Supplementary Material: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
Stanford University, Stanford, CA 94305, USA
richard@socher.org,{aperelyg,jcchuang,ang}@cs.stanford.edu
{jeaneis,manning,cgpotts}@stanford.edu

Overview

This supplementary material contains four sections. The first section details the design and methods we used to create the Sentiment Treebank. The second provides an analysis of how much this treebank helps performance on recursive models. The third section gives several examples of the data we performed the negation analysis on and the fourth shows the most negative/positive n-grams in the dev set selected by the RNTN.

1 The Stanford Sentiment Treebank

Recursive Neural Networks are based on the idea of learning compositional semantics. For sentiment analysis the pervasive annotation scheme however is to have only a single label for a sentence or entire document. Such an annotation scheme makes training for compositional effects hard and evaluating them close to impossible. Hence, we collected a dataset that explicitly represents the compositional semantic structure present in sentences. The Stanford Sentiment Treebank introduced in the accompanied paper gives a new level of detail by providing labels for every single node of the top parse tree. We used the corpus of movie review excerpts from the Rotten Tomatoes website originally collected and published by Pang and Lee (2005) for this purpose. This set was chosen because several methods for sentiment polarity classification had been compared on it previously (Nakagawa et al., 2010; Socher et al., 2011).

1.1 Text Pre-processing

The models we present in this work are built on top of parse trees. To achieve the most accurate parse of a sentence, we pre-process the original html files[1] that Pang and Lee (2005) collected from Rotten Tomatoes website. The main problems with the already processed documents were:

1. Missing capitalization. While downcasing a dataset helps remove sparseness, it could be problematic for a movie review corpus where mentions of proper nouns are ubiquitous. In addition, the parser we use is case-sensitive.

2. Tokenization. The original dataset is tokenized on whitespace. However, to facilitate the use of the Stanford Parser, input texts need to be tokenized in the Treebank-style.

3. HTML markup. The original dataset contains remnants of HTML mark-up.

We processed the data as follows:

1. We scraped the same 10,662 sentences (5,331 sentences for each polarity according to the original label decision) that appeared in sentence polarity dataset v.1[1].

2. HTML tags were isolated, and HTML characters such as `<` and `&#8217;` were mapped to a corresponding unicode symbol.

3. We maintained all accented characters

4. We filled in the missing and mismatched quotation marks that were left out from the excerpt. For example, the following review: *Red Dragon" never cuts corners.* would be corrected to *"Red Dragon" never cuts corners.*

[1] www.cs.cornell.edu/people/pabo/movie-review-data/

5. There were three one-word review snippets to which we added a sentence final punctuation mark, in order to obtain a proper binarized parse tree.

6. Marked curses like *'a**holes'* are kept as they appear.

7. We removed 57 reviews that were written entirely in a foreign language.

1.2 Parsing

Once the review texts were prepared through the aforementioned steps, we proceeded to obtain a binarized PCFG parses using the Stanford Parser (Klein and Manning, 2003). Some review snippets may be composed of multiple sentences, thus resulted in multiple trees. We obtained 11,855 parse trees from the 10,605 prepared review snippets. These consist of 215,154 unique constituents in total.

1.3 Crowdsourcing for Sentiment Annotation

Pang and Lee's dataset assumes the snippets marked as "fresh" by the Rotten Tomatoes website are positive, and "rotten" are negative. However, these labels do not always accurately reflect the opinion represented by the corresponding snippet. Take for example the following excerpt:

"Ultimately feels empty and unsatisfying, like swallowing a Communion wafer without the wine."

This snippet was marked as "fresh" by the Rotten Tomatoes website because the corresponding full-length review gave this movie a score of 3 out of 5 possible stars (with a half-star increment.) However, most people would agree that the extracted snippet does not suggest a positive opinion. Additionally, binary values are not expressive enough to capture the opinion distribution of these review snippets. Consider the following two snippets, both given a positive label:

