Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations

Jason S. Kessler
Dept. of Computer Science
Indiana University
Bloomington, IN 47405, U.S.A.
jaskessl@cs.indiana.edu

Nicolas Nicolov
J.D. Power and Associates, McGraw-Hill
4888 Pearl East Circle
Boulder, CO 80301, U.S.A.
nicolas_nicolov@jdpa.com

Abstract

User generated content is extremely valuable for mining market intelligence because it is unsolicited. We study the problem of analyzing users' sentiment and opinion in their blog, message board, etc. posts with respect to topics expressed as a search query. In the scenario we consider the matches of the search query terms are expanded through coreference and meronymy to produce a set of mentions. The mentions are contextually evaluated for sentiment and their scores are aggregated (using a data structure we introduce call the *sentiment propagation graph*) to produce an aggregate score for the input entity. An extremely crucial part in the contextual evaluation of individual mentions is finding which sentiment expressions are semantically related to (target) which mentions — this is the focus of our paper. We present an approach where potential target mentions for a sentiment expression are ranked using supervised machine learning (Support Vector Machines) where the main features are the syntactic configurations (typed dependency paths) connecting the sentiment expression and the mention. We have created a large English corpus of product discussions blogs annotated with semantic types of mentions, coreference, meronymy and sentiment targets. The corpus proves that coreference and meronymy are not marginal phenomena but are really central to determining the overall sentiment for the top-level entity. We evaluate a number of techniques for sentiment targeting and present results which we believe push the current state-of-the-art.

1. Introduction

User generated content is extremely valuable for mining market intelligence because it is unsolicited. The focus of this paper is on analyzing comments by users about companies, their products and services. We retrieve a large set of documents, including blogs and message board postings that mention a company name, its products and its services. After we filter spam documents (Nicolov and Salvetti 2007; Kolari, Finin, and Joshi 2006) we apply a combination of data-driven clustering, and manual specifications of clusters which result in cluster descriptions which are essentially a very large IR search query. There are alternative ways of

classifying documents to a topic. We are interested in the demographics and sentiment on a certain topic or subtopic. The IR query would retrieve relevant documents on a topic together with matching key phrases. These matching key phrases serve as input for the sentiment analysis component which we focus on in this paper.

Conceptually we view the input phrases recovered from the IR search (we call these anchors) as identifying few of the mentions of an entity in a document. We expand the input set of mentions through co-reference resolution (Ng and Cardie 2002; Luo et al. 2004; NIST Speech Group 2006). Because often when people express sentiment, they express sentiment toward parts and features of a product (Hu and Liu 2006) we further expand the current set of mentions with additional ones reachable from the current ones through meronymy relations. We further look for coreferential mentions to the newly introduced ones and then seek additional mentions connected to them through meronymy, etc. Essentially we find the transitive closure of the coreference and meronymy relations on the initial set of mentions.

Next, we wish to determine which of these mentions are being evaluated and the polarity of these evaluations. Much of the evaluation will be rooted in words or phrases which directly evaluate these mentions. We refer to these as **sentiment expressions** (SEs) and to the mentions they evaluate as their targets. The polarity and intensity of these SEs can be found through their semantic properties and by taking into account other elements such as negators and intensifiers that target them. The evaluations rendered by these SE's to their target mentions propagate through coreference and meronymy chains, and aggregately form what we call entity-level sentiment.

The focus of our paper is on the problem of linking sentiment expressions to the mentions they target. We present an approach where potential target mentions of an SE are ranked using supervised machine learning (Support Vector Machines) where the main features are the syntactic configurations (typed dependency paths) connecting the SE and the mention. We have created a large English corpus of blog entries containing critical discussions of products and includes annotations of semantically typed mentions, their coreference and meronymy relationships, and sentiment expressions linked to their targets. The corpus proves that coreference and meronymy are not marginal phenomena but

are really central to determining the overall sentiment for the top-level entity. We evaluate a number of techniques for sentiment targeting and present results which we believe push the current state-of-the-art.

2. Related work

The problem of finding targets of objects similar to our sentiment expressions has received considerable recent attention.

In our previous work (Nicolov, Salvetti, and Ivanova 2008) we have used token proximity as a proxy for semantic relatedness. While the technique is fast, however, error analysis has revealed cases where sentiment expressions are in proximity of a key phrase match but they are not semantically related. It is exactly these cases that motivated us to develop the current approach of identifying the targets of the sentiment expressions.

Hu and Liu (2004), working to find opinions about movie reviews, provide one of the earliest approaches to the problem of targeted sentiment analysis. Having automatically annotated adjectival opinion terms, these terms are linked to adjacent features they modify. Our approach does not require mention/SE adjacency or limit the syntactic types of sentiment expressions.

