IDAT@FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets

Bilal Ghanem[1], Jihen Karoui[2], Farah Benamara[3], Véronique Moriceau[3], and Paolo Rosso[1]

[1] PRHLT Research Center, Universitat Politècnica de València, Spain
bigha@doctor.upv.es, prosso@dsic.upv.es
[2] AUSY R&D, Paris
jkaroui@ausy.fr
[3] IRIT-CNRS, Université de Toulouse, France
{farah.benamara,veronique.moriceau}@irit.fr

Abstract. This overview paper describes the first shared task on irony detection for the Arabic language. The task consists of a binary classification of tweets as ironic or not using a dataset composed of $5,030$ Arabic tweets about different political issues and events related to the Middle East and the Maghreb. Tweets in our dataset are written in Modern Standard Arabic but also in different Arabic language varieties including Egypt, Gulf, Levantine and Maghrebi dialects. Eighteen teams registered to the task among which ten submitted their runs. The methods of participants ranged from feature-based to neural networks using either classical machine learning techniques or ensemble methods. The best performing system achieved F-score value of 0.844, showing that classical feature-based models outperform the neural ones.

Keywords: Irony detection · Arabic language · Social media

1 Aims and Motivations

Irony is a complex linguistic phenomenon widely studied in philosophy and linguistics. In the standard pragmatic model [11], 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 $Not(P)$. For example, if one says to his colleague "Congratulation for your great presentation" after a disappointing talk. 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 [31, 6, 13, 2, 33] that respectively describe irony in terms of echoic mention, allusional pretense, predicate and propositional negations, relevant inappropriateness, and implicit display. Irony is used here as an umbrella term that covers a variety of other figurative devices such as satire, parody, and sarcasm [6, 10].

Irony detection has gained relevance recently, due to its importance in various NLP applications such as sentiment analysis, hate speech detection, author profiling, fake news detection, and crisis management (e.g., terrorist attacks, public disorder). For example, recent studies on irony show that the performances of sentiment analysis systems drastically decrease when applied to ironic texts [3, 8, 15, 34]. This is mainly due to the complexity of ironic contents that make use of figures of speech to convey non-literal meaning.

Most state of the art approaches to irony detection consider social media data and tweets in particular, as specific hashtags (#irony, #sarcasm) are often employed by users to help readers understand their ironic contents. These hashtags are used as gold labels to detect irony in a supervised learning setting. Most related work concern English [14] with some efforts in French [18], Portuguese [4], Italian [9], Dutch [22], Hindi [32] and Arabic [17]. Also, many shared tasks on irony have been proposed, such as SemEval 2018 task 3 for English [14], DEFT 2017 for French [3], IronITA@Evalita 2018 for Italian [5], and IroSvA@IberLEF-2019 for Spanish variants [28] (from Spain, Cuba and Mexico). As far as we know, this is the first shared task on irony for the Arabic language and will be a good opportunity to compare the performances of Arabic irony detection to those reported in recent shared tasks in other languages.

2 Processing Arabic Tweets: Main Challenges

Computational processing of the Arabic language has received a great attention in the literature for over a twenty years[4]. Several resources and tools have been built to deal with Arabic nonconcatenative morphology and Arabic syntax [24]. There is also a wide range of Arabic NLP (ANLP) applications including question answering [26], automatic translation [30] and sentiment analysis [21]. However, the field of ANLP is still very vacant at the layer of pragmatics. As far as we know, the sole effort towards Arabic irony detection was done by Karoui et al. [17] who proposed a supervised approach to detecting ironic tweets. The performance of several groups of features (like surface, sentiment, shifter and contextual features) have been assessed achieving an accuracy of 72.36% on a dataset composed of 3,466 tweets among which 50% were ironic.

Detecting irony in Arabic poses a significant challenge, as the Arabic language is mainly characterized by the lack of diacritics (dedicated letters to represent short vowels), complex agglutination, pro-drop structure, and free order word structure. For instance, [7] estimated that the average number of ambiguities for a token in Arabic can reach 19.2, compared to 2.3 in most other languages. Also, short vowels are not often explicitly marked in writing. Indeed, they are neither written in the Arabic handwriting of everyday use nor in general publications.

