ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-765

Detecting Ironic Intent in Creative Comparisons

Tony Veale and Yanfen Hao[1]

Abstract. Irony is an effective but challenging mode of communication that allows a speaker to express sentiment-rich viewpoints with concision, sharpness and humour. Irony is especially common in online documents that express subjective and deeply-felt opinions, and thus represents a significant obstacle to the accurate analysis of sentiment in web texts. In this paper we look at one commonly used framing device for linguistic irony – the simile – to show how irony is often marked in ways that make it computationally feasible to detect. We conduct a very large corpus analysis of web-harvested similes to identify the most interesting characteristics of ironic comparisons, and provide an empirical evaluation of a new algorithm for separating ironic from non-ironic similes.

1 INTRODUCTION

Irony is a curious form of double-speak in which a speaker implies the opposite of what is said [5], or expresses a sentiment in direct opposition to what is actually believed [6]. Intriguingly, an ironic speaker does this in the hope that the audience will actually see past this artifice to comprehend the speaker's actual meaning. On the surface, this seems a most irrational, round-about and risky way to communicate meanings [13]. But on closer analysis, irony reveals itself to be anything but round-about: it is, in fact, a very compact way of saying or doing two useful things at once. Irony can be used to divide an audience into those who "get it" and those who don't; it can be used to soften a criticism with humour, or more often, to salt a wound by cloaking it in an apparent compliment that is quickly dashed; and most concisely of all, it can echo a viewpoint that is advanced by another while simultaneously undermining that viewpoint [13]. For instance, the ironical comparison "*you are about as tough as a marshmallow cardigan*" (from our web-corpus) integrates the expectation that the audience ("you") is believed to be "tough" with a comparison that utterly undermines this expectation.

Veale and Hao [16] note that irony is pervasive in the sentiment-rich texts of the web. In their analysis of similes harvested from the web, they report that a surprising 18% of unique simile types (such as "*as private as a park-bench*") are ironical. While some are formulaic [9], such as "*as crazy as a fox*", they find that most are ad-hoc, creative and laden with negative sentiment disguised in superficially uncritical terms. Veale *et al*. [17] perform a comparable analysis of similes in Chinese and report the incidence of irony in Chinese similes to be 3% to 4%, indicating that irony is not just a linguistic phenomenon, but a cultural one too.

Irony is a problem for human and computers alike because, as noted by Sperber and Wilson [13], "*the linguistic form of an utterance grossly underdetermines its interpretation*". Both commonsense and world knowledge are required to separate the overt content of irony from its true meaning. In many cases, knowledge of cultural and social stereotypes is also required, even if this knowledge is technically incorrect. One needs this knowledge to know that e.g., "*as tanned as an Irishman*" means "*very pale*" and that "*as sober as a Kennedy*" has a meaning much closer to "*dissolutely drunk*" than "*sober*". But irony is not quite as misleading as outright lying, and speakers do not craft ironies that cannot be understood. Though speakers do not explicitly mark their use of irony – to do so would defeat the purpose of irony [6] - they do often signal the use of irony through intonation and choice of linguistic construction (e.g., "*a fine X, indeed*!" is unlikely to describe something "fine"). Ultimately, however, speakers signal irony by using descriptions that are conceptually unsuited to their targets, and this unsuitability often manifests itself as a violation of category-membership norms.

The problem of irony thus cuts through every aspect of language, from pronunciation to lexical choice, syntactic structure, semantics and conceptualization. As such, it is unrealistic to seek a computational silver bullet for irony, and a general solution will not be found in any one technique or algorithm. Rather, we must try to identify specific aspects and forms of irony that are susceptible to computational analysis, and from these individual treatments attempt to synthesize a gradually broader solution. In this paper we concentrate on one common form of ironic description – the humorous simile – and develop a multi-pronged approach to separating ironic from non-ironic instances of similes. In section 2 we consider the relevance of irony to the computational problem of sentiment analysis, and briefly survey past computational work on irony. In section 3 we collect and analyze a very large corpus of ironic similes from the web. This corpus analysis not only demonstrates the extent to which irony is used on the web, but also allows us to test a key hypothesis as to how speakers signal their use of irony. In section 4 we then exploit this hypothesis in a two-pronged approach to irony detection, one that combines the use of subtle markers with a corpus-based model of conceptual category membership. A large-scale empirical evaluation of this approach – on real ironies gathered from the texts of the web – is then described in section 5, before we conclude with some closing arguments in section 6.

