Coherence Relations

Predicting the Helpfulness of Product Reviews Using Discourse Relations

Sebastian Golly

(761737)

August 22, 2017

Abstract Previous studies on automatically assessing the helpfulness of product reviews have largely neglegted the impact of discourse structure. This paper investigates whether the distribution of discourse relations in a product review, that is the presence and occurrence frequencies of different discourse-relation types, can serve as an indicator of its helpfulness. To this end, a probabilistic baseline model is enriched by discourse-relation features. Experimental results show that the presence of certain discourse-relation types in product reviews has, in fact, an effect on their perceived helpfulness to other users.

Reference Style: APA

University of Potsdam

Summer 2017

1 Introduction

Many e-commerce websites contain large amounts of user-authored product reviews. They are an "important source of information for making informed purchase decisions" (Almagrabi et al. 2015: 55). Some of them are more helpful to other users, while others are less.

Although most people seem to know intuitively what it means to be a *helpful* product review, there is no generally accepted theoretical definition of review helpfulness. Following Mudambi & Schuff (2010: 186), a helpful review is "a peer-generated product evaluation that facilitates the consumer's purchase decision process." Along these lines, review helpfulness can be seen as "a measure of perceived value in the decision-making process" (ibid.).

Online retailers have an interest in displaying more helpful reviews more prominently than less helpful ones in order to create a benefit for their users. For this purpose, they require some procedure for evaluating review helpfulness. Currently, helpfulness is mainly assessed manually by other users who are asked for their votes by posing a question like "Was this review helpful to you?" This kind of manual assessment, however, suffers from a number of limitations (cf. Almagrabi et al. 2015: 49f.), including:

- *Sparseness:* Many reviews have received no helpfulness votes at all; many more have not received enough votes to reliably compute their helpfulness.
- *No instant evaluation:* After a review has been posted, it takes a significant amount of time until a sufficient number of other users have evaluated it. As a consequence, it is impossible to immediately rank it according to its helpfulness.
- *Biases:* Human evaluations are subject to a number of biases. For instance, reviews with many positive helpfulness votes are displayed prominently, which results in being read by many users and receiving even more helpfulness votes. This "winner circle" makes it hard for new reviews to be ranked appropriately.

To overcome these limitations, it makes sense to find solutions for evaluating review helpfulness automatically. By using a probabilistic model specializing in this task, websites would be able to obtain a helpfulness assessment for every single review, in the moment it is posted, and at least less affected by human biases.

Previous studies on automatically predicting review helpfulness have taken into account different kinds of indicators. However, discourse structure has largely been neglected. The present study aims at filling this

gap by investigating whether the distribution of discourse relations in a product review can serve as an indicator of its helpfulness. Concretely, a state-of-the-art baseline model is enriched by discourse-relation features to examine whether the additional features improve its performance.

Having introduced the topic, section 2 outlines relevant previous studies, section 3 presents the research objective of this work, and section 4 describes the methodology to investigate this objective. The experimental results are presented in section 5 and discussed in section 6. Finally, section 7 briefly summarizes and concludes the study.

2 Previous Work

There exist different frameworks for capturing discourse structure. The one adopted in this work is the Penn Discourse Treebank 2.0 (PDTB) framework, as presented in Prasad et al. (2008). PDTB provides a lexically grounded approach, that is, each relation is signaled by a discourse connective. In case a relation holds between discourse units that is not explicitly signaled, an implicit connective is annotated.

PDTB only describes low-level relations between atomic discourse segments, not relations between larger spans of text. Also, it does not try to cover the complete discourse, but only those parts of it where an explicit connective can be found or an implicit one can be hypothesized.

As discourse connectives are often ambiguous in meaning (*since*, for example, can have both temporal and causal semantics), PDTB provides an annotation of discourse-relation types, so-called *senses*, used for disambiguation. These sense tags are structured in a three-level hierarchy where nested tags inherit the properties of their parents (cf. Prasad et al. 2008: 5). The subtypes *Reason* and *Result*, for instance, both inherit from the type *Cause*, which, in turn, is a child of the *Contingency* class.

There have been a number of studies working towards the goal of automatically assessing the helpfulness of product reviews (cf. Almagrabi et al. 2015 for a survey). A very influential one is that by Kim et al. (2006) who propose a Support Vector Machine (SVM) regression model for predicting review helpfulness. They systematically investigate how well different types of features capture the helpfulness of a review, considering the following classes:

- *structural* features (e.g., the number of tokens in the review)
- *lexical* features (e.g., its *n-gram* statistics)

- *syntactic* features (e.g., part-of-speech analyses)
- *semantic* features (e.g., sentiment analyses)
- *meta-data* features (e.g., the overall product rating the review gives)

In their experiments, Kim et al. (2006) evaluate different combinations of features. The feature combination that performs best captures the length of the review, its unigram statistic, and its product rating.

