Building Corpora for Figurative Language Processing: The Case of Irony Detection

Antonio Reyes[1 2], **Paolo Rosso**[2]

[1] Language Technology Lab
Instituto Superior de Intérpretes y Traductores, Mexico
antonioreyes@isit.edu.mx
[2] Natural Language Engineering Lab — ELiRF
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract

Figurative language is one of the most arduous topics that natural language processing (NLP) has to face. Unlike literal language, the former takes advantage of linguistic devices, such as metaphor, analogy, ambiguity, irony, sarcasm, and so on, in order to communicate more complex meanings, which usually represent a serious problem, not only for computers, but for humans as well. In this article we describe the problem of figurative language processing concerning corpus-based approaches. This type of language is quite common in web contents; however, its automatic processing entails a huge challenge, both theoretically as pragmatically. Here we describe the need of automatically building training corpora with objective and reliable data. In this respect, we are focused on addressing a quite complex device: irony. Such linguistic phenomenon, which is widespread in web content, has important implications for tasks such as sentiment analysis, opinion mining, or even advertising.

1. Introduction

Language, in all its forms, is the most natural and important mean of conveying information. However, given its social nature, it cannot be conceptualized only in terms of grammatical issues. In this respect, while it is true that grammar regulates language in order to have a non-chaotic system, it is also true that language is dynamic, and accordingly, a live entity. This means that language is not static, rather it is in constant interaction between the rules of its grammar and its pragmatic use. For instance, the idiom *"all of a sudden"* has a grammatical structure which is not made intelligible only by knowledge of the familiar rules of its grammar (Fillmore et al., 1988), but by inferring pragmatic information as well. This latter provides the knowledge that, in the end, gives sense to the idiom.

Emphasizing the social aspect of language, modern linguists deem language as a continuum of symbolic structures in which lexicon, morphology, and syntax form a continuum which differs along various parameters but can be divided into separate components only arbitrarily (Langacker, 1991). Language, thus, is viewed as an entity whose components and levels of analysis cannot be independent nor isolated. On the contrary, they are embedded in a global system that depends on cognitive, experiential, and social contexts, which go far beyond the linguistic system proper (Kemmer, 2010). Let us consider the following example:

1. "I really need some antifreeze in me on cold days like this".

Example 1 is fully understandable only within a context in which the sense is given by figuring out the analogy between $antifreeze$ (referential knowledge: antifreeze is a liquid) and liquor (inferential knowledge: antifreeze is a liquid, liquor is a liquid, antifreeze is a liquor).

In this context, the following sections introduce the theoretical background concerning figurative language (Section 2.), describe the problem of dealing with figurative language in a technological framework (Section 3.), report on how we approach the task of automatically building a training corpus for figurative language processing (Section 4.), and conclude with some final remarks about our approach and its further implications (Section 5.).

2. Literal and Figurative Language

Traditionally, language has been described from dichotomous points of view: *langue* $vs.$ *parole*, signifier $vs.$ signified, synchrony $vs.$ diachrony, paradigmatic $vs.$ syntagmatic, oral $vs.$ written, an so on. In this section, another dichotomy will be discussed: literal language $vs.$ figurative language. The simplest definition of literal language is related to the notion of $true$, $exact$ or $real$ meaning; i.e. a word (isolated or within a context) conveys one single meaning (the one conventionally accepted), which cannot be deviated. In Saussure's terms, literal meaning is corresponded with a perfect dichotomy of signifier and signified (cf. (de Saussure, 1974)). Some experts, in addition, have noticed certain properties of literal meaning: it is direct, grammatically specified, sentential, necessary, and context-free (see (Katz, 1980; Searle, 1978; Dascal, 1987)). Hence, it is assumed that it must be invariant in all contexts. For instance, the word $flower$ can only refer to the concept of plant, regardless of its use in different communicative acts or discourses (e.g. botany, evolution, poetry).

