The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain

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

Miriam Eckert, Lyndsie Clark, and **Nicolas Nicolov**
J.D. Power and Associates, McGraw-Hill
4888 Pearl East Circle
Boulder, CO 80301, U.S.A.
{*miriam_eckert*, *lyndsie_clark*, *nicolas_nicolov*}*@jdpa.com*

Abstract

This paper presents a rich annotation scheme for mentions, co-reference, meronymy, sentiment expressions, modifiers of sentiment expressions including neutralizers, negators, and intensifiers, and describes a large corpus annotated with this scheme. We describe how this corpus relates to recent, state-of-the-art work in sentiment analysis, and define the various annotation types, provide examples, and show statistics on occurrence and inter-annotator agreement. This resource is the largest sentiment-topical corpus to date and is publicly available. It helps quantify sentiment phenomena, and allows for the construction of advanced sentiment systems and enables direct comparison of different algorithms.

Introduction

The expression of sentiment is a complex phenomenon which is intertwined into the semantic structure of text (Polanyi and Zaenen 2006). A document-level label, such as positive or negative, does not present a full representation of all sentiment present in a document. Sentiment, which we define as evaluation, is expressed toward discourse entities by means of individual expressions of sentiment targeted at mentions of those entities. These expressions of sentiment are often rooted in single or multi-word units, whose positive or negativeness may be impacted by the context. Elements in the context that can alter the polarity include negations and terms which can alter the truth-value of an expression of sentiment, as well as less understood phenomena such as sarcasm and tone. While sentiment toward individual mentions of an entity contribute to its overall sentiment, sentiment toward another, related entity such as a part or a feature may also contribute. Sentiment directed toward individual entities can also effect other entities when comparisons among entities are made. An additional dimension of the phenomena is that certain expressions of sentiment may be attributed to discourse participants other than the speaker.

Our goal is to annotate structures pertinent to sentiment that can be combined to formally explain the sentiment that occurs in a document.

The J.D. Power and Associates (JDPA) Corpus consists of user-generated content (blog posts) containing opinions about automobiles. They have been manually annotated for named, nominal, and pronominal mentions of entities. We define entities as discourse representations of concrete objects (e.g., car, door) and non-concrete objects (e.g., handling, power). For some entities that are prominent topics in the discourse, a single mention from the co-reference chain is selected and marked with entity-level sentiment. This aggregates all sentiment toward that entity.

The examples we give, unless otherwise specified, are taken directly from the corpus and have not been edited.

Mentions referring to the same entity are marked as co-referential. Mentions are assigned semantic types consisting of the Automatic Content Extraction (ACE) (NIST Speech Group 2006) mention types and additional domain-specific types. Meronymy (part-of and feature-of) and instance relations are also annotated. Expressions that convey sentiment toward an entity are annotated with the polarity of their prior and contextual sentiment and are linked to the mentions they target. The following modifiers are annotated. These may target other modifiers or sentiment expressions.

- negators (expressions that invert the polarity of a sentiment expression or modifier)
- neutralizers (expressions that do not commit the speaker to the truth of the target sentiment expression or modifier)
- committers (expressions that shift speaker's certainty toward a sentiment expression) or modifier)
- intensifiers (expressions that shift the intensity of a sentiment expression or modifier)

Additionally, we have annotated when the opinion holder of a sentiment expression is someone other than the author of the blog by linking the expression to the holder. We also annotate when two entities are compared on a particular dimension.

In this overview of the corpus, we aim to not only present the nature of the annotations we have added, their examples, numbers, and inter-annotator agreement, but also to highlight problems/tasks in sentiment analysis and natural language processing that can be addressed using this corpus.

The data was gathered manually by annotators by conducting web searches using a variety of car-related search

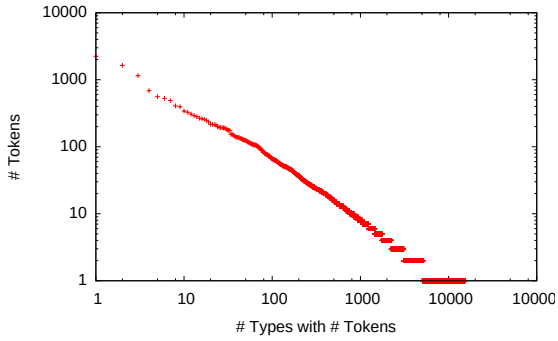

Figure 1: Types vs. tokens of mentions. The power law exponent is -0.84, with $R^2 = 0.93$.

