1 Introduction and Motivation

The research objective of this proposal 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.

From the perspective of Army operational capabilities, the value of models that can infer whether an individual intends statements ironically or not from raw text is obvious: the intended meaning of an utterance can only be discerned in this way, and considering surface text alone is not sufficient for the types of sociological analyses that are of primary interest to the Army. For example, failing to recognize ironic statements during community detection may result in an author being assigned to the exact opposite sub-community to which they actually belong. Consider a scientist commenting, ironically, 'we all know global warming is a hoax'; he or she may be (wrongly) assigned to the sub-community of climate change skeptics. More generally, irony detection is necessary for any computational sociolinguistic task that requires analyzing the *content* latent in data. Models that can infer extra-linguistic information (such as social structure) from natural language have many potential Army applications, but these cannot be fully realized if irony cannot be reliably discerned.

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 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 we argue that these simple representations will often be insufficient to infer ironic intent. Progress on (potentially more sophisticated) methods for automated irony detection has been hindered by a dearth of resources. No high-quality, large corpus of annotated examples of irony presently exists, despite obvious interest in the task. In light of these observations, we propose the following research agenda.

- First, *we propose to collect and annotate a high-quality corpus to facilitate research on irony detection* (Section 2). Currently, no such datasets exist, and the datasets that have been used by the NLP and ML communities to evaluate approaches to irony detection are problematic, as we shall highlight. This a major obstacle to progress on automated irony detection, and the creation of a new corpus would be a huge boon to the research community.
- A second proposed contribution is *to analyze when existing ML and NLP technologies fail to detect ironic intent* (Section 3) empirically. We are particularly interested in assessing quantitatively whether *context* is necessary to discern ironic intent (and how often this is the case). We will use the dataset assembled in the above aim to assess the limits of existing methods.
- Our third aim is to *develop a new approach to irony detection that instantiates sociolinguistic conceptions of irony within a modern, probabilistic machine learning framework* (Section 4). The proposed approach is (1) informed by theoretical sociolinguistic perspectives on irony (and

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

thus likely capable of discerning ironic utterances missed by existing computational models), and, (2) practical enough to be operational. Developing this approach will require careful attention to metrics, both in terms of measuring distance (similarity) between individuals, and with respect to evaluation.

Our team at Brown is uniquely positioned to tackle these aims. Wallace, the PI, holds a PhD in machine learning and is a member of the Brown Laboratory for Linguistic Information Processing (BLLIP). He has written a synthesis on irony detection, surveying sociolinguistic and computational vantages [36]. Charniak, Professor in Computer Science and Cognitive Science, heads BLLIP; he is a Fellow of the American Association of Artificial Intelligence (AAAI) and was previously a Councilor of the organization. His research has always been in the area of language understanding or technologies which relate to it. Kertz is an assistant Professor in Cognitive, Linguistic and Psychological Sciences. Her research focusses on expectations in discourse modeling; she thus brings linguistic expertise. Trikalinos heads the Brown Center for Evidence-Based Medicine (CEBM) and brings expertise in statistical methods and study design.

1.1 Sociolinguistic Constructs of Irony

The ironic device is well-studied from sociolinguistic and philosophical perspectives [4, 10]. The observation from this body of work with the most import for the present proposal is that *expectations about speakers – and the context that induces them – are crucial to irony detection.* A popular, albeit incomplete, definition of verbal irony is something like: *a rhetorical trope in which the speaker says the opposite of what they mean*. While this is a very simplified definition (in that it misses many examples of irony), it captures the most obvious instances of verbal irony.

According to Grice [17, 18], speakers convey irony by flouting *conversational maxims* tacitly obeyed by interlocutors. Specifically, Grice supposes that the *maxim of Quality* prescribes that speakers are not to say what they believe to be false. Receivers then infer irony when they divine that it cannot possibly be the case that their interlocutor is being sincere. Clark and Gerrig [9] extended this vein by proposing the *pretense* theory of irony, wherein ironists are assumed to be effectively mocking a person who would sincerely articulate the proposition(s) that they are ostensibly communicating. The pretense theory postulates two audiences: those equipped to detect the ironic voice and decode the intended, latent meaning of an utterance, and those who will accept it literally.

Similar to the pretense account of irony is the *allusional pretense* theory developed by Kumon-Nakamura et al. [23]. On this view, an ironic voice is inferred if an utterance is perceived as being (1) insincere and (2) an allusion to a failed expectation. The former condition is to account for utterances in which the literal proposition expressed is not necessarily false, but the comment is intended ironically nonetheless. The authors provide an instructive example: imagine that a very knowledgable student is arrogantly dominating a classroom discussion when a classmate remarks to him "boy, you sure know a lot". The proposition may be literally true, but the remark is interpreted ironically because of inferred pragmatic insincerity.

Following Grice's maxim-violation hypothesis, Attardo [2] considers irony a form of 'relevant inappropriateness', where the speaker utters an intentionally contextually inappropriate, though relevant, remark – confident that the recipient will reject the literal meaning (on account of its being so obviously inappropriate). Note the close relatedness of this theory to Grice's original maxim-violation account [17]. This is somewhat similar to Wilson and Sperber's take, in which they re-casted ironic utterances as cases of *echoic mention* [31, 40]. On their conception, ironic

statements are always implicitly alluding to some real or hypothetical proposition, typically to demonstrate its absurdity. Utsumi [34, 35] outlined an abstract formulation of irony interpretation, where ironic utterances are understood as such only when embedded within a proper context.

Common to these theories is the idea that people infer irony when they recognize an incongruity between an utterance and what is known (or expected) about the speaker (and/or the environment); something we will refer to as the *pragmatic context*. Only listeners with a sufficient grasp of this context will discern irony, unless the speaker signals ironic intent in other ways, e.g., via *surface cues*. With respect to how the above theories might inform computational models for irony detection, the differences between these accounts thus hardly matter; the point is that all of them, from the seminal work of Grice [17] onward, imply that for any inferential system to decide if an utterance is intended ironically, it must take into account its model of the utterer. To quote Wilson and Sperber: "... the choice between literal and ironical interpretation must be based on information external to the utterance — contextual knowledge and other background assumptions ..." [31]. Thus any classification approach that operates only in the space of word counts without an explicit model of the speaker will (at least sometimes) fail. Yet existing ML/NLP methods rely more or less exclusively on surface features of text, ignoring the speaker.

