Artif Intell Rev (2015) 43:467–483
DOI 10.1007/s10462-012-9392-5

# Computational irony: A survey and new perspectives

**Byron C. Wallace**

Published online: 8 January 2013

**Abstract** Irony is a fundamental rhetorical device. It is a uniquely human mode of communication, curious in that the speaker says something other than what he or she intends. Recently, computationally detecting irony has attracted attention from the natural language processing (NLP) and machine learning (ML) communities. While some progress has been made toward this end, I argue that current machine learning methods rely too heavily on shallow, unstructured, syntactic modeling of text to consistently discern ironic intent. Irony detection is an interesting machine learning problem because, in contrast to most text classification tasks, it requires a *semantics* that cannot be inferred directly from word counts over documents alone. To support this position, I survey the large body of existing philosophical/literary work investigating ironic communication. I then survey more recent computational efforts to operationalize irony detection in the fields of NLP and ML. I identify the disparities of the latter with respect to the former. Specifically, I highlight a major conceptual problem in all existing computational models of irony: none maintain an explicit model of the speaker/environment. I argue that without such an internal model of the speaker, irony detection is hopeless, as this model is necessary to represent *expectations*, which play a key role in ironic communication. I sketch possible means of embedding such models into computational approaches to irony detection. In particular, I introduce the *pragmatic context model*, which looks to operationalize computationally existing theories of irony. This work is a step toward unifying work on irony from literary, empirical and philosophical perspectives with modern computational models.

**Keywords** Irony · Representation · Machine learning

## 1 Introduction

Inferring that utterances are intended ironically and subsequently re-constructing their latent meaning is a complex task, yet it is one that humans seem to perform with relative ease. Indeed,

B. C. Wallace (✉)
Center for Evidence-Based Medicine, Brown University, Providence, RI, USA
e-mail: byron_wallace@brown.edu

the ability to discern verbal irony appears to develop early on in life: children are able to do so at the age of six (Pexman and Glenwright 2007). Illuminating the underlying mechanisms that facilitate ironic communication thus has the potential to improve our understanding of human thought processes, language and communication.

Aside from its philosophical import, the pragmatic benefits of computationally detecting verbal irony have recently been recognized by the machine learning community (Davidov et al. 2010; Tepperman et al. 2006). Perhaps the most immediate benefit would be the improvement of models for *sentiment analysis* (Pang and Lee 2004), in which the aim is to automatically assess the latent, subjective emotion expressed in a text or utterance. For example, a standard sentiment analysis task is to induce a model capable of classifying movie reviews as 'positive' or 'negative' (see, e.g., Kennedy and Inkpen 2006; Pang and Lee 2004; Kanayama and Nasukawa 2006; Sindhwani and Melville 2009). The presence of verbal irony is responsible for many of the errors made by state-of-the-art methods for this task (Carvalho et al. 2009); automatically detecting ironic intent on the part of the reviewer would thus improve said methods. More broadly, any model that hopes to make sense of human communication or expression must be able to discern the ironic from the sincere.[1]

The task of inferring irony is academically interesting in part because existing machine learning techniques (discussed at length in Sect. 3.2) are not very good at it. In particular, it seems that the classic 'bag-of-words' representation typically used in text classification, in which each text document is mapped to an indicator vector that encodes the words, or *n*-grams, present therein, is insufficient for the task.[2] This is in contrast to the somewhat surprising success that this simple representation has achieved for other text classification tasks, provided enough training data, a phenomenon Halevy et al. refer to as "the unreasonable effectiveness of data" (Halevy et al. 2009).

The ironic/unironic distinction is thus a unique classification task: unlike general text classification (and the closely related sentiment analysis problem), I argue that *word counts alone are an insufficient representation for verbal irony detection*. In particular, building on existing theories of verbal irony (discussed in the following section) I argue that a model of the speaker (or, broadly, the contextual environment) is a *necessary* condition for irony detection, making it a formidable (and interesting) task from a machine learning perspective. I discuss previous attempts to formalize irony detection, both abstractly (in Sect. 3.1) and more concretely, via recent machine learning approaches (in Sect. 3.2). I discuss what I believe to be theoretical deficiencies of the latter methods in light of the aforementioned theories of verbal irony. Finally, in Sect. 4, I introduce an operational account of irony to mitigate these shortcomings. This model looks to be 1) informed by theoretical perspectives on irony (and thus capable of discerning ironic utterances missed by existing models), and, 2) practical enough to be operational for computational irony detection.

## 2 Theories of verbal irony

Entire books have been written on the ironic device (e.g., Booth 1975; Colebrook 2004), and I do not attempt to provide a thorough general treatment of the subject here, as that would be beyond the scope of this work. Instead, I attempt to delineate a theory of verbal irony and

[1] A robot incapable of recognizing irony would be unable to communicate with humans naturally.

[2] More complex representations exist, but this is the canonical scheme for text classification. Furthermore, all text representations of which I am aware are, ultimately, functions over word counts.

how it is recognized from the existing literature, in order to later juxtapose this with formal symbolic attempts at capturing it.

A popular, albeit incorrect, definition of verbal irony is something like: *a rhetorical trope in which the speaker says the opposite of what they mean*. This definition captures the most obvious sorts of ironic utterances. For example, imagine that Kelly's liberal-democrat friend Joe remarks to her: "Sarah Palin sure is smart". It is likely that in this case Joe means to express the exact opposite of this statement, and Kelly will probably infer this with ease. How might Kelly perform this task?

On the theory put forth by Grice (1975), the speaker (Joe) is here intentionally violating the *maxim of Quality*, one of four 'rules' of conversation tacitly obeyed by interlocutors. Grice refers to these as conversational maxims.[3] The maxim of Quality prescribes that speakers are not to say what they believe to be false. On this view, Kelly recognizes ironic intent on Joe's part because she knows that her (liberal) friend cannot possibly be sincere in what he has said. Moreover, she knows that he knows that she is aware of this, and she must therefore assume that he is intentionally flouting the conversational maxim of sincerity. This obvious violation forces Kelly to reject the literal interpretation of what he has said. Kelly is then left with the task of constructing what Joe *actually* intended; and here Grice suggests that the negation of the ironic proposition is the most obvious interpretation.