Bloom07 consider the task of linking targets of attitudes, within the appraisal expression framework. Attitudes are similar to our SEs. An attitude links to a target when two conditions are met: the typed dependency path between the two elements matches one of 41 hand-crafted paths, and no path with a higher rank in the path list links the attitude to another target. Accuracy in the middle 70% range is recorded for this task over 100 sentences of manually annotated data. We test a machine learning based approach over a much larger annotated corpus. Similar to Bloom, Garg, and Argamon, Popescu and Etzioni (2005) use 10 "syntactic dependency rule templates" over a dependency tree to relate identified product features to potential opinion words. Kessler (2008) also uses a database of typed dependency paths to find intrasentential subjective relationships. Unlike our approach, Kessler finds relationships that denote the declaration of belief and only considers relationships that link expressions to declarative finite clauses.

Zhuang, Jing, and Zhu (2006), working on the same problem as Hu and Liu (2004), automatically induce a list of dependency path templates (decorated with part-of-speech tags) by collecting frequently occurring paths between opinion words and features of movies which co-occur in the same sentence. Zhuang, Jing, and Zhu do not directly evaluate the linking component of their system, instead showing that their combined opinion-feature linker and opinion-term/feature annotator outperform that of Hu and Liu.

Bethard et al. (2004) consider the task of finding targets of opinion denoting verbs that are propositional. In contrast, we find targets of all sentiment expressions, regardless of their constituent, part-of-speech, or length in tokens. However, we only consider sentiment expressions which target explicit mentions.

Kim and Hovy (2006) link opinion-bearing adjectives and verbs (equivalent to our sentiment expressions) to topics (our mentions) using automatic semantic role labeling (SRL) as an alternative to using purely syntactic relationships. SRL is the identification of which semantic roles (e.g., Agent, Patient) some constituents take and the linking of these constituents to predicates. When linked, these constituents are called frame elements. Their hypothesis is that certain frame elements are always sources or targets of their predicates. This is in contrast to syntactic relations which are not sufficient to determine source or target relations. Kim and Hovy give the examples:

(1)   a. He *condemned* [$_{topic}$ the lawyer].
       b. [$_{topic}$ Her letter] *upset* me.
       c. Her [$_{target}$ letter] *upset* me.
       d. [$_{topic}$ Her letter and attitude] *upset* me.
       e. Her [$_{target}$ letter] and [$_{topic}$ attitude] *upset* me.

(1-a-b) demonstrating that, although *condemned* and *upset* are both verbs, they have different arguments as topics– *condemned* (1-a) linking to its direct object and *upset* (1-b) to its subject. However, both topics are in the stimulus frame element. These variations indicate that global syntactic pattern-based approaches such as Bloom, Garg, and Argamon are inadequate. While we do not employ a semantic role labeler in our study, we do wish to capture these relations. Another difference is that we do not consider mentions to always be at the chunk level. For example, (1-c) shows how we would annotate Kim and Hovy's (1-b). (1-d) (our annotation would be (1-e)) shows how annotating below the chunk level increases the difficulty of the task by allowing multiple targets in the same frame element when coordinates are separated.

Kobayashi et al. (2006) presents an approach to targeting attitudes in Japanese. Their study is over a corpus of product reviews annotated with, among other things, linked aspect-evaluation. Their "aspects" are equivalent to our mentions while their "evaluations" are equivalent to our sentiment expressions. They identify these linkages by ranking candidate aspects by their likelihood of being linked to by a given evaluation. Using the method presented in Iida et al. (2003), the ranking is computed by training a classifier to determine which of two candidate aspects is more likely to be linked. The winner of each comparison goes to be classified against the next unseen aspect until a global winner is found. The classifier uses features derived from a shallow syntactic parse as well as the semantic type of each candidate aspect. We explore a similar approach using similar features though our corpus is in English. The results of a study comparing dependency parser performance among different languages suggests that Japanese could be correctly parsed more reliably than the 13 other studied languages (Nivre 2008). This leads us to believe that English, not sharing Japanese's case-markers, may be more difficult to parse and present more of a challenge.

Ruppenhofer, Somasundaran, and Wiebe (2008) discusses difficulties in finding sources and targets of subjective expressions. Subjective expressions, which include expressions of belief, speculation, emotion, and evaluation, encompass more than the sentiment expressions which target mentions of physical entities we consider. Their treatment

of targets is more general than ours, allowing for a target to be whatever the subjective expression is about. Wilson (2008) presents the MPQA 2.0 corpus which, among other things, contains annotations of subjective expressions and their targets. Section 9 compares the size of the corpus we developed with MPQA 2.0.

Stoyanov and Cardie (2008a) present a similar corpus to MPQA 2.0 (also based on MPQA 1.2.) which they call the MPQA$_{TOPIC}$ corpus. This is annotated with, among other things, topic spans of opinions that are sentiment-bearing, which are equivalent to what we call targets. They use the term "target" to refer to the content of the opinion itself, for instance the complement clause of a speech event verb. Implicit topics are also annotated. Stoyanov and Cardie (2008b) present an approach, based on this corpus, for finding coreferent topics.

