OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is a publisher’s version published in: https://oatao.univ-toulouse.fr/22188

Official URL:
https://doi.org/10.1016/j.procs.2017.10.105

To cite this version:

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr
1. Introduction

Irony is a complex linguistic phenomenon widely studied in philosophy and linguistics [17, 31, 34]. Despite theories differ on how to define irony, they all commonly agree that it involves an incongruity between the literal meaning of an utterance and what is expected about the speaker and/or the environment. For example, to express a negative opinion towards a cell phone, one can either use a literal form using a negative opinion word, as in "What an excellent phone!!", or a non-literal form by using a positive word, as in "This phone is a disaster, or a non-literal form by using a positive word, as in What an excellent phone!!.

For many researchers, irony overlaps with a variety of other figurative devices such as satire, parody, and sarcasm [9, 15]. In computational linguistics, irony is often used as an umbrella term that includes sarcasm, although some researchers make a distinction between irony and sarcasm, considering that sarcasm tends to be harsher, humiliating, degrading, and more aggressive [10, 22]. In this paper, we use irony as an umbrella term that covers satire, parody and sarcasm.

Irony detection is quite a hot topic in the research community also due to its importance for efficient sentiment analysis [13, 19, 24]. Indeed, although current systems achieve relatively good results on objective vs. subjective classification, polarity analysis still need further improvement to address implicit polarity reversal. Examples (1) and (2) from Twitter illustrate this phenomena. In (1), the author ironically employs several positive opinion words ("character of the year", "perfect") towards Morsi, the former president of Egypt to express
a clearly negative opinion. The tweet (2) is also negative towards Salim Cheboub, daughter’s husband of the former Tunisian president Ben Ali, although there are any negative opinion words. The author here describes an ironic situation due to a sudden change of topic (Ali Ben Salem’s Uncle vs. Salim Cheboub) and a false assertion (“Salim Cheboub says that he has no relationship with Ben Ali” is not true in reality).

(Morsi will be named the character of the year for 2012 believe we are ungrateful this man is perfect #sarcasm)

(While Ali Ben Salem’s Uncle is crying on Tunisia 7, Salim Cheboub says that he has no relationship with Ben Ali on #Ettounisia #Ironically #peaceful #Tunisia)

As shown in the previous examples, in social media such as Twitter, users tend to utilize specific hashtags (#irony, #sarcasm, #sarcastic) to help readers understand that their message is ironic. These hashtags are often used as gold labels to detect irony in a supervised learning setting (i.e., learning whether a text span is ironic/sarcastic or not). In doing so, systems are not able to detect irony without explicit hashtags, but on the positive side, it provides researchers with positive examples with high precision.

This binary classification task relies on a variety of features. Most of them are gleaned from the utterance internal context going from n-grams models, stylistic (punctuation, emoticons, quotations, etc.), to dictionary-based features (sentiment and affect dictionaries, slang languages, etc.). These features have shown to be useful to learn whether a text span is ironic/sarcastic or not [3, 5, 16, 32, 33]. Most related work concern English data. We note however some efforts to detect irony and sarcasm in French [20], Portuguese [7], Italian [14], Dutch [23] and Japanese [35]. In this paper, we focus for the first time on the automatic detection of irony in Arabic tweets.

In addition to misspellings and grammatical errors, detecting irony in Arabic tweets poses a significant challenge. Indeed, Arabic tweets are often characterized by non-diacerised texts, a large variations of unstandardized dialectal Arabic, and finally linguistic code switching between Modern Standard Arabic and several dialects (cf. example (1)), and between Arabic and other languages like English and French (cf. example (2)) [12, 25]. Due to the difficulty of the task and the lack of freely dedicated tools to process Arabic social media, we propose as a first step to detect irony relying on features that do not require any preprocessing step such as morpho-syntactic parsing. Our aim is three folds: (1) Test whether state of the art features whose efficiency have been empirically proved in other languages such as French, English, Italian, Dutch and Japanese, are also valid for Arabic, (2) Find among these features, the most productive ones, and (3) Analyze main source of errors.

This paper is organized as follows. The next section describes main existing work on irony in Arabic literature. Section 3 presents our corpus. Section 4 details our model for automatic detection of irony. We conclude this paper in Section 5.