Grice's theory works nicely for simple cases like the one above. However, as pointed out by Wilson and Sperber (1992), among others (e.g., Utsumi 1996), the simple definition of verbal irony as a trope in which the ironist says the literal opposite of what they intend is far from comprehensive. To illustrate its inadequacy, consider *ironic understatement*, in which the speaker highlights the intensity of something by deliberately and markedly understating it. For example, imagine that you encounter a woman on the sidewalk during a torrential downpour and she says "it's raining just a bit". One would interpret this as an amusing allusion to the downpour. However, the literal negation of this statement is that it is *not* raining a bit, which of course is not what the speaker intended. Wilson and Sperber (1992) provide additional examples of verbal irony that this definition does not capture. Moreover, the problem isn't only that this definition is not sufficient (in the sense that it does not cover all cases of verbal irony), it would also seem to suggest that certain decidedly un-ironic statements in fact are Wilson and Sperber (1992).

To overcome this deficiency, Wilson and Sperber re-casted ironic utterances as cases of *echoic mention* (Sperber and Wilson 1981; Wilson and Sperber 1992). On their conception, ironic statements are always implicitly alluding to some real or hypothetical proposition, typically to demonstrate its absurdity. Thus to say "it's raining just a bit" in a downpour is to mock that very proposition; given the circumstances (downpour), it would be absurd to make such a statement, and it is exactly this that the speaker is pointing out. (Of course, highlighting the absurdity of the notion that it is raining "just a bit" also serves as an indirect way of communicating "it is down-pouring".) The irony-as-echoic-mention theory views utterances as being *about* their literal proposition. This is known as the *use-mention* distinction: genuine statements *use* propositions, whereas ironic utterances *mention* them. Note that this is consistent with the "air-quotes" gesticulation commonly associated with verbal irony, which explicitly conveys that an utterance is about what is being said.

[3] The other three are: Quantity, Relation and Manner; these are not immediately relevant to the discussion here.

Elsewhere, (Clark and Gerrig 1984) proposed the *pretense* theory of irony, extending the work of (Grice 1975, 1978). On this account, ironists are affecting a pretense when speaking; i.e., they are acting a part, 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 ironical voice and decode the intended, latent meaning of an utterance, and those who will accept it at its surface meaning. This theory eschews the *use-mention* distinction central to Wilson and Sperber's view; the ironist is not *mentioning* a literal proposition but rather is *pretending* to be someone who would use this proposition. Absurdity plays an important part in this theory; the ironist is pretending to communicate something manifestly absurd, thereby mocking the position he or she ostensibly espouses. Note that unlike other theories of irony, the pretense view accounts for sustained exercises in irony, e.g., Jonathan Swift's famous Modest Proposal (Swift 1955).

Similar to the pretense account of irony is the *allusional pretense* theory developed by Kumon-Nakamura and Glucksberg (1995). 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. (Interestingly, recent work Colston 2001 has shown empirically that pragmatic insincerity may not be a necessary condition for verbal irony comprehension, though of course it may be sufficient.)

Following Grice's maxim-violation hypothesis, (Attardo 2000) considers irony a form of 'relevant inappropriateness'. On this model, 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). Consider the above schoolboy example: here the remark in question is is judged to be relevant, though inappropriate (with respect to the environment in which it was uttered) and its literal implication is hence rejected. Note the close relatedness of this theory to Grice's original maxim-violation account (Grice 1975).

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 I 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 1975) onward, imply that for any inferential system to decide if an utterance is ironic it must take into account its model of the utterer. To quote Wilson and Sperber: "... it is clear that the choice between literal and ironical interpretation must be based on information external to the utterance—contextual knowledge and other background assumptions—rather than the form or content of the utterance itself" (Sperber and Wilson 1981). Indeed, (Clark and Gerrig 1984) provide many examples in which the same sentence can (and should) be interpreted as ironic in some contexts, and genuine in others. It follows that any classification approach that operates only in the space of word counts, with no external knowledge or semantics, will fail. Yet as we shall see, existing methods rely more or less exclusively on shallow, syntactic features of text (utterances).

## 3 Previous work on formalizing irony

I will now turn from general theories of irony to more precise (symbolic) formalisms. In particular, I review two disparate bodies of work, broad philosophical formalisms and more pragmatic, operationalized attempts to computationally detect verbal irony. I discuss the limits of current work in the latter category, particularly in light of the work reviewed above.

### 3.1 Philosophical formalisms of irony

The first attempt to symbolically formalize irony is due to Utsumi (1996, 2000). He first argues that an ironic utterance must be communicated in a proper context, i.e., the speaker of an ironic utterance must be embedded in an *ironic environment*. This environment is realized under the following three conditions:

- The speaker has an expectation **E** at time $t_0$.
- This expectation **E** fails (is violated) at time $t_1$.
- Consequently, the speaker is displeased in this violation of their expectation (i.e., the disconnect between their belief of what would happen and the reality of what took place).

On Utsumi's theory, an ironic statement made by the speaker in this environment would allude to their failed expectation. Consider a simple example to illustrate this model. Suppose Jonathan and Stephen are on vacation. They wake one morning to a dreary, overcast day. Looking out the window, Jonathan remarks: "What a lovely day". On Utsumi's interpretation, Jonathan here is alluding to his failed expectation that it would be nice out. Stephen then infers Jonathan's (ironic) intent by recognizing this allusion.

While their theories are ostensibly similar, Utsumi claims that Wilson and Sperber's (1992) (see Sect. 2) echoic-mention account is too narrow: ironic utterances sometimes allude to expectations *derived* from failed expectations, rather than only the failed expectation itself. Utsumi implicitly assumes that the speaker of an ironic utterance is (in actuality) expressing a negative sentiment. This is in line with empirical evidence that irony is generally used to convey negative sentiment (Colston 1997).

