Decision Support Systems 53 (2012) 754–760

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

# Making objective decisions from subjective data: Detecting irony in customer reviews

Antonio Reyes *, Paolo Rosso

*Natural Language Engineering Lab, ELiRF, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Camino de Vera, s/n 46022, Valencia, Spain*

ARTICLE INFO

Available online 23 May 2012

*Keywords:*
Irony detection
Natural language processing
Web text analysis

ABSTRACT

The research described in this work focuses on identifying key components for the task of irony detection. By means of analyzing a set of customer reviews, which are considered ironic both in social and mass media, we try to find hints about how to deal with this task from a computational point of view. Our objective is to gather a set of discriminating elements to represent irony, in particular, the kind of irony expressed in such reviews. To this end, we built a freely available data set with ironic reviews collected from Amazon. Such reviews were posted on the basis of an online viral effect; i.e. contents that trigger a chain reaction in people. The findings were assessed employing three classifiers. Initial results are largely positive, and provide valuable insights into the subjective issues of language facing tasks such as sentiment analysis, opinion mining and decision making.



## 1. Introduction

Verbal communication is not a trivial process. It implies sharing a common code as well as being able to infer information beyond the semantic meaning. A lot of communicative acts imply information not grammatically expressed to be able to decode the whole sense: if the hearer is not capable inferring that information, the communicative process is incomplete. Let us consider a joke. The amusing effect sometimes relies on not given information. If such information is not filled, the result is a bad, or better said, a misunderstood joke. This information, which is not expressed with "physical" words, supposes a great challenge, even from a linguistic analysis, because it points to social and cognitive layers quite difficult to be computationally represented. One of the communicative phenomena which better represents this problem is irony. According to Wilson and Sperber [34], irony is essentially a communicative act which expresses an opposite meaning of what was literally said.

Because irony is common in texts that express subjective and deeply-felt opinions, its presence represents a significant obstacle to the accurate analysis of sentiment in such texts (cf. Councill et al. [10]), in particular, when its presence may represent valuable information to make the best decision. For instance, consider the goodness of a product or the quality of a service (restaurant, hotel, etc.). In this context, this research work aims at gathering a set of discriminating elements to represent irony. We especially focus on analyzing a set of customer reviews (posted on the basis of an online viral effect) in order to obtain a set of key components to face the task of irony detection. Such reviews have been taken as ironic by people, both in social and mass media (Youtube, Wikipedia, BBC, ABC). Our objective thus consists of defining a feature model in order to represent part of the subjective knowledge which underlies such reviews. In this respect, the relevance of this work lies in the fact that such model might imply direct and indirect knowledge in tasks as diverse as sentiment analysis (cf. [24] about the importance of determining the presence of irony in order to set a fine-grained polarity), opinion mining (cf. [28], where the authors note the role of irony for minimizing the error when discriminating negative from positive opinions), or even advertising (cf. [19], about the function of irony to increase message effectiveness).

This paper is organized as follows. Section 2 introduces the theoretical problem of irony. Section 3 presents the related work as well as the evaluation corpus. Section 4 describes our model and the experiments that were performed. Section 5 evaluates the model and presents the discussion of the results. Section 6 re-evaluates the model on a corpus of news articles. Finally, Section 7 draws some conclusions and addresses the future work.

## 2. Pragmatic theories of irony

Literature divides two primary classes of irony: verbal and situational. Most theories agree on the main property of the former: verbal irony conveys an opposite meaning; i.e. a speaker says something that seems to be the opposite of what s/he means [9]. In contrast, situational irony is a state of the world which is perceived as ironic [2]; i.e. situations that should not be [21]. Our work focuses on verbal irony. This kind of irony is defined as a way of intentionally denying what is literally expressed [11]; i.e. a kind of indirect negation [15]. On the basis of some pragmatic frameworks, authors focus on certain fine-grained aspects of this term. For instance, Grice [16] considers that an utterance is ironic if it intentionally violates some conversational maxims. Wilson and Sperber [34] assume that verbal irony

* Corresponding author at: Esfuerzo 72, Depto. 4501, C.P. 07530, Mexico, D.F., Mexico. Tel.: +52 5559121923.
*E-mail addresses:* areyes@dsic.upv.es (A. Reyes), prosso@dsic.upv.es (P. Rosso).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.05.027

must be understood as echoic; i.e. as a distinction between use and mention. Utsumi [31], in contrast, suggests an ironic environment which causes a negative emotional attitude. According to these points of view, the elements to conceive a verbal expression as ironic point to different ways of explaining the same underlying concept of opposition, but specially note, however, that most of them rely on literary studies [2]; thus, their computational formalization is quite challenging. Furthermore, consider that people have their own concept of irony, which often does not match the rules suggested by the experts. For instance, consider the following expressions retrieved from the web:

1. "It's not that there isn't anything positive to say about the film. There is. After 92 minutes, it ends".
2. "Difference between a virus and Windows? Viruses rarely fail".
3. "The room at the hotel was clean and quiet. Pity that it cost only 200 Euros per night."

These examples, according to some user-generated tags, could be either ironic, or sarcastic, or even satiric. However, the issue we want to focus on does not lie in what tag should be the right one for every expression, but in the fact that there is not a clear distinction about the boundaries among these terms. Where does irony end, and where does sarcasm (or satire) begin? For Colston [8], sarcasm is a term commonly used to describe an expression of verbal irony; whereas for Gibbs [13], sarcasm along with jocularity, hyperbole, rhetorical questions, and understatement, are types of irony. Attardo [2] in turn, considers that sarcasm is an overtly aggressive type of irony. Furthermore, according to Gibbs and Colston [14], irony is often compared to satire and parody.

