Identifying and Investigating the Feasibility of Cross-Domain Authorship Analysis

Daway Chou-Ren1, Dr. Byron Gao2
1Princeton University 2Texas State University

Introduction

- Traditional Problem
  - Given an anonymous document, can we identify which candidate’s writings samples it most closely resembles?
- Solution: Extract stylistic features from writing samples, use statistical or machine learning algorithms to classify unknown document
- Applications: the Federalist Papers, Shakespeare plays, poetry, newspaper articles, novels
- Commonality? Short samples, noisy, many candidates, but single-domain

An End-to-End System

- Extract Feature Vectors
- Model Validation
- Results and Discussion

Feature Set

- Word/sentence-based frequencies
  - 23 tokens
- Character-based frequencies
  - 63 a-z, 0-9
- Vocabulary richness metrics
  - 4 Schiffr’s S
- Capitalization types
  - 4 ALL CAPS
- Function word frequencies
  - 160 a, an, and
- Internet lingo frequencies
  - 116 lol, haha
- Part of speech tags and bigrams
  - 53 MN NNPS
- Syntactic parent-child pairs
  - 769 VB VBD
- Total
  - 1290

Results and Discussion

Model Validation

- Corpus # of Suspects Tokens per Suspect Accuracy Dummy Classification
- Federalist Papers 4 9,000 – 150,000 97% 11/12
- Sports Columns 6 2000 x 10 = 20,000 93% –
- Research Papers 3 7500 x 15 = 110,000 100% 11/15
- College Assignments 10 25,000 x 6 = 150,000 88% –

Defining Domains

- Same student may turn in a term paper similar to the Federalist Papers and lab report similar to a research paper
- Predicting College Assignments from each other is actually a cross-domain problem
- Two documents may be considered to exist in separate domains when required document structure, purpose, or audience changes dramatically, syntactic, or lexical patterns, but not content.
- Often form, audience, and purpose are intertwined–eg. blog posts vs online messaging vs academic essays
- Other times, only one of the three may change: emails to a friend vs to a coworker
- Abbasi et al’s VQ and clustering technique can be seen as attempting to find a single-domain solution from a cross-domain problem

Domain-Independent Feature Set

- Better experimental results are achieved using a neural network, though any classifier can be used. An aggregate ensemble fast correlation based filter works well for feature selection.

Initial Results

- Corpus # of Suspects Tokens per Suspect Dummy classification
- Facebook Posts from Facebook Messages 8 250 – 1500 5/8

Conclusion

- This study investigated authorship analysis from a new direction focusing on cross-domain analysis
- 1. We identified and defined cross-domain analysis as a future direction in authorship studies
- 2. We validated a single-domain model and demonstrated relative failure for cross-domain applications
- 3. We achieved positive initial results on a small sample set, demonstrating feasibility of a potential solution

Future Research

- Experiment with balanced feature set
- Expand cross-domain corpus
- Increase length of documents and number of samples
- More pre- and post-processing
- Test other domain combinations
- Blogs, essays, emails, tweets

Acknowledgements

This work was made possible by the Texas State University Computer Science Department, Benjamin Fung, and Noel Gong. This research was funded by NSF REU award #1358593.

References


Contact Information
Daway Chou-Ren: dchouren@princeton.edu