Utsumi presented a symbolic formalism to capture his conception of irony, exploiting the machinery of *situation calculus*. Crucially for modeling the expectations integral to Utsumi's account, this calculus provides a mechanism for representing an agent's mental state. Utsumi's model is an ideal abstract symbolic formalism for irony; it integrates agents, beliefs, propositions and time, and posits that agents exploit this context to appreciate irony. Unfortunately, it is defined at to too high a level to operationalize computationally. For example, automatically classifying the example utterances he presents in his article as ironic would require first *grounding* text, a task beyond the capabilities of modern natural language processing technologies. I will later revisit this issue and Utsumi's theory.

### 3.2 Machine learning approaches to irony detection

There has recently been a flurry of methodological development for the task of computational irony detection (Carvalho et al. 2009; Burfoot and Baldwin 2009; Davidov et al. 2010; Tepperman et al. 2006) in the machine learning (ML) and natural language processing (NLP) communities. I will be focussing on what I view as the insufficiencies of the computational models reviewed below, in light of the theoretical accounts reviewed above. However, I should note that the researchers who developed these methods were not aspiring to develop general models of irony, but rather pragmatic algorithms for irony detection. Furthermore, the proposed models provide useful machinery for inferring irony from surface cues; I will argue that these shallow methods should thus be combined with models that exploit contextual information. In all of the computational works I review, irony detection has been treated as a particular instance of *text classification*, a standard problem in machine learning that I will now briefly review.

### *3.2.1 Text classification*

Text classification is the problem of inducing a model, or *classifier*, capable of categorizing documents (e.g., newspaper articles) into one of $k$ categories or *classes* (e.g., 'sports' and 'world news'). The aim is thus to induce a function that maps (representations of) documents into their respective classes. Popular text classification algorithms include Naive Bayes (Lewis 1998) and Support Vector Machines (SVMs). The latter is widely accepted to be the 'state-of-the-art' in terms of text classification (Joachims 1998).

To apply an SVM or any other inductive learning algorithm to a set of documents (i.e., a *corpus*), the documents must first be mapped into vector representations. These vectors are points in what is referred to as the *feature-space*. The canonical encoding for text is known as 'bag-of-words' (BOW), a simple scheme that works as follows. First, an $n$-dimensional space is defined over the corpus, where $n$ is the total number of words therein (usually uninformative *stop-words* such as 'the' and 'or' are first removed). Each document $d$ is then represented as a point $\mathbf{x^d}$ in this space, where $\mathbf{x^d_j}$ is 1 if the $j$th word exists in document $d$ and 0 otherwise.[4] The features here are (functions over) word counts.

Perhaps surprisingly, this simple binary representation for text has proven sufficient even for seemingly complex classification tasks. For example, SVMs induced over BOW-encoded movie reviews can distinguish between 'positive' and 'negative' reviews (as labeled by humans) with 80–90 % accuracy (see, e.g., Pang and Lee 2004). It is tempting to conclude that similar results using BOW might eventually be obtained for any task, given enough training data. But results reported thus far for the task of automated irony detection have not been nearly as good, suggesting that irony may not be discernible via this shallow representation. Indeed, in a few cases (Burfoot and Baldwin 2009; Davidov et al. 2010, discussed below) additional representation resulted in substantial performance gains, even though the learning algorithm was not changed. This suggests that representational issues are integral to the irony detection task.

### *3.2.2 Machines learning irony*

In a recent approach to computational irony detection typical of machine learning approaches, Davidov et al. investigated the problem of recognizing sarcastic sentences in 'tweets'[5] and Amazon[6] book/product reviews (Davidov et al. 2010). In particular, they experimented with one dataset of  6 million tweets and another of 66,000 Amazon product reviews. They created gold standard subsets of these comprising manually labeled data, i.e., subsets labeled as 'ironic' or not by humans. Their algorithm, called SASI (Davidov et al. 2010), is a *semi-supervised* strategy, i.e., it exploits both labeled and unlabeled instances/examples when

[4] Variants on this method exist, including the popular term frequency/inverse document frequency (TF-IDF) scheme, but these, too, are ultimately some function over word counts in documents.

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

[6] Amazon is an online marketplace.

inducing a model. This is in contrast to standard *supervised* machine learning, which learns only over labeled instances.

Their approach is outlined as follows. First, they identify patterns, i.e., template sentences. These patterns are automatically extracted from online texts, using an algorithm they proposed elsewhere (Davidov and Rappoport 2006). They then construct feature-vectors comprising indicators of the presence (or absence) of these patterns in the corresponding input texts. Additionally, they extract punctuation-based features (e.g., the number of exclamation marks in a sentence). Notice that their approach is entirely shallow; they represent texts with features reflecting surface patterns and punctuation alone. Nonetheless, they achieved reasonable, if underwhelming, results with this simple representation; 91 % precision at 76 % recall on the Amazon corpus and 72 % precision at 44 % recall on the Twitter dataset.[7] An interesting note here is that their approach will, by construction, only be able to detect cliche ironies (i.e., ironies following formulaic construction). Novel ironies would therefore escape detection.

A similar—though fully- rather than semi-supervised—approach to irony detection was recently proposed by Carvalho et al. (2009). 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 my view is that emoticons[8] are the single best grammatical indicator of irony on web posts (at least, in their corpus).

In an interesting work, (Tepperman et al. 2006) exploited the audio features of speech to detect irony. In particular, they restricted themselves to individuals uttering "yeah right", both ironically and in earnest. They encoded various spectral properties of the sound associated with each utterance (e.g., pitch, rising/falling frames, etc.), and their classification was performed in this space. This is a direct analogue of the shallow approach for text; their classification is based solely on surface structures of speech. They achieve reasonable results, obtaining 87 % accuracy an F-measure of 70 %,[9] though one should keep in mind their restricted application to utterances of "yeah right". Interestingly, however, these results were only achieved when they supplemented the audio features with rudimentary 'contextual' features, such as the sex of the speaker; using audio features alone performed worse. This is in line with our central thesis that context and semantics are a pre-requisite to reliable irony detection. Note that this work is in line with empirical observations that voice modulations accompany ironic utterances in speech (Scharrer and Christmann 2011).