3. The larger problem

The focus of this paper is on the problem of finding the targets of individual expressions of sentiment. In this section we discuss how this task fits into the larger problem of opinion mining and show how it can be subdivided into approachable tasks.

The larger goal in annotating this corpus is to determine how sentiment is expressed toward entities (e.g., cars or cameras) that are being evaluated. We annotate the corpus under the hypothesis that one can recover this sentiment by determining which sentiment expressions target mentions of the entity itself and the entity's parts and features. Once the contextual polarity of each of these sentiment expressions is found, the polar sentiment can be propagated to the top-level entity. Based on the sentiment toward mentions of parts and features, and their parts and features, etc... we can determine part and feature level sentiment, and propagate sentiments up to their parent parts, eventually coming to an aggregate, document level sentiment toward a relevant, parent entity. This approach, if done correctly, will ensure that sentiment directed toward irrelevant entities will not be considered.

We consider **sentiment expressions** to be any evaluative expression that targets a mention of an entity. These exclude mentions that target entities not explicitly mentioned or toward events. A **mention** is defined to be any concrete object or feature of an object that could be evaluated. For instance, consider the following invented car review. Sentiment expressions are italicized while mentions are underlined.

(2) a. I *like* my new Honda Civic.
     b. It gets *great* gas mileage, has a *powerful* engine, and a *nice* interior!
     c. Sadly, the seats and upholstery are not *durable*.
     d. I also looked at the Toyota Corolla, but its engine seemed sluggish.
     e. My mechanic told me yesterday the Civic had *good* handling.

Figure 1 describes the meronymy, coreference, and sentiment targeting relations that directly or indirectly affect the Honda Civic entity. Mentions and sentiment related to the Corolla entity are not pictured because they are not connected to the graph. Each entity is represented by a rectangle. Each mention of an entity is drawn as an egg within that entity's box. Contextual sentiment toward entities as well as their referent objects are listed at the top of each box. Part-of and feature-of relations are drawn between entity boxes. Sentiment expressions and their modifiers occur outside the entity boxes and link to the mentions they target. Immediate opinion holders are treated like any other entity, except they link to the sentiment expressions which they hold.

While having such a structured document representation in place makes the task of finding the sentiment toward the top-level entity considerably easier, it opens up many avenues for opinion mining. For example, queries posed to a question-answering system such as "What do people not like about product X?" or "What other features do users who dislike the camera's zoom lens feel strongly about?" can be trivially answered with this data structure. Framing the opinion mining problem in this manner allows for the use of a wide array of previous work in computational linguistics.

There have been studies on the problems of finding opinion holders (Kim and Hovy 2006; Bethard et al. 2004), the contextual polarities of sentiment expressions (Choi and Cardie 2008; Wilson, Wiebe, and Hoffmann 2005), and coreference resolution among topics (Stoyanov and Cardie 2008b) and meronymic relations (Girju, Badulescu, and Moldovan 2006) between mentions. We will consider approaches to these important problems in future work. Polanyi and Zaenen (2006) discuss the problem of determining valences (what we call contextual polarity or sentiment) of mentions and opinion bearing terms. The issue of determining how sentiment propagates up the subtopic-topic chain is also discussed. For the remainder of the paper, we will concentrate on the problem of determining targets of sentiment expressions.

4. Data

To perform this study, we manually collected 194 blog entries that contained a paragraph or longer evaluative passage about a car or a digital camera. See Figure 1 for detailed information about the corpus composition.

The data was tokenized, sentence-split, part-of-speech tagged with SVMTool (Giménez and Márquez 2004) and parsed with the Stanford dependency parser (de Marneffe, MacCartney, and Manning 2006).

Domain	Docs.	Sents.	Tokens	SEs	Mentions
Car	111	4,493	80,560	3,353	16,953
Camera	96	3,527	65,226	2,724	16,193
All	207	8,020	145,786	6,077	33,146

Table 1: Corpus size statistics.

4.1. Annotation process

Our annotation philosophy was to allow the annotators to annotate expressions of sentiment as they intuitively viewed them. However, each sentiment expression annotation was to cover the smallest amount of text possible while still obeying token boundaries.
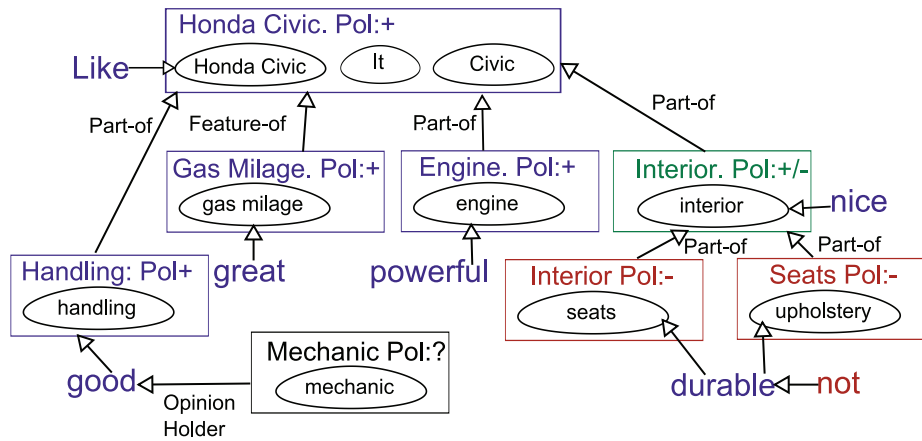
Figure 1: The sentiment propagation graph of example 2

