Authorship Attribution

Efstathios Stamatatos

Πανεπιστήμιο Αιγαίου
University of the Aegean

Tutorial Layout

- Introduction
- Tasks, applications
- Stylometry
- Attribution paradigms
- Evaluation, resources
- Summary

Introduction

- Plethora of electronic texts in Internet media
- Need for efficient handling of this information
- Boost in research:
  - Information Retrieval
  - Machine Learning
  - Natural Language Processing
- Text mining
  - Text categorization
  - Text clustering
  - …

Text Categorization

- The task of approximating the target function $\Phi : D \times C \rightarrow \{T, F\}$
  - *D*: documents
  - *C*: categories
- Binary vs. multi-class
- Single-label vs. multi-label
- Closed-set vs. open-set
- Hierarchical vs. flat
- Crisp vs. ranking

Text Categorization Criteria

- Topic
  - Filtering of newswire stories
  - Indexing of scientific articles
  - Spam filtering
  - … [Sebastiani, 2002]
- Opinion
  - Sentiment analysis [Pang and Lee, 2008]
- Style
  - Authorship
  - Genre

Style-based Text Categorization Tasks

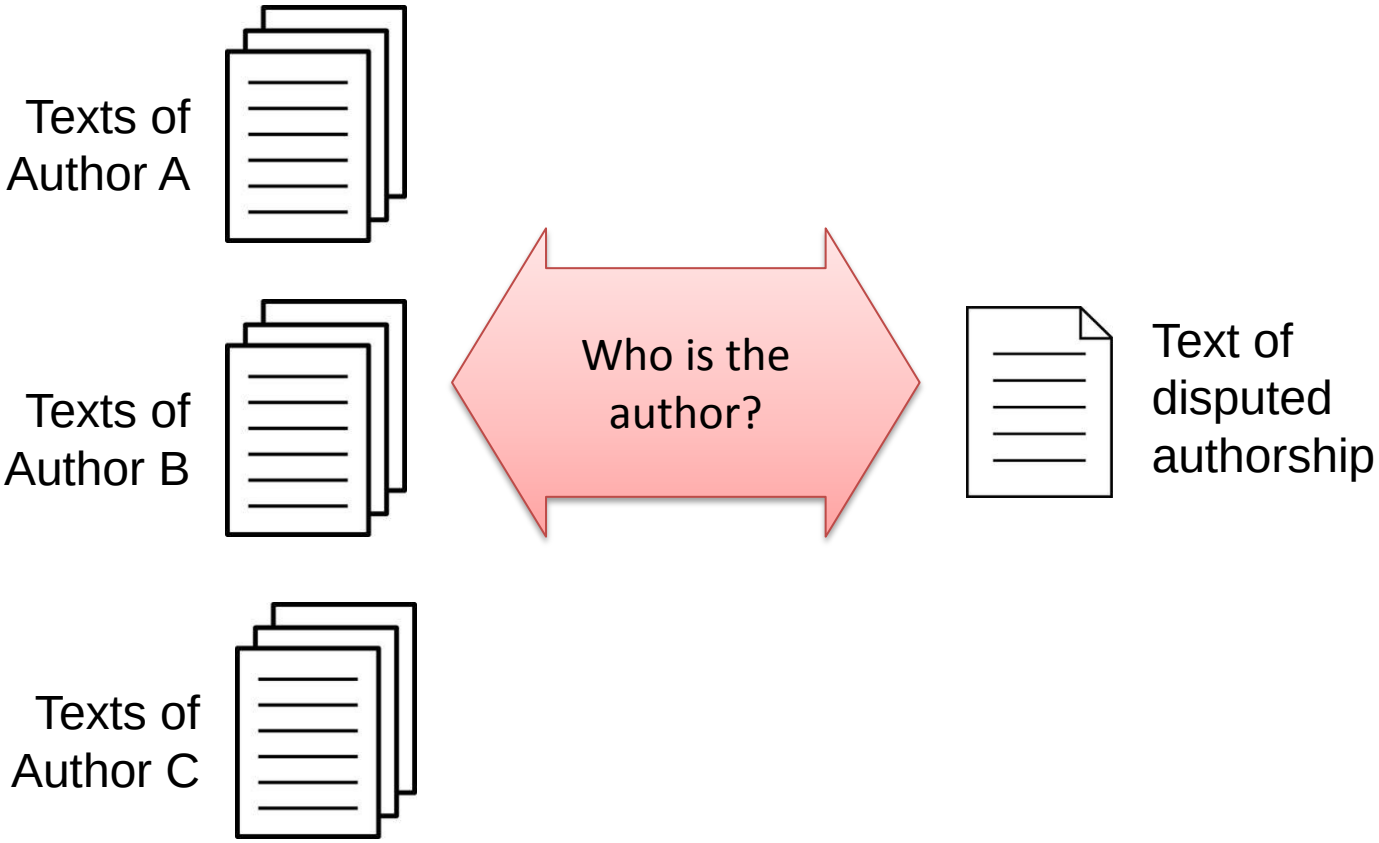
- Authorship analysis
  - Deals with the personal style of the authors
- Genre analysis
  - Deals with the form and communicative purpose of documents

Authorship Analysis

- It has a long history [Mendenhall, 1887]
- A seminal study by [Mosteller & Wallace, 1964] introduced *non-traditional* approaches and provided evidence on the *Federalist Papers* case
- By late 1990s the focus was on examination of literary cases of unknown or disputed authorship [Holmes, 1998]
- During the last decade, it is applied to modern genres (online newspaper, blogs, forum messages, emails, tweets, etc.) [Stamatatos, 2009]

Authorship Analysis Tasks

- Author identification (aka authorship attribution)
  - Given a set of candidate authors and some texts by them, to attribute an unseen text to one of them

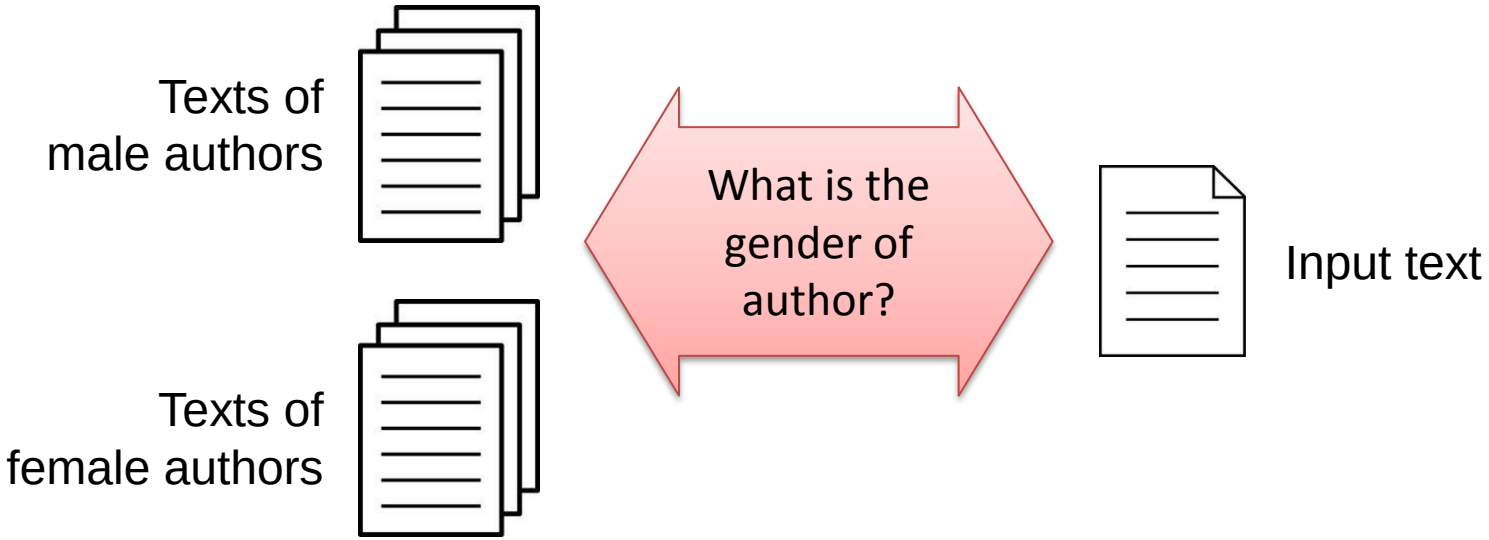


Authorship Analysis Tasks

- Author verification
  - Given texts of a certain author, to decide whether an unseen text was written by that author or not

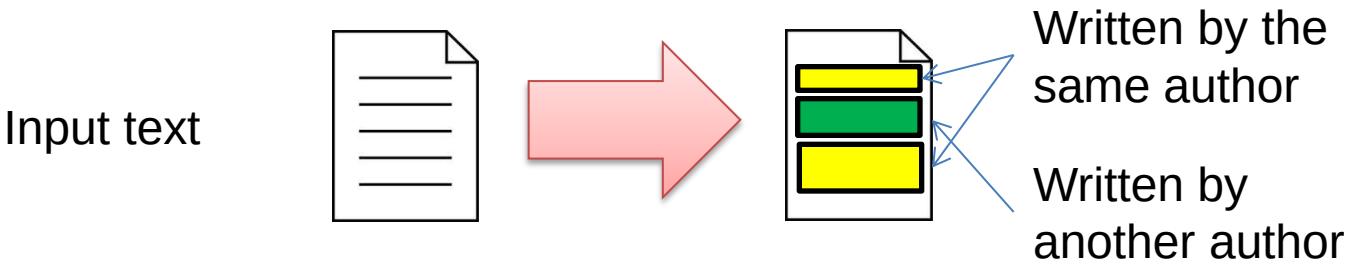
Authorship Analysis Tasks

- Author profiling or characterization
  - Extraction of information about the age, gender, educational level, dialect, personality, etc. of the author



Authorship Analysis Tasks

- Author diarization
  - Decompose a multi-author document into authorial components



Applications

- Forensics
  - Intelligence
    - attribution of messages or proclamations to known terrorists
  - Criminal law
    - identifying writers of harassing messages, verifying the authenticity of suicide notes, etc.
  - Civil law
    - copyright disputes
  - Plagiarism detection
- Humanities
  - Literary research
    - attributing anonymous or disputed literary works to known authors
  - Historical research
    - identifying the role of political figures in certain historical periods
- Decision making
  - Marketing based on demographics
  - Personalized product advertisement

Tutorial Layout

- Introduction
- Tasks, applications
- Stylometry
- Attribution paradigms
- Evaluation, resources
- Summary

Stylometry

- The line of research dealing with the quantification of writing style
- Style is more difficult than topic
- We need measures:
  - Stable throughout text-length
  - Stable in topic shifts
  - Stable in genre variations
  - Able to capture information unconsciously used by the authors

Stylometric Features

- More than 1,000 different features
  [Rudman, 1998]
- Lexical features
- Character features
- Syntactic features
- Semantic features
- Application-specific features

Lexical Features

- A text is a sequence of tokens (perhaps grouped into sentences)
  - each token corresponds to a word, number, or a punctuation mark
- Require tokenizers (and sentence splitters)
  - In some languages it is not a trivial procedure
- May also require stemmers, detection of homographic forms etc.