Mertz et al. (2014) build upon the model proposed by Kim et al. (2006), evaluating the use of discourse connectives as an additional feature. In their experiment, however, this does not yield a significant improvement in predicting review helpfulness. This might be because they set up their experiment as a classification task in which their baseline model already obtains "classification accuracies between eighty and ninety percent" (Mertz et al. 2014: 8). They suggest that a less restrictive setup might yield more convincing results.

Another shortcoming in the work of Mertz et al. (2014) is that they use simple regular-expression matching for extracting discourse connectives. They state that the "frequencies of the individual discourse connectives are not used directly as features, but aggregated into senses first" (Mertz et al. 2014: 7), corresponding to the PDTB hierarchy of sense tags. Considering the vast ambiguity in the mapping from discourse connectives to senses, however, it remains unclear how they perform this aggregation without applying proper discourse parsing.

3 Goal of this Study

Previous work on predicting the helpfulness of product reviews has used different classes of features, including structural, lexical, syntactic, semantic, and meta-data features (cf. Kim et al. 2006: 424f.; Almagrabi et al. 2015).

With the exception of Mertz et al. (2014), discourse structure has not been taken into account as an indicator for this task. However, it seems plausible that the way statements in a review are presented (e.g., justified, elaborated, contrasted) and interconnected (e.g., by causal or conceding relations) affects as how comprehensible and credible it is considered by other users, and consequently has an impact on its perceived value in their purchase-decision making process – its helpfulness.

Concretely, the aim of this study is to build upon the best-performing model of Kim et al. (2006), enrich

it with features capturing the presence and occurrence frequencies of different discourse-relation types, and evaluate whether the additional features yield a significant improvement.

Applying these operationalizations to the aforementioned expectations, the hypothesis of this study can be phrased as follows:

> The performance of a probabiblistic model predicting the helpfulness of product reviews can be improved by adding features capturing the distribution of discourse relations, that is the presence and occurrence frequencies of different discourse-relation types.

4 Method

This section descibes the dataset used in the experiment, the learning task and features of the probabilistic model, as well as the overall experimental setup.

4.1 Data

For learning and evaluating the probabilistic model, I used the 5-core corpus of *Amazon.com* reviews on products from the category "Electronics" created by He & McAuley (2016)[1]. For each of its 1,689,188 reviews spanning May 1996 to July 2014, it includes, among others:

- the text of the review,
- its overall product rating (1 to 5 stars),
- the numbers of positive and total helpfulness votes it has received.

Listing 1 shows a shortened sample review from the corpus.

In order to use only informative items, reviews with less than 10 helpfulness votes or with a text shorter than 20 characters have been filtered out. From the remaining 107,035 items, 20,000 reviews have been randomly sampled and partitioned into a development dataset containing 2,000 reviews and a cross-validation dataset containing 18,000 reviews.[2]

[1] Available at `http://jmcauley.ucsd.edu/data/amazon/`.
[2] Available at `https://drive.google.com/open?id=0B4FHGozCmQFEQkJmeWFTeUJKVEE`.

```
{
   "asin": "B000W9OJVA",
   "helpful": [
      103,
      105
   ],
   "overall": 4.0,
   "reviewText": "I recently purchased a set of these. Personal video
      glasses are a difficult item to purchase because it's very hard to
      find a store that carries them, let alone lets you try them on
      before purchase, so I wasn't quite sure what to expect. [...]",
   "reviewTime": "03 27, 2008",
   "reviewerID": "AVPNQUVZWMDSX",
   "reviewerName": "esanta \"esanta\"",
   "summary": "Fantastic!",
   "unixReviewTime": 1206576000
}
```


Listing 1: Sample review (shortened).

4.2 Learning Task

Helpfulness prediction is sometimes seen as a classification problem, labeling each review as either *helpful* or *not helpful* (cf. Almagrabi et al. 2015: 3ff.). In order to allow for ranking reviews according to their helpfulness, however, it makes sense to rather see the problem as a regression task in which the probabilistic model assigns a score to each review that captures its helpfulness.

In order to determine this helpfulness score, Kim et al. (2006: 424) propose the simple helpfulness function (1) that computes the helpfulness of a review r as the ratio of its positive helpfulness votes to its total helpfulness votes.

$$h(r \in R) = \frac{rating_{+}(r)}{rating_{+}(r) + rating_{-}(r)} \tag{1}$$

This helpfulness ratio is defined for every review with at least one helpfulness vote. It ranges between 0 (minimum helpfulness) and 1 (maximum helpfulness). For training and evaluation, the gold standard of helpfulness scores is determined based on the manually assessed helpfulness votes indicated in the corpus.