In accordance with these statements, the limits among these figurative devices are not clearly differentiable. Their differences rely indeed on matters of usage, tone, and obviousness, which are not so evident in ordinary communication acts. Therefore, if there are no formal boundaries to separate these concepts, even from a theoretical perspective, people will not be able to produce ironic expressions as the experts suggest. Instead, there will be a mixture of expressions intending to be ironic but being sarcastic, satiric, or even humorous. This gets worse when dealing with non prototypical examples. Observe the following fragment from our corpus:

4. "I am giving this product [a t-shirt] 5 stars because not everyone out there is a ladies' man. In the hands of lesser beings, it can help you find love. In the hands of a playa like me, it can only break hearts. That's why I say use with caution. I am passing the torch on to you, be careful out there, folks."

In this text irony is perceived as a mixture of sarcasm and satire, whose effect is not only based on expressing an opposite or negative meaning, but a humorous one as well. Taking into account these assumptions, we begin by defining irony as a *verbal subjective expression whose formal constituents attempt to communicate an underlying meaning, focusing on negative or humorous aspects, which is opposite to the one expressed*. On the basis of this definition, we consider sarcasm, satire, and others forms of figurative language, such as the ones suggested in [13] (jocularity, hyperbole, rhetorical questions, and understatement), as specific extensions of a general concept of irony.

## 3. Approaching irony detection

Ironic statements can be found in almost every web site; they impose a big challenge since they come along with unique characteristics compared to other text types. If ironic texts were discriminated accurately, they would be of great value for different tasks (cf. Section 1).

On this subject, as far as we know, very few attempts have been carried out in order to integrate irony in a computational framework. The research described by Utsumi [31] was one of the first approaches to computationally formalize irony. However, his model is too abstract to represent irony beyond an idealized hearer–listener interaction. Recently, from a computational creativity perspective, Veale and Hao [32] focused on studying irony by analyzing humorous similes. Their approach gives some hints to explain the cognitive processes that underlie irony in such structures. In contrast, Carvalho et al. [7] suggested some clues for automatically identifying ironic sentences by means of identifying features such as emoticons, onomatopoeic expressions, punctuation and quotation marks. Furthermore, there are other approaches which are focused on particular devices such as sarcasm and satire, rather than on the whole concept of irony. For instance, Tsur et al. [30], as well as Davidov et al. [12], address the problem of finding linguistic elements that mark the use of sarcasm in online product reviews and tweets, respectively. Finally, Burfoot and Baldwin [6] explore the task of automatic satire detection by evaluating features related to headline elements, offensive language and slang.

Although these approaches have shown that irony, as well as linguistic devices related to it, can be handled in terms of computational means, it is necessary to improve the representation of its characteristics, and especially, to create a feature model capable of symbolizing, in the least abstract way possible, both linguistic and social knowledge in order to describe deeper properties of irony. Therefore, our objective is to identify some salient components of irony[1] by means of formal linguistic arguments; i.e. words and sequences of them, in order to gather a set of discriminating items to automatically differentiate an ironic review from a non-ironic one. To this end, we have defined six categories of features which attempt to represent irony from different linguistic layers. They are assessed on the basis of the examples found in our corpus, which is described in the following section.

### 3.1. Evaluation corpus

Due to the scarce work on automatic irony processing, and on the intrinsic features of irony, it is quite difficult and subjective to obtain a corpus with ironic data. Therefore, we decided to rely on the wisdom of the crowd and use a collection of customer reviews from the Amazon web site. These reviews are considered ironic by customers, as well as by many journalists, both in mass and social media. According to such means, all these reviews deal with irony, sarcasm and satire (hence, they are consistent with our definition of irony). All of them were posted by means of an online viral effect, which in most cases, increased the popularity and sales of the reviewed products. The *Three Wolf Moon T-shirt* is the clearest example. This item became one of the most popular products, both in Amazon as in social networks, due to the ironic reviews posted by people.[2] For instance, consider the effect caused by this t-shirt in the following web sites: Youtube,[3] Wikipedia,[4] BBC,[5] or ABC.[6] This viral effect shows the power of irony and the need to automatically detect it.

The importance of Amazon in electronic commerce is well known. However, this importance is not supported by only its business schema, but also by trusting in the opinions posted by its customers. Those opinions impact, either positively or negatively, on other customers interested in the products offered by Amazon. The fact of considering such opinions in order to mine deeper information and to be able to detect irony, could be capitalized on for labeling opinions beyond a

[1] In the terms by which we defined it at the end of Section 2.

[2] According to results obtained with Google, apart from the more than one million results retrieved when searching this product, there are more than 10,000 blogs which comment on the effect caused by these reviews.

[3] http://www.youtube.com/watch?v=QPB45AUmchM.

[4] http://en.wikipedia.org/wiki/Three_Wolf_Moon.

[5] http://news.bbc.co.uk/2/hi/8061031.stm.