[4] For a detailed description of Modern Standard Arabic and an overview of Arabic NLP, see [12].

Non diacritized texts are highly ambiguous. For example, the word علم can be diacritized in 9 different forms [23]: عِلْم (*science*), عَلَم (*flag*), عَلَّمَ (*He was taught*), etc.

In addition to the specificities of Modern Standard Arabic (MSA) discussed above, dialects pose a number of challenges including a large variations of unstandardized dialectal Arabic, and linguistic code switching between MSA and several dialects, and between Arabic and other languages like English and French. For example, the English word *Table* can be translated as طاولة (*tawela*) in Egyptian dialect, طابلة (*tabla*) in Algerian (Maghrebi) dialect, طبلية (*tabliah*) in Levantine dialect, and ماسه (*masa*) in some of the Gulf dialect speaking countries. Finally, there are also problems with the extensive use of transliterated words, such as the French word *Automobile* (*Automotive*) that becomes طونوبيل (*tonobil* which means *car*).

3 Data and Annotation

3.1 Data Collection

The collected dataset is composed of tweets posted on Twitter during the years 2011 to 2018 about different political issues and events related to the Middle East and the Maghreb. A set of predefined keywords is used to collect tweets, which targeted specific political figures (e.g., هيلاري (*Hillary*), ترامب (*Trump*), السيسي (*Al-sissi*), مبارك (*Moubarak*), مرسي (Morsi),بن علي (BenAli) , بشار (Bachar), etc.) which were the subject of the Arab spring and the presidential elections of Egypt and US. From these retrieved tweets, we selected those containing or not the Arabic ironic hashtags #سخرية, #مسخرة, #تهكم, #استهزاء[5]. Before starting with any selection process, we discarded tweets that are duplicated or tweets that depend on external links, images or videos to understand their context.

The collection process resulted in a set of 22,318 tweets (6,809 ironic tweets and 15,509 are not). These tweets are written using standard (formal) and dialectal Arabic, as shown in the examples below. Dialectal tweets contain different Arabic language varieties: Egypt (cf.(1)), Gulf (cf.(2)), and Levantine dialects (cf.(3)).

(1) #هتنتخب_السيسي_ليه عشان يقف ورا مكتب ترامب يقدم القهوة ويلمع الجزمة

(#I_will_vote_El_Sisi_why so that he stands behind Trump's office serving coffee and shining boots)

[5] All of these words are synonyms meaning "Irony".

(2) #فشلتونا احنا الزهرة نشوفها نجمة و اهما امس شافو زحل قمر .. صكيتوا ع القذافي #استهزاء

(#we_failed We see Venus as a star, and yesterday they saw Saturn as a moon .. You look like Kadhafi #irony)

(3) عبتزكر أيام زمان لما كانوا يقولوا أنو بشار منيح بس اللي حوليه هنن الفاسدين....هههههههه #مسخره

(I remember the old days when they said that Bachar is perfect but the people around him are corrupted...hahahahaha #irony)

3.2 Annotation Procedure and Agreement Study

We took a sample of 6,000 tweets to investigate the validity of using the original tweets labels. This sample consists of 3,000 tweets as ironic and 3,000 as not. It has been manually annotated by two Arabic native speakers following a three-steps procedure where an intermediate analysis of agreement and disagreement between the annotators was carried out. Annotators were first trained on 100 tweets, then were asked to annotate separately the 6,000 tweets (this step allows to compute inter-annotator agreements, cf. below). The final step was adjudication where the main case of disagreements was discussed and solved. Tweets that are either duplicates, do not contain enough context (not a clear sentence or just a link) or where annotators failed to agree have been discarded.

During the annotation process, we found 124 Farsi language tweets in the non-ironic tweets part. These tweets were retrieved automatically because both Arabic and Farsi languages use the same character encoding. We deleted these tweets since each language uses completely different meaning and words.