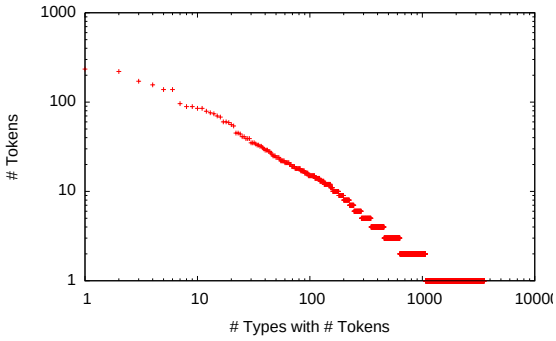

Figure 2: Types vs. tokens of sentiment expressions. The power law exponent is -0.77, with $R^2 = 0.91$.

terms and restricting the retrieved results to certain blog-host sites. The personal blog posts in particular are different in style and sentence structure from professionally edited news texts, with a higher frequency of emotional and colloquial expressions. However, unlike data from Twitter or other microblogging sites, we found the data to adhere for the most part to standard grammatical rules, and disfluencies or incomplete sentences are rare.

We have annotated 335 blog-posts, covering 13,126 sentences and 223,001 tokens.

In this paper, we will cover the annotation of mentions of entities and their semantic relations, the annotation of sentiment expressions and their modifiers, the annotation process, how we judged inter-annotator agreement, and directions of future work.

Annotation types

Evaluative discourse has two, sometimes overlapping components: references to the entities that are being evaluated and terms that are used to express evaluation, or modify its intensity or polarity. We annotate entities that occur in each document, regardless of whether they have any sentiment associated with them. Each entity is represented by coreferring mention span annotations. Furthermore, entities can have relations between each other.

We first discuss our annotation of mentions and the entities they refer to, as well as semantic relations between entities: part-of, feature-of, instance-of, and member-of. Next, we discuss sentiment expression annotations and their modifiers: negators, neutralizers, committers, and intensifiers.

Entities and their relations

Entities are defined as discourse representations of concrete objects (e.g., car, door) and non-concrete objects (e.g., handling, power).

We annotate for four other relations between mentions.

The most basic relation is **refers-to**. It links together two mentions that are coreferring.

Additional additional relations expression semantic relations between the entities the mentions refer to, as opposed to the mentions themselves. However, these relations are annotated between mentions of the entities linked. The specific mentions do not change the meaning of the relation annotations.

Winston, Chaffin, and Herrmann (1987) presents six relationships between entities that encompass what humans would consider to be a "part-of" relationship. We annotated for three of these that were found applicable to the automotive domain.

What we call the **part-of** relation encompasses the relationship of one entity being a concrete part of another. This is Winston et al.'s "component/integral object" relation. He gives the examples of "handle-cup; and "punchline-joke". Some of the part-of relationships that we found in the corpus are:

(1) a. Center console$_1$ Kleenex holder$_{\text{PART-OF-1}}$; I cannot find a tissue box that size to fit in it.
b. The 2009 Mercedes-Benz S600$_2$ is equipped with a twin-turbocharged 5.5 - liter V-12 engine$_{\text{PART-OF-2}}$...

The **feature-of** relation also connects entities, but deals with more abstract entities, where one entity is a property of another. This corresponds to Winston et al.'s "feature-activity" relation. His examples are "paying-shopping" and "dating-adolescence". In our corpus:

(2) a. I love the comfort$_{\text{FEATURE-OF-1}}$ of interior seating$_1$
b. The speed and fuel gauges$_2$ are very hard to see$_{\text{FEATURE-OF-2}}$

Sometimes entities are defined being a type of or equivalent to another entity. These definitional and hypernymic relations that we call **instance-of** relations do not appear in Winston, Chaffin, and Herrmann (1987). Some examples are:

(3) a. Hyundai's futuristic proposal$_{\text{INSTANCE-OF-1}}$ for a small three-door crossover$_1$...
b. Cadillac has launched the 2009 Escalade Platinum Hybrid$_{\text{INSTANCE-OF-2}}$, the most technically advanced large luxury SUV$_2$ yet.

Member-of relations exist between an entity that is part of a group represented by another entity. That entity could be a group (e.g., "they") or part of a product-line, as in Example (4). These correspond to Winston's "member/collection" relations, his examples are "tree-forest" and "card-deck". An example is:

