Author Identification Using Semi-supervised Learning

Ioannis Kourtis and Efstathios Stamatatos

University of the Aegean

Outline

- Introduction
- The proposed method
  - Common n-grams
  - SVM
  - Semi-supervised learning
- Evaluation
  - Tuning the model parameters
  - Results
- Conclusions

Author Identification

- Authorship attribution vs. authorship verification
- Closed-set vs. open-set classification
- Text representation
  - Low-level (e.g., char n-grams) vs. high-level (e.g., syntactic) features
- Classification method
  - Profile-based vs. instance-based paradigm

One Text vs. Groups of Texts

- Most author identification methods are based on a fixed and stable training set
- There are many cases where we need to decide about the authorship of groups of texts
  - Alternatively, a long text (a book) of unknown authorship can be segmented into multiple parts
- Test sets can be used as unlabeled examples
- Semi-supervised learning methods can then be used

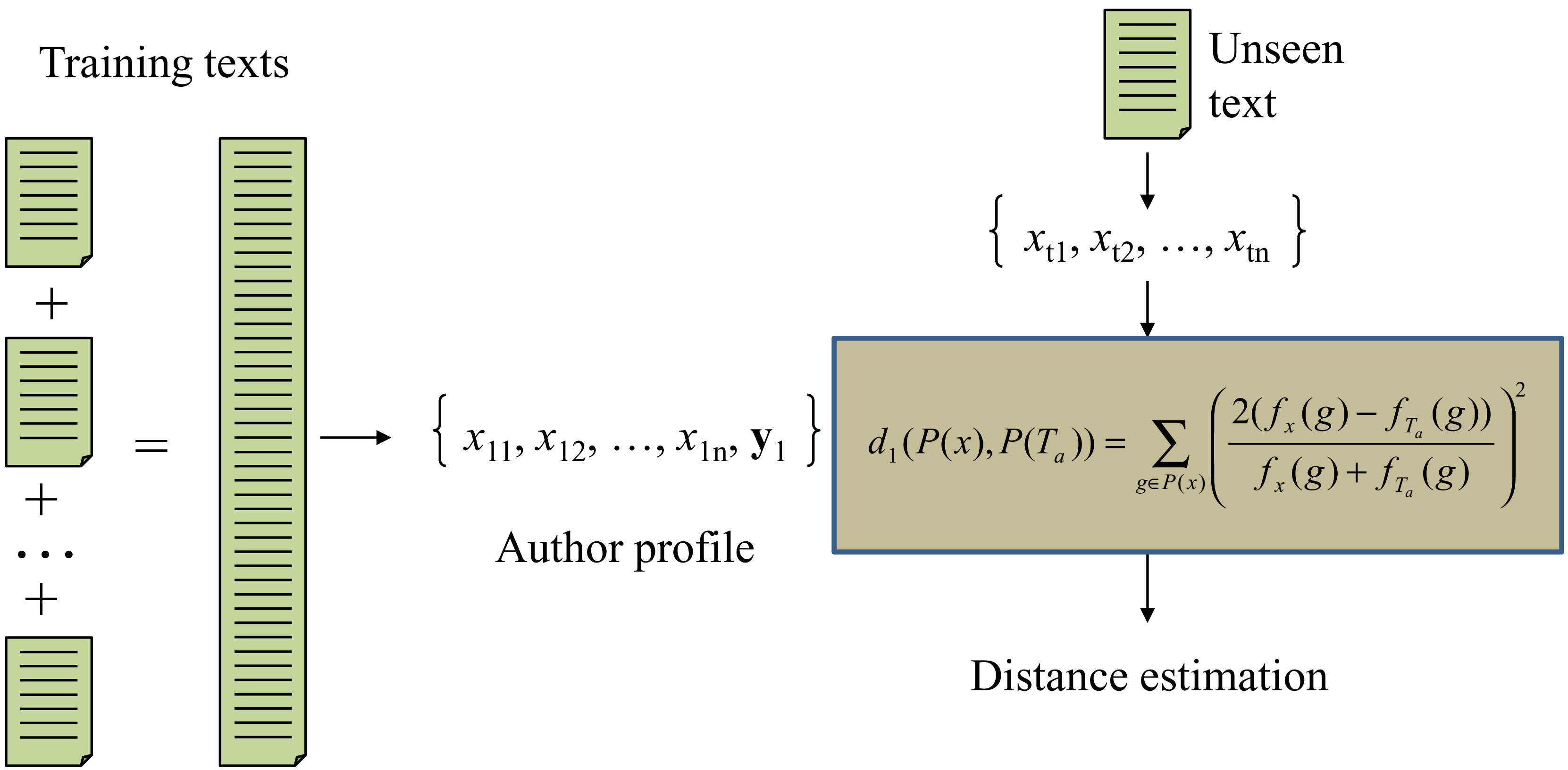
- Guzman-Cabrera et al. (2009) proposed the use of unlabeled examples found in the Web to enrich the training set

