Multilingual Irony Detection in Social Media

Farah Benamara
MCF-HDR Informatique

IRIT-CNRS
Univ. de Toulouse
farah.benamara@irit.fr

My research activities within MELODI

NLP at the semantics-pragmatics interface

- Study context-dependent aspects of meaning that arise within sentences as well as contextual phenomena that operate beyond the sentence:
  - Question answering
  - Evaluative language detection (sentiment analysis, hate speech detection, figurative language)
  - Intent detection (dialogues and texts)
  - Discourse processing
  - Relation extraction
  - Multilinguality and linguistic resources

My research activities within MELODI

NLP at the semantics-pragmatics interface

- Study context-dependent aspects of meaning that arise within sentences as well as contextual phenomena that operate beyond the sentence:
  - Question answering
  - Evaluative language detection (sentiment analysis, hate speech detection, figurative language)
  - Intent detection (dialogues and texts)
  - Discourse processing
  - Relation extraction
  - Multilinguality and linguistic resources

What is irony?

- Irony is a form of figurative language that can be defined as an incongruity between the literal meaning of an utterance and its intended meaning (Grice, 1975; Sperber andWilson, 1981; Utsumi, 1996).
- For example, to express a negative opinion towards a cell phone, one can either employ:
  - a literal form using a negative opinion word: *This phone is a disaster*
  - or a non-literal form by using a positive word: *What an excellent phone!!*

What is irony?

Irony in tweets: explicit vs. implicit oppositions

- **Explicit opposition**
  - The speaker intentionally creates an explicit juxtaposition of incompatible actions or words that can either have opposite polarities, or can be semantically unrelated.
    - *The Voice is more important than Fukushima tonight*
    - Ben non ! Matraquer et crever des yeux, ce n'est pas violent et ça respecte les droits !!! #ironie
  - The explicit positive/negative contrast between a subjective proposition and a situation that describes an undesirable activity or state.
    - *I love when my phone turns the volume down automatically*.

What is irony?

Irony in tweets: explicit vs. implicit oppositions

- **Implicit opposition**
    - The opposition between an assertion P describing an event or state and a pragmatic context external to the utterance in which P is false or is not likely to happen.
        - *The #NSA wiretapped a whole country. No worries for #Belgium: it is not a whole country.*
        - *#Hollande is really a good diplomat #Algeria.*

Irony in Computational Linguistics

- Binary classification task classification (Gonzaloz-Ibanez et al.,2011;Reyes et al.,2013;Joshi et al., 2016).....
  - Context-agnostic vs. Context-aware approaches
- Many shared tasks in different languages: SemEval 2020-Task 7, IDAT 2019, Evalita 2018, IroSVA 2019, ...
- Main goal: Improve polarity detection, hate speech detection, ...

Our objectives

- Focus for the first time on irony in French tweets.
    - Use irony as an umbrella term that covers irony, sarcasm and even humor.
    - Compared to English, French irony hashtags (#ironie, #sarcasme, #sarcastique, etc.) are not widely used.
- Study portability to Indo-European languages (English, Italian, Spanish) as well as less culturally close languages (Arabic).

Methodology

(a) Study how irony is expressed in social media

- Can the types of irony studied in linguistic state of the art be found in social media such as Twitter?
- If yes, what are the most frequent types? Are these types explicitly marked?
- What are the correlations between irony types and these markers?
- See if different languages have a preference for different categories.

Methodology

(b) How can we exploit these correlations in a purpose of automatic detection?

