UFPelRules to irony detection in Spanish variants

Diulia Justin Deon[1] and Larissa Astrogildo de Freitas[1]

UFPel, Gomes Carneiro 1, Porto, 96010-610, Rio Grande do Sul, Pelotas, Brazil
`djdeon,larissa@inf.ufpel.edu.br`

Abstract. Figurative language is one of the most difficult challenge to Natural Language Processing. In this study, we propose a strategy to irony detection in Spanish based in linguistic patterns. After studying the state of the art, we implement seven linguistic patterns in 9000 texts written in three different linguistic variants.

Keywords: Irony detection · Spanish variants · Patterns

1 Introduction

Usually, irony is known as a way of communicating the opposite of the literal meaning [3]. Commonly known as, irony is a figure of speech that seeks to express a word or text with distinct meaning from the original.

The detection of ironic declarations represent a major challenge to Natural Language Processing. In this study, we analyzed and implemented some linguistic patterns that may be associated with ironic declarations in Spanish language varieties.

Irony can be seen as a complex communication mechanism that is governed by pragmatic principles. Observing the uses of the term, it is observed that irony is often mistaken for sarcasm, satire or parody. In this study, we based the task of irony detection in a general concept for this phenomenon, since there are no consensuses on a rigid definition of irony. [8] and [4] define irony as an apparent violation of pragmatic principles in an utterance. [1], on the other side, define irony as a contradictory property in a given context or event. [7] states that the presence of irony conveys a pragmatic meaning when alluding to expectations (failures or not).

The elaboration of patterns involving possible evidence of ironic declarations considers the following elements: syntactic rules, static expressions, lists of laughter expressions, specific scores, and symbolic language. The implementation of the proposed patterns is based on the work of [6], and [3].

Six linguistic patterns classified into three groups were implemented: (C1) lists; (C2) exact expressions; and (C3) symbols (Table 1). Finally, we applied these patterns in 9000 texts and analysed the results.

2   Related Works

The detection of irony has been increasingly worked and explored by the scientific community.

The use of linguistic patterns to detect irony has already been used by other researchers, such as [3] and [6].

[3] proposed 13 linguistic patterns to detect irony in a corpus of tweets in Brazilian Portuguese about End of the World.

[6] proposed an investigation from a corpus of comments. In this study, the authors concluded that the most productive patterns for the detection of irony are related to the use of punctuation marks, emoticons and quotes.

3   Linguistic Patterns

List of patterns were built in the three varieties of the Spanish language that are present in the corpus: Mexican, Spanish and Cuban [5], [2]. In this study six linguistic patterns were implemented for all varieties. They are:

Table 1. List of patterns.

Categories	Patterns	Expression
C1	P1	List of Laughter Expression
C2	P2	"no te creas","no es cierto", "es broma"
C3	P3	...<EXPR>!
	P4	Uppercase
	P5	* | * | !* | ?* | ** | ** | !*?* | ?*!*
	P6	Quotations Marks

P1- List of Laughter Expression: Social network's users, like twitters, usually use expressions like "jajaja" and "hehehe" to express that they are kidding or to reverse the utterance's meaning. For example, "Jajajajajajajajaja tendencioso, a mi primero que me explique la falta de capacidad de sus candidatos y luego hablamos de conflicto de intereses.".

P2 - "No te creas", "No es cierto", "Es broma": In Brazilian Portuguese, people use the expression "Só que não" to add a negative meaning to all that said before. In Spanish, we found three equivalent expressions: "No te creas", "Es broma" and "No es cierto". For example,"Pues no te creas que es tan jovencita.".

P3 - ...<EXPR>! : Expressions accompanied by an exclamation point after an ellipsis are considered ironies, because the punctuation (...) indicates that the

phrase is over. However, when such a pattern is used, it would be interpreted as an attempt by the author to add something, finish his sentence, which is supposed to be an irony. For example, "Ojalá Carmen Calvo saliendo a decir 'Pero no os habéis dado cuenta? Todo este tiempo... el relator... eras tú!' y se va volando agarrada a un paraguas.".

P4 - Uppercase: To give greater intonation and prominence to the texts, users choose to use words with uppercase formatting in text points, or all of it, indicating that his speech would change, producing the feeling that this is an irony. For example: "LA TV DIGITAL ESTA SIENDO MUY EFICIENTE.".

P5 - * | * | !* | ?* | ** | ** | !*?* | ?*!* : People end up not using more final scores in sentences, and in the case of Twitter, they use even less because the network counts and only allows tweets with a certain number of characters. Analyzing this, it is noted that the score is no longer used in the application in question. Thus, when it is present in some phrase in a large number followed, called repetition, it brings indicators that the phrase contains irony. For example: "entonces la tv cubana PAGA los derechos? !VIVA EL pqt!!!!!!!!!!!!!!!!!!!!!!".