On the other hand, figurative language could be regarded as the simple oppositeness of literal language. Thus, whereas the latter is assumed to communicate a direct meaning, the former is more related to the notion of conveying indirect or veiled meanings. For instance, the word $flower$, which literally refers only to the concept of plant, speaking figuratively can refer to several concepts, which not necessarily are linked to plants. Thereby, it can be used instead of concepts such as beauty, peace, purity, life, and so on, in such a way its literal meaning is *intentionally* deviated in favor of secondary interpretations[1].

Although, at first glance, this distinction seems to be clear and sufficient on its own, figurative language involves basic cognitive processes rather than only deviant usage (Peters, 2004). Therefore, it is necessary going deeper into the mechanisms and processes that differentiate both types of languages.

In accordance with classical perspectives, the notions of literalness and figurativity are viewed as pertaining directly to language; i.e. words have literal meanings, and can be used figuratively (Katz, 1980; Searle, 1978; Dascal, 1987). Figurative language, therefore, could be regarded as a type of language that is based on literal meaning, but is disconnected from what people learn about the world [or about the words] based on it [them] (Bergen, 2005). Thus, by breaking this link, literal meaning loses its primary referent and, accordingly, the interpretation process becomes senseless. Let us consider Chomsky's famous example to explain this issue:

2. "Colorless green ideas sleep furiously" (Chomsky, 1957).

Beyond grammatical aspects, in example 2 is possible to observe that, either phonologically or orthographically, Chomsky's example is fully understandable in terms of its linguistic constituents[2]. However, when interpreting such constituents in context, its literal meaning is completely nonsensical. For instance, the bigrams [colorless green] or [green ideas] are sufficiently disconnected from their conventional referents for being able to produce a coherent interpretation. Thus, in order to make the example understandable, secondary interpretations are needed. If such interpretations are successfully activated, then figurative meaning is triggered[3] and, accordingly, a more coherent interpretation can be achieved. Based on this explanation, literal meaning could be deemed as denotative, whereas figurative meaning, connotative; i.e. figurative meaning is not given a priori, rather, must be implicated. Furthermore, in (Lönneker-Rodman and Narayanan, 2008), authors point out that figurative language can tap into conceptual and linguistic knowledge (as in the case of idioms, metaphor, and some metonymies), as well as evoke pragmatic factors in interpretation (as in indirect speech acts, humor, irony, or sarcasm). In accordance with the assumptions, an expected conclusion is to conceive the processes of interpreting figurative language much more complex than the ones performed when interpreting literal language.

3. Figurative Language and Web Content

Web-based technologies have become a significant source of data in a variety of scientific and humanistic fields. Such technologies provide a rich vein of information that is easily mined. User-generated content (such as text, audio and images) provides knowledge that is topical, task-specific, and dynamically updated to broadly reflect changing trends, behavior patterns and social preferences. In this context, figurative language can be found on almost every web site in a variety of guises and with varying degrees of obviousness. For instance, when analyzing instances of irony, one of the most important micro-blogger sites: Twitter, allows its users to self annotate their posts with user-generated tags (or hashtags according to Twitter's terminology). Thus, the hashtag #irony is used by people in order to self-annotate all varieties of irony, whether they are chiefly the results of deliberate word-play or merely observations of the humor inherent in everyday situations (e.g. "Sitting in the eye-doctor's office, waiting for the doctor to see me"), or simply sarcastic expressions (e.g. "I thank God that you are unique!").