1.2 ML and NLP Approaches to Irony Detection

There has recently been a flurry of methodological development for the task of computational irony detection [8, 6, 33, 14, 32, 16, 30] in the ML and NLP communities [36]. In these works, irony detection has been treated as a particular instance of *text classification*, a standard ML problem. Much of the existing work concerns identifying good (discriminating) features from text, e.g., the presence of multiple exclamation points, which seems to be used to indicate irony [13].

For example, in a recent ML approach to irony detection, Davidov et al. investigated the problem of recognizing sarcastic sentences in tweets and within Amazon[2] book/product reviews [14]. They manually labeled a small set of ironic sentences in the reviews as a 'seed set'. Their algorithm is *semi-supervised*, i.e., it exploits both labeled and unlabeled instances when inducing a model. This is in contrast to standard *supervised* machine learning, which learns only from manually labeled instances. Their approach relies on exploiting surface patterns of text, i.e., 'template sentences'. These patterns are automatically extracted from online texts, using an algorithm they proposed elsewhere [13]. They also exploit punctuation-based features. They achieve reasonable results, but their approach fails to identify many instances of irony (recall of 76% and 44% on Amazon and Twitter datasets, respectively). We discuss these datasets further in Section 2.

A similar (though fully- rather than semi-supervised) approach to irony detection was recently proposed by Carvalho et al. [8]. Their approach exploits various shallow features extracted from text (e.g., punctuation) along with word counts to discriminate ironic from genuine user generated posts taken from a news website. Their reported results are somewhat hard to interpret, however the most interesting observation to be gleaned from their work in our view is that emoticons[3] are the single best grammatical indicator of irony on web posts (at least, in their corpus).

Elsewhere, Hao and Veale [20] recently investigated the task of classifying similes as ironic or not. For example, the simile "as subtle as a freight-train" is ironic (freight-trains are decidedly unsubtle). They proposed a classification model that exploits both heuristic clues in sentences (e.g., "about as" indicates an ironic simile) and semantic relationships between the two words comprising

[2] Amazon is an online marketplace.

[3] 'emoticons' are character patterns used in text communication to indicate emotions, e.g., ':)'.

the simile, as gleaned from WordNet (http://wordnet.princeton.edu/). They also exploit the existence of certain *precedent*, or template, similes by looking for inverted variations of these. This is an interesting and promising approach (they achieved 90% recall and 60% precision with respect to recognizing ironic similes), but the specific task of ironic *simile* detection is a distinct problem from that of verbal irony detection *in general*, because the irony in such similes arises from disagreement *internal* to a given text (sentence). This obviates the need for contextual information.

Burfoot and Baldwin [6] investigated the task of classifying news articles as satirical or genuine. Their corpus comprised news articles from the satirical new source The Onion and the Associated Press (AP). The task was then to induce a classifier that could automatically discern to which of these two sources a given article belonged. Their baseline approach was a standard classifier (SVM) induced over a *Bag-of-Words* (BoW) representation of the news articles.[4] They demonstrated that this strategy fares relatively poorly, correctly identifying only 50% of the ironic articles. Their primary contribution was the introduction of novel features (beyond word counts) specific to the ironic news article detection task to improve model performance for this case. For example, they encoded the presence or absence of profanity and slang in documents (real news articles are unlikely to contain such language).

While some of these methods have shown promise, irony detection has been successful only in limited, closed domains, and even in such domains performance remains relatively poor. A solution to the general task of automated irony detection thus remains elusive. We hypothesize that this is at least in part because existing approaches rely only on surface cue features to identify irony (e.g., punctuation), and *do not model the speaker*. Such methods will succeed only when the structure of an ironic utterance is sufficiently paradigmatic to convey the speaker's ironic intent. But many instances of verbal irony will convey no such cues (or, they will contain some, but too few for the receiver to reliably infer irony). *One aim of the proposed work is to characterize when existing 'shallow' approaches will fail to detect irony*. But to accomplish this, we will need a sufficiently rich annotated dataset in which naturally occurring ironic utterances are tagged. That no such dataset presently exists greatly hinders research on models for irony detection. Hence we propose to create one, thus developing a valuable community resource that will remove a fundamental barrier to research in this field; the absence of annotated data.

2 Development of a new Corpus

In our view, an ideal irony detection system should recognize irony usage as it naturally occurs (at least online), e.g., on forums. Assuming this aim, existing evaluations of computational approaches to irony have been rather limited, or else have had a decidedly commercial focus. We briefly discuss existing datasets and highlight the need for a new annotated corpus to move research on irony detection forward.

The most common data source used to evaluate irony detection systems is Twitter [30, 16, 14, 28]. This is for practical reasons; researchers have relied on *irony* and *sarcasm* hashtags ('#') that are provided by users (tweeters) as a proxy for annotations, rather than investing the effort to manually classify tweets as 'ironic' or not. This is problematic, because tweets tagged by their authors as *irony* are often pointers to content that the tweeter considers somehow ironic (e.g., a picture), rather than actual ironic utterances. *Such tweets are not instances of verbal irony*. Indeed, automatically identifying these kinds of tweets is really an entirely different task that looks to

[4] Bag-of-Words is an unstructured representation of text.

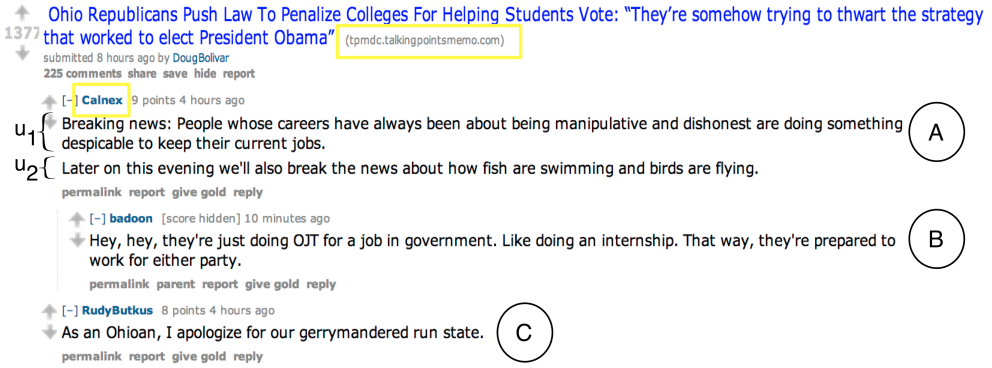

Figure 1: Illustrative comments from reddit. At the top, we see the title of the thread (this is a link to external content, which we have highlighted). Underneath, there are three posts *(or comments), marked A, B and C. Each post comprises sentences, which we will refer to as* utterances*. Both sentences in A (marked as u_1 and u_2) are examples of ironically intended utterances. Note that B is nested underneath A, whereas A and C are at the same 'level'. Usernames are associated with each comment (we have highlighted the first).*

recognize when a user will use the #irony hash tag. Moreover, irony will not always be explicitly marked as such by the speaker. (And when it is, we hardly need a system to then tell us so.)