(b.1) Monolingual detection (focus here on French)
(b.2) Multilingual detection

⟶ role of cross-lingual word representations
⟶ role of syntax

Outline

1. Impact of pragmatic phenomena on irony detection

2. Monolingual irony detection: Towards a contextual model

3. Multilingual irony detection
   - On the role of cross-lingual word representations
   - On the role of syntax

Motivation

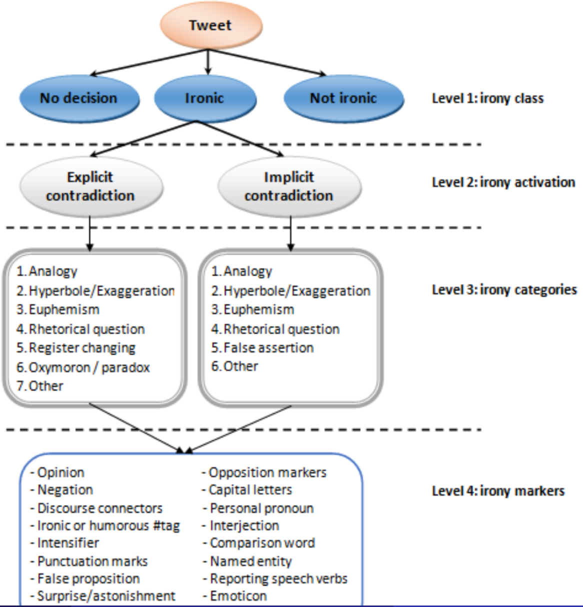
- Different categories of irony have been studied in the linguistic literature.
    - Hyperbole, exaggeration, repetition or change of register, etc.
- These categories were mainly identified in literary texts (books, poems).
- **Are these categories still valid in social media contents?**

Our approach

Informed by well-established linguistic theories of irony, we proposed (Karoui et al, EACL 2017):

- A multi-layered annotation schema in order to:
  - Measure the impact of a wide-range of pragmatic phenomena in the interpretation of irony
  - Investigate how these phenomena interact with the local context of the tweet.
- A multilingual corpus annotated according to this schema.

A multi-layered annotation scheme



Tweet

No decision

Ironic

Not ironic

Level 1: irony class

Explicit contradiction

Implicit contradiction

Level 2: irony activation

1. Analogy
2. Hyperbole/Exaggeration
3. Euphemism
4. Rhetorical question
5. Register changing
6. Oxymoron / paradox
7. Other

1. Analogy
2. Hyperbole/Exaggeration
3. Euphemism
4. Rhetorical question
5. False assertion
6. Other

Level 3: irony categories

- Opinion
- Negation
- Discourse connectors
- Ironic or humorous #tag
- Intensifier
- Punctuation marks
- False proposition
- Surprise/astonishment

- Opposition markers
- Capital letters
- Personal pronoun
- Interjection
- Comparison word
- Named entity
- Reporting speech verbs
- Emoticon

Level 4: irony markers

A multi-layered annotation scheme

8 irony categories

(1) Analogy: *Sunday is like Benzema in the French team. It is useless... :D*

(2) Hyperbole/Exaggeration

(3) Euphemism

- *The PS was so successful that all is less well: polution, housing, security #Parisledebat #Paris2014*

(4) Rhetorical question: *"Miss France is a competition" No seriously? because I didn't know!*

(5) Register changing: *Duflot left the governement. In the middle of Lent, we can not even celebrate it. Really, she bothers until the end ... *sigh**

(6) False assertion: *The #NSA wiretapped a whole country. No worries for #Belgium: it is not a whole country.*

(7) Oxymoron/paradox: *It is obvious that every whistler was here for November 11th and not to whistle François Hollande's politics.*

(8) Other: *Polution alert: it is not recommanded to take your bike to go work at 9am...but not your diesel car !*

Data

- Tweets about hot topics discussed in the media.
  - The pragmatic context needed to infer irony is more likely to be understood by annotators compared to tweets that relate personal content.
- Three corpora in French, English and Italian
- Selection of ironic vs. non-ironic tweets
  - Partly different criteria for the three addressed languages in order to tackle their features.

Corpus	*Ironic*	*Not Ironic*
French	2,073	16,179
English	5,173	6,116
Italian	806 (Sentipolc) + 2,273 (TW-SPINO)	5,642 (Sentipolc)

Annotation procedure

- Ironic hashtags removed.
- First, the annotation of the French data with three French native speakers:
  - Training: 100 tweets
  - Adjudication stage: 300 tweets
    - Ironic/Not ironic classification: Cohen's Kappa =0.69
    - Irony activation: Cohen's Kappa =0.65
    - Irony category identification: Cohen's Kappa = 0.56 (Cohen's Kappa= 0.60 when similar devices are grouped together)
  - Effective annotation: 1,700 tweets
  - Distribution of ironic tweets in each stage: 80%.

