Robust Features for Computational Stylometry

Walter Daelemans
University of Antwerp
CLiPS, department of linguistics
walter.daelemans@ua.ac.be

RANLP 2009

"Computational Stylometry"

- Combinations of linguistic invariants in text, not under the conscious control of the author, can be used to determine
  - Individual authors (authorship attribution)
  - Characteristics of authors
    - Gender detection
    - Region, age, education level detection
    - Personality
    - Period detection (dating)
- As opposed to topic / register / genre …

Contents

- Applications of stylometry
- The current standard model: automatic text categorization
- Ideology from text
- Personality from text
- Robust features for Authorship attribution
  - Many authors
  - Short texts

Stylometry in Antwerp

- FWO project
  - Goals
    - Find methodology that is suited to find these invariants and use them in prediction
      - Many potential authors
      - Small sized training data (few paragraphs)
    - Develop software package / library (TACTICS)
    - Attract (humanities) students
  - People
    - Kim Luyckx (PhD)
    - Mihai Tolea
    - Guy De Pauw

Stylometry in Antwerp

- PhD project Mike Kestemont
  - Scribe / author detection in medieval manuscripts
- Advanced MA project Senja Pollak
  - Distinguishing Kenyan from Western media writing in English about the Kenyan elections
- Using stylometry techniques to check for signs of Alzheimer disease in later work of Hugo Claus
- Diagnostic tests for schizophrenia
- Disputed authorship in French theatre
- …




Applications

- Forensic uses
- Customer Relations Management
- Literary and philological studies
- Pragmatics studies (culture and ideology)
- Semantic Web automatic meta-information assignment
- Plagiarism detection (?)

Plagiarism Detection

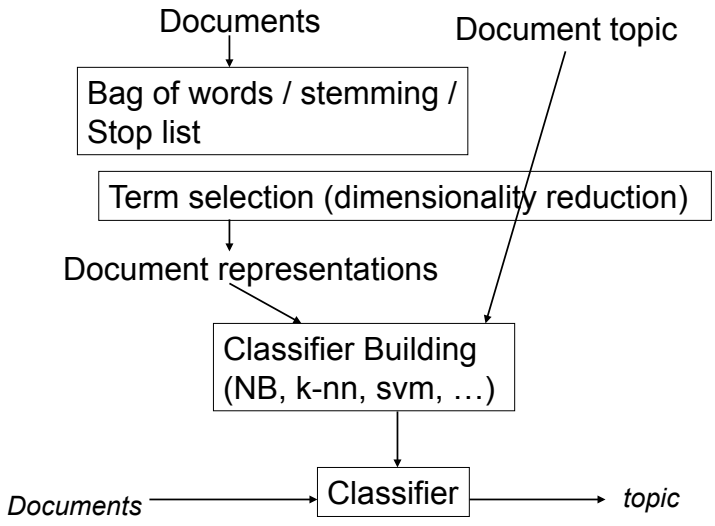
- Current plagiarism detection software
  - based on string matching
  - has severely limited usefulness
    - only works when the plagiarized text is on the WWW or in a user database
- Solution: *Linguistic profiling* (Van Halteren, 2007, ACM TSLP) of an author
  - based on texts known to be written by him/her
  - text that doesn't match the author's linguistic profile is suspect
  - plagiarism detection = authorship attribution

Current approach

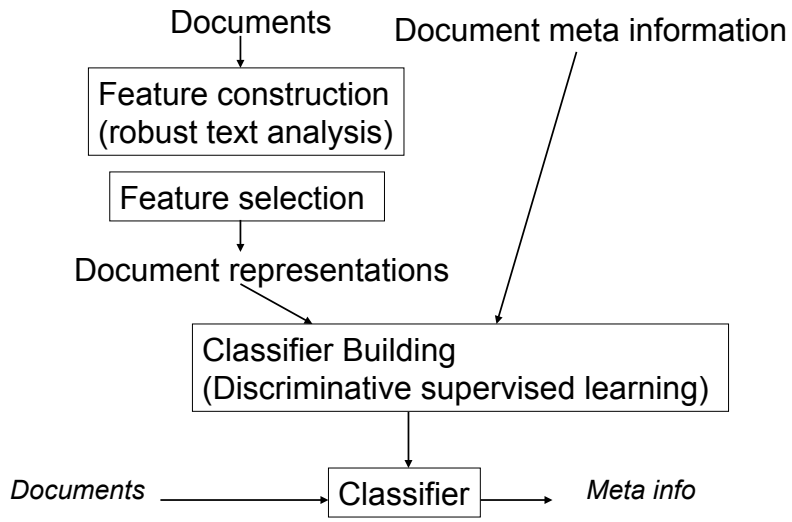
Automatic text categorization

Automatic Text Categorization

(e.g. Sebastiani, 2002, Computing Reviews)