The Proposed Method

- We propose a combination of two well-known classification methods
  - Common n-grams
  - Support Vector Machines
- Both methods are based on character n-gram representation
- Test texts are used as unlabeled examples
- A semi-supervised learning method enrich the training set
- Applied to closed-set classification tasks

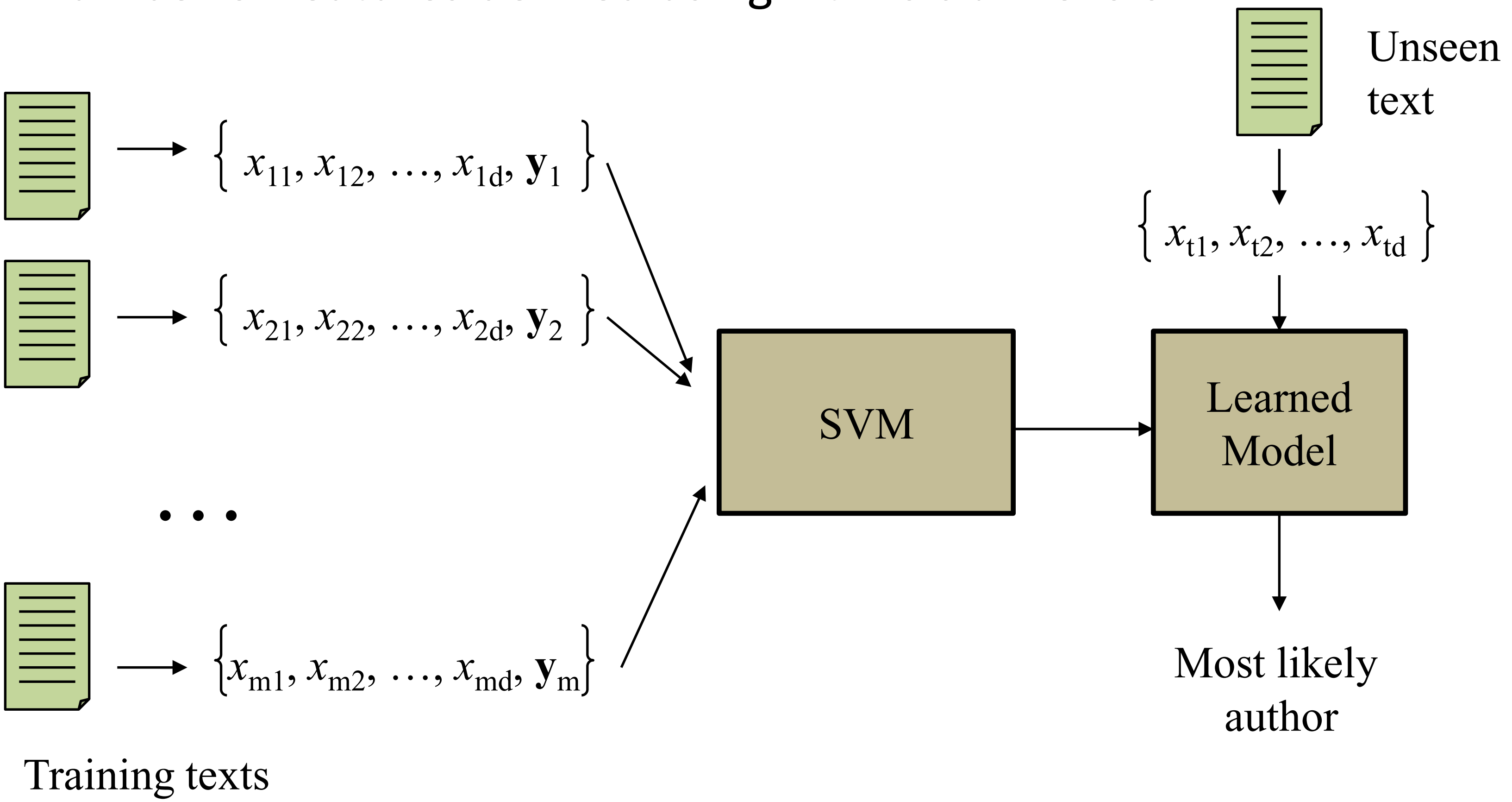
Common n-grams

- A profile-based method
- Originally proposed by Keselj et al. 2003
- Alternative dissimilarity measure proposed by Stamatatos, 2007



SVM

- Well-known and effective algorithm
- Character 3-gram representation
- Number of features defined using *intrinsic dimension*



Comparison

- CNG
  - Robust in class imbalance
  - Vulnerable when there are many candidate authors
