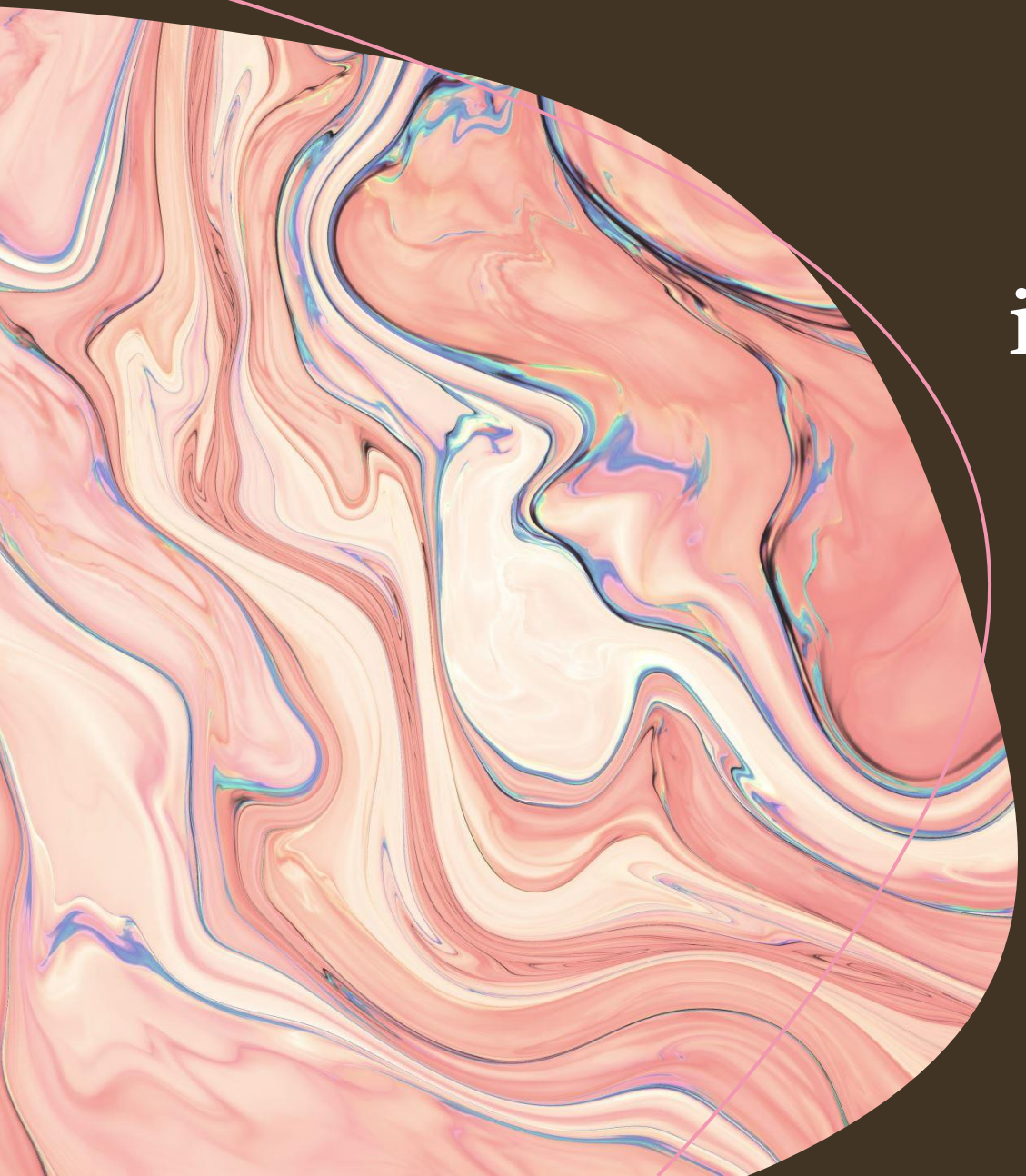

Annotation of *you*-pronouns in Early Modern English texts using machine learning techniques

A project by Maria Irena Szawerna

for Machine learning for statistical NLP: Advanced LT2326

What is this project about? I

- Copious amounts of literature on second-person singular EME or Shakespearean pronouns:
  - Whether *you* forms are singular or plural is either disregarded or manually annotated by every researcher separately.
  - Busse (2002) tries to extrapolate his results using estimates from a smaller corpus, but admits that in reality those results are not necessarily true to the source material (pp. 30, 40-42).

What is this project about? II

- My own research for an MA thesis at a different university forced me to hand-annotate Shakespeare's plays for this feature:
  - Software that annotated pronouns for that particular feature could be very useful for potential future quantitative research in the field.
- Goal: developing a method for *you*-pronoun annotation using Machine Learning techniques.

Background information I: previous research

- Many inquiries into whether taggers developed on modern languages work on their historical variants:
  - Spelling, punctuation need standardization.
  - OOV tokens.
  - Lower accuracy due to older grammar rules.
  - Sparsity of source material.
- Not always tagging for features like number.

