REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
13-04-2015	Final Report	1-Sep-2013 - 30-Nov-2014

4. TITLE AND SUBTITLE
Final Report: Sociolinguistically Informed Natural Language Processing: Automating Irony Detection

5a. CONTRACT NUMBER
W911NF-13-1-0406

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Byron C Wallace

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Brown University
Office of Sponsored Projects
Box 1929
Providence, RI 02912 -9093

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
64481-MA.3

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Irony detection is an important problem in computational sociolinguistic tasks, such as automatic community detection on the web. But the ironic/sincere distinction has proven to be a particularly difficult classification problem. Existing Machine Learning (ML) and Natural Language Processing (NLP) approaches, which tend to rely on simple statistical models built on top of word counts, are not very good at it. We hypothesize that this is because, in contrast to most text classification problems, word counts and syntactic features alone do not constitute an adequate representation for verbal irony detection. Indeed, sociolinguistic theories of verbal irony imply that a

15. SUBJECT TERMS
natural language processing, machine learning, irony, sarcasm, social media

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Byron Wallace
UU	UU	UU	UU		19b. TELEPHONE NUMBER 512-471-3821

Report Title

Final Report: Sociolinguistically Informed Natural Language Processing: Automating Irony Detection

ABSTRACT

Irony detection is an important problem in computational sociolinguistic tasks, such as automatic community detection on the web. But the ironic/sincere distinction has proven to be a particularly difficult classification problem. Existing Machine Learning (ML) and Natural Language Processing (NLP) approaches, which tend to rely on simple statistical models built on top of word counts, are not very good at it. We hypothesize that this is because, in contrast to most text classification problems, word counts and syntactic features alone do not constitute an adequate representation for verbal irony detection. Indeed, sociolinguistic theories of verbal irony imply that a model of the speaker is a necessary condition for irony detection, at least in certain cases.

In this project, we have collected a new corpus of online comments and annotated the sentences therein "ironic" (or not). We used this dataset to empirically demonstrate that human annotators require context to infer irony. Moreover, we have shown that the classification errors made by standard machine learning approaches tend to be on those same examples for which humans require context. We have begun to develop methods that capitalize on contextual information to infer verbal irony.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper	
08/21/2014	1.00	Byron Wallace, Do Kook Choe, Laura Kertz, Eugene Charniak. Humans Require Context to Infer Ironic Intent (so Computers Probably do, too), The 52nd Annual Meeting of the Association for Computational Linguistics . 22-JUN-14, . : ,
08/21/2014	2.00	Byron Wallace, Laura Kertz. Can Cognitive Scientists Help Computers Recognize Irony?, The Annual Meeting of the Cognitive Science Society (CogSci 2014). 23-JUL-14, . : ,

TOTAL: **2**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Joern Klinger	0.00	
FTE Equivalent:	**0.00**	
Total Number:	**1**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Byron Wallace	0.22	No
Thomas Trikalinos	0.02	
Eugene Charniak	0.05	
Laura Kertz	0.10	
FTE Equivalent:	**0.39**	
Total Number:	**4**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
Zoe Downes	1.00	
Zoe Fieldsteel	1.00	
Gabrielle Frampton	0.02	
Isue Shin	0.56	
Sarah Palasits	0.36	
FTE Equivalent:	**2.94**	
Total Number:	**5**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:..... 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME
Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See attachment

Technology Transfer

Sociolinguistically Informed Natural Language Processing: Automating Irony Detection (Proposal Number: 64481-MA, Agreement Number: W911NF-13-1-0406)

Scientific Progress and Accomplishments ("Final Report")

Byron C. Wallace

2015-04

1 Foreward

This "final report" actually describes a project still in progress, having recently transitioned from Brown University to the University of Texas at Austin (which Wallace, the PI, joined in fall 2014). The project has transitioned smoothly. Wallace is currently working with one PhD student (in computational linguistics) and an undergraduate in Computer Science on furthering the methodological research for this project. Moreover, David Beaver, Professor in the Linguistics and Philosophy at UT Austin, has joined the project, replacing Brown linguist Laura Kertz. Brown Professor in Computer Science Eugene Charniak remains on the project to provide guidance on natural language processing methods.