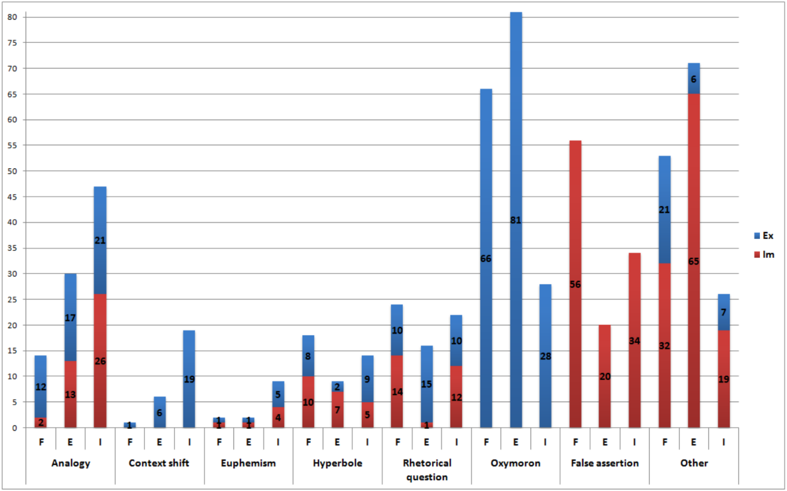
Quantitative results

Number of tweets in annotated corpora in French, English and Italian

	Ironic		**Non Ironic**	**No decision**	*Total*
	explicit	implicit			
French	394	1066	380	160	2000
English	144	283	99	24	550
Italian	260	140	100	–	500

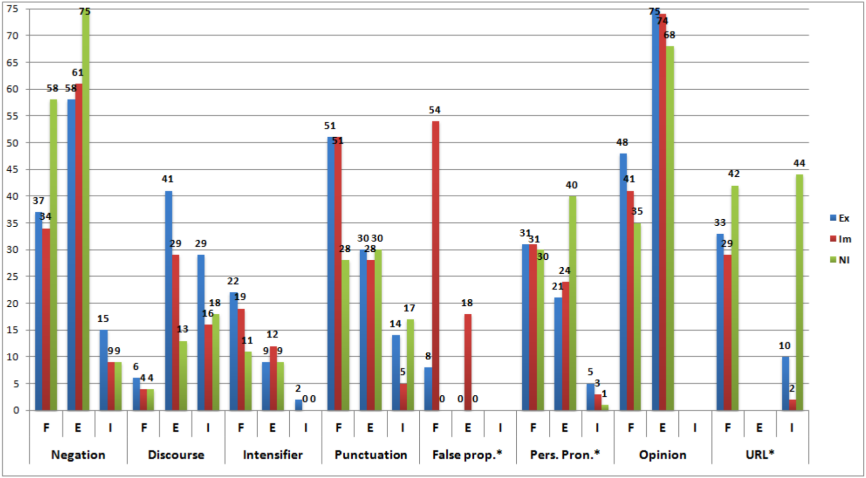
Quantitative results

Categories in explicit (*Ex*) or implicit (*Im*) activation in French (F), English (E) and Italian (I) (in %)



Quantitative results

Markers in ironic (*Exp* or *Imp*) and non ironic (*NI*) tweets in French, English and Italian (in %). Markers with an * have not been studied in irony literature



Main conclusions

- The results show that our schema is reliable for French and that it is portable to English and Italian, observing relatively the same tendencies in terms of irony categories and markers.
- We observed correlations between markers and ironic/non ironic classes, between markers and irony activation types (explicit or implicit) and between markers and irony categories.

- The annotated multilingual corpora are available upon request
- The French corpus has been used for the first French shared task on irony detection (Benamara et al, DEFT@TALN2017)
- Extension of this study to Italian, see (Cignarella et al., LREC 2018).

