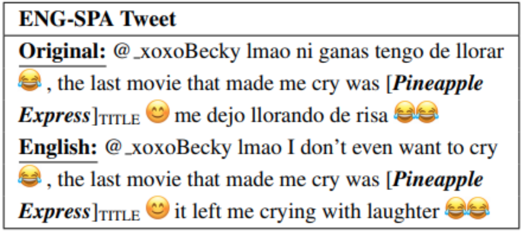
Seminar Topics: Information Extraction

English topics!

Viktor Hangya

hangyav@cis.lmu.de

IE from Code-Switched Data



ENG-SPA Tweet

Original: @_xoxoBecky lmao ni ganas tengo de llorar 😂 , the last movie that made me cry was [***Pineapple Express***]TITLE 😊 me dejo llorando de risa 😂😂

English: @_xoxoBecky lmao I don't even want to cry 😂 , the last movie that made me cry was [***Pineapple Express***]TITLE 😊 it left me crying with laughter 😂😂

- Code-switched data:
  - mix of multiple languages in sentences
  - hard to process with IE approaches
  - small code-switched training data
- Project:
  - What are the difficulties when processing such data?
    - introduce the problem, datasets, evaluation methodology, is the problem solved or are there open issues
  - What techniques can be applied to deal with these issues?
    - pick a one (or more) papers and describe their motivation, approach and findings
  - focus on part-of-speech tagging, named entity recognition or sentiment analysis
- Resources:
  - Aguilar et al., 2020, **LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation** *Proceedings of LREC-2020*

Target- and Aspect-Level Sentiment Analysis

- Sentiment analysis: extract sentiment polarity of opinions:
  - Positive: I'm happy.
  - Negative: I'm sad.
  - Neutral: The sky is blue.
- Target-level: Opinions can be different given the target entity:
  - Android is better than iOS.
  - The food was great but the service was awful.
- Project:
  - focus on sentiment polarity detection (there could be other tasks as well: e.g. category or target/aspect detection)
  - introduce the task and describe a few interesting approaches
- Resources:
  - Pontiki et al., 2016, **SemEval-2016 Task 5: Aspect Based Sentiment Analysis** *Proceedings of SemEval-2016*
  - https://github.com/songyouwei/ABSA-PyTorch

Toxic Span Detection

- Toxic/hate speech detection:
  - important task to protect people online
  - usually text classification task: is a given text toxic?
- Span detection:
  - extract the toxic expressions in texts
  - more precise aid for moderators

  This is a **stupid ass** example, so thank you for nothing **a!@#!@**.
- Project:
  - Why is the task important?
  - Is it easy to decide what is toxic, even for humans?
  - Describe a few approaches, highlight their most interesting aspects and compared to other systems.
- Resources:
  - Pavlopoulos et al., 2021, **SemEval-2021 Task 5: Toxic Spans Detection** *Proceedings of SemEval-2021*

Relation Extraction and Classification in Scientific Documents

- Automatically identify relevant domain-specific semantic relations in scientific publications, e.g.:
  - a new **method** is proposed for a **task**
  - a **phenomenon** is found in a certain **context**
  - **results** of different **experiments** are compared to each other
- Used for e.g.:
  - build knowledge-graphs
  - do a more detailed search
- Project:
  - Cover both relation identification and relation type classification!
  - What are the challenges of the task? Are there relation types that are harder to detect? Why?
- Resources:
  - Gábor et al., 2018, **SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers** *Proceedings of the 12th International Workshop on Semantic Evaluation*

Questions?

hangyav@cis.lmu.de