General Methodology



Milestone: Gender assignment

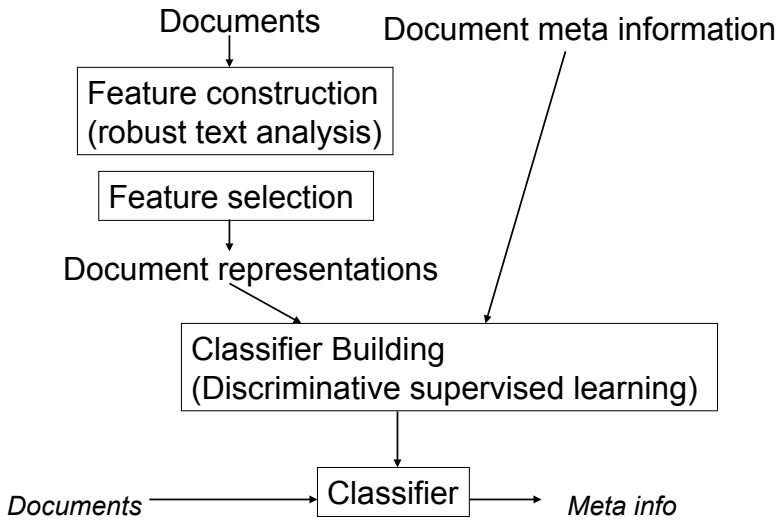
(Koppel, Argamon, Shimoni, Literary and Linguistic Computing, 2002)

- Documents: British National Corpus (fiction and non-fiction)
- Meta-data: gender of author
- Feature construction:
  - lexical (Function Words)
  - POS (Function Words)
- Supervised learning: linear separator
- Results: gender ~ 80% predictable

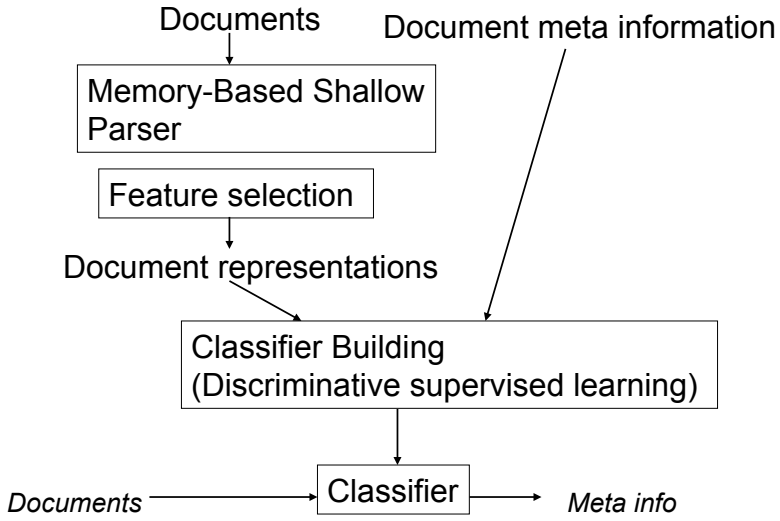
Gender Differences

- Use of pronouns (more by women) and some types of noun modification (more by men)
  - *"Male" words: a, the, that, these, one, two, more, some*
  - *"Female" words: I, you, she, her, their, myself, yourself, herself*
- More "relational" language use (by women) and more "informative" (descriptive) language use by men
- Even in formal language use!
- Strong correlation between male language use and non-fiction, and female language use and fiction