Outline

1. Impact of pragmatic phenomena on irony detection

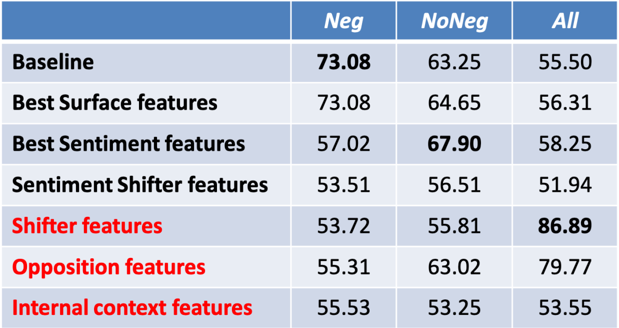
2. Monolingual irony detection: Towards a contextual model

3. Multilingual irony detection
   - On the role of cross-lingual word representations
   - On the role of syntax

Our approach

- Our results are interesting in a perspective of pragmatically and linguistically informed automatic irony detection, since it brings out the most discriminant features.
- French irony detection focusing on irony expressed using an implicit activation (Karoui et al., ACL 2015). **Two complementary models**:
  - **Model 1**: A supervised learning method relying exclusively on the information internal to the tweet.
  - **Model 2**: A query-based method that corrects the misclassified ironic instances of the form $Not(P)$ by looking for P in reliable external sources of information on the Web.

Model 1 Results based on tweet linguistic content only



	Neg	*NoNeg*	*All*
Baseline	**73.08**	63.25	55.50
Best Surface features	73.08	64.65	56.31
Best Sentiment features	57.02	**67.90**	58.25
Sentiment Shifter features	53.51	56.51	51.94
Shifter features	53.72	55.81	**86.89**
Opposition features	55.31	63.02	79.77
Internal context features	55.53	53.25	53.55

Model 2 based on context outside the tweet

Example:

- **Topic**: Valls
- **Tweet**: #Valls has learnt that Sarkozy was wiretapped in newspapers. Fortunately he is not the interior minister.

Steps:

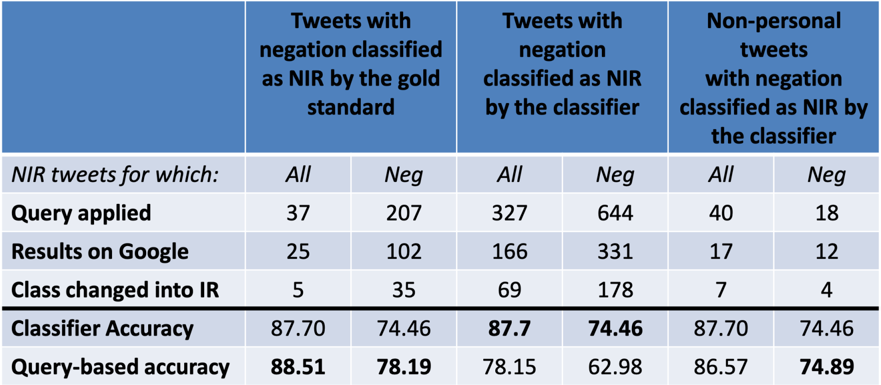
- Sentences segmentation:
  - S1: #Valls has learnt that Sarkozy was wiretapped in newspapers.
  - S2: Fortunately he is not the interior minister.
- From S2, we remove the negation word “not”, isolate the negation scope P={interior, minister} and generate the query Q1 = {Valls interior minister}.
- **Result**:

  <Title>Manuel **Valls** - Wikipedia, the free encyclopedia</Title>

  <Snippet>... French politician. For the Spanish composer, see Manuel **Valls** (composer). .... **Valls** was appointed **Minister** of the **Interior** in the Ayrault Cabinet in May 2012.</Snippet>

➔ **All query keywords were found in this snippet, we can then conclude that the tweet is ironic.**