Mention annotations were performed over dozens of semantic types, a subset of which originated in the ACE guidelines (NIST Speech Group 2006). These include everything from people to geo-political entities to times, rates, ages to locations and facilities. Each mention was annotated with a semantic type. We included custom, domain specific types including Car, SUV, CarPart, Truck, Camera and Camera-Part. Pronouns that referred to any of these entities were marked with their referent's semantic type.

Some verbs whose nominalizations are features of objects (e.g., "handles" and "handling") are also annotated as mentions. For example

(3) a. The <u>acceleration</u> is *good*.
     b. It <u>accelerates</u> *well*.

(3-a-b) are clearly paraphrases. Both express positive sentiment toward an object's acceleration. Some denominal verbs may also be marked. For instance, "zoom" in (4-b) will be annotated as a mention because it refers to the function of a concrete entity.

(4) a. The <u>camera</u> has *good* [target <u>zoom</u>].
     b. The <u>camera</u> let me [target <u>zoom</u>] in on the <u>police car</u> *perfectly*.

It should be noted that sentiment expressions can have multiple targets. In (5-a), this occurs through conjunctive coordination. In (5-b), it is a semantic property of the sentiment expression.

(5) a. <u>Coffee</u> and <u>Red Bull</u> are *essential* to graduate student life.
     b. [target <u>Bill</u>] *bragged* about <u>his</u> [target <u>Mustang</u>].

In (5-b), *bragged* targets <u>Bill</u> with negative sentiment and targets <u>Mustang</u> with positive sentiment held by Bill.

Six car documents were used as annotator training documents and annotated by four annotators. Five more were marked by three, and 25 by two. One camera document was annotated by every annotator as a training exercise for the domain. The remainder of the camera documents were annotated by only one person. Conflicts in the documents annotated multiple times were addressed in annotator conferences.

To evaluate inter-annotator agreement, we adopted the agr metric from Wiebe, Wilson, and Cardie (2005) which measures the recall of one annotator's work using the other's annotations as a gold standard. The positions of the two annotators are reversed, and the harmonic mean of the two precision values is taken, rendering a single performance metric. Formally

$$agr(A||B) = \frac{|A \cap B|}{|A|}$$

where A and B are the annotation sets produced by different annotators on the same document, and $A \cap B$ is the set of overlapping annotations. Table 2 shows promising agreement statistics. The harmonic mean agr reported for expressive subjective elements in (Wiebe, Wilson, and Cardie 2005) of 72% and is 82% agreement for direct subjective expressions and speech events. Wilson (2008) reports 85% agreement for target annotation. Our scores are similar.

Note that contextual polarity of sentiment expressions, intensifiers, committers, shifters, and neutralizers were also annotated. In addition, opinion holder information, comparisons, coreference, and meronymic relations were also annotated. We will discuss these annotations in more detail in future work.

We intend to make the corpus available for joint projects and possibly wider distribution in the future.

5. Problem statement

We define the problem of targeting sentiment expressions as follows. Assume a document is annotated with mentions and sentiment expressions as described above. Further, assume we have access to the types of mentions that have been annotated. The task is to identify which mention or mentions a sentiment expression targets. Given the annotation scheme, we can assume that every sentiment expression has a target.

Example (6) illustrates the difficulty and importance of this problem.

(6) a. Unlike the *sluggish* <u>Chevy</u>, the <u>Audi R8</u> performed like a *dream*.
     b. The <u>Dodge Ram</u>, which <u>I</u> purchased from the <u>Washington, D.C.</u> <u>Craigslist</u> used auto section,

A vs. B	# Docs	Sent. Exps.			Targets			Mentions		
		$agr(A||B)$	$agr(B||A)$	Mean	$agr(A||B)$	$agr(B||A)$	Mean	$agr(A||B)$	$agr(B||A)$	Mean
A_1 vs. A_2	5	67.9	84.3	75.2	89.8	88.1	90.0	90.1	93.1	91.6
A_1 vs. A_3	6	81.3	85.2	83.2	84.3	80.8	82.6	92.3	92.8	82.6
A_1 vs. A_4	5	86.1	83.9	85.0	79.9	83.6	81.8	93.7	91.3	92.5
A_2 vs. A_4	17	78.0	80.1	79.1	79.6	80.9	80.3	87.5	91.7	89.6
A_2 vs. A_3	34	73.7	71.9	72.8	80.6	83.1	81.9	84.7	91.8	88.1
A_3 vs. A_4	28	78.4	78.3	78.3	80.9	79.4	80.1	90.9	77.8	83.9

