Humans Require Context to Infer Ironic Intent (so Computers Probably do, too)

Byron C. Wallace, Do Kook Choe, Laura Kertz and **Eugene Charniak**
Brown University
{byron_wallace, do_kook_choe, laura_kertz, eugene_charniak}@brown.edu

Abstract

Automatically detecting verbal irony (roughly, sarcasm) is a challenging task because ironists say something other than – and often opposite to – what they actually mean. Discerning ironic intent exclusively from the words and syntax comprising texts (e.g., tweets, forum posts) is therefore not always possible: additional contextual information about the speaker and/or the topic at hand is often necessary. We introduce a new corpus that provides empirical evidence for this claim. We show that annotators frequently require context to make judgements concerning ironic intent, and that machine learning approaches tend to misclassify those same comments for which annotators required additional context.

1 Introduction & Motivation

This work concerns the task of detecting verbal irony online. Our principal argument is that simple bag-of-words based text classification models – which, when coupled with sufficient data, have proven to be extremely successful for many natural language processing tasks (Halevy et al., 2009) – are inadequate for irony detection. In this paper we provide empirical evidence that *context* is often necessary to recognize ironic intent.

This is consistent with the large body of pragmatics/linguistics literature on irony and its usage, which has emphasized the role that context plays in recognizing and decoding ironic utterances (Grice, 1975; Clark and Gerrig, 1984; Sperber and Wilson, 1981). But existing work on automatic irony detection – reviewed in Section 2 – has not explicitly attempted to operationalize such theories, and has instead relied on features (mostly word counts) intrinsic to the texts that are to be classified as ironic. These approaches have achieved some success, but necessarily face an upper-bound: the *exact same sentence* can be both intended ironically and unironically, depending on the context (including the speaker and the topic at hand). Only obvious verbal ironies will be recognizable from intrinsic features alone.

Here we provide empirical evidence for the above claims. We also introduce a new annotated corpus that will allow researchers to build models that augment existing approaches to irony detection with contextual information regarding the text (utterance) to be classified and its author. Briefly, our contributions are summarized as follows.

- We introduce the first version of the *reddit irony corpus*, composed of annotated comments from the social news website reddit. Each sentence in every comment in this corpus has been labeled by three independent annotators as having been intended by the author ironically or not. This dataset is publicly available.[1]
- We provide empirical evidence that human annotators consistently rely on contextual information to make ironic/unironic sentence judgements.
- We show that the standard 'bag-of-words' approach to text classification fails to accurately judge ironic intent on those cases for which humans required additional context. This suggests that, as humans require context to make their judgements for this task, so too do computers.

Our hope is that these observations and this dataset will spur innovative new research on methods for verbal irony detection.

[1] https://github.com/bwallace/ACL-2014-irony

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 512–516,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

2 Previous Work

There has recently been a flurry of interesting work on automatic irony detection (Tepperman et al., 2006; Davidov et al., 2010; Carvalho et al., 2009; Burfoot and Baldwin, 2009; Tsur et al., 2010; González-Ibáñez et al., 2011; Filatova, 2012; Reyes et al., 2012; Lukin and Walker, 2013; Riloff et al., 2013). In these works, verbal irony detection has mostly been treated as a standard text classification task, though with some innovative approaches specific to detecting irony.

The most common data source used to experiment with irony detection systems has been Twitter (Reyes et al., 2012; González-Ibáñez et al., 2011; Davidov et al., 2010), though Amazon product reviews have been used experimentally as well (Tsur et al., 2010; Davidov et al., 2010; Reyes et al., 2012; Filatova, 2012). Walker et al. (2012) also recently introduced the Internet Argument Corpus (IAC), which includes a *sarcasm* label (among others).

Some of the findings from these previous efforts have squared with intuition: e.g., overzealous punctuation (as in "great idea!!!!") is indicative of ironic intent (Carvalho et al., 2009). Other works have proposed novel approaches specifically for irony detection: Davidov et al. (2010), for example, proposed a semi-supervised approach in which they look for sentence *templates* indicative of irony. Elsewhere, Riloff et al. (2013) proposed a method that exploits contrasting sentiment in the same utterance to detect irony.

To our knowledge, however, no previous work on irony detection has attempted to leverage *contextual* information regarding the author or speaker (external to the utterance). But this is necessary in some cases, however. For example, in the case of Amazon product reviews, knowing the kinds of books that an individual typically likes might inform our judgement: someone who tends to read and review Dostoevsky is probably being ironic if she writes a glowing review of *Twilight*. Of course, many people genuinely do enjoy *Twilight* and so if the review is written subtly it will likely be difficult to discern the author's intent without this background. In the case of Twitter, it is likely to be difficult to classify utterances without considering the contextualizing exchange of tweets (i.e., the conversation) to which they belong.