Lexical Features

- Sentence length counts, word length counts [Mendenhall, 1887]
- Vocabulary richness functions are attempts to quantify the diversity of the vocabulary of a text [Yule, 1944]
  - type-token ratio V/N, hapax legomena
  - Unstable over text-length
- Word frequencies
  - The most frequent words are the most useful
  - In topic-based TC these words are removed

Lexical Features: Word frequencies

- Function words
  - How are they defined?
  - [Abbasi & Chen 2005]: 150 words
  - [Argamon, et al., 2003]: 303 words
  - [Zhao & Zobel, 2005]: 365 words
  - [Koppel & Schler, 2003]: 480 words
- The most frequent words [Burrows, 1992]
  - How many? (50, 100, 250, 1000, …)
  - The larger the frequent word set, the more likely to include content-specific words

Lexical Features: Word *n*-grams

- Take advantage of contextual information
- The dimensionality of the representation increases exponentially with *n*
  - Sparse data
- It is quite likely to capture content-specific rather than stylistic information

Lexical Features: Error-based

- Spelling errors are characteristic of the author's style [Koppel & Schler, 2003]
- Letter omissions and insertions
- Formatting errors (all caps words)
- An accurate spell checker is needed

Character Features

- A text is viewed as a mere sequence of characters
- Language independent measures:
  - alphabetic characters count
  - digit characters count
  - uppercase and lowercase characters count
  - letter frequencies
  - punctuation marks count

Character *n*-grams

- Simplistic but quite effective approach
- Able to capture
  - lexical information (e.g., |_in_|, |text|),
  - hints of contextual information (e.g., |in_t|),
  - use of punctuation and capitalization
- Tolerant to noise
  - *simplistic* vs. *simpilstc*
- Suitable to oriental languages

Character *n*-grams

- How to define the order (*n*)?
- Small *n* (2 or 3)
  - Not able to adequately represent the contextual information.
- Large *n* (>3)
  - better captures lexical and contextual information
  - increases substantially the dimensionality
- The selection of the best *n* value is a language-dependent procedure
- Variable-length *n*-grams [Houvardas & Stamatatos, 2006]

Character Features

- Compression-based approaches use the model acquired from one text to compress another text
  - Based on off-the-shelf text compression tools

  [Benedetto, et al., 2003] [Khmelev & Teahan, 2003]
- No concrete representation
- Essentially they are based on repetitions of character sequences

Syntactic Features

- Authors tend to use similar syntactic patterns unconsciously
- Function words are related with syntactic patterns
- We need robust and accurate NLP tools to perform syntactic analysis
  - A language-dependent procedure
  - Noisy measures

NLP Tools Providing Syntactic Features

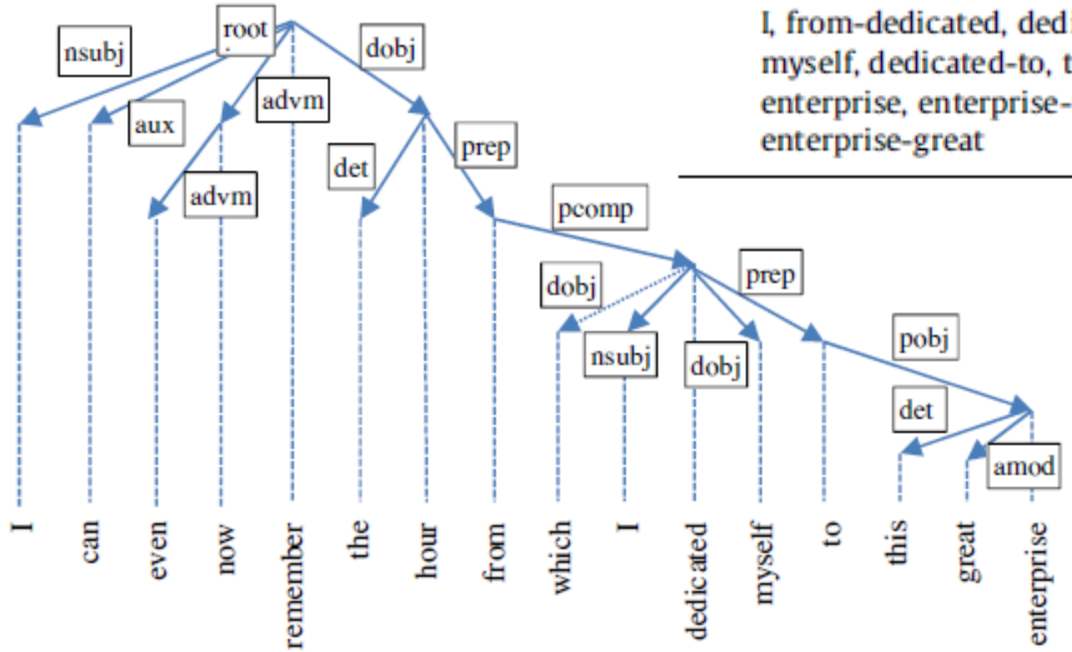
- POS tagging [Kukushkina, et al. 2001]
- Morpho-syntactic tagging [van Halteren, 2007]
- Text chunking [Stamatatos, et al., 2000]
- Partial parsing [Luyckx & Daelemans, 2005]
- Full-parsing [Gamon, 2004] [Sidorov, et al., 2014]
- Spell checking [Koppel & Schler, 2003]

Syntactic Features: Examples

- Rewrite rule frequencies: [Baayen, et al., 1996]
  - A:PP → P:PREP + PC:NP
- Text chunks: [Stamatatos, et al., 2000]
  - PP[*On the other hand*], NP[*this method*] VP[*requires*] NP[*accurate NLP tools*].
- Partial parsing bigrams: [Hirst & Feiguina, 2007]:
  - NX DT JJ NN
- Unigrams, bigrams, and trigrams of morpho-syntactic tags, *n*-grams of rewrite rules [van Halteren, 2007]
  - 900K features!

Syntactic Features: Examples

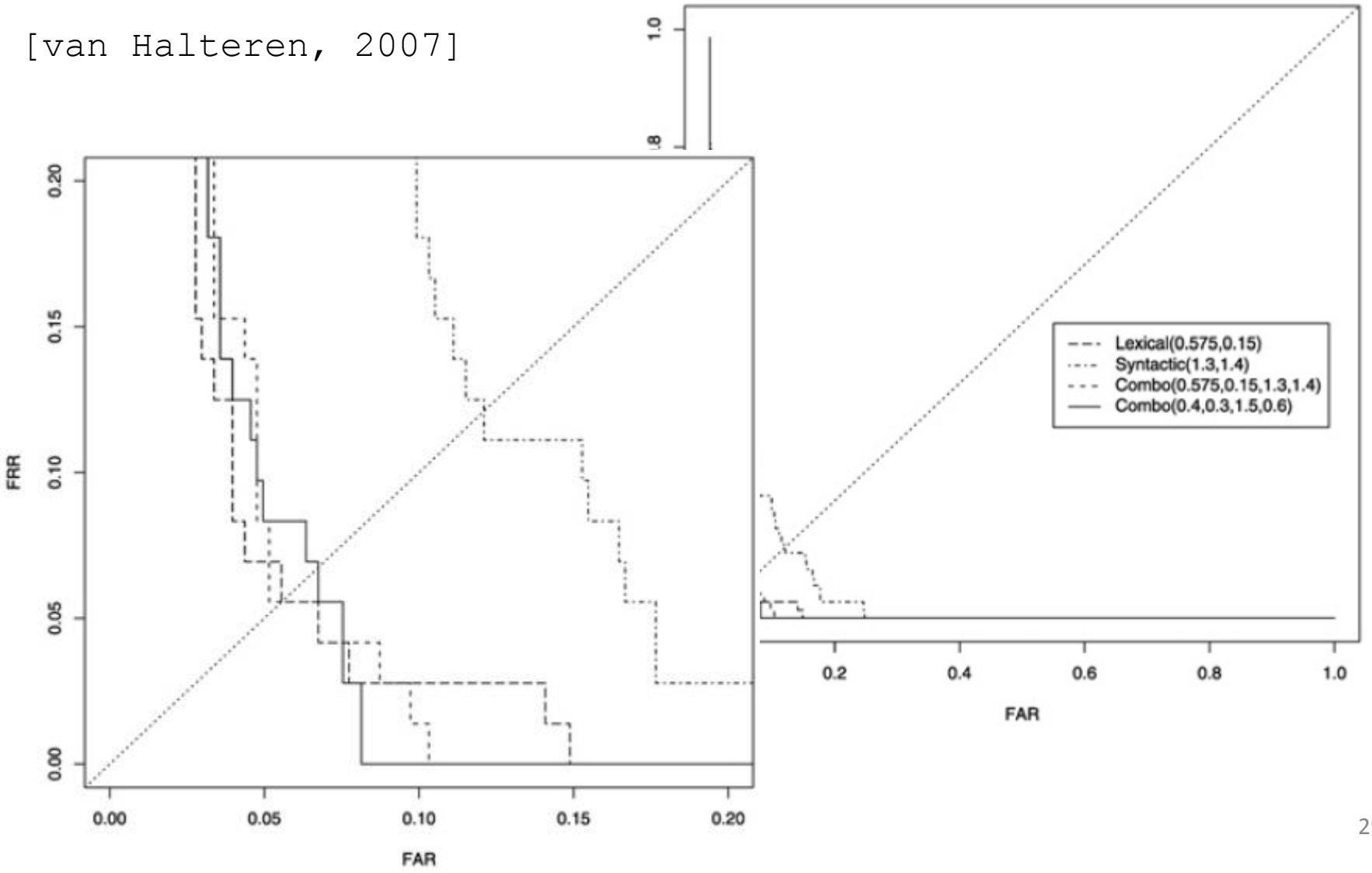
- Syntactic n-grams: `[Sidorov, et al., 2014]`



Syntactic bigrams	Traditional bigrams
Remember-now, now-even, remember-hour, remember-I, remember-can, hour-the, hour-from, dedicated-which, dedicated-I, from-dedicated, dedicated-myself, dedicated-to, to-enterprise, enterprise-this, enterprise-great	I-can, can-even, even-now, now-remember, remember-the, the-hour, hour-from, from-which, which-I, I-dedicated, dedicated-myself, myself-to, to-this, this-great, great-enterprise

Syntactic vs. Lexical Features

[van Halteren, 2007]



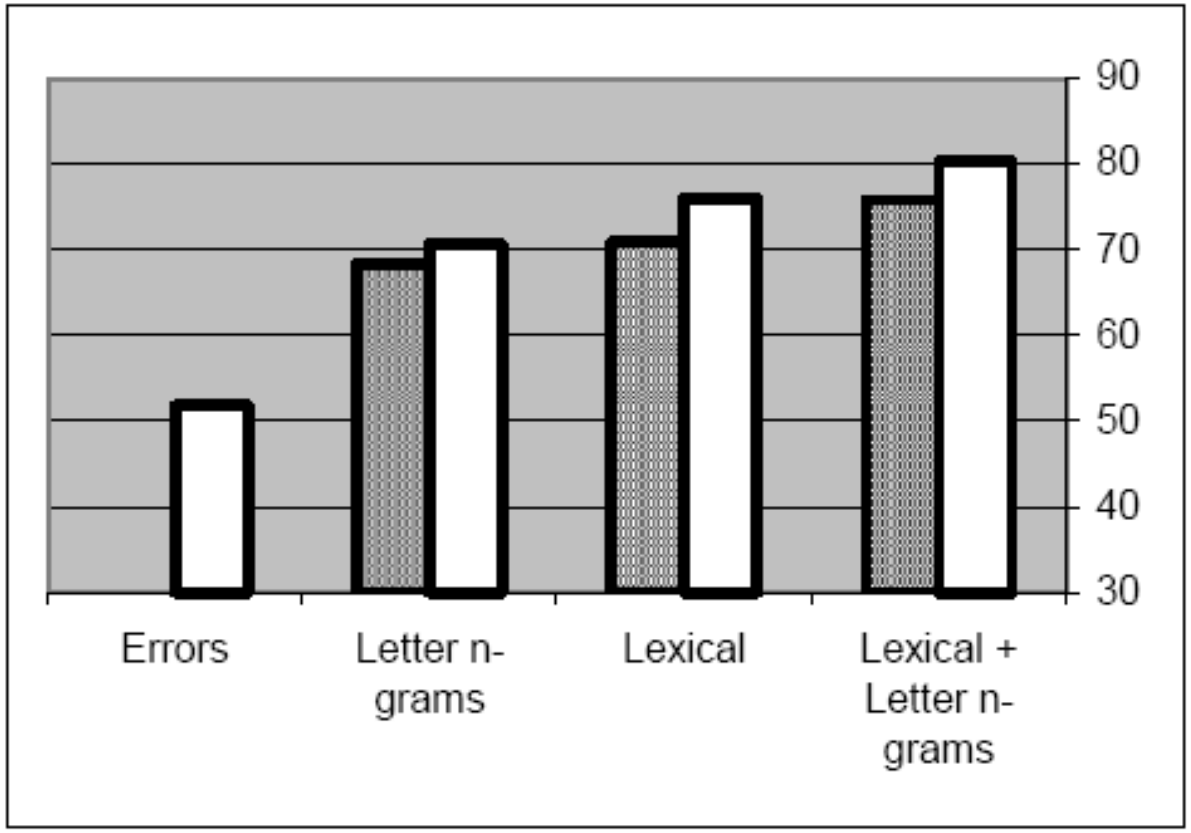
Syntactic Features: Error-based

- Syntactic errors are useful style indicators
  - sentence fragments, run-on sentences, mismatched tense, etc.
- This type of information is similar to that used by human experts when they attempt to analyze style.
- A powerful spell checker should be available.
  - Noisy measures requiring manual modification

