Polarity Contrast for the Detection of Verbal Irony

Alessandro Valitutti, Nicole Novielli

University of Bari
{alessandro.valitutti, nicole.novielli}@uniba.it

Abstract. In this paper, we propose two metrics capable of modeling forms of polarity contrast: *polarity divergence* and *polarity dimorphism*. To explore their potential usefulness to the detection of verbal irony and sentiment polarity, we performed an exploratory text analysis on a corpus of figurative tweets annotated by sentiment. The results of a text analysis show that, employing two different types of valenced lexicon, we can improve performance in polarity classification.

1 Introduction

One of the most challenging issues in sentiment analysis is the classification of sarcastic texts. In this context, the term *sarcasm* indicates a type of verbal irony where the polarity of the literal meaning (or *literal polarity*) is positive, and the polarity of the intended meaning (or *ironic polarity*) is negative. The capability to recognize sarcasm (and, more generally, verbal irony) is necessary to avoid the misclassification induced by considering literal polarity instead of ironic polarity. During the last few years, a good amount of research on sarcasm detection has been focused on tweets, since the features of verbal irony are concentrated on a relatively short text and are more likely to be modeled [14]. Initial studies took into account features such as specific punctuation, exclamations, hashtags (e.g., *#sarcasm* or *#yeahright*, or emoticons) (see [4] for a broader overview). The emotional meaning expressed in the text and emotional categories were used to extract features relevant to irony and sarcasm [12]. Other works focused instead on more subtle features representing polarity contrast between different parts of the text, and relying on the distinction between different types of valenced lexicon [9].

In the present research, we wonder to what degree it is possible to detect verbal irony and retrieve ironic polarity from a single sentence, using only information about literal polarity. To address this challenge, we combined some of the intuitions discussed by Riloff et al. [9] about the contrast between sentiment and situation, and the distinction between *irony markers* and *irony factors* proposed by Attardo [1]. While irony markers are simply meta-communicative clues telling the reader that the text containing them is ironic, irony factors additionally provide operational information for identifying the ironic meaning.

Therefore, we defined two metrics, capable of modeling two corresponding types of polarity contrast, and meant to improve irony detection and sentiment

analysis of sentences with verbal irony. They are called *polarity divergence* and *polarity dimorphism*. Polarity divergence is defined as the difference between the maximum positive and maximum negative sentiment strength. By contrast, polarity dimorphism is based on the assumption that most ironic sentences can be separated into two parts: one referring some topic or situation with either a positive or a negative "stereotypical" polarity, and the other one expressing the author's sentiment about the topic. The proposed metrics are, on the one hand, general enough to be applied to different types of verbally-ironic texts, since they do not depend on a specific syntactic pattern. On the other hand, they are specific enough to separate texts with verbal irony from other types of texts.

To explore the potential usefulness of the proposed metrics to the detection of verbal irony and sentiment polarity, we performed a case study on a corpus of figurative tweets annotated by sentiment and provided as part of the SemEval-2015 Task 11 [5] on sentiment analysis of figurative tweets. The results of the text analysis show that employing two types of valenced lexicon (i.e., emotion words and non-emotion sentiment words), we can improve performance in polarity classification

2 Polarity Contrast and Related Metrics

Attardo [1] discussed the distinction between *irony markers* and *irony factors*. Irony factors are meta-communicative clues. They tell the reader that the text presented to them is ironic. However, they indicate verbal irony but do not necessarily show how to retrieve the ironic meaning. On the other hand, irony factors are the real "ingredients" of verbal irony. They provide operational instructions for identifying the ironic meaning.

For example, in the following tweet:

> *Love watching news stories about plane issues while waiting at the airport #sarcasm*

we have two irony markers: the hashtag *#sarcasm* and the polarity contrast between '*Love*' and '*watching news stories about plane issues while waiting*'. However, polarity contrast additionally indicates that the ironic polarity can be obtained by reversing the polarity of '*Love*', so also playing the role of irony factor.

Over the last years, several markers of irony and sarcasm have been identified, such as interjections or scare quotes (see for example [3] and [8]). Nevertheless, almost all of them can be classified, according to the Attardo's distinction as "irony markers" and none of them as "irony factors". In other words, these features indicate that some text is ironic, but they are not sufficient to provide operational information to extract the ironic polarity. For instance, if some hashtags such as "#notreally" or "#yeahright" are contained in a tweet, they indicate that there is verbal irony, but cannot tell us where exactly polarity reversal occurs.

In this research, we focus on polarity contrast expressed in the tweet content, thus deprived of markers such as the above hashtags. In other words, we do not consider the contrast between the ironic hashtags and the remaining text. The advantage to define metrics representing this "internal" polarity contrast is the possibility to apply them to a more general class of ironic texts.