1. *"One of the greatest family-oriented, fantasy-adventure movies ever."*

2. *"Light, cute and forgettable."*

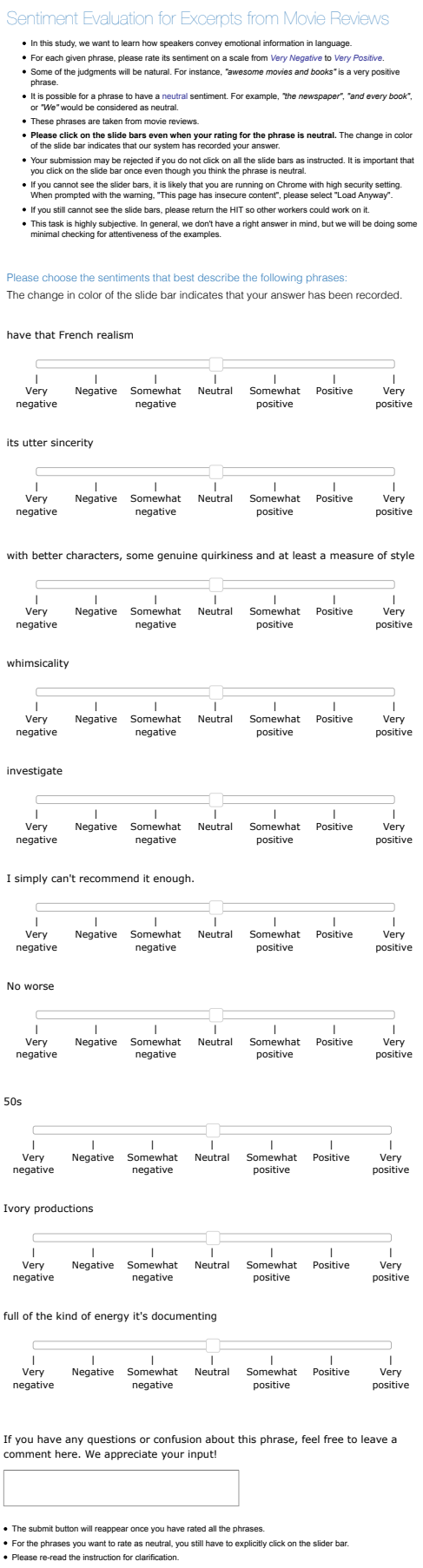

Figure 1: A sample of a complete task seen by a human annotator.

The first snippet, for which we collected an average rating of 0.93 (out of 1) was mapped to *very positive*. It clearly indicates a stronger positive sentiment than the latter, where the average rating is closer to *positive or even neutral* with a positivity score of 0.6.

We utilized Amazon Mechanical Turk's crowdsourcing platform to collect a finer sentiment annotation. In addition to the sentiment score of the sentence, we also collect the score for the phrases spanned by every node of the parse tree.

Note that this also differs from the previous labels which were given by the writer of the review. In contrast, our labels come from the reader and their perceived sentiment of the phrases and sentences.

We use Amazon Mechanical Turk to annotate all 215,154 unique phrases obtained from the parse trees. Each annotation task consists of 10 different phrases chosen at random. It was then independently shown to 3 human annotators, who were asked to rate each phrase using a multi-stop slider bar. Each of these slider bars has 7 tick marks with 3 soft stops in-between, and is initially set to neutral. Fig. 1 shows an an example of a complete annotation task. These collected annotations are mapped to numerical values ranging from 1 (*very negative*) to 25 (*very positive*). The average variance of the collected data is 9.7238. See Fig. 2 for an illustration of a parse tree with the human annotated sentiment labels.

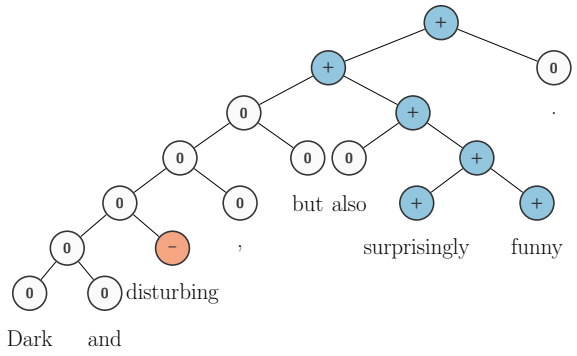

Figure 2: A sample of an annotated parse tree.

