Identifying and Investigating the Feasibility of Cross-Domain Authorship Analysis

Daway Chou-Ren[1], Dr. Byron Gao[2]
[1] Princeton University [2] Texas State University

Introduction

Traditional Problem

- Given an anonymous document, can we identify which candidate's writings samples it most closely resembles?
- Solution: Extract stylometric features from writing samples, use statistical or machine learning algorithms to classify unknown document[1]
- Applications: the Federalist Papers, Shakespeare plays, poetry, newspaper articles, novels[2]
- Commonality? All print-based, large samples available, well-formed writing, same topic, few candidate authors

Contemporary Problem

- Can we identify shorter, noisier electronic documents that have more candidate authors?
- Solution: Increase feature sets, incorporating misspellings, emoticons, document structure, Internet lingo, etc. [3]
- Applications: chat logs, forum posts, emails, tweets[4]
- Commonality? Short samples, noisy, many candidates, *but* single-domain

Our New Cross-Domain Problem

- Is it possible to use writing samples to identify an unknown message from a different domain? Can a blog post be used to identify an email? Or a Facebook message a tweet?

Why?

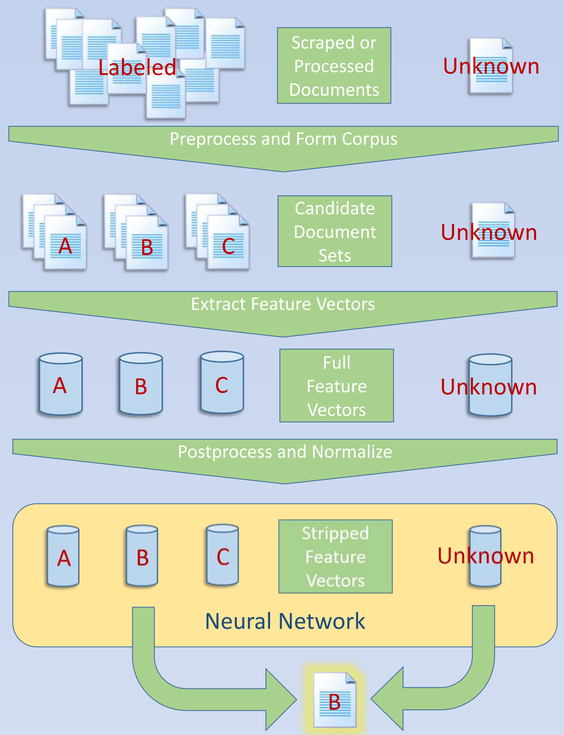
- Online domains allow for anonymity
- No way to get labeled posts from anonymous forum, email account, Facebook account, etc.
- Can hopefully find labeled text from another domain– emails from court injunction, old schoolwork, etc.

Model and Methodology

Feature Set

Feature	Count	Example
Word/sentence-based frequencies	23	# tokens
Character-based frequencies	63	a-z, 0-9
Vocabulary richness metrics	4	Sichel's S
Capitalization types	4	ALL CAPS
Function word frequencies	260	a, an, and
Internet lingo frequencies	116	lol, haha
Part of speech tags and bigrams	51	NN NNPS
Syntactic parent-child pairs	769	VB VBD
Total	**1290**	

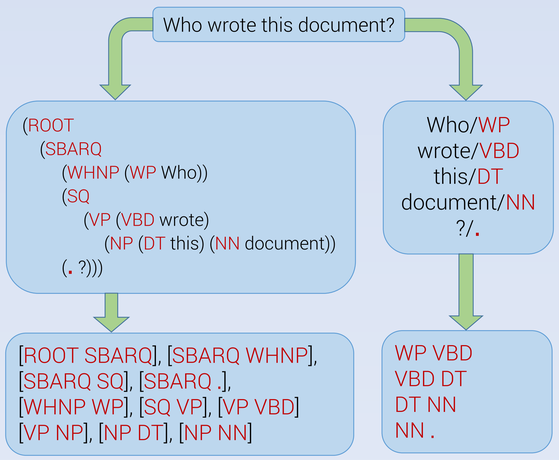
An End-to-End System



Best 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.

A Closer Look at Feature Extraction

Documents are split on multiple levels: by character, word, sentence, and by line. They are also tokenized for part of speech tagging and syntactic parsing through the Stanford NLP toolkit.



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 = ~100,000	100%	11/15
College Assignments	10	25,000 x 6 = 150,000	88%	–

Federalist Papers	Sports Columns	Research Papers
In the extent and proper structure of the Union, therefore, we behold a republican remedy for the diseases most incident to republican government.	June 2011: Detroit, $325 million October 2011: Philly, $280 million June 2012: New Orleans, $338 million October 2012: Memphis, $377 million	[3] proposes a proba- bilistic framework based on Hidden Markov Random Fields, incorporating supervision into k- clustering al- gorithms. [8]

- High accuracies for traditional problems
- High accuracies for contemporary problems
- Handles noise very well
- 5/9 misclassified documents for College Assignments were from author whose document set was split between journal entries and essays

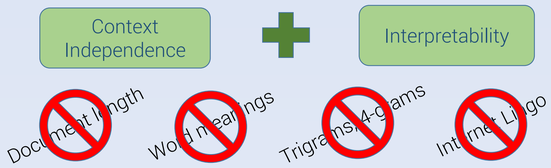
Defining Domains

- Same student may turn in a term paper similar to the Federalist Papers and a 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 structural, 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 Writeprint clustering technique can be seen as attempting to find a single-domain solution from a cross-domain problem[5]

Domain-Independent Feature Set



Initial Results

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

- Posts: brief, public reactions
- Messages: possibly length and private conversations
- Sample size too small, but tokens per suspect also small
- Additional difficulty dealing with insufficient tokens per suspect[6]

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 Neil Gong. This research was funded by NSF REU award #1358939.

References

1. Rudman, Joseph. "The state of authorship attribution studies: Some problems and solutions." Computers and the Humanities 31, no. 4 (1997): 351-365.
2. Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, *60*(1), 9-26.
3. Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556.
4. Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. Annual review of information science and technology, 43(1), 1-43.
5. Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS), 26(2), 7.
6. Layton, R., Watters, P., & Dazeley, R. (2010, July). Authorship attribution for twitter in 140 characters or less. In Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second (pp. 1-8). IEEE.

Contact Information

Daway Chou-Ren: dchouren@princeton.edu