Internship
JokeR: Automatic Wordplay and Humour Translation

Duration: 5 months

Starting date: February 2022

Organization: Université de Bretagne Occidentale, HCTI EA-4249

Contact: Liana Ermakova liana.ermakova@univ-brest.fr

Context:

This project is a part of the JokeR project: Automatic Pun and Humour Translation

JokeR website: http://www.joker-project.com/

Humour and irony studies are now crucial when it comes to social listening (Ghanem et al., 2019; Karoui et al., 2017; Reyes et al., 2009), dialogue systems (chatbots), recommender systems, reputation monitoring, and the detection of fake news (Guibon et al., 2019) and hate speech (Francesconi et al., 2019). To understand humour, one often has to grasp implicit cultural references and/or capture double meanings, which of course raises the question of the (un)translatability of humour. While modern translation is heavily aided by technological tools, virtually none has any specific support for humour and wordplay.

Pun is a type of wordplay that exploits multiple meanings of a term or of similar-sounding words.

Humorous wordplay often exploits the confrontation of similar forms but different meanings. On the one hand, Machine Translation is generally ignorant of pragmatics and assumes that words in the source text are formed and used in a conventional manner. Machine Translation systems fail to recognize the deliberate ambiguity of puns or the unorthodox morphology of neologisms, leaving such terms untranslated or else translating them in ways that lose the humorous aspect (Austrian Research Institute for Artificial Intelligence (OFAI), 1010 Vienna, Austria & Miller, 2019). Most AI-based translation tools require a quality and quantity of training data (e.g., parallel corpora) that has historically been lacking of humour and wordplay. To construct such a corpus automatically one should (1) detect wordplay instances (2) align them with their translations.

Candidate profile

We are looking for highly-motivated **computer-science** candidates (bac+4, bac+5) who are creative, autonomous and offer initiatives. Ideally, a candidate should have an experience in *deep learning* and *natural language processing* (NLP).

Programing languages: Python (prototyping), Java (integration into PunCAT)

Knowledge of Pandas (https://pandas.pydata.org/), NLTK (https://www.nltk.org/), WordNet (https://wordnet.princeton.edu/), XML libraries is an advantage.

Regular communication is required. Suggestions, questions and problems should be reported immediately.

Tasks

The candidate will be involved in the following activities (non-exhaustive list):

- automatic pun detection and localization with deep learning methods (e.g. Google's Multilingual T5 model (Xue et al., 2021, p. 5))
- pun generation
- pun interpretation (e.g. searching for WordNet meaning)
- pun classification
- automatic parallel text alignment
- integration of develop prototypes into the PunCAT tool
- website construction

Useful links:

- Paper "SemEval-2017 Task 7: Detection and Interpretation of English Puns": https://aclanthology.org/S17-2005.pdf
- Dataset: https://tudatalib.ulb.tu-darmstadt.de/bitstream/handle/tudatalib/2445/semeval2017_task7.tar.xz
- SemEval-2017 results: https://alt.qcri.org/semeval2017/task7/index.php?id=results
- Multilingual T5 model: https://github.com/google-research/multilingual-t5

References

Austrian Research Institute for Artificial Intelligence (OFAI), 1010 Vienna, Austria, & Miller, T. (2019). The Punster's Amanuensis: The Proper Place of Humans and Machines in the Translation of Wordplay. *Proceedings of the Second Workshop Human-Informed Translation and Interpreting Technology Associated with RANLP 2019*, 57–65. https://doi.org/10.26615/issn.2683-0078.2019_007

Francesconi, C., Bosco, C., Poletto, F., & Sanguinetti, M. (2019). *Error Analysis in a Hate Speech Detection Task: The Case of HaSpeeDe-TW at EVALITA 2018*. 7.

Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., & Rosso, P. (2019). *IDAT@FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets*. 11.

Guibon, G., Ermakova, L., Seffih, H., Firsov, A., & Noé-Bienvenu, G. L. (2019, April 7). *Multilingual Fake News Detection with Satire*. CICLing: International Conference on Computational Linguistics and Intelligent Text Processing. https://halshs.archives-ouvertes.fr/halshs-02391141

Karoui, J., Benamara, F., Moriceau, V., Patti, V., Bosco, C., & Aussenac-Gilles, N. (2017). Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. *15th European Chapter of the Association for Computational Linguistics (EACL 2017), 1-long pap*, 262–272. https://oatao.univ-toulouse.fr/18921/

Reyes, A., Buscaldi, D., & Rosso, P. (2009). An Analysis of the Impact of Ambiguity on Automatic Humour Recognition. In V. Matoušek & P. Mautner (Eds.), *Text, Speech and Dialogue* (pp. 162–169). Springer. https://doi.org/10.1007/978-3-642-04208-9_25

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41