P6 - Quotations Marks: Quotes are used to give prominence, show that it would be something similar or derived from, causing the reader a sense of insecurity referring to what is being read, and thus indicating, that it may be an irony, a phrase using figured sense. For example: "Entendi 'perfectamente' , ms claro no pudieron ser. Que cosa es servicio de mensajeria ??".

4 Training set

In this work, we used the training set, available in IroSvA. This database contains 7200 texts with the following annotation: identification code (ID), topic (text subject), classification of ironic or non-ironic.

UFPelRules classified 964 Cuba texts, 864 Mexico texts and 749 Spain texts as ironic (Table 2a, 2b and 2c) on the training set.

There were 2400 texts for each spanish variation, of these 800 were ironic. UFPelRules was able to identify 964 text with irony in the Cuban variety (Table 2a), 1463 in the Mexican variety (Table 2b) and 743 in the Spanish variety (Table 2c). Comparing the results of UFPelRules and the training set, the tool identified and marked more texts as ironic in Table 2a and 2b, and less in Table 2c, in relation to those predefined by the training set.

These results demonstrate that the tool can detect patterns, with half of the marked texts being identified as originally ironic. It was also observed that the P4 pattern was the one that most identified irony, achieving expressive numbers in the three variations, standing out among the six linguistic patterns. This proves that in most cases, Spanish speakers use this language feature to express irony.

Table 2. Confusion Matrix with Training Set.

	Real Class		
Predicted Class	Not Ironic	Ironic	Score
Not Ironic	966	634	1600
Ironic	470	330	800
Score	1436	964	2400

(a) **Cuba.**

	Real Class		
Predicted Class	Not Ironic	Ironic	Score
Not Ironic	1017	583	1600
Ironic	519	281	800
Score	1536	864	2400

(b) **Mexico.**

	Real Class		
Predicted Class	Not Ironic	Ironic	Score
Not Ironic	1094	506	1600
Ironic	557	243	800
Score	1651	749	2400

(c) **Spain.**

5   Test set

After the analysis with the training set, we applied the patterns in test set. This database contains: code identifier (ID) and the message to be analyzed. This time, the texts were not annotated for ironic or non-ironic. The results are shown in the Table 3.

The P4 repeated the result of the training set, and identified irony with a great prominence in relation to the others, being able to identify a large group of texts in the three variations. Second with the best result was the P6 and the third was the P5.

Although P3 does not present a significant result, we still believe that − given the intuition of language users − that it would be possible to find better results in other.

Table 3 presents the final average obtained in the analyzes of the test set, as well as the average of other resources used by the authors of Irony Corpus [5],[2].

Table 3. F1-score with Test Set.

	ES	MX	CU	AVG
LDSE	0.6795	0.6608	0.6335	0.6579
W2V	0.6823	0.6271	0.6033	0.6376
Word nGrams	0.6696	0.6196	0.5684	0.6192
Our Approach	**0.5088**	**0.5464**	**0.5620**	**0.5391**
MAJORITY	0.4000	0.4000	0.4000	0.4000

6   Final remarks

During this work, it was noted that irony is expressed in complex ways, and in similar ways even with different languages. Although intelligent methods are used for this analysis, irony is shown as a major challenge.

Interpreting, understanding and detecting social networks requires multi-analysis on linguistic concepts and understanding of how people use virtual communication. It is concluded that the linguistic patterns established by the study offer satisfactory results, in addition to allowing the detection in more than one language, demonstrating a high similarity between the structures of different languages.

Based on observing language behavior in other corpora, it is possible that new patterns can be added to the UFPelRules, as well as some existing ones can be excluded if they provide irrelevant results.

References

1. Antonio Reyes, Paolo Rosso, D.B.: From humor recognition to irony detection: The figurative language of social media. Data Knowledge Engineering 74, 1–12 (Apr 2012)
2. Francisco Rangel, Paolo Rosso, M.F.S.: A low dimensionality representation for language variety identification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624), pp. 156-169 (2018)
3. Larissa A. de Freitas, Aline A. Vanin, D.N.H.M.N.B.R.V.: Pathways for irony detection in tweets. In: Proceedings of the 29th Annual ACM Symposium on Applied Computing. pp. 628–633. SAC '14, ACM, New York, NY, USA (2014)
4. Grice, H.P.: Logic and conversation. Academic Press (1975)
5. Ortega-Bueno, R., Rangel, F., Hernández Farías, D.I., Rosso, P., Montes-y-Gómez, M., Medina Pagola, J.E.: Overview of the Task on Irony Detection in Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS.org (2019)
6. Paula Carvalho, Luís Sarmento, M.J.S.E.d.O.: 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. pp. 53–56. TSA '09, ACM, New York, NY, USA (2009)
7. R. J. Kreuz, S.G.: How to be sarcastic: The choice reminder theory of verbal irony. Journal of Experimental Psychology: General 118, 374–386 (1989)
8. Searle, J.R.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press (1969)