2. Irony in Arabic texts

In the standard pragmatic model [17], irony is viewed as an apparent violation of the maxim of quality, stating that the speaker does not say what he believes to be false. In this model, when one ironically utters P, one conversationally implicates its opposite, that is $\neg(P)$. This vision has been criticized by several authors who pointed out that logical opposition between what is said and what is intended captures only one type of irony. To overcome this deficiency, different theories have been proposed to deal with the multi-dimensional nature of opposition. Among them, we cite [1, 9, 18, 31, 35] that respectively describe irony in terms of echoic mention, allusional pretense, predicate and propositional negations, relevant inappropriateness, and implicit display.

These theoretical models have been used as a basis to automatically detect irony and sarcasm in texts. As far as we know, no work has investigated irony detection in Arabic social media. As we aim to test whether irony markers used in irony detection in other languages (such as English) also work for Arabic, we focus in the remaining of this section

1 #Ettounisia is a Tunisian TV channel.
on linguistic studies that provide comparative analysis between irony expressions in English and Standard Arabic in written texts.

Among them, Chakhchiro [8] proposed a model for the analysis and translation of irony in political commentary texts from English into Arabic. This model, which relies on the presence of three groups of ironic devices in both languages (see below), shows that Arabic and English texts share several similarities in the rhetorical, grammatical and lexical use of devices, text strategies and rhetorical meaning. However, the differences are, as expected, striking at the textual realisation level. The devices are as follows:

- **lexical devices**: synonymy and near synonymy (as in (3) with the use of *our souls* and *inner-selves*), metonymy, binomials2, compound words, lexical choice and idiomatic expression.

  
  
  (3) 
  
  "...and we say that our souls and inner-selves could not suppress the crying laughter when they knew and heard that..." [8]

- **grammatical devices**: redundancy, free indirect speech, reflexive pronoun, conjunction of supposition, adjunct adverb, inversion

- **rhetorical devices**: parallel structure, overstatement, rhetorical questions3 (as in (4)) and stylistic placing.

  
  
  (4) 
  
  "...فَلَأَلْوِنْ وَحْدَهُ بِيَبَيِّ الْإِنْسَان...وَهَلَّ سَعَةُ الْحُبُّ تَعْنِي سَعَةَ الْمَعْرَفَةَ؟ تِسْقِيمَ الْمِدْلَالُ وَهُلْ أَذَا تُؤْلُي رَمَآءًَا أُحْسَابَ النَّاجِم؟" (Can man live by colour alone, and do big pockets (financial wealth) mean wide knowledge? Would justice become right (be served) if it was controlled by mine owners?) [8]

Sigar et al. [30] also studied the difference between literal and ironic expressions in both English and Arabic in texts extracted from books, articles, the internet and everyday situations. Four main conclusions resulted from the analysis of the data: (1) Irony is a universal phenomenon since it has been examined in English and Arabic as an example for its universality; (2) In most cases, ironic expressions are similar to each other in both languages, the differences fall in the usage of some utterances that are exclusively related to a certain culture whether English or Arabic; (3) Ironic expressions in Arabic are more figurative than those in English, and finally (4) Opposition, humour and exaggeration are the most common devices to form ironic expressions in both languages.

As there are many similarities between the irony expressions in English and Standard Arabic, the use of features commonly employed to detect irony in English data to learn ironic Arabic statements is thus linguistically motivated. We details our features in Section 4.

3. Data

The dataset used in this study is about political tweets. We made this choice because politics is among the most discussed and criticized topics in social networks. Our corpus was built as follows. We selected a set of politicians’ names like Hillary, Trump, Al-sissi, Mubarak and Morsi which were the subject of the US and Egyptian presidential elections. The politicians’ names are used as keywords to collect the tweets. Then, we selected ironic tweets containing the topic keywords and the translation of #irony and #sarcasm. We also selected non ironic tweets that contained only the keywords. We removed duplicates, retweets and tweets containing pictures which would need to be interpreted to understand the ironic content. Arabic irony hashtags (#استمرون, #تيمك, #مسخرة, #عفرية) are removed from the tweets for the following experiments.

This procedure resulted in a set of 5,479 tweets distributed as follows: 1,733 ironic tweets and 3,746 non-ironic tweets. The collected corpus consists of tweets written in standard Arabic, dialectal Arabic or a mix of standard and dialectal Arabic (most cases). Since Twitter does not distinguish between standard Arabic, dialectal Arabic and different dialects, we obtained a corpus of tweets that mixes many dialects, most of which are written in Egyptian, Syrian and Saudi dialect. Other dialects have been rarely used such as the Tunisian and Algerian dialects. The corpus will be made freely available for research purposes.