3.1. The Core of the Problem

Although the arguments given in the previous sections provide some elements to determine what figurative language is, a major question still remains: how to differentiate between literal language and figurative language (theoretically and automatically)? The examples given so far have shown some of their main characteristics; however, based on that information, there is not way of totally affirming that example 1 is more figurative than example 2. Finally, both examples could be expressing, either of literal or figurative language. To be able to provide arguments for differentiating both linguistic realities, a crucial extra-linguistic element (with linguistic repercussion) must be highlighted: **intentionality**. Beyond mechanisms to explain why figurative language requires much more cognitive efforts to correctly interpret its meaning, the most important issue is that the previous examples are simply sequences of words with semantic meaning. Perhaps, such meaning is very clear (literalness), or perhaps is senseless (figurativity), but they could be explained in terms of performance and competence or even as a matter of correctness . However, such difference could be motivated by the need of maximizing a communicative success (cf. (Sperber and Wilson, 2002)). Such need would be then the element that will determine what type of information has to be profiled. If a literal meaning is profiled, then certain intention will permeate the statement. This intention will find a linguistic repercussion by selecting some words or syntactic structures to successfully communicate what is intended. In contrast, if the figurative meaning is profiled, then the intention will guide the choice of others elements to ensure the right transmission of its content. It is likely that such content cannot be ac-

[1] It is worth noting that such secondary interpretations are not guaranteed. Their success will depend on several factors, both linguistic as extra-linguistic.

[2] It is worth stressing that this sentence is an intentional example of semantic senseless, whose meaning (either literal or figurative) is supposed to not exist. However, here is used to precisely exemplify the nonsensical effect produced by figurative contents. Most of them, finally, are senseless on their own, and need a pragmatic anchor to correctly interpret their meanings.

[3] According to (Sikos et al., 2008), understanding figurative language often involves an interpretive adjustment to individual words; i.e. not all the constituents of the example trigger a figurative meaning on their own, rather, this is usually triggered by manipulating individual words.

complished, but in this case, the failure will not lay on the speaker's intention, rather, on the hearer's skills to interpret what is communicated figuratively. Let us observe the following sentences to clarify this point.

3. "The rainbow is an arc of colored light in the sky caused by refraction of the sun's rays by rain" (cf. WordNet (Miller, 1995) v. 3.0).
4. "The rainbow is a promise in the sky".

Whereas in example 3 the intention is to describe what a rainbow is, in example 4 the intention is to communicate a veiled meaning, motivated and understandable by a specific conceptual context. In each statement the speaker has a communicative need, which is solved by maximizing certain elements. Thus, in the first example, the communicative success is based on making a precise description of a rainbow (note that all the words in this context are very clear in terms of their semantic meaning), whereas in the second, is based on deliberately selecting elements that entail secondary and nonliteral relations: [rainbow - promise], [promise - sky].

4. User-Generated Tags: Explicit Intentionality

Once argued that intentionality is one of the most important mechanisms to differentiate literal from figurative language, it is worth noting that user-generated tags provide specific elements to deliberately express different types of figurative contents: metaphor, allegory, irony, similes, analogy, and so on. In this respect, we are focused on the case of irony.

Irony (and most figurative language) is very subjective and often depends on personal appreciation[4]. Therefore, the task of collecting ironic examples (positive data) is quite challenging. In addition, as noted in (Reyes and Rosso, 2011; Reyes et al., 2012), the boundaries to differentiate the different types of irony (mostly verbal irony and situational irony) are very fuzzy indeed: non-expert people usually use an intuitive and unspoken definition of irony rather than one sanctioned by a dictionary or a text-book. Hence, such task becomes any harder.

4.1. A Basic Sample

Although a manual annotation is supposed to be the best way of obtaining reliable information in corpus-based approaches, in tasks like this one, such approach is hard to be achieved. First, there are not formal elements to accurately determine the necessary components to label any text as ironic. Then, in the case that we had a prototype of ironic expressions, its discovery is a time-consuming manual task[5]. Finally, linguistic competence, personal appreciation, moods, and so on, make irony quite subjective; therefore, any annotation agreement faces the complexity of standardizing annotation criteria. That is why we decided to use examples labeled with user-generated tags, which are **intentionally** focused on particular topics[6]. By opting for this approach, we eliminate the inconveniences above mentioned: such examples are self-annotated (thus, it is not necessary the presence of "human annotators" to manually (and subjectively) collect and label positive examples). In addition, positive examples can be retrieved effortless taking advantage of their tags (thus, it is likely having thousands of examples in a short time).