Table 2: Interannotator agreement. Mean refers to harmonic mean. Target agreement is assessed only when sentiment expressions match. We do not take into account mention types when calculating agreement.

turned out to be *excellent*.

in (6-a), the negative sentiment expression *sluggish* should target Chevy. However, it is only three tokens away from Audi R8, which is being praised. A proximity based approach that would assign sentiment to all mentions within a window would likely cause *sluggish* to be considered as targeting. As a simple alternative, an approach that would simply select a sentiment expression's closest mention would work.

In (6-b), the sentiment expression *excellent* and its target, Dodge Ram, are at least 16 tokens away. This presents two problems for an n-token window approach. First, the target is so far away from the sentiment expression that it would likely fall outside of the window. Second, there are four mentions in the relative clause that are closer than the target which are not targeted by the sentiment expression. A system using a shallow parse to evaluate targets for sentiment expression would be at an advantage. It would prevent the mentions within the relative clause from linking to a constituent in the main verb phrase.

In Example (1), we have seen that a shallow parse is not sufficient for determining targeting. Lexico-semantic knowledge is required.

6. Statistics about the problem: Why the task is hard

So far, we have presented a few examples that present challenges for existing methods. The following statistics suggest that some of the assumptions on which these methods are built may be problematic.

51% of targets appeared to the left of the sentiment expression and 49% appeared to the right. This means that assuming a target will appear on one side more often than another will not improve performance.

While the median number of tokens between a sentiment expression and its target is 2, the mean is 6.21. Making matters worse, 41% of sentiment expressions have a mention they do not target at least as close to them as a target. 30% have one that is closer. This indicates that a proximity based targeting algorithm will have a difficult time achieving a high precision.

91% of targets are in the same sentence as their sentiment expression. This indicates that high performance is possible without having to perform difficult inter-sentential target classifications.

70% of noun phrase targets are common noun phrases, 14% are pronouns, and 16% are proper names. While this does not have direct implications for target classification, it does motivate the need for robust semantic type detection and coreference.

1,002 unique shortest paths along dependency parses connect sentiment expressions to intrasentential targets. 49% of targets are linked to their sentiment expression by one of the ten most frequent paths. 58% of targets follow one of the top 20 paths. 65% of targets follow one of the top 40 paths, and 67% are in the top 50. 73% are in the top 100. This means that even having a large (> 100) set of patterns has a substantial ceiling on recall. Table 3 shows how some of the most frequent paths have very poor precision if they are treated as targeting whenever they appear between a sentiment expression and a candidate target. For instance, the path ↑prep,↑pobj is the fourth most frequently targeting pattern in the corpus. However, it only leads to a correct target 32% of the time.

7. Approaches to the problem

When solving this problem, we do not account for intersentential SE-mention linkages. We evaluate four approaches to finding targets on the car and camera domains both separately and concatenated.

Our first approach is the proximity baseline. We call this approach **Proximity**. We select the candidate mention closest to the sentiment expression. Ties are broken by picking the mention to the right. Ideally, an approach that properly uses syntactic information will do better than this approach.

The next three approaches make use of dependency parse trees. However, before the trees can be used they have to be transformed to accommodate mention and sentiment expression annotations. Following Kessler (2008), we make each annotation its own node in the tree. The terms covered by each annotation and the arcs between them are subsumed by the annotation node. Incoming and outgoing arcs to subsumed nodes are migrated to the annotation nodes. This has the effect of treating annotations as single words.

The second, **Heuristic Syntax**, is a baseline syntactic approach. This approach takes advantage of the dependency parse tree of the sentence where the sentiment expression occurs. The parse tree is first transformed by merging all nodes that occur in mentions or sentiment expression together into mention or sentiment expression-level. Next, every node that governs or is governed by a sentiment expression and is a mention is considered to be a target. Six of the 20 most frequently occurring relations are longer than one dependency hop. Also, many single hop relations are unreliable (e.g., ↑dep and ↓dep have lower than 50% precision).

Pattern	# Targeting	Precision	Example
↓ amod	894	87.4%	...changed to *beefier* <u>stock</u>
↑ nmod	475	60.8%	<u>It</u> was *loud*
↑ dobj	168	68.3%	I *hated* the <u>3 speed transmission</u>
↑ prep,↑ pobj	163	32.0%	I'm not *letting go* of the <u>Cressida</u>
↓ advmod	94	86.2%	it *sure* did <u>run</u>
↓ nn	65	80.2%	*convenience* <u>items</u>
↑ dep	63	47.0%	*Fast*, powerful <u>3S-GTE engine</u>
↑ pobj	60	56.1%	I really *like* the <u>base car</u>
↓ rcmod	58	61.1%	'72 <u>Gremlin</u> that *needs a lot of work*.
↑ cop	47	90.4%	<u>it's</u> *only one year old*

Table 3: Ten most frequently targeting paths. The paths follow the dependency links (labeled with type and direction) from sentiment expressions to targets. Precision is computed by $\frac{\text{\# times targeting}}{\text{\# times leading to a mention}}$.