---

2 A lexical repetition in the form of near-synonymy is used aesthetically for emphasis.

3 A rhetorical question aims to ask a question in order to make a point rather than to elicit an answer.
4. A binary classifier for irony detection in tweets

4.1. Features set

We represent each tweet with a vector composed of four groups of features: surface, sentiment, shifter and internal context features. Most of features have been successfully used for irony detection in French [20], English [3, 6, 16, 27, 33], Dutch [23] and Japanese [35]. We made this choice to test the performance of related work features for the Arabic language and to verify whether some features are language independent. The features are as follows:

(1) Surface features: These are the typical surface features that can be irony markers. For example, the presence of opposing words in a tweet may express an incongruity between two propositions according to irony definition [17]. Similarly, quotations may indicated the presence of an echoic mention [31]. These features check for the presence or absence of:

- punctuation marks (such as ..., ?, !) [16],
- emoticons [6, 16]. To implement this feature, we used a lexicon of 681 emoticons which we have collected from Twitter,
- quotations [27, 33],
- opposition words such as لكن (but) and بالرغم من (although) [35]. To implement this feature, we used a French opposition word lexicon (23 entries)[28] that we translated into Arabic.
- sequence of exclamation or question marks [7],
- combination of both exclamation and question marks [6],
- discourse connectives that do not convey opposition [20], since we assume that non ironic tweets are likely to be more verbose. To implement this feature, we used a lexicon of 416 Arabic discourse connectives [21] such as لذاك (in addition), لذاك (so), etc.
- interjections [6, 16]. In Arabic ironic tweets, we noticed that interjections that express laugh such as #مغامرات_السعوديات مع قانون_ولي_الأمر_يا_نتاونا_أحمدوا_ربي_ليل_وهيار_هههههه #تاتير (#Suffering of Saudian women with guardian law. O Tunisian women, pray God day and night hhhhhh #irony #sarcasm)

We also take into account features that count the number of emoticons, laugh and the tweet length in words [33].

(2) Sentiment features: As irony is often used to criticize or make fun, we look for features that check for the presence of positive/negative opinion words [26]. We also account for features that count the number of positive and negative opinion words [3]. To get these features we tested several existing Arabic opinion lexicons:

1. The Arabic translation of Bing Liu’s Lexicon which contains 2,006 positive terms and 4,783 negative terms built by translating the Bing Liu Sentiment Lexicon into Arabic using Google Translate.
2. The Arabic translation of MPQA Subjectivity Lexicon obtained by translating the MPQA Sentiment Lexicon into Arabic using Google Translate. It contains 2,718 positive words, 570 neutral and 4,911 negative words.
3. The Arabic Emoticon Lexicon automatically built from tweets where each word is associated to an opinion score between -7 and 7. It contains respectively 22,962 and 20,342 positive and negative words.
4. The Arabic Hashtag Lexicon (dialectal) also built from tweets by crowdsourcing. It contains 11,941 positive terms and 8,179 negative terms.

The last two lexicons [29] were exploited in the 7th task of the SemEval’2016 campaign on determining sentiment intensity of English and Arabic phrases ⁵. Among these four lexicons, the best results were achieved by combining the last two which resulted in a lexicon of 22,239 negative words and 26,777 positive words after removing

⁴ The lexicons are available at http://saifmohammad.com/WebPages/ArabicSA.html
⁵ http://alt.qcri.org/semeval2016/task7/
duplicates. This was expected as the first two lexicons were extracted from review style data while the two others were automatically collected from Twitter. In addition, for the same entry, these lexicons provide morphological variations (e.g. عاصمة which means capital and عاصمة التي which means city) as well as dialect variations (e.g. متميز which means excellent). This is important since we do not rely on any morpho-syntactic tool to extract our features.

(3) Shifter features: They allow to detect the following phenomena:

- **False assertion**: It indicates that a proposition, fact or an assertion fails to make sense against the reality, as in an example (6) where the assertion (!) "do not contradicts the law of God") is not true. We focus in particular on false assertions that are trigged by negation words like لا (no), ليس (not) 6.

