Irony Detection

Pierre EPRON, Maxime RENARD, Shu ZHANG

Irony

Irony is generally defined as a difference between the literal meaning and the intended meaning [1], i.e. there exists a **contradiction** between elements of an utterance.

- The **contradiction** can resides at a semantic, veracity or intention level.
- It is either **explicit**: contrast between the two propositions, or **implicit**: contrast between what is claimed and some context that is external to the utterance
- The perception of irony changes based on the cultural background of the person [2].

Motivations

Irony, a complex linguistic phenomenon with varying interpretations, poses a challenge for both humans and automated systems. Recognizing irony is crucial, especially in the context of harmful behavior on social media [3],[4]. Leveraging recent advancements and a rich irony dataset, this project aims to enhance irony detection using state-of-the-art language models.

Existing datasets for Irony detection

- *Sarcasm as Contrast between a Positive Sentiment and Negative Situation* [5],
- Task 3 from *Evalita 2014 SENTIment POLarity Classification Task* [6],
- *Sarcasm Detection on Czech and English Twitter* [7],
- *SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter* [8],
- *DEFT 2017: Analyse de sentiments et détection de l'ironie* [9],
- *SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media* [10],
- *A Large Self-Annotated Corpus for Sarcasm* [11],
- Task 3 from *SemEval-2018* [12]

Dataset

EPIC (English Perspectivist Irony Corpus) [13] is a disaggregated English corpus for irony detection, containing 3,000 pairs of short conversations (*Posts-Replies*) from Twitter and Reddit, along with the demographic information of each annotator (age, nationality, gender, and so on)

- **Reddit** comments were collected from selected subreddits between January 2020 and June 2021, with subsequent processing to ensure English language and data integrity by removing irrelevant pairs and identifying language by using **LangID Python library**[14].
- The **Twitter** data collection utilized geolocation through the Twitter API to discern English varieties by validating the country associated with tweet pairs. Queries to the **Twitter Stream API**[15] were made to gather English tweets from the specified five countries, with a focus on "conversation starting" tweets excluding replies or quotes. The collection included (*Post, Reply*) pairs where the *Post* served as the conversation initiator or a direct reply.

Dataset

Annotation

- Annotations were collected through the **Prolific**[16] platform, involving English-speaking annotators from diverse backgrounds.
- Each pair has been annotated by multiple annotators that were asked to provide a **binary label** (either Irony or not-Irony) for the *Reply* text given the context provided by *Post*.
- A total of 76 annotators, native English speakers, were hired and each annotated 200 instances from the dataset. Instances were balanced across annotators from five English-speaking countries, promoting a diverse perspective on irony perception. Resulting in a final set of 74 valid annotators for the dataset.

Fine-tuning Text Classification

RoBERTa based model [17] fine-tuned on irony detection task from TWEETEVAL benchmark [18]. Task 3 from SemEval 2018.

- Compare RoBERTa-base and RoBERTa-irony (with and without fine-tuning on EPIC dataset).
- Cross Entropy vs Matthews Correlation Coefficient loss [19].

In-context Text Classification

Use a pre-trained Large Language Model (LLM) to solve a task.

- Llama 2 [20]
- Zero shot prompt vs Few shot prompt.
- Dealing with potential hallucinations.
- Is "World Knowledge" useful for irony detection ?
- Does including the source (reddit or twitter) in the prompt have an impact ?

Prompt tuning "leaderboard": `https://www.promptingguide.ai/`

Conclusion

- Don't try to define irony.
- Do individuals develop a similar sense of irony when they relocate to a different place?
- Do LLMs perform more effectively on doxa compared to conventional language models?
- Can LLMs capture and differentiate various "web cultures"?

Conclusion

- Prepare the data for training and evaluation
- Fine-tune and evaluate RoBERTa.
- Define a minimal prompt and evaluate it with Llama 2.
- Analyze results and establish next steps

References I

[1] Jihen Karoui et al. "Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 262–272. URL: https://aclanthology.org/E17-1025.