We measured the inter-annotator agreement using Cohen's Kappa and obtained a score of 76%, which referred to a strong agreement. This score is inline with agreements reported in annotating irony in tweets from other languages such as English (e.g., $Kappa = 0.72$ in the SemEval-2018 task 3) [14]), French ($Kappa = 0.69$ in the Deft 2017 shared task [3]), and Spanish ($Kappa = 0.67$ in IroSvA@IberLEF [28]) . The disagreement between the annotators is due to two main factors: (1) the misinterpretation or comprehension of some dialectal words; and (2) the lack of context knowledge to understand the ironic sense of the tweet. The examples (4) and (5) below respectively illustrate each of these two cases above.

(4) ترى ماصارت ينونا مرة مات حسني مبارك مرة غيبوبه

(They will make us crazy!, sometimes Hosni Moubarak is dead and sometimes he is in coma)

In this example, the irony is triggered by the word ينونا (made us crazy), written in the Gulf dialect which makes one of the two annotator having limited knowledge of this dialect miss the ironic meaning.

(5) حفل تخريج طيارين في السودان في هَذا اليَوم المبارك اي مبَارك الله يهدِيك الناس طَارَت هههه
(Graduation ceremony of pilots in Sudan on this blessed day, Moubarak God guides you .. the people flew haha)
The author of this tweet was offending a graduation ceremony of pilots in Sudan country, while in that event an helicopter landed in the middle of the ceremony and the air generated by it made the chairs and tents flying away. The lack of context knowledge by one of the annotators made her/him annotate it as not ironic.

We also measured the agreement score between the annotators' labels and the original labels and obtained a kappa score of 0.60, which is moderate. The example (6) shows an ironic tweet (that is containing an ironic hashtag) where both annotators considered it as non ironic.

(6) القذافي: الاطاحـــة بنظامـــي مدعـــاة للضحـــك - فيديو #مسخرة
(Kadhafi: the overthrow of my system is laughable - video #irony)

After the adjudication phase, we got a total of 5,030 tweets among which 2,614 were ironic tweets and 2,416 not ironic. We got a total of 5,030 tweets instead of 6,000 after removing the 124 Farsi non ironic tweets and 846 tweets (22 ironic and 824 not ironic) containing a single word and/or several hashtags that make the tweet difficult to understand. Table 1 summarizes some statistics of our dataset.

Table 1. Statistics of the irony sample corpus

#Ironic class (reference labels)	# Not-Ironic class (reference labels)
3,000	3,000
Human annotations	**Human annotations**
Ironic: 2,231 Not Ironic: 769	Ironic: 405 Not Ironic: 2,471
Total ironic tweets	**Total non-ironic tweets**
2,614	2,416
Total: 5,030	

3.3 IDAT Dataset

The distribution of tweets in the final IDAT dataset is given in Table 2. The class distribution (ironic vs. non ironic) is quite similar, with a proportion of ironic tweets of about 52% in both train and test.

Table 2. Tweet distribution the IDAT dataset.

	#Ironic	# Not-Ironic	Total
Train	2,091	1,933	4,024
Test	523	483	1,006
Total	2,614	2,416	5,030

4 Task Description and Evaluation Measures

The task consists of classifying a tweet as ironic or not ironic. The IDAT training set has been released on May 31^{th} and participants had one month and a half to train their systems. The test was then released on July 15^{th} and each participant was allowed to submit a maximum of 3 runs within 10 days.

Participating systems were evaluated using standard evaluation metrics, namely accuracy and F-score as follows. Official rankings is given according to F-score.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ number\ of\ instances} \tag{7}$$

$$Precision = \frac{True\ Positives}{True\ Positives + True\ Negatives} \tag{8}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{9}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

5 Methods of Participants and Results

Eighteen teams have registered to the shared task among which **ten** submitted their runs. Participants were from 7 different countries: Algeria, Canada, Egypt, India, Jordan, Pakistan and UK. All team members were from public entities (Universities, Research Centers).

