Sarcasm Detection by Measuring Sincerity

Eric Dagobert
PhD Computer Science / GC CUNY
edagobert@gc.cuny.org

Introduction

The aim of sarcasm is to express a criticism unbeknownst to the target while the funny facet is revealed to the audience. Sarcasm uses figurative devices such as metaphor, analogy and irony to express a negative feeling wrapped into a seemingly flattering statement, and thus is hard to detect not only by human let alone by computers. Because of its elusive aspect, sarcasm is not formally defined. For [1], sarcasm is ironical and funny, whereas, for [2], irony suffices to define sarcasm. We agree that humor is part of sarcasm, but we also admit that humor is subjective and thus hard to detect. Plus it seems that irony, by the ambiguous meaning it carries, is intrinsically humorous.

So why detecting sarcasm? First, because it's funny. More seriously, understanding mechanisms of irony is not only a support to better understand how language works, but also can be applied to broader subjects such as humor or manipulation. Interpretation of figurative languages such as slang could also be analyzed with the same mechanisms. A more pragmatic aspect of irony is that in sentiment analysis a sarcasm might be literally interpreted instead of figuratively thus be considered as a positive utterance whereas it is just the opposite. That is why Amazon for instance is building tools to detect presence of sarcasm in product reviews in order to increase the accuracy of customers sentiments. In this paper we would like to demonstrate that Irony can be detected using the concept of Sincerity: apparent positive statement that hide a negative one. To do so we will describe a way to evaluate the sincerity of a text with semantic features.

1 Background

Elena Filatova [2] created a corpus dedicated to sarcasm and irony. She rightly argues that systems of opinion mining and sentiment analysis can be improved by correctly identifying sarcastic utterance. For her, building a corpus is the necessary first step towards any scientific study because it materializes an agreement on what is irony, in absence of formal definition. This corpus is made of Amazon product reviews manually labeled by regular people via Mechanical Turk, another Amazon achievement. Asking internet users to voluntarily participate to experiments like this is now called crowd-sourcing. This corpus has been widely used by different researchers since its creation.

Sarcasm has already been the object of different studies looking at different approaches: [4] uses prosody features to detect sarcasm in speech with an accuracy of about 80%. Interestingly, in that study, supervised training is built from acted performances in a TV show. This is unexpected because of the time non-actors provide better emotion expressions than actors.

About written text, a wide range of NLP techniques has been more or less successfully employed. The most common approach it seems is described in [5] where sarcasms are detected through lexical and non-verbal written features. Davidof uses punctuation marks (exclamation, dots, interrogation, quotes) as features and as pattern objects, patterns that are made, among other things, of High Frequency Words and punctuation. We could make a parallel with the use of non-verbal devices, for instance locutions such as lol or emoticons or also punctuation, with the envelopes of speech used in [5]. In both cases, the meaning is occulted to the profit of the communication aspects. And in both studies, the accuracy is close to 80%. But Davidof obtains this type of accuracy from Twitter only; his team tested their algorithm on Amazon reviews with moderate success (F-score of 0.5). We would explain this with the fact that Twitter is actually closer to spoken that written language, allowing only brief sentences exchanged in almost real time and using lexical features to express emotions in a similar way the tone or the pitch of the voice help carrying feelings.

An advanced work on semantic analysis of irony (and humor) is provided in [1]. In this paper figurative features of language are emphasized, these features being classified in four categories: Ambiguity (in three layers: structural, morphosyntactic and semantic), Polarity, Unexpectedness and Emotional scenarios. Such classification is inspired from Utsumi[3], a first attempt to formalize irony. Ambiguity is determined from the probability of senses given to a word (computed from synsets or ontological markers). Polarity is the sentiment expressed in a sentence, good or bad, and measured from the number of positive and negative words. The contextual imbalance is a function of the similarity measurement among the words, and emotional scenarios are evaluated from categories provided by Whissel's Dictionary of Affect in Language. Results are very good concerning humor detection, with an accuracy of more than 90% in most cases, but, in their experiments, irony remains hard to detect (F-score of 0.60) without the ambiguity feature. Adding ambiguity consequently improve accuracy (F-score=0.9) of irony detection in Humor, Politics or Technology, but remain weak in General topic text with a F-score of 0.65.

