Breaking Sentiment Analysis of Movie Reviews

Ieva Staliūnaitė and **Ben Bonfil**
Utrecht University
i.r.staliunaite@students.uu.nl

Abstract

The current paper covers several strategies we used to 'break' predictions of sentiment analysis systems participating in the BLGNLP2017 workshop. Specifically, we identify difficulties of participating systems in understanding modals, subjective judgments, world-knowledge based references and certain differences in syntax and perspective.

1 Introduction

Participants in the BLGNLP2017 shared task were invited to either build sentiment analysis systems (as a *Builder* team) or break them, by compiling linguistically motivated test cases that result in false predictions (as a *Breaker* team). A data set of movie reviews was provided as the domain for participating systems and as a source for generating breaking test cases. As a Breaker team, our goal was to construct minimal pairs consisting of a review from the source data set, and a modified version of the review that would be used to evaluate the robustness or sensitivity of the participating systems predictions. The modified version of each review could either preserve the sentiment of the original review, or reverse it.

Movie reviews from *Rotten Tomatoes* are a good source for comments full of sentiment, as the informal setting provides for humor, pathos, wild comparisons, sarcasm, artistic expressions and the like. Hence, it was probably not an easy task for the Builder systems to analyze sentiments to begin with, and we tried to make it even harder. Based on the sentiment analysis of our linguistic examples it seems like there are several ways to trick the Builder systems.

In our own judgments of the provided items, we followed a positive/negative sentiment dichotomy, which was not always straightforward given the complexity of the data. However, even if a neutral sentiment option was included (as found in the predictions of some of the Builder system) it would not have accounted for the whole variation, as some items could have multiple plausible interpretations, affecting their perceived valence. Thus, it is important to bear in mind that the judgments provided by us might not always coincide with those of other people.

We begin this paper by describing the general rationale we had employed in creating our test cases. We then present some examples of sentences that broke the Builder systems and discuss the nature of the errors, and the main difficulties in analyzing sentiment. In addition, we discuss the linguistic processes that take place in inferring sentiment from the various examples.

2 Breaking Strategy

Our approach to judging sentiment was based on two implicit questions: "Would I watch this movie based on the comment?" and "Would the comment be likely accompanied by a five star evaluation?" Thus, our judgment relied on reviewers' description of enjoyment and quality as measurements of sentiment. In making the minimal pairs we employed a number of different strategies. We used our intuitions and knowledge of syntax, semantics and pragmatics in order to make big differences in meaning with superficially small changes. We tried to make realistic examples of movie evaluations, focusing on linguistic inferences that a machine might not be able to do.

When it comes to breaking predictions, the largest number of errors appeared with examples involving words that bear judgment, but are not inherently positive or negative on their own. We used modals and opinion adverbs to contribute to

Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, pages 61–64
Copenhagen, Denmark, September 8, 2017. ©2017 Association for Computational Linguistics

the meaning of a phrase by providing information about the speaker's subjective stance. For instance, adverbs such as 'too', 'enough', 'hardly', 'supposedly', 'barely', 'seldom', 'rarely' and 'finally' all convey a relative stance in certain contexts. The use of such expressions changes the construal of the sentence so that the perspective of the subject of consciousness is foregrounded (Verhagen et al., 2007). Therefore, including such an adverb can change the valence of the sentence, such as in examples (1) and (2) below. While the truth-conditions of (2) would not change if 'hardly' was substituted with 'a little', the judgment of the speaker would disappear. Hence, the sentiment in this example is expressed by foregrounding the speaker's evaluation of the extent of the difference between the two types of movies. While most Builder systems classified (1) as positive, just about half of them classified (2) as negative:

(1) Munich is more measured and classy than Spielberg's action-adventures.

(2) Munich is hardly more measured and classy than Spielberg's action-adventures.

The examples above point to another tactic found in our test items. Namely, besides the valence that the adverb contributes in these examples, world-knowledge is also necessary to properly infer the speaker's meaning. Since the sentence uses a proper noun and refers to a well-known figure, it can bear great influence on the valence of the sentence as a whole. We used this strategy in making minimal pairs that proved to confuse the participating systems. It has been claimed in the literature that proper nouns are mostly used in objective or neutral sentences (Pak and Paroubek, 2010). However, proper nouns can also carry sentiments in certain contexts. For instance, while Shakespearean is always a compliment, E.L. James-ian might not be. Most systems categorized (3) as positive, however a few of them missed the negative connotations of (4).

(3) Shakespearean in its violence, Oldboy also calls up nightmare images of spiritual and physical isolation that are worthy of Samuel Beckett or Dostoyevsky.

(4) EL James-ian in its violence, Oldboy also calls up nightmare images of spiritual and physical isolation that are worthy of Paulo Coelho quotes.

We think that world-knowledge could be included in the sentiment analysis systems and it would benefit the judgment of examples such as the one above. Even though this might appear as a non-linguistic issue, references and comparisons with well known directors or actors are found in many of the original reviews and play a role in determining the sentiment.

We have identified another difficulty in pragmatics that is prominent in movie reviews. In examples (5) and (6) below, the mention of the reader's expectations can mean very different things depending on the context:

(5) Sharp dialogue and detailed observations make it a good deal funnier than you might expect.

(6) Horrible dialogue and abysmal acting make it a good deal funnier than you might expect.

The minimal pair of (5) and (6) sheds light on the issue of whether calling a movie funny is a positive comment. This brings us to the discussion of the multi-layered sentiment structure. That is, while 'funny' refers to a positive emotion experienced by someone watching the movie, that might not be a positive comment on the movie, if it is the poor quality of acting that causes one to laugh, such as in example (6). We constructed a similar example where 'emotional pain' was experienced when watching the movie, which could be used to either admire or ridicule the movie. These examples show that the meaning of positive or negative adjectives can change with varying circumstances, such as expectations.