We also note that this project has recently spurred collaboration with researchers in Portugal, specifically with Dr. Paula Carvalho [1] and her team. With these researchers, Wallace joined a proposal that just awarded to the team through the Scientific Research and Technological Development Project in Interactive and UT Austin-Portugal Program Digital Media and Emerging Technologies program, entitled "Expression and Recognition of Irony in Multicultural Social Media". This will provide opportunity for furthering the work and to disseminate the ideas and methods produced through the current work.

In this report, however, we focus mainly on the progress made on the project prior to its transfer to UT Austin.

2 Background

The research objective of this project is to develop resources and novel computational methods to advance automated irony detection (i.e., identification of the ironic voice in online content). This is a challenging task because the meaning of natural language is not captured by words and syntax alone. Rather, utterances (tweets,[1] sentences in forum posts, etc.) are embedded within a specific context. The ironic voice is an important example of this phenomenon: to appreciate a speaker's intended meaning, it is crucial to first infer if he or she is being ironic or sincere.

Existing computational approaches to irony detection leverage statistical natural language processing (NLP) and machine learning (ML) methods. These models tend to be relatively 'shallow' in that they operate only over simple, unstructured representations of data. For example, in the case of natural language (text), one might encode documents with word counts or functions thereof, and in the case of network-based

[1] 'tweets' are short messages posted to the internet for the consumption of 'followers' via the web service Twitter.

data (e.g. social networks) one might rely on analogously simple functions of link counts. Classification would then be performed by algorithms operating over these encodings. But these simple representations will often be insufficient to infer ironic intent [6]. In this project we therefore aim to explore novel approaches to irony detection that are motivated by linguistic principles.

To this end, we have brought together a diverse team comprising members with unique expertise. Senior personnel on this project included PI Wallace and Charniak, both of whom have substantive expertise in statistical natural language processing; we have combined this with important expertise from Kertz in linguists and from Trikalinos in statistics and experimental design. This interdisciplinary team has been a crucial property of our approach: in our view previous efforts to identify irony relied too heavily on standard computer science methods, largely ignoring the perspectives of, e.g., linguistics and cognitive scientists. Indeed, as part of this broad effort of facilitating interdisciplinary communication around this important problem, **we organized and ran a workshop at CogSci 2014 this past July**, which included speakers and attendees from both computer and cogntive science (`https://sites.google.com/a/brown.edu/irony/`) [8]. We now continue this interdisciplinarity at UT Austin by involving Professor David Beaver (from linguistics and philosophy).

In the next section we review the specific objectives of this project and then discuss our progress toward meeting them in the relevant period.

3 Specific Objectives

The broad specific objectives of this project were as follows:

1. First, *to collect and annotate a high-quality corpus to facilitate research on irony detection.* Prior to this project, no such high-quality dataset existed. This has been a major obstacle to progress on automated irony detection.
2. Second, *to analyze when existing ML and NLP technologies fail to detect ironic intent* empirically. We specifically proposed to assess quantitatively (using the collected dataset) whether *context* is necessary to discern ironic intent (and how often this is the case).
3. Finally, we aimed to *develop a new approach to irony detection that instantiates sociolinguistic conceptions of irony within a modern, probabilistic machine learning framework.* The idea was that this approach would be informed by theoretical sociolinguistic perspectives on irony (and thus likely capable of discerning ironic utterances missed by existing computational models), while also being practical enough to be operational.

Below we enumerate our findings thus far regarding these objectives.

