Identifying Expressions of Opinion in Context*

Eric Breck and Yejin Choi and Claire Cardie
Department of Computer Science
Cornell University
{ebreck,ychoi,cardie}@cs.cornell.edu

Abstract

While traditional information extraction systems have been built to answer questions about facts, subjective information extraction systems will answer questions about feelings and opinions. A crucial step towards this goal is identifying the words and phrases that express opinions in text. Indeed, although much previous work has relied on the identification of opinion expressions for a variety of sentiment-based NLP tasks, none has focused directly on this important supporting task. Moreover, none of the proposed methods for identification of opinion expressions has been evaluated at the task that they were designed to perform. We present an approach for identifying opinion expressions that uses conditional random fields and we evaluate the approach at the expression-level using a standard sentiment corpus. Our approach achieves expression-level performance within 5% of the human interannotator agreement.

1 Introduction

Traditional information extraction tasks can be viewed as beginning with a set of questions about facts, such as *who?*, *where?*, or *how many?*. Researchers then build systems to extract the answers to these questions from text [MUC7, 1998; NIS, 2005; DUC2005, 2005]. More recently, researchers have investigated the construction of language-processing systems that extract answers to subjective questions, such as *how does X feel about Y?* (see Section 4). Intelligence analysts, for example, could use such opinion-oriented systems to monitor trouble spots around the world while marketing researchers could use them to understand public opinion about their product.

As with factual information extraction and question answering, subjective information extraction and question answering will require techniques to analyze text below the sentence level. For example, Stoyanov et al. [2005] show that identifying opinion expressions is helpful in localizing the answers to opinion questions.

*This work was supported by the Advanced Research and Development Activity (ARDA), by NSF Grants IIS-0535099 and IIS-0208028, and by gifts from Google and the Xerox Foundation.

Consider the following sentences, in which we denote two kinds of opinion expression in boldface and italic (described below).

1: Minister Vedrine **criticized** the White House reaction.

2: 17 persons were killed by sharpshooters **faithful to** the president.

3: Tsvangirai **said** the election result was "*illegitimate*" and a clear case of "*highway robbery*".

4: Criminals have been *preying* on Korean travellers in China.

To understand, extract, or answer questions about the opinions in these sentences a system must (minimally) determine the basic attributes of the opinion: Is its polarity positive, negative, or neutral? With what strength or intensity is the opinion expressed: mild, medium, strong or extreme? Who or what is the source, or holder, of the opinion? What is its target, i.e. what is the opinion about? The opinion expressions marked in the sentences above are the key to answering these questions. In particular, the marked phrases denote the polarity of the opinion: for example, "criticized" and "faithful to" (examples 1 and 2) denote negative and positive attitudes, respectively. The opinion expressions also often provide linguistic anchors for the automatic extraction of the source and target of the opinion. The predicate "criticized", for example, organizes the semantic roles that denote the source of the opinion (the agent role = "Minister Vedrine") and the target of the opinion (the object/theme role = "White House reaction"). Similarly, the opinion expression "faithful to" organizes the semantic roles associated with the source (the agent role = "sharpshooters") and the target (the object/theme role = "the president") of the opinion in example 2.

Wiebe et al. [2005] distinguish two types of opinion expressions, and we follow their definitions here. *Direct subjective expressions* (DSEs), shown in boldface, are spans of text that explicitly express an attitude or opinion. "Criticized" and "faithful to" (examples 1 and 2), for example, directly denote negative and positive attitudes towards the "White House reaction" and "the president", respectively. Speech events like "said" in example 3 can be DSEs if they express subjectivity. In contrast, *expressive subjective elements* (ESEs), shown in italics, are spans of text that indicate, merely by the specific choice of words, a degree of subjectivity on the part of the speaker. The phrases "illegitimate" and "highway robbery",

IOB . . . **faithful/B to/I** the/O president/O ./O
IO . . . **faithful/I to/I** the/O president/O ./O

Figure 1: How to encode the class variable: The IOB method and the IO method. IO is used in this paper.

for example, indirectly relay Tsvangirai's negative opinion of "the election result" (example 3), and the use of "preying on" (instead of, say, "mugging") indicates the writer's sympathy for the Korean travellers in example 4.