2 IRONY AND SENTIMENT ANALYSIS

Irony presents a significant double-pronged challenge to the automatic classification of sentiment in texts. On one hand, irony cleverly disguises the true attitudes of a speaker, hiding them behind an utterance which superficially promises a radically different meaning. To misclassify an irony is not just to underestimate its sentiment, it is to completely misunderstand the speaker's intent. Yet, because ironic statements typically convey a speaker's most critical viewpoints, they are precisely the statements that sentiment classifiers should attempt to seek out and analyze. As Sperber and Wilson [13] note, "*irony* ...

[1] School of Computer Science and Informatics, University College Dublin, Ireland, email: {tony.veale, yanfen.hao}@ucd.ie

crucially involves the evocation of an attitude – that of the speaker to the proposition mentioned". If a key aim of sentiment analysis is to separate purely propositional content from a speaker's attitude to this content [2], then irony is simply too rich a vein of attitude for the analysis to ignore. However, while irony is widely recognized as a problem of sentiment recognition, it still lies beyond the capabilities of current sentiment analysis approaches [11].

At its simplest, sentiment analysis considers the sentiment of a text to be a function of the sentiment levels of its component words and phrases, and these can be determined either by direct assignment from an resource like Whissell's [19] *dictionary of affect*, or via corpus-derived weights that reflect a degree of association to anchor words of strong sentiment like "excellent" and "poor" [15]. But there is great danger in treating a text as a simple bag of words, since perceived sentiment is often subject to the workings of *valence shifters* [7], special words that can reduce, enhance or even invert the sentiment levels of other words that fall within their scope. For instance, words like "not", "never", "reject" and "avoid" may not contribute much sentiment in themselves, yet they cause the sentiment of the phrases they govern to be reversed. Since irony is commonly seen as a form of indirect negation [5], its presence in a text has the same effect as a negation marker, but one that is implied rather than explicitly stated. But if we view irony an implicit valence shifter, its workings are more nuanced than that of explicit negation. While words like "not" always invert the sentiment of the phrases they negate, whether positive or negative, irony prefers to invert positive meanings to obtain a critical meaning with a negative sentiment. Unlike simple negation then, irony typically works as a valence *downshifter*, turning positive sentiments into criticisms while leaving most negative sentiments untouched. This is an important characteristic of irony, and one that we shall empirically demonstrate in next section, since it allows sentiment analysis to focus its concerns about irony on overtly positive descriptions.

In textual irony, readers are usually expected to recognize that a statement is ironic because it violates expectations raised by the surrounding text; in other words, the literal content of the ironic statement does not cohere with the rest of the text. It follows that a technique like *Latent Semantic Analysis* (LSA), which has been used to measure coherence in text and more generally for measuring the semantic similarity or distance between different text fragments [4], might offer some traction in the detection of irony. This is the approach pursued by Rubinstein [12], who uses LSA to detect ironic distance between paired fragments of text from the same newspaper headlines (e.g., between "Afghanistan" and "a touristy leisure getaway"). LSA uses a bag-of-words model of each text, so examples were carefully chosen to be free of negation and other valence inverters. Rubinstein's results are mildly encouraging, but do not appear scalable for a number of reasons, not least the small size of the evaluation (7 ironic headlines were used) and the inherent problems of using LSA. Furthermore, irony is not the only phenomenon that exploits semantic distance between what is said and what is meant; jokes and highly creative metaphors and similes also create a semantic tension via incongruity, so distance alone cannot separate irony from non-irony. Finally, Rubinstein disputes the signal characteristic that makes irony easier to recognize – its strong preference for expressing critical views in uncritical terms [5].