To extract this feature, we used a negation word lexicon (15 entries) built manually.

```
(What is the problem when 75,634 are stuck in some place, and they make a mistake?!
This, of course, does not contradict the law of God! IsThen #Morsi #irony #freedom #revolution)
```

- **Exaggeration**: Usually, irony is triggered when someone expresses an idea or a feeling with an exaggerated way using intensifier such as جدا (very), أيضا (too) [3, 23]. We used an intensifier lexicon (25 entries) translated from a set of intensifiers used for the French language [20] (cf. tweet (7)).

```
كلامون جدا أولاد الذين يلحونذاعن أنه في غيابكم أو تقصيكم سبأ من يُخذم مكانكم. نادرن فخافضو عليهما! #استهزا

(Very loyal to those who always hint that in your absence or shortcoming an other person will take your place.
They are rare so keep them #sarcasm)
```

- **Reported speech**: We hypothesize that ironic tweets are more likely to contain repeated statements said by someone else using reporting speech verbs such as قال (say), أعلن (announce), etc. (cf. tweet (8)) [20]. We used an Arabic reported speech verb lexicon (119 entries) [21].

```
مرسي قال لو مليون نمو ضد هستقبل #طهرية

(Morsi said if a million show up against me I will resign #irony)
```

(4) Contextual features. The utterance context is important to understand the ironic meaning of a statement. These features check the presence/absence of internal contextual clues such as the presence of personal pronouns (أنا (I), أيتها (he), and named entities (cf. tweet 9) [20].

```
لماذا أنت حزين ؟! #凭什么 #السياسة السيسى سيقضي على الآخوان #فرد كأنك انتطاحي #فخافض #المرسي

(Why are you sad ?! We don’t understand politics Sisi will eliminate #Islamists #chirilikeGrovel #sarcasm #farce #reduce)
```

For named entity lexicon (4,501 entries), we used the ANERGazet Gazetteers [4] that contains a collection of 3 Gazetteers: locations (2,181 entries), persons (2,309 entries) and organizations (403 entries), to which we added the named entities used to collect our data, such as مرسى (Morsi), كلينتون (Clinton), etc.

4.2. Experiments and results

For the tweet classification task, we used several classifiers under the Weka toolkit with standard parameters: Support Vector Machine (SMO), Naive Bayes (NB), Logistic Regression (Logistic), Linear Regression (Logistic simple), 6

---

6 In our corpus, around 16.5% of tweets contain negation clues.
Random Tree and Random Forest. Among these classifiers, the last one was the best. We then present the results of our experiments as given by Random Forest.

We trained the classifier with a balanced corpus of 1,733 ironic tweets and 1,733 non-ironic tweets. Since the size of the corpus is limited, we opt for two experiments: (1) A train/test configuration where we used 80% of the corpus for training and 20% for testing with an equal distribution between ironic (IR) and non ironic (NIR) instances; (2) A 10-cross validation configuration on a balanced subset. In both configurations, all surface features were used as baseline.

We first wanted to test the performance of each group of features. We have therefore built four Random Forest models, each one with a dedicated group of features. Table 1 gives the results. When testing using a 10-cross validation, the best groups in terms of accuracy are sentiment features (57.15%) and contextual features (65.37%). However, the features groups taken separately are not sufficient to classify the NIR and IR tweets.

### Table 1. Classification results of ironic (IR)/non-ironic (NIR) tweets obtained by Random Forest classifier in terms of Precision (P), Recall (R), F-score (F) and Accuracy (A) by group of features.

<table>
<thead>
<tr>
<th>Group of Features</th>
<th>10-Cross validation</th>
<th>All features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>0.566</td>
<td>0.669</td>
</tr>
<tr>
<td>NIR</td>
<td>0.530</td>
<td>0.605</td>
</tr>
<tr>
<td>Sentiment features</td>
<td>0.566</td>
<td>0.488</td>
</tr>
<tr>
<td>IR</td>
<td>0.510</td>
<td>0.616</td>
</tr>
<tr>
<td>NIR</td>
<td>0.482</td>
<td>0.502</td>
</tr>
<tr>
<td>Contextual features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>0.544</td>
<td>0.618</td>
</tr>
<tr>
<td>NIR</td>
<td>0.511</td>
<td>0.639</td>
</tr>
</tbody>
</table>

In the second experiment, we tested the performance of Random Forest when trained using all groups of features. The results presented in Table 2 show that our approach outperforms the baseline in both configurations. In particular, 10-Cross validation achieves the best results when using the four groups of features with an Accuracy = 72.36% and F-score = 72.70% for the ironic class and 72.10% for the non-ironic class.

### Table 2. Classification results of ironic (IR)/non-ironic (NIR) tweets obtained by Random Forest in terms of Precision (P), Recall (R), F-score (F) and Accuracy (A) using All features.