While some previous work identifies opinion expressions in support of sentence- or clause-level subjectivity classification [Riloff and Wiebe, 2003; Wiebe and Wilson, 2002], none has directly tackled the problem of opinion expression identification, developed methods for the task, and evaluated performance at the expression level. Instead, previous work in this area focuses its evaluation on the sentence-level subjectivity classification task.

In this work, we treat the identification of opinion expressions as a tagging task, and use conditional random fields to address this problem. For comparison, we chose two lists of clues to subjectivity from previous work [Wilson *et al.*, 2005b; Wiebe and Riloff, 2005]. These clues have previously been evaluated only for their utility in clause- or sentence-level classification. Here we interpret the clues as expressions of opinion and compare them to our results. We achieve F-measures of 63.4% for ESEs and 70.6% for DSEs, within 5% of the human interannotator agreement for DSEs and within 10% for ESEs.

The rest of this paper proceeds as follows. In Section 2, we discuss our approach to identifying direct subjective expressions and expressive subjective elements. In Section 3, we present experimental results. In Section 4, we present related work and in Section 5, we conclude and discuss future work.

2 The approach

As the example sentences in Section 1 suggest, identifying subjective expressions is a difficult task. The expressions can vary in length from one word to over twenty words. They may be verb phrases, noun phrases, or strings of words that do not correspond to any linguistic constituent. Subjectivity is a realm of expression where writers get quite creative, so no short fixed list can capture all expressions of interest. Also, an expression which is subjective in one context is not always subjective in another context [Wiebe and Wilson, 2002]. Therefore, we present in this section our machine learning approach for the identification of direct subjective expressions and expressive-subjective elements. We treat the task as a tagging problem, and use conditional random fields [Lafferty *et al.*, 2001][1]. Below we discuss the encoding of the class variable (Section 2.1), the features (Section 2.2), and the learning method (Section 2.3).

[1]CRFs allow the use of a large, diverse set of features, while offering the choice of modeling individual tokens or sequences. Margin-based methods would be another natural option, but pilot experiments on development data suggested the performance of SVMs was inferior to CRFs for this task.

2.1 The class variable

A common encoding for extent-identification tasks such as named entity recognition is the so-called IOB encoding. Using IOB, each word is tagged as either Beginning an entity, being In an entity (i.e. an opinion expression), or being Outside of an entity (see Figure 1). While we initially used this encoding, preliminary investigation on separate development data revealed that a simpler binary encoding produces better results. We suspect this is because it is rarely the case in our data that two entities are adjacent, and so the simpler model is easier to fit. Thus, we tag each token as being either In an entity or Outside of an entity. When predicting, a sequence of consecutive tokens tagged as In constitutes a single predicted entity.

2.2 Features

In this section, we describe the features used in our model. We include features to allow the model to learn at various levels of generality. We include lexical features to capture specific phrases, local syntactic features to learn syntactic context, and dictionary-based features to capture both more general patterns and expressions already known to be opinion-related. The same feature set is used for identifying both types of subjective expression. For pedagogical reasons, we present the features as categorically valued, but in our model we encode all features in binary (i.e. a feature (f, v) is 1 for a token t if $f(t) = v$, and 0 otherwise).

Lexical features We include features lex_i, defined to be the word at position i relative to the current token. We include $lex_{-4}, lex_{-3}, \ldots lex_4$. These are encoded into about 18,000 binary features per position (i.e. the vocabulary size).

Syntactic features We include a feature *part-of-speech*, defined to be the part of speech of the current token according to the GATE part-of-speech tagger [Cunningham *et al.*, 2002] (encoded into 45 binary features). We also include three features *prev, cur,* and *next*, defined to be the previous, current, or following constituent type, respectively, according to the CASS partial parser [Abney, 1996][2]. These are encoded into about 100 binary features each.