This last point is much too important to remain a simple matter of opinion. In the next section we harvest a very large corpus of creative similes from the texts of the web, to empirically determine the preferred affective signature of ironic comparisons. It has been observed [8] that the prefix "about" often signals the use of irony in similes, such as in this example from Raymond Chandler: "*He looked about as inconspicuous as a tarantula on a slice of angel food*". In fact, Moon [8] goes as far as to say that "about" always signals the use of irony in similes, a claim we dispute with our analysis here, but we do agree that "about" is a useful signal of ironic intent, and show in section 4 how it can be exploited in the detection of ironies.

3 A CORPUS OF IRONIC SIMILES

We consider in this paper ironic similes with explicit grounds, of the form "as GROUND as a VEHICLE". Using the wildcarded query "as * as a *" on a search engine like Google reveals that the internet is awash with instances of this basic simile pattern, such as "*as strong as an ox*", "*as cool as a cucumber*" and "*as dead as a doornail*". Note that our query pattern has no wildcard for the topic of the simile - the entity that is actually described – since the topic is often undetermined in real texts (e.g., given by a pronoun).

The irony of a textual description can manifest itself at different levels of a text. In text-*external* irony, a description is ironic with respect to how the text as a whole is situated in the outside world; for instance, a no-smoking sign in the foyer of a tobacco company, or an advert for beefsteak in a vegetarian newsletter. In contrast, text-*internal* ironies can be recognized wholly within the text they appear in, as when a description is ironic relative to other information imparted in the same text (e.g., consider Rubinstein's [12] ironic headlines from section 2). Even in text-internal cases, an irony can be self-contained within a given description or apparent only in relation to other descriptions in a text. For instance, if we describe someone in a text as "a hero" shortly after describing this person's craven and cowardly behaviour, then this characterization is likely an example of description-*external* (but text-*internal*) irony. The kind we concern ourselves with here is description-internal irony, in which a description is ironic with respect to other information that the description itself conveys. This is precisely the kind of irony we find in explicit similes such as the following corpus examples: "*as useful as a chocolate teapot*", "*as tough as a marshmallow cardigan*" and "*as welcome as a root-canal without anesthesia*". Rubinstein's headlines also fall into the category of description-internal ironies, and we note that the notion of *internal* refers only to the structure of a text and its use. Description-internal ironies often rely for their resolution on common-sense knowledge of the world outside the text, but the key point here is that the distinct textual elements that exhibit ironic distance from each other are nevertheless found inside the same description.

Similes with explicit grounds exhibit description-internal irony when the vehicle exemplifies a property that is ironically opposed to the stated ground, as in another of our attested corpus examples: "*as modern as a top-hatted chimneysweep*". Similes like this are very common indeed, and we can compile a large corpus by trawling the internet for matches to the query pattern "*about as * as **". Recall that Moon [8] predicts that the "about" prefix always signals the use of irony in similes, and a large corpus constructed around this query will allow us to put this prediction to the test. We use the Google API as our interface to the texts of the web, and overcome Google's limitations on the number of hits/snippets returned per query by generating a different query for each ground property. Thus, to find similes that accentuate strength, or lack thereof, we generate the queries "*about as a strong as **" and "*about as weak as **". To fully automate the harvesting process, we use WordNet [3] as a source of adjectival grounds, and focus on antonymous adjective pairs such as "strong" and "weak", since such pairs typically define property

scales on which similes mark out extreme points. In all, we generate over 2000 queries and request 200 snippets for each from Google.

This harvesting process yields 45,021 instances of the pattern "*about as * as **". Most (85%) are found just once by the harvester, suggesting that "about" comparisons are usually bespoke one-offs. Of course, not every comparison is a simile. As Ortony [10] points out, many comparisons serve a simple correlative function, as in "*a wart about as a big as a plum*", while a simile is a special form of comparison in which the vehicle is used to exemplify a property for which it is highly representative (or strongly opposed, in the case of irony). Thus, "*as brave a knight*" is a simile but "*about as brave as my science teacher*" is likely just a comparison. Because many examples are not self-contained, with pronouns and other strong links to surrounding text, separation of comparisons from similes must be done by hand, and yields a corpus of 20,299 distinct simile types.