<table>
<thead>
<tr>
<th>Group of Features</th>
<th>Train/test</th>
<th>10-Cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>0.713</td>
<td>0.679</td>
</tr>
<tr>
<td>NIR</td>
<td>0.669</td>
<td>0.705</td>
</tr>
<tr>
<td>All surface features (baseline)</td>
<td></td>
<td>68.84</td>
</tr>
<tr>
<td>IR</td>
<td>0.566</td>
<td>0.488</td>
</tr>
<tr>
<td>NIR</td>
<td>0.550</td>
<td>0.626</td>
</tr>
<tr>
<td>All features</td>
<td></td>
<td>55.68</td>
</tr>
<tr>
<td>IR</td>
<td>0.729</td>
<td>0.713</td>
</tr>
<tr>
<td>NIR</td>
<td>0.718</td>
<td>0.728</td>
</tr>
</tbody>
</table>

In order to improve these results, we applied the GainRatio algorithm for features selection under the Weka toolkit. It allows to have the descending order of the most relevant features which will give the combination of the best features by considering their individual predictive capacity with the degree of redundancy between them. The GainRatio algorithm gave the combination of the following features: Number of emoticons, number of laugh, number of named entities, presence of exclamation marks, presence negation words, presence of named entities. However, when we run the classifier with this combination, the results did not improve.

Therefore, we tested another approach by learning the Random Forest classifier by adding features one by one (following the order given by the selection algorithms) in order to identify the most discriminant features and to have the features subset that maximizes the classifier performance. The results, presented in Table 3, show that the use of all features except reporting speech verbs slightly improves the classification task by 1.06% in terms of accuracy (Accuracy = 72.76% and F-score = 73% for the ironic class and 72.50% for the non-ironic class).

### Table 3. Classification results of ironic (IR)/non-ironic (NIR) tweets obtained by Random Forest in terms of Precision (P), Recall (R), F-score (F) and Accuracy (A) using All features except reporting speech verbs.

To conclude, although we have used a set of features most of which are surface features, the results for Arabic tweet classifications in ironic/non-ironic are very encouraging in comparison to those obtained for other languages. For ironic class, we obtained a precision of 72.4%, while for the French language, the precision is 93% [20], 30% for
Table 3. Classification results of ironic (IR)/non-ironic (NIR) tweets obtained by Random Forest classifier in terms of Precision (P), Recall (R), F-score (F) and Accuracy (A) using the best combination of features (All features except reporting speech verbs)

<table>
<thead>
<tr>
<th></th>
<th>Train/test</th>
<th>10-Cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>IR</td>
<td>0.723</td>
<td>0.696</td>
</tr>
<tr>
<td>NIR</td>
<td>0.709</td>
<td>0.736</td>
</tr>
</tbody>
</table>

4.3. Error analysis

An error analysis shows that classification errors are mainly due to two factors:

- **Absence of context**: This is the main cause of errors. Indeed, interpretation of misclassified tweets requires external contextual knowledge to tweets. For example, in (10), the reader has to know that the Egyptian President bought weapons and that he was accused to use them against his people (according to the author’s view, money belongs to the people).

- **Wrong pre-annotation**: The absence (presence) of #irony or #sarcasm in tweets containing ironic (non ironic) utterances, as the ironic tweet in (11) and the non ironic tweet in (12). These cases are a minority (around 3%) as it is for other languages like French and English where irony hashtags have been shown to be quite reliable as good inter-annotator agreements (kappa around 0.75) between annotators irony label and the reference irony hashtags have been reported (Karoui et al., 2015). We plan to fully analyze those cases in the future so that wrong pre-annotations will be corrected before making the corpus available.

5. Conclusions

For the first time, this paper proposed an approach to irony detection in a corpus of Arabic tweets. Our approach is supervised and rely on a group of four features, that have been used for detecting irony in other languages (English, French, Dutch and Japanese). We evaluate the performance of each group (among surface, sentiment, shifter and contextual features) and conclude that all features were important for our task. We achieved an accuracy of 72.36% which is good given the difficulty of processing Arabic social media texts and the lack of dedicated tools to deal with code switching.

This study is a first step towards an Arabic irony detection system. In the future, we plan to manually check the reliability of hashtags and include pragmatic features that help to infer the context needed to understand ironic reading. For example, recent studies have shown that features like the use of common vs. rare synonyms [3], discussion thread [2] and external sources of knowledge [36] are quite effective to infer irony in other languages. It would be interesting to investigate similar features for the Arabic language.

Acknowledgements

This work was partially funded by the French FUI (Fond d’Investissements d’Avenir de 2014) SparkinData project (http://spardindata.org/).
References