[6] http://abcnews.go.com/WN/story?id=7690387&page=1.

positive or negative polarity, and for making a fine-grained analysis to allow, for instance, better decision making.[7]

In this context, our positive data are thus integrated with reviews of five different products published by Amazon. All of them were posted through the online viral effect. The list of products includes:

- Three Wolf Moon T-shirt. Amazon product id: B002HJ377A
- Tuscan Whole Milk. Amazon product id: B00032G1S0
- Zubaz Pants. Amazon product id: B000WVXM0W
- Uranium Ore. Amazon product id: B000796XXM
- Platinum Radiant Cut 3-Stone. Amazon product id: B001G603AE

A total of 3163 reviews were retrieved. Then, in order to automatically filter the ones more likely to be ironic without performing a manual annotation, we removed the reviews whose customer rating, according to the Amazon rating criteria, was less than four stars. The assumptions behind this decision rely on two facts: i) the viral purpose, and ii) the ironic effect. The former causes that people to post reviews whose main purpose, and perhaps the only one, was to exalt superficial properties and non-existent effects; thus the possibilities of finding *real* reviews were minimal. Considering this scenario, the latter supposes that, if someone ironically wants to reflect properties and effects such as the previous ones, s/he will not do it by rating the products with one or two stars, instead, s/he will rate them with the highest scores. After applying this filter, we obtained an ironic set integrated with 2861 documents. On the other hand, three negative sets were automatically collected from the following sites: Amazon.com, Slashdot.com, and TripAdvisor.com. Each contains 3000 documents. The products selected from Amazon (AMA) were[8]: Bananagrams (toy), The Help by Kathryn Stockett (book), Flip UltraHD Camcorder (camera), I Dreamed a Dream (CD), Wii Fit Plus with Balance Board (videogame console). The subset collected from Slashdot (SLA) contains web comments categorized as funny in a community-driven process. Finally, the last subset was taken from the TripAdvisor (**TRI**) data set [3], which contains opinions about hotels. The whole evaluation corpus is integrated with 11,861 documents. It is available at: http://users.dsic.upv.es/grupos/nle/.

## 4. Model

According to the arguments given in Section 3, we consider that the task of defining irony features in terms of linguistic elements seems to be the most viable approach. Nonetheless, some fine-grained theoretical concepts, such as the ones described in Section 2, cannot be directly mapped to our framework due to the idealized communicative scenarios which they suppose, and that do not completely match the ones found in our data. Hence, our approach focuses on obtaining the underlying core from those concepts in order to represent it in our model. By mapping this core through words, we expect to be able to represent some profiled characteristics of irony. To this end, we defined the following six features: *n-grams, POS n-grams, funny profiling, positive/negative profiling, affective profiling*, and *pleasantness profiling*.

The first one attempts to find frequent sequences of words considering n-grams of different orders. The second one tries to find morphosyntactic templates given the part of speech (POS) tags. The third feature evaluates a selection of the best-performing humor features found in the literature. The fourth one assesses, from a sentiment analysis point of view, the polarity profiled. The fifth one represents attitudes, emotions, moods, etc., by means of analyzing affective elements in the reviews. The last one measures the degree of pleasantness produced by every review.

[7] See Kim et al. [18] and [17] about the role of trust on decision making.
[8] Being one of the top best sellers was the only criterion to select them.

### 4.1. N-grams

This feature focuses on representing the ironic documents in the simplest way: with sequences of n-grams (from order 2 up to 7) in order to find a set of recurrent words which might express irony. Note that all the documents were preprocessed. Firstly, the stop words were removed, and then, all the documents were stemmed. The next processing consisted of removing irrelevant terms by applying a *tf–idf* measure. The measure is calculated according to Formula 1:

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot log = \frac{|D|}{\left|\left\{d_j \middle| t_j \in d_j\right\}\right|} \quad (1)$$

where $|D|$ is the number of documents in $D$, and $|\{d_j|t_j \in d_j\}|$ is the number of documents in $D$ containing $t_i$. This measure assesses how relevant a word is, given its frequency both in a document as in the entire corpus. Irrelevant words such as *t-shirt*, *wolf*, *Tuscan*, *milk*, etc., were then automatically eliminated. The complete list of filtered words, stop words included, contains 824 items. Examples of the most frequent sequences are given in Table 1.

### 4.2. POS n-grams

The goal of this feature is to obtain recurrent sequences of morphosyntactic patterns. According to our definition, irony looks for expressing an opposite meaning; however, the ways of transmitting that meaning are enormous. Therefore, we intend to symbolize an abstract structure through sequences of POS tags (hereafter, POS-grams) instead of only words. It is worth highlighting that a statistical substring reduction algorithm [20] was employed in order to eliminate redundant sequences. For instance, if the sequences "he is going to look so hot in this shirt" and "he is going to look hot in this shirt" occur with similar frequencies in the corpus, then, the algorithm removes the last one because it is a substring of the first one. Later on, we labeled the documents employing the FreeLing resource [1]. The N-best sequences of POS-grams, according to orders 2 up to 7, are given in Table 2.