Participants used either traditional machine learning approaches (SVM, Multimodel Naive Bayes, Logistic Regression, Ensemble models) and/or deep learning methods (CNN, RNN, LSTM, Gated Recurrent Unit, Transformers). The tweet contents are represented by traditional bag of words (*YOLO* [20], *SSN-NLP* [19]), n-grams (*BENHA* [27]) eventually weighted with TF-IDF (*BENHA, YOLO, PITS* [16]), emotion features (*Kinmokusu* [25], *PITS, YOLO*) and word embeddings (*Kinmokusu, Ali_Allaith* [1], RGCL [29], *Amrita_CEN, Tha3aroon*). Embeddings were obtained using different models such as Word2Vec, FastText and BERT. Prior to learning, some participants employed well known preprocessing steps such as removing punctuations, usernames, URLs, multiple spaces and letter duplicates (*BENHA, RGCL, Amrita_CEN, PITS, Kinmokusu, Ali_Allaith*) while others used Arabic specific cleaning to account for incorrect word spellings and reduce out of vocabulary terms (*RGCL, YOLO, Tha3aroon*).

This includes byte pair encoding, replacing some Arabic letters (e.g., hamza with ء, ى with ي) and removing diacritics and Arabizi characters (the Arabic chat alphabet).

Table 3 presents participants' results for each submitted run. The results are ranked according to the F-score. For each system, best run is given in bold font. We also compare the results with those of two baselines: SVM with unigrams term frequency (BOW) and a random baseline.

Table 3. Participants results ranked in terms of F-score. Baselines are in italic font.

Team	F-score			
	1	2	3	Rank
YOLO	**0.844**	0.833	0.823	**1**
Chiyu_Zhang_UBC	0.819	**0.824**	0.811	**2**
BENHA	0.816	0.811	**0.821**	**3**
RGCL	**0.818**	0.804	0.816	4
Ali_Allaith	**0.817**	0.794	–	5
SSN_NLP	**0.816**	0.793	0.709	6
PITS	**0.807**	–	–	7
Tha3aroon	**0.794**	0.75	–	8
BOW Baseline	*0.793*			
Kinmokusu	**0.695**	0.687	0.689	9
Amitra_CEN	**0.687**	0.534	0.434	10
Random Baseline	*0.496*			

The best 3 runs are obtained by the following systems:

1. *YOLO* using an ensemble model (based on 3 classifiers: Gradient Boosting, Random Forest and Multilayer Perceptron) relying on surface features (bag of words, TF-IDF, topic modeling). This classical ensemble outperforms both word-level Bi-LSTM ensemble with the same features set (run 2) and an hybrid ensemble that combines the first two runs (run 3);
2. *Chiyu_Zhang_UBC* using BERT in a multi-task learning configuration, BERT being pre-trained on a dialectal Twitter dataset. Several gold data were used to train the model: sentiment analysis, gender detection, age detection, dialect identification, and emotion detection. This is the sole model that views dialects as constituting different domains and therefore proposed an in-domain pre-training model with dialectal data rather than exclusively on MSA.
3. *BENHA* using an ensemble model (based on 4 classifiers: Random Forest, SVM, linear and multinomial Bayes) relying on TF-IDF and n-grams.

Neural networks have also been used by *Ali_Allaith* with Arabic FastText embeddings, and *RGCL* where six different architectures were evaluated: pooled Gated Recurrent Unit (GRU) (run 2), Long Short-Term Memory (LSTM), GRU with Attention, 2D Convolution with Pooling (run 1), GRU with Capsule (run 3) and LSTM with Capsule and Attention. Among them, 2D Convolution with Pooling was the best with an F-score of 0.818.

Similarly to *Chiyu_Zhang_UBC*, *SSN_NLP* used a deep learning approach using transformers architecture which achieved better compared to GRU with Scaled Luon attention (run 2) and a Multi-Layer Perceptron using a 300 dimensions vector as given by the AraVec pre-trained word embeddings (run 3).

PITS submitted one run consisting of a voting system with three classifiers (Multimodel Naive Bayes, Support Vector Machine, and Logistic Regression) employing a combination of frequency-based and emotion-based features. The latter were obtained by the Deepmoji tool after translating tweets from Arabic to English via the Google translation API.

Finally, *Tha3aroon* and *Amitra_CEN* used FastText word embeddings for representing tweets while *Kinmokusu* experimented with CNN and a combination of subword embeddings (obtained with Word2vec CBOW) with surface (word count, presence of hashtags, etc.) and sentiment features as given by external lexicons.

