Inducing an Ironic Effect in Automated Tweets

Alessandro Valitutti, Tony Veale
School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin D4, Ireland
Email: {Tony.Veale, Alessandro.Valitutti}@UCD.ie

***Abstract*—Irony gives us a way to react creatively to disappointment. By allowing us to speak of a failed expectation as though it succeeded, irony stresses the naturalness of our expectation and the absurdity of its failure. The result of this playful use of language is a subtle valence shift as listeners are alerted to a gap between what is said and what is meant. But as irony is not without risks, speakers are often careful to signal an ironic intent with tone, body language, or if on Twitter, with the hashtag #irony. Yet given the subtlety of irony, we question the effectiveness of explicit marking, and empirically show how a stronger valence shift can be induced in automatically-generated creative tweets with more nuanced signals of irony.**

I. Introduction

Verbal irony is a powerful communicative device. It can be used to express sentiments and opinions in a surprising and subtle way. When a statement is recognized as ironic, its meaning is perceived as different from (and often opposite to) the one it would express if non-ironic. In the specific case of affective meanings such as sentiments and opinions, verbal irony can modify their valence and polarity. For example, the ironic comparison *"as useful as a chocolate teapot"* [1] induces polarity inversion in the word 'useful'.

In the context of NLP, there is an increasing interest in the automated detection and generation of ironic texts [2]. The capability to detect verbal irony could improve the performance of text mining tasks such as sentiment analysis. On the other hand, the controlled production of verbal irony could increase the creative expressiveness of natural language generation.

In this work, we studied the extent to which a computer-generated sentence, recognized as ironic, can change the affective valence typically attributed to specific words and invert their polarity[3], [1]. We call *valence effect* the capability of verbal irony to produce valence shifting or polarity inversion. In particular, we used valence effect to perform an indirect measurement of verbal irony.

As a source of ironic statements generated automatically, we used *@MetaphorMagnet*, a Twitterbot described by Veale [4]. The system can generate a rich range of creative statements, which are regularly posted on Twitter. A subsets of the tweet patterns are designed to be intentionally ironic. We used *@MetaphorMagnet* as a test bed, to generate randomized sets of outputs according to several experimental settings. Then, we evaluated them with human judges using a crowdsourcing platform.

In particular, we studied two types of features employed in the generation of irony: (1) adjectives with opposite polarity and contrastive comparison, used to produce ironic incongruity, and (2) irony markers such as scare quotes and hashtag *#irony*.

The results of the empirical evaluation show that the perception of irony is correlated to two complementary elements: on one hand, the semantic contrast induced by adjectives and comparisons and, on the other hand, the irony clue provided by scare quotes. Furthermore, they produce statistically significant variation of both valence and polarity rates on specific target words.

II. Background

A. Semantic Opposition, Irony Markers and Echo

Verbal irony is a rhetorical device in which the intended meaning of statements is different from (and typically opposite of) the literal meaning. Encyclopaedia Britannica defines verbal irony[1] as a "language device [...] in which the real meaning is concealed or contradicted by the literal meanings of the words".

In his account of linguistic theories of irony[2], Wilson [5] emphasizes three frameworks as key steps. The first theoretical framework, coming from classical rhetoric, describes irony as a form of figurative communication. According to this interpretation, ironic meaning is detected by recipients through a process of inference from the literal meaning and its underlying grammatical structure. A limitation of this approach is that it does not give account of the several cases in which the literal content of the utterance is not sufficient to infer the ironic interpretation. An important theoretical change can be ascribed to Grice [6], [7], which reframed irony as a pragmatic phenomenon and, as such, it should include the communicative intentions of the writer.

A more recent theoretical description was proposed by Sperber and Wilson [8], according to which an ironic utterance is characterized as "echoic". This term is used to mean that the utterance alludes to some previous remark or a familiar fact, not necessarily expressed by the literal meaning. The intent is to express a sort of dissociation respect to the fact being echoed. This particular attitude is, thus, the motivation to express a remark in an ironic way. In the automatic generation of ironic statements, we identify a specific feature for each of the above theoretical approaches. The first framework is

[1] `http://www.britannica.com/EBchecked/topic/294609/irony` – retrieved 30 April 2015.