These 20,299 "about" similes employ vehicles with a mean length of three words, excluding initial determiner, such as "*[as intelligible as] a gorilla directing traffic*". A substantial number – 30% – use vehicles that are compositions of two or more noun-concepts connected by a preposition, such as "*[as soothing as] a cat in a blender*", while 12% of similes make reference to a topical proper-named entity such as *Karl Rove*, *Paris Hilton* or *Michael Moore*. However, the most interesting statistic concerns the prevalence of irony. Annotating each simile by hand, we find that 15,502 simile-types (76%) are ironic while just 4797 simile types (24%) are non-ironic. This remarkable imbalance generally supports the predication that "about" often signals the use of irony, but the 24% of similes that are non-ironic refutes Moon's [8] strong claim that "about" always signals irony. Neither vehicle length, syntactic complexity or the use of proper-named entities offer any statistically-significant predictors of whether a simile is ironic.

Whissell's (1989) *dictionary of affect* is an inventory of over 8000 English words with pleasantness scores that are statistically derived from human ratings. These scores range from 1.0 (most unpleasant) to 3.0 (most pleasant), with a mean score of 1.84 and a standard deviation of 0.44. To determine whether ironic similes possess a clearly-defined affective signature, we use Whissell's dictionary to automatically classify each "about" simile into one of three categories: those with clearly positive grounds (such as "beautiful", "brave", etc.); those with clearly negative grounds (such as "ugly", "dumb", etc.); and those with grounds that cannot easily be classified according to Whissell's resource. A ground is considered negative if it possesses a pleasantness score less than one standard deviation below the mean (≤ 1.36), and positive if it has a pleasantness score greater than one standard deviation above the mean (≥ 2.28). A breakdown of similes that match these criteria is shown in Table 1. From Table 1 is it clear that ironic similes have a strong preference for disguising negative sentiments in positive terms, while only a small minority of "about" similes (8%) attempt to convey a positive message in an ironically negative guise.

Table 1. "about" similes categorized by irony and affect. Total is 100%

	Straight	Ironic
Positive Ground	9%	71%
Negative Ground	12%	8%

4 COMPUTATIONAL IRONY DETECTION

Given these empirical findings, a sentiment analyzer presented with a simile of the form "about as GROUND as a VEHICLE" can generate a strong initial hypothesis before it even considers the linguistic makeup of the vehicle. If the ground is clearly positive, then the description has a significant chance ($> 70\%$) of being ironic and negative. But what of descriptive similes that are not helpfully prefixed by the "about" marker, or the many more similes for which the ground is not an obviously positive or negative word?

The corpus analysis of section 3 reveals both a high frequency of hapaxes and a preference for long vehicles in "about" marked similes, which suggests that this marker is typically used to minimize the risk of misinterpretation for newly coined creative similes. But there is a substantial grey area between these creative one-offs and the stock similes that one finds in printed resources [9]. This grey area is populated with a large number of similes that have enough linguistic currency to be somewhat familiar, but not enough to be seen as formulaic and deserving of explicit representation in the lexicon. Given an ironic but untelegraphed simile of the form "as GROUND as a VEHICLE", it is likely that the variation "*about* as GROUND as a VEHICLE" has already been used by another speaker in another time and place, when the simile was riskier and less familiar. A mechanism for irony detection can exploit the web as a source of past uses of a simile. Using the web (and a search API like that offered by Google), we can distinguish between three kinds of similes: those that have never been used with "about" on the web; those that have, but not predominantly so; and those appear with "about" more frequently than they appear without this marker. These three categories provide, respectively, weak evidence *against* irony, weak evidence *for* irony, and strong evidence for irony.

But in the final analysis, helpful markers like "about" are no more than heuristic clues that direct an audience to look for irony; these clues do not contain the substance of the irony, nor do they obviate the need to understand the irony in conceptual terms. A mechanism for irony detection must thus grapple with thorny issues of categorization – category structures, norms, boundaries and membership criteria – if detection is to be based on appreciation as well as informed guesswork. Generally speaking, a simile of the form "as GROUND as a VEHICLE" claims that a topic is as much a member of the conceptual category "*things that are* GROUND" as the landmark concept VEHICLE. If the stated vehicle is a clear member of this ad-hoc category, then the simile is a straight description; but if the vehicle strongly resists categorization in this way, it is likely ironic. In the simile "as subtle as a freight-train", we see that freight-trains are very difficult to conceptualize as "things that are subtle" and so we consider the description to be ironic. The category "things that are subtle" is an ad-hoc category in the sense of Barsalou [1], inasmuch as is task-specific and cuts across conventional taxonomic boundaries. While a resource like WordNet provides useful knowledge for irony detection via its synonymy, antonymy and hyponymy relations, it lacks these cross-cutting structures.