We also report, for each participant's best run, results in terms of accuracy, precision and recall (Table 4). Best recall (and accuracy) was obtained by *YOLO* while best precision by *Chiyu_Zhang_UBC*'s transformer model.

Table 4. Participants' best run in terms of accuracy, precision, recall and F-score. Baselines are in italic font.

Team	Accuracy	Precision	Recall	F-score
YOLO	**0.830**	0.808	**0.881**	**0.844**
Chiyu_Zhang_UBC	0.818	**0.828**	0.82	0.824
BENHA	0.805	0.785	0.86	0.821
RGCL	0.797	0.766	0.877	0.818
Ali_Allaith	0.806	0.802	0.832	0.817
SSN_NLP	0.800	0.781	0.855	0.816
PITS	0.786	0.761	0.859	0.807
Tha3aroon	0.786	0.795	0.794	0.794
BOW Baseline	*0.780*	*0.777*	*0.811*	*0.793*
Kinmokusu	0.689	0.701	0.688	0.695
Amitra_CEN	0.684	0.708	0.667	0.687
Random Baseline	*0.484*	*0.504*	*0.488*	*0.496*

Overall, IDAT results (best F-score= 0.844) are higher compared to the one reported in other irony detection shared tasks in other languages. For instance, macro F-score= 0.783 at the French DEFT2017 [3], F-score= 0.705 at Task 3@SemEval2018 [14], and macro F-score= 0.716 at IroSvA@IberLEF-2019 for Spanish [28].

6 Conclusion

This paper overviews the first shared task on irony detection in Arabic social media that aims at classifying a tweet as ironic or not. Eighteen teams participated in the task and a total of ten teams submitted their runs. Systems have been trained on a nearly balanced dataset composed of ironic and non ironic tweets about political issues that raised between 2011 and 2018 in the Middle East and Maghreb. The dataset has been manually annotated and inter-annotator agreement was good ($Kappa = 0.76$). The methods proposed by participants ranged from traditional features-based approaches relying on bag of words features to neural methods using pre-trained word embeddings. Several neural architectures were tested such as CNN, LSTM and Transformers. Ensemble methods have also been used. The best system achieved an F-score of 0.844 showing that classical features-based models outperform deep learning methods when applied to the IDAT dataset.

Acknowledgments

This publication was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the last author. The work of Paolo Rosso was also partially funded by Generalitat Valenciana under grant PROMETEO /2019/121.

References

1. Allaith, A., Shahbaz, M., Alkoli, M.: Neural Network Approach for Irony Detection from Arabic Text on Social Media. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)
2. Attardo, S.: Irony as Relevant Inappropriateness. Journal of Pragmatics **32**(6), 793–826 (2000)
3. Benamara, F., Grouin, C., Karoui, J., Moriceau, V., Robba, I.: Analyse d'opinion et langage figuratif dans des tweets présentation et résultats du Défi Fouille de Textes DEFT2017. In: Actes de DEFT@TALN2017 (2017)
4. Carvalho, P., Sarmento, L., Silva, M.J., Oliveira, E.D.: 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. ACM (2009)
5. Cignarella, A.T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P.: Overview of the EVALITA 2018 task on irony detection in italian tweets (ironita). In: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018. (2018)