To summarize, given a collection of features of a product review, the task of the probabilistic regression model will be to predict its helpfulness score.

4.3 Features

In order to answer the research question whether discourse-relation features improve the performance of a model predicting review helpfulness, I implemented a baseline model with a standard set of features, enriched it by features capturing discourse relations, and finally compared the results of the models. The individual features are described below.

The baseline model implements the combination of features that had been shown to perform best according to Kim et al. (2006: 428):

- `STR`: the overall rating the review gives to the product (1-5 stars),
- `LEN`: the total number of tokens in the review text,
- `UGR`: the *tf-idf* statistic of each token occurring in the review.

In order to extract discourse relations from the product reviews, the *PDTB-styled end-to-end discourse parser* (Lin et al. 2014) was used. It can be configured to extract sense tags at any of the three levels of the PDTB sense hierarchy (cf. Prasad et al. 2008: 5). As a reasonable compromise between accuracy and parsing performance, sense tags from the second (*type*) level have been extracted (e.g., `Comparison.Contrast` or `Contingency.Cause`). Because of the unsatisfactory performance of the non-explicit classifier (cf. Lin et al. 2014: 175), only explicitly signalled relations were used in this experiment.

Based on the extracted occurrences of discourse relations in each review, two variants of discourse-relation features have been aggregated:

- `REL-CNT`: occurrence frequencies of explicit discourse-relation types in each review (listing 2 shows an instance of this feature for a sample review),
- `REL-PRS`: presence of explicit discourse-relation types in each review (with values `1` or `0` according to whether or not a certain discourse-relation type occurs in the review).

4.4 Experimental Setup

The setup of this experiment is similar to that of Kim et al. (2006). It relies on a Support Vector Machine (SVM) regression model using the radial basis function (RBF) kernel. The hyperparameters of the model, C (the penalty parameter) and γ (the kernel-width parameter), have been tuned performing full-grid search on the development dataset. Each feature has been scaled to the range $[-1, 1]$.

```
{
    'Comparison.Concession': 0,
    'Comparison.Contrast': 13,
    'Comparison.Pragmatic concession': 0,
    'Comparison.Pragmatic contrast': 0,
    'Contingency.Cause': 8,
    'Contingency.Condition': 2,
    'Contingency.Pragmatic condition': 0,
    'Expansion.Alternative': 0,
    'Expansion.Conjunction': 17,
    'Expansion.Exception': 0,
    'Expansion.Instantiation': 0,
    'Expansion.List': 0,
    'Expansion.Restatement': 0,
    'Temporal.Asynchronous': 8,
    'Temporal.Synchrony': 6
}
```

Listing 2: REL-CNT feature for a sample review.

In contrast to the experiment of Kim et al. (2006) that uses helpfulness scores only to learn the *ranks* of each review on a product, this study aims at learning the helpfulness scores themselves, as the rankings can be easily derived from them and, from an application perspective, a model predicting scores rather than ranks is more practical when inserting a newly created review into a list of existing reviews on the same product.

For evaluating the results, 10-fold cross-validation has been used, where each model was trained using 9 folds, and its performance was evaluated on the remaining test fold. As an evaluation metric, the Pearson correlation coefficient (*Pearson's r*) between the predicted helpfulness scores and the gold standard (based on the *Amazon.com* helpfulness votes) has been computed. It ranges between -1 (total negative correlation) and +1 (total positive correlation).

5 Results

Table 1 shows the evaluation results of the different feature combinations in terms of their Pearson correlation coefficients.

The baseline model including the overall rating (STR), review length (LEN), and unigram statistic features (UGR) already yields convincing results with a clearly positive correlation. Adding the discourse-relation counts feature (REL-CNT) does not affect the performance of the model at all. However, adding the *presence*

Feature Combination	Pearson's r[a]
STR, LEN, UGR	0.560 (± 0.042)
STR, LEN, UGR, REL-CNT	0.560 (± 0.041)
STR, LEN, UGR, REL-PRS	**0.574 (± 0.040)**

[a] 95% confidence bounds are calculated using 10-fold cross-validation.

Table 1: Evaluation results of the baseline model and the two variants including discourse-relation features.

of discourse-relation types instead (REL-PRS) outperforms the baseline model. This difference between the baseline and the REL-PRS model turns out to be significant, with $p < .05$[3].