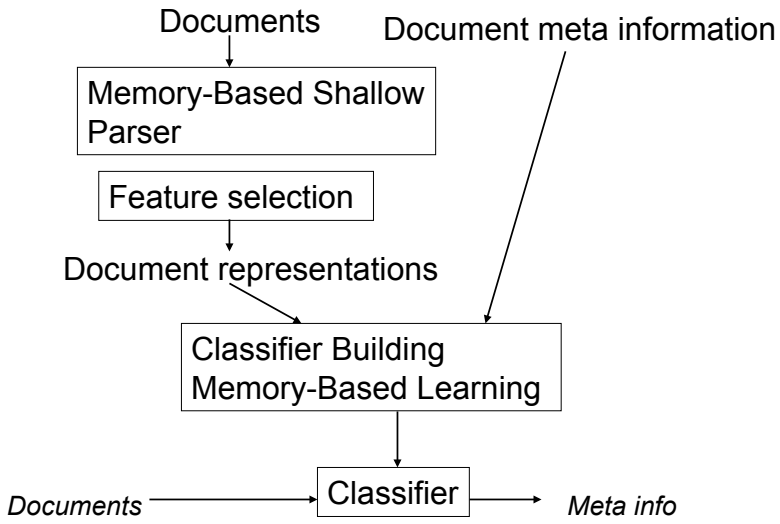
CLiPS Stylometry Environment

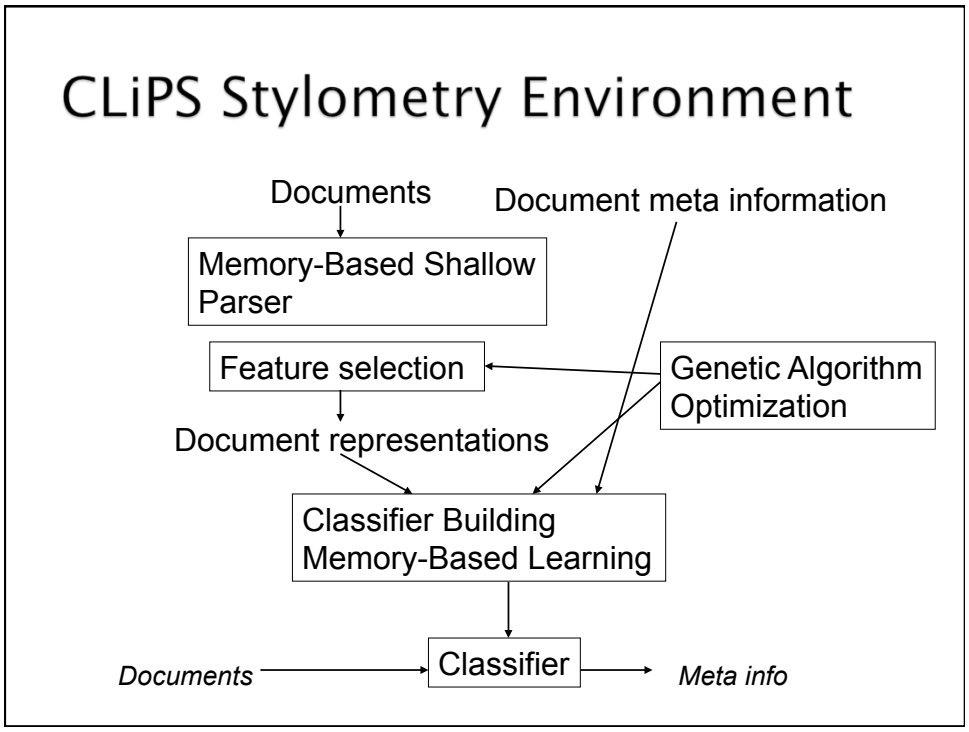


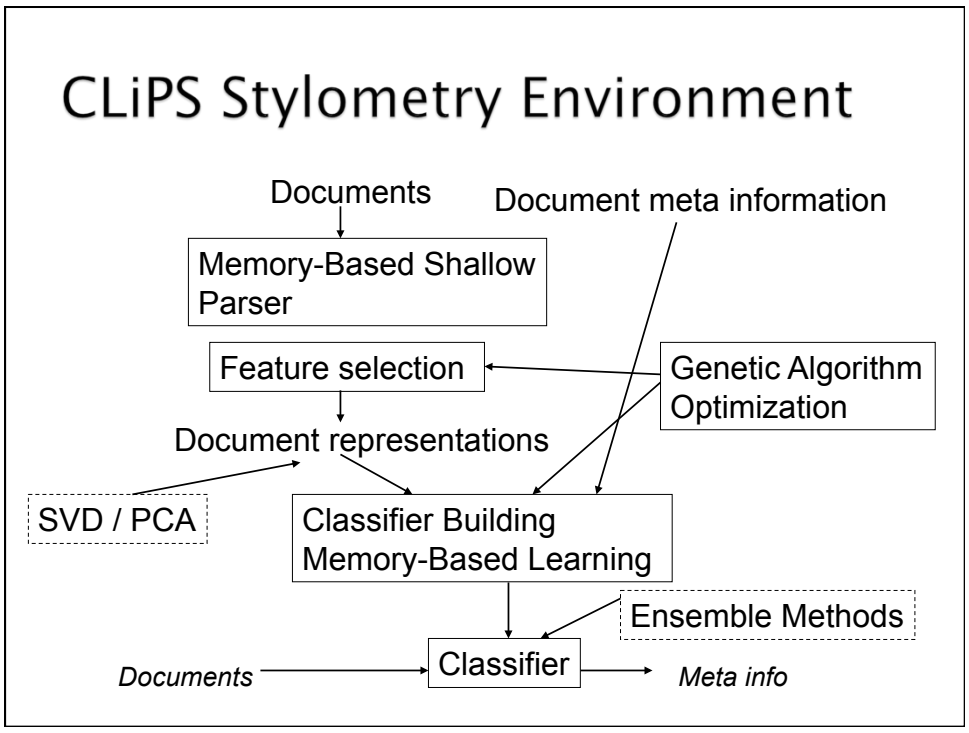
CLiPS Stylometry Environment



CLiPS Stylometry Environment




CLiPS Stylometry Environment
Documents
Document meta information
Memory-Based Shallow Parser
Feature selection
Genetic Algorithm Optimization
Document representations
Classifier Building Memory-Based Learning
Documents
Classifier
Meta info


CLiPS Stylometry Environment
Documents
Document meta information
Memory-Based Shallow Parser
Feature selection
Genetic Algorithm Optimization
Document representations
SVD / PCA
Classifier Building Memory-Based Learning
Ensemble Methods