Dictionary-based features We include features from four sources. We include a feature *WordNet*, defined to be all synsets which are hypernyms of the current token in the WordNet hierarchy [Miller, 1995]. This is encoded into 29,989 binary features, many of which may be 1 for a given token. We include a feature *Levin*, defined to be the section of Levin's [1993] categorization of English verbs in which a given verb appears, and a feature *Framenet*, defined to be the category of a word in the categorization of nouns and verbs in Framenet[3]. Finally, we include a feature that specifically targets subjective expressions. Wilson et al. [2005b] identify

[2]Cass is available for download at `http://www.vinartus.net/spa/`.

[3]`http://www.icsi.berkeley.edu/~framenet/`

number of sentences	8297
number of DSEs	6712
number of ESEs	8640
average length of DSEs	1.86 words
average length of ESEs	3.33 words

Table 1: Statistics for test data

a set of clues as being either strong or weak cues to the subjectivity of a clause or sentence. We identify any sequence of tokens included on this list, and then define a feature *Wilson* that returns the value '-' if the current token is not in any recognized clue, or **strong** or **weak** if the current token is in a recognized clue of that strength. This clue is encoded into two binary features (the '-' case is not encoded).

2.3 The learning method

We chose to use a linear-chain conditional random field model for all of our experiments, using the MALLET toolkit [McCallum, 2002]. This discriminatively-trained sequence model has been found to perform extremely well on tagging tasks such as ours [Lafferty *et al.*, 2001]. Based on pilot experiments on development data, we chose a Gaussian prior of 0.25.

3 Experiments

In this section, we describe the data and evaluations used in our experiments, describe the baselines we compare to, and present our results.

3.1 Data

The MPQA corpus [Wiebe *et al.*, 2005][4] consists of 535 newswire documents annotated with a variety of annotations of interest for subjectivity research. In particular, all DSEs and ESEs in the documents have been manually identified. In this work, we used 135 documents for development of our features and determination of parameters, and kept the remaining 400 blind for evaluation. All of our evaluation results use 10-fold cross-validation on the 400 documents. Table 1 presents some statistics on these 400 documents.

3.2 Evaluation

As with other information extraction tasks, we use precision, recall and F-measure to evaluate our method's performance. Precision is $\frac{|C \cap P|}{|P|}$ and recall is $\frac{|C \cap P|}{|C|}$, where C and P are the sets of correct and predicted expression spans, respectively. F is the harmonic mean of precision and recall, $\frac{2PR}{P+R}$. Our method often identifies expressions that are close to, but not precisely the same as, the manually identified expressions. For example, for the expression "roundly criticized," our method might only identify "criticized." We therefore introduce softened variants of precision and recall as follows. We define soft precision as $SP^a = \frac{|\{p|p \in P \wedge \exists\ c \in C\ s.t.\ a(c,p)\}|}{|P|}$ and soft recall as $SR^a = \frac{|\{c|c \in C \wedge \exists\ p \in P\ s.t.\ a(c,p)\}|}{|C|}$, where $a(c,p)$ is a predicate true just when expression c "aligns" to expression p in a sense defined by a. We report results according to two predicates: $exact$ and $overlap$. $exact(c,p)$ is true just when c and p are the same spans - this yields the usual notions of precision and recall. A softer notion is produced by the predicate $overlap(c,p)$, which is true when the spans of c and p overlap[5].

3.3 Baselines

For baselines, we compare to two dictionaries of subjectivity clues identified by previous work [Wilson *et al.*, 2005b; Wiebe and Riloff, 2005]. These clues were collected to help recognize subjectivity at the sentence or clause level, not at the expression level, but the clues often correspond to subjective expressions. Each clue is one to six consecutive tokens, possibly allowing for a gap, and matching either stemmed or unstemmed tokens, possibly of a fixed part of speech. In the following experiments, we report results of the *Wiebe* baseline, which predicts any sequence of tokens matching a clue from Wiebe and Riloff [2005] to be a subjective expression, and the *Wilson* baseline, using similar predictions based on clues from Wilson et al. [2005b]. When predicting DSEs, we remove all clues from the list which never match a DSE in the test data, to make the baseline's precision as high as possible (although since many potentially subjective expressions are often not subjective, the precision is still quite low). We similarly trim the lists when predicting the other targets below. Apart from this trimming, the lists were not derived from the MPQA corpus. Note that the higher-performing of these two baselines, from Wilson et al. [2005b], was incorporated into the feature set used in our CRF models[6].