  - Robust when distribution of training and test sets are not similar
- SVM
  - Vulnerable in class imbalance
  - Robust when there are multiple candidate authors
  - Robust when distribution of training and test sets are similar
  - Better exploitation of very high dimensionality

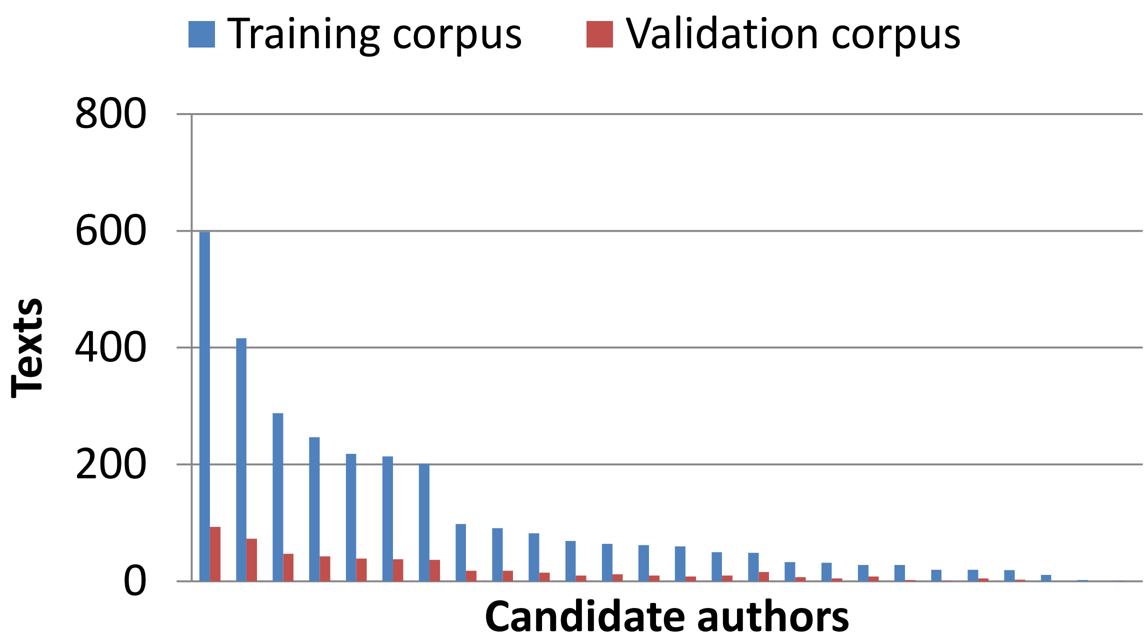
Semi-supervised Learning Algorithm

- Inspired by co-training (Blum & Mitchell, 1998)
- Given:
  - a set of training documents (labeled examples)
  - a set of test documents (unlabeled examples)
- Repeat
  - Train CNG and SVM models on the training set
  - Apply CNG and SVM models on the test set
  - Select test texts that CNG and SVM predictions agree
  - If text size is larger than a threshold move texts from test to training set
- Use SVM as default classifier for the remaining test texts

