Breaking NLP: Using Morphosyntax, Semantics, Pragmatics and World Knowledge to Fool Sentiment Analysis Systems

Taylor Mahler, Willy Cheung, Micha Elsner,
David King, Marie-Catherine de Marneffe,
Cory Shain, Symon Stevens-Guille and **Michael White**
The Ohio State University
mahler.38@osu.edu

Abstract

This paper describes our "breaker" submission to the 2017 EMNLP "Build It Break It" shared task on sentiment analysis. In order to cause the "builder" systems to make incorrect predictions, we edited items in the blind test data according to linguistically interpretable strategies that allow us to assess the ease with which the builder systems learn various components of linguistic structure. On the whole, our submitted pairs break all systems at a high rate (72.6%), indicating that sentiment analysis as an NLP task may still have a lot of ground to cover. Of the breaker strategies that we consider, we find our semantic and pragmatic manipulations to pose the most substantial difficulties for the builder systems.

1 Introduction

This paper describes our submission to the 2017 EMNLP "Build It Break It" shared task on sentiment analysis, in which we constructed minimal pairs of sentences designed to fool sentiment analysis systems that would participate in the task. One member of the pair existed in the blind test data, and the other member was a minimally edited version of the first member designed to cause the systems to make an incorrect prediction on exactly one of the two. The edits were made according to four broad, linguistically interpretable strategies: altering syntactic or morphological structure, changing the semantics of the sentence, exploiting pragmatic principles, and including content that can only be understood with sufficient world knowledge. Some of our changes were designed to fool bag-of-words models, others used more complex structures to try to fool more sophisticated systems relying on parsing and/or compositional methods. Our submitted pairs broke the builder systems at a high rate (72.6%) on average, and our overall weighted F_1 score as defined by the shared task (28.67) puts us in second place out of the four breaker submissions.

2 Strategies

Our edits to the original sentences can be categorized under four broad categories: morphological and syntactic change, semantic change, pragmatic change, and use of world knowledge to determine the meaning. This categorization scheme draws on the definitions used across the field of linguistics; we give a more precise definition of each category below.

In each example, we indicate how our team judged the sentence in terms of sentiment ('+' for positive sentiment and '–' for negative sentiment); these labels were viewed as "gold" by the organizers. In each pair, the first sentence is the original one, the second our constructed test case. The test cases highlighted below were especially effective at breaking builders' systems (i.e., most or all of the systems predicted the wrong sentiment, where superscripts ‡, †, and ⋆ indicates that all but 2, 1, and 0 systems predicted the wrong sentiment).

Figure 1 shows a histogram of minimal pairs by strategy in our submission.

2.1 Morphological and Syntactic Strategies

Edits involving syntactic and morphological changes included the addition or removal of negation, as well as comparatives. Both syntactic negation and comparatives exhibit co-occurrence restrictions, one of the canonical diagnostics for syntactic properties (Dawson and Phelan, 2016). These restrictions can be seen in (1) for lexical negation. *Not* in this case syntactically selects a verb phrase (VP); the VP can stand alone as it

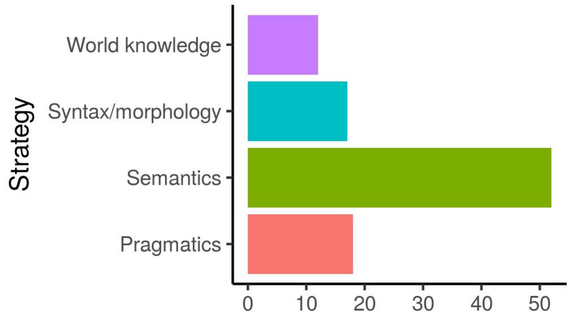

Figure 1: Number of submitted test pairs by strategy.

does in our edited version of the sentence precisely because *not* imposes a co-occurrence restriction on the VP rather than vice versa. Moreover, except in sentence-final emphatic cases, *not* imposes linear order restrictions, appearing prior to the VP.[1] Although there is some dialectical variation, our edited version of the sentence that uses *quite* without negation is slightly degraded, frequently judged as archaic or pretentious, and would also receive some level of additional phonological emphasis on *quite*, which is unavailable from text alone.[2] By contrast, *not quite* is a common expression arguably far less subject to judgment variation.