[2] From here on, we use the terms *verbal irony* and *irony* interchangeably.

based on the inferential connection between literal and ironic meaning. We employ a form of semantic opposition based on affective polarity. The second framework emphasizes the need to make the ironic intention recognizable by the reader. According to it, we make use of irony markers. Finally, an implication of the third framework is that ironic statements typically make reference to some familiar or common-sense fact to be echoed. Hence we consider ironic statements produced as modification of the familiar fact using semantic opposition and irony markers. In this way, the familiar fact is referenced by the ironic statement since it is embedded in it.

Semantic opposition and irony markers are different types of features. Features of the first type are called *irony factors* (or *contextual markers*). According to Attardo [9], contextual markers should be considered factors representing the "content of irony". On the other hand, *irony markers* (also called *co-textual markers*) are used as meta-communicative clues helping the reader to identify the ironic intent and the related content. In a research by Carvalho et al. [10] a number of textual elements such as interjections or scare quotes were identified as irony markers.

In this study, we focus on ironic features such as contrastive expressions (i.e., pairs of words or phrases with opposite polarity) and special punctuation (e.g. scare quotes).

B. Valence Shifting

According to Gardiner and Dras [11], valence shifting consists of "rewriting a text to preserve much of its meaning but alter its sentiment characteristics". In the context of this paper, we use the term *valence* to indicate the evaluative or affective attitude (e.g. opinion, judgment, feeling, or emotion) intended by the author of the text and perceived by the reader. Moreover, we assume the terms *sentiment*, *connotation*, and *semantic orientation* as synonyms of valence. Although valence and polarity are sometimes used as equivalent terms in literature, a few authors use them as distinct concepts [12] and discuss the relatedness between irony and polarity inversion [13]. According to Moreno-Ortiz et al. [12], the usage of polarity is restricted to *"non-graded, binary assignment, i.e., positive / negative"*, whereas valence *"is used to refer to a rating on an n-point semantic orientation scale"*.

As we consider polarity and intensity as two dimensions of valence, they can be used to identify different types of valence shifting. For example, valence shifting could occur as a variation of intensity, thus without changing of polarity. On the other hand, polarity inversion corresponds to a more constrained type of valence shifting characterized by change of polarity.

According to literature, there are at least three different ways in which valence shifting can be induced computationally. One approach is based on the use of *valence shifters*. This term indicates words capable of modifying the sentiment expressed by other words in the text, such as negatives and intensifiers [14]. A number of previous works have been done on using valence shifters to improve the performance of sentiment analysis at the sentence level [3], [15].

A different approach consists of the substitution of single words with synonyms having different connotation. Using this form of word replacement, called *lexical slanting*, the overall sentiment of the text containing the original word is modified accordingly. Guerini et al. [16] implemented lexical slanting in the *Valentino* tool. Gardiner and Dras [11] employed this system to perform an evaluation of valence shifting with human raters.

Finally, a third line of research (related to what we call here *valence effect*) is based on automated generation of ironic statements. Typically, ironic sense and literal one have opposite polarities, even though there are cases in which they have same polarity and different only by intensity of valence [1].

III. Automated Irony in @*MetaphorMagnet*

Metaphor and irony each hinge upon a provocative contrast. Metaphors allow us to view an entity T as though it were a member of category V, where V and T share some salient similarities (e.g. think of gas-guzzling cars as alcoholics) *and* some striking dissimilarities (e.g. unlike alcoholics, cars are neither living nor sentient). If an apt context can lend more weight to similarities than to dissimilarities, an effective albeit provocative metaphor ensues. In contrast, irony prefers to emphasize dissimilarity, to make sport of contrast and to stress the failure of reasonable expectations. If a computational system can imagine scenarios that juxtapose similarity *and* dissimilarity, it can build metaphorical and ironic observations upon the very same relational chassis.