While some of the above methods have shown promise, it should be noted that successful detection of irony has been achieved only in limited, closed domains. A solution to the general task of automated irony detection remains elusive; the performance of existing classifiers is relatively poor. In my view, this is because all of the methods reviewed exploit only essentially shallow features to identify irony. Such methods will succeed, of course, only when the structure of an ironic utterance is sufficiently paradigmatic to convey the speaker's ironic intent. However, many instances of verbal irony will convey no such clues (or, they will contain some clues, but too few for the listener/receiver to infer irony with any degree of confidence). In light of the discussion in Sect. 2, it is clear that such shallow models will not capture even straight-forward ironies, e.g., your liberal-leaning friend singing Sarah

[7] *Recall* here quantifies the total fraction of ironic sentences identified by the algorithm; *precision* refers to the fraction of sentences classified by the algorithm as ironic that in fact were.

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

[9] *Accuracy* is the total fraction of utterances correctly classified. *F-measure* is a harmonic mean of precision and recall.

Palin's praises. Indeed, on the pretense theory of irony (Kumon-Nakamura and Glucksberg 1995), a skilled ironist will adopt the syntactic structure of their target, thereby rendering it undetectable at this shallow level.

Consider another example; restaurant reviews. You cannot possibly ascertain whether a stranger is being ironic when they proclaim "I love McDonald's". By contrast, should your friend utter this exact same sentence, you very well may be able to infer whether he or she is being ironic (your friend is a health conscious food snob) or sincere (your friend loves junk food), often with high confidence. It is clear that, because the sentence is *exactly the same*, neither word counts nor shallow grammatical clues will suffice. Context, too, is key. Someone who writes "... and the food is so healthy!" in a review of a fast-food restaurant is presumably being ironic, whereas if he or she wrote this in a review of a health-food oriented cafe it would probably be intended sincerely. Despite this, none of the existing computational methods attempt to model the speaker or environment in order to infer irony. That said, there have been models that have incorporated interesting features beyond word counts and other shallow elements. I now review two of these, and discuss what I view as their contributions.

In a departure from the shallow, syntactic models reviewed thus far, (Hao and Veale 2010) recently investigated the task of classifying similes as ironic or not. For example, consider the simile "as subtle as a freight-train". This is clearly intended ironically, as freight-trains are decidedly unsubtle. Hao remarks: "... we see that freight-trains are very difficult to conceptualize as 'things that are subtle' and so we consider the description to be ironic." Hao and Veale (2010). They proposed a classification model that exploits both heuristic clues in sentences (e.g., the presence of "about as" is a strong indicator of an ironic simile) and semantic relationships between the two words comprising the simile, as gleaned from WordNet. They also exploit the existence of certain *precedent*, or template, similes by looking for inverted variations of these.

Hao and Veale (2010) achieve relatively promising empirical results (90 % recall of ironic similes a 60 % precision). Moreover, theirs is the most *semantic* of the computational models reviewed here, i.e., they do not rely on bag-of-words text representation. However, the specific task of ironic *simile* detection is an easier problem than the verbal irony detection task *in general*, because the irony in such similes arises from disagreement *internal* to a given text (i.e., sentence); Hao and Veale refer to this as *text-internal* irony (Hao and Veale 2010). Thus, critically, no model of the speaker or environment is necessary to infer ironic similes of this type, as the irony is specified completely by the text therein, by definition. What is needed is a similarly semantic approach to irony detection in the general, *text-external* case. This, however, will require a model of the agent doing the uttering—a strategy not yet adopted in any of the irony detection algorithms proposed thus far.

In other work that I view as being in the right direction, (Burfoot and Baldwin 2009) investigated the task of classifying news articles as satirical or genuine. Their corpus comprised news articles from The Onion[10] and the Associated Press (AP). The task was then to induce a classification model that could automatically discern to which of these two sources a given article belonged. Their baseline approach was a standard SVM induced over a BOW representation of the news articles. 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.

In particular, they exploited a few intuitions about news articles. First, they encoded the presence or absence of both profanity and slang in documents, as real news articles are

[10] The Onion is a satirical news source.

unlikely to contain such language (thus a new feature is introduced into the vector representation of a document to indicate if profanity is present therein). Second, they explicitly encoded the words present in the article's headline, as humans are typically capable of recognizing sarcastic articles from the headline alone. Most interestingly (in my view) was their introduction of a *validity* feature, which attempts to measure the absurdity of a news article.

Consider, for example, The Onion's running portrayal of Joe Biden, which has provided coverage of the vice president stirring controversy by appearing in risque Hennessy ads. This news coverage is immediately recognized as ironic by humans because of the absurdity of the proposed situation. Burfoot and Baldwin (2009) proposed a practical method to capture this intuition; they used Named Entity Recognition to (automatically) identify the proper nouns in each article (say, Joe Biden and Hennessy), and then performed a web search via Google of the conjunction of these entities ("Joe Biden AND Hennessy"). They added to the feature set the number of matching documents returned by Google. The intuition is that unlikely combinations of entities will return fewer search results, indicating 'absurdity' and suggesting irony. Note that this notion of absurdity is consistent with the 'relevant inappropriateness' model of irony discussed by Attardo (2000), in which a violation of contextual appropriateness signals ironical intent.

One can also re-cast this model in terms of failed expectations. In particular, theirs is a discriminative model that discerns ironic from real news articles; the latter category may be viewed as the *expectation* of how news articles are written, the former being the case when this fails. Hence the inclusion of, for example, profanity as a feature. One does not expect that real news articles would contain profanity, and thus its presence is indicative of verbal irony. Similarly, the attempt to quantify absurdity may be viewed in the context of expectations. In light of the pretense theory of irony (Clark and Gerrig 1984), I might conclude that if an article seems too absurd to be genuine, then someone is adopting the *pretense* of a news writer penning the story, presumably in order to mock the ostensible content or stance of the article. A straight-forward modification to this approach would be to include features encoding the presence (or absence) of hyperbole (e.g., words such as *amazing*), and other such verbal cues outlined by Utsumi (2000). In any case, it is notable that the 'feature engineering' (including profanity, etc.) done by Burfoot and Baldwin drastically improved results. Indeed, the classifier induced on instances represented with these additional features achieved ∼68 % recall and ∼ 96% precision, as compared to the baseline classifier induced over binary BOW alone, which managed only 50 % recall and a ∼94 % precision.