1.4 Corpus Composition

The Sentiment Treebank we created from the method described above consists of 11,855 sentences and 215,154 unique phrases. The distribution of sentiment labels is shown in Fig. 3. The sentiments at sentence-level are bimodal, as most reviews tend to have have a clear positive or negative sentiment. In contrast, a large fraction of the sentiments across all phrases appear to be neutral—examples include words like *"but"*, *"the"*, *"and"*, as well as longer phrase that have no context, such as *"his earlier English-language movie"*.

This new dataset has the potential for a much finer analysis and allows classification with multiple classes as well as binary labels.

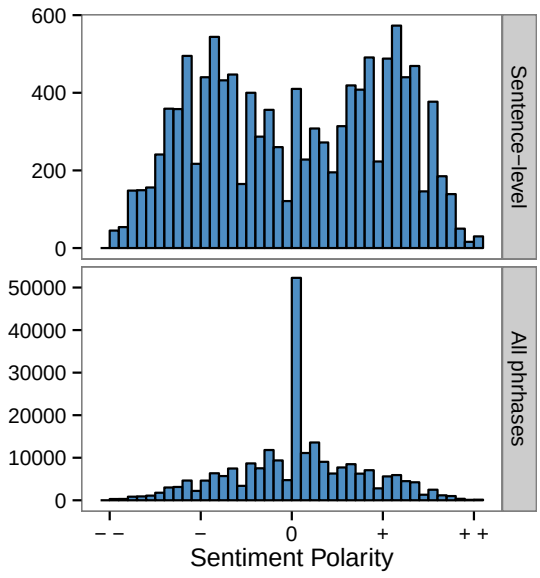

Figure 3: Histograms of human-annotated sentiment labels ranging from *most negative* (– –) to *most positive* (+ +). **Top:** Labels of the sentence-level nodes. Notice that the distribution is bimodal, indicating that sentences tend to express a clear sentiment polarity. **Bottom:** Sentiment labels across all phrases in the dataset. This distribution has a peak in the middle, indicating a large percentage of phrases with no strong polarity sentiment.

2 Effect of Annotation Level on Model Performance

Previous approaches, including previous recursive neural models, used datasets consisting only of sentence-level annotations of sentiment values (Socher et al., 2011). To properly compare the added value of having a dataset with an explicit annotation of compositionality effects, we trained our best neural models with only sentence-level annotations and

Model	Positive/Negative Accuracy	
	Sentence-level	All Node
RNN	78.2	82.4
MV-RNN	80.0	82.9
RNTN	79.8	85.4

Table 1: Comparison of positive/negative sentiment prediction accuracy for neural models trained with only sentence-level annotations and trained with all annotations.

compared the results with those obtained by training with the full set of annotations. The model trained with sentence-only annotation achieved worse performance than the models trained with the new treebank. The RNTN and MV-RNN get around 80% performance. Table 1 shows the accuracy of positive/negative classification for all of the neural models with sentence-level and all-node sentiment annotations. The significant difference is likely due to the inability of the neural models to properly learn compositional effects given only sentential annotations as there is no way to explicitly represent how the different parts of a sentence interact. With the fully labeled dataset, the neural models can learn at all levels of the parse tree and compose the meaning of the entire sentence much better.

3 Negation Dataset

We additionally constructed a special dataset of sentences and their negations for the purpose of determining how well the various models were able to detect their linguistic effects, as described in Section 5 of the main paper. This dataset consisted of a set of 21 positive and 21 negative sentences from the dev set. The sentences are modified to be easily negatable, and each sentence comes with its negation. Table 2 shows several example sentences from this dataset. Note, that in some cases just one or two words differ, yet the meaning is much more negative.

4 Querying for Most Positive and Negative Phrases

We queried the model for its predictions on what the most positive or negative n-grams are in the dev set, measured as the highest activation of the most negative and most positive classes. Table 3 shows the most positive and negative phrases the RNTN could find in the dev set, all of which indeed depict strong sentiment.

References

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *ACL*.

T. Nakagawa, K. Inui, and S. Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *NAACL*, HLT.

B. Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, pages 115–124.

R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *EMNLP*.