    [Koppel and Schler, 2003]

Error-based Features

- Performance on native language identification
  [Koppel, et al., 2005]

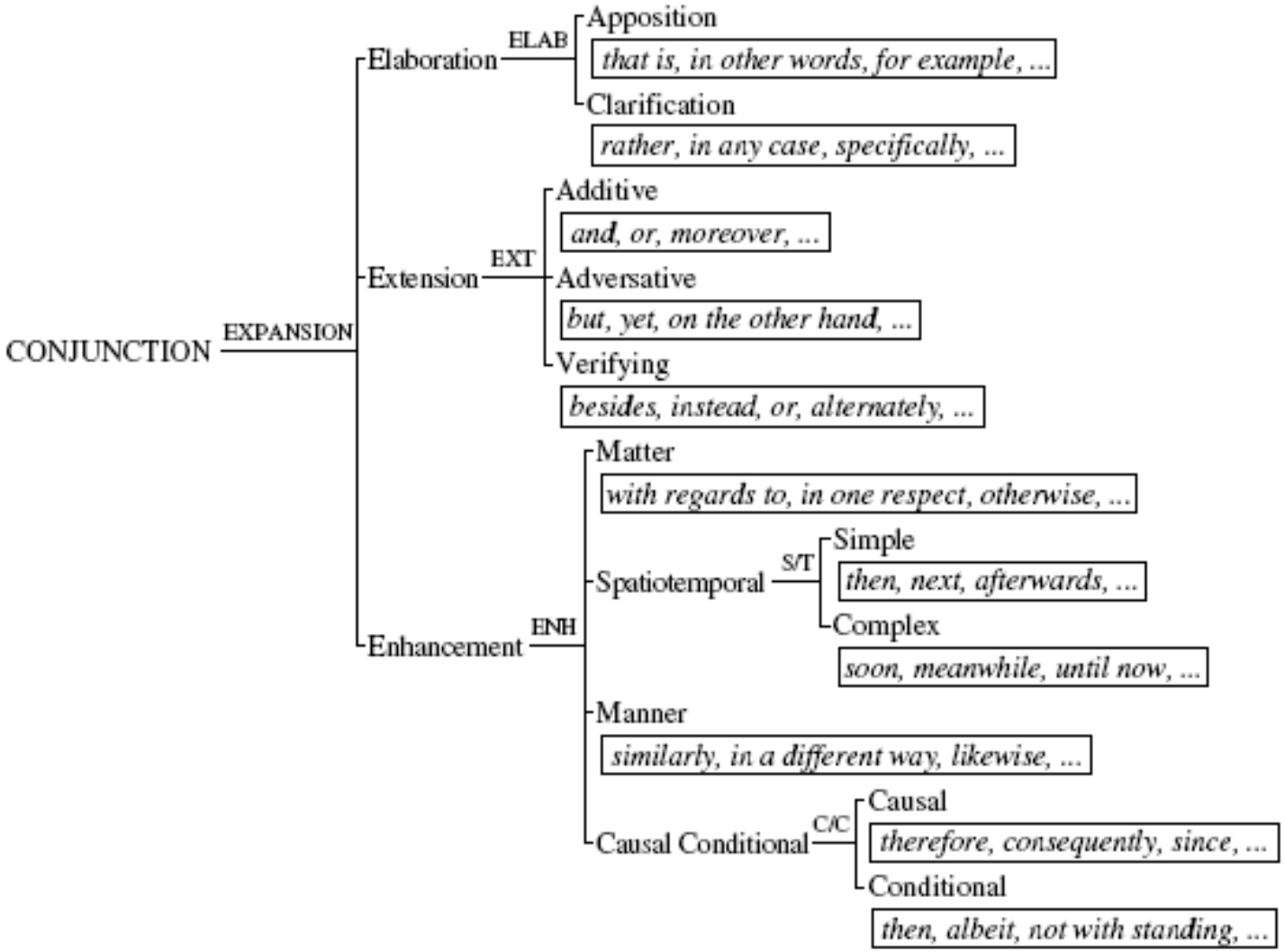


Semantic Features

- More complicated NLP tools are needed
  - More noise in measures
- They can be an important complement to other more powerful features
- Examples:
  - Semantic dependencies [Gamon, 2004]
  - Synonyms using Wordnet [McCarthy, et al., 2006]
  - Systemic Functional Grammar [Argamon, et al., 2007]
  - Semantic frames [Hedegaard, et al., 2011]

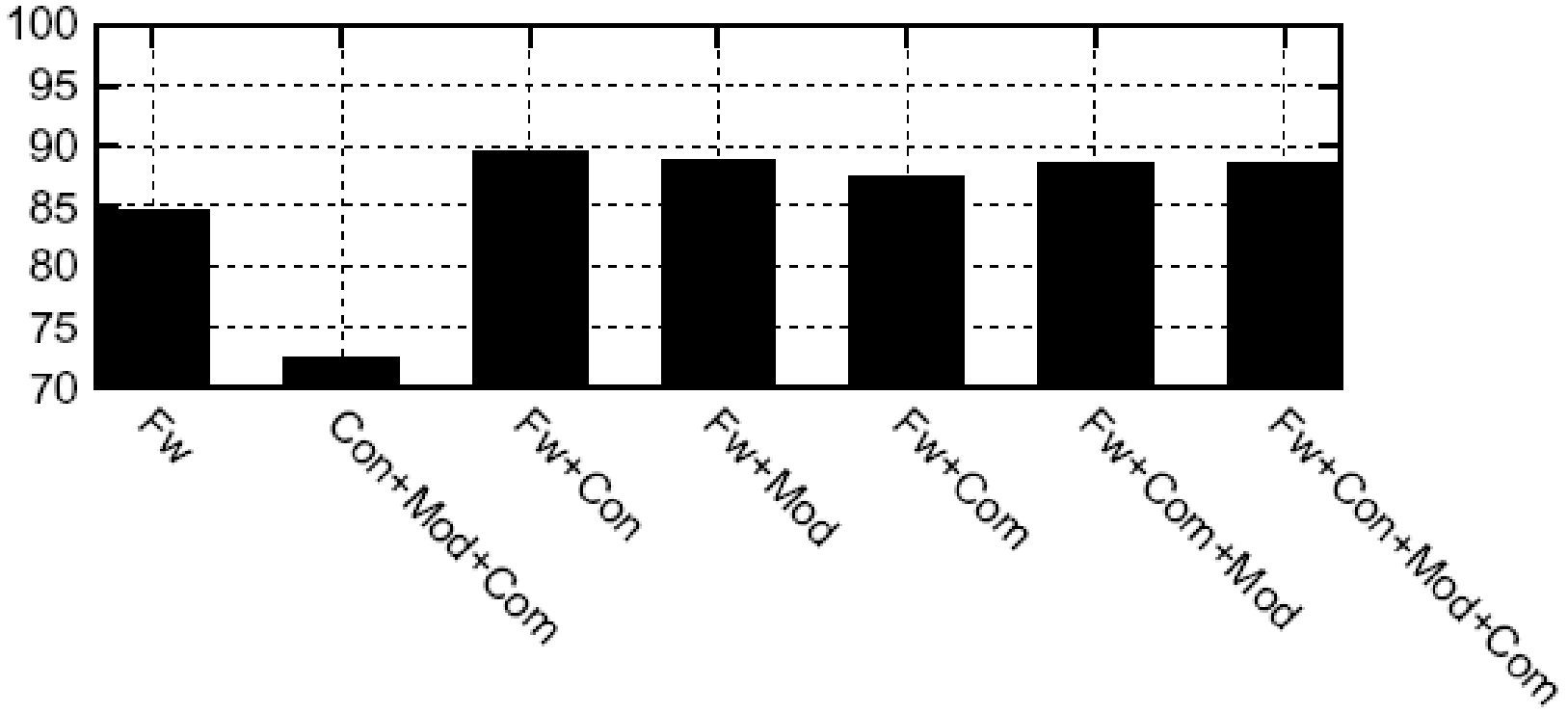
Functional Lexical Features

[Argamon, et al., 2007]



Functional Lexical Features
Performance on Authorship Attribution

[Argamon, et al., 2007]



Application-specific Features

- Can only be defined in specific domains
- Document type
  - Emails (greetings, farewells)
- Document format
  - HTML documents (font color, font size)
- Document topic
  - *misc.forsale.computers* (deal, sale, obo)
- Document language
  - Modern Greek (diglossia)

Heterogeneous Feature Sets

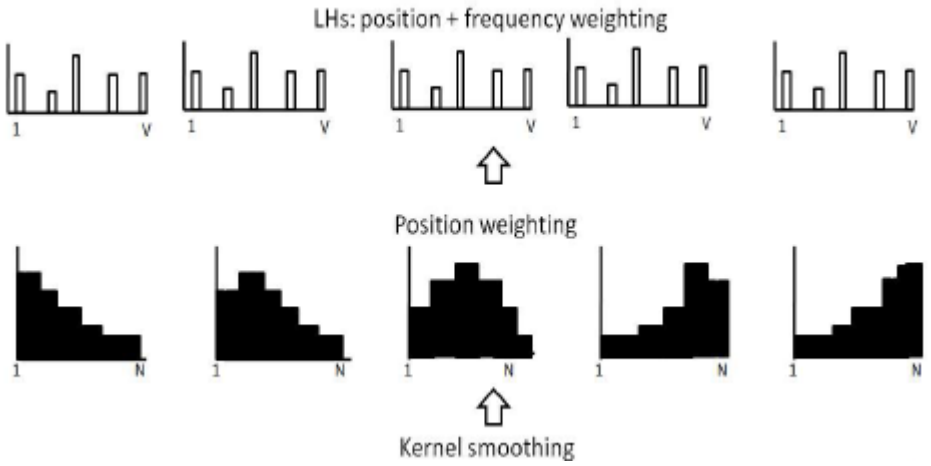
- Several feature types are usually combined

[de Vel, et al. 2001]

Style Marker Attribute Type
Number of blank lines/total number of lines
Average sentence length
Average word length (number of characters)
Vocabulary richness i.e., V/M
Total number of function words/M
Function word frequency distribution (122 features)
Total number of short words/M
Count of hapax legomena/M
Count of hapax legomena/V
Total number of characters in words/C
Total number of alphabetic characters in words/C
Total number of upper-case characters in words/C
Total number of digit characters in words/C
Total number of white-space characters/C
Total number of space characters/C
Total number of space characters/number white-space characters
Total number of tab spaces/C
Total number of tab spaces/number white-space characters
Total number of punctuations/C
Word length frequency distribution/M (30 features)

Alternative Representations

- Instead of extracting global histograms
  - Extract local histograms [Escalante, et al., 2011]