As far as we are aware, there exist only two small datasets for which researchers have manually designated utterances as ironic (or not). The first comprises Amazon product reviews [33, 14, 27]. This is a very small dataset (fewer than 500 labeled sentences in all). Moreover, this dataset is focussed on consumer product reviews, while we are interested in more general irony usage. The second dataset comprises a small number of user posts from a Portuguese newspaper website about local politicians that were manually classified as 'ironic' [8]. These examples were sampled in a selective, non-i.i.d. way to rapidly identify examples of irony. The absence of annotated data is a major obstacle to progress on the task of automated irony detection [29].

We thus propose creating a new, large, high-quality corpus to facilitate work on automated irony detection. To capture common, social usage of the ironic voice online, we propose gathering data from *reddit* (`http://reddit.com`), a social-news website. Reddit is a very popular website (Alexa rank 118 globally, 52 in the USA [1]) 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. These comments come from registered users, and we can therefore read though previous posts by the same author; these past posts and the content of the *thread* to which a comment belongs will thus provide context. A thread comprises a title, other posts (which are nested) and an external link to content; an example is shown in Figure 1. We propose to harvest reddit comments to identify a diverse set of ironically intended utterances. This will be feasible because reddit provides an Application Programming Interface (API) that allows programs to automatically download content from the site (`http://www.reddit.com/dev/api`).

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. Consider this comment from the *politics* sub-reddit in response to an article about stock markets reaching record highs: "I thought Obama was a socialist who hated profits? Now I'm all confused." It is reasonable to surmise that the author intended this comment ironically, and indeed perusing this user's comment history supports this supposition. Comments like this are not uncommon on reddit. And given its size, social nature, the diversity of communities, and the availability of user

sub-reddit (URL)	active members (approximate)	description
politics (r/politics)	2,895,000	Political news and editorials; focus on the US.
worldnews (r/worldnews)	3,557,500	News from around the world
technology (r/technology)	3,047,900	Technology news
conservative (r/conservative)	23,000	A community for political conservatives.
progressive (r/progressive)	25,000	A community for political progressives (liberals).

Table 1: Sub-reddits from which we propose to harvest comments.

histories (which sometimes span years and comprise hundreds to thousands of comments), reddit provides rich material to study the use of irony on the internet (and hence how best to identify it). Moreover, this represents a 'data-tractable' scenario, in that similar information (i.e. post history) is likely available on many web-based platforms hosting user-generated content.

Table 1 lists the sub-reddits from which we propose to download comments. We propose sampling and downloading comments from threads belonging to each of these sub-reddits. To allow us to exploit relevant *context* (see Section 4), for each comment we will also download: (1) the thread in which the comment is embedded, (2) the history (past comments) of the commenter, and, (3) the thread's title and the associated link (see Figure 1). As mentioned above, reddit provides an API that will facilitate our scraping this data.

More specifically, our proposed sampling procedure (subject to operational changes as necessary) is as follows. For each sub-reddit (Table 1) we will retrieve all threads in order of their current *rank* on reddit at the time of sampling. Rank refers to, roughly, the current popularity of a thread, as decided by the community: users vote submissions (threads) up or down and rank is a function of the number of upvotes received and the amount of time passed since the thread was submitted. Thus newer threads with many upvotes will be highly ranked. We do not believe there is any reason to suppose that irony usage will be different on newer/popular (highly ranked) threads compared to less popular (lower ranked) threads. We will therefore iterate over these threads in rank order, sampling thread-level information and comments. The former will comprise the title of the thread and the content to which the thread links (i.e., the web-page). With respect to posts (comments), we will sample up to 100 comments that are at the first or second level in the thread's comment hierarchy (see Figure 1); lower-level comments are often tangential/off-topic posts. If there are more than 100 such high-level comments, we will sample 100 of them uniformly at random. This limit will preclude the possibility of a few large threads dominating the sample of posts we draw for a given sub-reddit. For each comment, we will also sample the 50 most recent comments posted by the same user (on any sub-reddit). Finally, we will maintain a pointer from every comment to the thread in which it was embedded. These latter attributes will constitute our 'context'.

We propose the following scheme for annotation, which we may alter if necessary. We will hire 3 annotators (Brown humanities students) to label the sentences comprising the sampled comments as *ironic* or not. All 3 annotators will label all sentences in the corpus, for a total of 5 sub-reddits $\times$ 5,000 comments (comprising ≥ 1 sentence) $\times$ 3 annotators = $\geq$75,000 labels. Such a large dataset is needed given the complexity of the task and the relative scarcity of irony. The PI and Kertz will label an additional subset of these to ensure that the annotators are trained properly and for quality control. Acquiring multiple labels for each utterance will be crucial here because discerning irony is a somewhat difficult task. Labeling (i.e., classifying utterances) will be carried out using a custom web-based tool developed by our team. We have experience developing such annotation tools [39]. Annotators will be shown utterances (and comments) embedded in their context, similar to how they appear on the reddit website (Figure 1), and will be asked to label them as ironic or not. If they cannot confidently make a decision, they will be asked to leverage available context by

considering: (1) the most recent 50 posts from the same user, (2) the content of the link associated with the thread. We will pilot the annotation process by having all annotators (and the PI) label the sentences in 200 randomly sampled comments. We will make refinements (and/or repeat the pilot round) as needed.

Finally, for every utterance annotators deem 'ironic', they will be asked to designate in the software whether they discerned irony: *(A) from the utterance (and overall comment) alone (using surface cues)*; *(B) by considering the commenter's history*; *(C) using the content of the link associated with the thread*; or, *(D) via a combination of (B) and (C)*. This information regarding how annotators came to their decision will provide insights into the question of how it is that humans are able to infer ironic intent in text-only (web-based) mediums. This information will also inform our analysis of when it is that existing ML/NLP methods fail (Section 3). We note that we may modify this operational taxonomy.