Background information II: previous research

- Older literature focuses on testing modern taggers and determining what is causing issues:
  - Adesam, Y., & Bouma, G. (2016), , Bollmann, M. (2013), Hiltunen, T., & Tyrkkö, J. (2013), Hupkes, D., & Bod, R. (2016), Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007), Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011).
- Newer research focuses on the use of Feature or Word Embeddings:
  - Kulick, S., Ryant, N., & Santorini, B. (2022), Yang, Y., & Eisenstein, J. (2016).

Background information III: data

- Modernized spelling versions from The Folger Shakespeare.
- Manually trimmed and annotated versions of *As You Like It* and *Hamlet*.
  - _SG, _PL, _UNK tags.

```
Who's there?

Nay, answer me. Stand and unfold yourself_SG.

Long live the King!

Barnardo?

He.

You_SG come most carefully upon your_SG hour.

'Tis now struck twelve. Get thee to bed, Francisco.

For this relief much thanks. 'Tis bitter cold,
And I am sick at heart.

Have you_SG had quiet guard?

Not a mouse stirring.

Well, good night.
If you_SG do meet Horatio and Marcellus,
The rivals of my watch, bid them make haste.
```

Methods and implementation I

- Annotating only the number of a pronoun:
  - Not fully a POS-tagging problem, could be viewed as a classification problem.
  - Classes should be represented equally and _UNK is very rare – binary classification problem (_SG, _PL).
- Utilizing BERT embeddings but not simply fine-tuning it to the task:
  - Kulick, Ryant, & Santorini (2022) argue in favor of word embeddings.
  - Fine-tuning BERT does not make for a good project.

Methods and implementation II

- Determining the number of a pronoun:
  - No morphological distinction between *you_SG* and *you_PL* except for *yourself* and *yourselves*.
  - Contextual clues (nouns/NPs of address, previously addressing with *thou*, etc.).
- Bidirectional LSTM:
  - Takes the given context (sentence) into the account.
  - Timestep representations can be accessed at the index of a pronoun making it possible to classify more than one pronoun per sentence.

Methods and implementation III

- Implementation:
  - Jupyter Notebook for easy step-by-step execution.
  - Custom functions for extracting the annotated data and turning it into samples.
  - BERT embeddings sourced from the penultimate layer thereof.
  - PyTorch for Dataloaders and the model architecture itself.
    - Bidirectional LSTM
    - Classification layers: Dropout (0.05), Linear, LeakyReLU, Linear, Sigmoid.
    - BCELoss, Adam

Methods and implementation III

- Evaluation:
  - Accuracy, recall, precision, F1 (sklearn.metrics).
  - DataFrame with decoded classes.
  - Annotating another play (*Macbeth*).

Results I

- The project has been executed and is available on GitHub:
  - It has been re-run multiple times to try to pick the best hyperparameters.
  - The best performing model has been saved.
- The following measures have been recorded for this model:
  - Accuracy = 0.8
  - Recall = 0.7692307692307693
  - Precision = 0.8108108108108109
  - F1 = 0.7894736842105263

Results II

- Qualitative analysis: shorter sentences are more likely to be misclassified.
  - In testing data:
    - Beggar that I am, I am even poor in thanks; but I thank you, and sure, dear friends, my thanks are too dear a halfpenny.
      - Predicted: PL, true: PL
    - To you I give myself, for I am yours.
      - Predicted: PL, true: SG
  - In *Macbeth:*
    - Kind gentlemen, your_PL pains Are registered where every day I turn The leaf to read them.
    - [Enter Messenger.] What is your_PL tidings?

Conclusions

- Sentence-level context may not be enough:
  - Short sentences do not contain the key clues.
  - Sometimes out-of-utterance information could be relevant.
- Word embeddings:
  - A promising strategy as long as the source material is similar enough.
  - Still a fair bit of OOV and mis-parsed words.

Bibliography

- Adesam, Y., & Bouma, G. (2016). Old Swedish Part-of-Speech Tagging between Variation and External Knowledge. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 32–42.
- Bollmann, M. (2013). POS Tagging for Historical Texts with Sparse Training Data. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse,* 11–18.
- Busse, U. (2002). *Linguistic Variation in the Shakespeare Corpus: Morpho-syntactic variability of second person pronouns.* John Benjamins Publishing Company.
- Hiltunen, T., & Tyrkkö, J. (2013). Tagging Early Modern English Medical Texts (1500-1700). *Corpus Analysis with Noise in the Signal 2013* (conference).
- Hupkes, D., & Bod, R. (2016). POS-tagging of Historical Dutch. *Proceedings of the Tentch International Conference on Language Resources and Evaluation (LREC'16),* 77-82.
- Kulick, S., Ryant, N., & Santorini, B. (2022). Parsing Early Modern English for Linguistic Research. *Proceedings of the Society for Computation in Linguistics 2022*, 143–157.
- Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora. *Proceedings of Corpus Linguistics 2007*.
- Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011). Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 19–23.
- The Folger Shakespeare. N.d. *Download Shakespeare's Plays, Sonnets, and Poems.* Available from: https://shakespeare.folger.edu/download/
- Yang, Y., & Eisenstein, J. (2016). Part-of-Speech Tagging for Historical English. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1318–1328.