3.4 Results

Tables 2 and 3 present experimental results on identifying direct subjective expressions and expressive subjective elements in several settings, as well as presenting the two baselines for comparison purposes. We experiment with two variants of conditional random fields, one with potentials (features) for Markov order 1+0 (similar to the features in a hidden Markov model, labeled crf-1 in the tables), and one with features only for order 0 (equivalent to a maximum entropy

[4] Available at `http://www.cs.pitt.edu/mpqa/`. We use version 1.1 of the corpus. Code and data used in our experiments are available at `http://www.cs.cornell.edu/~ebreck/breck07identifying`.

[5] A potential issue with soft precision and recall is that the measures may drastically overestimate the system's performance. A system predicting a single entity overlapping with every token of a document would achieve 100% soft precision and recall with the overlap predicate. We can ensure this does not happen by measuring the average number of expressions to which each correct or predicted expression is aligned (excluding expressions not aligned at all). In our data, this does not exceed 1.13, so we can conclude these evaluation measures are behaving reasonably.

[6] The CRF features based on the Wilson dictionary were based on the entire dictionary, including clues not relevant for the particular problem being tested. Also, the choice to use only the Wilson dictionary and not the Wiebe for features was made during development of the model on a separate development dataset. So the model tested was in no way developed using the test data.

	overlap			exact		
method	recall	precision	F	recall	precision	F
Wiebe baseline	$45.69^{2.4}$	$31.10^{2.5}$	$36.97^{2.3}$	$21.52^{1.8}$	$13.91^{1.4}$	$16.87^{1.4}$
Wilson baseline	$55.15^{2.2}$	$30.73^{1.9}$	$39.44^{1.9}$	$25.65^{1.7}$	$13.32^{1.0}$	$17.52^{1.2}$
crf-1-DSE	$60.22^{1.8}$	$\mathbf{79.34^{3.2}}$	$68.44^{2.0}$	$42.65^{2.9}$	$\mathbf{57.65^{2.8}}$	$\mathbf{49.01^{2.8}}$
crf-1-DSE&ESE	$62.73^{2.3}$	$77.99^{3.1}$	$69.51^{2.4}$	$\mathbf{43.23^{2.9}}$	$55.38^{2.8}$	$48.54^{2.8}$
crf-0-DSE	$65.48^{2.0}$	$74.85^{3.5}$	$69.83^{2.4}$	$39.95^{2.4}$	$44.52^{2.2}$	$42.10^{2.2}$
crf-0-DSE&ESE	$\mathbf{69.22^{1.8}}$	$72.16^{3.2}$	$\mathbf{70.65^{2.4}}$	$42.13^{2.3}$	$42.69^{2.5}$	$42.40^{2.3}$

Table 2: Results for identifying direct subjective expressions. Superscripts designate one standard deviation.

	overlap			exact		
method	recall	precision	F	recall	precision	F
Wiebe baseline	$56.36^{2.1}$	$43.03^{4.5}$	$48.66^{3.3}$	$15.09^{1.1}$	$9.91^{1.6}$	$11.92^{1.4}$
Wilson baseline	$\mathbf{66.10^{2.6}}$	$40.94^{4.7}$	$50.38^{4.0}$	$17.23^{1.9}$	$8.76^{1.5}$	$11.56^{1.6}$
crf-1-ESE	$46.36^{4.1}$	$\mathbf{75.21^{6.6}}$	$57.14^{3.6}$	$15.11^{1.7}$	$\mathbf{27.28^{2.3}}$	$19.35^{1.5}$
crf-1-DSE&ESE	$48.79^{3.2}$	$74.09^{6.7}$	$58.70^{3.7}$	$15.58^{1.1}$	$26.18^{2.1}$	$\mathbf{19.46^{0.8}}$
crf-0-ESE	$61.22^{3.4}$	$64.84^{5.4}$	$62.82^{3.3}$	$18.31^{1.7}$	$17.11^{3.0}$	$17.61^{2.2}$
crf-0-DSE&ESE	$63.46^{3.3}$	$63.76^{5.7}$	$\mathbf{63.43^{3.3}}$	$\mathbf{18.96^{1.4}}$	$16.79^{2.5}$	$17.74^{1.8}$