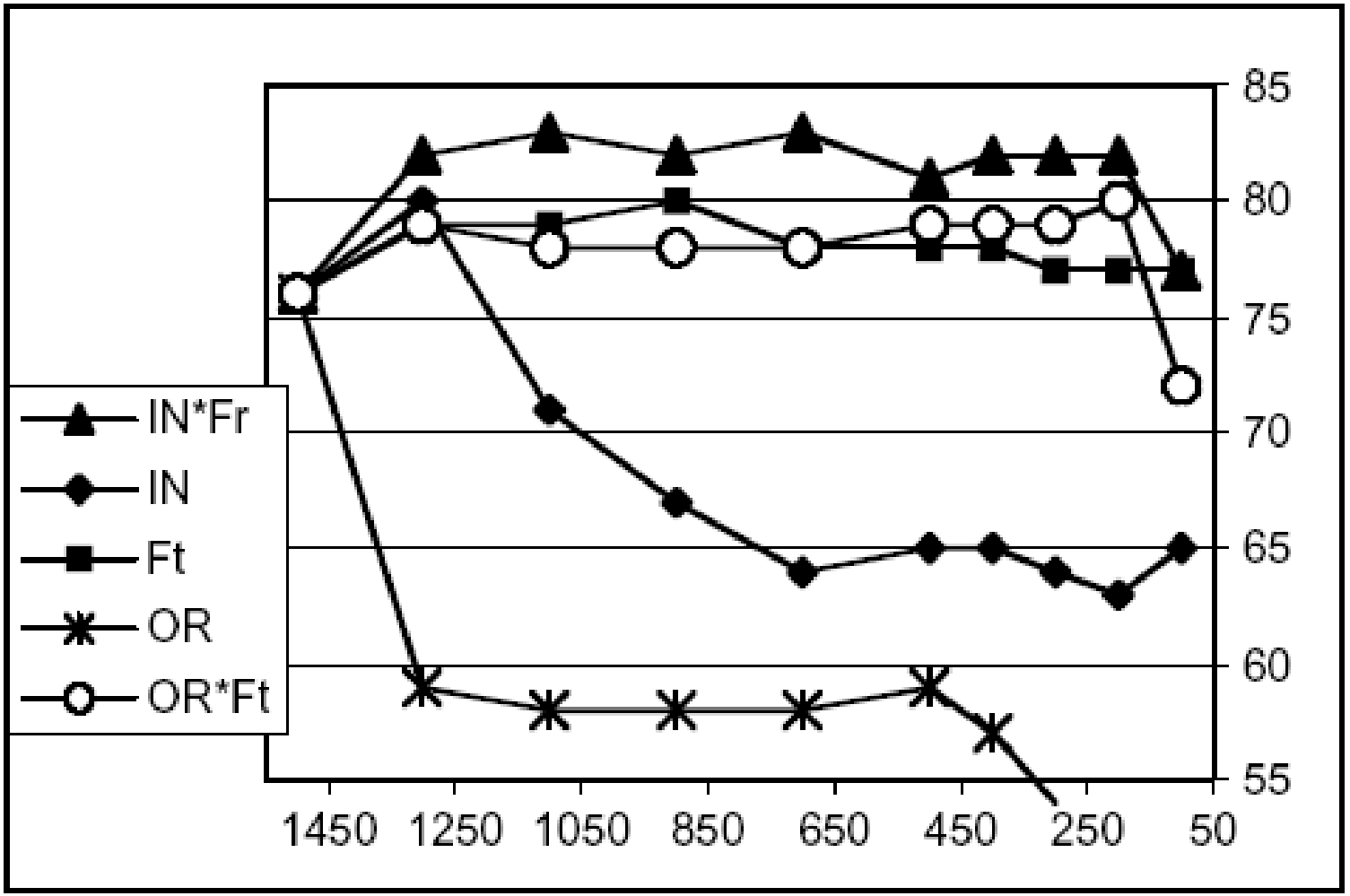


  - Use topic-models
    [Seroussi, et al., 2012] [Savoy, 2013]
  - Construct a graph [Arun, et al., 2009]

Feature Selection

- Feature (subset) selection can be used to reduce dimensionality
- Applying feature selection based on distinctiveness of features may be misleading
  - Due to content-specific choices
  - Corpus-dependent features
- In authorship analysis tasks frequency is more important than distinctiveness
  - Frequency vs. InfoGain [Houvardas & Stamatatos, 2006]
  - Frequency vs. OddsRatio [Koppel et al, 2006]
- The most frequent features can be extracted from a general-purpose corpus
  - Corpus-independent features

Feature Selection Performance



[Koppel, et al., 2006]

Feature Requirements 1/2

[Stamatatos, 2009]

Features		Required tools and resources
Lexical	Token-based (word length, sentence length, etc.)	Tokenizer, [Sentence splitter]
	Vocabulary richness	Tokenizer
	Word frequencies	Tokenizer, [Stemmer, Lemmatizer]
	Word *n*-grams	Tokenizer
	Errors	Tokenizer, Orthographic spell checker
Character	Character types (letters, digits, etc.)	Character dictionary
	Character *n*-grams (fixed-length)	-
	Character *n*-grams (variable-length)	Feature selector
	Compression methods	Text compression tool

Feature Requirements 2/2

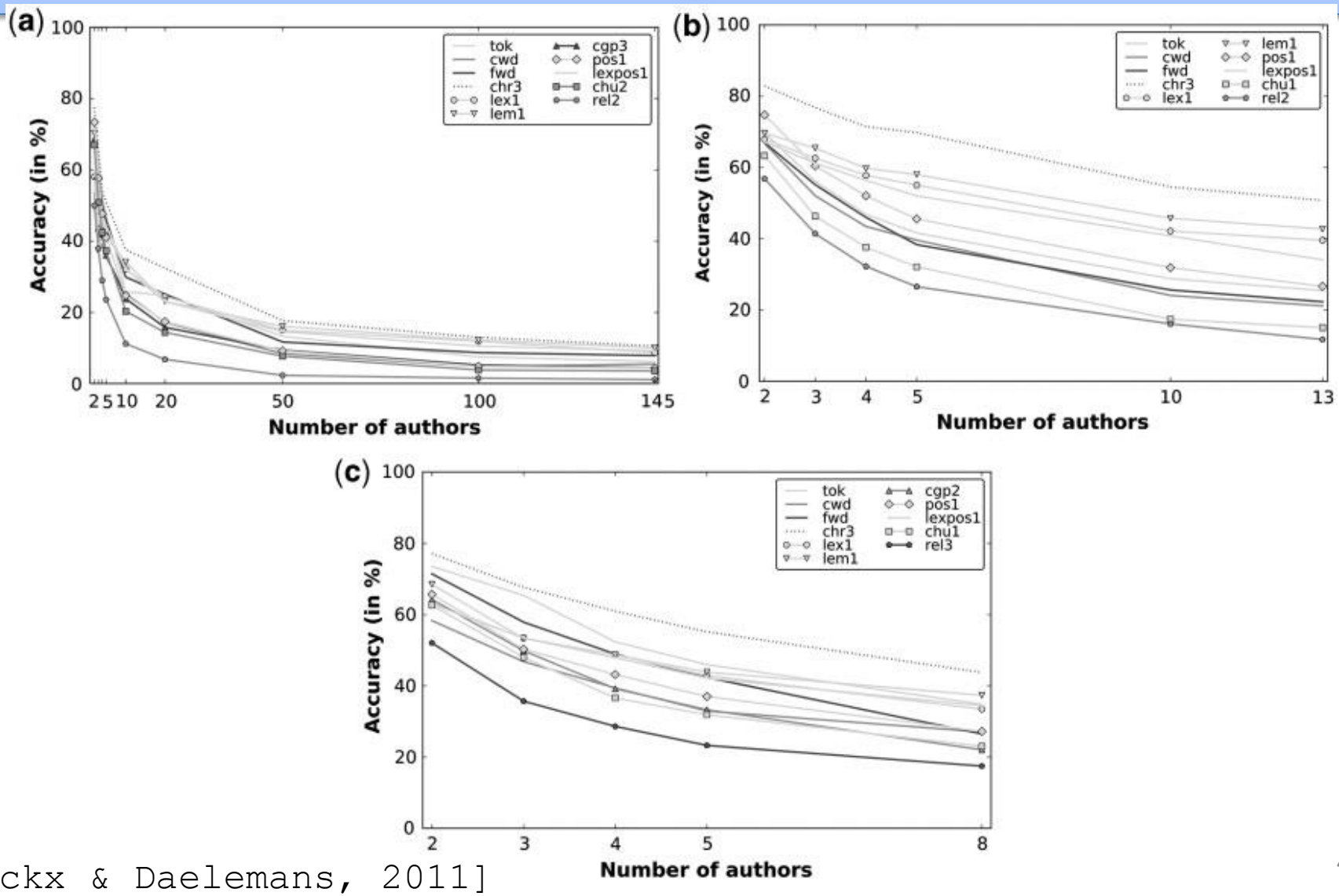
[Stamatatos, 2009]

Features		Required tools and resources
Syntactic	Part-of-Speech	Tokenizer, Sentence splitter, POS tagger
	Chunks	Tokenizer, Sentence splitter, [POS tagger], Text chunker
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser
	Errors	Tokenizer, Sentence splitter, Syntactic spell checker
Semantic	Synonyms	Tokenizer, [POS tagger], Thesaurus
	Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser
	Functional	Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries
Application-specific	Structural	HTML parser, Specialized parsers
	Content-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries
	Language-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries

What Stylometric Features are the Most Effective?

- In several studies character n-grams provide the best results
- Function words (or frequent words) are also very effective
- Higher-level (syntactic or semantic) features are too noisy
  - They are useful as complement
- When possible to apply, structural or application-specific features are valuable

Author Identification



[Luyckx & Daelemans, 2011]

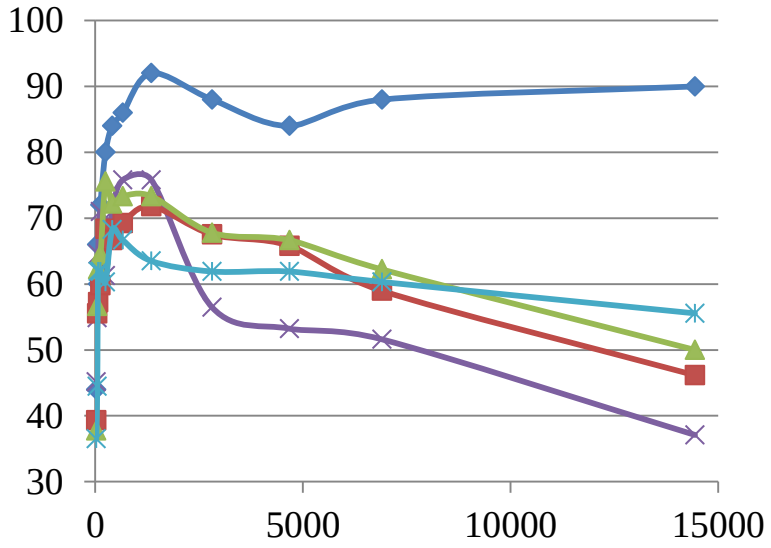
Author Identification

[Grieve, 2007]

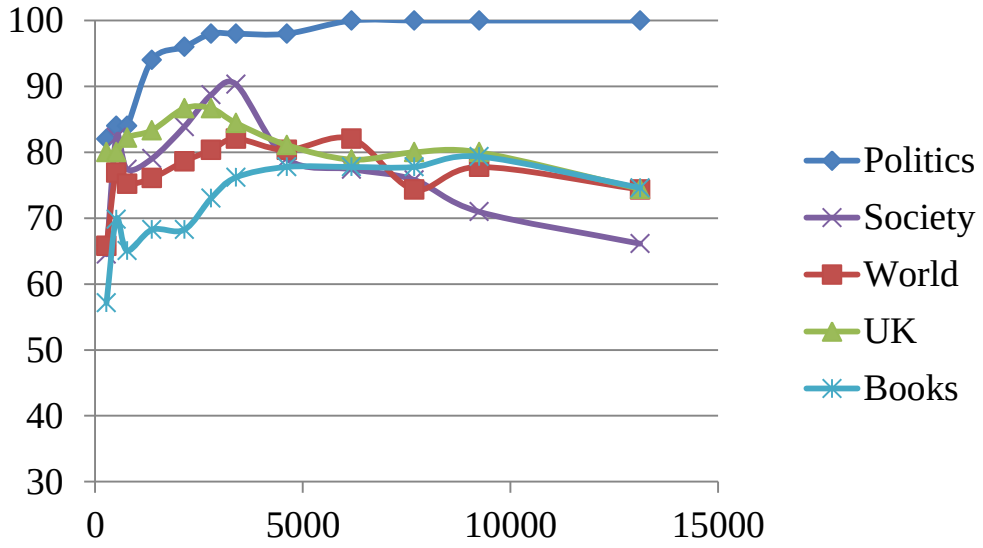
Textual measurement (Variant)	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94
3-gram profile (10-limit)	61	72	78	85	88	91	94
4-gram profile (10-limit)	55	64	73	83	85	89	93
Grapheme and punctuation mark profile	50	60	70	81	84	87	93
Multiposition graph profile (first and last six in word)	49	58	68	79	82	86	92
Word profile (5-limit)	48	57	67	77	80	85	88
5-gram profile (10-limit)	47	55	66	76	79	84	90
Multiposition grapheme profile (first six in word)	43	53	64	76	79	84	90
Multiposition grapheme profile (last six in word)	42	52	63	74	79	83	90
Punctuation mark profile (by character)	34	46	58	72	76	80	89
6-gram profile (10-limit)	35	45	56	68	72	78	86
Word-internal grapheme profile	28	39	51	65	70	76	85
Single-position grapheme profile (last in word)	27	36	49	63	68	73	84
Grapheme profile	25	35	47	62	67	74	83
7-gram profile (2-limit)	34	42	45	59	64	69	81