## 4 New perspectives on computational irony

The fundamental shortcoming in all of the existing machine learning models for recognizing verbal irony (reviewed in Sect. 3.2) is their failure to explicitly model *expectations* regarding utterances. More broadly, this shortcoming can be viewed as a want of a *pragmatic context* in which to perform inference. Current statistical approaches have no contextual model on which they can condition their decisions regarding a given utterance. Instead, they have relied almost exclusively on shallow, unstructured features alone, using only the semantics implicitly gleaned from the Bag-of-Words model. These approaches will work insofar as verbal and grammatical cues are necessary and sufficient conditions of irony.

By contrast, the literary and philosophical theories of verbal irony, reviewed in Sect. 2, emphasize the role of the environment and the receiver's model of his or her interlocutor in discerning the ironic from the sincere. As put by Schaffer (1982):"Recognition of irony

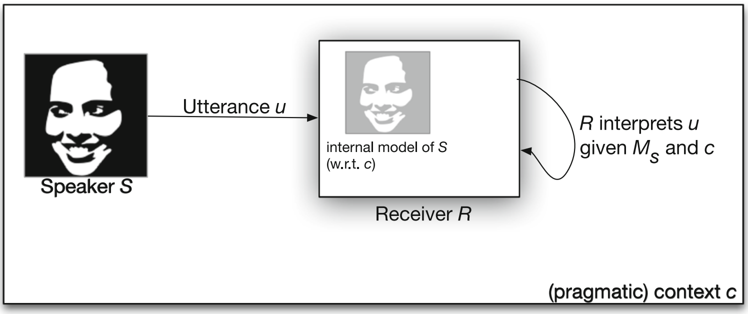
Utterance $u$
internal model of $S$ (w.r.t. $c$)
Speaker $S$
Receiver $R$
$R$ interprets $u$ given $M_S$ and $c$
(pragmatic) context $c$

**Fig. 1** The pragmatic context model. The receiver (computational model) contains a model of the speaker that informs its decision regarding whether or not an utterance $u$ is intended ironically

rarely comes from the words themselves". Context is everything in irony detection, yet existing machine learning approaches to the problem essentially ignore it.

### 4.1 Pragmatic context: a proposed computational framework for detecting verbal irony

What is needed is a machine learning system that incorporates a model of the environment and the speaker sufficiently rich to recognize $u$ as ironic. Such a system is depicted schematically by Fig. 1. On this view, an utterance can be understood only with respect to its *pragmatic context*. This context comprises the elements, or *aspects*, that are sufficient to infer that an utterance is intended ironically. These aspects, which reflect beliefs about a user and/or the broader environment, encode expectations by capturing what we believe the speaker would likely (not) say. Aspects will sometimes be propositional ("it is raining today") but more often will be estimates along a spectrum. For example, *political leaning* is an aspect. Many of us contain internal estimates of our friend's locations along this spectrum. Indeed, in the example above it is what enabled Kelly to infer that her friend's remark about Sarah Palin was intended ironically.

The ironist assumes their target audience is sufficiently cognizant of the pragmatic context to decode the true intent of their utterance, i.e., that they have good-enough internal estimates of the relevant aspect to do so. This may be (the speaker's) political orientation or the weather, for example. In any case, the message can be decoded *only* with the requisite knowledge regarding the relevant aspect. On the pragmatic context model, the recipient of an utterance 1) decides which aspect[11] constitutes its pragmatic context, and then 2) projects the utterance's (semantic) contents onto this aspect's spectrum, comparing this projected location to their internal estimate of the utterer's position with respect to the corresponding aspect (see Fig. 2).

A discrepancy between the internal model and the inferred placement of the utterance suggests ironic intent on the part of the speaker. Syntactic cues provide additional evidence: the posterior likelihood that the speaker intended an utterance ironically is a function of (a) the discrepancy between the internal model of speaker, with respect to the relevant pragmatic context, and (b) the syntactic cues present in the remark. The former should be weighted by the confidence we have in our internal model of the speaker, with respect to the relevant aspect: the more we know someone, the better able we are to infer when they are being ironic. Conversely, the less we know someone (the weaker our internal model of them, with

[11] I will restrict the discussion to utterances that tacitly address only a single aspect.

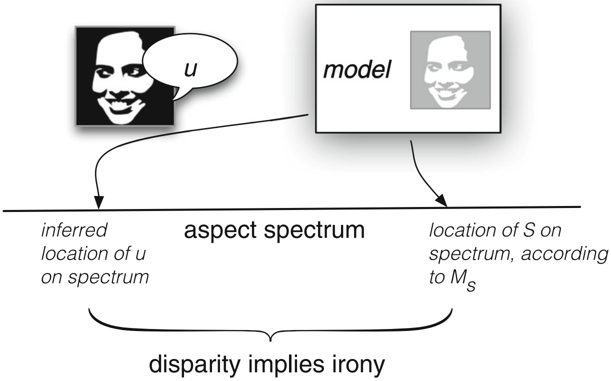
u
model
inferred location of u on spectrum
aspect spectrum
location of S on spectrum, according to M_S
disparity implies irony

**Fig. 2** The model receives an utterance and extracts the salient aspect. The model must then project the utterance along this aspect's spectrum. This estimated location is compared to the location of the speaker on this spectrum, according to the internal model

respect to a given aspect) the more we rely on syntactic cues and other properties internal to an utterance.

We can formalize the above 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 by $Var\{M_s(a)\}$ the 'variance' of this model, i.e., some measure of inverse confidence in our model. Finally, denote by $S(u)$ the syntactic/internal cues extracted from $u$. We have:

$$P\{\mathcal{I}(u)\} \propto \left(\frac{P\{\mathcal{I}(u)|M_s(a)\}}{Var\{M_s(a)\}}\right) \cdot P\{\mathcal{I}(u)|S(u)\} \quad (1)$$

This expression breaks 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). This decomposition is useful because it allows us to leverage the previously developed machine learning methods that have focussed on the latter. (The computational models reviewed in Sect. 3.2.2 drop the first term, and rely exclusively on properties internal to the utterance.) In practice, the first term, which relies on the estimate $M_s(a)$ and our confidence therein, could be estimated in many different ways, with various degrees of accuracy and effort. To make the approach more concrete, I will consider an example in Sect. 4.3. First, however, I will consider the relationship of the pragmatic context model to other conceptualizations.