Fortunately, languages like English provide a number of constructions to easily specify ad-hoc sets, and these constructions can be mined from a corpus, or from the web, to learn precisely the ad-hoc categories that are needed to appreciate irony. For instance, the construction "*hot environments such as saunas, kitchens and locker-rooms*" specifies a partial extension for the ad-hoc category "environments that are hot", and this category in turn allows us to recognize that the simile "as hot as a sauna" is not ironic. Veale, Li and Hao [18] show how a large knowledge-base of fine-grained categoriza-

tions like these can be bootstrapped from an initial seed set of adjective:noun pairs, and further show how seeds of different size can be acquired from resources like WordNet. But these ad-hoc categories can be also learned on the fly by an irony detection mechanism, by selectively mining the web for constructions that involve a specific ground property. Thus, given the simile of the form "as GROUND as a VEHICLE", we look for web patterns of the form "GROUND * such as VFORM", where VFORM is either VEHICLE itself, a synonym of VEHICLE in WordNet, or a hyponym of some sense of VEHICLE in WordNet. The presence of any of these patterns on the web is strong evidence for the belief that GROUND is an apt predicate for VEHICLE and so the simile is *not* ironic.

We should note that novelty is a gradable quality in similes, since some are truly novel whiles others merely offer a variation on an existing theme. As speakers are exposed to more examples of the former, they should become more adept at recognizing variations of the latter kind. Computationally, this requires a form of unsupervised learning, wherein a system notes what it believes to be clear examples of ironic and non-ironic similes, to enable it to determine at some future stage whether an apparently novel simile is simply a variation on one of these noted precedents. We conceive of two kinds of variation, *direct* and *inverse*. The simile "as G-VARIATION as a V-VARIATION" is a direct variation of the simile "as GROUND as a VEHICLE" iff: G-VARIATION = GROUND or G-VARIATION is a synonym or hyponym of (some sense of) GROUND in WordNet, or if the mutual-support pattern "as G-VARIATION and GROUND as *" is found more than once on the web; and V-VARIATION = VEHICLE, or V-VARIATION is a synonym or hyponym of (some sense of) VEHICLE in WordNet. In contrast, the simile "as G-VARIATION as a V-VARIATION" is an inverse variation of "as GROUND as a VEHICLE" iff V-VARIATION is a direct variation of VEHICLE and some sense of G-VARIATION is an antonym of some sense of GROUND in WordNet.

To classify "as GROUND as a VEHICLE", we follow this 9-step sequence:

1. A simile is classified as *non-ironic* if there is lexical/morphological similarity between: i) the vehicle and the ground (e.g., *as manly as a man*); ii) between the vehicle and a synonym of the ground (e.g., *as masculine/manly as a man*); or iii) between the vehicle and an adjective that is a frequently conjoined with the ground as a co-descriptor (e.g., *as cold [and snowy] as snow*).

2. If the web frequency of "*about* as GROUND as a VEHICLE" is more than half that of "as GROUND as a VEHICLE" (i.e., the "about" form is predominant), then the simile is classified as *ironic* and noted as an ironic precedent.

3. If this simile is recognizable as a direct variation of an ironic precedent (see 2 above), then this simile is also classified as *ironic*.

4. If this simile is recognizable as an inverse variation of an ironic precedent (see 2 above), then this simile is inversely classified as *non-ironic*.

5. If the ad-hoc category pattern "GROUND * such as VEHICLE" is found on the web, then the simile is considered *non-ironic* and is noted as a non-ironic precedent.

6. If the simile is a direct variation of a non-ironic precedent, it is deemed *non-ironic*.

7. If the simile is an inverse variation of a non-ironic precedent, it is deemed *ironic*.

8. If the simile has a web-frequency of 10 or more, it is classified as *non-ironic* and is also noted as a non-ironic precedent.