@*MetaphorMagnet* is an automated generator of figurative tweets (a Twitterbot) that finds and frames the contrasts that arise when aspects of its knowledge-base are bisociated [17]. Its knowledge comes from a variety of Web services: the *MetaphorMagnet* Web service of Veale [18], which models the stereotypical properties of familiar ideas and attests, via corpus evidence from the Google n-grams [19], as to how these properties might be applied to other targets; the *Metaphor Eyes* Web service of Veale & Li [20] which models the relational structure of concepts by harvesting generic relationships from WH-questions in Web query logs, and which supports analogical mapping over its relational structures; and the *Thesaurus Rex* Web service of Veale & Li [21], which models the category structure of concepts by harvesting fine-grained categorization statements on the Web. *MetaphorMagnet* thus suggests e.g., that priests are stereotypically pious and compassionate, while *Metaphor Eyes* indicates that they lead religious services and deliver sermons, while *Thesaurus Rex* reveals that they are seen as *religious leaders* and *trusted individuals*.

Consider the contrast framed in this tweet:

To some students, learning is a rewarding investment. To others, it is an unrewarding chore.
#Learning= #Investment #Learning= #Chore

@MetaphorMagnet contrives to bisociate two contrasting views on *learning* here, each of which is derived from Google n-grams (the 4-grams "*learning is an investment*" and "*learning is a chore*"). The knowledge that investments are typically rewarding while chores are typically unrewarding is provided by the *MetaphorMagnet* Web service. The resulting contrast of rewarding : unrewarding is then framed in this tweet not as a single metaphor but as a clash of two figurative world-views.

How does a contrast rise to the level of irony, so it can be seen as a failure of reasonable expectations? *@MetaphorMagnet* uses a gambit favored by many Twitter users: it affixes an *#irony* tag:

#Irony: *When the remembrances that are facilitated by moving memorials are facilitated by fixed monuments*. #Moving= #Fixed #Remembrance

Here the system searches its knowledge-base to identify possible scenarios (e.g. involving memorials *and* remembrances) that can give rise to striking contrasts. The *#irony* tag encourages readers to view this not-uncommon scenario as an instance of situational irony, and thus question whether all memorials are *moving*, or indeed if all monuments are truly *fixed*. To subject a specific part of a contrast to ironic scorn, *@MetaphorMagnet* uses yet another proven gambit on Twitter, "scare" quotes:

#Irony: *When some chefs prepare "fresh" salads the way apothecaries prepare noxious poisons*. #Chef= #Apothecary #Salad= #Poison

An analogy to the preparation of dangerous chemicals encourages readers to doubt the freshness of the stereotypically *healthy* salad options on menus. The targeted use of scare quotes means that it is the freshness of the salad, rather than the noxiousness of the poison, that is drawn into question here. To generate bisociative scenarios, *@MetaphorMagnet* looks for contrasts that are hidden in plain sight, and which only emerge when two otherwise banal facts are connected in unassuming ways, as in:

#Irony: *When some activists promote "enduring" principles the way originators promote temporary fads*. #Activist= #Originator #Principle= #Fad

So when tweets about people (activists, chefs, etc.) are framed as ironic insights, the irony may carry an added charge of hypocrisy. Scare quotes add to this air of deliberate pretense [22], and, as some principles are not truly enduring, they are little better than passing fads. Wit is commonly used to puncture such pretensions, and irony is one of its sharpest and most precise tools.

IV. Experiments

We performed three experiments with human raters, in order to study the contribution of the irony indicators, described in the previous sections, to the ironic effect. In particular, we investigated the role of contrastive expressions (i.e. adjectives and comparisons with opposite polarity) as irony factors, and the complementary role of scare quotes and `#irony` tag as irony markers.

In the first experiment, we used irony recognition as an explicit measurement of the ironic effect. In the second experiment, we studied the effect of specific words on valence and polarity, and its correlation with irony recognition. As reported below, the results show that both valence shifting and polarity inversion are good indicators of irony perception. Finally, in the third experiment we used only valence shifting and polarity inversion as indirect measurement of the ironic effect.

In all experiments, we used CrowdFlower as a web platform to recruit subjects and collect their judgments[3]. Each experiment consists of one or more CrowdFlower tasks.

