LING 131A
Prof. Marcus Verhagen
Group member: Xiaoyu Lu and Yonglin Wang
Nov 26, 2019

Final Project Proposal: Sentence-level Formality Classifier

Description

This project is inspired by the formality classifier modeling in Pavlick and Tetreault (2016) and the formality rules for ESL learners in Despres-Berry (2019). In this project, **we will train a text analyzer to determine the formality (binary: formal or informal) of a given sentence in four different genres (answers, blogs, emails, and news)**. We will train a ridge regression model using cross validation on the training data. Features we plan to use for training include, but are not limited to, case, lexical, length, POS, punctuation, readability, etc.

Dataset

- Primary source: UPenn formality corpus
  - annotated, containing both formal and informal sentences but no rewriting; medium-sized, ~6,700 sentences; safest option)
- Fancier source: he GYAFC corpus
  - annotated, informal sentences with their re-written formal counterparts; could be much larger (~10K pairs?); pending 2-step access request, ½ approved by now)
- Fallback option: see Sheikha & Inkpen (2010)

Packages Used

Scikit-learn, Stanford CoreNLP, TextBlob

Project Timeline

- Data gathering and package/feature decisions – now to Dec 2nd (~Assignment 5 due)
- Coding & Implementation: Dec 3rd - deadline

References (title hyperlinked to the papers)

Despres-Berry, C. (2019). Introduction to Advanced Communicative English. Lawrence University.

Pavlick, E., & Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4, 61-74.

Rao, S., & Tetreault, J. (2018). Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535.*

Sheikha, F. A., & Inkpen, D. (2010). Automatic classification of documents by formality. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)* (pp. 1-5). IEEE.

Tetreault, J. (2018). Under the Hood at Grammarly: Transforming Writing Style with AI. Retrieved November 25, 2019, from https://www.grammarly.com/blog/transforming-writing-style-with-ai/.