Type	# Mentions	# Named	# Nominal	# Pronominal	# Coreference groups
CarPart	14128	1704	11791	633	11705
Vehicles.Cars	8729	4259	2723	1747	3618
Person	7407	764	1487	5156	2593
CarFeature	6263	264	5930	69	5804
Organization	4910	4092	346	472	2164
Vehicles.SUVs	2052	1115	567	370	837
Time.Year	1208	928	258	22	1136
Units.Money	813	177	628	8	616
Units	796	246	536	14	763
Vehicles	770	243	431	96	432
Units.Rate	741	298	436	7	720
Facility	649	147	464	38	512
Time	568	347	211	10	549
Vehicles.Trucks	466	228	172	66	205
Time.Duration	315	78	236	1	303
GeoPolitical.City	251	191	56	4	206
GeoPolitical.Countries	184	156	18	10	130
Location	157	22	133	2	148
GeoPolitical.Nationalities	131	127	4	0	115
GeoPolitical.USStates	98	89	7	2	81
Time.Month	87	74	13	0	84
GeoPolitical	82	51	29	2	70
Time.Date	56	44	11	1	55
Units.Age	41	10	26	5	38
Time.DaysOfTheWeek	36	36	0	0	36
Time.OClock	13	10	3	0	13

Table 1: Distribution of mention annotations

(4) The peeled back headlamps$_{\text{MEMBER-OF-1}}$, tight front grille$_{\text{MEMBER-OF-1}}$, and stylized tail lamps$_{\text{MEMBER-OF-1}}$ are some of its attractive features$_1$.

The corpus has 61,284 mentions which comprise 42,763 coreference groups (or entities), averaging 1.43 mentions per group. See Table 3 for inter-annotator agreement among mentions and their relations.

Sentiment

Sentiment Expressions. Sentiment expressions are single or multi-word phrases that evaluate an entity. They are linked to the mention they modify through the "target" relation. Our corpus contains 10,425 sentiment expressions, covering 3,545 unique types. 49% of sentiment expressions are headed by adjectives, 22% by nouns, 20% by verbs, and 5% by adverbs. This leads to a diversity of syntactic configurations where sentiment expressions are linked to their target mentions (Kessler and Nicolov 2009). 13% of sentiment expressions are two or more words long.

In general, sentiment expressions convey positive or negative evaluations. We use the term **prior polarity** to refer to whether a sentiment expression is positive or negative. The prior polarity is inferred from the meaning of the sentiment expression, given its target, as opposed to its entire context. "Prior polarity" is a term from Wilson, Wiebe, and Hoffmann (2009); we allow it to depend on a sentiment expression's sense, figurativeness, and a target. Prior polarity contrasts with **contextual polarity** (another term from Wilson, Wiebe, and Hoffmann (2009)) in that contextual polarity is the polarity of the sentiment expression given any modifiers or contextual information that doesn't change its inherent meaning or sense. For example, the prior polarity of "good" in Example (5-a,b,c) (invented) is always positive, while its contextual polarity is respectively positive, negative, and positive. See Table 3 for inter-annotator agreement. We do not annotate contextual polarity directly. Our goal is to make it inferable from modifiers that have been annotated such as negators and other that we discuss below.

(5) a. The car is *good*.
b. The car is not *good*.
c. Only an idiot would think the car is not *good*.

The distribution of prior polarities is skewed toward positive, with 74% positive, 24% negative, 1% neutral and and well less than 1% of mixed prior polarity. Sentiment expressions having "mixed" prior polarity simultaneously express a positive and negative evaluation. These include "pimped-out", "gangsta", "usable", "subtle," and "curious". Neutral sentiment expressions evaluations that are not clearly positive or negative, such as "as expected", "average", "conventional", "so-so", and "different". A next step in expanding this corpus is correcting for the skew in positive and negative sentiment expressions.

Table 2(b) shows the 20 most frequently annotated sentiment expressions in the corpus.

Some sentiment expression types have been marked with different prior polarities when they occur in different contexts. For example, the term "increasing" is marked positive in Example (6-a,b) but negative in Example (6-c).

(6) a. . . . an electric motor that reduces the load on the engine, *increasing* efficiency.
b. . . . *increasing* combustion efficiency and the torque. . .
c. . . . *increasing* gas prices and stricter federal emissions regulations. . .

While prior polarity of "interesting" depends on its topic, other sentiment expressions like "excellent" have a constant prior polarity. Although only 6% of sentiment expression types have tokens with conflicting prior polarities, these account for 25% of sentiment expression tokens in the corpus, making polarity-based disambiguation an important task. Reasons for conflicting prior polarities other than annotator error were the sense of the sentiment expression. For instance, "safe" in Example (7-a) is positive, referring to a vehicle's protectiveness, while "safe" in Example (7-b) is negative, inferring its targets' design is traditional.