### 4.3. Funny profiling

Irony takes advantage of funny aspects to produce its effect. This feature intends to characterize the documents in terms of humorous properties. In order to represent this feature, we selected some of the best humor features reported in the literature: *stylistic features*, *human centeredness*, and *keyness*. The stylistic features, according to the experiments reported in [22], were obtained by collecting all the words labeled with the tag "sexuality" in WordNet Domains [4]. The second feature focuses on social relationships. In order to retrieve these words, the elements registered in WordNet [23], which belong to the synsets *relation*, *relationship* and *relative*, were retrieved. The last feature is represented by obtaining the keyness value of the words (cf. [26]). The words considered are supposed to have a sufficiently high keyness value to be representative of the ironic documents. This value is calculated by comparing the word frequencies in the ironic documents against their frequencies in a reference corpus. Google N-grams [5] was used as the reference corpus. The process consisted of building

**Table 1**
Statistics of the most frequent word n-grams.

| Order | Sequences | Examples |
|---|---|---|
| 2-grams | 160 | opposit sex; american flag; alpha male |
| 3-grams | 82 | sex sex sex; fun educ game |
| 4-grams | 78 | fun hit reload page; remov danger reef pirat |
| 5-grams | 76 | later minut custom contribut product |
| 6-grams | 72 | fals function player sex sex sex |
| 7-grams | 69 | remov danger reef pirat fewer shipwreck surviv |

**Table 2**
Statistics of the most frequent POS-grams.

| Order | Sequences | Examples |
|---|---|---|
| 2-grams | 300 | dt nn; nn in; jj nn; nn nn |
| 3-grams | 298 | dt nn in; dt jj nn; jj nn nn |
| 4-grams | 282 | nn in dt nn; vb dt jj nn |
| 5-grams | 159 | vbd dt vbg nn jj |
| 6-grams | 39 | nnp vbd dt vbg nn jj |
| 7-grams | 65 | nns vbd dt vbg nn jj fd |

two word lists, one for the ironic documents, and one for the reference corpus, then we compared both data applying the Log likelihood ratio. Only the words whose keyness was ≥100 were kept.

### 4.4. Positive/negative profiling

As we have already pointed out, one of the most important properties of irony relies on the communication of negative information through positive one. This feature intends to be an indicator about the correlation between positive and negative elements in the data. The Macquarie Semantic Orientation Lexicon (MSOL) [27] was used to label the data. This lexicon contains 76,400 entries (30,458 positive and 45,942 negative ones).

### 4.5. Affective Profiling

In order to enhance the quality of the information related to the expression of irony, we decided to represent information linked to psychological layers. The affective profiling feature is an attempt to characterize the documents in terms of words which symbolize subjective contents such as emotions, feelings, moods, etc. The WordNet-Affect resource [29] was employed for obtaining the affective terms. This resource contains 11 classes to represent affective content (attitude, behavior, cognitive state, edonic signal, emotion, mood, physical state, emotional response, sensation, emotion-eliciting situation, and trait). According to the authors, these classes represent how speakers convey affective meanings by means of selecting certain words and not others.

### 4.6. Pleasantness profiling

The last feature is an attempt to represent ideal cognitive scenarios to express irony. This means that, like words, the contexts in which irony appears are enormous. Therefore, since it is impossible to make out all the possibilities, we intend to define a schema to represent favorable and unfavorable ironic contexts on the basis of pleasantness values. In order to represent those values, we used the Dictionary of Affect in Language [33]. This dictionary assigns a score of pleasantness to ~9000 English words. The scores were obtained from human ratings. The range of scores goes from 1 (unpleasant) to 3 (pleasant).

## 5. Evaluation

In order to verify the effectiveness of our model, we evaluated it through a classification task. Two underlying goals were analyzed: a) feature relevance; and b) the possibility of automatically finding ironic documents.

The classifiers were evaluated by comparing the positive set against each of the three negative subsets (AMA, SLA and TRI, respectively).[9] All the documents were represented as frequency-weighted term vectors according to a representativeness ratio. This ratio was estimated using Formula 2:

$$\delta(d_k) = \frac{\sum_{i,j} fdf_{i,j}}{|d|} \quad (2)$$

where $i$ is the i-th feature ($i = 1\ldots6$); $j$ is the j-th word of $i$; $fdf_{i,j}$ (*feature dimension frequency*) is the frequency of words $j$ of feature $i$; and $|d|$ is the length of the k-th document $d_k$. For features such as funny, positive/negative, affective, and pleasantness, we decided an empirical threshold of representativeness ≥0.5. A document was assigned the value = 1 if its $\delta$ exceeded the threshold, otherwise a value = 0 was assigned. For instance, the text "I was searching for clothes that speak to me… These pants not only spoke to me, they entered my soul and transformed me" contains the words *pant* and *soul* which belong to the feature *funny*; *cloth*, *speak* (twice), *enter*, and *transform*, which belong to the feature *pleasantness*; and *search*, *speak* (twice), *pant*, *enter*, *soul*, and *transform*, which belong to the feature polarity. After summing the words ($j$) of all the features ($i$) that appear in the text ($d_k$), we obtain a frequency of 14, which is then normalized relative to the length of the text. Its $\delta$, thus, is 0.60.

A different criterion was determined for the n-grams and POS-grams because we were not only interested in knowing whether or not the sequences appeared in the corpus, but also in obtaining a measure to represent the degree of similarity among the sets. In order to define a similarity score, we used the Jaccard similarity coefficient. According to Formula 3, the similarity was obtained on the basis of comparing the overlapping between two sets given the union of both sets:

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

The classification accuracy was assessed employing three classifiers: Naïve Bayes (NB), support vector machines (SVM), and decision trees (DT). The sets were trained with 5861 instances (2861 positive and 3000 negative ones). 10-fold cross validation method was used as test. Global accuracy is shown in Table 3, whereas detailed performance, in terms of *precision*, *recall*, and *F-measure*, is given in Table 4.