### 4.2 Relationship to other models

The pragmatic context model may be viewed as a practical computational instantiation of many of the models discussed in Sect. 3.1. It is perhaps most directly a realization of Utsumi's *failed expectations* model (Utsumi 1996, 2000): the pragmatic context may be viewed as a more concrete example of what Utsumi referred to as the *ironic context*. However, Utsumi specifies only that the speaker have some sort of expectation at a given time, whereas on our view inferring the salient aspect of an utterance is an explicit computational task. Nor does Utsumi consider the contributions of syntactical cues to the inference. Finally, Utsumi does not offer any feasible computational way of modeling expectations (of course, his aim was not to operationalize irony detection, but rather to give a theoretical account thereof). Here I suggest leveraging existing machine learning technologies to build a probabilistic user-model that captures expectations.

The proposed model may also be viewed from the vantage of Grice's account (Grice 1978). Here the aspect modeling pertinent to the utterance (which provides the pragmatic context for the exchange) provides the means to assess whether the speaker is intentionally flouting the maxim of Quality (i.e., to infer if the speaker is saying something he or she cannot literally mean). Under Wilson and Sperber's *echoic mention* theory (Wilson and Sperber 1992), meanwhile, the pragmatic context contains the information necessary to infer if a proposition is being *mentioned* or *used*. If the communicated proposition is clearly not something the speaker would say (according to our model), then we must conclude that the speaker is merely alluding to someone who *would* believe such a thing. Finally, the pragmatic context model may also be interpreted as a natural operationalization of Attardo's *relevant inappropriateness* model (Attardo 2000): it is with respect to the aspect that an utterance is judged to be relevant yet unbelievable (taken at face value), and hence inappropriate.

### 4.3 An example

Suppose we are interested in classifying individual sentences in restaurant reviews as ironic or sincere. If a history of individual users is available (as it is on many online review sites, e.g., Yelp!),[12] one could construct a model of the types of places a person tends to like, and from these one could estimate the user's position with respect to various relevant aspects. In addition, one might build a language model that captures the writing style of the individual: this, too, would contribute to the first term in Eq. 1. Given an utterance, the task is then first to infer the salient aspect therein. The strategy would then be to estimate the disparity between the sentiment regarding said aspect (ostensibly) communicated by their utterance and the sentiment expected, according to the internal model of the utterer. Concretely, in the case of Yelp!, this internal model $M_s(a)$ might be induced using the user's past likes/dislikes, and the opinions of similar users to which we have access. This information could be used to construct a simple probabilistic estimate $M_s(a)$, e.g., using collaborative filtering techniques. The confidence in our estimate would be a function of the number of past reviews from said user to which we have access.

Imagine, for example, that we encounter a review of a BBQ restaurant written by a (vegetarian) user who typically frequents health-food restaurants. Suppose that he or she begins this review by writing "We all know how much I love meat...". Notice first that the reviewer is postulating an in-group and an out-group—those who know the writer is in fact a vegetarian, and thus discern the irony, and those who do not. In-group/out-group dynamics play a role in many theories of irony (e.g., Clark and Gerrig 1984; Attardo 2000). According to the proposed model, the in-group comprises those with a sufficient grasp of the pragmatic context. The aim of an irony detection system, then, is to become a member of this in-group, i.e. by recognizing the ironic voice. The fact that has the most import for our discussion here is that *no syntactic features extracted from this sentence would by themselves imply irony*;[13] many people, of course, genuinely do love meat. Thus any machine learning system that relies only on (some function of) word counts over sentences (and/or other syntactic features, such a grammar) would necessarily fail to recognize this sentence as ironic. However, using the

[12] http://yelp.com

[13] This is not to say that it will *never* be possible to infer irony from syntactic cues alone. However, subtle forms of irony may well be completely devoid of such cues.

user history to construct a coarse model of the reviewer, we might be able to recognize this sentence as an instance of irony.

Indeed, we might do so by first classifying the *affect* of the sentence ("love") with respect to its aspect ("meat"). The former task is within reach of modern NLP/ML techniques, at least in restricted domains. The latter, i.e. discerning the aspect, may be (imperfectly) accomplished using modern NLP methods (Yi et al. 2003). Using the (ostensible) sentiment classification of this sentence, we could essentially pose the computational question: "what is the probability that this user would [affect] [aspect]?" —in this case the relevant spectrum would be sentiment (regarding meat). This projection would constitute our $M_s(a)$.

The answer to this could be inferred in various ways. One might locate historically similar users (via clustering, perhaps) and assess their general sentiment regarding *subject* by inferring the affect of sentences containing it. Alternatively, one could build upon recent work by Guerra et al. (2011), in which they developed a sentiment analysis algorithm that exploits the structure of 'endorsements' in social networks to infer latent bias. The latter approach is promising also because social networks may provide opportunity to explicitly construct the in- and out-groups posited by most theories of irony. In any case, the (inverse-) confidence we have in our model ($Var\{M_s(a)\}$) will be informed by the number of relevant data points to which we have access. However the sentiment is inferred, if it seems to contradict the ostensible feeling expressed in the given sentence, then there is a disparity between our expectation (projection) and what the utterance communicates, implying ironic intent (Fig. 2). (We might also inform the probability estimate regarding a statement's likelihood of being ironic by its surface sentiment, recalling that irony is typically employed to express actually negative/ostensibly positive sentiment Wilson and Sperber 1992).

The work of Grice (1978) (later extended by Clark and Gerrig 1984) offers an additional insight that might be exploited in estimating the first term of Eq. 1: irony as *pretense*, or affectation. On this view, the ironist is adopting a voice not their own. It might therefore be fruitful to first construct a *language model* of the speaker, and then when deciding if a new utterance $u$ from said speaker is or is not intended ironically, the relative likelihood of $u$ under the induced language model could be assessed. If $u$ is unlikely, i.e., does not fit with our model of the speaker, then the posterior probability that he or she is being ironic increases. On our model, a disparity between the internally estimated language model of the speaker and $u$ offers additional evidence that the utterance is not something the speaker would say; this sort of estimate can easily be incorporated into the first term of Eq. 1.