Comparison with Co-training

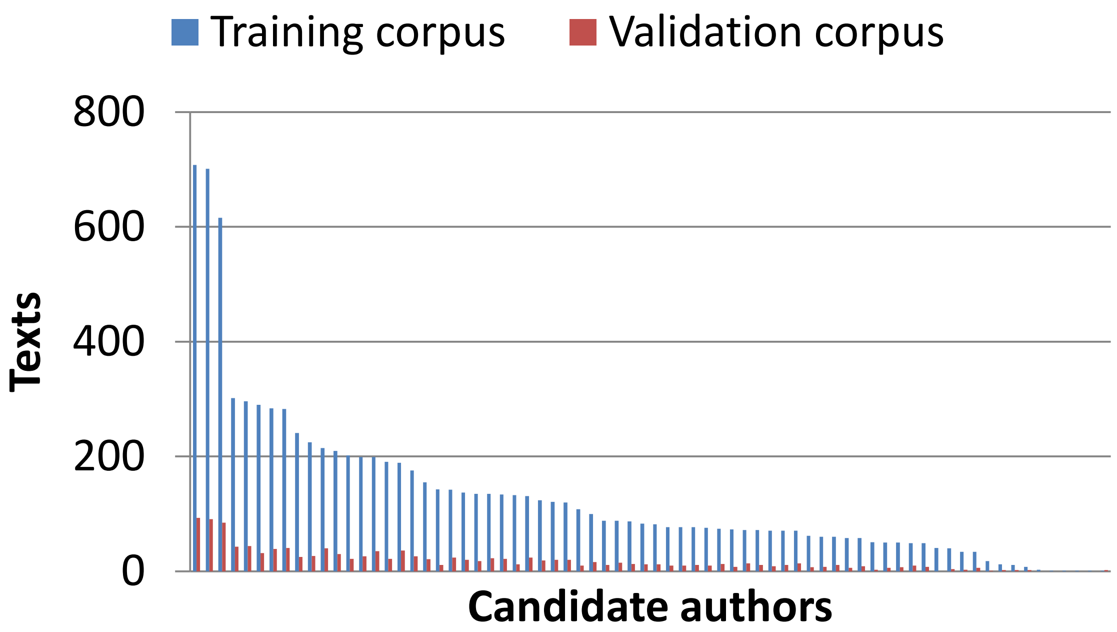
- Proposed algorithm:
  - Based on heterogeneous classifiers
  - Common feature types
  - Uses cases where the 2 classifiers agree
- Co-training:
  - Based on homogeneous classifiers
  - Non-overlapping feature sets
  - Uses cases where the 2 classifiers are most confident