6. Clark, H.H., Gerrig, R.J.: On the Pretense Theory of Irony. Journal of Experimental Psychology: General **113**(1), 121–126 (1984)
7. Farghaly, A., Senellart, J.: Intuitive Coding of the Arabic Lexicon. In: Proceedings of the Workshop Machine Translation for Semitic Languages associated to the Association for Machine Translation in the Americas. AMTA (2003)
8. Ghosh, A., Veale, D.T.: Fracking Sarcasm using Neural Network. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 161–169 (2016)
9. Gianti, A., Bosco, C., Patti, V., Bolioli, A., Caro, L.D.: Annotating Irony in a Novel Italian Corpus for Sentiment Analysis. In: Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals, Istanbul, Turkey. pp. 1–7 (2012)
10. Gibbs, R.W.: Irony in Talk Among Friends. Metaphor and symbol **15**(1-2), 5–27 (2000)
11. Grice, H.P.: Logic and Conversation. In: Cole, P., Morgan, J.L. (eds.) Speech Acts. Syntax and Semantics, Volume 3, pp. 41–58. Academic Press, New York (1975)
12. Habash, N.: Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2010)
13. Haverkate, H.: A speech act analysis of irony. Journal of Pragmatics **14**(1), 77 – 109 (1990)
14. Hee, C.V., Lefever, E., Hoste, V.: SemEval-2018 Task 3: Irony Detection in English Tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018. pp. 39–50 (2018)
15. Hernańdez Farías, D.I., Patti, V., Rosso, P.: Irony Detection in Twitter: The Role of Affective Content. ACM Transactions on Internet Technology (TOIT) **16**(3), 1–24 (2016)
16. Kanwar, N., Kumar, R., Mundotiya, Agarwal, M., Singh, C.: Emotion based voted classifier for Arabic irony tweet identification. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)
17. Karoui, J., Benamara, F., Moriceau, V.: SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media. In: Third International Conference On Arabic Computational Linguistics, ACLING 2017, November 5-6, 2017, Dubai, United Arab Emirates. pp. 161–168 (2017)
18. Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., Belguith, L.H.: Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing : Volume 2 short papers. pp. 644–650. ACL-IJCNLP'15 (2015)
19. Kayalvizhi, S., Thenmozhi, D., aKumar, B.S., Aravindan, C.: SSN_NLP@IDAT-FIRE-2019 : Irony Detection in Arabic Tweets using Deep Learning and Features-based Approaches. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)
20. Khalifa, M., Hussein, N.: Ensemble Learning for Irony Detection in Arabic Tweets. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)

21. Kiritchenko, S., Mohammad, S.M., Salameh, M.: Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In: Proceedings of the International Workshop on Semantic Evaluation. SemEval '16 (2016)
22. Liebrecht, C., Kunneman, F., van den, B.A.: The Perfect Solution for Detecting Sarcasm in Tweets# Not. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 29–37 (2013)
23. Maamouri, M., Bies, A., Kulick, S.: Diacritization: A Challenge to Arabic Treebank Annotation and Parsing. In: Proceeding of the British Computer Society Arabic NLP/MT Conference (2006)
24. Marton, Y., Habash, N., Rambow, O.: Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features. Computational Linguistics **39**(1), 161–194 (2013)
25. Moudjari1, L., Akli-Astouati, K.: An embedding-based approach for irony detection in Arabic tweets. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)
26. Mozannar, H., Hajal, K.E., Maamary, E., Hajj, H.M.: Neural arabic question answering. CoRR **abs/1906.05394** (2019), http://arxiv.org/abs/1906.05394
27. Nayel, H.A., Medhat, W., Rashad, M.: BENHA@IDAT: Improving Irony Detection in Arabic Tweets using Ensemble Approach. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)
28. Ortega-Bueno, R., Rangel, F., Hernández Farıas, D., 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)
29. Ranasinghe, T., Saadany, H., Plum, A., Mandhari, S., Mohamed, E., Orasan, C., Mitkov, R.: RGCL at IDAT: Deep Learning models for Irony Detection in Arabic Language. In: Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. (2019)
30. Sadat, F., Mohamed, E.: Improved Arabic-French machine translation through preprocessing schemes and language analysis. In: Proceedings of the Canadian Conference on Artificial Intelligence. pp. 308–314. AI (2013)
31. Sperber, D., Wilson, D.: Irony and the use-mention distinction. Radical pragmatics **49**, 295–318 (1981)
32. Swami, S., Khandelwal, A., Singh, V., Akhtar, S.S., Shrivastava, M.: A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection. 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing) (2018)
33. Utsumi, A.: Stylistic and Contextual Effects in Irony Processing. In: Proceedings of the 26th Annual Meeting of the Cognitive Science Society. pp. 1369–1374 (2004)
34. Zhang, S., Zhang, X., Chan, J., Rosso, P.: Irony detection via sentiment-based transfer learning. Information Processing Management **56**(5), 1633 – 1644 (2019)