Cross-topic and Cross-genre Authorship Attribution

Words



Char 3-grams



Corpus of 13 authors
Training based on texts about Politics

[Stamatatos, 2013]

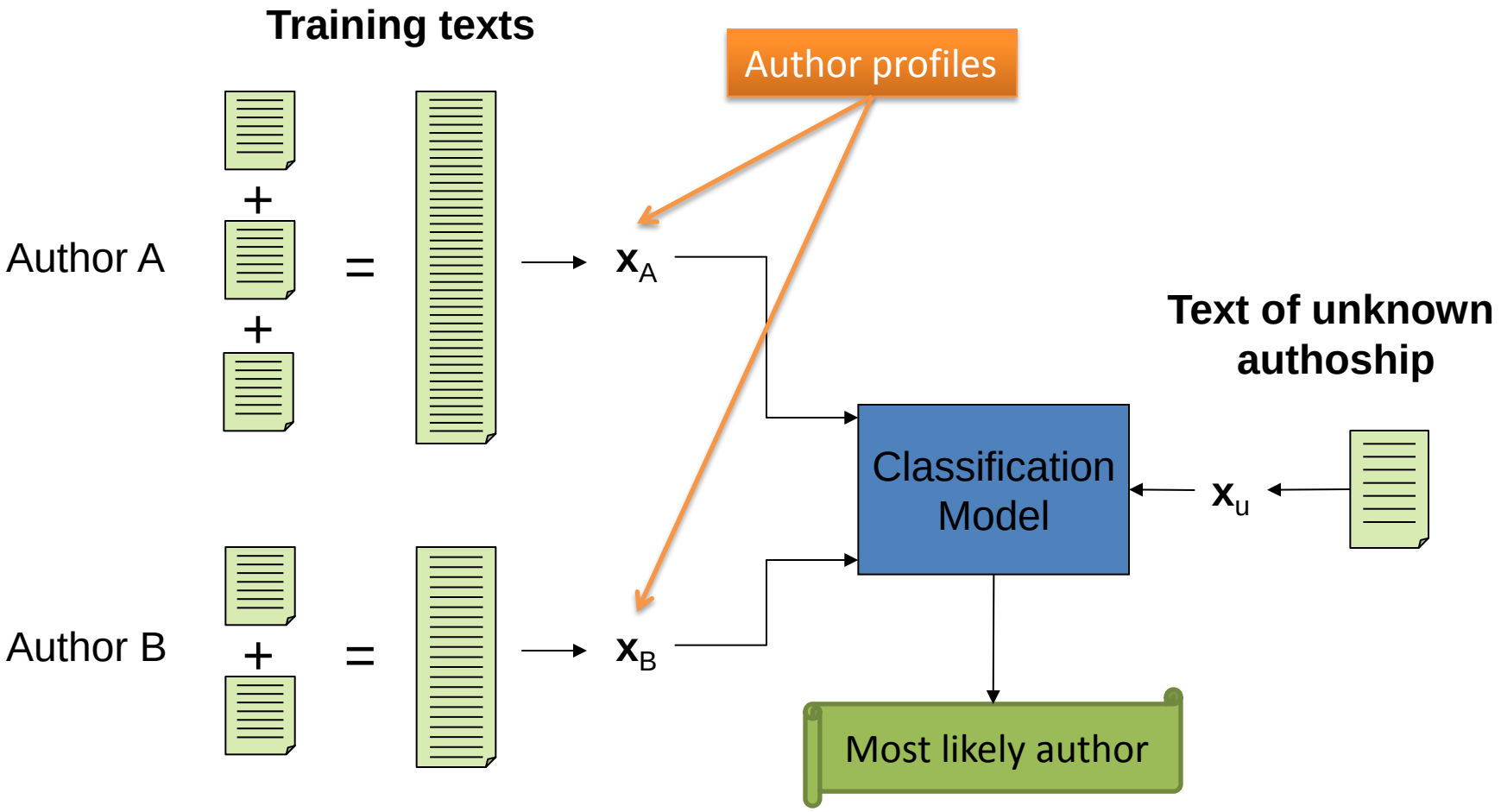
Tutorial Layout

- Introduction
- Tasks, applications
- Stylometry
- Attribution paradigms
- Evaluation, resources
- Summary

Attribution Paradigms

- Profile-based
  - All the available texts per class are concatenated and then a profile is extracted
  - Author-centric: style of an author
  - Classification of generative nature
- Instance-based
  - Each text of known authorship provides a separate training instance
  - Document-centric: style of a document
  - Classification of discriminatory nature

Profile-based Paradigm



Profile-based Paradigm

- The differences between the training texts by the same author are disregarded
- The stylometric measures of the concatenated file may be quite different than each of the original training texts
- Very simple training process
- Distance-based attribution:

$$author(x) = \arg\min_{a \in \mathbf{A}} d(PR(x), PR(x_a))$$

Profile-based Paradigm: Probabilistic Approach

- Probabilistic: [Peng, et al., 2004]

$$author(x) = \arg\max_{a \in \mathbf{A}} \log_2 \frac{P(x \mid a)}{P(x \mid \bar{a})}$$

- Naïve Bayes can be augmented with statistical language models
  - Allows local Markov chain dependencies in the observed variables to capture contextual information
  - Can be applied to both character and words

Profile-based Paradigm: Compression-based Approach

[Khmelev & Teahan, 2003] [Marton, et al., 2005]

- $PR(x)=x$
- $d(x, x_a)=C(x_a+x)-C(x_a)$
- C(.) provided by RAR, LZW, GZIP, BZIP2, 7ZIP, …

- *Prediction by partial matching* (used by RAR) works practically the same as the method of [Peng, et al., 2004]
  - But the models describing x_a are adaptive (not static) with respect to x and slower
  - Can be applied only to characters

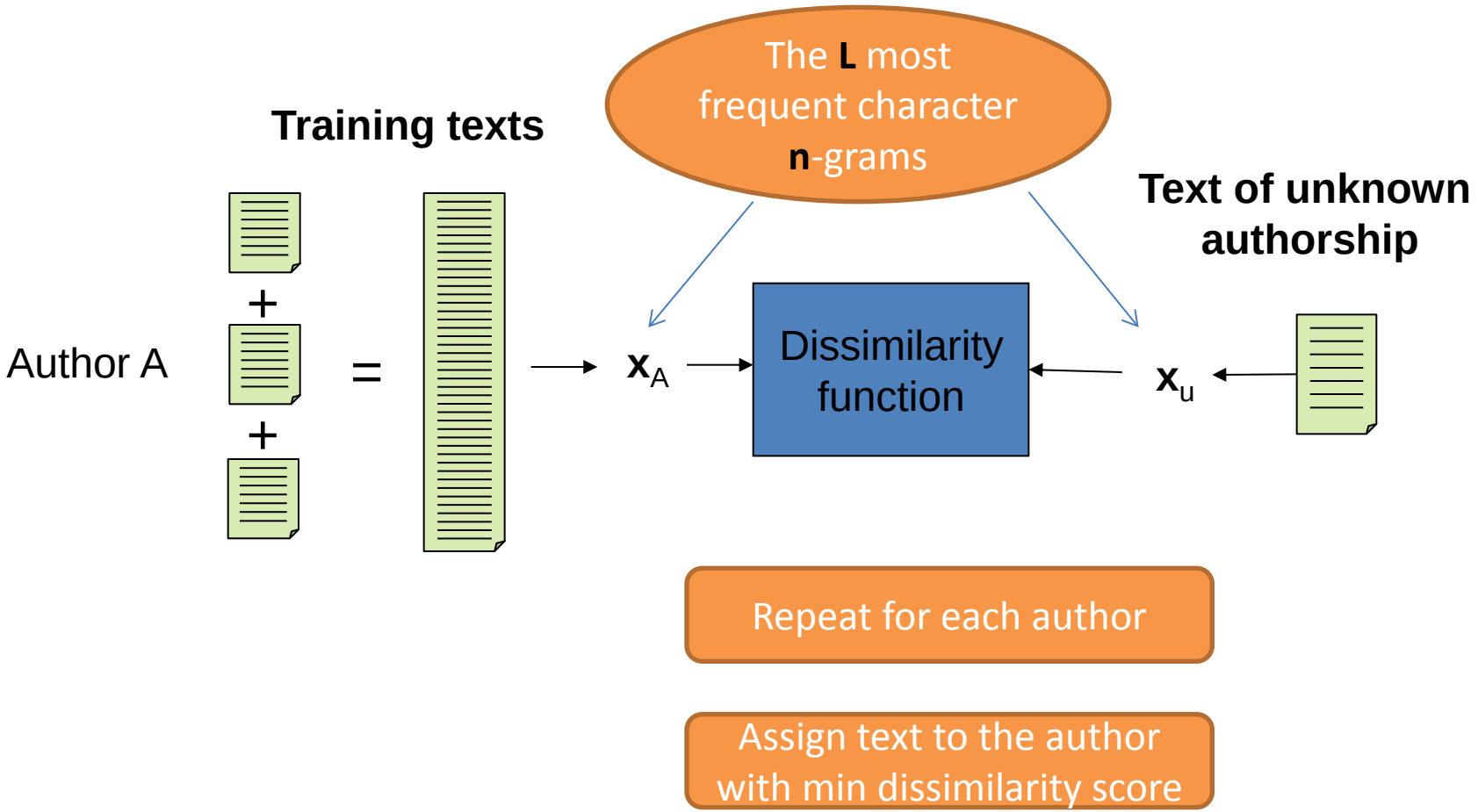
Compression-based TC

- Authorship attribution performance

  [Khmelev and Teahan, 2005]

Method	$R < 0.25$	$R < 0.5$	$R < 0.75$	$R < 1.00$	$R \leq 1.0$
R-measure	82.1	86.4	87.1	87.8	89.0
Multi-SVM	80.6	83.4	83.5	84.6	85.0
Bzip2	56.9	55.2	45.9	51.9	48.2
Gzip	55.7	53.5	53.9	50.1	59.4
Markov Chains, order 1	62.3	64.6	63.2	64.3	66.1
Markov Chains, order 2	60.9	64.4	61.8	64.7	64.5
Markov Chains, order 3	48.6	60.3	59.3	61.7	63.3
RAR	**84.3**	**86.9**	**87.3**	88.5	**89.4**
PPMD, order 2	77.8	79.1	79.4	80.5	81.3
PPMD, order 3	80.6	82.3	84.0	85.0	86.4
PPMD, order 4	82.5	85.4	86.0	87.7	88.4
PPMD, order 5	82.2	86.1	86.3	**88.8**	89.2

CNG [Keselj, et al., 2003]