### 5.1. Result discussion

Regarding the first goal (feature relevance), our a-priori aim of representing irony in terms of six general features seems to be acceptable. According to the results depicted in Table 3, the proposed model achieves good rates of classification which support this assumption: from 72% up to 89%, whereas a classifier that labels all texts as non-ironic would achieve an accuracy around 54%.

Moreover, both precision and recall, as well as F-measure rates corroborate the effectiveness of such performance: most of classifiers obtained scores >0.7. This means that, at least regarding the data sets employed in the experiments, the capabilities for differentiating an ironic review from a non-ironic one are satisfactory. However, it is important to note how the model is not constant with the three negative subsets. For instance, the TRI subset achieves the best results with all classifiers. In contrast, both AMA and SLA subsets obtain the worst ones. This behavior impacts on the learning. For instance, note in Fig. 1

**Table 3**
Classification results.

| | NB | SVM | DT |
|---|---|---|---|
| AMA | 72.18% | 75.75% | 74.13% |
| SLA | 75.19% | 73.34% | 75.12% |
| TRI | 87.17% | 89.03% | 89.05% |

[9] A preliminary evaluation of the comparison of the positive set against just two negative sets (AMA and SLA) was described in [25].

**Table 4**
Precision, recall and F-measure.

| | | Precision | Recall | F-measure |
|---|---|---|---|---|
| NB | AMA | 0.745 | 0.666 | 0.703 |
| | SLA | 0.700 | 0.886 | 0.782 |
| | TRI | 0.853 | 0.898 | 0.875 |
| SVM | AMA | 0.771 | 0.725 | 0.747 |
| | SLA | 0.706 | 0.804 | 0.752 |
| | TRI | 0.883 | 0.899 | 0.891 |
| DT | AMA | 0.737 | 0.741 | 0.739 |
| | SLA | 0.728 | 0.806 | 0.765 |
| | TRI | 0.891 | 0.888 | 0.890 |

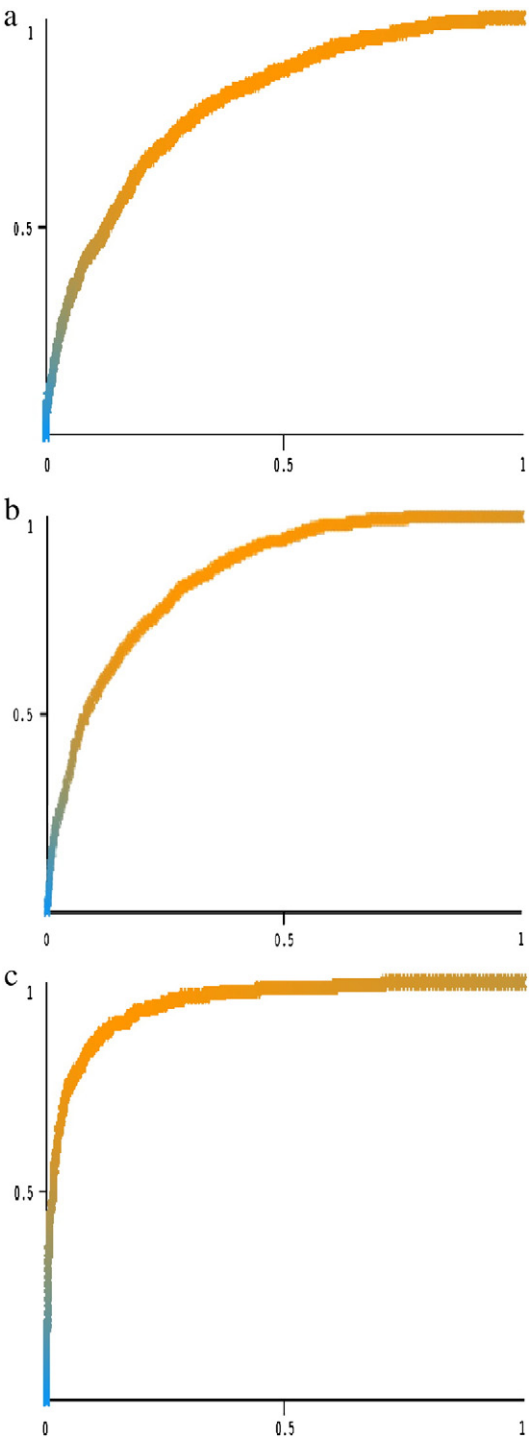


**Fig. 1.** Learning according to AMA (a), SLA (b), and TRI (c) subsets.

how the learning is achieved with less instances regarding the TRI subset, whereas the AMA and SLA ones require many more examples.

With respect to the second goal (the possibility of automatically finding ironic documents), an information gain filter was applied in order to verify the relevance of the model for finding ironic documents regarding the different narrative *discourses* profiled in each negative subset. In Table 5 we detailed the most discriminating features per subset according to their information gain scores. On the basis of the results depicted in this table, it is evident how the relevance of the feature varies in function in the negative subset. For instance, when classifying the AMA subset, it is clear how the POS-grams (order 3), pleasantness and funny features, are the most informative ones; in contrast, the pleasantness, n-grams (order 5) and funny features, are the most relevant ones regarding the SLA subset, whereas the n-grams (order 2, 3 and 4) are the most discriminating ones when the TRI subset is classified. Moreover, it is important to note how the negative words, without being the most differentiable ones, function very well as discriminating elements.