Annotators will be told to spend no more than 3 minutes per comment. In our estimation, it takes about 30 seconds, on average, to label the sentences in a comment (thus representing a feasible workload of about 630 person-hours for the numbers given above). In practice, these estimates may need to be adjusted, but we are optimistic that we may in fact be able to acquire even more labels within the allocated budget. By attaining and recording multiple annotations for each sentence, we will be able to assess agreement with respect to irony judgements. This exercise will be valuable in and of itself, as it will provide a measure of agreement regarding irony online detection (amongst humans), and a reasonable upper-bound in terms of how well we can expect a computer to perform. *This corpus will be hugely valuable in terms of facilitating research in computational sociolinguistics because no high-quality, large irony detection corpus currently exists (despite obvious interest in the task).*

This corpus will allow us to address our second research aim – quantifying when existing approaches fail to detect irony – empirically (as discussed in the following section). Moreover, the collected and annotated corpus will allow us to shed empirical insight on the issue of which metrics work well empirically. Specifically, the parametric model we propose below relies on quantifying 'distances' between individuals, but it is not clear how this distance should be measured. Distance metrics should be sociolinguistically justified and empirically supported.

3 When do Existing Method Fail?

We have argued that existing ML/NLP methods are insufficient to infer certain varieties of irony. But which, exactly? This is a question we propose to address. To do so, we will use the labeled corpus discussed above. As a first step, we will implement existing state-of-the-art statistical irony detection methods [14, 32, 16] and perform qualitative assessments of when they fail. We will also perform regression analysis to quantitatively explore when existing methods make incorrect predictions. Specifically, we will test the hypothesis that existing methods will fail on those examples for which humans require context to infer irony. This analysis will be possible because the annotators will have recorded *how* they decided that a given utterance was intended ironically (as described above), i.e., if they required context to do so.

We will also explore the empirical 'upper limits' of existing approaches. In many classification tasks, 'throwing more training data' into a learning algorithm eventually results in very good performance, even when using simple models [19]. Machine translation is a classic example of this; very large annotated corpora have allowed automated systems to perform reasonable translations,

despite using quite simple underlying models. By contrast, we hypothesize in the case of irony detection – which requires some sort of model of the speaker – the performance of simple statistical NLP methods will plateau relatively quickly, e.g., training on 10,000 rather than 5,000 examples will not make a terribly big difference.

4 Developing a new Generative Model of Irony

In this section we sketch a new model of irony that we propose to develop. This model incorporates sociolinguistic precepts regarding irony usage into a unified generative framework, while also exploiting the surface cues on which existing NLP/ML methods rely. We begin (Section 4.1) by motivating the intuition behind our model, which looks to pragmatically capture what we expect an individual to say about a given aspect, or topic. In Section 4.2, we instantiate this general approach as a probabilistic graphical model.

4.1 The *Pragmatic Context* Framework

Existing ML/NLP models do not explicitly model *expectations* regarding utterances (posts) by individuals. But sociolinguistic theories of verbal irony emphasize the receiver's model of his or her interlocutor in discerning the ironic from the sincere. Thus a primary aim of ours is to develop a model that exploits a sufficient amount of 'context' to decode irony (while simultaneously exploiting surface cues). We will refer to the information external to an utterance but necessary to infer irony as the *pragmatic context*.

We hypothesize that ML/NLP approaches need to incorporate a model of the speaker to recognize certain ironically intended utterances as such. Put algorithmically, the recipient of an utterance (1) decides what the utterance is about (we will refer to this as its *aspect*), and then (2) projects the utterance's ostensible meaning onto a spectrum concerning this aspect, and, (3) compares this projected location to the internal estimate of the utterer's position. A discrepancy between the internal model and the ostensibly conveyed sentiment in the utterance suggests ironic intent on the part of the speaker (as depicted in Figure 2). Consider an illustrative example. Suppose you know that Jim dislikes spicy food. You invite Jim out for Indian food, and he replies "Sure, since I love spicy food so much". Here the aspect would be *spicy food* and the spectrum would be *sentiment*; you may infer irony because you have an internal model of Jim that says he dislikes spicy food, but he has ostensibly communicated otherwise. Another example of a spectrum is *political polarity*; we may infer irony if a conservative friend says something ostensibly left-wing.

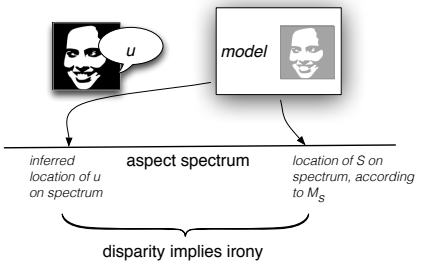

Figure 2: The listener receives an utterance and extracts the salient aspect and spectrum, then projects the utterance along this spectrum. This estimated location is compared to the location of the speaker on this spectrum, according to the internal model.

In addition to this internal model of a speaker, we also rely on the surface cues previously discussed. In conversation, these may be cues such as hyperbolic language ("I am *extremely* happy to see *her*!"). In the case of text, these are the features that are easily exploited by existing ML/NLP methods, such as emoticons and extraneous exclamation points. Depending on how well we know someone, we may place more or less weight on the surface cues in discerning irony. Subtle irony may be devoid of surface cues altogether, relying entirely on the audience's knowledge of the speaker. We can formalize this symbolically as follows. Denote by: $\mathcal{I}(u)$ the proposition 'the

utterance u was intended ironically'; $M_s(a)$ the (internal) model of the speaker s with respect to aspect a; and $S(u)$ the syntactic/internal cues extracted from u. We have:

$$P\{\mathcal{I}(u)\} \propto \lambda P\{\mathcal{I}(u)|M_s(a)\} + (1-\lambda)P\{\mathcal{I}(u)|S(u)\} \quad (1)$$

Where λ represents a scaling parameter that controls the relative contribution of the term based on our expectations regarding the speaker. For example, λ might express inverse-confidence in our user-based prediction, and thus would be small when we have access to a very limited user history; this would upweight the contribution of the surface cues. In general, the less we know about s, the smaller λ should be; we would want Equation 1 to tend toward the 'shallow' model based on surface cues only as less contextual data about the speaker is available. This high-level expression decomposes the likelihood of an utterance having been intended ironically into two parts: the first reflects what we know about the speaker (i.e., our model) while the second operates over properties internal to the utterance (e.g., syntactic features), and reflective of general patterns of irony usage. This decomposition allows us to leverage the previously developed ML/NLP methods that have focussed on the latter [8, 6, 14, 32, 16, 30].