The third approach is to employ the pattern list from Bloom, Garg, and Argamon to predict targets. We call this approach **Bloom**. The list of 42 hand-crafted dependency parse patterns is ordered with the intention of having the first pattern found be labeled as targeting.

The fourth approach is **RankSVM**. The objective is to learn a model that ranks mentions occurring in the same sentence as a sentiment expression such that the ones which are highest ranked are likely to be targeted. Because we allow sentiment expressions to target multiple mentions, we also want to allow multiple mentions to be tied for as the highest ranking candidate mentions. Casting this as a bipartite ranking problem (Freund et al. 2003) addresses these concerns.

In this formalism, the instance set is partitioned into two sets: one that is ranked higher and one that is ranked lower. Here, each sentiment expression has an instance set where each instance corresponds to a mention in the same sentence. The instances corresponding to mentions which are targets are given a rank of 1 while the others are given a rank of 0. We use SVMLight's implementation of RankSVM (Joachims 2002) to train a model over these ranked instances. RankSVM trains a preference function with the objective of scoring higher ranking instances higher than lower ranking instances. The default parameters are used. The feature vectors are formed based on the syntactic and semantic relationship between the sentiment expression and candidate mention. This approach allows for multiple targets to be ranked equally. See Table 4 for the features used. We evaluated this approach using 10-fold cross validation and default SVMLight parameters.

A benefit of using this approach over Kobayashi et al.'s instance-to-instance competition is that RankSVM can give many instances the same score, allowing for multiple targets to be annotated per sentiment expression. The competition ensures exactly one target is found and thus cannot handle cases such as conjunctively coordinated targets.

8. Results

We performed experiments with the approaches mentioned above. The results are shown in Table 5.

As expected, the Proximity approach performed worse than Hueristic Syntax and RankSVM.

The biggest surprise was the poor recall of Bloom. Many of the over 40 patterns used did not connect any sentiment expression to its target. In contrast to Bloom, Heuristic Syntax did extremely well. Eight of the top ten paths are singletons and thus picked up by Heuristic Syntax. In fact, the performance of Bloom was worse than the Proximity approach which did not use syntax at all.

Clearly, RankSVM outperformed all other approaches we evaluated. It won all categories except recall among adjectives in the camera corpus. Adjectives almost always target the nouns they modify and thus tend to employ short and regular linking paths. This case is perfectly suited to the Heuristic Syntax approach, leaving it the winner in this category.

Both Heuristic Syntax and RankSVM could target multiple mentions per sentiment expression. This is reflected in the larger number of targets each annotated.

We did not notice significant differences in results across domains. Treating both domains as part of the same data set generally reduced the performance of RankSVM. This may indicate that not only general syntactic properties of the problem were learned in each domain but that the semantic properties of domain specific sentiment expressions were also learned by the specialized classifiers.

While we did not test our data on the same corpus as used in Kim and Hovy, we did include statistics for sentiment expressions headed by verbs and adjectives. Like Kim and Hovy, we found that adjectives were easier to parse than verbs. With their system achieving an F-score of 66.5% for verbs and 70.3% for adjectives, it is likely the tournament model outperforms their semantic role labeler-based system. This is particularly significant because the tournament approach does not use any external lexical resources or unannotated data.

9. Discussion

In this paper we do not address how sentiment expressions are identified. We assume there is a component which performs this task. Many systems use a lexicon of sentiment expressions that have collected in an offline process (e.g., (Stone et al. 1966; Turney 2002; Nicolov, Salvetti, and Ivanova 2008)).

Our results are lower than have been reported in experiments by Bloom, Garg, and Argamon. This is due to a number of factors. For one, we were working on blog data, which is notoriously difficult to parse. Furthermore, we had

Feature Name	Description	Example
Lexical Distance	# of tokens separating the mention from the SE	3
Lexical Path	The tokens between the mention and SE	to drive the
Lexical Stem Path	Stems of Lexical Path	to drive the
Lexical POS Path	Parts of speech of Lexical Path	TO VBP DT
Dependency Path	Shortest dep. path between SE and mention	↓xcomp↓prep↓pobj
Sent Exps in Path	# of sentiment expressions along Dependency Path	0
Mentions in Path	# of mentions along Dependency Path	0
Mention type	Semantic type of mention	Vehicle
POS Relation	Parts of speech of the SE and mention heads	VBP-NN
Stem Dependency Path	Stem of SE concatenated to Dependency Path	like↓xcomp↓prep↓pobj

Table 4: Features used in the ranking classifier. "Dep. path" refers to typed-dependency path. Examples are based on the sentence, "I *like* to drive the <u>car</u>," evaluating *like* and <u>car</u>.