9. If the simile has a web-frequency less than 10, it is classified as *ironic*.

Steps 8 and 9 are catch-alls for those similes that remain unclassified after steps 1 – 7. In (8), we exploit the fact that irony is a mode of creative expression, reasoning that common similes (e.g., those with web-frequencies ≥ 10) are less likely to be creative and thus less likely to be ironic. Finally, in (9), we conversely use low frequency as a crude indicator of creative novelty, and thus, indirectly, of irony. These two steps are crude, to be sure, but they apply late in the process, only after more subtle means have failed. Steps 2, 5 and 8 provide opportunities for the unsupervised learning of precedents: though fallible, these steps are sufficiently accurate to serve as a strong basis for variation detection. This is similar to how humans deal with irony: classification is fallible and misclassification is not always corrected when it occurs [13].

5 EMPIRICAL EVALUATION

Our test set should offer a balanced sample of both highly creative and highly formulaic similes, with the bulk of the test data comprising similes drawn from the grey area between these poles. To construct a convincingly large test set, we return to the web to harvest all similes of the form "*as ADJ as* *" (note the absence of an explicit "about" marker). Once again we use the Google API and analyze the first 200 snippets retrieved for each query, only keeping those matches where the vehicle is lexicalized in WordNet as a single-noun or as a noun compound (i.e., similes with syntactically complex vehicles are overlooked). Once these candidate comparisons are hand-annotated to find ironic and non-ironic similes, the resulting test-set comprises 2787 unique ironic similes and 12,252 non-ironic similes. Because the retrieval query is not explicitly biased toward the harvesting of ironic similes, this roughly 80:20 breakdown is broadly representative of the distribution of ironic similes on the web.

As an initial baseline, consider that a detection system which simply identifies all similes as straight (*non*-ironic) will achieve complete recall on non-ironic similes and no recall at all on ironic similes; by incorrectly classifying all 2787 ironic similes as non-ironic, this baseline system still achieves a F-score of 0.89 for non-ironic similes, a micro-accuracy (the accuracy micro-averaged across all classifications of individual similes) of .81 and a macro-accuracy (the mean of the accuracy for the ironic and non-ironic classes) of .50. Such a system has no irony detection capabilities at all, yet achieves reasonable performance because of the imbalance of irony to non-irony among similes.

Now consider a slightly more complex system that just uses the "about" marker as an indicator of ironic intent. None of the 15,039 similes in our test set are explicitly marked with "about" – because we neither used it in the harvesting query nor sought it when parsing the harvested snippets – but a detection system can easily determine how likely a given simile is to occur with or without "about" by obtaining web-counts for the corresponding strings via the Google API. For instance, the string "*as strong as an ox*" has a web-count of 21,000 occurrences, while "about *as strong as an ox*" has just 3 occurrences; since the "about" form is clearly not dominant for the

pairing of strong and ox, a system may infer that this pairing is non-ironic. In all, while 37% of the ironies in our test set are predominately used with the "about" form, this form is preferred by just 2% of the 12,252 straight similes in the test set. Overall then, the simple addition of the "about" heuristic of step (2) yields an F-score of 51% while detecting 1 in 3 ironies. The "about" heuristic is our most precise cue for recognizing irony directly, but it clearly lacks the scope to achieve significant levels of recall. The other steps in our 9-step algorithm are not so much concerned with detecting irony as in eliminating non-irony.

5.1 Error Analysis

We now consider an error analysis of each of these 9 steps in turn. Since each step concerns itself with just one category of simile, ironic or non-ironic, the precision and recall for that one category is provided for each step in isolation:

Step 1, which exploits morphological similarity between the vehicle and a variant of the ground, is limited but very precise: it correctly identifies 339 similes as non-ironic, and misclassifies just 7 ironies as non-ironic (non-ironic R =.03, P = .98).

Step 2 uses the predominance of the "about" marker to correctly classify 1030 ironies while misclassifying 200 non-ironic similes as ironic. As we saw in table 1 of section 3, the "about" marker is used in about 20% of cases to indicate a creative, and often directly critical, perspective on a vehicle (ironic R = .37, P = .84).