A major research objective of this proposal is to realize sociolinguistic concepts within the framework of modern, probabilistic generative models of text. Specifically, we propose to develop and evaluate novel statistical models that estimate the probability of ironic intent given an utterance, the speaker's identity and associated contextual content. This requires addressing the following sub-tasks: (1) identifying, in a pragmatic sense, 'what' the utterance u is about (its aspect), (2) given this, inferring the sort of utterance we would *expect* of s, given whatever contextual information about them is available – e.g., past utterances, and, (3) combining these to make a prediction regarding the likelihood of irony, by inferring the 'distance' between our expected and the observed utterance, with respect to the aspect of interest (also factoring in any syntactic cues in u). We next outline a proposed generative model that looks to realize these aims within a unified probabilistic model.

4.2 Latent Aspect Expectation Model

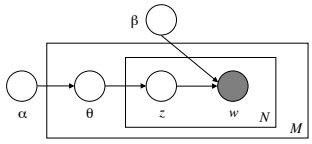

Figure 3: The graphical model of latent Dirichlet allocation (LDA) [3]. θ parameterizes the multinomial governing topics (the zs); here we treat topics as aspects. The observed words w are then assumed to be drawn from a multinomial conditioned on z. (β and α are hyper-parameters). Here the plates denote that there are N (observed) words and M topics.

Here we present our proposed approach, which we call the *latent aspect expectation* model. *This is a flexible probabilistic generative approach that jointly incorporates surface cues, latent aspects and our expectations regarding how we expect individuals to talk about them*. This approach thus instantiates the pragmatic context model discussed above. Here we make many operational decisions in terms of the specifics of the model; in part the approach we sketch here is meant as a skeleton demonstrating a concrete realization of our framework. We view investigating different assumptions as an important piece of the present proposal.

We build on top of *probabilistic topic modeling* [3], and leverage recent component-based generative models of text [15, 25, 26]. (We have recently used such models to explore patient sentiment across different aspects of outpatient care [26].) For context, the standard topic model (termed *latent Dirichlet allocation*, or LDA) is depicted graphically in Figure 3. The 'generative story' is as follows. When writing a document, an author first draws a mixture of topics θ, each of which corresponds to different distributions over the vocabulary. To generate each word in a document, the author

samples a topic z from this mixture, then samples a word w from this topic. We will sometimes refer to words as *tokens*, because we model the generation of punctuation and other features in addition to words. For our purposes, we can view 'topics' as the aforementioned 'aspects'. Given a collection of documents, one can use sampling procedures to automatically infer the latent topics.

Algorithm 1 A generative story of irony

1: Draw various hyper-parameters for all β, ω and α from $\mathcal{N}(0, \mathbf{I}\sigma)$
2: **for all** $post \in posts$ **do**
3: Choose aspect distribution $\theta_{post} \sim \text{Dir}(\alpha)$
4: **for all** utterances $u \in post$ **do**
5: draw aspect of post $a \sim \text{Mult}(\theta_{post})$
6: $\text{logit}(\pi) = \beta_0 + \beta_s + \beta_a + \beta_{\text{post}} + \boldsymbol{\beta X}$
7: draw *irony indicator* $I \sim \text{Bernoulli}(\pi)$
8: $\omega \propto \begin{cases} \omega^b + \omega^a + \omega^s + \omega^{s,a} \text{ if not } I \\ \omega^b + \omega^a + \omega^{\bar{s}} + \omega^{\bar{s},a} + \omega^I \text{ otherwise} \end{cases}$
9: draw token distribution for u $\phi \sim \text{Dir}(\omega)$
10: **for all** tokens $w \in u$ **do**
11: sample $w \sim \text{Mult}(\phi)$
12: **end for**
13: **end for**
14: **end for**

The basic idea of our proposed model is to *perturb distributions over words to reflect our expectations regarding the speaker* in addition to the aspect (topic) that a sentence is about. If an individual violates our expectations regarding what she is likely to say about an aspect, we infer irony. Furthermore, the proposed model simultaneously exploits surface cues, e.g., grammatical markers that commonly communicate ironic intent. If a speaker is being ironic, this changes the likelihood of certain tokens, e.g., exclamation points and extremely positive words may become more probable. The 'generative story' of our proposed model is told by Algorithm 1 and graphically by Figure 4. It goes as follows. For each utterance (e.g., sentence in a forum post), an individual selects an aspect (topic) according to some latent distribution over aspects. The number of aspects is a parameter of the model, but we believe that as long as this is sufficiently high, the model will not be overly sensitive to changes in this parameter.

As shown in line 6 of Algorithm 1, given the drawn aspect a, the individual decides whether or not to be ironic with a probability π that is a (linear) function of the following terms. (1) Overall prevalence of ironic usage in the corpus, expressed by β_0. (2) Their personal tendency to use irony (i.e., β_s). (3) The aspect a, captured by β_a; e.g., people may tend to be ironic more often when talking politics than when talking about technology. And, (4) the post to which an utterance belongs (β_{post}). We assume that a post spans one or more utterances (sentences) but has a single θ_{post}. β_{post} thus captures the intuition that if one sentence in a post is intended ironically, it is more likely that others in the same post are, too. Finally, $\boldsymbol{\beta X}$ represents additions to the model we may wish to later incorporate.

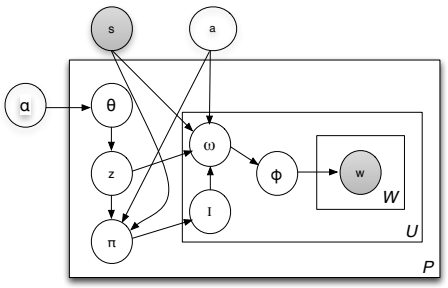

Figure 4: A graphical representation of our proposed model (see Algorithm 1 for the corresponding generative story). Shaded nodes are observed. An aspect index z is drawn for each post from a multinomial θ. This aspect and the speaker (s), affect the probability π of an ironic utterance. Together, the speaker, aspect and irony indicator (I) define the parameter to a multinomial ϕ from which tokens are drawn.