Data Set	# targets	Proximity					Heuristic Syntax					Bloom					RankSVM				
		N	C	P	R	F	N	C	P	R	F	N	C	P	R	F	N	C	P	R	F
both all	6691	5742	3663	0.638	0.547	0.589	4586	3234	0.705	0.483	0.574	3717	2314	0.623	0.346	0.445	**5849**	**4377**	**0.748**	**0.654**	**0.698**
verb	1350	1157	618	0.534	0.458	0.493	**1277**	697	0.546	0.516	0.531	807	398	0.493	0.295	0.369	1178	**866**	**0.735**	0.641	**0.685**
adj	3059	2824	2031	0.719	0.664	0.690	2316	1929	**0.833**	0.631	0.718	2270	1591	0.701	0.520	0.597	**2860**	**2329**	0.814	**0.761**	**0.787**
car all	3699	3131	1973	0.630	0.533	0.578	2399	1658	0.691	0.448	0.544	1902	1160	0.610	0.314	0.414	**3206**	**2357**	**0.735**	**0.637**	**0.683**
verb	779	699	397	0.568	0.510	0.537	**750**	418	0.557	0.537	0.547	481	251	0.522	0.322	0.398	710	**533**	**0.751**	0.684	**0.716**
adj	1485	1385	979	0.707	0.659	0.682	1083	892	**0.824**	0.601	0.695	1072	730	0.681	0.492	0.571	**1408**	**1135**	0.806	**0.764**	**0.785**
camera all	2992	2615	1693	0.647	0.566	0.604	2186	1579	0.722	0.528	0.610	1819	1160	0.638	0.388	0.482	**2643**	**2009**	**0.760**	**0.671**	**0.713**
verb	571	465	228	0.490	0.399	0.440	**531**	283	0.533	0.496	0.514	331	152	0.459	0.266	0.337	468	**332**	**0.709**	0.581	**0.639**
adj	1574	1437	1050	0.731	0.667	0.697	1233	1037	**0.841**	0.659	0.739	1197	861	0.719	0.547	0.621	**1452**	**1186**	0.817	**0.753**	**0.784**

Table 5: Results of target classification experiments. Results in the top partition are annotated over all documents. Those in the lower partitions are from the camera and car corpora. N is # of targets classified, C is # correct targets, P is precision, R is recall, and F is F-Score.

a looser constraint on what, syntactically, could constitute a sentiment expression or mention.

While we do not address the problem of identifying sentiment expressions or determining their polarity, there are instances where knowing the target of a potential sentiment expression is essential to these decisions. These tend to occur in polar facts or utterances that do not contain subjective language but imply sentiment. Consider,

(7)   a.   <u>I</u> *broke* the handle.
       b.   I *broke* the *poorly made* <u>handle</u>.

In (7-a), the author is criticizing himself for breaking something. However, in (7-b), the author is effectively heaping criticism on the handle.

Another challenging aspect of our corpus was that we allowed for sentiment expressions to target multiple mentions. This problem does not appear to be addressed in similar work.[1] We plan to work on including conjunctive coordinates of the retrieved target as immediate future work.

Another aspect of future work is to create features that are more amenable to mention-to-mention comparison.

The corpus can also be used to train mention detection, coreference, and meronymy systems.

We know of two other publicly available corpora that contain opinion-related information in English that include targets of opinions.

The first was presented in Hu and Liu (2004), in which

---

[1]Agarwal, Prabhakar, and Chakrabarty (2008) performs sentence level sentiment analysis, using conjunctions to determine which clauses contribute most to the overall polarity of the sentence. This work does not address targeted sentiment analysis.

the topic of each sentence is annotated and its contextual sentiment value is given. The sentences are drawn from on line reviews of five consumer electronics devices. It contains 113 documents spanning 4,555 sentences and 81,855 tokens. While our corpus is larger and contains much richer annotations, it does not contain annotations for implicit sentiment expressions which are indirectly covered by their approach.

The second is the subset of the MPQA v2.0 corpus containing target annotations (Wilson 2008). The documents are mostly news articles. It contains 461 documents spanning 80,706 sentences and 216,080 tokens. It contains 10,315 subjective expressions (annotated with links) that link to 8,798 targets. These subjective expressions are annotated with "attitude types" indicating what type of subjectivity they invoked. 5,127 of these subjective expressions convey sentiment.

We are continuing the annotation effort adding documents in diverse domains. By the end of 2009 we expect to have annotated more than 300,000 tokens. We hope to make the corpus freely available for research purposes.