In this context, we here describe how we have taken advantage of the user-generated tags in order to build a training corpus for the irony detection task. To this end, we are focused on one of the current trendsetters in social media: the Twitter micro-blogging service. We first determine a membership criterion for including a tweet in the corpus: each should contain a specific $hashtag$ (i.e. the user-generated tag according to Twitter's terminology). The hashtags selected are #irony, in which a tweet explicitly declares its ironic nature, as well as #education, #humor, and #politics, to provide a large sample of potentially non-ironic tweets. These hashtags are selected because when using the #irony hashtag, people employ (or suggest) a family-resemblance model of what it means (cognitively and socially) for a text to be ironic. In this respect, a text so-tagged may not actually be ironic by any dictionary definition of irony, but the tag reflects a tacit belief about what constitutes irony. Based on these criteria, we collect a training corpus of 40,000 tweets, which is divided into four parts, comprising one self-described positive set and three other sets that are not so tagged, and thus assumed to be negative. The final corpus contains 10,000 ironic tweets and 30,000 largely non-ironic tweets. Some statistics are given in Table 1. It is worth noting that all the hashtags were removed. No further preprocessing was applied at this point.

Table 1: Statistics in terms of tokens per set.

	#irony	#education	#humor	#politics
Vocabulary	147,671	138,056	151,050	141,680
Nouns	54,738	52,024	53,308	57,550
Adjectives	9,964	7,750	10,206	6,773
Verbs	29,034	18,097	21,964	16,439
Adverbs	9,064	3,719	6,543	4,669

Due to the intrinsic characteristics concerning writing habits in technological platforms such as blogs, cell phones, etc., it is very likely the presence of many errors in the documents, as well as the presence of duplicate documents, or even pointless information. In order to minimize such errors, several measures can be applied. Here we outline just one of them: the Jaccard distance. Such metric measures the dissimilarity between two samples, and is calculated according to Formula 1. The Jaccard distance is here used to estimate the overlap between the ironic set and each of the three non-ironic ones. In addition, it should help mini-

[4] That is why the importance of considering both linguistic as paralinguistic features when modelling this complex device.

[5] According to (Peters and Wilks, 2003), this is a reason for the restricted number of attested instances of figurative language in texts. In addition, it is worth noting that irony appears quite often in discourse. For instance, in (Carvalho et al., 2011), authors indicate that irony is present in approximately 11% of their data.

[6] Recall the role of intentionality in the process of communicating the figurative intent.

mizing the likelihood of noise arising from the presence of typos, common misspellings, and the abbreviations that are endemic to short texts.

$$J_{\delta}(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (1)$$

Results in Table 2 suggest a significant difference between the vocabularies of the four tweet sets. As one might expect, this difference is least pronounced between sets #irony and #humor. After all, irony is most often used to communicate a humorous attitude or insight, as in examples 5 and 6, in which the tweet was tagged as #irony:

Table 2: Jaccard distance among sets.

	J_{δ}(A,B)
(irony, education)	0.8233
(irony, humor)	0.8565
(irony, politics)	0.8246

5. Just think: every time I breathe a man dies. —A friend: Have you tried to do something about bad breath?

6. I find it humorously hypocritical that Jeep advertises on TV about how we shouldn't watch tv in favor of driving their vehicles.

Finally, it is worth noting that this approach is useful to the spread of researches related to figurative language, as well as to palliate the lack of resources for figurative language processing, and especially, to face tasks in which the scarcity of data, the subjectivity of the task, or the impossibility of making personal interviews, are challenges to be tackled[7].