Documents
Classifier
Meta info

ML is NOT easy

- What influences outcome of ML experiment?
  - Information sources
    - Feature construction, selection and representation
  - Training data
    - Size
    - Training data properties
  - ML algorithm
    - Bias
    - parameters
- Interactions:
  - E.g. Feature selection and algorithm parameters

GA results on authorship attribution

	Timbl memory-based learner
Default	69.4
Feature selection (fw-bw)	72.6
Parameter Optimization	70.5
GA joint FS and PO	

GA results on author detection

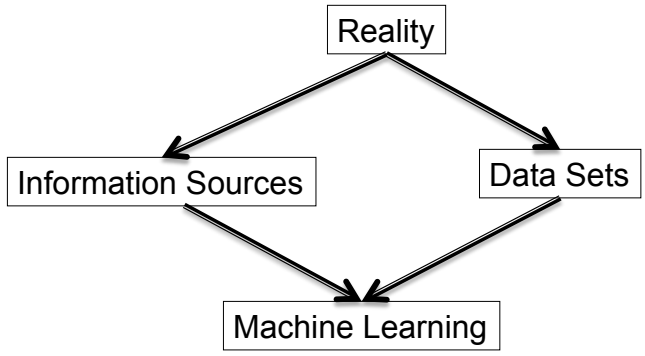
	Timbl memory-based learner
Default	69.4
Feature selection (fw-bw)	72.6
Parameter Optimization	70.5
GA joint FS and PO	80.1