Previous studies on irony and sarcasm analysis have been centered types of semantic contrast. For instance, Karoui et al. [7] claim that in all ironic and sarcastic tweets there is a contradiction between two phrases or words. Moreover, they distinguish between *explicit activation* (when the incongruity is internal to the tweet text) and *implicit activation* (when the contrast occurs between the tweet content and some background context). According to this terminology, we focus on the explicit activation.

We defined two metrics representing polarity contrast: *polarity divergence* and *polarity dimorphism*. Polarity divergence measures the rate of polarity opposition in the text. It is obtained calculating the sum of the absolute values of positive and negative sentiment. For example, a text with null positive and negative sentiment has null polarity divergence, while a text with +1 positive polarity and -2 negative polarity will return 3 as polarity divergence. To implement polarity divergence, we used the lexicon underlying *SentiStrength* [13] to calculate sentiment polarity. We hypothesize that both verbal irony and situational irony correspond to high values of polarity divergence. In other words, we typically have mixed polarity in both cases, but with a different function. In the case of situational irony, mixed polarity expresses a contrast of situations with opposite polarity. In the case of verbal irony, the conflict of polarities is used as a clue that polarity reversal is occurring in a portion of the text.

A specific type of polarity divergence is what seems to distinguish verbal irony from situational irony. A good number of verbally ironic sentences are structured in such a way to express a positive evaluation on a (typically) negative topic or, less often, a negative evaluation about a positive topic. For example:

I just love when students don't do their homework!

He's as nice as a lion to his prey.

In the above sentences, we can distinguish between the negative topic and the positive evaluation. We refer to the polarity expressed in the evaluation by the author of the sentence as *evaluative polarity* and denotes the polarity typically attributed, in the common-sense knowledge, to the topic, as *stereotypical polarity*. Therefore, we call *polarity dimorphism* the polarity divergence between the evaluative polarity and the stereotypical polarity. Unlike polarity divergence, for the implementation of polarity dimorphism we employed two different lexicons. To measure evaluative polarity, we used *WordNet-Affect* [11]. To detect stereotypical polarity, we used words included in SentiStrength but not in WordNet-Affect.

3 Ironic Tweets and Text Analysis

To evaluate the potential usefulness of the proposed metrics for improving polarity classification of verbally-ironic texts, we carried out an exploratory text analysis on a collection of tweets previously annotated by sentiment. We employed the test set provided at the *SemEval-2015 Task 11* on *Sentiment Analysis of Figurative Tweets* [5]. The dataset consists of a list of tweet IDs, each annotated with a value of positive valence (also known as *sentiment strength*) and a value of negative valence. For privacy reasons, the text of the tweets was not published. Using the Twitter APIs, we retrieved about 6000 tweets. Next, we filtered tweets where the text is followed by one or more "irony hashtags", that is a list of hashtags we assumed to be clues of verbal irony (e.g. `#sarcasm`, `#sarcastic`, `#irony`, `#ironic`, `#yeahright`, `#not`, etc.). Moreover, we extracted the last sentence of the tweet text, assuming that it could likely be a verbally ironic sentence. The filtering was performed automatically and returned 3927 items, thus obtaining the dataset we used for the text analysis.

The text analysis was aimed to test if the two metrics (i.e., polarity divergence and polarity dimorphism) can improve the performance of an available sentiment polarity classifier. We used SentiStrength as baseline classifier and implemented two simple modified versions based on polarity divergence and polarity dimorphism, respectively. The two metrics-based classifiers are defined as follows: given an input sentence, SentiStrength is applied, and the metric is calculated. Next, if the value of the feature is non-null, then the sentence polarity is assumed to be the opposite of the overall polarity by SentiStrength (i.e., the algebraic sum of the positive and negative strength divided by 4 to normalize it). The reason is that a non-null value of the metric is interpreted as the occurrence of polarity reversal.

Table 1 reports the results of the evaluation. For each of the three classifiers (i.e., the baseline and the metrics-based ones), we calculated recall, precision, and F-score for recognition of positive polarity (first three columns), recognition of negative polarity (next three columns), and general case (i.e. means of corresponding scores for the positive and negative case – last three columns). The results confirm that polarity dimorphism is the metric that behaves better. In particular, it gives an F-score outperforming the corresponding values of the other two classifiers. In particular, in negative classification polarity dimorphism produces an impressive increase of performance since it outperforms baseline recall by 42% and F-score by 34%.