Taking into account these remarks, we could conceive the model as a *local optimum* model instead of a *global optimum* one; i.e. the model is a good solution for some data sets but it is not for all the possible data sets, hence, its efficacy to find ironic documents will depend on the kind of data.

### 5.2. Feature analysis

In this section we would like to stress some observations with respect to each feature.

Regarding the *n-grams*, it is important to note the presence of some interesting sequences which are not common to all three subsets. For instance: *pleasantly surprised*. However, we cannot define irony only in terms of these sequences because they might represent domain-specific information such as the bigram: *customer service*.

With respect to the *POS-grams*, the fact of focusing on morpho-syntactic templates instead of only on words seems to be more effective. For instance, the sequence *noun + verb + noun + adjective* would represent more information than the sum of simple words: *[grandpa/hotel/bed] + [looks/appears/seems] + [years/days/months] + [younger/bigger/dirtier]*. The sequences of POS tags show how an abstract representation could be more useful than a simple word representation. However, the relevance of such sequences might be language-dependent; i.e. the POS-grams intend to represent prototypical templates given POS information, but POS information is obtained by means of applying either a deep or shallow syntactic parser, hence, their relevance could be co-related to syntactic restrictions.

The *funny* feature seems to be a relevant element to express irony. However, its relevance might be supported by the kind of information profiled in the positive set. Considering the comic trend in the reviews posted by Amazon's customers, it is likely that many of the words belonging to this feature appeared in such reviews. For instance, in the following example the words in italics represent funny elements: "I cannot write this review and be any happier with my purchase. It replaced at least one or two of my *family guy* t-shirts and is perfectly

**Table 5**
Most discriminating features regarding the information gain filter.

| AMA | SLA | TRI |
|---|---|---|
| 3POS-grams | Pleasantness | 2-grams |
| Pleasantness | 5grams | 3-grams |
| Funny | Funny | 4-grams |
| 2POS-grams | Affectiveness | Pleasantness |
| 4POS-grams | 6-grams | 5-grams |
| Positive words | 2POS-grams | Funny |
| Negative words | 3POS-grams | Negative words |
| Affectiveness | Negative words | Positive words |
| 5POS-grams | 4-grams | 6-grams |
| 7POS-grams | 7-grams | 7-grams |

designed to hide my pit *stains* after *playing* twelve hours of xbox. I am an attractive *guy*. Slender, weak, and I have never shaved in my 19 years, but *sexy* as hell, and I cannot tell you how many *women* have flocked to me since my purchase". However, it is important to stress that this feature is equally discriminating for all sets, funny web comments included.

Concerning the *positive/negative profiling*, it is necessary to emphasize that, despite the greater number of negative words in the MSOL (more than 15,000 words of difference; cf. Section 4.4), the positive elements are the most representative in the ironic documents. This fact corroborates the assumption about the use of positive information in order to express an underlying negative meaning: "The $cool_{POS}$, $refreshing_{POS}$ $taste_{POS}$ of the $milk_{POS}$ washed away my $pain_{NEG}$ and its $kosher_{POS}$ $source_{POS}$ of $calcium_{POS}$ wash away my $fear_{NEG}$".

Regarding the *affective* feature, its relevance is not as important as we have a-priori considered, despite its being one of the features used to discriminate the SLA subset: "Man, that was *weird* …I think it is *funny*, because there's a *good* overlap". However, if we take into account the whole accuracy for this subset, then we can conclude that its relevance is minor. Nonetheless, we still consider that the affective information is a valuable factor which must be taken into account in order to provide rich knowledge related to subjective layers of linguistic representation.

The role played by the *pleasantness* feature on the classifications is significant. Despite the feature's not being the most discriminating, its effectiveness for increasing the classification accuracy is remarkable. For instance, consider the following ironic sentence: "I *became* the *man* I *always dreamed* I *could be all* those *nights staying up* late watching *wrestling*", where most of its constituents are words whose pleasantness score is $\geq 2.5$; i.e. these words (in italics) should communicate information related to favorable pleasant contexts.

## 6. Re-evaluating the model

We have highlighted throughout the previous sections the difficulty of capturing, by means of linguistic elements, the essence of irony. Phenomena such as linguistic and social factors impact on the perception of irony, making the task of automatically identifying ironic documents quite complex. Nonetheless, despite these issues, we have suggested a model which seems to be efficient to describe salient irony attributes beyond a purely theoretical framework. However, could this model be useful beyond the data sets we have employed, especially if we take into account the way in which we obtained the features? (They were not obtained by manually annotating the ironic data, but by trying to represent the core of this concept with general categories). In this section we intend to provide arguments to answer this question.

To this end, we employed the corpus described in [6]. This corpus was firstly used to perform experiments on automatic satire detection. It contains 4233 news articles, of which 233 are satiric articles. We decided to assess the capabilities of our model on this corpus due to the two following reasons: i) as we have stressed in Section 3.1, there are not available corpora with ironic examples to learn from; thus, the possibility to compare our method with a baseline is, so far, null; ii) according to our definition of irony, stated in Section 2, figurative devices such as satire or sarcasm are means to express, as well as to contain, ironic content; hence, in the absence of an ironic baseline, the satiric content of this corpus represents ad hoc instances to evaluate the model.