(1) - In the structure of his screenplay Ross has taken a risk, and he has **not** quite brought it off.
    + ‡In the structure of his screenplay Ross has taken a risk, and has quite brought it off.

Comparatives also impose co-occurrence restrictions, subcategorizing for both an object to be compared and another to be compared to. In the literature on comparatives, the *-er* morpheme is often taken to affect scopal relations beyond its surface position; moreover, *-er* is taken to be analogous to *more* in its semantics and to likewise subcategorize for a(n expression of) degree (Kennedy, 2002; Bhatt and Takahashi, 2011). Nonetheless, the present case is still morphological insofar as it is the distribution of *-er*, as opposed to *less*, that distinguishes the members of the pair—the former must morphologically compose with another expression, the latter need not.[3] Removing an adjective for the morpheme to compose with is then predicted to produce different corresponding semantic–and consequently sentiment—effects, as in (2):

(2) + School of Rock made me laugh **harder** than any movie I've seen this year.
    – ⋆School of Rock made me laugh **less** than any movie I've seen this year.

Finally, we introduced negation morphologically, by the addition of derivational morphemes, as in (3) and (4):

(3) + [A] great big ball of entertainment ...
    – ‡[A] great big ball of **anti-**entertainment ...

(4) + A remarkably convincing examination of heroism, hero worship, and the seductive allure of villainy.
    – ‡A remarkably **un**convincing examination of heroism, hero worship, and the seductive allure of villainy.

We hypothesized that minimal edits to these constructions could introduce semantic scope resolution difficulties for NLP systems and cause them to mis-classify the overall sentiment. Our intuition is that NLP applications can perform sentiment analysis reasonably well on the original sentences. By only editing words which carry semantic operators, a sentiment analysis system with no model of semantics or the scope of semantic operators would be unable to capture the change in sentiment.

[1] Compare with an emphatic case, which is also usually accompanied by emphatic phonological stress on the expression of negation:

He quite brought it off, not!

[2] By dialectical variation we have in mind the differences between e.g., certain American and Canadian dialects of English, as opposed to British dialects, for which some of the present authors have personal attestation of such sorts of utterance.

[3] Compare:

**School of Rock made me laugh lesser than any movie I've seen this year.*

School of Rock gave me fewer laughs than any movie I've seen this year.

Note that according to (Kennedy, 2002)[527] '*less* and *as* differ from *more* only in the nature of the ordering relation they impose', where the ordering relation is over degrees.

2.2 Semantic Strategies

Semantic edits are those that affect the truth conditions of the expression. One might object that all of the examples in the other strategies have a semantic component.[4] This is true, but our semantics-specific strategy targets semantic information that is independent of the morphology or the syntax of the expressions, while the other strategies explicitly exploit morphological and semantic information that may e.g., alter scope information.

Most edits involving semantic changes altered the sentiment by introducing or modifying an operator that is not straightforward negation, such as *too*, *enough* and *only*. Since these words shift a sentiment's polarity without altering the rest of the sentence, we hypothesized that sentiment analysis systems that are not sensitive to these shifts would mislabel sentences with these edits:

(5) – Aiming to join the Jerry Bruckheimer/Michael Bay school of American movie war games, Stealth is just **too** dumb to make the grade.
+ †Aiming to join the Jerry Bruckheimer/Michael Bay school of American movie war games, Stealth is just dumb **enough** to make the grade.

Another strategy that shifts sentiment polarity without modifications to the original sentence involves embedding clauses or predicates under various semantic operators. In (6), for example, embedding the original clause under *tell* diminishes the author's commitment to that clause. Further, adding *I simply can't see why* reverses the positive sentiment of that original clause.