# Tokens	Type
2238	i
1639	it
1153	car
687	my
559	engine
527	you
490	its
407	power
395	we
341	vehicle
327	cars
307	one
291	2009
282	interior
266	me
261	they
256	2008
248	ford
235	toyota
216	honda

(a) Mentions

# Tokens	Type
234	good
220	new
171	great
156	like
138	comfortable
138	better
96	love
89	problems
89	fun
85	well
85	unique
79	nice
76	best
74	excellent
70	difficult
68	smooth
60	powerful
60	expensive
59	easy
56	poor

(b) Sentiment expressions

# Tokens	Type
18	seems
18	felt
16	still
16	think
14	seemed
14	feel
14	definitely
13	looks
13	feels
12	certainly
12	may
11	actually
11	might
10	really
10	probably
9	sure
8	seem
8	overall
8	looks like
7	always

(c) Committers

# Tokens	Type
63	if
34	would
27	should
18	could
14	want
13	when
11	optional
10	can
9	needs
8	how
8	may
7	?
6	might
6	or
6	expected
6	need
5	wanted
4	feels
4	expect
4	supposed

(d) Neutralizers

# Tokens	Type
325	very
227	more
122	most
111	much
84	really
77	so
76	top
64	too
58	pretty
39	extremely
38	quite
36	enough
35	even
32	!
32	just
28	less
28	bit
28	a bit
27	completely
26	a little

(e) Intensifiers

# Tokens	Type
299	not
122	no
45	doesn't
44	without
36	don't
28	never
27	isn't
26	didn't
20	don t
20	wasn't
19	doesn t
19	can't
14	aren't
13	won't
13	didn
13	wouldn't
13	nothing
8	lack
8	wasn
8	isn t

(f) Negators

# Tokens	Type
71	like
69	says
66	told
52	owner-reported
30	according
26	ranked
23	said
21	top-ranked
20	ranks
19	according to
12	from
9	reported
9	say
9	calls
8	think
8	rated
7	love
6	rating
6	likes
5	ranking

(g) OPOs

Table 2: Top 20 annotated items in different categories.

(7) a. My family and friends feel extremely *safe* in our Hummer.
b. I saw two VW Eos last week. . . . . and both looked good, albeit in a *safe*, conservative Solara-sort-of-ways.

Much work (Ding, Liu, and Yu 2008; Fahrni and Klenner 2008; Choi, Kim, and Myaeng 2009) has focused on identifying the target-dependent polarity of sentiment expressions[1], while Wiebe and Mihalcea (2006) and Su and Markert (2008) have looked at the problem of polysemy from the perspective of disambiguating subjective and objective senses. Some expressions are only sentiment-bearing when in the right context. For example the term "usable" occurs nine times in the corpus, four of which are annotated as sentiment expressions. Example (8-a) illustrates an example of "usable" being a sentiment expression, and Example (8-b) illustrates a case where it is not.

(8) a. . . . a comfortable and *usable* interior. . .
b. . . . 5,800 pounds (2,631 kg) of *usable* towing capacity. . . .

In fact, 44% of sentiment expression types occurring in the corpus also match non-sentiment bearing sequences of words. These account for 74% of all sentiment expression tokens, motivating the need for sentiment expression detection which can disambiguate candidates based on their context. However, 10% of sentiment multi-word units types have a non-sentiment bearing occurrence but are observed to be sentiment-bearing more than half the time. These account for a substantial 40% of all sentiment expression tokens. 34% of sentiment expression types are not sentiment-bearing in more than half their occurrences. These account for 34% of all sentiment-expression tokens.

Breck, Choi, and Cardie (2007) has applied sequence labeling techniques to the similar task of identifying subjective expressions, a problem which involves the contextual disambiguation of sentiment bearing and non-sentiment bearing phrases.

Sentiment expressions are linked to the mention they describe through the **target** relation. This forms an important connection between sentiment expressed in a document and entities discussed. For inter-annotator agreement purposes, we treat this relation as span-entity link, although annotators are instructed to link to the mention that is directly targeted.

Figures 1 and 2 show the comparative types vs. tokens distributions of mentions and sentiment expressions. Both are nearly similar but sentiment expressions, having a larger exponent, have a fatter tail and, therefore, might be more difficult to recognize automatically.

Contextual polarity and modifiers

There has been considerable work on identifying the contextual polarity of sentiment expressions (Kim and Hovy 2004; Choi and Cardie 2008; Wilson, Wiebe, and Hoffmann 2009; Wiegand and Klakow 2009; Moilanen and Pulman 2009).

A sentiment expression's context can change or modify its polarity, as illustrated by Example (5). We annotate several types of modifiers, which act to change the polarities of sentiment expressions and change the properties of other modifiers. Similar sets of modifiers have been discussed in the literature, but ours is the first attempt at manually annotating occurrences of these modifiers (Polanyi and Zaenen 2006; Shaikh, Prendinger, and Ishizuka 2008; Choi and Cardie 2008; Moilanen and Pulman 2009).