Model 2 based on context outside the tweet



	Tweets with negation classified as NIR by the gold standard		**Tweets with negation classified as NIR by the classifier**		**Non-personal tweets with negation classified as NIR by the classifier**	
NIR tweets for which:	*All*	*Neg*	*All*	*Neg*	*All*	*Neg*
Query applied	37	207	327	644	40	18
Results on Google	25	102	166	331	17	12
Class changed into IR	5	35	69	178	7	4
Classifier Accuracy	87.70	74.46	**87.7**	**74.46**	87.70	74.46
Query-based accuracy	**88.51**	**78.19**	78.15	62.98	86.57	**74.89**

Outline

1. Impact of pragmatic phenomena on irony detection

2. Monolingual irony detection: Towards a contextual model

3. Multilingual irony detection
   - On the role of cross-lingual word representations
   - On the role of syntax

Plan

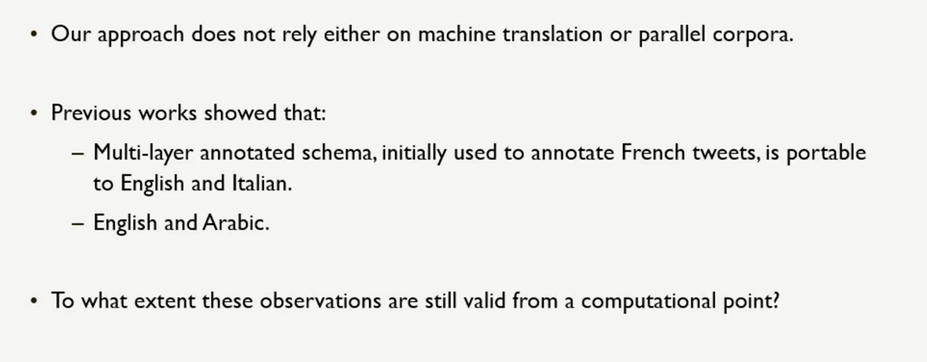
1. Impact of pragmatic phenomena on irony detection

2. Monolingual irony detection: Towards a contextual model

3. Multilingual irony detection
    - On the role of cross-lingual word representations
    - On the role of syntax

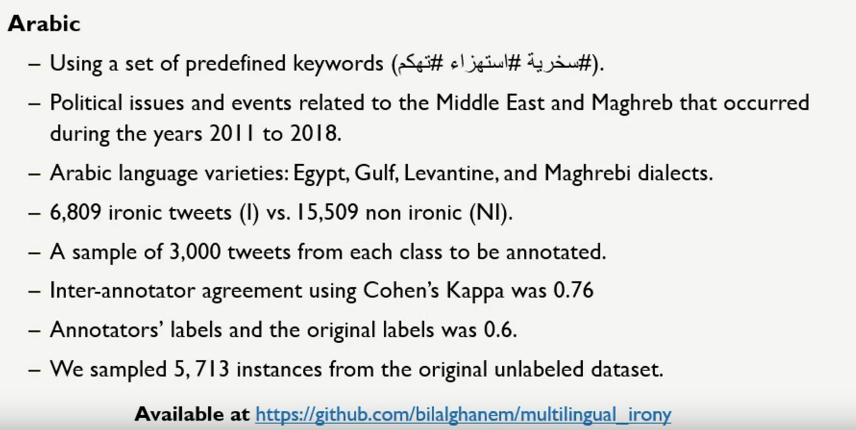
Motivations

(Ghanem et al., ECIR 2020)



- Our approach does not rely either on machine translation or parallel corpora.
- Previous works showed that:
  - Multi-layer annotated schema, initially used to annotate French tweets, is portable to English and Italian.
  - English and Arabic.
- To what extent these observations are still valid from a computational point?