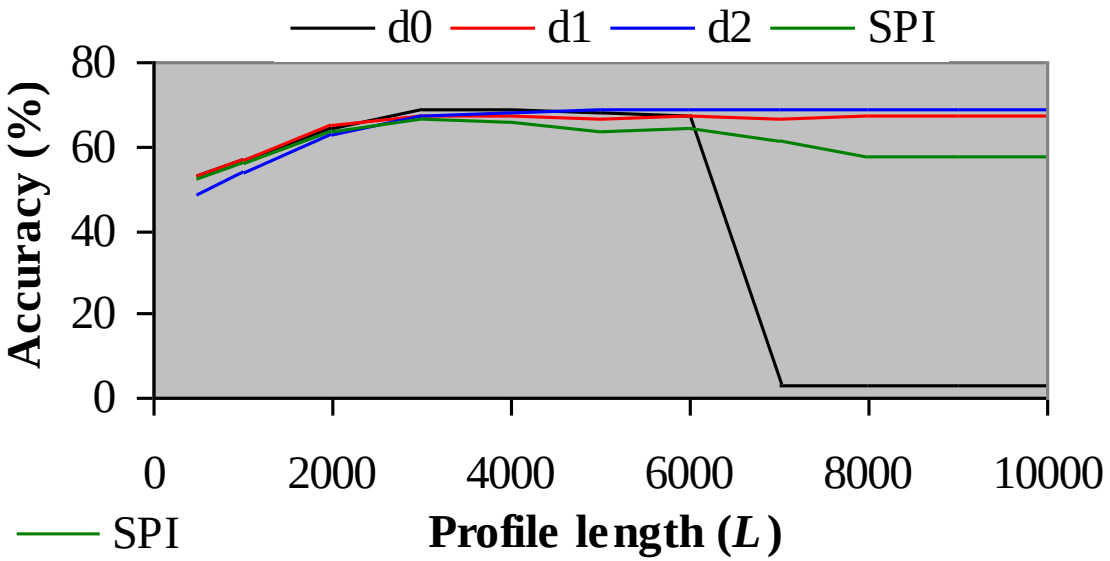
Variants of CNG

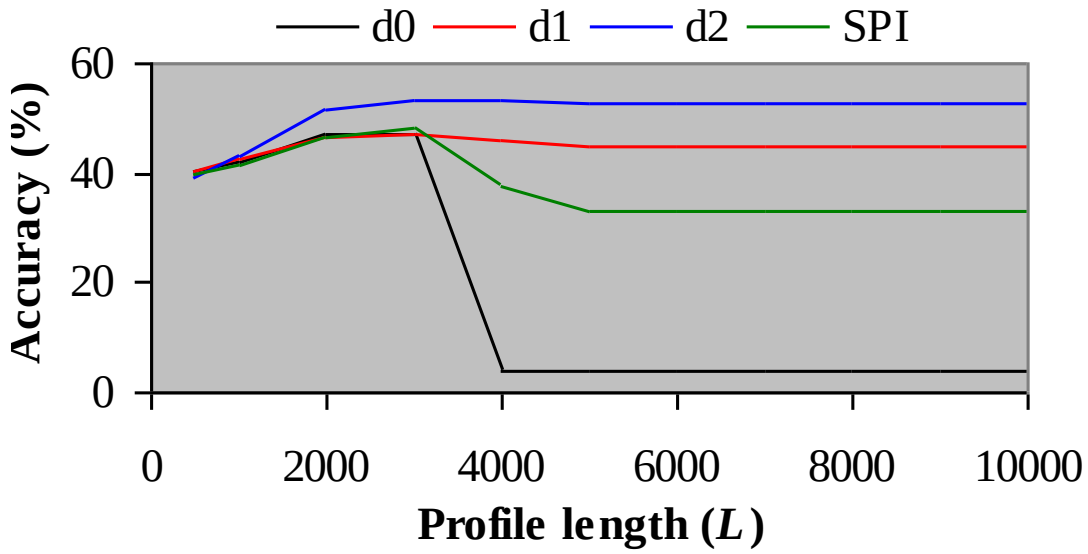
- Original measure:
$$d(PR(x), PR(y)) = \sum_{g \in P(x) \cup P(y)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2$$
  - Unstable when classes are imbalanced
- SPI [Frantzeskou, et al., 2006]
$$SPI(SP(x), SP(T_a)) = \left| SP(x) \cap SP(T_a) \right|$$
  - Good results for source code authorship attribution
- Other similarity functions: [Stamatatos, 2007]
$$d_1(PR(x), PR(y)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2$$
$$d_2(PR(x), PR(y), PR(N)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \cdot \left(\frac{2(f_x(g) - f_N(g))}{f_x(g) + f_N(g)} \right)^2$$
  - Stable with class imbalance and limited data

CNG and Variants: Performance

[Stamatatos, 2007]

50 authors,
50 texts per author





50 authors,
10 texts per author

Ensemble Method

[Koppel, et al., 2011]

Given: snippet of length L1; known-texts of length L2 for each of C candidates

1. **Repeat** *k1* times
   a. Randomly choose some fraction *k2* of the full feature set
   b. Find top match using cosine similarity

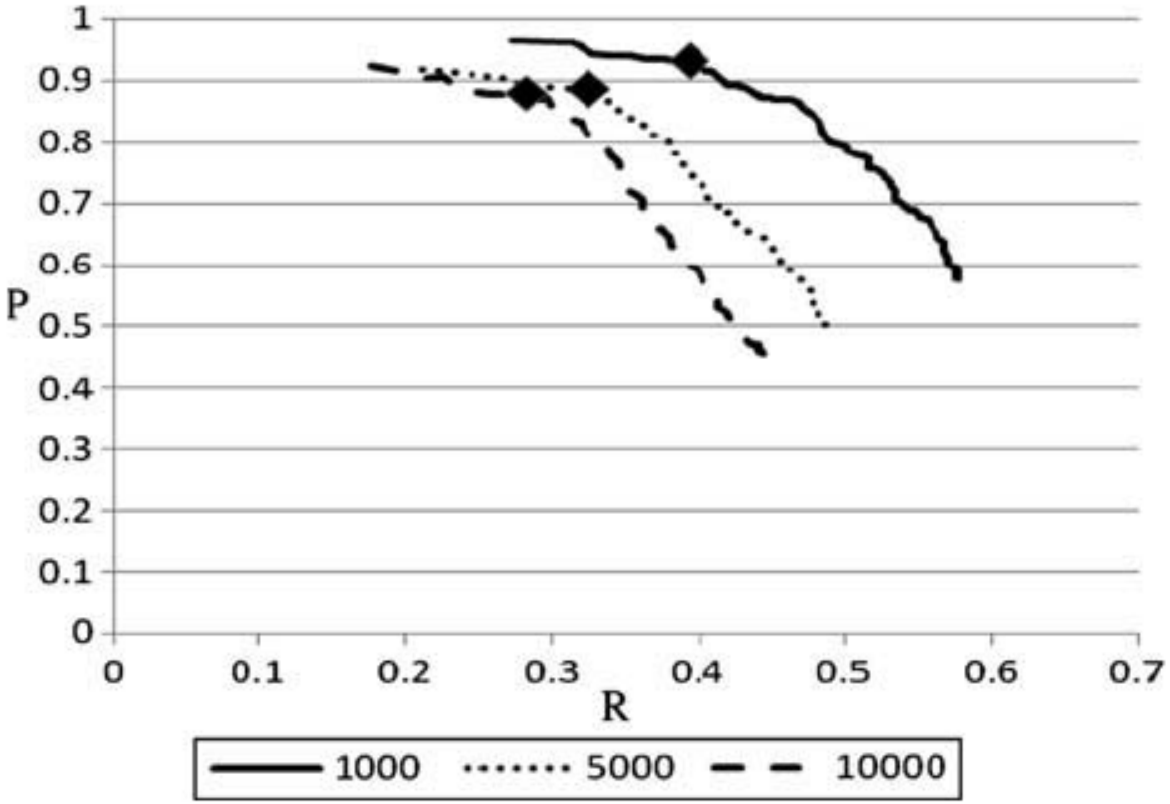
2. **For each** candidate author A,
   a. Score(A) = proportion of times A is top match

Output: arg $\max_A$ Score(A) **if** max Score(A) > σ^*; **else** *Don't Know*

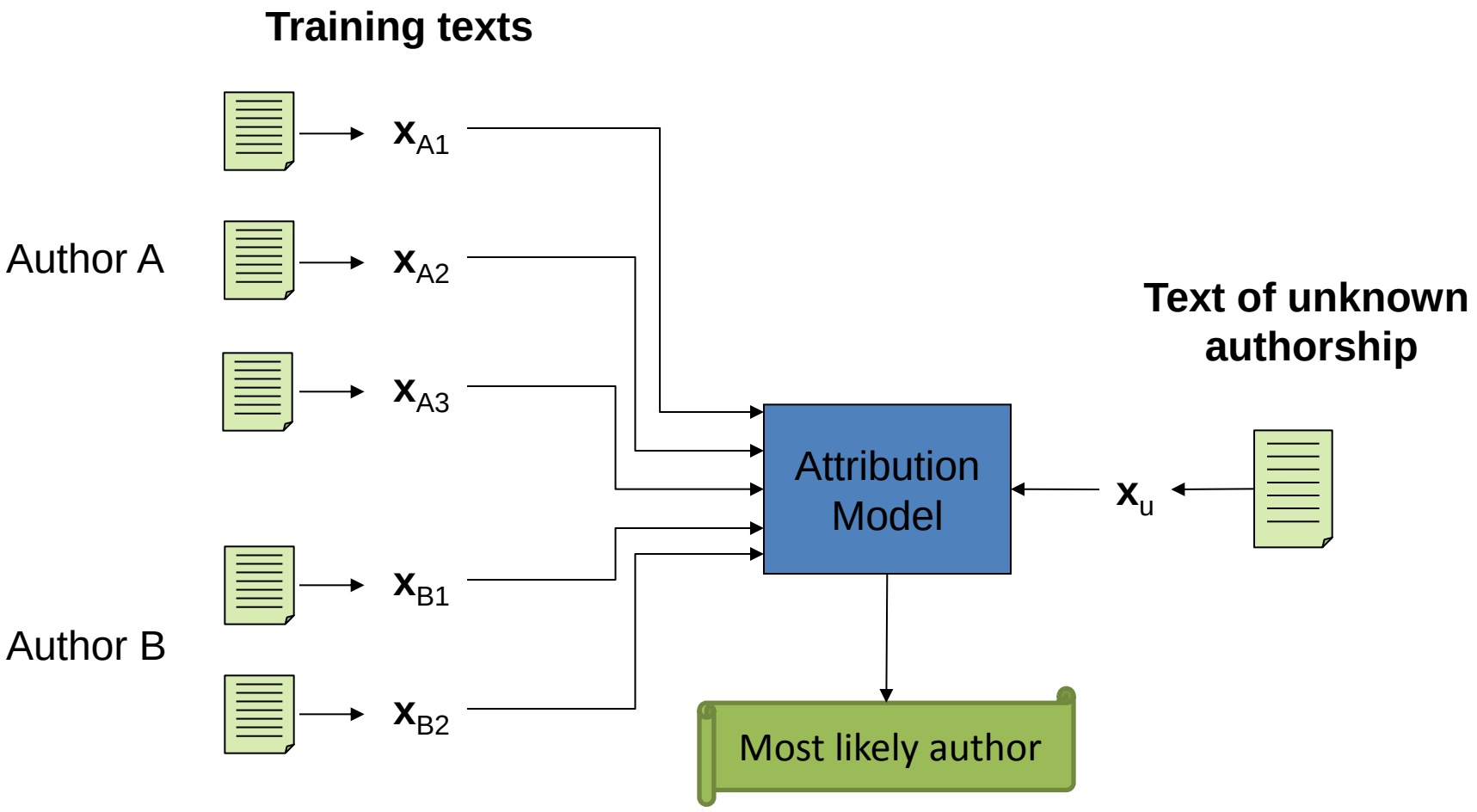
- Open-set approach
- Able to handle large sets of candidate authors
- Effective for short texts

Ensemble Method Performance

- Recall-Precision for various candidates set size



Instance-based Paradigm



Instance-based Paradigm

- Requires multiple training instances per author
- It may require segmentation of training texts
  - Long training texts (books)
  - Training texts of variable-length
  - Segments of equal length?
  - How long? Difficult decision

    `[Sanderson and Guenter, 2006]` : chunks of 500 characters

    `[Koppel, et al., 2007]` : chunks of 500 words

Instance-based Paradigm: Vector Space Models

- Powerful machine learning algorithms can be used:
  - SVM, Neural nets, Discriminant analysis, …

    [de Vel, et al, 2001] [Diederich, et al, 2003] [Sanderson & Guenter, 2006] [Zheng, et al., 2006]
- Can effectively handle high-dimensional, noisy, and sparse data
- Allow more expressive (heterogeneous features) representations of texts
- Affected by the class imbalance problem

  [Stamatatos, 2008]