Furthermore, we used another strategy that is based on expressing expectations. A concessive relation, as found in (7) and (8), expresses a contrast between two statements. One of the statements in each sentence is positive and the other one is negative, however the overall sentiment of the two sentences differs. This is achieved by the fact that concessive relations have an expectation in the first component and deny that expectation in the second (Izutsu, 2008). This denial of expectation puts argumentative emphasis on the second

part of the sentence, making the second judgment of the sentence stronger. This is why (7) is negative, while (8) is positive. However, many of the Builder systems had difficulty categorizing both sentences, as they include both positive and negative statements.

(7) It's harmless, sure, but it's also charmless.

(8) It's not harmless, sure, but it's also not charmless.

Another factor we found to affect the valence of the whole sentence, is the use of positive or negative adjectives to refer to a character in the movie or to the plot, but not to the movie itself. For example, the 'smoldering, humorless intensity' in (9) and (10) is a negative attribute of a person, but it might make a great character, such as in (10). However, a few of the Builder systems did not recognize it as a positive review.

(9) [Bettis] has a smoldering, humorless intensity that's unnerving.

(10) [Bettis] has a smoldering, humorless intensity that's hilarious.

As can be seen from the example above, treating words as separate entities with emotional valence can sometimes fail in analyzing sentiment of complete sentences. This leads to another strategy, which is changing the structure of the sentence with minimal changes in the lexical items used. For example, the sentences in (11) and (12) differ minimally in terms of the words used, but they have completely different syntactic structures. The syntactic dependencies determine what is the subject of the sentence and thus who is the savior and who we are saved from.

(11) Someone has to save us from Lawrence's onslaught of cinematic dross.

(12) Lawrence is someone who has saved us from an onslaught of cinematic dross.

Furthermore, syntactic structures can also introduce implicatures. For instance, we changed a sentence into a question or added a tag question and it resulted in Builder system errors. It can be seen from examples (13) and (14) that the sentences are nearly the same, except one of them is declarative and the second one is interrogative. Especially in combination with the use of ellipsis, sentence (14) implies doubt by the speaker, since they are asking a rhetorical question, provided the context is a movie review. Even though there is no explicit negation, the speaker explicitly does not commit to a positive statement. Implicatures are derived from the fact that the speaker did not use a more informative or stronger expression when they could have (Potts, 2015). In this case, if the speaker had found the movie exceptional, they would have said so. Many Builder systems did not recognize it as carrying negative sentiment.

(13) An exceptional science fiction film...

(14) Is this an exceptional science fiction film...?

We also employed ellipsis to change perspective and imply different content in the omitted part. In elliptical sentences, a part of the syntactic structure is missing, as demonstrated in examples (15) and (16) (the part in brackets was omitted in the items). The addition of 'please' to sentence (16) changes it from a declarative sentence to an imperative one. Elliptical utterances are reduced, therefore knowing the discourse goal of the speaker would facilitate the interpretation of the utterance (Carberry, 1989). Hence, the difference between sentences (15) and (16) can be inferred from the fact that one is a claim and the other is a request. Many of the Builder systems did not perform well on sentence (16).

(15) [This is] more of the same...

(16) [I want/give me] more of the same, please!

In addition, a couple of hypothetical sentences with implied content also confused the Builder systems. For example, the difference between (17) and (18) is simply the mood of the verb. The hypothetical in (18) implies that in fact the movie is not a good adaptation, as reality is different from what could have been. In other items, we used the verb 'to try' for an analogous effect, as claiming that someone tried to achieve something, implies that they did not succeed. In both cases almost all

of the systems predicted the direct statement correctly, but did not register the implicature.

(17) Pride and Prejudice is a gorgeous and well-acted adaptation.

(18) Pride and Prejudice could have been a gorgeous and well-acted adaptation.

A final strategy that we adopted in developing our examples is the use of special characters and punctuation marks to affect meaning. In example (19), we used an explicit 'A+' grade, which frames the comment as positive feedback, even if it is preceded by a proposition that is negative on its own.

(19) Ridiculous, confusing, vaguely noir-ish nonsense. A+

All Builder systems failed to recognize it as a good movie mark, probably because such characters are filtered from input. Similarly, the quotation marks in (21) embed the speaker's statement as said by someone else, which in turn, together with an opposing comment, contests the original negative review. This was also not caught by the Builder systems. The change of subject of consciousness or speaker could even be done without the quotation marks, as the very contradictory statements could not both be held by one person, and the second phrase in (21) is clearly a retort.

(20) Flawed, clich, contrived, and poorly developed...

(21) "Flawed, clich, contrived, and poorly developed..." What do they know.

3 Conclusion

To conclude, we have shown how the rich and informal domain of movie reviews allows for sentences that are difficult to analyze for valence. Further manipulation had succeeded in creating items that are not properly understood by the participating systems. In particular, our results suggest that the context of a movie review allows for pragmatic and stylistic manipulations that pose difficulties to current systems. The identification of some of those difficulties might contribute to the improvement of sentiment analysis systems.

References

Sandra Carberry. 1989. A pragmatics-based approach to ellipsis resolution. *Computational Linguistics*, 15(2):75–96.

Mitsuko Narita Izutsu. 2008. Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics*, 40(4):646–675.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10.

Christopher Potts. 2015. Presupposition and implicature. *The handbook of contemporary semantic theory*, 2:168–202.

Arie Verhagen et al. 2007. Construal and perspectivisation.