2 Method

We have used Elena Filatova Sarcasm Corpus[1] for the purposes of both training and testing our hypothesis. The data set consists of 1200 reviews of Amazon non-specific products, separated into two categories: Regular and Ironic. That classification has been made using 'crowdsourcing', i.e. asking people to volunteer to review the texts and label them. Reviews are chosen by pair: for the same product we find one standard review and one ironical. Then every review is annotated by five other volunteers and the classification decision is taken from the majority of votes (votes are weighted by Krippendorf's alpha coefficient to distinguish reliable annotators from non-reliable).

In this paper we aim to characterize sarcasm (or irony) in written text by finding the "amount" of sincerity, an aspect not mentioned previously but present in [3].
We think that, in opposition to [1], the sarcasm is intrinsically funny, and there is no need to find humor to find sarcasm. Indeed, there is several levels of wittiness in irony: First of all it let suppose the object of sarcasm is the only one not be aware of the incoherence it is the center of, and turns into a target of teasing (If the object is not a human, it is implicitly extended to human in the same situation or related somehow to the object). Secondly, in the case of excessive sincerity, the finality is to show absurdity as well, at a different level this time, the absurdity being in the answer. For instance, in french, a sarcastic way to express something is boring or stupid is the expression "I'll talk about this with my horse", which is absurd, but seems positive.

Another point evoked in [3] is ambiguity, but we claim presence of ambiguity is simply to characterize an under sincere statement. For example the following sarcasm :"You're so mature". At first this seems globally positive, articulated as a compliment. Now if we extend the adjective "mature" semantically, WordNet gives us "old" as the next equivalent term. The sentence is then similar to "you're so old", and it is not flattering any more. This statement appeared sincere, but a small semantic shift makes it negative. Over sincerity is also a way to express irony. Here again, the sarcasm is efficient if it underlines some humor, but this would be the case if the absurdity is present, and also if the dosage of insincerity is small enough to subtly hid the negative intention. For example the locution "Yeah sure !" presents an overstatement by repeating the acknowledgment combined with the exclamation point. A simple statement "sure" or "why not" is usual to approve and contains the "correct" amount of sincerity. We can remark the punctuation features mentioned earlier can be seen as markers of over sincerity, thus help detect irony but they are not primary causes. To summarize, we are looking forward to detecting sarcasm through a measure of sincerity.It appears ambiguity and absurdity can be vectors of under sincerity and punctuation characterizes over sincerity but in our claim, these are not primary factors of sarcasm.

3 Evaluation and Implementation

To validate the assumption described above, Elena Filatova's corpus has also been used as a reference data set from which we could compare the results of our study with other papers ([1],[6] and [5]).
We have determined the following semantic features to express a multidimensional vector of Sincerity:

- Polarity : number of positive words versus negative words.
- Hidden Polarity: polarity of semantic equivalences (maximal negative polarity).
- Incoherence : "semantic drift", variation of similarity within the sentence computed from sliding window of words.
- Emphasis : number of adjectives or nouns semantically close that are repeated. Considered but not used.

Each review has been represented by a feature vector; values have been normalized and scaled with Inverse Hyperbolic Tangent. Feature vectors have been used to feed a supervised learning algorithm. After different trials, it seems SVM with Radial Kernel gives the best results.
At first, texts have been parsed with nltk default

[1] http://storm.cis.fordham.edu/ filatova/SarcasmCorpus.html

tagger (Brown corpus) in order to extract adjectives, verbs and nouns. Afterwards, polarity has been computed with VADER [7]: 'good','bad' and 'neutral' parameters having values comprised between 0 and 1 express the intensity of the sentiment. VADER Sentiment analysis evaluation uses valence-based lexicons and lexical analysis features, combined with machine-learning, to create a method of evaluation taking into account context-awareness and typical lexical patterns associated to social network (such as emoticons). Choosing this way of evaluating polarity was decided because research papers show this method is relatively accurate, and also makes sense as the Amazon reviews contains many of these social networks lexical patterns. As a complement of polarity, some people are using the global rating (1-5 stars) of the product, but we thought it would be probably insincere in ironical reviews, and it cannot be considered as a semantic feature and we discarded it. VADER has provided us with four polarity features : negative value, neutral value, positive value and compound value.