(6) + An exceptional science fiction film.
– †**Many have told me this is** an exceptional science fiction film**, but I simply can't see why.**

In (7), changing the modal *could* to *should* subtly reverses the sentiment.

(7) – This quirky, snarky contemporary fairy tale could have been a family blockbuster.
+ †This quirky, snarky contemporary fairy tale **should** have been a family blockbuster.

In (8), we embedded the verb phrase from the original sentence under *keep trying*, thus implying that the event described by the complement of *keep trying* has not happened.

(8) + The two featured females offset these distractions by having so much apparent fun that it becomes contagious.
– ‡The two featured females **keep trying to** offset these distractions by having so much apparent fun that it becomes contagious.

Finally, some edits were purely lexical and thus belong to the domain of lexical semantics. In these cases, a single word or multi-word expression carrying the sentiment was changed, as in (9) where we used an antonym.

(9) – This movie plays like they were reading [Roger Ebert's] little movie glossary and they took every cliche in there.
+ ⋆This movie plays like they were reading [Roger Ebert's] little movie glossary and they **avoided** every cliche in there.

In some cases where the genre of the film was mentioned, we simply changed it. Since different genres are intended to have different effects, what counts as positive and negative depends on the genre of the movie. For instance, in (10), the description of the experience of the film does not match the intended effects of a romantic comedy, but it does match those of a horror film.

(10) – The Break-Up, a grim excuse for a romantic comedy, is basically an hour and 45 minutes spent in the company of two unpleasant people during a miserable time in their lives.
+ †The Break-Up, **a grimly compelling horror film**, is basically an hour and 45 minutes spent in the company of two unpleasant people during a miserable time in their lives.

While it might be argued that manipulation of

[4] This also goes some way to explaining the success of the semantic strategies in general, since they are in part exploited by the other strategies.

genre is a world knowledge strategy, since the sentiment of these sentences depends crucially on understanding the lexical meaning of the word that indicates the genre, we classify genre manipulation as a semantic strategy.

2.3 Pragmatic Strategies

Pragmatic strategies make use of inferences which go beyond the literal compositional meaning of the words, relying on knowledge of general principles of human communication, but *not* on extra-linguistic and contextual knowledge. Since most NLP applications lack the information necessary to make use of pragmatics as robustly as humans do, we exploited a variety of pragmatic principles to either create or convey an impression of sarcasm. In the simplest case, we used scare quotes to convey sarcasm, changing the sentiment from positive to negative, as in (11).

(11) + Russell is terrific as coach Herb Brooks.
– *Russell is “terrific” as coach Herb Brooks.

This seemingly simple manipulation actually proved quite difficult for the builder systems. Both pairs we submitted that used this strategy broke all six builder systems.

In other examples, we created Gricean conversational implicatures (Grice, 1975). For instance, our constructed sentence in (12) flouts the Gricean maxim of quantity by providing too little information, implicating that a more informative statement praising the film could not be made because it would be false, and violate the maxim of quality. While there’s nothing overtly negative in our constructed sentence in (12), it nonetheless conveys a negative sentiment.

(12) + I think it’s a sweet film.
– †I think it’s a film.

Our edited sentence in (13) flouts the maxim of relation by providing information that is not relevant in a movie review, implicating that a relevant, positive statement could not be made because it would be false (again violating the maxim of quality).

(13) + The **performances** are uniformly superb.
– *The **marketing** was uniformly superb.

A final pragmatic strategy involved cases where two phrases were conjoined with *but*. Often the sentiment of the second conjunct is also the sentiment of the entire sentence. In such cases, reversing the order of the conjuncts can also reverse the sentiment of the entire sentence, as in the constructed example in (14).

(14) – The sentiments are right on the money, but the execution never quite filled me with holiday cheer.
+ ‡The execution never quite filled me with holiday cheer, but the sentiments are right on the money.