Ideology from text

Ideology from Text

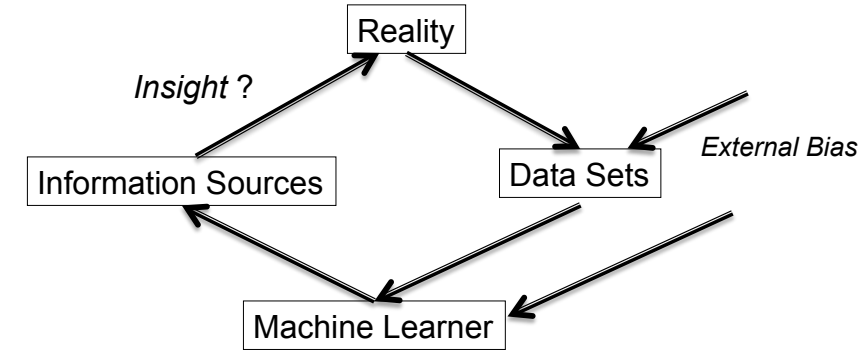
- AMA thesis Senja Pollak
- 2008 post-election crisis Kibaki versus Odinga
- Western versus local media coverage
- Goals
  - Can we predict the source of a news article?
  - Do we find cultural or ideological differences in analyzing informative features?

How to (ab)use ML



Accuracy

How to (ab)use ML



Experiment

- 464 documents
  - Same period
  - 50-50 western versus local
- Features
  - Unigrams, bigrams, trigrams
  - 500 best features using chi-square
  - Binary or frequency vector representation
- Rule Induction techniques

Results

- Accuracy 90-95% (10-fold CV)
- Differences:
  - Referring to the candidates
    - Local: ODM leader (Raila) (Odinga)
    - Western: Opposition leader
  - Referring to tribal divisions
    - Western: tribe, tribal, Kikuyu
  - Referring to (primitive) violence
    - Western: machetes, sticks, burned
  - Use of titles
    - Local: Mr., Dr., Prof.,…

Personality from Text

Personae Corpus

- Collected November 2006
- 200,000 words, Dutch
- 145 BA students (from a population of ~200) in a course on interdisciplinary linguistics
- Voluntarily watched the same documentary on Artificial Life (but received 2 cinema tickets as incentive)
  - Topic, genre, register, age held constant

Personae Corpus

- Wrote a text of ~ 1200 words
  - Factual description + Opinion
- Did an on-line personality test
- Submitted their profile, the text and some user information via a web-site
- All text processed with MBSP (memory-based shallow parser)
  - Tokenizer / Tagger / Chunker / Relation Finder

Personality Detection

- Are personality traits such as extraversion reflected in writing style?
- Seminal work by Gill & Oberlander on extraversion and neuroticism
  - Not in a prediction (classification) context but in a descriptive statistics context
    - Disregards effect of combinations of features
  - Based on e-mail
- Parallel work on prediction
  - Argamon et al., 2005; Nowson & Oberlander, 2007; Mairesse et al., 2007

Previous hypotheses and observations

- Extraverts
  - Use fewer hedges (confidence)
  - More verbs, adverbs and pronouns (vs. nouns, adjectives, prepositions)
  - Less formal
  - Fewer negative emotion words more positive emotion words
  - Fewer hapaxes
  - More present tense verbs
  - Fewer negation and causation words
  - Fewer numbers and less quantification
  - Less concrete

Meyers-Briggs

- Forced-choice test
- < Carl Jung's personality typology
- Categorization according to 4 preferences:
  - **I**ntroversion & **E**xtraversion (attitudes)
  - i**N**tuition & **S**ensing (information-gathering)
  - **F**eeling & **T**hinking (decision-making)
  - **J**udging & **P**erceiving (lifestyle)

Meyers-Briggs

- Leads to 16 types: ENTJ (1.8%) … ESFJ (12.3%)
- Mental functions: ST, SF, NT, NF
- Attitudes: TJ, TP, FP, FJ
- Temperaments: SP (artisan), SJ (guardian), NF (idealist), NT (rational)
- MBTI correlates with "Big Five" (OCEAN) personality characteristics extraversion and openness, to a lesser extent with agreeableness and conscientiousness, but not with neuroticism
- Validity and reliability have been questioned

Participant characteristics

- Too homogeneous for some experiments
  - 77% female
  - 97% native speaker of Flemish-Dutch
  - 77% from Antwerp region
- MBTI dichotomies:
  - E 80 vs. I 65
  - N 78 vs. S 67
  - F 105 (72%) vs. T 40
  - J 117 (81%) vs. P 28