The preceding sketches a few possible approaches to estimating the first term in Eq. 1. For the second term—which estimates the probability of an utterance being ironic given its surface cues (word counts, grammatical features, etc.)—one could exploit any of the machine learning methods reviewed in Sect. 3.2.2. Note that this component of the model would be *general*, i.e., would be applicable to all users, while the first term would be contingent on the specific speaker (and to a lesser extent, the specific aspect). Thus, for example, if a user has no previous history, $Var\{M_s(a)\}$ will effectively be infinity and the first term will drop out; in such cases the classifier will have to rely exclusively on the syntactic elements of $u$. On the other hand, the lower the variance (the more user-history available, say) the more we will emphasize our internal model over the syntactic cues.

The approaches sketched above are meant as illustrative examples. Unfortunately, it seems that it will be necessary to construct aspect- and user-specific models for ironic inference to operationalize the pragmatic context model. This will undoubtedly be challenging, but not impossible. Indeed, the popularity of social networks continues to surge, and preferences can be inferred with relatively little effort via social graphs (Hogg 2010). Exploiting networks is a general approach to the task, and hence partially mitigates the need to be build specific models for aspects. Furthermore, as natural language processing systems become more sophisticated, this area of user-modeling will almost certainly receive increased attention, as speaker-models are integral to nearly every sort of interaction.[14] I re-iterate that this internal speaker-model should be combined with syntactic cues in the utterance when deciding whether or not it was intended ironically. I have provided an operational framework for accomplishing this (expressed in Eq. 1), though how accurate a user-model can be constructed for given aspects (and how well these salient aspects can be extracted from utterances) remains an empirical question at present. Nonetheless, I believe doing so may be within reach of modern NLP/ML tools, as evidenced by the preceding example.

## 5 Harder problems of irony

The discussion thus far has been concerned with the discernment of irony in utterances. A natural corollary to this is the task of re-constructing the latent, intended meaning of a given ironic utterance. To my knowledge, this task has not been addressed by the NLP/ML communities, perhaps due to its difficulty. Nonetheless, it is instructive to consider what solving this task would involve.

Note first that de-coding ironic intent is only coherent in the case of what Booth refers to as ‘stable irony’ (Booth 1975), i.e., irony in which there actually exists an underlying meaning. This is in contrast to ‘post-modern’ modes of irony, in which no such ground truth exists. Indeed, given that ironic speech is first and foremost an act of disassociation from the literal statement (Wilson and Sperber 1992), the task would perhaps better be thought of *interpreting*, rather than re-constructing, the intended meaning of a given remark. In any case, the problem is easiest when considering simple ironic utterances, in which the speaker’s intended meaning is the exact opposite of their literal remark. As discussed, this naive (though common) account of verbal irony is not nearly sufficient to account for the varieties of irony used in practice (Sperber and Wilson 1981; Booth 1975), but nonetheless captures many common cases (consider product reviews, in which perhaps the majority of ironic comments are affected praise, e.g., a dissatisfied customer may follow a litany of complaints with “In short, a great product!”).

For these simple ironic utterances, the task is clear: extract the (most salient) proposition from an utterance and negate it. Proposition detection in free-text has received some attention, e.g., Bethard et al. developed a method for opinion proposition detection in the context of mining the opinions held by speakers, achieving passable, if not compelling, performance (51inary BOW alone, whi% recall, 58inary BOW alone, whi% precision) (Bethard et al. 2004). Unfortunately this method merely detects the presence or absence of propositions; extracting the propositional *c*ontent would require additional computational machinery.

The grand aim would be to extract propositions into the situational calculus algebra used by Utsumi, or a similar formalism (Utsumi 1996, 2000). Solving the task of mapping free-text to symbolic expressions (situational calculus) representing agents, propositions and expectations would be a huge step forward for artificial intelligence, as it subsumes issues of grounding and entailment. It is thus unlikely to be feasible in the immediate future. Whether or not automatic interpretation of ironic statements is viable without such an intermediate step remains an interesting open question.

A second task that I have not seen addressed is automated *generation* of ironic statements, i.e., a generative model for irony. This would be in the vein of existing work on ‘computational

[14] Very early work on this sort of thing exists, see, e.g., (Davey 1978).

creativity' problems, including for example generative methods for automatically producing music (Puckette 1996) and poetry (Greene et al. 2010). I have argued that a model of the speaking agent is necessary for the task of irony detection. This is also true for the case of irony generation, though in this case the modeling requirement is once removed: the computer must model the person's model of itself (the computer). For example, a user's computer may 'know' that said user is a Republican. The computer might then monitor news feeds, and perform an ironic speech act regarding a political outcome to its owner, e.g., perhaps sending the (ironically intended) message "great news: the Democrats won!" to its Republican owner. In addition to requiring a model of the user (to many people this genuinely would be great news!), information regarding entities and their relationships, as well as a natural language generation system (Reiter and Dale 2000) would be required to accomplish this. Existing techniques might work for restricted domains, though not for general discourse, as this would involve first solving natural language generation.

## 6 Discussion

The aim of this work was to tease out the philosophical issues inherent to computational irony tasks, thereby elucidating shortcomings in current approaches and guiding future work. More specifically, I reviewed theoretical accounts of irony and juxtaposed these with modern, pragmatic approaches to detecting ironic statements recently investigated by the natural language processing and machine learning communities. I have argued that the latter must re-visit the former if computational approaches are to succeed at robust irony detection.

More specifically, computer scientists have thus far approached the irony detection task as a particular instance of the standard text classification problem, in which the aim is to induce a model capable of discriminating 'ironic' from 'unironic' utterances. Some progress has been made using these existing machine learning methods, but modern classifiers remain substantially poorer at discerning irony than are humans. I have argued that the fundamental issue with the machine learning techniques that have been proposed to date is that they operate almost exclusively over the space of textual or syntactic features, and these will never be sufficient to detect all ironic utterances. Indeed, this is necessarily true, given that the exact same sentence can be intended ironically by one speaker and genuinely by another. It thus follows that a model of the speaker is imperative for automated irony detection.

