Identifying Types of Claims in Online Customer Reviews

Shilpa Arora, Mahesh Joshi and **Carolyn P. Rosé**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University, Pittsburgh PA 15213
{shilpaa,maheshj,cprose}@cs.cmu.edu

Abstract

In this paper we present a novel approach to categorizing comments in online reviews as either a *qualified claim* or a *bald claim*. We argue that this distinction is important based on a study of customer behavior in making purchasing decisions using online reviews. We present results of a supervised algorithm for learning this distinction. The two types of claims are expressed differently in language and we show that syntactic features capture this difference, yielding improvement over a bag-of-words baseline.

1 Introduction

There has been tremendous recent interest in opinion mining from online product reviews and it's effect on customer purchasing behavior. In this work, we present a novel alternative categorization of comments in online reviews as either being *qualified claims* or *bald claims*.

Comments in a review are claims that reviewers make about the products they purchase. A customer reads the reviews to help him/her make a purchasing decision. However, comments are often open to interpretation. For example, a simple comment like *this camera is small* is open to interpretation until qualified by more information about whether it is small in general (for example, based on a poll from a collection of people), or whether it is small compared to some other object. We call such claims *bald claims*. Customers hesitate to rely on such bald claims unless they identify (from the context or otherwise) themselves to be in a situation similar to the customer who posted the comment. The other category of claims that are not bald are *qualified claims*. Qualified claims such as *it is small enough to fit easily in a coat pocket or purse* are more precise claims as they give the reader more details, and are less open to interpretation. Our notion of qualified claims is similar to that proposed in the argumentation literature by Toulmin (1958). This distinction of qualified vs. bald claims can be used to filter out bald claims that can't be verified. For the qualified claims, the qualifier can be used in personalizing what is presented to the reader.

The main contributions of this work are: (i) an annotation scheme that distinguishes qualified claims from bald claims in online reviews, and (ii) a supervised machine learning approach that uses syntactic features to learn this distinction. In the remainder of the paper, we first motivate our work based on a customer behavior study. We then describe the proposed annotation scheme, followed by our supervised learning approach. We conclude the paper with a discussion of our results.

2 Customer Behavior Study

In order to study how online product reviews are used to make purchasing decisions, we conducted a user study. The study involved 16 pair of graduate students. In each pair there was a customer and an observer. The goal of the customer was to decide which camera he/she would purchase using a camera review blog[1] to inform his/her decision. As the customer read through the reviews, he/she was

[1] http://www.retrevo.com/s/camera

Proceedings of NAACL HLT 2009: Short Papers, pages 37–40,
Boulder, Colorado, June 2009.

asked to think aloud and the observer recorded their observations.

The website used for this study had two types of reviews: expert and user reviews. There were mixed opinions about which type of reviews people wanted to read. About six customers could relate more with user reviews as they felt expert reviews were more like a 'sales pitch'. On the other hand, about five people were interested in only expert reviews as they believed them to be more practical and well reasoned.

From this study, it was clear that the customers were sensitive to whether a claim was qualified or not. About 50% of the customers were concerned about the reliability of the comments and whether it applied to them. Half of them felt it was hard to comprehend whether the user criticizing a feature was doing so out of personal bias or if it represented a real concern applicable to everyone. The other half liked to see comments backed up with facts or explanations, to judge if the claim could be qualified. Two customers expressed interest in comments from users similar to themselves as they felt they could base their decision on such comments more reliably. Also, exaggerations in reviews were deemed untrustworthy by at least three customers.

3 Annotation Scheme

We now present the guidelines we used to distinguish bald claims from qualified claims. A claim is called qualified if its validity or scope is limited by making the conditions of its applicability more explicit. It could be either a fact or a statement that is well-defined and attributed to some source. For example, the following comments from our data are qualified claims according to our definition,

1. *The camera comes with a lexar 16mb starter card, which stores about 10 images in fine mode at the highest resolution.*
2. *I sent my camera to nikon for servicing, took them a whole 6 weeks to diagnose the problem.*
3. *I find this to be a great feature.*

The first example is a fact about the camera. The second example is a report of an event. The third example is a self-attributed opinion of the reviewer.