Evaluation Corpora - Small



- 26 authors
- Imbalanced
- Similar distribution in training and validation sets

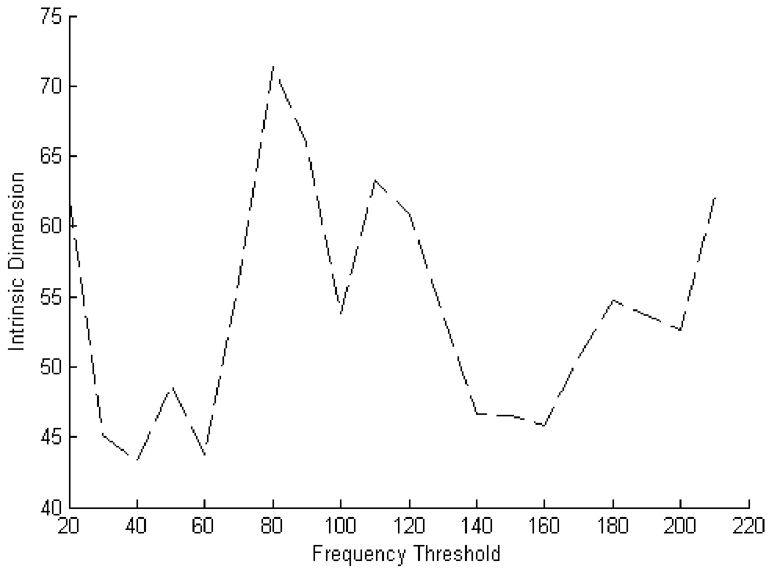
Evaluation Corpora - Large



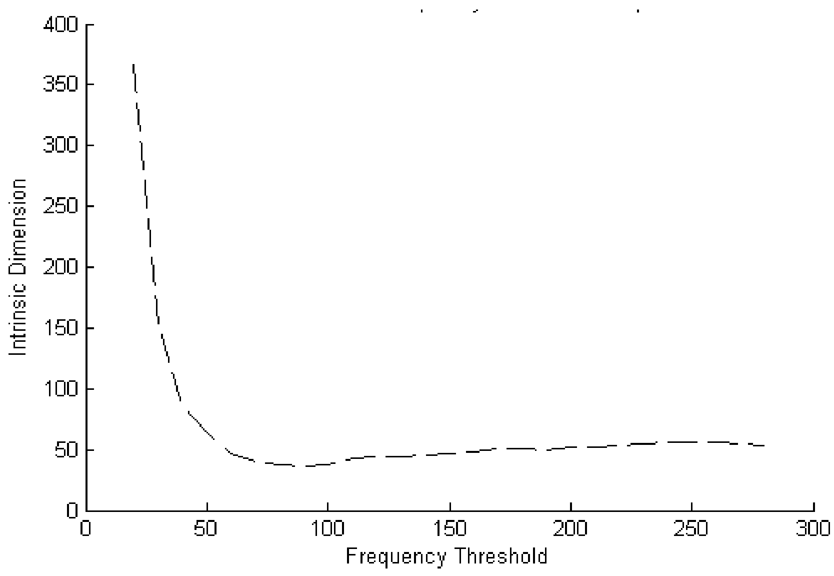
- 72 authors
- Imbalanced
- Similar distribution in training and validation sets

Frequency Threshold (SVM model)

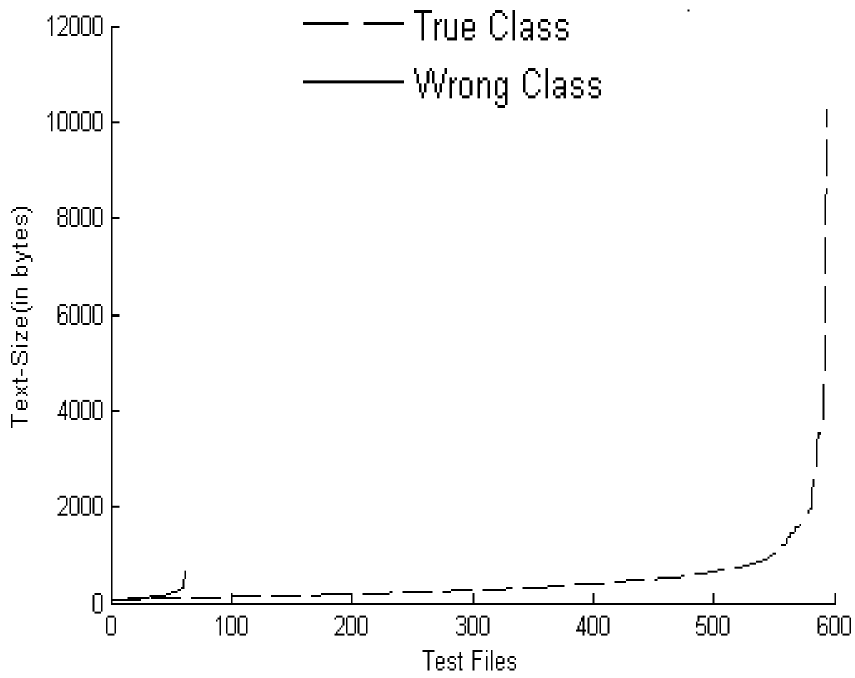
Small



Large



Text-size Threshold



- A threshold of 500 bytes excludes most of the cases where the two models agree but the predicted author is not the correct answer

Settings

- Labeled examples:
  - Training and validation sets
- Unlabeled examples:
  - Test set
- CNG
  - n=3, L=3,000
- SVM
  - n=3, max intrinsic dimension

Performance

Corpus	MacroAvg Prec.	MacroAvg Recall	MacroAvg F1	MicroAvg accuracy	Rank
Small	0.476	0.374	0.38	0.638	7/17
Large	0.549	0.532	0.52	0.658	1/18

Conclusions

- First attempt to apply semi-supervised learning to author identification
- Encouraging results for closed-set tasks
- Character n-gram representation proves to be very effective
- More diversity is needed in the classifier decisions
- Plan to extend this approach to open-set tasks