Negators invert the polarity of the sentiment expression

[1] We draw the distinction between the immediate target of a sentiment expression and a document-level topic. Other work, such as Nowson (2009), has addressed the problem of developing topic-dependent feature-sets for supervised classification of document-level polarity.

they target.[2] While "not" is the most well known negator, many other expressions act the same way toward sentiment expressions and other modifiers. For example, in Example (9) "avoids" acts to invert the polarity of the sentiment expression "reduction". Other counter-factives, like "pretend", would also be marked as negators.[3]

(9) This layout *avoids* any *reduction* in the interior space...

In addition to targeting sentiment expressions, negators can also target other modifiers (see Example (10-a)) and even mentions as indicating the absence of an entity. For example, in Example (10-b) "suppressed" indicates the absence of the entity invoked by the mention "noise".

(10) a. ...*not* a *very* quick car.
b. Road and engine noise have been *suppressed*...

The negator "not" in Example (10-a) targets an intensifier, pragmatically acting to negate the sentiment expression (i.e., "quick") the intensifier targeted.

1,014 negator annotations appear in the corpus, tokens of 160 unique types.

Intensifiers act to amplify or dampen the intensity of the sentiment expressed by a sentiment expression or the force of another modifier. Unlike other annotation schemes (Wiebe, Wilson, and Cardie 2005; Hu and Liu 2004) which record the intensity of sentiment, we include these not to assess intensity per se, but for their interaction with other modifiers. These interactions (cf. Example (10-b)) might alter the polarity of sentiment, a process we aim to capture with this annotation scheme.

The direction property can be set to strengthen or weaken. "Considerable" in Example (11-a) would have a direction strengthen, and Example (11-b)'s direction would be weaken.

(11) a. ...it also adds *considerable* benefits...
b. It is *kind of* fun to drive

The direction strengthen is far more common than weaken, with 2,159 occurrences (84%) of strengthening intensifiers (covering 396 types) and 422 occurrences (16%) of weakening intensifiers, accounting for 155 types.

Committers are used to express the author's certainty toward a modifier or sentiment expression.[4] They often express epistemic modality (as in the case of Examples (12-a,b,d)) or hedges ((12-c)). Committers have a property, direction, upward or downward, indicating whether the commitment is being strengthened or weakened. Examples (12-a,b) are all labeled as upward committers, while (12-d) is downward.

(12) a. It was discovered that the switch itself was *DEFINITELY* cracked...
b. I'm *sure* this will drive well...
c. A good looking car *in itself*...
d. The interior *looks* to be in nice condition...

The distribution of direction is relatively even with 417 upward committers (covering 202 types) and 379 downward committers (covering 235 types). The high types-to-tokens ratio and sparsity of the annotation type indicates that this type may be difficult to recognize automatically.

Some committers have been marked as neutralizers or intensifiers and vice versa. In fact, "may" occurs in the top 20 neutralizers and committers (Table 2).

Neutralizers are used to place sentiment expressions or other modifiers into a context where their truth-value is unknown, as occurs in hypothetical or conditional sentences.[5] For the purposes of simplification, in our annotation scheme, neutralizers only target sentiment expressions and not states or events. The targets of the neutralizers in Examples (13) have been shown for clarity. Example (13-a) shows a hypothetical neutralizer, "if" targeting the sentiment expression "poor". That sentiment expression now has a neutral contextual polarity. The neutralizer in (13-b) is a verb that neutralizes the veridicity of the its complement clause, headed by the sentiment expressions "like". (13-c) is similar, except the neutralized argument is in a prepositional phrase.

(13) a. ...*if*$_{\text{TARGET-1}}$ ...the interior is *poor*$_1$...
b. I *tried*$_{\text{TARGET-2}}$ to get used to it and *like*$_2$ it...
c. Aimed at young couples and families who *look*$_{\text{TARGET-3}}$ for a higher level of *performance*$_3$...

437 neutralizers (covering 150 types) are annotated in the corpus.

Entity and mention-level sentiment

Sentiment is marked for certain mentions. Most sentiment is inferable from the structure of sentiment expressions and their modifiers, as all sentiment expressions target mentions. However, in the case where sentiment expressions of conflicting contextual polarities target a mention or in similarly ambiguous cases, annotators mark the **ContextualSentiment** property of mentions. Other mentions carry some inherent sentiment, which we refer to as **MentionPriorPolarity**. For example, referring to a car as a "lemon" would convey a negative mention prior polarity.

Entities that were judged to be prominent were assigned an **EntityLevelSentiment**, which summarized the author's sentiment toward that entity and its meronyms. A mention of a prominent entity is annotated for entity-level sentiment. 873 entities were assigned entity-level sentiment.

[2] Called "negatives" in Polanyi and Zaenen (2006)

[3] The TimeML corpus (Pustejovsky et al. 2003) has explicit annotations for counter-factive events and treats negation as a property of an event. We believe that both act the same way w.r.t. contextual polarity.