[2] Reynier Ortega Bueno et al. "Overview of the Task on Irony Detection in Spanish Variants". In: *IberLEF@SEPLN*. 2019. URL: https://api.semanticscholar.org/CorpusID:199448552.

[3] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. "From humor recognition to irony detection: The figurative language of social media". In: *Data & Knowledge Engineering* 74 (2012), pp. 1–12. ISSN: 0169-023X. DOI: https://doi.org/10.1016/j.datak.2012.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0169023X12000237.

References II

[4] María Estrella Vallecillo Rodrguez et al. "SINAI at SemEval-2023 Task 10: Leveraging Emotions, Sentiments, and Irony Knowledge for Explainable Detection of Online Sexism". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 986–994. DOI: 10.18653/v1/2023.semeval-1.136. URL: https://aclanthology.org/2023.semeval-1.136.

[5] Ellen Riloff et al. "Sarcasm as Contrast between a Positive Sentiment and Negative Situation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 704–714. URL: https://aclanthology.org/D13-1066.

References III

[6] Valerio Basile, Andrea Bolioli, and Viviana Patti. "Overview of the Evalita 2014 SENTIment POLarity Classification Task". In: *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa* (2014). URL: https://api.semanticscholar.org/CorpusID:247071464.

[7] Tomáš Ptáček, Ivan Habernal, and Jun Hong. "Sarcasm Detection on Czech and English Twitter". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 213–223. URL: https://aclanthology.org/C14-1022.

References IV

[8] Aniruddha Ghosh et al. "SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 470–478. DOI: 10.18653/v1/S15-2080. URL: https://aclanthology.org/S15-2080.

[9] Amira Barhoumi et al. "«~L'important c'est de participer~» : positive ironie. Analyse de sentiments et détection de l'ironie Les systèmes du LIUM et d'OCTO.". In: *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Actes de l'atelier « Défi Fouille de Textes » (DEFT 2017)*. LIUM-OCTO Results - DEFT 2017. Orléans, France: Association pour le Traitement Automatique des Langues, June 2017, pp. 42–54. URL: http://talnarchives.atala.org/ateliers/2017/DEFT/4.pdf.

[10] Jihen Karoui, Farah Benamara, and Véronique Moriceau. "SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media". In: *International Conference on Arabic Computational Linguistics*. 2017. URL: https://api.semanticscholar.org/CorpusID:207429781.

References V

[11] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. "A Large Self-Annotated Corpus for Sarcasm". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: https://aclanthology.org/L18-1102.

[12] Cynthia Van Hee, Els Lefever, and Véronique Hoste. "SemEval-2018 Task 3: Irony Detection in English Tweets". In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 39–50. DOI: 10.18653/v1/S18-1005. URL: https://aclanthology.org/S18-1005.

[13] Simona Frenda et al. "EPIC: Multi-Perspective Annotation of a Corpus of Irony". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 13844–13857. DOI: 10.18653/v1/2023.acl-long.774. URL: https://aclanthology.org/2023.acl-long.774.

References VI

[14] LangIDPythonLibiary. *LangIDPythonLibiary*. Year. URL: https://github.com/saffsd/langid.py.

[15] Twitter. *TwitterAPI*. Year. URL: https://developer.twitter.com/en/docs/twitter-api.

[16] Prolific. *Prolific*. Year. URL: https://www.prolific.com/.

[17] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv* abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[18] Francesco Barbieri et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. DOI: 10.18653/v1/2020.findings-emnlp.148. URL: https://aclanthology.org/2020.findings-emnlp.148.

References VII

[19] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PLoS ONE* 12 (2017). URL: https://api.semanticscholar.org/CorpusID:10830747.

[20] Hugo Touvron et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *ArXiv* abs/2307.09288 (2023). URL: https://api.semanticscholar.org/CorpusID:259950998.

Thanks

Thanks for watching!