10. Conclusions

In this paper we have considered the problem of finding the semantic relation between sentiment expressions and their target mentions. This problem arises in the context of sentiment analysis for (top-level) entities identified with a few mentions (in our case these input mentions come about as matches from an IR query). We have motivated that it is crucial to augment the initial set of mentions through the transitive closure of coreference and meronymy relations. We have annotated a large English corpus with product discussions containing semantic types of mentions, coreference,

meronymy and sentiment target relations. We have considered the sentiment targeting using supervised machine learning (ranking mentions with SVMs) and have also implemented previously introduced techniques to compare our approach. Our approach is language independent and can be applied to other languages given availability of a parser and an annotated corpus. We have improved on the state-of-the-art doing experiments on a larger dataset (in particular, we are referring to the techniques suggested by Bloom, Garg, and Argamon (2007) which we implemented. Yet, the task remains quite challenging. We continue to work actively on this problem — we aim to explore a more sophisticated set of features. We are also growing the sentiment corpus adding documents in diverse domains.

Acknowledgments

References

Agarwal, R.; Prabhakar, T. V.; and Chakrabarty, S. 2008. "I know what you feel": Analyzing the role of conjunctions in automatic sentiment analysis. In *GoTAL*.

Bethard, S.; Yu, H.; Thornton, A.; Hatzivassiloglou, V.; and Jurafsky, D. 2004. Automatic Extraction of Opinion Propositions and their Holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Bloom, K.; Garg, N.; and Argamon, S. 2007. Extracting Appraisal Expressions. In *NAACL-HTL*.

Choi, Y., and Cardie, C. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *EMNLP*.

de Marneffe, M.; MacCartney, B.; and Manning, C. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*.

Freund, Y.; Iyer, R.; Schapire, R. E.; and Singer, Y. 2003. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.* 4.

Giménez, J., and Márquez, L. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *LREC*.

Girju, R.; Badulescu, A.; and Moldovan, D. 2006. Automatic discovery of part-whole relations. *Comput. Linguist.* 32(1).

Hu, M., and Liu, B. 2004. Mining and Summarizing Customer Reviews. In *KDD*.

Hu, M., and Liu, B. 2006. Opinion Extraction and Summarization on the Web. In *AAAI-2006 Nectar Paper Track*.

Iida, R.; Inui, K.; Takamura, H.; and Matsumoto, Y. 2003. Incorporating Contextual Cues in Trainable Models for Coreference Resolution. In *EACL Workshop on the Computational Treatment of Anaphora*.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *KDD*.

Kessler, J. S. 2008. Polling the Blogosphere: a Rule-Based Approach to Belief Classification. In *ICWSM*.

Kim, S.-M., and Hovy, E. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *ACL Workshop on Sentiment and Subjectivity in Text*.

Kobayashi, N.; Iida, R.; Inui, K.; and Matsumoto, Y. 2006. Opinion Mining on the Web by Extracting Subject-Attribute-Value Relations. In *AAAI-CAAW*.

Kolari, P.; Finin, T.; and Joshi, A. 2006. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *Computational Approaches to Analyzing Weblogs*.

Luo, X.; Ittycheriah, A.; Jing, H.; Kambhatla, N.; and Roukos, S. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *ACL*.

Ng, V., and Cardie, C. 2002. Improving machine learning approaches to coreference resolution. In *ACL*.

Nicolov, N., and Salvetti, F. 2007. Efficient Spam Analysis for Weblogs through URL Segmentation. In *RANLP*, volume 292 of *Current Issues in Linguistic Theory (CILT)*.

Nicolov, N.; Salvetti, F.; and Ivanova, S. 2008. Sentiment Analysis: Does Coreference Matter? In *AISB 2008 Convention Communication, Interaction and Social Intelligence*.

NIST Speech Group. 2006. The ace 2006 evaluation plan: Evaluation of the detection and recognition of ace entities, values, temporal expressions, relations, and events.

Nivre, J. 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Comput. Linguist.* 34(4).

Polanyi, L., and Zaenen, A. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications*.

Popescu, A.-M., and Etzioni, O. 2005. Extracting product features and opinions from reviews. In *HLT-EMNLP*.

Ruppenhofer, J.; Somasundaran, S.; and Wiebe, J. 2008. Finding the sources and targets of subjective expressions. In *LREC*.

Stone, P. J.; Dunphy, D. C.; Smith, M. S.; and Ogilvie, D. M. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Stoyanov, V., and Cardie, C. 2008a. Annotating Topics of Opinions. In *LREC*.

Stoyanov, V., and Cardie, C. 2008b. Topic Identification for Fine-Grained Opinion Analysis. In *COLING*.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL*.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating Expressions of Opinions and Emotions in Language. In *LREC*.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP*.

Wilson, T. A. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. Dissertation, University of Pittsburgh.

Zhuang, L.; Jing, F.; and Zhu, X.-Y. 2006. Movie review mining and summarization. In *CIKM*.