Authorship Identification with Modality Specific Meta Features

Thamar Solorio, Sangita Pillay, Manuel Montes,

Natural Language Processing Lab
University of Alabama at Birmingham

Introduction

- Authorship attribution assumes unique and identifiable writeprints in text.
- But similarities exist among authors **across specific linguistic dimensions**.
- We want to take advantage of these similarities to improve prediction accuracy.

Proposed approach

- **Idea:** Exploit independent clustering of linguistic modalities to generate meaningful meta features
- **Assumption:** The individual processing of linguistic modalities will allow the extraction of relations in the writeprint of authors, and these relations will be unique for each author.

More specifically

1 Document representation

A document $\mathbf{x}$ is represented as $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ where m is the number of modalities, and each $\mathbf{x}_i$ is a vector with $|\mathbf{x}_i|$ features in modality i

Note that

- $\textbf{union}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m) = \mathbf{x}$
- $\textbf{intersection}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m) = \emptyset$

2 Generating meta features

- Each of the m different vectors are input to a clustering algorithm
- Output= m clustering solutions for the training data with k clusters each
- Note this is an unsupervised step, **no class information is included**

More specifically

2 Generating meta features

- From each cluster c_j in each of the m clustering solutions, we compute a centroid by averaging all the feature vectors in that cluster.

$$\textbf{centroid}_{m_j} = \frac{1}{\mid c_{m_j} \mid} \sum_{x_i \in c_{m_j}} \mathbf{x}_i \tag{1}$$

  where j above ranges from 1 to k, the number of clusters.
- Meta features = the *similarity* of each instance to these centroids using the cosine function.
- Each instance $\mathbf{x}$ is now represented by the original set of first level features $\langle \mathbf{x}_{i_1}, ..., \mathbf{x}_{i_{|x_i|}} \rangle$ in combination with the meta features $\langle \mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k} \rangle$ generated for each modality j.

First level features

Four linguistic modalities:

1. Lexical features
2. Stylistic features
3. Perplexities from language models
4. Syntactic features

Note that these features were selected for AA in posts from web forums[1], **no customization** was performed for the PAN data.

[1] Solorio et al. (to appear in IJCNLP'11)

First level features

Modality	Features
Stylistic	Total number of words Average number of words per sentence Binary feature indicating use of quotations Binary feature indicating use of signature Rate of all caps words Rate of non-alphanumeric characters Rate of sentence initial words with first letter capitalized Rate of digits Number of new lines in the text Average number of punctuations (!?.;:,) per sentence Rate of contractions (won't, can't) Rate of two or more consecutive non-alphanumeric characters
Lexical	Bag of words (freq. of unigrams)
Perplexity	Perplexity values from character 3-grams
Syntactic	Part-of-Speech (POS) tags Dependency relations Chunks (unigram freq.)

Table: Feature breakdown by modality

Experimental settings

- We used WEKA's implementation of SVMs
- For clustering we used CLUTO
  - Parameter for the number of clusters
  - k =number of authors $\times 15$
- **Baseline system**: training and testing the model with only first level features (FLF)
- No out of training author experiments

Results

System	TestSet	MacroAvg Precision	MacroAvg Recall	MacroAvg F1	MicroAvg Precision	MicroAvg Recall	MicroAv F1
Baseline	Large	0.119	0.054	0.041	0.155	0.155	0.155
MSMF	Large	0.171	0.084	0.066	0.148	0.148	0.148
	Change	43.6%	55%	60.9%	-4.5%	-4.5%	-4.5%
Baseline	Small	0.440	0.152	0.148	0.384	0.384	0.384
MSMF	Small	0.415	0.205	0.185	0.440	0.440	0.440
	Change	-5.6%	34.8%	25%	14.5%	14.5%	14.5%

Table: Comparison of micro and macro averaged precision, recall, and F1 values in two PAN'11 test sets. MSMF stands for our modality specific meta features approach.

Concluding remarks

Lessons learned

- Meta features helped improve accuracy, for the most part
- Feature selection is a must

Current work

- Understand better the role of the meta features
- Need to handle out of training authors
- Evaluate the influence of modality specific features
- Develop new approaches to exploit the linguistic modalities

Thank you for your attention!
And many thanks to the PAN organizers