The experiment consisted of representing the 233 satiric articles, as well as 700 randomly selected non satiric ones (or *real*, following the terminology employed by the authors)[10] by means of the features previously described. The aim was focused on assessing the relevance of the model to accurately retrieve the satiric instances on the basis only of such representation. The same processing was applied to the 933 instances; i.e. they were stemmed, stop words were removed, and finally, they were transformed into term vectors. The vectorization was performed by assigning a value = 1 every time a word (or sequence of them, or their POS tags) appears in the document, regardless of the feature it belongs to. These values were summed and divided by the number of features of the model[11] in order to obtain the documents whose probability to be considered as satiric was greater. The final target was focused on retrieving as many satiric articles as possible.

**Table 6**
Most informative features regarding the re-evaluation.

| Ranking | Feature | Ranking | Feature |
|---|---|---|---|
| 1 | 3POS-grams | 9 | 4POS-grams |
| 2 | 2POS-grams | 10 | 3-grams |
| 3 | Funny | 11 | 2-grams |
| 4 | Affectiveness | 12 | 6POS-grams |
| 5 | Pleasantness | 13 | 5POS-grams |
| 6 | Positive words | 14 | 4-grams |
| 7 | Negative words | 15 | 5-grams |
| 8 | 7POS-grams | 16 | 6-grams |
| | | 17 | 7-grams |

The results are very interesting. Considering 233 as the maximum of documents to retrieve, the model predicted 193 satiric articles, failing in 40 articles; i.e. the accuracy is 82.83%. Moreover, after applying an information gain filter to these results, we could corroborate some of the observations discussed in Section 5.1. For instance, the ranking of the most informative features, presented in Table 6, shows the practically null relevance of the n-grams in the task, whereas the rest of the features keep a similar relevance to the one registered with our data sets. According to these results, we can infer the applicability of the model. If the accuracy achieved is similar in all the experiments (cf. Table 3), it means that some underlying patterns to express what people consider the core of figurative contents (either with respect to irony or satire), are adequately represented by these features. Concluding, due to the difficulty of the irony detection task, we considered the accuracy obtained to be promising.

## 7. Conclusions and future work

Irony is one of the most subjective phenomena related to linguistic analysis. Its automatic processing is a real challenge, not only from a computational perspective but from a linguistic one as well. The linguistic and social factors which impact on the perception of ironic utterances make the task of automatically detecting ironic documents quite complex. However, in this work we have suggested, beyond a theoretical framework, a model which attempts to describe salient characteristics of irony. According to our definition of irony (Section 2), we have established a model to represent verbal irony in terms of six categories of features: n-grams, POS-grams, funny profiling, positive/negative profiling, affective profiling, and pleasantness profiling. They intend to symbolize low and high level properties of irony on the basis of formal linguistic elements. A freely available data set with ironic reviews was created to assess our initial assumptions. Two goals were considered in the evaluation: feature relevance and capability of finding ironic documents. The results achieved with three different classifiers are satisfactory, both in terms of classification accuracy, as well as precision, recall, and F-measure. At this point, it is worth mentioning that, although the learning examples focus on some products which could not be in vogue anymore, the underlying mechanism (viral effect) to produce their popularity is currently one of the most employed in the web. Thus, the fact of considering them is not trivial, since the same effect can be extrapolated to many other products and situations, thereby

[10] In this case we focused on keeping a relation 1 to 3 because the figurative contents (either ironic, satiric, or sarcastic) do not appear in real contexts in a relation 1 to 1.

[11] A total of 17 features (n-grams from 2 to 7; POS-grams from 2 to 7; funny, negative, positive, affective, and pleasantness profiling).

achieving important implications for tasks where irony plays an important role. For instance, companies can have direct access to negative information and, on the basis of that information, plan actions in order to reverse the negative image. However, when the information implies more than a positive or negative opinion, it is more difficult to make a correct decision.

Finally, an evaluation with new and unseen data (Section 6) showed the relevance of the model for retrieving figurative content. In the near future we plan to manually annotate the ironic samples in order to compare the results with the ones presented in this paper. Furthermore, new features will be studied in order to come up with an improved model capable of detecting better ironic patterns in different kinds of texts.

## Acknowledgments

The National Council for Science and Technology (CONACyT — Mexico) has funded the research of the first author. The European Commission as part of the WIQEI IRSES-Project (grant no. 269180) within the FP 7 Marie Curie People Framework has partially funded this work. This work was carried out in the framework of the MICINN Text-Enterprise (TIN2009-13391-C04-03) research project and the Microcluster VLC/Campus (International Campus of Excellence) on Multimodal Intelligent Systems.