Step 3 recognizes a further 226 ironies as variations on these "about" precedents, at a cost of 315 further misclassifications. Direct variation is clearly a subtle phenomenon that is beyond the abilities of WordNet to detect with high precision (ironic R = .08, P = .42).

Step 4, however, indicates that inverse variations that exploit antonyms are more reliably detected. This step correctly classifiers 312 similes as non-ironic, while misclassifying just 6 ironies as non-ironic (non-ironic R = .025, P = .98).

Step 5 is the most significant step in the identification of non-irony, and uses the ad-hoc categorization pattern "GROUND * such as VEHICLE" to correctly classify 3439 similes as non-ironic. Just 161 ironies are misclassified, indicating that this categorization pattern is very rarely used for ironic purposes in English (non-ironic R = .28, P = .96).

Step 6 seeks variants of similes that were attested as non-ironic via the "such as" pattern in (5). An additional 2900 non-ironic similes are correctly classified, while another 186 ironies are misclassified (non-ironic R = .24, P = .94).

Step 7 seeks inverse variants of the similes classified as non-ironic in (5) and (6), but this kind of variation proves very infrequent: just 35 further ironies are correctly identified in this step, while 35 non-ironies are misclassified (ironic R = .01, P = .50).

Step 8 comes into play when most common ironies have already been detected, and so assumes that any remaining simile with a web-frequency of 10 or more must be non-ironic. This proves to be a safe assumption, for 3856 similes are correctly classified as non-ironic while no ironies at all are misclassified (non-ironic R = .31, P = 1.0)

Step 9 assumes that most non-ironies have already been identified. Remaining similes have a web-frequency of 9 or less, and are classified as ironic, to yield 1136 correct and 856 incorrect classifications (ironic R = .41, P = .57).

In all, these 9 steps identify 87% of the ironies with 63% precision, and 89% of the non-ironies with 97% precision. The F-score for irony detection is thus 73%; for non-irony detection it is 93%. The model achieves a micro-accuracy of .88 and a macro-accuracy of .88 for the recognition of both ironic and non-ironic similes together.

6 DISCUSSION AND CONCLUSIONS

Irony is a most vexing form of communication because – superficially, at least – it uses imagination and ingenuity to artfully disguise the expression of a negative sentiment. Consider this extract from an online discussion of the rules of baseball [14]:

"*[B]aseball's rules structure has remained remarkably steady for more than 100 years. While basketball fiddles with 3-point lines and football puts its pass-interference, overtime and ref-upstairs rules in a Cuisinart each offseason, baseball rules remain as suggestible as a glacier*."

While one can try to analyze the underlined simile (our marking) in isolation, it is clear that the take-home message is consolidated over the entire paragraph. Note how the ground of the simile, "suggestible", contrasts sharply with the property "steady" that is highlighted in the first sentence, and note how the second sentence uses "While" to establish a contrast between baseball and the more changeable games of basketball and football. Moreover, the extreme changeability of football is conveyed metaphorically, via the exaggerated claim that the football rulebook is shredded in a food processor at the end of each season. These are vexing rhetorical strategies in their own right, and offer little in the way of easily exploitable cues for irony detection.

However, while Google identifies just one documentary source for the novel combination "*as suggestible as a glacier*" [14], "glacier" is a commonly used vehicle in similes. For instance, our test set from section 5 contains 20 non-ironic examples, highlighting the properties *cold, cool, strong, fresh, impressive, unstoppable, pure, gradual, slow, slick, relentless, unwieldy, irresistible, frozen, frosty, implacable, impenetrable, unforgiving, forceful* and *implacable*. Glaciers are also used ironically in our test set, to highlight the lack of the following properties: *mobile, erotic, excitable, speedy* and, of course, *suggestible*. Using the web query "*as steady and * as*" to find co-descriptors that are lexically primed by "steady", we find that these properties of "glacier" are primed: *strong, slow, cool, cold, implacable* and *unstoppable*. It follows that when a system has already acquired a rich feature description of a vehicle from similes that were previously encountered and classified as non-ironic, it can choose to ignore the explicit ground in a new simile if it is not lexically primed by its context, and rely instead on those features of the vehicle that are primed. In this case, the features *slow* (to change) and *implacable* (in the face of change) are most appropriate to the topic of baseball rules. In effect, an informed system that learns from the similes that it identifies as non-ironic can sometimes correctly interpret an ironic simile without having to first recognize it as ironic.