Table 3: Results for identifying expressive subjective elements. Superscripts designate one standard deviation.

	overlap			exact		
method	recall	precision	F	recall	precision	F
Wiebe baseline	$51.59^{2.0}$	$61.35^{4.6}$	$55.99^{2.8}$	$17.70^{0.8}$	$19.61^{2.0}$	$18.58^{1.2}$
Wilson baseline	$61.23^{2.1}$	$58.48^{4.7}$	$59.73^{3.1}$	$20.61^{1.4}$	$17.68^{1.5}$	$19.00^{1.3}$
crf-1-DSE+ESE	$64.77^{2.2}$	$81.33^{4.4}$	$72.03^{2.2}$	$26.68^{2.7}$	$39.23^{2.6}$	$31.70^{2.4}$
crf-1-DSE&ESE	$62.36^{2.1}$	$\mathbf{81.90^{4.1}}$	$70.74^{2.2}$	$28.24^{2.7}$	$\mathbf{42.64^{1.9}}$	$\mathbf{33.92^{2.3}}$
crf-0-DSE+ESE	$\mathbf{74.70^{2.5}}$	$71.64^{4.5}$	$\mathbf{73.05^{2.8}}$	$\mathbf{30.93^{2.5}}$	$28.20^{2.3}$	$29.44^{2.0}$
crf-0-DSE&ESE	$71.91^{2.2}$	$74.04^{4.5}$	$72.88^{2.6}$	$30.30^{2.2}$	$29.64^{2.3}$	$29.91^{1.8}$

Table 4: Results for identifying expressions which are either DSEs or ESEs. Superscripts designate one standard deviation. DSE&ESE indicates a model trained to make a three-way distinction among DSEs, ESEs, and other tokens, while DSE+ESE indicates a model trained to make a two-way distinction between DSEs or ESEs and all other tokens.

	overlap			exact		
feature set	recall	precision	F	recall	precision	F
base	$47.14^{2.6}$	$70.91^{4.4}$	$56.60^{3.0}$	$30.55^{2.7}$	$\mathbf{45.12^{3.1}}$	$36.41^{2.8}$
base + Levin/FN	$50.57^{3.1}$	$70.51^{4.1}$	$58.86^{3.3}$	$32.20^{3.1}$	$44.11^{3.3}$	$37.20^{3.1}$
base + Wilson	$54.92^{2.4}$	$70.73^{4.0}$	$61.81^{2.9}$	$34.61^{2.5}$	$43.60^{2.9}$	$38.57^{2.5}$
base + Wilson + Levin/FN	$57.21^{2.6}$	$70.79^{4.1}$	$63.26^{3.0}$	$35.77^{2.4}$	$43.42^{2.8}$	$39.21^{2.5}$
base + WordNet	$68.29^{2.4}$	$71.82^{3.5}$	$70.00^{2.8}$	$41.80^{2.5}$	$42.71^{2.5}$	$42.24^{2.4}$
base + Wilson + WordNet	$68.93^{2.1}$	$72.06^{3.3}$	$70.45^{2.6}$	$42.10^{2.5}$	$42.71^{2.6}$	$42.40^{2.5}$
base + Levin/FN + WordNet	$68.48^{2.4}$	$71.87^{3.3}$	$70.13^{2.8}$	$41.92^{2.2}$	$42.80^{2.5}$	$42.34^{2.3}$
base + Levin/FN + WordNet + Wilson	$\mathbf{69.22^{1.8}}$	$\mathbf{72.16^{3.2}}$	$\mathbf{70.65^{2.4}}$	$\mathbf{42.13^{2.3}}$	$42.69^{2.5}$	$\mathbf{42.40^{2.3}}$

Table 5: Results for feature ablation for identifying DSEs. FN is the FrameNet dictionary features. "base" indicates the lexical features and the syntactic features. The bottom line represents the same model as CRF-0-DSE&ESE in Table 2.

model, labeled crf-0 in the tables). Orthogonally, we compare models trained separately on each task (classifying each token as in a DSE versus not or in an ESE versus not, labeled just DSE or ESE in the tables) to models trained to do both tasks at once (classifying each token into one of three classes: in a DSE, in an ESE, or neither[7], labeled DSE&ESE in the tables).