Motivations



Arabic

- Using a set of predefined keywords (#سخرية #استهزاء #تهكم).
- Political issues and events related to the Middle East and Maghreb that occurred during the years 2011 to 2018.
- Arabic language varieties: Egypt, Gulf, Levantine, and Maghrebi dialects.
- 6,809 ironic tweets (I) vs. 15,509 non ironic (NI).
- A sample of 3,000 tweets from each class to be annotated.
- Inter-annotator agreement using Cohen's Kappa was 0.76
- Annotators' labels and the original labels was 0.6.
- We sampled 5, 713 instances from the original unlabeled dataset.

Available at https://github.com/bilalghanem/multilingual_irony

⟶ An extended version of this Arabic dataset has been used for the first shared task on Arabic irony detection (Ghanem et al, IDAT@FIRE 2019).

Data

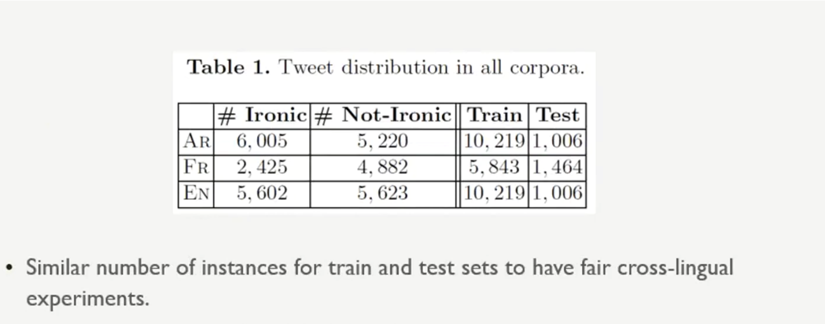
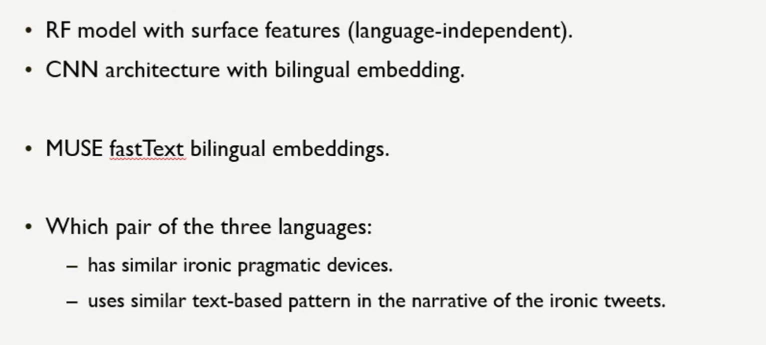


Table 1. Tweet distribution in all corpora.

	# Ironic	# Not-Ironic	Train	Test
AR	6, 005	5, 220	10, 219	1, 006
FR	2, 425	4, 882	5, 843	1, 464
EN	5, 602	5, 623	10, 219	1, 006

- Similar number of instances for train and test sets to have fair cross-lingual experiments.

Models



- RF model with surface features (language-independent).
- CNN architecture with bilingual embedding.
- MUSE fastText bilingual embeddings.
- Which pair of the three languages:
  - has similar ironic pragmatic devices.
  - uses similar text-based pattern in the narrative of the ironic tweets.

Results

- Semantic perspective:
  - Arabic and French pairs.
  - Arabic and English pairs.
- Word embeddings low coverage.
- Arabic dialects

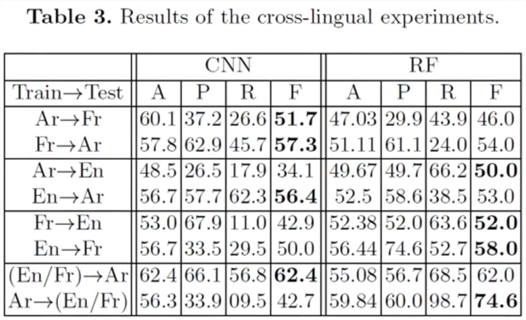
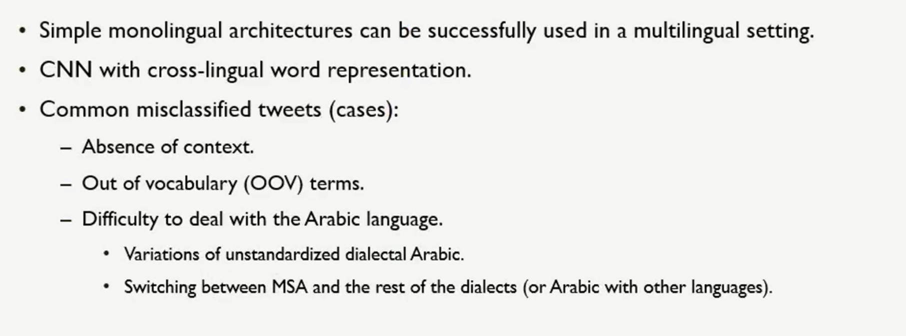