Bald claims on the other hand are non-factual claims that are open to interpretation and thus cannot be verified. A straightforward example of the distinction between a bald claim and a qualified claim is a comment like *the new flavor of peanut butter is being well appreciated* vs. *from a survey conducted among 20 people, 80% of the people liked the new flavor of peanut butter.* We now present some examples of bald claims. A more detailed explanation is provided in the annotation manual[2]:

- **Not quantifiable gradable**[3] words such as *good*, *better*, *best* etc. usually make a claim bald, as there is no qualified definition of being good or better.
- **Quantifiable gradable** words such as *small*, *hot* etc. make a claim bald when used without any frame of reference. For example, a comment *this desk is small* is a bald claim whereas *this desk is smaller than what I had earlier* is a qualified claim, since the comparative *smaller* can be verified by observation or actual measurement, but whether something is *small* in general is open to interpretation.
- **Unattributed opinion or belief:** A comment that implicitly expresses an opinion or belief without qualifying it with an explicit attribution is a bald claim. For example, *Expectation is that camera automatically figures out when to use the flash*.
- **Exaggerations:** Exaggerations such as *on every visit, the food has blown us away* do not have a well defined scope and hence are not well qualified.

The two categories for gradable words defined above are similar to what Chen (2008) describes as *vagueness, non-objective measurability and imprecision*.

4 Related work

Initial work by Hu and Liu (2004) on the product review data that we have used in this paper focuses on the task of opinion mining. They propose an approach to summarize product reviews by identifying opinionated statements about the features of a product. In our annotation scheme however, we classify

[2] www.cs.cmu.edu/~shilpaa/datasets/opinion-claims/qbclaims-manual-v1.0.pdf

[3] http://en.wikipedia.org/wiki/English_grammar#Semantic_gradability

all claims in a review, not restricting to comments with feature mentions alone.

Our task is related to opinion mining, but with a specific focus on categorizing statements as either bald claims that are open to interpretation and may not apply to a wide customer base, versus qualified claims that limit their scope by making some assumptions explicit. Research in analyzing subjectivity of text by Wiebe et al. (2005) involves identifying expression of private states that cannot be objectively verified (and are therefore open to interpretation). However, our task differs from subjectivity analysis, since both bald as well as qualified claims can involve subjective language. Specifically, objective statements are always categorized as qualified claims, but subjective statements can be either bald or qualified claims. Work by Kim and Hovy (2006) involves extracting pros and cons from customer reviews and as in the case of our task, these pros and cons can be either subjective or objective.

In supervised machine learning approaches to opinion mining, the results using longer n-grams and syntactic knowledge as features have been both positive as well as negative (Gamon, 2004; Dave et al., 2003). In our work, we show that the qualified vs. bald claims distinction can benefit from using syntactic features.

5 Data and Annotation Procedure

We applied our annotation scheme to the product review dataset[4] released by Hu and Liu (2004). We annotated the data for 3 out of 5 products. Each comment in the review is evaluated as being qualified or bald claim. The data has been made available for research purposes[5].

The data was completely double coded such that each review comment received a code from the two annotators. For a total of $1,252$ review comments, the Cohen's kappa (Cohen, 1960) agreement was 0.465. On a separate dataset (365 review comments)[6], we evaluated our agreement after removing the borderline cases (only about 14%) and there was a statistically significant improvement in kappa to 0.532. Since the agreement was low, we resolved our conflict by consensus coding on the data that was used for supervised learning experiments.

6 Experiments and Results

For our supervised machine learning experiments on automatic classification of comments as qualified or bald, we used the Support Vector Machine classifier in the MinorThird toolkit (Cohen, 2004) with the default linear kernel. We report average classification accuracy and average Cohen's Kappa using 10-fold cross-validation.

6.1 Features

We experimented with several different features including standard lexical features such as word unigrams and bigrams; pseudo-syntactic features such as Part-of-Speech bigrams and syntactic features such as dependency triples[7]. Finally, we also used syntactic scope relationships computed using the dependency triples. Use of features based on syntactic scope is motivated by the difference in how qualified and bald claims are expressed in language. We expect these features to capture the presence or absence of qualifiers for a stated claim. For example, "*I didn't like this camera, but I suspect it will be a great camera for first timers.*" is a qualified claim, whereas a comment like "*It will be a great camera for first timers.*" is not a qualified claim. Analysis of the syntactic parse of the two comments shows that in the first comment the word "great" is in the scope of "suspect", whereas this is not the case for the second comment. We believe such distinctions can be helpful for our task.