To this end, I have proposed the *pragmatic context* model, which operationalizes many earlier theories of irony discernment in a probabilistic framework. This model outputs the probability of an utterance *u* having been intended ironically as a function of two components. The first is the discrepancy between the position along the spectrum (sentiment regarding) the salient *aspect* of *u* and an internal estimate of the speaker's location along this spectrum; this is term is weighted by the confidence we have in the model, i.e., how much data we have about the speaker. The second component assesses the probability of an ironic utterance as a function of shallow features, e.g., word counts and grammatical features: the methods proposed in the machine learning literature thus far restrict themselves to this component. These estimates are synthesized into a single classification decision regarding whether *u* is intended ironically or not. Unlike previous computational models, this estimate would incorporate contextual knowledge about the user, and thus be able to infer subtle forms of irony.

While the *pragmatic context* model is computationally practical, I have only provided a sketch of how it might operationalized. The primary aim of this work was to discuss the shortcomings of existing computational methods, in light of theoretical accounts of irony, and to suggest how these might be remedied. Further explorations will require better datasets on which irony detection methods may be tested. These datasets must include some sort of information about the users, however, if irony detection is to be successful. As I have argued, this is because unlike most classification tasks, simply 'throwing more data' at existing machine learning algorithms will not do the trick. And, indeed, current approaches do not perform all that well. A good irony detection model requires some amount of semantics (i.e., regarding the literal proposition being expressed) and a model of the speaker (the expectation that he or she believes what is said). These points have been well covered in the theoretical literature on irony and its usage; the artificial intelligence community ought to take notice.

## References

Attardo S (2000) Irony as relevant inappropriateness. J Pragmat 32(6):793–826

Bethard S, Yu H, Thornton A, Hatzivassiloglou V, Jurafsky D (2004) Automatic extraction of opinion propositions and their holders. In: 2004 AAAI spring symposium on exploring attitude and affect in text, p 2224

Booth W (1975) A rhetoric of irony. University of Chicago Press, IL

Burfoot C, Baldwin T (2009) Automatic satire detection: are you having a laugh? In: Proceedings of the ACL-IJCNLP 2009 conference short papers, pp 161–164. Association for Computational Linguistics (2009)

Carvalho P, Sarmento L, Silva M, de Oliveira E (2009) Clues for detecting irony in user-generated contents: oh...!! it's so easy;-) pp 53–56

Clark H, Gerrig R (1984) On the pretense theory of irony. J Exp Psychol 113:121–126

Colebrook C (2004) Irony. Routledge

Colston H (1997) Salting a wound or sugaring a pill: the pragmatic functions of ironic criticism. Discourse Process 23(1):25–45

Colston H (2001) On necessary conditions for verbal irony comprehension. Pragmatics & # 38. Cognition 8(2):277–324

Davey A (1978) Discourse production: a computer model of some aspects of a speaker. Edinburgh University Press

Davidov D, Rappoport A (2006) Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words, pp 297–304

Davidov D, Tsur O, Rappoport A (2010) Semi-supervised recognition of sarcastic sentences in twitter and amazon. Conference on natural language learning (CoNLL) p 107

Greene E, Bodrumlu T, Knight K (2010) Automatic analysis of rhythmic poetry with applications to generation and translation. In: Proceedings of the 2010 conference on empirical methods in natural language processing, pp 524–533. Association for computational linguistics

Grice H (1975) Logic and conversation. pp 41–58

Grice H (1978) Further notes on logic and conversation pp 113–127

Guerra P, Veloso A, Meira Jr W, Almeida V (2011) From bias to opinion: a transfer-learning approach to real-time sentiment analysis. KDD

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12

Hao Y, Veale T (2010) An ironic fist in a velvet glove: creative mis-representation in the construction of ironic similes. Minds and Machines pp 1–16

Hogg T (2010) Inferring preference correlations from social networks. Electron Commer Res Appl 9(1):29–37

Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. Machine Learning: ECML-98 pp 137–142

Kanayama H, Nasukawa T (2006) Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of the 2006 conference on empirical methods in natural language processing, pp 355–363. Association for computational linguistics

Kennedy A, Inkpen D (2006) Sentiment classification of movie reviews using contextual valence shifters. Comput Intell 22(2):110–125

Kumon-Nakamura S, Glucksberg S (1995) How about another piece of pie: the allusional pretense theory of discourse irony. J Exp Psychol Gen 124:3–21

Lewis D (1998) Naive (Bayes) at forty: the independence assumption in information retrieval. Machine Learning: ECML-98 pp 4–15

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on association for computational linguistics, p 271. Association for computational linguistics

Pexman P, Glenwright M (2007) How do typically developing children grasp the meaning of verbal irony? J Neurolinguistics 20(2):178–196

Puckette M (1996) Pure data: another integrated computer music environment. In: Proceedings of the second intercollege computer music concerts pp 37–41

Reiter E, Dale R (2000) Building natural language generation systems

Schaffer R (1982) Vocal cues for irony in english. Diss, (unveroff.), The Ohio State University

Scharrer L, Christmann U (2011) Voice modulations in german ironic speech. Lang Speech 54(4):435–465

Sindhwani V, Melville P (2009) Document-word co-regularization for semi-supervised sentiment analysis. In: Data Mining, 2008. ICDM'08. Eighth IEEE international conference on, pp 1025–1030. IEEE

Sperber D, Wilson D (1981) Irony and the use-mention distinction

Swift J (1955) A modest proposal for preventing the children of poor people in ireland from being a burden to their parents or country; and for making them beneficial to the public (1729). Irish Tracts pp 1728–1733

Tepperman J, Traum D, Narayanan S (2006) "Yeah Right": sarcasm recognition for spoken dialogue systems

Utsumi A (1996) A unified theory of irony and its computational formalization pp 962–967

Utsumi A (2000) Verbal irony as implicit display of ironic environment: distinguishing ironic utterances from nonirony. J Pragmat 32(12):1777–1806

Wilson D, Sperber D (1992) On verbal irony. Lingua 87(1–2):53–76

Yi J, Nasukawa T, Bunescu R, Niblack W (2003) Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: ICDM, pp 427–434. IEEE