CS341 - Authorship Attribution

Tim Althoff Denny Britz Zifei Shan

{althoff, dbritz, zifei}@stanford.edu

1 Problem Description

We attempt to deanonymize authors and attributing authorship in scientific publications by their writing style through linguistic modeling.

Characterizing individuals through idiosyncratic behavior has further applications in identifying and linking users across online communities as well as detecting fraud by misrepresenting to be someone else (e.g. by overtaking someone's Twitter account).

We choose to focus on attributing authorship in scientific publications since large datasets are more readily available.

We ask questions such as: Are double-blind paper submission process really double blind? Or can we predict authorship of a paper just from the writing style? Do some scientific (sub)fields have particularly simple or complicated language or are particularly easy or hard for attributing authorship?

2 Datasets

We plan to use scientific publications available through arXiv ($>$900k submissions total) using Amazon S3[1]. In case we will need to use a subsample of the data we will use a subset of popular authors such that we have a history of prior publications for each scientist.

[1] `http://arxiv.org/help/bulk_data_s3`

In parallel, we are also considering to use the ACL anthology[2], the scientific publications at PLOS[3], or to obtain publications for a set of authors through Google scholar.

One potential confound of the analysis is that some high profile publications are edited by technical writers that will likely remove many linguistic fingerprints by the authors. We expect that this will only affect a small number of publications in our dataset though.

3 Related Work

3.1 Sample Related work

[1]

4 Proposed Model

4.1 Feature Engineering

The first step is extracting useful features from datasets. We will propose a taxonomy of knowledge that might help attributing authorship, which might include:

1. **Corpus statistics:** frequent and surprising words, K-grams, and sentences.
2. **Shallow NLP features:** distribution of POS tags, NER tags, etc.
3. **Deep NLP features:** parse tree structure, frequent dependency paths, etc
4. **Knowledge base:** entities and relations in the article; what kinds of relations are the author supporting or attacking.
5. **Domain-specific features:**
    - article structure and flow organization, citation patterns, tables and figures, order of authors, author affiliations, etc.
    - *Non-double-blind features:* features that are not applicable in double-blind review process, e.g. explicit self-citations ("our work", etc).

[2] `http://acl-arc.comp.nus.edu.sg/`
[3] `http://www.plos.org/`

With a knowledge taxonomy, we will extract features from our datasets. Extracting features will blow up the size of data.

We propose to conduct experiments to engineer useful features. This study might cast insight to authorship attribution as well as general knowledge-driven applications.

4.2 Supervised Classification

With extracted features, we propose to train several kinds of state-of-the-art supervised classifiers, e.g. Logistic Regression and SVM.

However, traditional supervised classification is not expressive enough to capture some of the available knowledge. For example,

TODO: Examples that joint inference help; motivation for next subsection

4.3 Probabilistic Joint Inference

5 Evaluation Metrics

Precision and Recall, we have the ground truth. Comparison with other approaches. What are the challenges? Which set of candidate authors to use?

6 Team Members Bios

6.1 Tim Althoff

Tim Althoff is a PhD Student at Stanford University working with Jure Leskovec. Tim has prior experience in data mining and machine learning as well as natural language processing and has published in ICWSM, ACM MultiMedia, and ECCV. He has taken CS224W and CS246 at Stanford as well as classes on scalable machine learning at UC Berkeley (CS281A/B).

6.2 Denny Britz

Denny Britz is a Master's student at Stanford University working with Chris Re on DeepDive, a probabilistic system for automatic Knowledge Base Con-

struction. Denny previously worked at the UC Berkeley AMPLab where he helped build Spark, a distributed framework for cluster-computing. Denny has taken CS224W and CS246 at Stanford.

6.3 Zifei Shan

Zifei Shan is a Master's student at Stanford University working with Chris Re. Zifei is interested in data mining, specifically using structural knowledge and probabilistic inference to improve complex computer tasks. He has taken CS224W, CS246 and CS229 at Stanford.

References

[1] Herzenstein, M., Sonenshein, S., and Dholakia, U. M. Tell me a good story and i may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research 48*, SPL (2011), S138–S149.