Instance-based Paradigm: The Delta Method [Burrows, 2002]

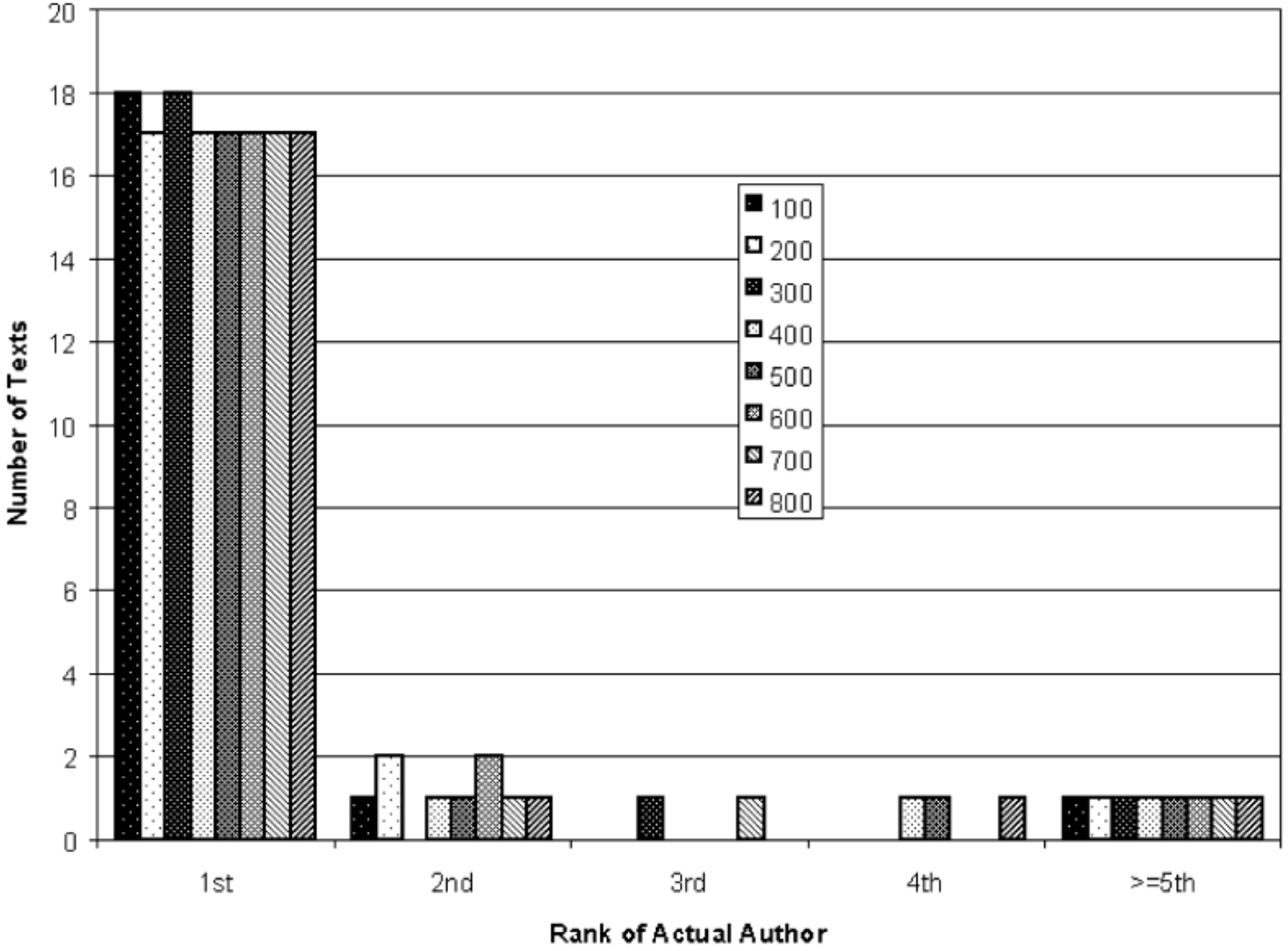
- It is based on pairwise similarity between the unseen text and each training text
- Calculates the deviation (*z*-score) of each word frequency from the norm
  - the 150 most frequent words
  - indicates whether it is used more or less times than the average
- Delta similarity: the mean of the absolute differences between the *z*-scores

The Delta Method

- Very popular in authorship of literary texts [Burrows, 2002]
- Texts of at least 1,500 words are required
- Using larger sets of words (500 words) improves performance [Hoover, 2004]
- Theoretical understanding of this method [Argamon 2008]
  - an axis-weighted form of nearest-neighbor classification

Delta Performance
Authorship Attribution on novels

[Hoover, 2004]



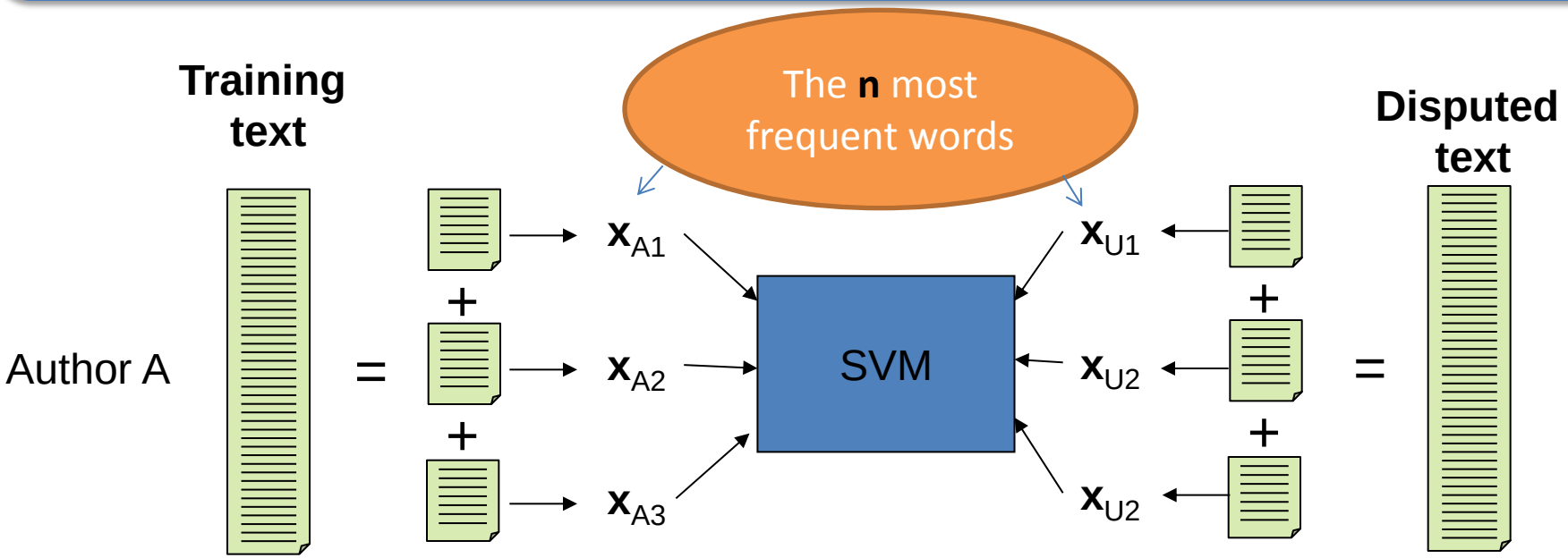
Instance-based Paradigm: Compression-based Methods

- *C(x)* : The compression of each training text using an off-the-shelf algorithm (GZIP)
- *C(x+y)* : The compression of the concatenation of each training text with the unseen text
- Similarity: *d(x,y)=C(x+y)-C(x)* [Benedetto, et al., 2002]
- Heavily critisized method
  - Computationally expensive
  - Sensitive to noise
  - GZIP takes into account only 32K of text
- Alternative: $NCD(x,y) = \frac{C(x+y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$ [Cilibrasi and Vitanyi, 2005]

Instance-based Paradigm: Unmasking [Koppel, et al., 2007]

- A meta-learning model for author verification
- There is no training phase
- One binary SVM classifier is built between the unknown text and the texts of each author
- In an iterative procedure, the most important features of the classifier are removed
- After a few iterations the accuracy of the classifier of the correct author would be too low
- It requires long texts

Unmasking [Koppel, et al., 2007]