2.4 World Knowledge Strategies

Most NLP applications have a limited understanding of world knowledge. To exploit this shortcoming, we edited sentences so that world knowledge crucially affected the sentiment of the sentence. Arguably, the world knowledge strategies are pragmatic in nature since pragmatics is typically taken to involve meaning that is contributed by context (Dawson and Phelan, 2016) . However, we categorize these strategies separately since the inferences exploiting world knowledge strategies crucially rely on extra-linguistic knowledge.

Many of the sentences we edited using this strategy involved a comparison. We edited such sentences so that knowledge about the standard of comparison was crucial for determining the sentiment. In some cases, the standard of comparison was a named entity, such as a film or an actor. In (15), the negative sentiment arises as a result of the comparison to a Jim Carrey film, which is not intended to be creepy and calibrated:

(15) + Unfolds with the creepy elegance and carefully calibrated precision of a Dario Argento horror film.
– *Unfolds with all the creepy elegance and carefully calibrated precision of a Jim Carrey comedy film.

In other cases, the comparison was metaphorical, and we manipulated the sentiment by altering the nature of the comparison itself. For instance, understanding that the constructed sentence in (16) is negative requires knowledge about the weight of bricks.

(16) + As pretty and light as a **feather** on the wind.

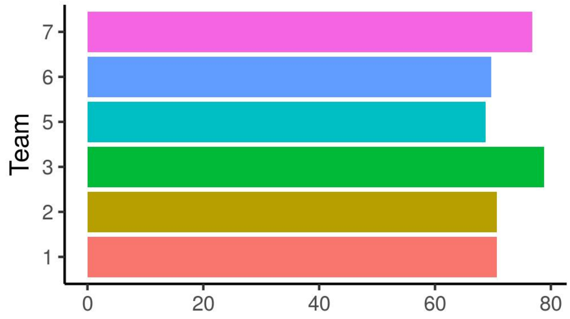

Figure 2: Percent break (out of all submitted pairs) by system.

– †As pretty and light as a **brick** on the wind.

We also manipulated the perspective from which a particular sentiment is conveyed. If the review praises something as being valued by individuals that are held in poor regard, then the sentiment is likely to be negative, despite the apparent praise. For example, understanding that the second sentence in (17) is negative requires knowing that racists are (generally) not well regarded.

(17) + An inspiring story for **teens and up**.
– *An inspiring story for **racists**.

Since most NLP applications do not know, for example, that Jim Carrey films are not intended to be creepy and calibrated, that bricks are heavier than feathers, and that one should not blindly follow the recommendations of racists, we predicted that computers would show lower performance when analyzing sentiment in these cases.

3 Results

Following the shared task's definition, a minimal pair is considered to "break" a builder system if the system makes a correct prediction for one member of the pair and an incorrect prediction for the other. The shared task also defines a weighted F_1 score for breaker teams as the F_1 of the builder system on the original sentences of the blind test set, multiplied by the percent of builder sentences on which the breaker team made an incorrect prediction.

We submitted 99 breaker pairs in total. We obtained a mean percent break across systems of 72.6%,[5] and the mean weighted F_1 across systems on our pairs was 28.67, placing us second in terms of this metric out of the four breaker teams in the shared task. Figures 2 and 3 shows percent break and weighted F_1 respectively by system. System 5 had the lowest percent break on our submitted test cases, while system 3 had the highest.

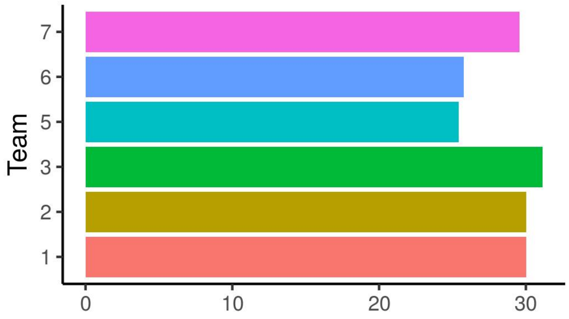

Figure 3: Weighted F_1 score by builder system on our 99 pairs.