We compute an approximation to the syntactic scope using dependency parse relations. Given the set of dependency relations of the form ≪relation, headWord, dependentWord≫, such as ≪AMOD,camera,great≫, an in-scope feature is defined as INSCOPE_headWord_dependentWord (INSCOPE_camera_great). We then compute a transitive closure of such in-scope features, similar to Bikel and Castelli (2008). For each in-scope feature in the entire training fold, we also create a corre-

[4] http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip

[5] www.cs.cmu.edu/~shilpaa/datasets/opinion-claims/qbclaims-v1.0.tar.gz

[6] These are also from the Hu and Liu (2004) dataset, but not included in our dataset yet.

[7] We use the Stanford Part-of-Speech tagger and parser respectively.

Features	QBCLAIM	HL-OP
Majority	.694(.000)	.531(.000)
Unigrams	.706(.310)	.683(.359)
+Bigrams	.709(.321)	.693(.378)
+POS-Bigrams	**.726*(.353*)**	.683(.361)
+Dep-Triples	.711(.337)	.692(.376)
+In-scope	.706(.340)	.688(.367)
+Not-in-scope	.726(**.360***)	.687(.370)
+All-scope	.721(.348)	.699(.396)

Table 1: The table shows accuracy (& Cohen's kappa in parentheses) averaged across ten folds. Each feature set is individually added to the baseline set of unigram features. * - Result is marginally significantly better than unigrams-only ($p < 0.10$, using a two-sided pairwise T-test). HL-OP - Opinion annotations from Hu and Liu (2004). QBCLAIM - Qualified/Bald Claim.

sponding not-in-scope feature which triggers when either (i) the dependent word appears in a comment, but not in the transitive-closured scope of the head word, or (ii) the head word is not contained in the comment but the dependent word is present.

We evaluate the benefit of each type of feature by adding them individually to the baseline set of unigram features. Table 1 presents the results. We use the majority classifier and unigrams-only performance as our baselines. We also experimented with using the same feature combinations to learn the *opinion* category as defined by Hu and Liu (2004) [HL-OP] on the same subset of data.

It can be seen from Table 1 that using purely unigram features, the accuracy obtained is not any better than the majority classifier for qualified vs. bald distinction. However, the Part-of-Speech bigram features and the not-in-scope features achieve a marginally significant improvement over the unigrams-only baseline.

For the opinion dimension from Hu and Liu (2004), there was no significant improvement from the type of syntactic features we experimented with. Hu and Liu (2004)'s opinion category covers several different types of opinions and hence finer linguistic distinctions that help in distinguishing qualified claims from bald claims may not apply in that case.

7 Conclusions

In this work, we presented a novel approach to review mining by treating comments in reviews as claims that are either qualified or bald. We argued with examples and results from a user study as to why this distinction is important. We also proposed and motivated the use of syntactic scope as an additional type of syntactic feature, apart from those already used in opinion mining literature. Our evaluation demonstrates a marginally significant positive effect of a feature space that includes these and other syntactic features over the purely unigram-based feature space.

Acknowledgments

We would like to thank Dr. Eric Nyberg for the helpful discussions and the user interface for doing the annotations. We would also like to thank all the anonymous reviewers for their helpful comments.

References

Daniel Bikel and Vittorio Castelli. *Event Matching Using the Transitive Closure of Dependency Relations*. Proceedings of ACL-08: HLT, Short Papers, pp. 145–148.

Wei Chen. 2008. *Dimensions of Subjectivity in Natural Language*. In Proceedings of ACL-HLT'08. Columbus Ohio.

Jacob Cohen. 1960. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement, Vol. 20, No. 1., pp. 37-46.

William Cohen. 2004. *Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data*. http://minorthird.sourceforge.net/

Kushal Dave, Steve Lawrence and David M. Pennock 2006. *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*. In Proc of WWW'03.

Michael Gamon. *Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis*. Proceedings of the International Conference on Computational Linguistics (COLING).

Minqing Hu and Bing Liu. 2004. *Mining and Summarizing Customer Reviews*. In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Soo-Min Kim and Eduard Hovy. 2006. *Automatic Identification of Pro and Con Reasons in Online Reviews*. In Proc. of the COLING/ACL Main Conference Poster Sessions.

Stephen Toulmin 1958 *The Uses of Argument*. Cambridge University Press.

Janyce Wiebe, Theresa Wilson, and Claire Cardie 2005. *Annotating expressions of opinions and emotions in language*. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.