Type	Original	Negated
Positive	I just loved every minute of this film.	I didn't like a single minute of this film.
	A strangely compelling and brilliantly acted psychological drama.	A strangely uncompelling and not brilliantly acted psychological drama.
	Preaches to two completely different choirs at the same time, which is a pretty amazing accomplishment.	Preaches to two completely different choirs at the same time, which is a failed accomplishment.
	If you enjoy more thoughtful comedies with interesting conflicted characters; this one is for you.	If you enjoy more thoughtful comedies with interesting conflicted characters; this one is not for you.
	It provides the grand, intelligent entertainment of a superior cast.	It does not provide the grand, intelligent entertainment of a superior cast.
	Like Mike is a winner for kids, and no doubt a winner for Lil Bow Wow.	Like Mike is not a winner for kids, nor a winner for Lil Bow Wow.
	Roger Dodger is one of the most compelling variations on this theme.	Roger Dodger is one of the least compelling variations on this theme.
	It's witty and inventive, and in hindsight, it isn't even all that dumb.	It's not witty or inventive and it is pretty dumb.
	One of the most significant moviegoing pleasures of the year.	Not one of the most significant moviegoing pleasures of the year.
	Easily my choice for one of the year's best films.	Definitely not my choice for one of the year's best films.
Negative	An instant candidate for the worst movie of the year.	Not a candidate for the worst movie of the year.
	The film seems a dead weight.	The film doesn't seem like a dead weight.
	I found it slow, drab, and melodramatic.	I didn't find it slow, drab, or melodramatic.
	The picture failed to capture me.	The picture didn't fail to capture me.
	It's a dumb action movie.	It's not a dumb action movie.
	The story is stupid and obvious.	The story is not stupid or obvious.
	The dialogue in this movie is rambling and repetitive.	The dialogue in this movie is neither rambling nor repetitive.
	Suffers from the visual drabness endemic to digital video.	Doesn't suffer from the visual drabness endemic to digital video.
	All of the characters are predictable stereotypes.	None of the characters are predictable stereotypes.
	It's just incredibly dull.	It's definitely not dull.

Table 2: Examples from our negation experiment set. Top: positive sentences that we negated. Bottom: negative sentences that were themselves negated.

n	**Most positive n-grams**	**Most negative n-grams**
1	engaging ; best ; powerful ; love ; beautiful ; entertaining ; clever ; terrific ; excellent ; great ;	bad ; dull ; boring ; fails ; worst ; stupid ; painfully ; cheap ; forgettable ; disaster ;
2	excellent performances ; amazing performance ; terrific performances ; A masterpiece ; masterful film ; wonderful film ; terrific performance ; masterful piece ; wonderful movie ; marvelous performances ;	worst movie ; bad movie ; very bad ; shapeless mess ; worst thing ; tepid waste ; instantly forgettable ; bad film ; extremely bad ; complete failure ;
3	an amazing performance ; a terrific performance ; a wonderful film ; wonderful all-ages triumph ; A masterful film ; a wonderful movie ; a tremendous performance ; drawn excellent performances ; most visually stunning ; A stunning piece ;	for worst movie ; A lousy movie ; most joyless movie ; a complete failure ; another bad movie ; fairly terrible movie ; a bad movie ; extremely unfunny film ; most painfully marginal ; very bad sign ;
5	nicely acted and beautifully shot ; gorgeous imagery , effective performances ; the best of the year ; a terrific American sports movie ; very solid , very watchable ; a fine documentary does best ; refreshingly honest and ultimately touching ;	silliest and most incoherent movie ; completely crass and forgettable movie ; just another bad movie . ; drowns out the lousy dialogue ; a fairly terrible movie ... ; A cumbersome and cliche-ridden movie ; a humorless , disjointed mess ;
8	one of the best films of the year ; simply the best family film of the year ; the best film of the year so far ; A love for films shines through each frame ; created a masterful piece of artistry right here ; A masterful film from a master filmmaker , ; 's easily his finest American film ... comes ;	A trashy , exploitative , thoroughly unpleasant experience ; this sloppy drama is an empty vessel . ; a meandering , inarticulate and ultimately disappointing film ; an unimaginative , nasty , glibly cynical piece ; bad , he 's really bad , and ; quickly drags on becoming boring and predictable . ; be the worst special-effects creation of the year ;

Table 3: Examples of n-grams ($n = 1, 2, 3, 5, 8$) from the dev set of the corpus for which our model predicts the most positive and most negative responses.