If the author decides to be sincere, then she selects each token comprising the utterance from a distribution reflecting general usage throughout the corpus, the drawn aspect, her usual 'voice' (i.e., way of talking or writing) and her typical way of talking about a; these correspond to the terms ω^b, ω^a, ω^s and $\omega^{s,a}$, respectively (line 8). Note that each ω term is a vector with dimensionality equal to the number of tokens (vocabulary size). If, on the other hand, she decides to generate an ironic statement, then the distribution over tokens from which she draws will be adjusted to account for (in addition to the aspect and baseline usage) a 'voice' that is different from the speaker's own ($\omega^{\bar{s}}$ and $\omega^{\bar{s},a}$, discussed below) and surface cues characteristic of irony (ω^I).

We have operationalized this by imposing a log-linear model over corresponding terms on to the parameter to the Dirichlet from which word distributions are drawn (line 8 of Algorithm 1; we follow typical probabilistic topic modeling distribution choices in our use of the Dirichlet). We will place priors around the ω variables drawn from independent normals centered around 0 [25]. This realizes our aim of the model regressing to the 'surface cue' based model when we do not have information about the speaker, i.e., when $\omega^{\bar{s}}$ is near 0 (similar to setting $\lambda = 0$ in Eq. 1).

Two important terms in the model are $\omega^{\bar{s}}$ and $\omega^{\bar{s},a}$. Here we use $\bar{s}$ to denote something like 'unlike speaker s'; in sociolinguistic terms, this corresponds to the group that the speaker is targeting with their irony [18, 2]. The idea behind this term is that *how* a speaker talks (and specifically how she talks about the aspect a) will tend to mimic the target group when she is being ironic [40], and this will be reflected in word usage. Roughly, $\omega^{\bar{s}}$ should encode a perturbation that reflects the usage of individuals 'distant from' the speaker s (and, similarly, the term $\omega^{\bar{s},a}$ should be a set of weights that reflects aspect-specific token usage by such individuals specifically when they talk about aspect a). A simple choice here would be $\omega^{\bar{s}} = -\omega^{s}$, $\omega^{\bar{s},a} = -\omega^{s,a}$. A more sophisticated model, however, would set this to explicitly reflect individuals different from the speaker s (with respect to the aspect a). For example, to detect political irony, we would want $\omega^{\bar{s}}$, $\omega^{\bar{s},a}$ to reflect the token-usage of someone on the other end of the political spectrum from s.

To accomplish this within the machinery of our model, we could express $\omega^{\bar{s}}$ ($\omega^{\bar{s},a}$) as weighted averages over $\omega^{s'}$ ($\omega^{s',a}$) terms corresponding to other individuals s', weighted by their distance to s (with respect to a). This would realize the construct that individuals assume a voice of a (dissimilar) target group, or at least one other than their own, when being ironic [40, 17, 2, 31]. Technically, this approach requires quantifying how 'distant' individuals are from one another. *Exploring different metrics to capture this distance is one of our research aims*. But we next discuss two concrete approaches that we will take initially.

If we have access to previous posts by individuals then one strategy to set the $\omega^{\bar{s}}$, $\omega^{\bar{s},a}$ terms would be to use metrics defined over the inferred aspects of previous posts by other users, under the assumption that similar users will post about similar things, in general. Specifically, we would have a vector for each user that captures the proportion of utterances across latent aspects, and could calculate a distance between two individuals using these vectors (e.g., the KL-divergence [24]). *We acknowledge that the specific distance metric used is important and 'just using', e.g., the Euclidean distance, is not satisfying. Identifying more appropriate metrics is a research problem we propose to address.*

Alternatively, if we have access to social network information, we may use *latent space network* (LSN) models [21, 22], which map observed links (adjacency graphs) into a lower-dimensional geometric space that captures user similarity. Specifically, we can take distances with respect user coordinates in the low-dimensional space induced by an LSN model. In the case of reddit data, we have access to the sub-reddits to which users post (and the frequency with which they do so). This information can be used as a proxy for network information. Of course, the aim of this proposal is *not* to identify irony on reddit, specifically, but to develop general approaches to the task that rely on data that will commonly be available. We think it is reasonable to assume that a proxy for social network information will often be available (and in many cases network data will be available explicitly).

Once the model is fully specified, we can perform inference via sampling methods [12, 25, 26]. This will also allow us to exploit available annotated data. And because the model is Bayesian, we will have full posterior distributions over parameters of interest (e.g., the probability that a

sentence was intended ironically, π), providing a means to assess uncertainty. This is an advantage over existing methods, which typically have provided only binary classifications regarding whether an utterance was intended ironically. But probabilities can feed into down-stream systems. For example, in systems wherein false positives (assuming ironic intent when the speaker was in fact being sincere) carry a high cost, we may only want to ultimately assume someone is being ironic when we are quite sure of it.

In summary, we have sketched a flexible, probabilistic model of language generation that captures surface cues in utterances *and* expectations regarding users. These expectations are based on previous posts (by the speaker and by similar users) and, if available, network structure. Of course, more details remain, some of which we address below.

5 Specific Research Aims

We summarize our research agenda here in three parts. These comprise the research aims and questions related to: corpus creation, empirical evaluations of existing approaches and the development and evaluation of a novel irony detection model.

5.1 Creation of a New Corpus

The first aim of this proposal is to create a high-quality corpus comprising labeled examples of ironically intended utterances. This corpus will also include contextual information in the form of user-histories. This is crucial to advancing research on computational approaches to irony detection because no such corpus presently exists. This task itself raises interesting questions we propose to address.

- **How much agreement is there amongst humans regarding the classification of online text as ironic?** We will answer this by having three annotators label all sentences in the retrieved corpus, and then assessing inter-rater agreement (e.g., via the kappa statistic [7]). Inter-annotator agreement establishes upper-limits on the performance of automated systems.
- **How do humans infer ironic intent in social media/online content? To what extent do they rely on available context?** As discussed in Section 2, we propose collecting information to answer this question during data annotation. We will develop our annotation tool in-house, and this tool will provide a mechanism for the labeler to indicate *how* they infer ironic intent when they do. Specifically, we will create an operational taxonomy that will include the following options (as discussed in Section 2): *(A) from the utterance (and overall comment) alone (using surface cues)*; *(B) by considering the commenter's history*; *(C) using the content of the link associated with the thread*; *(D) via a combination of (B) and (C)*. We will instruct annotators to provide this information for every utterance that they mark as ironic. This will thus provide valuable insight into how individuals discern irony, and in particular to what degree they rely on surface cues versus contextual information. This has implications for model development: surely if a human requires contextual information then so too will a model.
- **How often is irony used on social media platforms?** We will answer this question empirically, at least as it applies to the collected corpus. Our proposed sampling scheme draws from several 'sub-reddits' (Table 1), focussing on distinct interests (e.g., politics, technology), and is thus arguably representative of irony usage online in general.