The next feature, hidden polarity, has been computed from adjectives and verbs composing each review. For every term we looked at the hidden negative polarity using Wordnet Synsets and sentiment/opinion analysis library (SentiWordNet). To do so, we built an ontological graph (Synsets) recursively up to a given depth. Such a graph grows exponentially so it was pruned by keeping only words having a certain range of similarity with the reference word (here we used the Wu & Palmer measure of similarity). Then we selected the maximum negative sentiment within the graph. For example, the sentence 'you are so mature !' has many different equivalent representations but the 'you are so old!' is probably the most negative. At the end, the hidden negative sentiment for a given review was represented by two factors: the maximum negative value, as well as the sum of negative values. This to reflect hidden negative feelings either expressed with only a strong word, or a combination of weaker terms.

The last set of features used in this experiment was related to Incoherence, or Inconsistency. Indeed a statement expressing a very high positive sentiment and no hidden negative meaning may be classified as sarcastic if the content is the sightliest absurd (therefore probably funny), thus not sincere. To do so we first have created a matrix of cross-similarity from the verbs and nouns contained in every sentence.Similarity is measured by taking the maximum Wu & Palmer measure between the two synsets corresponding to the terms indexing the matrix. After filtering out zeros and ones (extreme values), we measured the level of coherence of a review by calculating the minimum, maximum, and the average similarity, which roughly define a 'coherence surface', the highest being the most semantically consistent sentence.

Next, we have proceeded to a binary classification of reviews (ironic/regular) using different algorithms, the most successful being the Support Vector Machine with radial kernel. [2]

Feature class	Method	Feature Value
Polarity	VADER	positive valence
Polarity	VADER	neutral valence
Polarity	VADER	negative valence
Polarity	VADER	compound
Hidden Negativity	Synsets	max (negativity)
Hidden Negativity	Synsets	sum (negativity)
Coherence	Similarity	Min(X-sim)
Coherence	Similarity	Max(X-sim)
Coherence	Similarity	Mean(X-sim)

Table 1: Feature table

4 Experiment and Results

Feature extraction has been achieved using Python and the nltk toolkit, including the Wordnet semantic package. Machine learning was based on sklearn Python libraries. Training and testing sets have been generated on a ratio of 80% for training and 20% for testing. The accuracy score has been computed using a 10-fold cross validation method. Accuracy represents the ratio of correctly classified items on total (weighted to have same precision and recall if the classification is pure random). We reached an accuracy average of 74 % +/- 0.02%. In order to compare this system with others, Table 2 recapitulates F1-scores for similar studies based on Amazon product reviews. Results found here are encouraging: if there is room for improvement, nevertheless the hypothesis claiming sarcasm can be defined from sincerity seems valid. Here are some examples of successful and unsuccessful classifications.

System	F1	P/ R	Type
This Paper	.71	.72 .70	Semantic
SASI [5]	.78	.77 .81	Lexical
[1]	.65	.78 .56	Semantic
[6]	.74	* *	Punctuation, BOW

Table 2: Some F1-scores / Amazon reviews

[2] sklearn provides all kind of ways to fine tune machine learning parametrization, from machine type to machine parameters. We tried : KNN, SVM with linear, rbf and polynomial kernels.

Successful classification:

- non-ironic: review on a Thanksgiving hat. "I wore this hat as part of my Thanksgiving outfit for school. Everyone loved the hat." has been classified as regular: indeed, it is slightly positive,rather neutral, no double meaning, coherence in the sentence and objectiveness.

- ironic: review on a Halloween mask: "This is a great mask to strike fear into those around you - an essentiality for the every day rapist or mugger" Detected as positive evaluation with a strong hidden negative aspect as well as a little incoherence .

Unsuccessful classification:

- Wrongly classified as ironic: "Terribly disapointing movie. Kevin Spacey mails it in, playing the same likable villain he played in The Client, minus the southern accent. Choopy, uneven, inplausible script. No character development. Just a waste of time." If we make abstraction of the fact that some vocabulary may not be present in the VADER lexicon, thus biases the evaluation, the main factor triggering irony classification is here the relative unsimilarity of terms employed: 'mails', 'villain', 'accent' have a very high semantic distance.