A. Experiment 1

The first experiment had exploratory character. We started from two observations: (1) The *@MetaphorMagnet* tweets meant to be ironic contain a pair of *contrastive adjectives* (i.e. adjectives with opposite polarity), and (2) the ironic tweets are generated according to different formats, some of which seem more effective.

In order to measure the effect of an ironic tweet on subjects, we considered four dimensions for the ironic effect: *irony perception*, *surprise*, *humor appreciation*, and *retweetability* (i.e., willingness to retweet it).

Then we stated the following research questions:

Q1: Do the contrastive adjectives significantly increase the ironic effect (according to each of the above dimension)?

Q2: Is there a significant difference in the ironic effect induced by different tweet formats?

Q3: Is there a correlation between the dimensions of the ironic effect?

We used *@MetaphorMagnet* to automatically generate ironic tweets in three formats. An example of tweet in each format is shown below. For each of them, the pairs of contrastive adjectives are emphasized in bold.

- *#Irony: When the thief that fences **vibrant** jewels is disguised with **lifeless** masks. #VibrantOrLifeless #Thief*
- *#Irony: When some athletes love "**vigorous**" sports the way mommas love **weak** infants. #AthleteOrMomma #SportOrInfant*
- *#Irony: **Beautiful** poets composing poems about **horrid** monsters. #BeautifulOrHorrid #PoemAboutMonster*

We randomly picked 80 tweets for each format. Then, we removed contrastive adjective from a half subset of tweets in each format. Finally, we removed the hash tags and expressed each tweet as an English statement.

In summary, the experiment is focused on six settings, each represented by a different set of statements. Table I shows an example for each experimental setting. We collected an overall randomized data set of 240 tweets.

[3] `http://crowdflower.com` – retrieved 13 February 2015.

	With Contrastive Adjectives (Adj)	Without Contrastive Adjectives (NoAdj)
Format 1	*Kindergartens are educating skinny kids about fat babies.*	[Rational] *sciences are discovering truth about* [irrational] *religion.*
Format 2	*The thief that fences vibrant jewels is disguised with lifeless masks.*	*The vows that are made by* [evangelical] *crusaders are made by the* [shyest] *nuns.*
Format 3	*Some astrologers study "beloved" stars the way entomologists study ugly spiders.*	*Some choreographers manage* ["free"] *dancers the way madams manage* [enslaved] *prostitutes.*

TABLE I: Examples of items for each experimental condition (Experiment 1).

Condition	Irony	Surprise	Humor	Retweet
Format 1	2.43	2.69	2.36	1.56
Format 2	2.53	2.76	2.45	1.59
Format 3	2.69	2.80	2.62	1.66
NoAdj	2.51	2.71	2.46	1.59
Adj	2.59	2.78	2.50	1.62

TABLE II: Average values of different factors for each experimental condition.

Comparison	Irony	Surprise	Humor	Retweet
Format 2 < Format 3	**0.16** $\mathbf{p < .001}$	0.04 $p > .005$	**0.17** $\mathbf{p < .001}$	**0.07** $\mathbf{p < .01}$
Format 1 < Format 2	**0.1** $\mathbf{p < .005}$	**0.07** $\mathbf{p < .002}$	**0.09** $\mathbf{p < .001}$	**0.03** $\mathbf{p < .05}$
Format 1 < Format 3	**0.26** $\mathbf{p < .001}$	**0.11** $\mathbf{p < .001}$	**0.26** $\mathbf{p < .001}$	**0.10** $\mathbf{p < .001}$
NoAdj < Adj	**0.08** $\mathbf{p < .002}$	**0.07** $\mathbf{p < .01}$	0.04 $p > .005$	0.03 $p > .005$

TABLE III: Differences of average values and corresponding p-values. The values in bold represent successful tests.

The task proposed to human raters consists of the evaluation of a group of 8 units (i.e. tweets), randomly collected from the dataset. Each unit was judged by 20 different subjects. A number of questions were asked about each tweet. The first one required to select the topic of the tweet from a list of 4 words. This question was used as test questions in order to identify scammers. If a subject failed to answer correctly more than three times to this question, all her judgments were removed[4]. The other questions were used to test the four factors introduced above: *irony*, *surprise*, *humor*, *retweetability*. They were measured as numeric variables (the first three assuming values between 1 and 5, and the last one with values between 0 and 2).