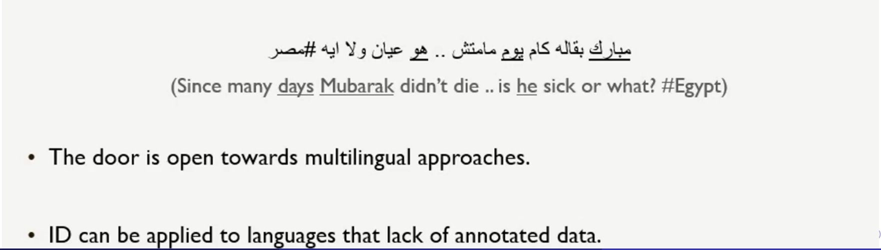
Table 3. Results of the cross-lingual experiments.

	CNN				RF			
Train→Test	A	P	R	F	A	P	R	F
Ar→Fr	60.1	37.2	26.6	**51.7**	47.03	29.9	43.9	46.0
Fr→Ar	57.8	62.9	45.7	**57.3**	51.11	61.1	24.0	54.0
Ar→En	48.5	26.5	17.9	34.1	49.67	49.7	66.2	**50.0**
En→Ar	56.7	57.7	62.3	**56.4**	52.5	58.6	38.5	53.0
Fr→En	53.0	67.9	11.0	42.9	52.38	52.0	63.6	**52.0**
En→Fr	56.7	33.5	29.5	50.0	56.44	74.6	52.7	**58.0**
(En/Fr)→Ar	62.4	66.1	56.8	**62.4**	55.08	56.7	68.5	62.0
Ar→(En/Fr)	56.3	33.9	09.5	42.7	59.84	60.0	98.7	**74.6**

Conclusions



- Simple monolingual architectures can be successfully used in a multilingual setting.
- CNN with cross-lingual word representation.
- Common misclassified tweets (cases):
  - Absence of context.
  - Out of vocabulary (OOV) terms.
  - Difficulty to deal with the Arabic language.
    - Variations of unstandardized dialectal Arabic.
    - Switching between MSA and the rest of the dialects (or Arabic with other languages).



مبارك بقاله كام يوم مامتش .. هو عيان ولا ايه #مصر

(Since many days Mubarak didn't die .. is he sick or what? #Egypt)

- The door is open towards multilingual approaches.

- ID can be applied to languages that lack of annotated data.

Plan

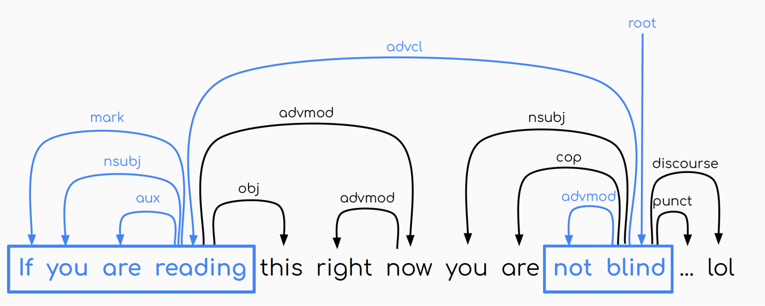
1. Impact of pragmatic phenomena on irony detection

2. Monolingual irony detection: Towards a contextual model

3. Multilingual irony detection
   - On the role of cross-lingual word representations
   - On the role of syntax

Motivations

- Most computational research on irony detection focuses primarily on semantic or pragmatic devices, neglecting syntax
- The deviation from syntactic norms has been reported in literature as a possible trigger of the phenomenon



Methodology

(Cignarella et al., Coling 2020)

(RQ-1) - Can morphological and syntactic knowledge be helpful in addressing the task of irony detection?