## References

[1] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, M. Padró, Freeling 1.3: syntactic and semantic services in an open-source nlp library, Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006, pp. 48–55.
[2] S. Attardo, Irony as relevant inappropriateness, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 135–174.
[3] S. Baccianella, A. Esuli, F. Sebastiani, Multi-facet rating of product reviews, Proceedings of the 31st European Conference on Information Retrieval, Lecture Notes in Computer Science, vol. 5478, Springer, 2009, pp. 461–472.
[4] L. Bentivogli, P. Forner, B. Magnini, E. Pianta, Revising the wordnet domains hierarchy: semantics, coverage and balancing, in: G. Sérasset (Ed.), Multilingual Linguistic Resources (COLING 2004), 2004, pp. 94–101.
[5] T. Brants, A. Franz, Web 1t 5-Gram Corpus Version 1, , 2006.
[6] C. Burfoot, T. Baldwin, Automatic satire detection: are you having a laugh? ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009, pp. 161–164.
[7] P. Carvalho, L. Sarmento, M. Silva, E. de Oliveira, Clues for detecting irony in user-generated contents: oh…!! It's "so easy";-), TSA'09: Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, ACM, Hong Kong, China, November 2009, pp. 53–56.
[8] H. Colston, On necessary conditions for verbal irony comprehension, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 97–134.
[9] H. Colston, R. Gibbs, A brief history of irony, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 3–24.
[10] I. Councill, R. McDonald, L. Velikovich, What's great and what's not: learning to classify the scope of negation for improved sentiment analysis, Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, University of Antwerp, Uppsala, Sweden, July 2010, pp. 51–59.
[11] C. Curcó, Irony: negation, echo, and metarepresentation, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 269–296.
[12] D. Davidov, O. Tsur, A. Rappoport, Semi-supervised recognition of sarcastic sentences in Twitter and Amazon, Proceeding of the 23rd International Conference on Computational Linguistics (COLING), July 2010.
[13] R. Gibbs, Irony in talk among friends, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 339–360.
[14] R. Gibbs, H. Colston, The future of irony studies, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 339–360.
[15] R. Giora, On irony and negation, Discourse Processes 19 (2) (1995) 239–264.
[16] H. Grice, Logic and conversation, in: P. Cole, J.L. Morgan (Eds.), Syntax and Semantics, vol. 3, Academic Press, New York, 1975, pp. 41–58.
[17] A. Jøsang, R. Ismail, C. Boyd, A survey of trust and reputation systems for online service provision, Decision Support Systems 43 (2) (2007) 618–644.
[18] D. Kim, D. Ferrin, H. Raghav, A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents, Decision Support Systems 44 (2) (2008) 544–564.
[19] R. Kreuz, Using figurative language to increase advertising effectiveness, Office of Naval Research Military Personnel Research Science Workshop, University of Memphis, Memphis, TN, 2001.
[20] X. Lü, L. Zhang, J. Hu, Statistical substring reduction in linear time, Proceedings of IJCNLP-04, Hainan Island, 2004.
[21] J. Lucariello, Situational irony: a concept of events gone awry, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 467–498.
[22] R. Mihalcea, C. Strapparava, Learning to laugh (automatically): computational models for humor recognition, Journal of Computational Intelligence 22 (2) (2006) 126–142.
[23] G. Miller, Wordnet: a lexical database for English, Communications of the ACM 38 (11) (1995) 39–41.
[24] A. Reyes, P. Rosso, Linking humour to blogs analysis: affective traits in posts, Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA), CAEPIA-TTIA Conference, Universidad de Sevilla, Sevilla, Spain, 13 November 2009, pp. 100–109.
[25] A. Reyes, P. Rosso, Mining subjective knowledge from customer reviews: a specific case of irony detection, Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), Association for Computational Linguistics, 2011, pp. 118–124.
[26] A. Reyes, P. Rosso, D. Buscaldi, Humor in the blogosphere: first clues for a verbal humor taxonomy, Journal of Intelligent Systems 18 (4) (2009) 311–331.
[27] M. Saif, D. Cody, D. Bonnie, Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus, Proceedings of the 2009 Conference on EMNLP, Association for Computational Linguistics, Morristown, NJ, USA, 2009, pp. 599–608.
[28] L. Sarmento, P. Carvalho, M. Silva, E. de Oliveira, Automatic creation of a reference corpus for political opinion mining in user-generated content, TSA '09: Proceedings of the 1st international CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, ACM, Hong Kong, China, November 2009, pp. 29–36.
[29] C. Strapparava, A. Valitutti, WordNet-affect: an affective extension of WordNet, Proceedings of the 4th International Conference on Language Resources and Evaluation 4 (2004) 1083–1086.
[30] O. Tsur, D. Davidov, A. Rappoport, ICWSM — a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews, in: W.W. Cohen, S. Gosling (Eds.), Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, The AAAI Press, Washington, D.C., 23–26 May 2010, pp. 162–169.
[31] A. Utsumi, A unified theory of irony and its computational formalization, Proceedings of the 16th Conference on Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, 1996, pp. 962–967.
[32] T. Veale, Y. Hao, Support structures for linguistic creativity: a computational analysis of creative irony in similes, Proceedings of CogSci 2009, the 31st Annual Meeting of the Cognitive Science Society, 2009, pp. 1376–1381.
[33] C. Whissell, The dictionary of affect in language, Emotion: Theory, Research, and Experience 4 (1989) 113–131.
[34] D. Wilson, D. Sperber, On verbal irony, in: R. Gibbs, H. Colston (Eds.), Irony in Language and Thought, Taylor and Francis Group, 2007, pp. 35–56.

**Antonio Reyes** is a Ph.D. student at Universidad Politécnica de Valencia, Spain. He is currently a member of the Natural Language Engineering and Pattern Recognition research group. His major interests are focused on figurative language processing; especially, on topics related to irony, sarcasm, and humor. He has published some papers in different conferences, workshops and journals being involved in many national and international research projects.

**Paolo Rosso** received his Ph.D. degree in Computer Science (1999) from the Trinity College Dublin, University of Ireland. He is currently an Associate Professor at Universidad Politécnica de Valencia, Spain, where he leads the Natural Language Engineering Laboratory of the Natural Language Engineering and Pattern Recognition research group. He has published over 200 papers in different conferences, workshops and journals being involved in many national and international research projects. His main research interests are mainly focused on irony detection, humor recognition, plagiarism detection and geographical information retrieval.