The CrowdFlower judgments were collected running four tasks, in order to achieve sufficient statistical power. As output, we obtained a total of 495 trusted subjects and 6482 trusted judgments. Table II reports the average values of each factor used to measure irony in different settings. In particular, the first rows show the averages for each subset of items corresponding to the three irony formats. The next two rows show the averages for the items with contrastive adjectives and the items without contrastive adjectives, respectively.

In order to check if the increase of the averages reported in Table II are statistically significant, we performed a set of hypothesis tests. We calculated the averages of each factor according to the subjects and focused on their dispersion along the sets of items. In particular, we applied the Wilcoxon Sum Rank Test to a group of condition pairs for each factor.

With the notation *dim(A) < dim(B)* we mean the test performed to check if the average of the factor "dim" increases from the condition A to the condition B.

With the notation $A < B$ we mean the set of tests performed to the two conditions A and B in all factors. The group of alternative hypotheses are:

1) Format 1 < Format 2
2) Format 2 < Format 3
3) Format 1 < Format 3
4) NoAdj < Adj

The first research question is represented by hypotheses 1-3 while the second research question is expressed as 4). Clearly, in cases when both 1) and 2) are confirmed, there is no need to test 3). The results of the test are shown in Table III.

To address the third research question, we performed measurements of correlation between the four factors. For each pair of factors, we calculated the Pearson's correlation coefficient. As a result, all factors results two-by-two significantly correlated ($p < 0.001$). More specifically, irony, surprise, and humor show a strong positive mutual correlation. Moreover, each of them shows a moderate positive correlation with retweetability.

In summary, the results provide a positive answer to all research questions. Specifically, the results show that (1) the contrastive adjectives are effective to induce surprise, and (2) the most effective format is the one using scare quotes as irony markers.

B. Experiment 2

We designed a second experiment aimed to measure the contribution of semantic opposition and irony markers. In this experiment, we extend the focus from the pair of contrastive adjectives to the contrastive comparison containing them. For example, last row of Table I shows the contrastive adjectives *rational* and *irrational* as part of the contrastive comparison *rational sciences discovering truth about irrational religion*.

Furthermore, we focus on a specific word contained in the ironic tweets generated by *@MetaphorMagnet*. It is the adjective in first member of the contrastive comparison, and characterized by a definite (typically positive) polarity. We call this adjective *focus word*. As a measurement of ironic effect to

[4]Some scammers might not have detected through this simple method. However, scammers typically give either random rates or the same values and do not contribute to the variation of the averages. They might have reduced the statistical power without affecting the correctness of the results.

Setting	Example
BASE	*The vegetables are mixed in healthful salads.*
QUOT	*The vegetables that are mixed in "healthful" salads.*
HASH	*#Irony: The vegetables are mixed in healthful salads.*
QUOT+COMP	*The vegetables that are mixed in "healthful" salads are treated with poisonous pesticides.*
QUOT+HASH	*#Irony: The vegetables that are mixed in "healthful" salads are treated with poisonous pesticides.*

TABLE IV: Examples of items for each experimental condition (Experiments 2 and 3).

be performed on the focus word, we consider valence shifting and a "shallow" form of polarity opposition. We hypothesize that irony indicators such as contrastive comparisons and scare quotes can alter the valence of the focus word and, in particular, tend to invert its polarity.

We define *valence shifting* of a focus word as the variation of its average valence according to two different experimental settings. The *polarity rate* of a focus word is defined as the rate of subjects judging it as positive. Then, we call *shallow polarity inversion* of a focus word the variation of its average polarity rate according to two different experimental settings. The word 'shallow' characterizes it as a weak form of polarity inversion. Indeed, it could occur without change the average polarity of the majority of subjects. However, it allows us to extract different and, to some extent, complementary information respect to valence shifting. Finally, we will use the term *valence effect* to indicate either valence shifting or shallow polarity inversion.

Using the above definition, we stated the following research questions:

Q4: Are contrastive comparisons and scare quotes capable of inducing a significant valence effect in the focus word?