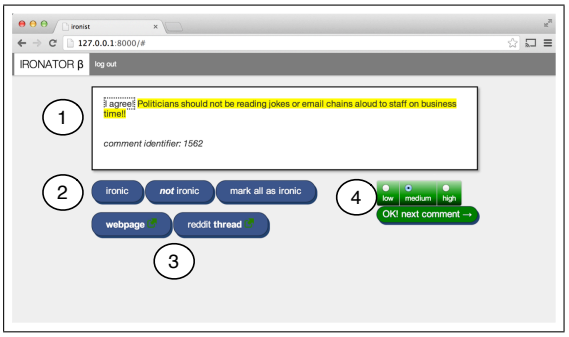

Figure 1: The web-based tool used by our annotators to label reddit comments. Enumerated interface elements are described as follows: **1** the text of the comment to be annotated – sentences marked as *ironic* are highlighted; **2** buttons to label sentences as *ironic* or *unironic*; **3** buttons to request additional *context* (the embedding discussion thread or associated webpage – see Section 3.2); **4** radio button to provide *confidence* in comment labels (*low*, *medium* or *high*).

3 Introducing the reddit Irony Dataset

Here we introduce the first version (β 1.0) of our irony corpus. Reddit (`http://reddit.com`) is a social-news website to which news stories (and other links) are posted, voted on and commented upon. The forum component of reddit is extremely active: popular posts often have well into 1000's of user comments. Reddit comprises 'sub-reddits', which focus on specific topics. For example, `http://reddit.com/r/politics` features articles (and hence comments) centered around political news. The current version of the corpus is available at: `https://github.com/bwallace/ACL-2014-irony`. Data collection and annotation is ongoing, so we will continue to release new (larger) versions of the corpus in the future. The present version comprises 3,020 annotated comments scraped from the six subreddits enumerated in Table 1. These comments in turn comprise a total of 10,401 labeled sentences.[2]

3.1 Annotation Process

Three university undergraduates independently annotated each sentence in the corpus. More specifically, annotators have provided binary 'labels' for each sentence indicating whether or not they (the annotator) believe it was intended by the author ironically (or not). This annotation was provided via a custom-built browser-based annotation tool, shown in Figure 1.

We intentionally did not provide much guidance to annotators regarding the criteria for what

[2]We performed naïve 'segmentation' of comments based on punctuation.

sub-reddit (URL)	description	number of labeled comments
politics (r/politics)	Political news and editorials; focus on the US.	873
conservative (r/conservative)	A community for political conservatives.	573
progressive (r/progressive)	A community for political progressives (liberals).	543
atheism (r/atheism)	A community for non-believers.	442
Christianity (r/Christianity)	News and viewpoints on the Christian faith.	312
technology (r/technology)	Technology news and commentary.	277

Table 1: The six sub-reddits that we have downloaded comments from and the corresponding number of comments for which we have acquired annotations in this β version of the corpus. Note that we acquired labels at the *sentence* level, whereas the counts above reflect *comments*, all of which contain at least one sentence.

constitutes an 'ironic' statement, for two reasons. First, verbal irony is a notoriously slippery concept (Gibbs and Colston, 2007) and coming up with an operational definition to be consistently applied is non-trivial. Second, we were interested in assessing the extent of natural agreement between annotators for this task. The raw average agreement between all annotators on all sentences is 0.844. Average pairwise Cohen's Kappa (Cohen, 1960) is 0.341, suggesting fair to moderate agreement (Viera and Garrett, 2005), as we might expect for a subjective task like this one.

3.2 Context

Reddit is a good corpus for the irony detection task in part because it provides a natural practical realization of the otherwise ill-defined *context* for comments. In particular, each comment is associated with a specific user (the author), and we can view their previous comments. Moreover, comments are embedded within discussion *threads* that pertain to the (usually external) content linked to in the corresponding submission (see Figure 2). These pieces of information (previous comments by the same user, the external link of the embedding reddit thread, and the other comments in this thread) constitute our context. All of this is readily accessible. Labelers can opt to request these pieces of context via the annotation tool, and we record when they do so.

Consider the following example comment taken from our dataset: "Great idea on the talkathon Cruz. Really made the republicans look like the sane ones." Did the author intend this statement ironically, or was this a subtle dig on Senator Ted Cruz? Without additional context it is difficult to know. And indeed, all three annotators requested additional context for this comment. This context at first suggests that the comment may have been intended literally: it was posted in the r/conservative subreddit (Ted Cruz is a conservative senator). But if we peruse the author's comment history, we see that he or she repeatedly derides Senator Cruz (e.g., writing "Ted Cruz is no Ronald Reagan. They aren't even close."). From this contextual information, then, we can reasonably assume that the comment was intended ironically (and all three annotators did so after assessing the available contextual information).