[4] Rubin (2007) presents a corpus containing "certainty markers", or expressions indicating commitment to a sentence or a clause and its level of certainty, on a scale from uncertain through absolute certainty. Our committers are judged on a binary scale: do they raise or lower the authors commitment to a sentiment expression or modification.

[5] The problem of determining when an event is asserted as true, false or unknown truth-value is called veridicity (Karttunen and Zaenen 2005). Kessler (2008) has developed a rule-based systems for recognizing the veridicity of some clauses which is tailored to the blogosphere and has released a lexicon which includes "neutral veridicality elements" which neutralize their argument clauses.

These entities had an average of eight either direct or indirect meronyms (e.g., the seats in a car's interior.) Many singletons and entities which are not invoked by many mentions exist in the corpus. Thus, the average prominent entity only had 13 mentions refer to it or one of its direct or indirect meronyms. An average of four sentiment expressions targeted any of these mentions.

Other person's opinions

Reported speech has been a prominent topic in subjectivity and sentiment analysis (Breck and Cardie 2004; Kim and Hovy 2006; Ruppenhofer, Somasundaran, and Wiebe 2008; Krestel, Witte, and Bergler 2008). In order to make the best use of annotation resources, we chose only to annotate in the case when the source of a modifier or sentiment expression was not the author of the document. This contrasts with the MPQA annotation scheme (Wiebe, Wilson, and Cardie 2005), where all reported speech and subjectivity attributed to a source, even if that source was the speaker. We annotate speech events or sentiment expressions that select for a source (i.e., Wiebe, Wilson, and Cardie (2005)'s direct subjective expressions) with the OPO or other person's opinion annotation. Example (14-a) gives an example of an objective speech event sourcing a sentiment expression to someone other than the author, while (14-b) shows an example of a speech event that is also a sentiment expression. In (14-b), "love" is annotated both as an OPO and as a sentiment expression. The sentiment expression targets "cars".

(14) a. The guards$_1$ at Indian Point *told*$_{\text{TARGET-2, SOURCE-1}}$ me *nice*$_2$ car...
b. My kids$_3$ *love*$_{\text{TARGET-SELF,SOURCE-3}}$ cars...

792 OPOs have been annotated in the corpus, covering 250 unique types.

Annotation process

Annotators were trained by reviewing written annotation guidelines and being trained on and having annotated a pilot project, and having their annotations be reviewed by a manager or experienced annotator. Annotators were instructed to mark up text that appeared to fit the criteria for a particular annotation regardless of its syntactic properties. The annotation scheme was developed by collectively annotating several documents. Seven annotators contributed to the corpus.

During the process of corpus creation, some annotation concepts became more concise, some proved to be not clearly enough defined to be accurately annotated, and others required the addition or deletion of slots. A new batch was started when a change to the annotation schema became necessary, or if an existing batch became too large. The following is a description of the individual batches.

- Batch 001: First batch. Size: 78,604 tokens.
- Batch 004: Addition of Mention.CarFeature to distinguish concrete, removable or purchasable CarParts from more abstract CarFeatures such as *power, acceleration* and *drive*. Size: 7,643 tokens.
- Batch 005: Batch consists of JDPower car review files. Size: 42,019 tokens.
- Batch 006: Addition of Mention.Descriptor[6] for adjectives preceding mention nouns, such as ***heated***, ***power*** *seats*; MemberOf slot added to link individual mentions to a plural mention. Size: 95,864 tokens.
- Batch 007: Removal of Mention.Descriptor and addition of Descriptor class to reflect the fact that descriptors do not refer to discourse entities. Size: 11,221 tokens.
- Batch 008: Same format as Batch 007. Size: 30,612 tokens.

The annotations are stored as XML-encoded, stand-off mark-ups produced by the Protégé plug-in Knowtator (Ogren 2006), the tool which as used to annotate documents.

Inter-annotator agreement. Because of the subjective nature of sentiment annotations, annotation quality cannot be judged accurately by similarity to a gold standard. Instead, we validate the quality of our annotations by measuring the similarity of the annotations made by two people marking up the same document independently. Assessing inter-annotator agreement on the corpus involves analyzing several types of annotations: *spans*, *properties*, *span-span-links*, *span-entity-links*, and *entity-entity-links*.

Spans are markings of consecutive sequences of tokens. Annotators assign these spans one of the annotation types (cf. Section Annotation Types). We consider two spans to match if they have one overlapping token and are of the same annotation type. Spans might be annotated with properties. Two spans can still match even if they have conflicting property annotations. We explain how we assess inter-annotator agreement on properties shortly. For example, the span annotations, denoted by underlines, in Examples (15) and (16) match while those in Example (17) do not.

(15) a. My Honda Civic coupe...
b. My Honda Civic coupe...