Q5: Is valence effect correlated to irony perception?

We arranged the dataset of statements[5] according to four conditions labeled as: *No Contrastive Comparison* (BASE), *Scare Quotes* (QUOT), *Contrastive Comparison* (COMP+QUOT), and *Contrastive Comparison + Scare Quotes* (COMP+HASH).

The judgment of valence about the focus word was provided as answer to three different questions. Firstly, subjects are required to read the word in isolation and judge it as either positive or negative. The second question is about the specific valence to attribute to the focus word. There are six possible values, between -3 to $+3$. Zero value was not considered. Finally, the third question is about the valence of the focus word inside the sentence. The reason for the first question is to be able to use gold items and, thus, detect possible scammers. If the subject selects the wrong polarity, it can be easily noticed as an error. Moreover, if answers to Question 1 and Questions 2 show opposite polarity, the corresponding judgment is removed as not trustworthy.

[5]In both Experiment 1 and Experiment 2, we performed the randomized selection of items as described for Experiment 1.

Setting	Valence	Polarity Rate
BASE	0.51 ± 0.38	0.91 ± 0.15
QUOT	0.41 ± 0.46	0.82 ± 0.13
COMP	0.29 ± 0.49	0.75 ± 0.15
QUOT+COMP	0.20 ± 0.54	0.64 ± 0.16

TABLE V: Mean valence and polarity rate, and their standard deviations, over texts under different experimental conditions (Experiment 2). The interval of values is normalized from $(-3, +3)$ to $(-1, +1)$.

	QUOT	COMP	QUOT+COMP
BASE	-0.10	-0.22	-0.31
QUOT	–	-0.12	-0.21
COMP	–	–	-0.09

TABLE VI: Values of valence shifting from the BASE to each other setting. Each value is associated to p-value $< .001$ (Experiment 2).

As a measure of valence of focus word in isolation and the same word in the sentence, we calculate the average values provided to Question 2 and Question 3, respectively. As a measure of shallow polarity inversion, we calculate the rate of judgments in which values provided to Question 1 and Question 2 have opposite polarities. The CrowdFlower task produced a total of 73 trusted subjects and 4000 trusted judgments. Each item was judged by 10 different subjects. Before applying the statistical tests, we removed the words recognized as negative in Questions 1 and Question 2. Most of them have no clearly recognizable polarity. Moreover, in current version of *@MetaphorMagnet* the ironic tweets use positive concepts as focus word.

We can summarize the empirical results as follows. According to Table V, Table VI and Table VII, irony indicators (i.e. contrastive comparisons and scare quotes), used as generative parameters, are capable of inducing the valence effect. In particular, the first row of Table VII shows that the contribution of QUOT and COMP is cumulated in QUOT+COMP. To test if valence shifting and shallow polarity inversion are correlated to irony perception, we applied the Wilcoxon Rank Sum test to the two subsets of data, in each condition, corresponding to "Yes" and "No" judgments on irony, respectively.

The results show that, in all conditions in which at least one irony parameter is applied (i.e. QUOT, COMP, and

	QUOT	COMP	QUOT+COMP
BASE	-0.09	-0.16	$\mathbf{-0.27}$
QUOT	–	-0.07	-0.18
COMP	–	–	-0.11

TABLE VII: Values of shallow polarity inversion from the BASE to each other setting. P-value for test comparing QUOT and COMP is $< .005$. P-value for every other comparison is $< .001$ (Experiment 2).

Setting	Valence	Polarity Rate
COMP	0.11 ± 0.60	0.52 ± 0.29
QUOT	-0.07 ± 0.60	0.35 ± 0.26
HASH	0.06 ± 0.59	0.50 ± 0.27
QUOT+HASH	-0.05 ± 0.61	0.35 ± 0.25

TABLE VIII: Mean valence and polarity rate, and their standard deviations, over texts under different experimental conditions (Experiment 3). The interval of values is normalised from $(-3, +3)$ to $(-1, +1)$.