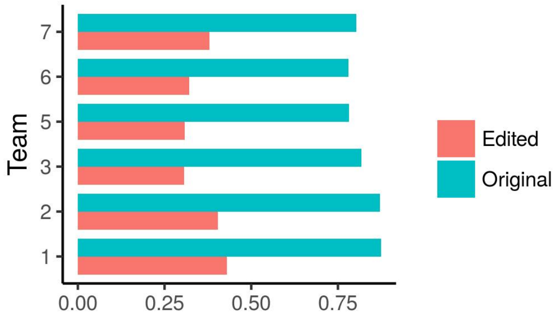

Figure 4: Raw F_1 by system on original vs. edited examples

In figure 4, we also present the raw F_1 scores by system on original vs. edited sentences. As is clear in the figure, our edits dramatically compromise classification accuracy across all systems.[6] Note that while Teams 5 and 6 perform well in terms of percent break shown in Figure 2, they have some of the lowest raw F_1 scores shown in Figure 4. This suggests that the strong break rate scores for these systems are driven by pairs in which both items are incorrectly classified, which are not considered to be breaks by the task definition.

Figure 5 provides overall percent break by strategy. Our pragmatic manipulations had the highest percent break while our world-knowledge-based manipulations had the lowest.

[5] I.e., the percent of all submitted pairs (99) that resulted in a break for that system as defined by the shared task.

[6] Although in principle breaker teams were allowed to submit edits designed to make classification easier, almost all of our submitted edits were designed to make classification harder.

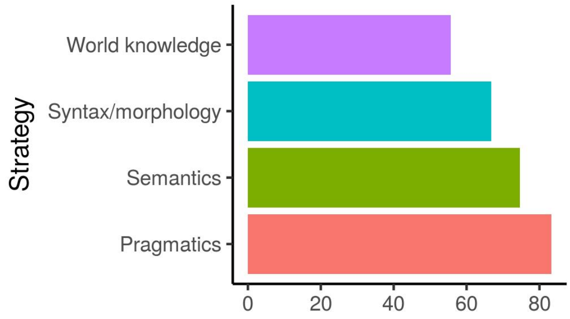

Figure 5: Percent break (out of all submitted pairs) by strategy.

In addition to breaks, there were also pairs on which the systems got both sentences wrong. For the most part, these appear to have been “neutral” labels (neither negative nor positive) used by the neural network teams (systems 5 and 6). Since neutral sentiment is not part of the gold label space, this appears to have been an error on the part of these systems. Note that as a consequence, the percent break for systems 5 and 6 was lower than it might have been otherwise.

We did not significance test differences in performance between systems or strategies because (1) the same items were shared across builder systems and were therefore not truly independent and (2) we were unsure whether our examples constituted a representative sample of naturally-occurring hard cases. Thus, while our findings are suggestive of the kinds of linguistic phenomena that pose difficulties for automatic sentiment analysis, we are unable to draw firmer conclusions based on this limited sample.

4 Ineffective test cases

In §2, we presented examples of test cases that tended to break the builder systems; here we briefly analyze the *ineffective* test cases (i.e., test cases that most or all of the systems got right) in hopes of evaluating where our test cases failed to break systems and/or where existing systems tended to predict the correct sentiment.

As shown in §3, our test cases yielded a generally high break rate across systems. In fact, of the 99 test cases we submitted, 72 broke more than half of the systems. Of the remaining 27 test cases on which at least half of the systems did not break, 12 were same-sentiment pairs (out of 12 total in our submission). In general these involved attempts to use one of the strategies discussed above to make a positive or negative classification more difficult. However, we appear to have left enough residual evidence of the source sentiment in the edited cases to allow most systems to make the correct decision. In addition, 8 of the 27 test cases involved lexical semantic manipulations, and 7 involved world knowledge, suggesting that these kinds of nuances may not have been as difficult for sentiment analysis systems as we had hypothesized.