4 Scientific Findings and Accomplishments

We have at least partially realized aims 1 and 2, slightly ahead of our slated timeline. Specifically, as described in detail in the following subsections, we have: (1) written code to scrape comments from *reddit*, a social-news website that we use as our corpus; (2) built a web-based tool to facilitate annotation of these comments; (3) assembled and trained a team of undergraduates to perform this annotaiton; (4) analyzed the resultant dataset. This analysis was summarized in our publication at this year's Association for Computational Linguistics [7], the premiere venue in natural language processing.

4.1 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`

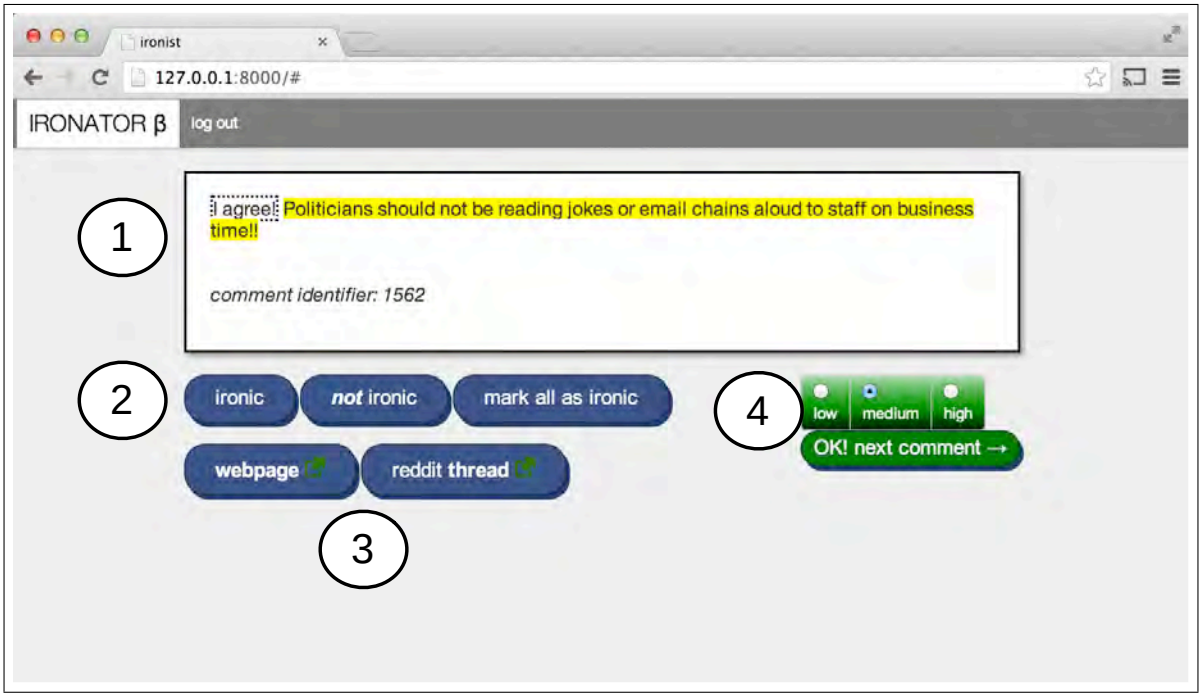

Figure 1: The web-based tool we built that was used by our annotators to label reddit comments. Enumerated interface elements are described as follows: **1** the text of the comment to be annotated – sentences marked as *ironic* are highlighted; **2** buttons to label sentences as *ironic* or *unironic*; **3** buttons to request additional *context* (the embedding discussion thread or associated webpage – see Section 4.1.2); **4** radio button to provide *confidence* in comment labels (Likert scale of *low*, *medium* and *high*).

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 respective numbers of 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.

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`. 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]

4.1.1 Annotation Process

Three Brown university undergraduates independently annotated each sentence in the corpus.[3] 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 facilitated via a custom-built browser-based annotation tool built as part of this project, shown in Figure 1.