5.2 Assess the Limits of Existing ('Shallow') Methods

Our second main aim is to address the question: **When do existing statistical ML/NLP approaches fail to detect irony? And what are the limits of existing approaches?**

- **How well do existing approaches fare in terms of detecting irony?** We will answer this empirically using the dataset we collect. Presently, we know only how well existing supervised ML/NLP approaches do with respect to datasets that are either relatively small (in the case of the Amazon product review data [33, 14, 27]) or labeled with non-standard definitions of what constitutes verbal irony, as in the twitter datasets often used [30, 16, 14, 28]. *The large, high-quality corpus of labeled irony examples we collect will afford the first possibility of assessing how well these supervised ML models can perform in detecting the ironic voice*.

- **What are the properties of the examples on which existing approaches fail?** We will qualitatively examine cases in which existing models fail – both in terms of false negatives and false positives – and from these observations deduce quantitative explanations for failures. This will allow us to understand when simple (or 'shallow') approaches work and when contextual information is necessary. Using data gathered from the above sub-aim (assessing the extent to which humans rely on context), *we will perform regression analyses to evaluate associations between the errors the model makes and the cases for which humans indicate that they require contextual information to classify an utterance*. Our hypothesis, of course, is that existing methods for irony detection fail on those cases on which humans are unable to make a decision without context.

- **Does annotating additional training data help? If so, how much?** We will assess the degree to which additional training data can improve the performance of existing methods by increasing the fraction of the annotated corpus available to the learning algorithms. *Our hypothesis is that existing ML/NLP approaches will quickly 'max out' in terms of prediction performance. That is, 'throwing more data' at the problem will not help here.*

5.3 A new Model for Irony Detection

The third aim of this work is to develop a new, sociolinguistically informed model for irony detection that improves on the state-of-the-art. We outlined this model above (Section 4.2). We will evaluate components of this sub-aim empirically using the corpus assembled for aim 1.

- **We will develop a probabilistic model that incorporates contextual expectations**. Specifically, we will implement a version of the model we sketched above (Algorithm 1). We will initially select simple operationalizations of the required variables and definitions. *We will develop a practical means of inferring model parameters from labeled data, likely using some version of Gibbs sampling [12].*

- **We will evaluate the performance of the developed model empirically, using the collected corpus**. In general, this evaluation will be conducted via standard cross-fold validation, wherein we will fit (estimate the parameters of) the model using a subset of the data and then assess its performance on the remainder of it. Initially, we will use standard metrics such as recall, precision, specificity and F-score to compare the developed approach to alternative methods. However, *these general metrics may be crude in that they require dichotomization of the model's*

probabilistic prediction. It may be important to assess the calibration of these probability estimates, because these could feed directly into a cost-sensitive system that decides what to do with ironic utterances. Calibration performance is important. For example, if one is performing community discovery, one could integrate the predicted probability that an utterance was intended ironically directly into the discovery model, ultimately begetting a confidence that a given individual belongs to a specific group. Thus one important point in evaluation will be to assess how accurate the predicted probabilities are (given the true labels). Because irony is relatively rare, however, metrics that quantify the performance of probability estimates such as the classic Brier-score [5] are inappropriate. This is because most metrics assess *overall* calibration, and thus consistently predicting low probabilities for *all* examples will result in ostensibly good performance. To address this issue, we have recently proposed *the stratified Brier-score* [37, 38] as a more appropriate metric for measuring the quality of probability estimates in imbalanced scenarios. Aside from the relative performance of the developed model in terms of prediction, in general, we are specifically interested in whether the developed model is able to discern irony where existing models fail to. We can test this directly by assessing the predicted probabilities of ironic intent over instances on which existing models make mistakes.

Evaluation will be an iterative process; we will continue to assess performance as we modify components of the model. Indeed, our model comprises several distinct components, and we will perform ablation experiments to assess the effects of each of these; e.g., how much does factoring in the corpus-wide prevalence of irony (β_0) help? And so on. However, *during model refinement, we will also hold-out a test set with which we will validate our 'final' proposed model* (perhaps around 10% of the entire corpus). This will remove the risk of indirectly over-fitting the collected corpus through model tuning.

- **We will experiment with variations of important model parameters**. The model we sketched in Section 4.2 makes several operational decisions (implicit or explicit) regarding definitions and functional forms. For example, we have assumed a simple linear form over individual terms corresponding to factors we believe likely to affect the decision to be ironic (or not). We will experiment with adding interaction terms (e.g., perhaps a speaker is often ironic about only a specific aspect), and potentially with using non-linear functions over said factors. This is just one example; we will iteratively refine and tinker with several components of our model. We next discuss at some length a particular facet of our model, the $\omega^{\bar{s}}$ and $\omega^{\bar{s},a}$ terms, which are meant to perturb the distribution over tokens to 'look like' the group the ironist is tacitly mocking.

- **What are 'good' distance metrics to quantify distances between individuals? (And what does 'good' mean here, anyway?)** Recall that an important component of our proposed model (Section 4) is to assume that, when employing an ironic voice, an individual will assume a vocabulary that reflects the usage of 'dissimilar' persons who she is tacitly mocking. Operationalizing this idea requires quantifying similarity. Ideally, this quantification will be done in a way that is both sociolinguistically and mathematically sound. At the same time, such a metric needs to be 'data-tractable', i.e., work with data likely to be available. We thus aim to exploit previous posts (i.e., user histories), which will often be available on social network platforms such as forums. *We believe this research on representations of individuals using practically available information (such as past posts and social network information), and on similarity metrics over such representations, will be of broad interest.*

More specifically, recall that the components $\omega^{\bar{s}}$ and $\omega^{\bar{s},a}$ are meant to capture what we do *not* expect s to say. One way of accomplishing this would be to set, e.g., $\omega^{\bar{s}} = -\omega^{s}$. But we also proposed defining these terms by taking weighted averages over over components $\omega^{s'}, \omega^{s',a}$ for individuals s' other than s where corresponding weights would be *the distance from s' to s*; this more directly captures the sociolinguistic theories that we have reviewed. But for this approach we need to quantify distances between individuals. We propose investigating (developing and evaluating) various metrics that capture this distance. Above (Section 4.2) we proposed two initial approaches with which we will experiment, one based only on previous posts and the other on social network information (or a proxy thereof) and leveraging latent space network (LSN) models [21, 22]. In addition to these two concrete metrics, we will work with team-member Kertz (whose research in sociolinguistics concerns how individuals choose to formulate utterances) to develop novel quantitative measures that capture 'distance' in a pragmatic sense.