Participant characteristics

28 ESFJ (provider)
23 ENFJ (teacher !)
16 ISFJ (protector)
15 INTJ (mastermind !)
15 INFJ
9 ENFP
8 ISTJ
8 ENTJ

6 ESFP
4 ISFP
4 INFP
4 ESTJ
3 INTP (architect)
1 ESTP (promoter)
1 ENTP (inventor)
0 ISTP (crafter)

Our typical student

- Flemish girl from around Antwerp who likes people and is warm, sympathetic, helpful, cooperative, tactful, down-to-earth, practical, thorough, consistent, organized, enthusiastic, and energetic. She enjoys tradition and security, and will seek a stable life that is rich in contact with friends and family
- (but she is not interested in Computational Linguistics) :-)

Features

- Feature selection: $\chi 2$ metric
- Binary or numeric
- Lexical
  - N-grams (n: 1-3)
  - Function word distributions
- Syntactic
  - N-grams (n: 1-3) of coarse-grained and fine-grained POS
- Readability
- Type / token (vocabulary richness)

Task	Feature set	Precision	Recall	F-score
Introverted	word 3-grams	87.69%	58.16%	69.94%
	random	*44.1%*	*46.2%*	
Extraverted	POS 3-grams	100.00%	56.74%	72.40%
	random	*54.6%*	*52.5%*	
iNtuitive	POS 3-grams	84.62%	64.71%	73.33%
	random	*48.7%*	*48.7%*	
Sensing	POS 3-grams	85.07%	56.44%	67.86%
	random	*40.3%*	*40.3%*	
Feeling	readability	100.00%	73.43%	84.68%
	random	*72.6%*	*73.3%*	
Thinking	word 2-grams	72.50%	39.19%	50.88%
	random	*28.2%*	*27.5%*	
Judging	word 3-grams	81.82%	100.00%	90.00%
	random	*77.6%*	*76.9%*	
Perceiving	word 2-grams	60.71%	36.96%	45.95%
	random	*6.9%*	*7.1%*	

Table: Results for the eight binary classification tasks with TiMBL

Predictive features per class

- I – ; conclusie principes misschien meer_inzicht
- E ! uitvoeren valt we zij zelf de_mens
- N aangezien simuleren term de_mogelijheid
- S gebeurt hersenen tastzin een_spontaan
- F ! beste denk ik toch een_beetje
- T : constructies stromingen theorie omdat de_socio-politieke
- J mechanisme proces systeem dankzij mijn_mening
- P ontwikkelingen symbiose tijdens van_levende

Discussion

- First two personality dimensions can be predicted fairly accurately
- Good results in 6 out of 8 binary classification tasks
- Even with skewed class distributions (.28 or .19 for positive class), still around 51% and 46% F-score
- Syntactic features work for personality prediction
- Unclear whether accuracy levels are high enough to make this useful beyond academic interest

Authorship Attribution

Robust Features

Author Attribution

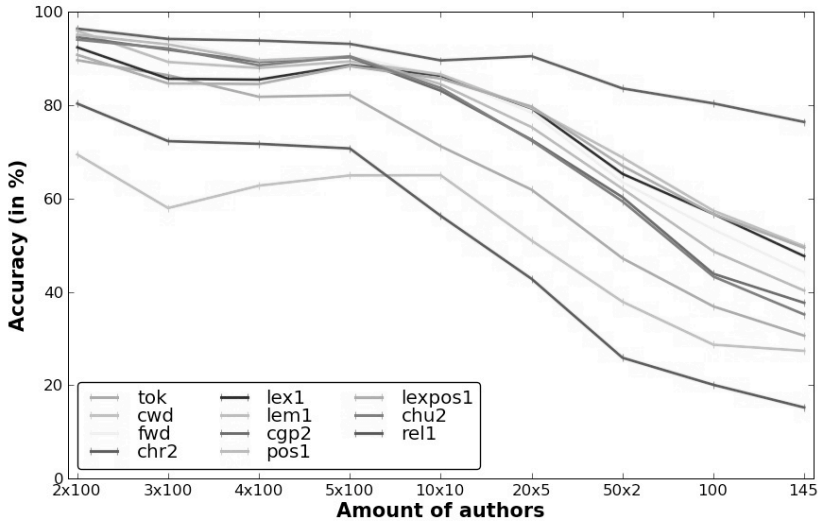
- Long tradition in "humanities computing" (Holmes, Burrows, Baayen, van Halteren, Juola, Hoover, …)
- Features:
  - Word length, sentence length, n-grams, distribution of POS tags, frequencies of rewrite rules / chunks, word frequency, vocabulary richness, …
- Mostly two or a few authors and long texts
- Generalizes to many authors and short texts?
  - Personae corpus allows study of distribution of features over large set of authors