Finally, we note that despite our experimental findings in section 3, which suggest that ironic similes typically use a ground with positive sentiment to impart a negative view of a vehicle, none of the steps in our 9 step algorithm use the relative sentiment of the ground and vehicle to detect irony. Unfortunately, our experiments with sentiment

resources like Whissell's dictionary of affect yielded no appreciable results that were worthy of inclusion here. Of course, we expect that sentiment analysis *should* play a significant role in the computational detection of irony, but only when positive and negative sentiment can be accurately assessed in the context of the author's specific goals and beliefs. For instance, in the baseball example above, it is clear that the author is criticizing the rules of baseball for not being "suggestible" enough, but in many other contexts "suggestible" has a negative connotation, suggesting laziness, weak-mindedness and uncertainty. So for now at least, it makes sense to continue to explore the related problems of irony detection and sentiment analysis in parallel, until such time that a partial solution to one can make an appreciable contribution to the solution of the other.

REFERENCES

[1] L.W. Barsalou, Ad hoc categories, *Memory and Cognition*, **11**, 211–227, (1983).

[2] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou and D. Jurafsky, Automatic extraction of opinion propositions and their holders, In *Proc. of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

[3] C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[4] P.W. Foltz, W. Kintsch and T.K. Landauer, The measurement of textual coherence with Latent Semantic Analysis, *Discourse Processes*, **25**, 285–307, (1998).

[5] R. Giora, On Irony and Negation, *Discourse Processes*, **19**, 239–264, (1995).

[6] H.P. Grice, Logic and conversation, In P. Cole and J.L. Morgan (Eds.), *Syntax and semantics: Vol. 9. Pragmatics*. New York: Academic, 1978.

[7] A. Kennedy and D. Inkpen, Sentiment classification of movie reviews using contextual valence shifters, *Computational Intelligence*, **22**(2), 110–125, (2006).

[8] R. Moon, Conventionalized as-similes in English: A problem case, *International Journal of Corpus Linguistics*, **13**(1), 3–37, (2008).

[9] N. Norrick, Stock Similes, *Journal of Literary Semantics*, **XV**(1), 39–52, (1986).

[10] A. Ortony, The role of similarity in similes and metaphors, in *Metaphor and Thought*, edited by A. Ortony, Cambridge, MA: Cambridge University Press, 1979.

[11] B. Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135, (2008).

[12] A. Rubinstein, Using LSA to detect Irony, In *Proc. of the Workshop on Corpus-Based Approaches to Figurative Language, at Corpus Linguistics*, Lancaster, UK, (2003).

[13] D. Sperber and D. Wilson, On verbal irony, *Lingua*, **87**, 53–76, (1992).

[14] A. Schwarz, The history of rule changes, *ESPN.com, February 4th*, 2003.

[15] P. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In *proc. of the 40^{th} Ann. Meeting of the Assoc. of Computational Linguistics*, 2002.

[16] T. Veale and Y. Hao, Making Lexical Ontologies Functional and Context-Sensitive, In *proc. of the 46^{th} Ann. Meeting of the Assoc. of Computational Linguistics*, 2007.

[17] T. Veale, Y. Hao and G. Li, Multilingual Harvesting of Cross-Cultural Stereotypes, *In proc. of the 46^{th} Ann. Meeting of the Assoc. of Computational Linguistics*, 2008.

[18] T. Veale, G. Li and Y. Hao, Growing Fine-Grained Taxonomies from Seeds of Varying Quality and Size, In *proc. of the Annual Meeting of the European Assoc. of Computational Linguistics*, 2009.

[19] C. Whissell, The dictionary of affect in language, In R. Plutchnik & H. Kellerman (Eds.) *Emotion: Theory and research*, New York: Harcourt Brace, 113–131, (1989).