The four cases below failed to break any system:

(18) – Unlike Raiders of the Lost Ark, which this movie wants so desperately to be, there’s nothing here to engage the brain along with the eyeballs.
– This movie is not like Raiders of the Lost Ark, which this movie wants so desperately to be.

(19) – This is one of the worst movies of the year.
– This is not one of the worst movies of the year.

(20) – Big on slogans, but low on personality.
– Low on personality, but big on slogans.

(21) + The less you know about this movie before seeing it — and you really should see it — the better.
– The less you know about this movie, the better.

Three out of four of these failed examples were same-sentiment (negative-negative) minimal pairs. The fourth removes the positive-sentiment parataxis *and you really should see it* to flip the overall sentiment. In all these cases, there remain words with likely negative sentiment that might short-circuit the difficulty that the edit was intended to introduce (*wants so desperately to*, *worst movies of the year*, *low on personality*, and *the less you know . . . the better*). Thus, in hindsight, it would have been better to exclude such examples, since it is not clear whether builder systems succeeded on them by correctly analyzing them or simply by detecting the negative-sentiment-bearing keywords.

5 Discussion

Our results, and those of the shared task in general, serve to highlight the distance which even sophisticated, modern sentiment analysis systems have yet to cover, particularly in terms of semantic and pragmatic analysis. Moreover, changes that broke the systems were often comparatively slight; just as image classification systems can be vulnerable to adversarial examples that look very similar to the originals (Szegedy et al., 2014), sentiment analysis systems may be fooled by changes to single words or morphemes. In many cases, of course, our strategies for constructing these examples drew on previous knowledge about hard problems, for instance in parsing (Kummerfeld et al., 2012) and the detection of irony in text (Wallace et al., 2014). Nonetheless, a concrete set of examples of these problems may help developers to create more robust systems in the future.

For sets of constructed examples like ours to be useful, they should contain enough instances of each construction to reliably indicate a system's capabilities. Looking towards the future, we hope that the next iteration of the contest will use a larger test section so that more examples can be created. Many of our strategies targeted particular constructions or idioms (for instance, right-node raising or concrete metaphors), and it was difficult to create many instances of these due to sparsity in the 521-example dataset. We found it difficult to create 100 examples as requested; in fact, two other breaker teams (including the one with the winning F-score) created only half as many.

A related issue is that of naturalness. Although we tried to make our examples sound like real sentences from movie reviews, we had no empirical way to check how well we did. It is probably easier to break NLP algorithms with *unnatural* or out-of-domain examples; although we hope we have not done so, in future, we would like to find better ways to make sure.

Acknowledgments

We would like to thank Julia Papke, Ramon Reyes-Padilla, and Marten van Schijndel for contributing minimal pairs, as well as three anonymous reviewers for helpful feedback and suggestions.

References

Rajesh Bhatt and Shoichi Takahashi. 2011. Reduced and unreduced phrasal comparatives. *Natural Language & Linguistic Theory* 29(3):581–620.

Hope C. Dawson and Michael Phelan, editors. 2016. *Language Files*. The Ohio State University Press, 12th edition.

H. Paul Grice. 1975. Logic and conversation. In P Cole and J Morgan, editors, *Syntax and Semantics 3: Speech Acts*, Academic Press, New York, pages 64–75.

Christopher Kennedy. 2002. Comparative deletion and optimality in syntax. *Natural Language & Linguistic Theory* 20(3):553–621.

Jonathan K Kummerfeld, David Hall, James R Curran, and Dan Klein. 2012. Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1048–1059.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *ICLR* https://doi.org/abs/1312.6199.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *ACL (2)*. pages 512–516.