Because the baselines were not designed to distinguish between DSEs and ESEs, we run another set of experiments where the two categories are lumped together. The rows labeled DSE&ESE use the models trained previously to distinguish three categories, but are here evaluated only on the binary decision of opinion expression or not. The rows labeled DSE+ESE are trained to classify a token as I if it is in either a DSE or ESE, or O otherwise. The results of these experiments are reported in Table 4.

Finally, to determine the effect of the various dictionaries, we examine all combinations of the various dictionaries - WordNet, Framenet, Levin, and the clues from Wilson et al. [2005b] (to save space, we combine the two smallest dictionaries, Framenet and Levin, into one). These results, on the DSE task, are reported in Table 5.

3.5 Discussion

We note that the order-0 models outperform the order-1 models according to overlap F-measure and recall, but by exact F-measure and either precision metric, the order-1 models are superior. The creators of the dataset state "we did not attempt to define rules for boundary agreement in the annotation instructions, nor was boundary agreement stressed during training." [Wiebe *et al.*, 2005, page 35]. For example, whether a DSE ought to be annotated as "firmly said" or just "said" is left up to the annotator. Therefore, we hypothesize that the model with greater capacity (the order 0+1) may overfit to the somewhat inconsistent training data.

The ablation results (in Table 5) indicate that the WordNet features are by far the most helpful. The other two dictionary sets are individually useful (with the Wilson features being more useful than the Levin/Framenet ones), but above the WordNet features the others make only a small difference. This is interesting, especially since the WordNet dictionary is entirely general, and the Wilson dictionary was built specifically for the task of recognizing subjectivity. Ablation tables for the other two targets (ESEs and DSE&ESE) look similar and are omitted.

In looking at errors on the development data, we found several causes which we could potentially fix to yield higher performance. The category of DSEs includes speech events like "said" or "a statement," but not all occurrences of speech events are DSEs, since some are simply statements of objective fact. Adding features to help the model make this distinction should help performance. Also, as others have observed, expressions of subjectivity tend to cluster, so incorporating features based on the density of expressions might help as well [Wiebe and Wilson, 2002].

[7] A small number of tokens are manually annotated as being part of both a DSE and an ESE. For training, we label these tokens as DSEs, while for testing, we (impossibly) require the model to annotate both entities.

Finally, we note that the interannotator agreement results for these tasks are relatively low; 0.75 for DSEs and 0.72 for ESEs, according to a metric very close to overlap F-measure[8]. Our results are thus quite close to the human performance level for this task.

4 Related work

Subjectivity research has been quite popular in recent years. While ultimately research in lexicon building [Hatzivassiloglou and McKeown, 1997, e.g.], and classification [Dave *et al.*, 2003, e.g.] may be relevant, we focus here on work in extracting sub-sentential structure relevant to subjectivity.

Bethard et al. [2004] address the task of extracting propositional opinions and their holders. They define an opinion as "a sentence, or part of a sentence that would answer the question 'How does X feel about Y?' " A propositional opinion is an opinion "localized in the propositional argument" of certain verbs, such as "believe" or "realize". Their task then corresponds to identifying a DSE, its associated direct source, and the content of the private state. However, in each sentence, they seek only a single verb with a propositional argument, whereas we may identify multiple DSEs per sentence, which may be multi-word expressions of a variety of parts of speech.

Another group of related work looks at identifying a class of expressions similar to the DSEs we identify [Wiebe *et al.*, 2003; Munson *et al.*, 2005; Wilson *et al.*, 2005a] [9]. We cannot compare our results to theirs because this previous work does not distinguish between DSEs and objective speech expressions, and because the prior results only address finding single word expressions.

