entertainment value** of that thing you do! **It's just about impossible to hate**. **It's an inoffensive, enjoyable piece** of nostalgia that is **sure to leave audiences smiling and humming, if not singing**, "that thing you do!" –quite possibly for days...

Method	Books	DVDs	Videos	Kitchen
NORATIONALES	80.20	80.95	82.40	87.40
OPINIONFINDER	81.65*	82.35*	84.00*	88.40
POLARITYLEXICON	82.75•	82.85•	84.55•	87.90
RANDOM	82.05•	82.10•	84.15•	88.00

Table 3: Experimental results for subset of Product Review data

– The numbers marked with • (or *) are statistically significantly better than NORATIONALES according to a paired t-test with $p < 0.05$ (or $p < 0.08$).

Notice that, although OPINIONFINDER misses some human rationales, it avoids the inclusion of "impossible to hate", which contains only negative terms and is likely to be confusing for the contrastive learner.

5 Related Work

In broad terms, constructing annotator rationales automatically and using them to formulate contrastive examples can be viewed as learning with prior knowledge (e.g., Schapire et al. (2002), Wu and Srihari (2004)). In our task, the prior knowledge corresponds to our operating assumptions given in Section 3. Those assumptions can be loosely connected to recognizing and exploiting discourse structure (e.g., Pang and Lee (2004), Taboada et al. (2009)). Our automatically generated rationales can be potentially combined with other learning frameworks that can exploit annotator rationales, such as Zaidan and Eisner (2008).

6 Conclusions

In this paper, we explore methods to automatically generate annotator rationales for document-level sentiment classification. Our study is motivated by the desire to retain the performance gains of rationale-enhanced learning models while eliminating the need for additional human annotation effort. By employing existing resources for sentiment analysis, we can create automatic annotator rationales that are as good as human annotator rationales in improving document-level sentiment classification.

Acknowledgments

We thank anonymous reviewers for their comments. This work was supported in part by National Science Foundation Grants BCS-0904822, BCS-0624277, IIS-0535099 and by the Department of Homeland Security under ONR Grant N0014-07-1-0152.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June. Association for Computational Linguistics.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. pages 169–184.

Yi Mao and Guy Lebanon. 2006. Sequential models for sentiment prediction. In *Proceedings of the ICML Workshop: Learning in Structured Output Spaces Open Problems in Statistical Relational Learning Statistical Network Analysis: Models, Issues and New Directions*.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 432–439, Prague, Czech Republic, June. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271, Morristown, NJ, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Robert E. Schapire, Marie Rochery, Mazin G. Rahim, and Narendra Gupta. 2002. Incorporating prior knowledge into boosting. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 538–545, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Maite Taboada, Julian Brooke, and Manfred Stede. 2009. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference*, pages 62–70, London, UK, September. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005a. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, Morristown, NJ, USA. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP '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.

Xiaoyun Wu and Rohini Srihari. 2004. Incorporating prior knowledge with weighted margin support vector machines. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 326–333, New York, NY, USA. ACM.

Omar F. Zaidan and Jason Eisner. 2008. Modeling annotators: a generative approach to learning from annotator rationales. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Morristown, NJ, USA. Association for Computational Linguistics.

Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL HLT 2007; Proceedings of the Main Conference*, pages 260–267, April.