- Irony not detected: "If you pay $250 for this blender you need your head examined. It's bulky, heavy and does the exact same thing as any $25 WalMart blender." Interestingly, the part "If you pay [..] for this blender you need your head examined. [..] has been classified as ironic! So far no explanation has been found on why the price ($250) here could have impacted the result, out of a probable bug in the implementation.

5 Analysis

Part of this study consists then in the analysis of results and the exploration of some paths to improvement:

Data

Irony is often defined as the expression of one's meaning by using language that normally signifies the opposite, typically for humorous or emphatic effect. But that definition is not always perfectly respected in the data set we have. We've seen things that were sometimes intended to be funny but clearly negative, as for instance: terrible swill. I can't belive this is from the company that WANTS to sell you their coffee maker. Try the Millstone or if you can find it the Target brands instead.. We recognize the manual aspect of the classification based on a concept difficult to define can lead to some uncertainty. Another point is that reviews are, most of the time, composed of several sentences, where irony is often present in one sentence only, mostly at the end of the text (to trigger a surprising comical effect for example). Having a mechanism scoring sentences rather than reviews would definitely be more accurate: Very often, irony is characterized by the presence of one particular term or group of terms and can be lost in noise especially in the case of long text. Therefore, a review could have been marked as sarcastic if at least one sentence was. The data set would have probably benefited from having ironical sentences classified, instead of entire text, and may have improved accuracy of semantic analysis and reduced the chances of overfitting on supervised learning. But this would necessitate the daunting task to re-evaluate the data set and manually isolate ironical sentences from the rest.

Spelling

Misspelling is also an issue when we need to rely on POS tagging, but this is an issue hard to address: the main goal is to detect irony automatically, so if some words are misspelled, manual correction is out of question. A solution involving an automatic spell check is not ideal neither, an taking care of all possible spellings in sentiment lexicons is almost impossible. At first glance, a misspelled word could be replaced by its closest Soundex.

Features

This approach is overall closer to a sketch of a solution rather than a complete study and it would necessitate some kind of fine tuning on the feature evaluation side. For instance about Emphasis we mentioned earlier on. We decided not to use it because this concept is not so simple to analyze after all. Indeed Emphasis can indicate sarcasm in some cases (locutions such as 'very good', exclamation mark, etc.) but it can also mean a good faith appreciation. It seems Emphasis will amplify either a sincere review or an ironical one. So it could be seen as a 'multiplier' of the writer's intent and not a feature coming along other factors. Nevertheless, the accuracy could be improved adding this parameter.

Last but not least the measure of Incoherence must also be refined, because as it is now, figurative language has a high probability to be recognized as incoherent, which is not always accurate. To avoid that, one approach would be either to use Vader-like contextual valuation of sentences parts and estimate the similarity of expressions or concepts, but this could be the object of another research paper.

6 Conclusion

In this paper we described a semantic approach of detecting irony inspired from Utsumi [3] definition of Sincerity, and we claimed that Humor, even if present in Irony, is rather a consequence than a cause. Even if the accuracy of our method can be improved, results can be considered as significant to validate our claim. Nevertheless it looks like raising levels of accuracy above 80% will necessitate a deeper and finer feature characterization.

References

[1] Reyes, Rosso and Buscaldi, From humor recognition to irony detection: the figurative language of social media. Data and Knowledge engineering, Elsevier, 2012.

[2] Filatova, Elena Irony and Sarcasm: Corpus Generation and Analysis Using Crowd sourcing. Fordham University, .

[3] Utsumi Akira, A Unified Theory of Irony and Its Computational Formalization. Association for Computational Linguistic, 1996.

[4] Rosenberg, Rakov,"Sure, I Did The Right Thing": A System for Sarcasm Detection in Speech. CUNY,2012.

[5] Davidov, Tsur, Rappoport Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon, Proceeding of the Fourteenth Conference on Computational Natural Language Learning,2010.

[6] Buschmeier, Cimiano, Klinger An Impact Analysis of Features in a Classification Approach to Irony Detection in Products Reviews, Bielefeld University,2014.

[7] Hutto, C.J. and Gilbert,Eric, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, Georgia Institute of Technology.