We intentionally did not provide much guidance to annotators regarding the criteria for what constitutes an 'ironic' statement, for two reasons. First, verbal irony is a notoriously slippery concept [4] 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 [2] is 0.341, suggesting fair to moderate agreement [5], as we might expect for a subjective task like this one. Still, ideally we would perhaps achieve better agreement: we plan on re-visiting issues of annotator agreement in our future work at UT Austin (at the very least this provides an upper-bound for what we can possibly expect from an automated approach).

4.1.2 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 (and the sentences they comprise). 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

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

[3]Additional undergraduates have since worked on this project, but only three annotators are included in our β 1.0 release of the corpus.

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. Because all of this information is readily available online, we think automated approaches ought to try and exploit it.

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).

4.2 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 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.

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

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.

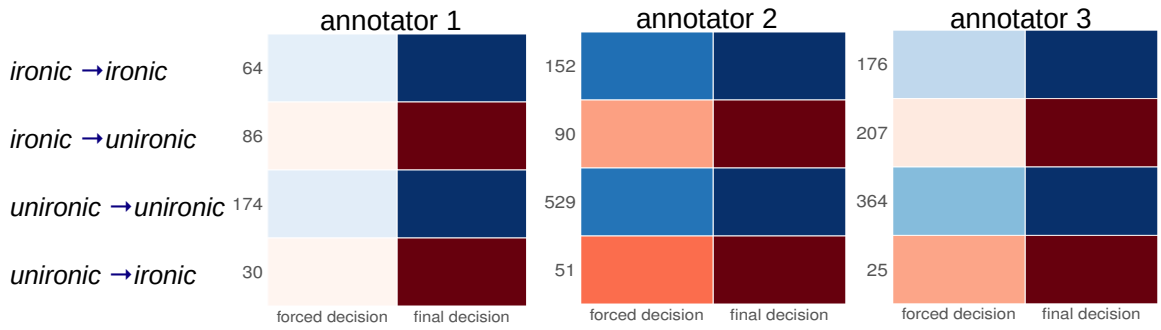

Figure 3: This plot illustrates the effect of viewing contextual information for three annotators (one table for each annotator). For all comments for which these annotators requested context, we show *forced* (before viewing the requested contextual content) and *final* (after) decisions regarding perceived 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.

4.3 Machines Probably do, too

To address research objective 2 above, we explored whether 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. It turns out that it does. This provides evidence that bag-of-words approaches are insufficient for the general task of irony detection: more context is necessary.

Specifically, 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 [3, 1] 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 (and the high ranking tokens are as expected); but obviously there is quite a bit of room for improvement.

We now explore empirically whether the 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 indicator 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:

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

Fitting this to the data, we estimated $\hat{\theta}_2 = 0.930$ with a 95% CI of (0.769, 1.093); $p < 0.001$. 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).

5 Progress summary and next steps

Toward realizing our first objective, we have collected and here described a new, publicly available corpus for the task of verbal irony detection online. 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 have 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.

Regarding our second objective, we have demonstrated that a standard token-based machine learning approach misclassified many of the same comments for which annotators tend to request context. Indeed we have shown that annotators rely on contextual cues (in addition to word and grammatical features) to discern irony and have argued that this implies computers should, too.

The obvious next step (toward realizing our third objective) 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).

We are now working on this at UT Austin. For example, we have developed a method that exploits the user-community (sub-reddit) to which a comment was posted to improve classification performance. We currently have a paper under review for potential presentation at the 2015 annual meeting of the Association for Computational Linguistics (ACL) describing this approach. We are also working on related problems of *regularization* in classification models, as we have discovered this to be particularly important for the task of irony detection. Finally, we are now collecting additional data from Twitter to be manually labeled to see if similar approaches are effective on this kind of data.

References

[1] P Carvalho, L Sarmento, MJ Silva, and E de Oliveira. 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, 2009.

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

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

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

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

[6] BC Wallace. Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, pages 1–17, 2013.

[7] BC Wallace, DK Choe, L Kertz, and E Charniak. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[8] BC Wallace and L Kertz. Can cognitive scientists help computers recognize irony? In *CogSci*, 2014.