(16) a. My Honda Civic coupe...
b. My Honda Civic coupe...

(17) a. My Honda Civic coupe...
b. My Honda Civic coupe...

To assess agreement on spans, we employ the agr metric, introduced by Wilson and Wiebe (2003), as a means of determining agreement of their subjective expression span annotations. $agr(A||B)$, where A and B are sets of spans marked by different annotators, gives the precision of A's annotations against B's. Formally, $arg(A||B) = \frac{|A \text{ matches } B|}{|A|}$.

Agreement on span properties (*properties*) is only measured on matching spans. Although Cohen's κ (Cohen 1960) has been used to measure inter-annotator agreement on nominal coding tasks such as this, our situation is complicated by heavily skewed distributions and the fact that multiple annotators have marked distinct sets of documents. Therefore, we only report observed agreement, or given annotators A and B, $obs(A, B) = \frac{|A \text{ matches } B|}{|A \cup B|}$. The final agreement score is the microaverage of all obs over all pairs of annotators, weighted by the number of properties annotated.

[6]Discussion of descriptors is omitted due to space constraints. See the annotation guidelines (Eckert et al. 2010) for details about this annotation.

Span-span-links are directed relations between spans (e.g., the target of a negator). Two span-links match if the sourced spans match and their destination spans match. Mismatches occur when there is a match between the originator spans, both spans have a span-link annotation of the same type, but link to non-matching spans. Agreement of one annotator, A, given another, B, is calculated by $agr(A||B) = \frac{|A \text{ matches } B|}{|A \text{ matches } B| + |A \text{ did not match } B|}$. To compute global statistics, we microaverage agr scores, weighting each by the number of times a relation occurred as a match or mismatch.

Span-entity-links are directed relations between a span and a co-reference group. For example, consider the target relation of a sentiment expression. While it is linked through the relation to a specific span, we are primarily interested in the co-reference group it targets and, thus consider the case when a matching span targets different mentions which belong to the same co-reference group as matching links. Mismatches occur when the entities linked are aligned but but the relation does not occur between them. Agreement of one annotator given another is calculated by $agr(A||B) = \frac{|A \text{ matches } B|}{|A \text{ matches } B| + |A \text{ did not match } B|}$. To compute global statistics, we microaverage agr scores, weighting each by the number of times a relation occurred as a match or mismatch.

Entity-entity-links function the same way as span-entity-links, however, it may match when the two source mentions are from matching coreference groups. For example, the part-of relation is treated as an entity-entity-link.

Comparison to other resources

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 the topic of each sentence is annotated and its contextual sentiment value is given. The sentences are drawn from online 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. They, as well as we, annotate sentences containing comparisons.

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.

Discussion

This corpus has been used to develop novel algorithms for finding targets of sentiment expressions (Kessler and Nicolov 2009) and we are aware of ongoing efforts for inducing co-reference systems (Sandra Keubler, p.c.; Shumin Wu, p.c.). Internally we have used this corpus to create statistical sentiment expression identification systems, a data-driven way for identifying topics and multi-word expressions associated with them.

The annotation types Descriptor and Comparisons are not discussed, and, while present in the annotation instructions, will be the topic of future papers. We currently have no way, besides the ContextualSentiment annotation of mentions, to account for issues such as tone and sarcasm. Recent work (Tsur, Davidov, and Rappoport 2010), makes inroads into addressing these difficult aspects of sentiment.

We are interested in annotating domains beyond automotive. So far we have annotated around 100,000 tokens in the consumer electronics domain (digital cameras) which we are also making available. We are also looking into creating multilingual resources within the same framework. We have meta-data about our documents including each post's URL and date which we will release in the future.

We have designed this corpus to be used as training and testing data for machine learning experiments. Detecting span annotations may be cast as sequence labeling (e.g., Breck, Choi, and Cardie (2007)) while detecting span properties may be simultaneously cast as an aspect of a sequence labeling problem (e.g., the semantic type of a named entity in named entity recognition) or as a separate task, along the lines of word-sense-disambiguation. Learning the refers-to relation can be cast as a coreference resolution problem (Ng and Cardie 2002). Systems to identify span-span-links can be be trained through supervised ranking. For example, Kessler and Nicolov (2009) used this technique to identify the targets of sentiment expressions in a previous version of the corpus, considering it a span-span relation. Entity-entity-links such as part-of relations can be identified through methods such as Girju, Badulescu, and Moldovan (2006).

Conclusion

In this paper we introduced a sentiment corpus with rich annotations, described the various annotation types and relations, presented statistics including inter-annotator agreement, and we cataloged components of sentiment that occur naturally. We also assessed their prevalence and found a very diverse form of linguistic expression that demonstrated many issues in semantics and discourse. We hope this corpus will be of interest to researchers building next-generation sentiment analysis systems.

