Project choices

Leon Derczynski

Innopolis University

Project

- 40% of the overall mark
- You’re welcome to work in groups of 1-3
  - Bigger group means you need a better assignment

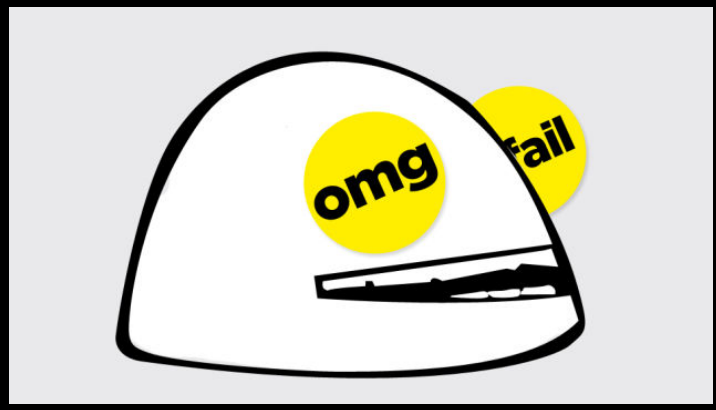
NER with gazetteers

- Summary
  - Named entity recognition, adding lists of some terms according to their type (cities, footballers, biscuits)

- Challenges
  - How to add gazetteer knowledge to normal features?
  - Feature sparsity could be an issue
- Input
  - CoNLL training data with NEs, Gazetteer lists
- Output
  - Test data with tokens labelled with entity type

NER for social media

- Summary
  - Find NEs in social media
- Challenges
  - See week 3 :) Unusual terms, unreliable case, mis-spelling
- Input
  - CoNLL-format training data with NEs, gazetteer lists
- Output
  - Entities from social media



Check reactions

- Formally called “Stance Detection”
- Support, deny, query or comment on a claim?
- LSTM classification could work (sample code for this from week 2)
- .. so could simpler methods: building lists of words that match each “stance”, then using these as features

- Data in “RumourEval” - Task A

PoS tagger for a new language

- Summary
  - Here's a new language (Dothraki?). Let's PoS tag it!
- Challenges
  - No data! That's OK; state-of-the-art accuracy can be attained with ~2 hours annotation

    “Learning a Part-of-Speech Tagger from Two Hours of Annotation”, Garrette and Baldridge, NAACL 2013
  - ...We don't need state-of-the-art :)
- Input
  - Text you've annotated with POS (following e.g. universal scheme)
- Output
  - A totally new tool for handling an “unresourced” language

Generative Eliza

- Build a copy of Eliza in Python
- Find some dialogues as training data
  - I have a lot of conversation scripts, ask!
- Learn a language model
- Output Eliza-like sentences
- One idea:
  - Train a seeded NLG system, like in the LSTM language model tutorial, based on some other conversation scripts
  - Make it talk with Eliza
  - Record the responses, so you have many Eliza conversations
  - Use this output, as a training set for Eliza responses
  - The resulting model can generate Eliza's side of the dialogue

Project format

- Write as an academic paper
  - Use the COLING 2020 style files
  - https://coling2020.org/pages/submission
  - Results will be published informally
  - You’re welcome to submit to COLING with my help
- Submit a project proposal first
  - Due ASAP
  - This describes the problem you’d like to work on
  - I’ll make sure you approach the right-sized problem

Project format

- Main sections:
  - Introduction
  - Background (literature, similar previous work)
  - Method
    - Baseline
    - Dataset
    - Your approach & why you chose it
  - Analysis
    - What worked, what didn’t work, and why
  - Conclusion

Course wrap-up

- Project assignment due **April 30**
- Mail me for any corpora/annotations, there are *many*

- Input:
  - Code or a link to a colab
  - Documentation (4-page paper)
- Output:
  - Sweet, sweet ECTS

- Thanks for participating!