To evaluate these metrics, we can first consider the relative change in the predictive performance of the model (with respect to predicting ironic intent); a 'good' metric would here be one that improves irony detection performance, with respect to the metrics already discussed. A more direct approach to empirically assessing the reliability of similarity measures will be to consider individuals that we know ought to be 'far' from one another, *a priori*, and ascertaining that they indeed are, according to the metric. We will also assess metrics in this direct way. To do so, recall that we have proposed collecting data from both the *conservative* and *progressive* sub-reddits (Table 1), representing posts from individuals at either end of the (US) political spectrum. We expect users that predominantly post the former to be 'dissimilar' with respect to those who post mostly in the latter (and, further, we expect those that post mostly in the *progressive* sub-reddit to be similar to other individuals who predominantly post to this sub-reddit). Hence we can confirm if the metrics that we have proposed (and others that we develop) agree with this *a priori* knowledge regarding individuals.

6 Dissemination Plan

We will make all developed code open-source and available online, via the PI's github account (github is a web-based platform for sharing source code; `http://github.com`). We will share this dataset and documentation online, both on github, the PI's academic homepage and the homepage of the Brown Laboratory for Linguistic Information Processing (`http://bllip.cs.brown.edu/`). We will publicize this dataset via publications and presentations.

7 Summary

Broadly, the major aims of this proposal are: (1) to collect and annotate a relatively large corpus comprising labeled examples of online irony use, (2) use this corpus to evaluate when existing methods for irony detection fail, and, (3) to develop and evaluate (using the collected corpus) a new model for irony detection that is motivated by sociolinguistic theory but is realized within the framework of probabilistic generative models of text. We believe that these aims align with the Army's, due to the proliferation of user-generated content online and the need to automatically make sense of it.

References

[1] Alexa: statistics summary for reddit.com (`http://www.alexa.com/siteinfo/reddit.com`), April 2013.

[2] Salvatore Attardo. Irony as relevant inappropriateness. *Journal of Pragmatics*, 32(6):793–826, 2000.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] Wayne C Booth. *A Rhetoric of Irony*. University of Chicago Press, 1975.

[5] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[6] Clint Burfoot and Timothy Baldwin. Automatic satire detection: are you having a laugh? In *Proceedings of the ACL-IJCNLP Conference: Short papers*, pages 161–164. Association for Computational Linguistics, 2009.

[7] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[8] Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio de Oliveira. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). *Proceeding of the CIKM workshop on Topic-Sentiment Analysis (TSA) for mass opinion*, pages 53–56, 2009.

[9] Herbert H Clark and Richard J Gerrig. On the pretense theory of irony. *Journal of Experimental Psychology*, 113:121–126, 1984.

[10] Claire Colebrook. *Irony*. Routledge, 2004.

[11] Herbert L Colston. On necessary conditions for verbal irony comprehension. *Pragmatics & Cognition*, 8(2):277–324, 2001.

[12] William M Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 642–647, 2011.

[13] Dmitry Davidov and Ari Rappoport. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 297–304. Association for Computational Linguistics, 2006.

[14] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics, 2010.

[15] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *International Conference on Machine Learning (ICML)*, 2011.

[16] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, volume 2, pages 581–586. Association for Computational Linguistics, 2011.

[17] Herbert P Grice. Logic and conversation. *Synatx and semantics: Speech arts*, 3:41–58, 1975.

[18] Herbert P Grice. Further notes on logic and conversation. *Syntax and Semantics*, 9:41–57, 1978.

[19] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[20] Yanfen Hao and Tony Veale. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds and Machines*, 20(4):635–650, 2010.

[21] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[22] Peter Krafft, Juston Moore, Bruce Desmarais, and Hanna Wallach. Topic-partitioned multi-network embeddings. In *Advances in Neural Information Processing Systems*, pages 2816–2824, 2012.

[23] Sachi Kumon-Nakamura, Sam Glucksberg, and Mary Brown. How about another piece of pie: the allusional pretense theory of discourse irony. *Journal of experimental psychology: General*, 124:3–21, 1995.

[24] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM, 2007.

[25] Michael Paul and Mark Dredze. Factorial LDA: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems 25*, pages 2591–2599, 2012.

[26] Michael J Paul, Byron C Wallace, and Mark Dredze. What affects patient (dis)satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, 2013 (forthcoming).

[27] Antonio Reyes and Paolo Rosso. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 118–124. Association for Computational Linguistics, 2011.

[28] Antonio Reyes and Paolo Rosso. Building corpora for figurative language processing: The case of irony detection. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (in conjunction with LREC 2012)*, pages 94–98, 2012.

[29] Antonio Reyes and Paolo Rosso. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, pages 1–20, 2013.

[30] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 2012.

[31] Dan Sperber and Deirdre Wilson. Irony and the use-mention distinction. *Radical pragmatics*, 49:295–318, 1981.

[32] Joseph Tepperman, David Traum, and Shrikanth Narayanan. "Yeah right": Sarcasm recognition for spoken dialogue systems. In *International Conference on Spoken Language Processing*, 2006.

[33] Oren Tsur, Dmitry Davidov, and Ari Rappoport. ICWSM – A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the fourth international AAAI conference on weblogs and social media*, pages 162–169, 2010.

[34] Akira Utsumi. A unified theory of irony and its computational formalization. 2:962–967, 1996.

[35] Akira Utsumi. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806, 2000.

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

[37] Byron C Wallace and Issa J Dahabreh. Class probability estimates are unreliable for imbalanced data (and how to fix them). In *International Conference on Data Mining (ICDM)*, pages 695–704. IEEE, 2012.

[38] Byron C Wallace and Issa J Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and Information Synthesis (KAIS)*, 2013.

[39] Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, and Thomas A Trikalinos. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 819–824. ACM, 2012.

[40] Deirdre Wilson and Dan Sperber. On verbal irony. *Lingua*, 87(1):53–76, 1992.