Figure 2: An illustrative reddit comment (highlighted). The title ("Virginia Republican ...") links to an article, providing one example of contextualizing content. The conversational thread in which this comment is embedded provides additional context. The comment in question was presumably intended ironically, though without the aforementioned context this would be difficult to conclude with any certainty.

4 Humans Need Context to Infer Irony

We explore the extent to which human annotators rely on contextual information to decide whether or not sentences were intended ironically. Recall that our annotation tool allows labelers to request additional context if they cannot make a decision based on the comment text alone (Figure 1). On average, annotators requested additional context for 30% of comments (range across annotators of 12% to 56%). As shown in Figure 3, annotators are consistently more confident once they have consulted this information.

We tested for a correlation between these requests for context and the final decisions regarding whether comments contain at least one ironic sentence. We denote the probability of at least one annotator requesting additional context for comment i by $P(\mathcal{C}_i)$. We then model the probability of this event as a linear function of whether or not

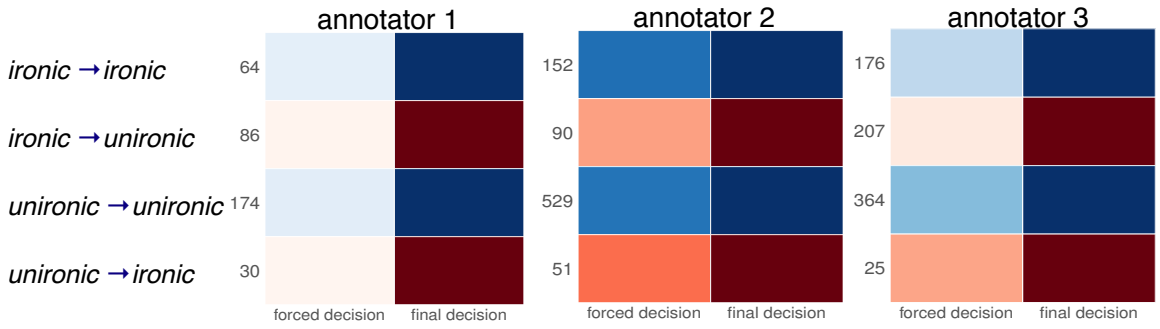

Figure 3: This plot illustrates the effect of viewing contextual information for three annotators (one table for each annotator). For all comments for which these annotators requested context, we show *forced* (before viewing the requested contextual content) and *final* (after) decisions regarding perceived ironic intent on behalf of the author. Each row shows one of four possible decision sequences (e.g., a judgement of *ironic* prior to seeing context and *unironic* after). Numbers correspond to counts of these sequences for each annotator (e.g., the first annotator changed their mind from *ironic* to *unironic* 86 times). Cases that involve the annotator changing his or her mind are shown in red; those in which the annotator stuck with their initial judgement are shown in blue. Color intensity is proportional to the average confidence judgements the annotator provided: these are uniformly stronger after they have consulted contextualizing information. Note also that the context frequently results in annotators changing their judgement.

any annotator labeled any sentence in comment i as ironic. We code this via the indicator variable $\mathcal{I}_i$ which is 1 when comment i has been deemed to contain an ironic sentence (by any of the three annotators) and 0 otherwise.

$$logit\{P(\mathcal{C}_i)\} = \beta_0 + \beta_1 \mathcal{I}_i \quad (1)$$

We used the regression model shown in Equation 1, where β_0 is an intercept and β_1 captures the correlation between requests for context for a given comment and its ultimately being deemed to contain at least one ironic sentence. We fit this model to the annotated corpus, and found a significant correlation: $\hat{\beta}_1 = 1.508$ with a 95% confidence interval of (1.326, 1.690); $p < 0.001$.

In other words, annotators request context significantly more frequently for those comments that (are ultimately deemed to) contain an ironic sentence. This would suggest that the words and punctuation comprising online comments alone are not sufficient to distinguish ironic from unironic comments. Despite this, most machine learning based approaches to irony detection have relied nearly exclusively on such intrinsic features.

5 Machines Probably do, too

We show that the misclassifications (with respect to whether comments contain irony or not) made by a standard text classification model significantly correlate with those comments for which human annotators requested additional context. This provides evidence that bag-of-words approaches are insufficient for the general task of irony detection: more context is necessary.