Another area of related work is reputation or sentiment analysis [Morinaga *et al.*, 2002; Nasukawa and Yi, 2003]. This work is in the context of marketing research, and involves identifying polarity and sentiment terminology with respect to particular products or product categories. Their notions of sentiment terms are related, but not identical, to ours. However, they do provide evidence that working at the expression level is of interest to consumers of opinion-oriented information extraction.

5 Conclusions and Future Work

Extracting information about subjectivity is an area of great interest to a variety of public and private interests. We have argued that successfully pursuing this research will require the same expression-level identification as in factual information extraction. Our method is the first to directly approach the task of extracting these expressions, achieving performance within 5% of human interannotator agreement. In the future, we hope to improve performance even further by

[8] Using the agr statistic, the interannotator agreement for ESEs on the MPQA data is 0.72 [Wiebe *et al.*, 2005, page 36], and for DSEs is 0.75 (Theresa Wilson, personal communication). agr is the arithmetic (rather than harmonic) mean of overlap recall and precision between two annotators.

[9] The category is referred to as an *on* or *private state frame*.

the methods discussed earlier, and build on our expression-level identification towards systems that present the user with a comprehensive view of the opinions expressed in text.

Acknowledgements

The authors would like to thank Oren Kurland, Alexandru Niculescu-Mizil, Filip Radlinski, and Theresa Wilson for helpful comments on earlier versions of this paper.

References

[Abney, 1996] Steven Abney. Partial parsing via finite-state cascades. *Journal of Natural Language Engineering*, 2(4):337–344, 1996.

[Bethard *et al.*, 2004] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004. March 22-24, 2004, Stanford.

[Cunningham *et al.*, 2002] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02)*, Philadelphia, July 2002.

[Dave *et al.*, 2003] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW2003*, 2003.

[DUC2005, 2005] *Proceedings of the Document Understanding Workshop*, Vancouver, B.C., Canada, October 2005. Presented at the HLT-EMNLP Annual Meeting.

[Hatzivassiloglou and McKeown, 1997] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *ACL97*, 1997.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[Levin, 1993] Beth Levin. *English Verb Classes and Alternations*. University of Chicago Press, 1993.

[McCallum, 2002] Andrew Kachites McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[Miller, 1995] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, November 1995.

[Morinaga *et al.*, 2002] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. Mining product reputations on the web. In *KDD 2002*, 2002.

[MUC7, 1998] *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Morgan Kaufman, April 1998.

[Munson *et al.*, 2005] M Arthur Munson, Claire Cardie, and Rich Caruana. Optimizing to arbitrary NLP metrics using ensemble selection. In *HLT-EMNLP05*, 2005.

[Nasukawa and Yi, 2003] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the international conference on Knowledge capture*, 2003.

[NIS, 2005] NIST. *Proceedings of The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005.

[Riloff and Wiebe, 2003] Ellen Riloff and Janyce M Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003.

[Stoyanov *et al.*, 2005] Veselin Stoyanov, Claire Cardie, Diane Litman, and Janyce Wiebe. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In Qu Shanahan and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Practice.* Springer, 2005.

[Wiebe and Riloff, 2005] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*. Springer-Verlag, 2005. Invited paper.

[Wiebe and Wilson, 2002] Jaynce Wiebe and Theresa Wilson. Learning to disambiguate potentially subjective expressions. In *Sixth Conference on Natural Language Learning*, Taipei, Taiwan, August 2002.

[Wiebe *et al.*, 2003] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury. Recognizing and Organizing Opinions Expressed in the World Press. In *Papers from the AAAI Spring Symposium on New Directions in Question Answering (AAAI tech report SS-03-07)*, 2003. March 24-26, 2003. Stanford.

[Wiebe *et al.*, 2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39(2-3):165–210, 2005.

[Wilson *et al.*, 2005a] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada, October 2005. Demo abstract.

[Wilson *et al.*, 2005b] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. Vancouver, Canada, 2005.