Record classifier accuracy

Remove the **k** most important features from **x**

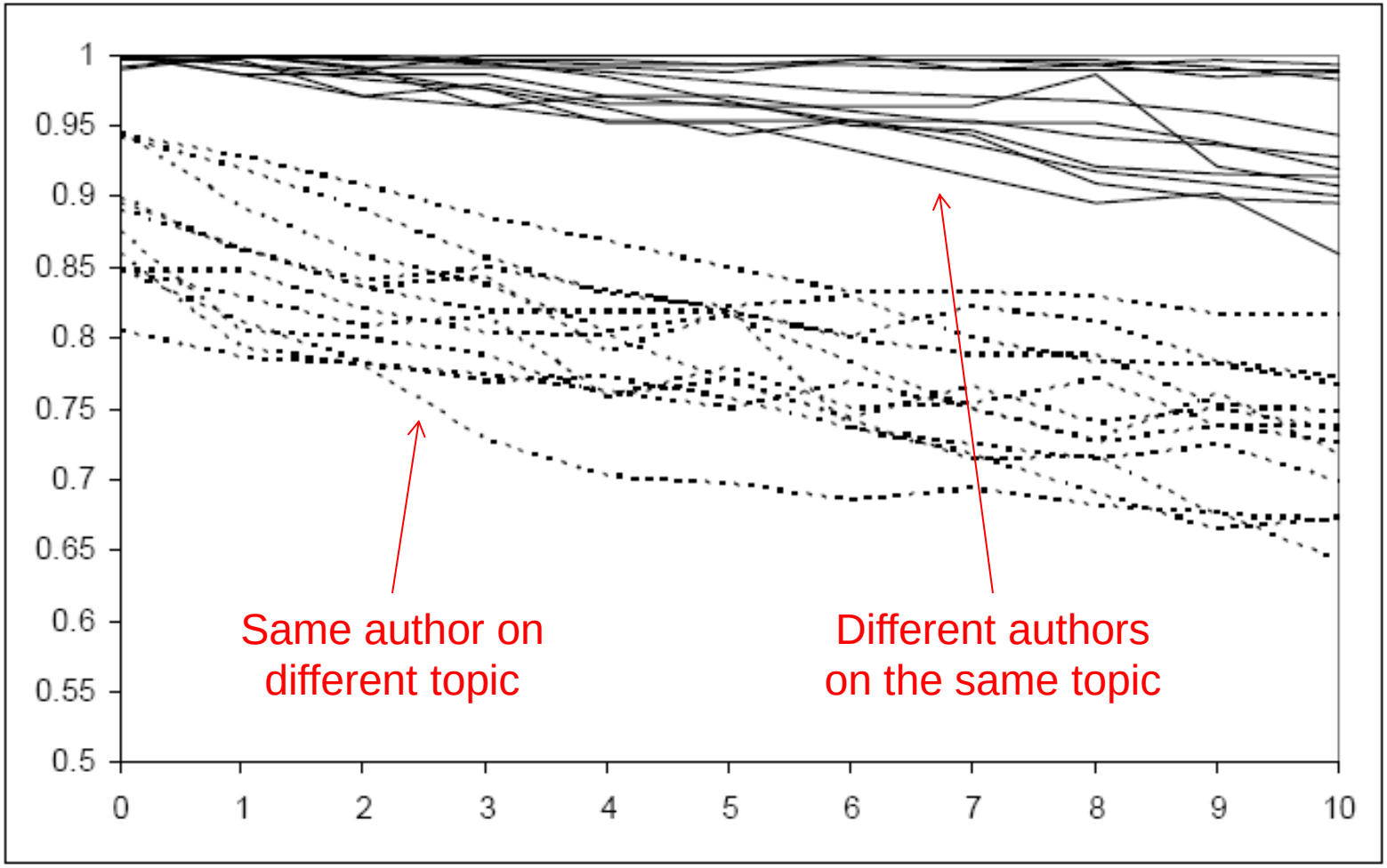
Repeat **r** times

Repeat for each author

Assign to author with largest drop in performance

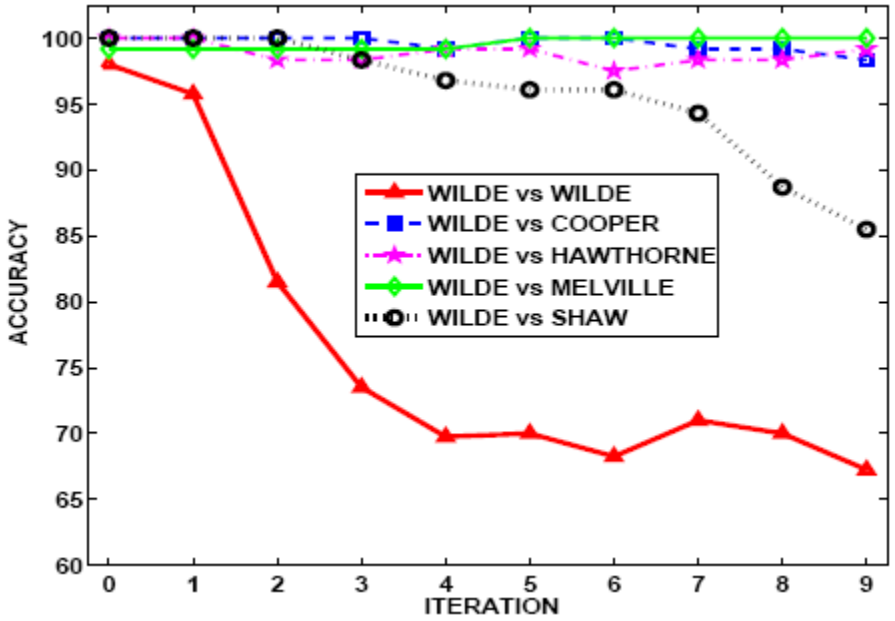
Unmasking: Performance

[Koppel, et al., 2007]

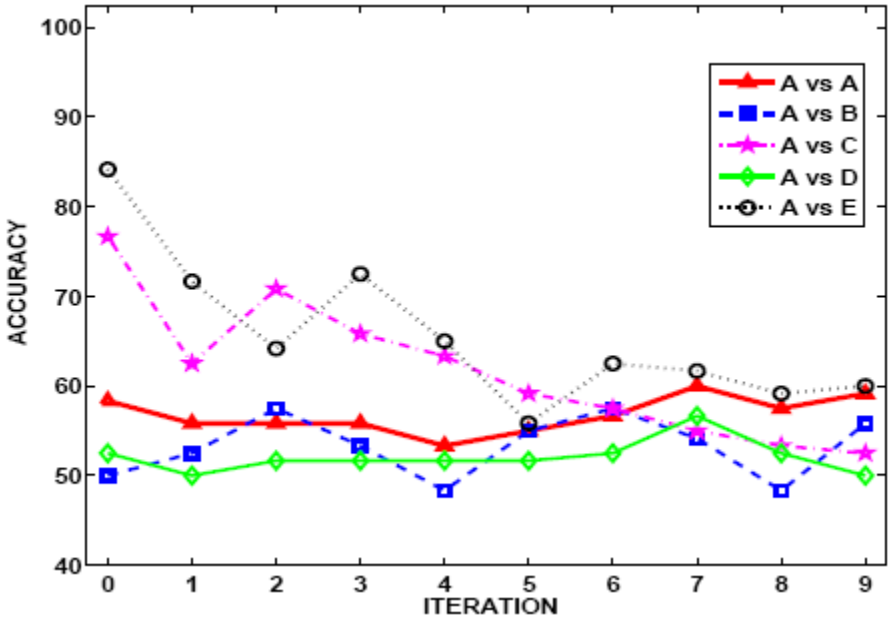


Unmasking: Performance

[Sanderson and Guenther, 2006]



Books



Newspaper articles

Hybrid Approaches

[Van Halteren, 2007] [Grieve, 2007]

- The training examples are represented separately
  - As it happens in the instance-based paradigm
- The representation vectors for each author are averaged feature-wise
  - As it happens in the profile-based paradigm

Comparison of Classification Paradigms

	Profile-based paradigm	Instance-based paradigm
Text Representation	One cumulative representation for all the training texts per class	Each training text is represented individually. Text segmentation may be required.
Stylometric features	Difficult to combine different features. Some (text-level) features are not suitable	Different features can be combined easily
Classification	Generative (e.g., Bayesian) models, Similarity-based methods	Discriminative models, Powerful machine learning algorithms (e.g., SVM), similarity-based methods
Training time cost	Low	Relatively high (low for compression-based methods)
Running time cost	Low (relatively high for compression-based methods)	Low (very high for compression-based methods)
Class imbalance	Depends on the length of training texts	Depends mainly on the amount of training texts

Tutorial Layout

- Introduction
- Tasks, applications
- Stylometry
- Attribution paradigms
- Evaluation, resources
- Summary

Evaluation Resources

- The *Federalist Papers* is popular in AA studies:
  `[Mosteller and Wallace, 1964]`
  - A well defined set of candidate authors
  - Sets of known authorship for all the candidate authors
  - A set of texts of disputed authorship
  - All the texts are of the same genre
  - All the texts are in the same thematic area
- But:
  - The set of candidate authors is too small
  - The texts are relatively long
  - The disputed texts may be the result of collaborative writing of the candidate authors

Evaluation Resources

- Many studies focus on literary works:
  - English literature
    [Burrows, 2002] [Hoover, 2004]
    [Argamon, et al., 2007]; [Koppel, et al., 2007]
    - Bronte sisters
      [Burrows, 1992] [Koppel, et al., 2006]
      [Hirst & Feiguina, 2007]
  - Russian literature [Kukushkina, et al., 2001]
  - Italian literature [Benedetto, et al., 2002]
- Long texts
- Small set of candidate authors

Evaluation Resources

- Corpora specifically-built for this task:
  - Online newspaper articles [Stamatatos, et al., 2000]
  - e-mail messages [de Vel, et al., 2001]
  - Online forum messages [Abbasi & Chen, 2005]
  - Newswire stories [Khmelev & Teahan, 2003]
  - Blogs [Koppel, et al., 2006]

- Relatively short texts
- Larger sets of candidate authors
- Modern genres related to certain applications

Evaluation Resources

- General-purpose corpora:
  - Reuters-21578 [Teahan & Harper, 2003]
  - Reuters Corpus Volume 1 [Khmelev & Teahan, 2003]
  - TREC corpus [Zhao & Zobel, 2005]
  - New York Times Annotated Corpus [Schein, et al., 2010]
- Many candidate authors
- Relatively short texts
- Authors are related to specific topics

Controlled Corpora

- To avoid any irrelevant stylistic changes, an ideal evaluation corpus should be controlled in:
  - Topic
  - Genre
  - Age
  - Education level
  - Nationality
  - Period
- Recent trend:
  - Cross-topic and cross-genre attribution

    [Kestemont, et al., 2012], [Stamatatos, 2013]

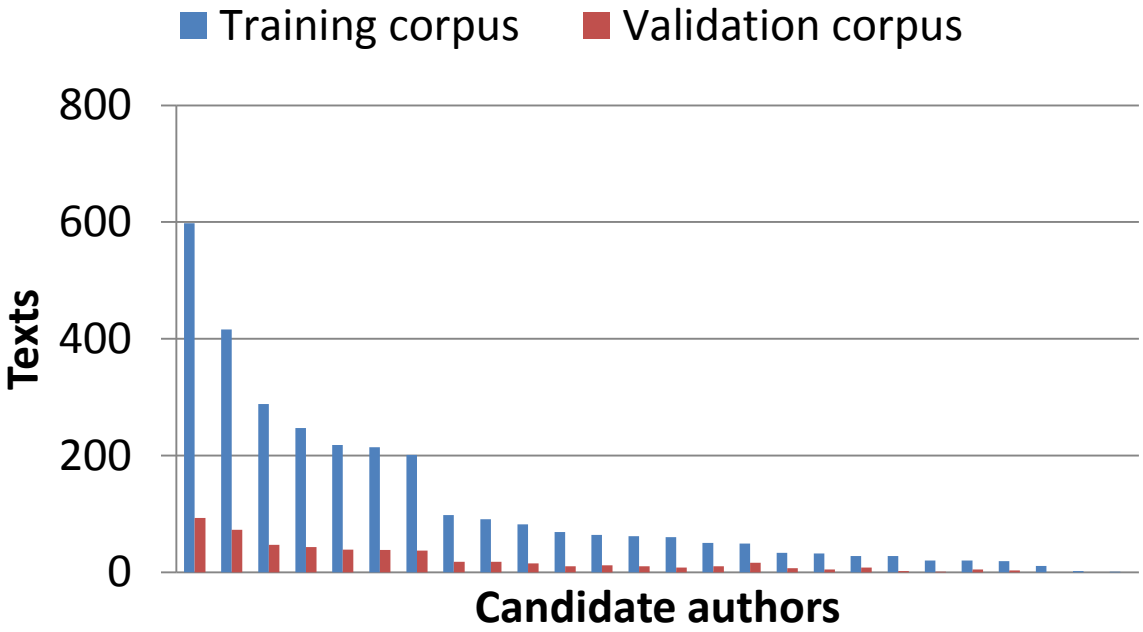
Controlled Corpora

- Forensic corpus [Chaski, 2001]:
  - Texts of 92 people on 10 common subjects (a letter of apology to your best friend, a letter to your insurance company, etc.)
- Movie reviews [Clement and Sharp, 2003]
  - 5 authors who review the same 5 movies
- Student essays [Baayen, et al., 2002]
  - 8 authors, 9 texts per author on specific topics covering three genres
- PAN corpora (pan.webis.de)

Authorship Attribution: Evaluation

- Evaluation is application-dependent:
  - Forensic applications
  - Text filtering
- In forensic applications the test set should always be balanced
  - The availability of texts of known authorship should not increase the likelihood of certain candidate authors
  - An important difference with topic-based TC

PAN11 – Imbalanced Evaluation Set



- 26 candidate authors
- Similar distribution in training and validation sets
  - Not appropriate for forensic applications

Benchmarks

- Ad-hoc authorship attribution competition
  - http://www.mathcs.duq.edu/~juola/authorship_materials2.html
- Blog corpus
  - http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
- PAN corpora:
  - http://pan.webis.de

Tools

- JGAAP (Java Graphical Authorship Attribution Program):
  - http://www.jgaap.com
- JStylo:
  - https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth
- Stylo in R:
  - https://sites.google.com/site/computationalstylistics/stylo

Tutorial Layout

- Introduction
- Tasks, applications
- Stylometry
- Attribution paradigms
- Evaluation, resources
- Summary

Topic-based Text Categorization

- Homogeneous features
- High dimensionality
- Feature selection
  - Distinctiveness
  - e.g., information gain, odds-ratio
- Evaluation
  - Similar distribution of training and test sets
  - e.g., stratified cross-validation
- Training set can be enriched
- Training and test sets follow the same properties and distribution

Authorship Attribution

- Heterogeneous features
- Relatively low dimensionality
- Feature selection
  - Frequency
- Evaluation
  - Test set should be balanced
- Training set can be extremely limited and imbalanced
- Training and test sets may not follow the same properties or distribution

Conclusion

- Authorship attribution should not be handled as yet another text categorization task
- Representing style is more difficult than topic
  - Simple features like char n-grams and function words are the most effective
- Current technology can handle cases with many candidate authors
  - Effectiveness is affected by text-length
  - Effectiveness is decreased when there are differences in topic and/or genre

Future Work Directions

- Investigating the relation between topic, genre, and authorship
  - How can we define features to tell them apart?
- How long should a text be so that we can adequately capture its stylistic properties?
  - Are there other factors (beyond text-length) that also affect this process?
- Transferability
  - Authorship attribution model trained on one genre and transferred to another genre
- Explainable stylometry
  - Useful in forensic applications

Recent Works with Promising Results

- Heterogeneous ensemble models

  [Moreau et al., 2015]

- Neural network language models

  [Bagnall, 2015]

- Distinguishing the most useful character n-grams

  [Sapkota et al., 2015]

More Info

- stamatatos@aegean.gr
- http://www.icsd.aegean.gr/Stamatatos
- http://pan.webis.de