6 Discussion

6.1 Implications

The results show that, also in this experimental setup, the established baseline model initially proposed by Kim et al. (2006) obtains a fairly good performance. Its features, however, do not reflect the way the statements in a review are presented and interconnected.

The proposed discourse-relation features are an attempt to capture these properties of discourse structure, at least to some extent. The evoked gain in prediction performance supports the hypothesis that the distribution of discourse relations in a product review is an indicator of its helpfulness to other users – at least if the term "distribution" is defined as the *presence* of certain relation types, rather than their occurrence frequencies.

In contrast to Mertz et al. (2014), adding discourse-relation features yields a statistically significant effect on the performance of the model. This disparity is likely due to differences in the experimental setups: While the task of predicting review helpfulness is considered a regression problem in this study, Mertz et al. (2014) treat it as a (simpler) classification task. They set up their experiment in a way that a "simple baseline already obtains incredibly good results" (Mertz et al. 2014: 8), making it hard for additional features to outperform the baseline.

The observation that relation-type *counts* do not yield any effect on the performance of the model is probably rather a consequence of technical circumstances (such as feature scaling) than deep theoretical

[3] Calculated with http://vassarstats.net/rdiff.html.

reasons.

The obtained results only represent a first step towards investigating the impact of discourse structure on review helpfulness. Applying more powerful discourse parsers and more sophisticated metrics for capturing discourse structure is likely to further improve prediction performance.

6.2 Limitations

While the results suggest that discourse structure has, in fact, an impact on the perceived helpfulness of a product review, it has to be stressed that, for two reasons, the signals analyzed in this study do not provide a comprehensive account of discourse structure.

First, implicitly signalled relations had to be ignored completely because of the poor performance of current discourse parsers in automatically detecting them. However, there is no theoretical foundation to the claim that implicit relations would have a smaller impact on review helpfulness than explicit ones.

Second, following the PDTB framework, only the most local level of discourse structure, the relations between atomic discourse segments, has been taken into account. Nonetheless, one might assume that higher-level discourse relations, holding between larger spans of text, equally affect the perceived value of a product review.

6.3 Future Work

This study provides evidence for the general claim that discourse structure influences the helpfulness of product reviews. However, there are a number of questions that remain unanswered and might serve as a starting point for further research.

One of them focuses on the impact of individual discourse-relation types on review helpfulness. Deeper insights about their particular contributions could be gained by implementing a probabilistic model that, other than the SVM regression model with a non-linear kernel used in this work, allows for feature-weight analysis.

Mudambi & Schuff (2010) find that reviews on different product types are evaluated differently with respect to their helpfulness. It would be interesting to investigate whether the effects of discourse structure vary across product categories, e.g., when comparing electronic products, as in this study, to books.

Another interesting research direction would be finding out about how exactly the presence of certain discourse relations influences the perceived value of a product review. A participant-based study might help to understand whether the use of specific discourse relations affects the review's comprehensibility, its credibility, or any other factors in the perception of other users.

Also, future studies in this field should take advantage of any improvements made in automatic discourse parsing, as these will allow to capture discourse structure more comprehensively.

7 Conclusion

Finding a reliable solution for automatically predicting the perceived helpfulness of product reviews is expected to bring about various advantages, both for online retailers and their customers. This study aimed to learn about the effect of discourse structure on review helpfulness, an aspect that had been largely neglected in previous work on this task.

In order to investigate whether the presence and occurrence frequencies of different discourse-relation types influcence review helpfulness, an established probabilistic model has been enriched by features capturing these properties. The results suggest that the distribution of discourse relations in a product review is, in fact, an indicator of its helpfulness to other users. How exactly discourse structure is related to review helpfulness will be subject of future research.

Both code and data used for this study are freely available for research purposes at

`https://github.com/s-go/cr-review-helpfulness`.

References

Almagrabi, H., Malibari, A., & McNaught, J. (2015). A Survey of Quality Prediction of Product Reviews. In *International Journal of Advanced Computer Science & Applications*, 1(6), 49-58.

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 507-517.

Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 423-430.

Lin, Z., Ng, H. T., & Kan, M. Y. (2014). A PDTB-styled end-to-end discourse parser. In *Natural Language Engineering*, 20(2), 151-184.

Mertz, M., Korfiatis, N., & Zicari, R. V. (2014). Using Dependency Bigrams and Discourse Connectives for Predicting the Helpfulness of Online Reviews. In *International Conference on Electronic Commerce and Web Technologies*, Springer International Publishing, 146-152.

Mudambi, S. M., & Schuff, D. (2010). What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. In *MIS Quarterly*, 34(1), 185-200.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). *The Penn Discourse Treebank 2.0.* Paper presented at the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesh, Morocco.