Acknowledgments

We would like to thank Prof. Martha Palmer, Prof. James Martin, Prof. Michael Mozer at University of Colorado, and Prof. Michael Gasser at Indiana University and Dr. William Headden at J.D. Power and Associates for their helpful discussions.

References

Breck, E., and Cardie, C. 2004. Playing the telephone game: Determining the hierarchical structure of perspective and speech expressions. In *COLING*.

Annotation	Property	Type	Agreement	Matched
Mention	—	span	83%	21,518
Mention	Semantic Type	property	83%	17,923
Mention	MentionPriorPolarity	property	100%	7
Mention	ContextualSentiment	property	95%	13
Mention	EntitySentiment[1]	property	85%	87
Mention	Inferred Contextual Sentiment[2]	property	87%	18,706
Mention	Refers-to	span-entity-link	68%	5,684
Mention	Part-of	entity-entity-link	35%	1,178
Mention	Feature-of	entity-entity-link	23%	294
Mention	Member-of	entity-entity-link	81%	34
Mention	Instance-of	entity-entity-link	73%	184
SentimentExpression	—	span	75%	3,976
SentimentExpression	PriorPolarity	property	95%	3,712
SentimentExpression	Target	span-entity-link	66%	2,879
Negator	—	span	66%	384
Negator	NegatorTarget	span-span-link	85%	335
Neutralizer	—	span	36%	70
Neutralizer	NeutralizerTarget	span-span-link	78%	64
Intensifier	—	span	60%	729
Intensifier	IntensifierDirection	property	96%	690
Intensifier	IntensifierTarget	span-span-link	95%	737
Committer	—	span	33%	93
Committer	CommitterDirection	property	91%	79
Committer	CommitterTarget	span-span-link	82%	75
OPO	—	span	33%	93
OPO	OPOTarget	span-span-link	66%	383
OPO	OPOSource	span-entity-link	88%	132

[1]Because this is a span property, matches are only counted when both annotators marked EntitySentiment toward matching mentions.

[2]This was automatically determined through a heuristic that accounted for targeting sentiment expressions, modifiers, and annotated prior polarity or contextual sentiment.

Table 3: Inter-annotator agreement on annotation types and their properties.

Breck, E.; Choi, Y.; and Cardie, C. 2007. Identifying expressions of opinion in context. In *IJCAI*.

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

Choi, Y.; Kim, Y.; and Myaeng, S.-H. 2009. Domain-specific sentiment analysis using contextual feature generation. In *TSA*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Ding, X.; Liu, B.; and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM*.

Eckert, M.; Clark, L.; Lind, H.; Kessler, J.; and Nicolov, N. 2010. *Structural Sentiment and Entity Annotation Guidelines*. J. D. Power and Associates Technical Report.

Fahrni, A., and Klenner, M. 2008. Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In *AISB*.

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*.

Karttunen, L., and Zaenen, A. 2005. Veridicity. In *Annotating, Extracting and Reasoning about Time and Events*.

Kessler, J. S., and Nicolov, N. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *ICWSM*.

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

Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. In *COLING*.

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*.

Krestel, R.; Witte, R.; and Bergler, S. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *LREC*.

Moilanen, K., and Pulman, S. 2009. Multi-entity sentiment scoring. In *RANLP*.

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

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

Nowson, S. 2009. Scary movies good, scary flights bad: Topic driven feature selection for classification of sentiment. In *TSA*.

Ogren, P. V. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *NAACL-HLT*.

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

Pustejovsky, J.; Hanks, P.; Sauri, R.; See, A.; Gaizauskas, R.; Setzer, A.; Radev, D.; Sundheim, B.; Day, D.; Ferro, L.; and Lazo, M. 2003. The timebank corpus. In *Corpus Linguistics*.

Rubin, V. L. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *NAACL-HLT*.

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

Shaikh, M. A. M.; Prendinger, H.; and Ishizuka, M. 2008. Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Appl. Artif. Intell.* 22(6).

Su, F., and Markert, K. 2008. From words to senses: a case study of subjectivity recognition. In *COLING*.

Tsur, O.; Davidov, D.; and Rappoport, A. 2010. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in product reviews. In *ICWSM*.

Wiebe, J., and Mihalcea, R. 2006. Word sense and subjectivity. In *ACL*.

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

Wiegand, M., and Klakow, D. 2009. Topic-related polarity classification of blog sentences. In *EPIA*.

Wilson, T., and Wiebe, J. 2003. Annotating opinions in the world press. In *SIGdial*.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2009. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics* 35(3).

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.

Winston, M. E.; Chaffin, R.; and Herrmann, D. 1987. A taxonomy of part-whole relations. *Cognitive Science* 11(4).