We implemented a baseline classification approach using vanilla token count features (binary bag-of-words). We removed stop-words and limited the vocabulary to the 50,000 most frequently occurring unigrams and bigrams. We added additional binary features coding for the presence of punctuational features, such as exclamation points, emoticons (for example, ';)') and question marks: previous work (Davidov et al., 2010; Carvalho et al., 2009) has found that these are good indicators of ironic intent.

For our predictive model, we used a linear-kernel SVM (tuning the C parameter via grid-search over the training dataset to maximize F1 score). We performed five-fold cross-validation, recording the predictions $\hat{y}_i$ for each (held-out) comment i. Average F1 score over the five-folds was 0.383 with range (0.330, 0.412); mean recall was 0.496 (0.446, 0.548) and average precision was 0.315 (0.261, 0.380). The five most predictive tokens were: *!*, *yeah*, *guys*, *oh* and *shocked*. This represents reasonable performance (with intuitive predictive tokens); but obviously there is quite a bit of room for improvement.[3]

We now explore empirically whether these misclassifications are made on the same comments for which annotators requested context. To this end, we introduce a variable $\mathcal{M}_i$ for each comment i such that $\mathcal{M}_i = 1$ if $\hat{y}_i \neq y_i$, i.e., $\mathcal{M}_i$ is an in-

[3]Some of the recently proposed strategies mentioned in Section 2 may improve performance here, but none of these address the fundamental issue of *context*.

dicator variable that encodes whether or not the classifier misclassified comment i. We then ran a second regression in which the output variable was the logit-transformed probability of the model misclassifying comment i, i.e., $P(\mathcal{M}_i)$. Here we are interested in the correlation of the event that one or more annotators requested additional context for comment i (denoted by $\mathcal{C}_i$) and model misclassifications (adjusting for the comment's true label). Formally:

$$logit\{P(\mathcal{M}_i)\} = \theta_0 + \theta_1 \mathcal{I}_i + \theta_2 \mathcal{C}_i \qquad (2)$$

Fitting this to the data, we estimated $\hat{\theta}_2 = 0.971$ with a 95% CI of (0.810, 1.133); $p < 0.001$. Put another way, the model makes mistakes on those comments for which annotators requested additional context (even after accounting for the annotator designation of comments).

6 Conclusions and Future Directions

We have described a new (publicly available) corpus for the task of verbal irony detection. The data comprises comments scraped from the social news website reddit. We recorded confidence judgements and requests for contextualizing information for each comment during annotation. We analyzed this corpus to provide empirical evidence that annotators quite often require context beyond the comment under consideration to discern irony; especially for those comments ultimately deemed as being intended ironically. We demonstrated that a standard token-based machine learning approach misclassified many of the same comments for which annotators tend to request context.

We have shown that annotators rely on contextual cues (in addition to word and grammatical features) to discern irony and argued that this implies computers should, too. The obvious next step is to develop new machine learning models that exploit the contextual information available in the corpus we have curated (e.g., previous comments by the same user, the thread topic).

7 Acknowledgement

This work was made possible by the Army Research Office (ARO), grant #64481-MA.

References

C Burfoot and T Baldwin. 2009. Automatic satire detection: are you having a laugh? In *ACL-IJCNLP*, pages 161–164. ACL.

P Carvalho, L Sarmento, MJ Silva, and E de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

HH Clark and RJ Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology*, 113:121–126.

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

D Davidov, O Tsur, and A Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. pages 107–116.

E Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*, volume 12, pages 392–398.

RW Gibbs and HL Colston. 2007. *Irony in language and thought: a cognitive science reader*. Lawrence Erlbaum.

R González-Ibáñez, S Muresan, and N Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *ACL*, volume 2, pages 581–586. Citeseer.

HP Grice. 1975. Logic and conversation. *1975*, pages 41–58.

A Halevy, P Norvig, and F Pereira. 2009. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12.

S Lukin and M Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. *NAACL*, pages 30–40.

A Reyes, P Rosso, and T Veale. 2012. A multidimensional approach for detecting irony in twitter. *LREC*, pages 1–30.

E Riloff, A Qadir, P Surve, LD Silva, N Gilbert, and R Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714.

D Sperber and D Wilson. 1981. Irony and the use-mention distinction. *1981*.

J Tepperman, D Traum, and S Narayanan. 2006. "Yeah Right": Sarcasm Recognition for Spoken Dialogue Systems.

O Tsur, D Davidov, and A Rappoport. 2010. ICWSM-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *AAAI Conference on Weblogs and Social Media*.

AJ Viera and JM Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.

MA Walker, JEF Tree, P Anand, R Abbott, and J King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.