4 Conclusions and Future Work

In this research, we explored the central role of polarity contrast in the characterization of verbal irony. Specifically, we defined two metrics for measuring polarity contrast at two different levels of granularity. The first metric – *polarity divergence* – is meant to represent the degree of polarity contrast in the more general way. The second metric – *polarity dimorphism* – employs the further distinction

	Positive			Negative			Mean		
	R	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
SentiStrength	0.29	0.03	0.05	0.18	0.95	0.30	0.23	0.49	0.17
Pol. Diver. Rate	0.78	0.04	0.07	0.17	0.95	0.29	0.48	0.49	0.18
Pol. Dimor. Rate	0.78	0.04	**0.07**	0.60	0.95	**0.74**	0.69	0.49	**0.41**

Table 1. Comparison between SentiStrength and classifiers using polarity divergence and polarity dimorphism.

between *evaluation polarity* and *stereotypical polarity*. Ironic utterances express, in most cases, an evaluation by the author about a target topic (e.g., a situation, a person, or an event). They play with the stereotypical polarity attributed to the topic in the common-sense knowledge, through the use of evaluation with opposite polarity. This contraposition induces a violation of readers' expectation and, according to the context, it achieves humorous or sarcastic effects.

The proposed approach can be summarized in the following points:

- The Attardo's distinction [1] between irony markers and irony factors identifies two different communicative functions: 1) the clue that the text is ironic/sarcastic, and 2) the operational information needed to locate polarity reversal and identify the ironic polarity.
- Polarity contrast is a central irony factor, which can be easily implemented through the use of sentiment lexicons. Our definition of polarity divergence provides a way to measure the degree of polarity contrast.
- We introduced the distinction between two types of polarity: *evaluative polarity* and *stereotypical polarity*. Accordingly, we defined polarity dimorphism as a particular type of polarity divergence.

The evaluation results show that the detection of literal polarity is an effective way to detect irony polarity. One implication is that available sentiment analyzers and lexicons originally developed for detecting literal polarity can be reused to detect ironic polarity. However, they should be enriched with the capability of separating the portion of text where polarity inversion occurs and, whenever possible, distinguishing evaluative and stereotypical polarity.

As next steps, we aim to extend the approach described in this paper to a larger variety of resources (such as the ones described in [6]). In particular, we will compare the effect of different sentiment lexicons for the measurement of evaluation and stereotypical polarity, and replicate the evaluation on more datasets of tweets annotated according to irony, sarcasm, and sentiment. Increasing the degree of granularity, we will consider other types of polarity contrast, such as *polypathy* (which occurs between different senses of the same word [2]), previously used in the automatic detection of verbal humor [10]. Finally, we will combine the proposed metrics to the features already employed in past works and study their performance with different classifiers.

References

1. Attardo, S.: Irony markers and functions: Towards a goal-oriented theory of irony and its processing. Rask - Internationalt tidsskrift for sprog og kommunikation 12, 3–20 (2000)
2. Basile, V., Nissim, M.: Sentiment analysis on italian tweets. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 100–107. Atlanta, United States (2013)
3. Carvalho, P., Sarmento, L., Silva, M.J., de Oliveira, E.: 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). pp. 53–56. ACM, Hong Kong, China (2009)
4. Farias, D.H., Rosso, P.: Irony, sarcasm, and sentiment analysis. In: Pozzi, F.A., Fersini, E., Messina, E., Liu, B. (eds.) Sentiment Analysis in Social Networks, chap. 7. Elsevier (2017)
5. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: SemEval-2015 Task 11: Sentiment analysis of figurative language in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado (4-5 June 2015)
6. Hernańdez Farías, D.I., Patti, V., Rosso, P.: Irony detection in twitter: The role of affective content. ACM Transactions on Internet Technology 16(3), 1–24 (2016)
7. Karoui, J., Benamara, F., Moriceau, V., Patti, V., Bosco, C., Aussenac-Gilles, N.: Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL17). vol. 1, pp. 262–272. Valencia, Spain (April 3-7 2017)
8. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language Resources and Evaluation 47(1), 239–268 (2013)
9. Riloff, E., Qadir, A., Surve, P., Silva, L.D., Gilbert, N., Huang, R.: Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). pp. 704–714. Seattle (18-21 October 2013)
10. Smigaj, A., Kovalerchuk, B.: Visualizing incongruity and resolution: Visual data mining strategies for modeling sequential humor containing shifts of interpretation. In: Proceedings of the 19th International Conference on Human-Computer Interaction (HCI International). Springer, Vancouver, Canada (9-14 July 2017)
11. Strapparava, C., Valitutti, A.: WordNet-Affect: an affective extension of WordNet. In: Proc. of 4^{th} International Conference on Language Resources and Evaluation (LREC 2004). Lisbon (May 2004)
12. Sulis, E., Hernández Farías, D.I., Rosso, P., Patti, V., Ruffo, G.: Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. Knowledge-Based Systems 108, 132–143 (2016)
13. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)
14. Wang, A.: #irony or #sarcasm – a quantitative and qualitative study based on twitter. In: Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation, pp. 349–356. National Chengchi University (2013)