	QUOT	HASH	QUOT+HASH
COMP	$\mathbf{-0.18}$ $\mathbf{p < .001}$	-0.05 $p > .05$	$\mathbf{-0.16}$ $\mathbf{p < .001}$
COMP	$\mathbf{-0.17}$ $\mathbf{p < .001}$	-0.02 $p > .05$	$\mathbf{-0.17}$ $\mathbf{p < .001}$

TABLE IX: Values of valence shifting (first row) and shallow polarity inversion (second row) from the COMP setting. Each value is associated to its p-value.

QUOT+COMP), both average valence and polarity rate are significantly lower ($p < .001$) in the case of statements recognized as ironic[6]. Instead, in the BASE condition, correctly, there is no significant variation of either average valence or polarity rate between ironic and non-ironic judgments. In other words, irony recognition can affect valence and polarity only in the case of irony factors.

Surprisingly, valence shifting and shallow polarity inversion occur also when the statements are not recognized as ironic (settings COMP and QUOT+COMP). This implies that irony perception is a moderator (and not a mediator) of valence effect, which can occur, to a smaller degree, even without the recognition of irony.

C. Experiment 3

In order to investigate the capability to use the valence effect as an indirect way to measure irony, we performed a new experiment in which irony was not mentioned and the subjects were asked to rate valence. We focused on the following research question:

Q6: Are scare quotes and hashtag *#irony* capable of inducing a significant valence effect in the focus word, even when irony is not explicitly mentioned?

In order to limit the expense for the experiment, we used the settings with the contrastive comparison as baseline and tested the valence shifting produced by the hashtag and the scare quotes, either in isolation or together. The experimental conditions are labeled as: *Contrastive Comparison* (COMP), *Contrastive Comparison + Scare Quotes* (QUOT), *Contrastive Comparison + Hashtag #irony* (HASH), and *Contrastive Comparison + Scare Quotes* (QUOT+HASH).

The CrowdFlower task, run with similar modality as Experiment 3 (i.e. 10 judgments per tweet) and scammer filtering produced a total of 1813 trusted judgments rated by 33 subjects. As in the previous two experiments, we applied Wilcoxon Rank Sum test to compare different settings. The results are shown in Table VIII and Table IX.

According to the results, the hashtag *#irony* induces a tiny valence shifting but, unexpectedly, it is not statistically significant ($p > 0.3$). The same result is achieved in the case of setting with both markers (i.e. scare quotes and *#irony* tag). On the other hand, the effect of scare quotes is statistically significant ($p < .001$).

[6]In both cases we applied Bonferroni correction for multiple testing

V. Conclusions

The results of the experiments described in the previous section show that contrastive comparisons and scare quotes, implemented in *@MetaphorMagnet* as generative parameters, are capable of achieving subtle forms of verbal irony. We aim to generalize these results and the experimental methodology used to produce them, to a larger class of irony factors and irony markers.

We identified a particular form of semantic slanting, called *valence effect*, induced on words that are typically recognized as positive. It is capable of modifying the intensity of their valence and to invert their polarity in a statistically significant way. We discovered that the valence effect can be used to perform an indirect measure of verbal irony in *@MetaphorMagnet* tweets. Specifically, this approach allows us to test the ironic effectiveness of hashtags used *#irony* tag as irony marker.

The empirical results show that we cannot make a tweet recognizable as ironic by simply adding the hashtag *#irony*. Instead, we need to employ a more subtle combination of features. The valence effect allows us to identify contrastive comparisons and scare quotes as more effective as ironic features. Furthermore, they can be combined to cumulate the effect induced by each of them. According to our data analysis of Experiment 3, the combined use of these features induced a negative valence shifting to 89% of target words.

An implication of the valence effect is its capability to disambiguate verbal irony from situational irony. Specifically, the valence shifting induced on target words shows that contrastive comparisons are not only capable to evoke incongruity (recognizable as situational irony). Indeed, they can also modify and, to some degree, subvert the meaning of words, achieving a form of verbal irony.

Another potentially interesting consequence is that irony indicators can be used as a way to modify the valence and, in particular, the polarity of words in a controlled way. To the best of our knowledge, this is the first time that a computational system is able to modify the sentiment of a word through the controlled use of irony indicators in the sentence context. We believe that this finding is step towards a stronger connection between sentiment analysis and methods for the automatic detection and generation of verbal irony.