- Universal Dependencies
- Experiments with UD-based word embeddings
- Experiments with “syntax-informed” BERT

Methodology

(RQ-2) - To what extent do *ad-hoc* resources in UD format (treebanks) improve irony detection performances?

Three experimental settings:

1. dependency-based syntactic features combined with classical ML classifiers
   → to find the best set of features
2. word-embedding models
   → UD-based
3. dependency-based syntactic features combined with Multilingual BERT

Methodology

(RQ-3) - Are results obtained using syntactic features stable across different languages?

Datasets made available from previous shared tasks:

- English (SemEval-2018 Task 3)
- French (DEFT 2017)
- Spanish (IroSvA 2019)
- Italian (IronITA 2018)

Models

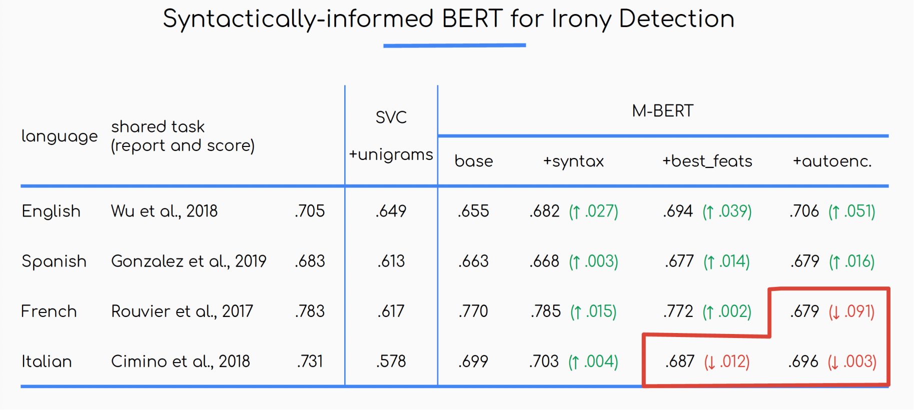
Three experimental settings:

1. classical ML
   → SVM, LR, RF and MLP

2. neural networks (GRU) and word embeddings
   → fastText
   → dependency-based word embeddings*

3. BERT + features
   → Multilingual BERT (M-BERT)

Results



Syntactically-informed BERT for Irony Detection

language	shared task (report and score)		SVC	M-BERT			
			+unigrams	base	+syntax	+best_feats	+autoenc.
English	Wu et al., 2018	.705	.649	.655	.682 (↑ .027)	.694 (↑ .039)	.706 (↑ .051)
Spanish	Gonzalez et al., 2019	.683	.613	.663	.668 (↑ .003)	.677 (↑ .014)	.679 (↑ .016)
French	Rouvier et al., 2017	.783	.617	.770	.785 (↑ .015)	.772 (↑ .002)	.679 (↓ .091)
Italian	Cimino et al., 2018	.731	.578	.699	.703 (↑ .004)	.687 (↓ .012)	.696 (↓ .003)

Conclusions

- Versatility of the UD format → language-independent approach
- Pre-trained word embeddings + dependency syntax for irony detection
- Syntax-based models outperform syntax-agnostic ones
- Our models overcome competitive baselines of the shared tasks and favorably compare with the best results
- We have enriched datasets with morphosyntactic information in UD

Many thanks to the amazing irony detection team!

J. Karoui

V. Moriceau

IRIT-UPS

V. Patti

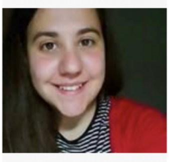
C. Bosco

V. Basile

Univ. Turin

P. Rosso

B. Ghanem



T. Cignarella