5. Final Remarks

In this article we have discussed the problem of figurative language and its automatic processing. In particular, we were focused on addressing the task of automatically building training corpora when facing one of the most complex figurative devices: irony. Although the approach here described is slightly theoretical, it has important implications for tasks such as sentiment analysis (cf. (Reyes et al., 2012) about the importance of determining the presence of irony in order to assign fine-grained polarity levels), trend discovery (cf. (Reyes and Rosso, 2011; Reyes and Rosso, In press), where authors note the impact of user-generated tags for discovering people's trends in ironic documents), or opinion mining (cf. (Sarmento et al., 2009), about the role of irony in discriminating negative from positive opinions). In the future, we plan to approach irony detection from each of its angles building corpora that could consider also valuable information such as gestural information, tone, paralinguistic cues, etc. (cf. (Cornejol et al., 2007)). Last but not least, it would be also interesting try to model irony taking into consideration the visual stimulus of brains responses when people have to process ironic statements (cf. (Mars et al., 2008)).

[7] The relevance of approaches like this one can be confronted in (Reyes and Rosso, 2011): in such work authors collected a corpus for irony detection only with reviews posted in Amazon.

Acknowledgements

This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+I.

6. References

B. Bergen. 2005. Mental Simulation in Literal and Figurative Language Understanding. In Seana Coulson, editor, *The Literal And Nonliteral in Language and Thought*, pages 255–280. Peter Lang Publishing, September.

P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 564–568.

N. Chomsky. 1957. *Syntactic Structures*. Mouton and Co, The Hague.

C. Cornejol, F. Simonetti, N. Aldunate, A. Ibáñez, V. López, and L. Melloni. 2007. Electrophysiological evidence of different interpretative strategies in irony comprehension. *Journal of Psycholinguist Research*, 36:411–430.

M. Dascal. 1987. Defending literal meaning. *Cognitive Science*, 11(3):259–281.

F. de Saussure. 1974. *Course in general linguistics*. Fontana, London.

C. Fillmore, P. Kay, and M. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

J. Katz. 1980. *Propositional structure and illocutionary force: A study of the contribution of sentence meaning to speech acts*. Harvard University Press.

S. Kemmer. 2010. About cognitive linguistics: Historical background. `http://www.cognitivelinguistics.org/cl.shtml`. Online on August 25, 2011.

R. Langacker. 1991. *Concept, Image and Symbol. The Cognitive Basis of Grammar*. Mounton de Gruyter.

B. Lönneker-Rodman and S. Narayanan. 2008. Computational approaches to figurative language.

R. Mars, B. Rogier, S. Debener, T. Gladwin, L. Harrison, P. Haggard, J. Rothwell, and S. Bestmann. 2008. Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise. *J. Neurosci.*, 28(47):12539–12545.

G. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

W. Peters and Y. Wilks. 2003. Data-driven detection of figurative language use in electronic language resources. *Metaphor and Symbol*, 18(3):161–173.

W. Peters. 2004. *Detection and Characterization of Figurative Language Use in WordNet*. Ph.D. thesis, University of Sheffield, Sheffield, England.

A. Reyes and P. Rosso. 2011. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 118–124.

A. Reyes and P. Rosso. In press. Making objective decisions from subjective data: Detecting irony in customers reviews. *Decision Support Systems*.

A. Reyes, P. Rosso, and D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*. DOI: 10.1016/j.datak.2012.02.005 `http://dx.doi.org/10.1016/j.datak.2012.02.005`.

L. Sarmento, P. Carvalho, M. Silva, and E. de Oliveira. 2009. Automatic creation of a reference corpus for political opinion mining in user-generated content. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 29–36.

J. R. Searle. 1978. Literal meaning. *Erkenntnis*, 13(1):207 – 224.

L. Sikos, S. Windisch Brown, A. Kim, L. Michaelis, and M. Palmer. 2008. Figurative language: “meaning" is often more than just a sum of the parts. In *Proceedings of the AAAI 2008 Fall Symposium on Biologically Inspired Cognitive Architectures.*, pages 180–185.

D. Sperber and D. Wilson. 2002. Relevance theory. *Handbook of Pragmatics*, 42(5):607–632.