Acknowledgments

This research was supported by the EC project *WHIM: The What-If Machine*. See http://www.whim-project.eu.

REFERENCES

[1] Yanfen Hao and Tony Veale, "An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes," *Minds and Machines*, vol. 20, no. 4, pp. 635–650, 2010.

[2] Antonio Reyes, Paolo Rosso, and Tony Veale, "A multidimensional approach for detecting irony in twitter," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[3] Alistair Kennedy and Diana Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[4] Tony Veale, "A service-oriented architecture for metaphor processing," in *Proceedings of the Second Workshop on Metaphor in NLP, at ACL 2014, the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA, 26 June 2014, pp. 52–60.

[5] Deirdre Wilson, "The pragmatics of verbal irony: Echo or pretence?," *Lingua*, vol. 116, pp. 1722–1743, 2006.

[6] Herbert Paul Grice, *Logic and Conversation. William James Lectures. Reprinted in: Grice, H.P., 1989, pp. 1-143*, Harvard University, 1967.

[7] Herbert Paul Grice, *Studies in the Way of Words*, Harvard University Press, Cambridge, MA, 1989.

[8] Dan Sperber and Deirdre Wilson, "Irony and the use-mention distinction," in *Radical pragmatics*, P. Cole, Ed., pp. 295–318. Academic Press, New York, 1981.

[9] Salvatore Attardo, "Irony markers and functions: Towards a goal-oriented theory of irony and its processing," *Rask - Internationalt tidsskrift for sprog og kommunikation*, vol. 12, pp. 3–20, 2000.

[10] Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira, "Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-)," in *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (TSA '09)*, Hong Kong, China, 2009, pp. 53–56, ACM.

[11] Mary Gardiner and Mark Dras, "Valence shifting: Is it a valid task?," in *Proceedings of Australasian Language Technology Association Workshop*, Dunedin, New Zealand, 2012, p. 4251.

[12] Antonio Moreno Ortiz, Chantal Pérez Hernández, and M. Ángeles Del-Olmo, "Managing multiword expressions in a lexicon-based sentiment analysis system for spanish," in *Proceedings of the 9th Workshop on Multiword Expressions (MWE)*, 2013.

[13] Luis Sarmento, Paula Carvalho, Mario Silva, and Eugenio de Oliveira, "Automatic creation of a reference corpus for political opinion mining in user-generated content," in *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, 2009, pp. 29–36.

[14] Livia Polanyi and Annie Zaenen, "Contextual valence shifters," in *Computing Attitude and Affect in Text: Theory and Applications*, Yan Qu James G. Shanahan, James G. Shanahan, Yan Qu, and Janyce Wiebe, Eds., vol. 20 of *The Information Retrieval Series*, pp. 1–10. Springer Netherlands, 2006.

[15] František Simančík and Mark Lee, "A ccg-based system for valence shifting for sentiment analysis," in *Advances in Computational Linguistics: Proceedings of 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009)*, Mexico City, Mexico, March 1-7 2009.

[16] Marco Guerini, Carlo Strapparava, and Oliviero Stock, "Valentino: A tool for valence shifting of natural language text," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2008, pp. 243–246.

[17] Arthur Koestler, *The Act of Creation*, Penguin Books, 1963.

[18] Tony Veale, "A service-oriented architecture for computational creativity," *Journal of Computing Science and Engineering*, vol. 7, no. 3, pp. 159–167, 2013.

[19] Thorsten Brants and Alex Franz, "Web 1t 5-gram ver. 1," Philadelphia: Linguistic Data Consortium, 2006.

[20] Tony Veale and Guofu Li, "Creative introspection and knowledge acquisition: Learning about the world thru introspective questions and exploratory metaphors," in *Proceedings of AAAI'2011, the 25th Conference of the Association for the Advancement of Artificial Intelligence*, 2011.

[21] Tony Veale and Guofu Li, "Creating similarity: Lateral thinking for vertical similarity judgments," in *Proceedings of ACL 2013, the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 4-9 August 2013.

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