Features

- Every essay divided into 10 parts, 8 in training, 2 in testing (5-fold cross-validation)
- Feature selection: $\chi 2$ metric
- Binary or numeric
- Character n-grams (n: 1-3)
- Lexical
  - Word N-grams (n: 1-3)
  - Lemma n-grams
  - Function and content word distributions
- Syntactic
  - N-grams (n: 1-3) of coarse-grained and fine-grained POS
- Readability
- Type / token (vocabulary richness)

Effect of number of authors

- Identifying one of 2, 5, 10 authors may be an easy task compared to identifying one of 145 (and more)
- Feature types working well for these easier cases may not work anymore in the more difficult case

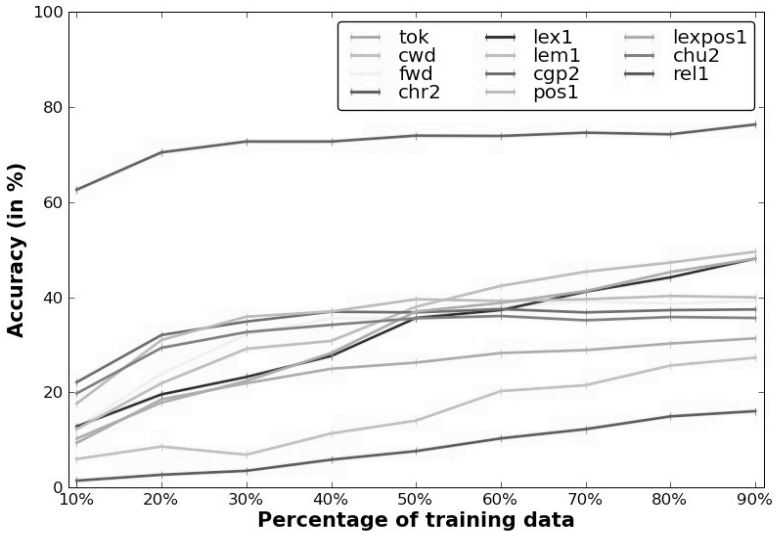


Observations

- Significant decrease of performance with more authors (96% to 76% for the best feature set)
  - Note: second-best, lemmas = 96% to 50%!
- Character n-grams are most robust
  - Syntactic features, function words, and lexical features start out fine but deteriorate quickly
- Robustness of features is robust itself (few crossing curves as number of authors increases)

Effect of training data size

- Identifying authors on the basis of large training data sets may be very easy compared to only small training snippets
- Feature types working well for small data may not work very well for large data and vice versa



Observations

- Significant increase of performance with more data (62% to 76% for the best features)
  - Note: second-best, lemma = 12% to 50%!
- Character n-grams are most robust
  - Syntactic features, function words, and lexical features start out poorly and increase less with more data
- Robustness of features is robust itself (few crossing curves as size of data increases)

No hope for linguistics?

- Although character n-grams are surprisingly robust and accurate over many authors and small datasets, and are language independent, all is not lost
- Character n-grams combined with linguistic features lead to large error decreases:
  - Character n-grams + pos information in 145 authors
    - 76% -> 89% (> 50% error reduction)
  - Character n-grams + pos information in 10% data
    - 62% -> 68% (> 15% error reduction)
    - 40% error reduction on 20% of the data

Why do char n-grams work so well?

- Good trade-off between sparseness and information
- Implicit punctuation, morphology, semantics, … style?
- Best character bigrams:
  - '- "- -" -' .l -, ,- a. .. _- -_ l. .i i. wh tm oq !. -! zv ik

Conclusions

- Robust text analysis + Machine learning in a text categorization framework is a powerful combination for inferring meta-data about text
- Ideology from text
  - Results are encouraging at best
- Personality from text
  - Results are encouraging at best

Conclusions

- Authorship attribution in the face of small datasets and many potential authors
  - Close to useful and usable
  - Character n-grams combined with robust text analysis (POS) leads to 90% with 145 authors and 70% with only a hundred words
- Basic research problem remains
  - Text characteristics are the result of many interacting factors. How do we factor out only those of interest?