Feature Bagging for Author Attribution

PAN - CLEF 2012

François-Marie Giraud / Thierry Artières
LIP6 – University Paris 6 - France

Motivation

- From the littérature on author attribution
  - Hard to beat a simple and efficient system

Linear SVM on bag of features

- Hypothetical explanations
  - Intrinsic difficulty to define relevant stylistic features
    - Stylistic individual features are embedded and hidden in a large amount of features
    - Stylistic features depend on the writer
  - Optimization concern
    - Undertraining phenomenon [McCallum et al., CIIR 2005]

Motivation

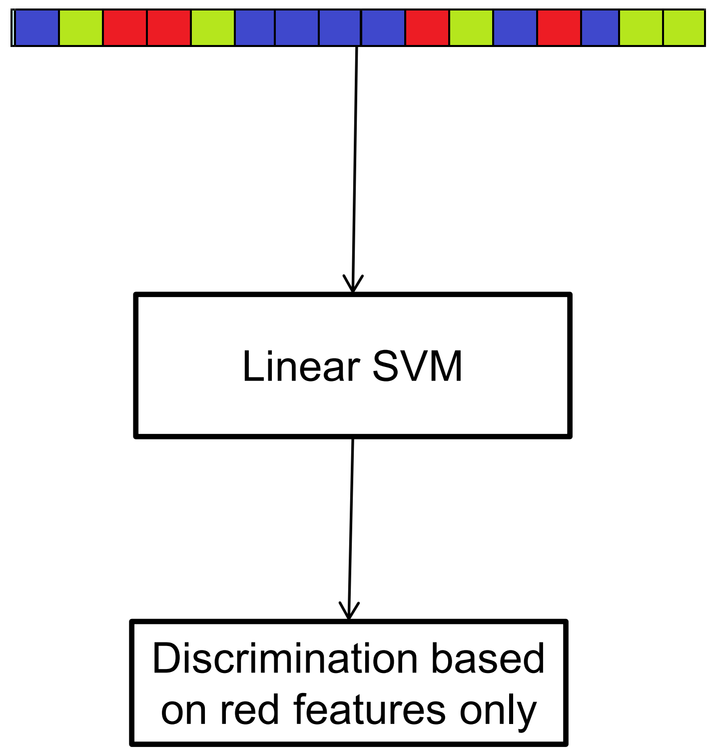
- Undertraining phenomenon

Training Document set: Bag of features
(words sorted most to less frequent)

Motivation

- Undertraining phenomenon

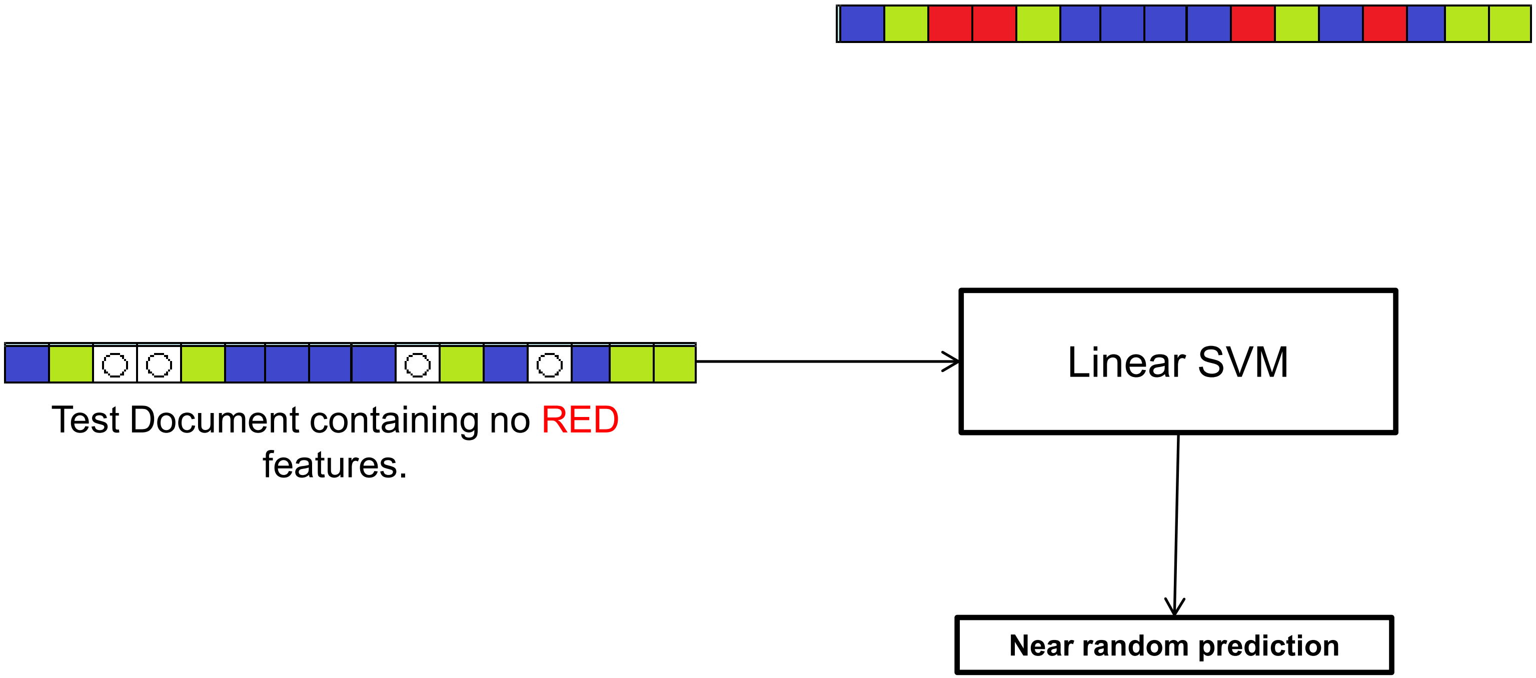
Training Document set: Bag of features
(words sorted most to less frequent)



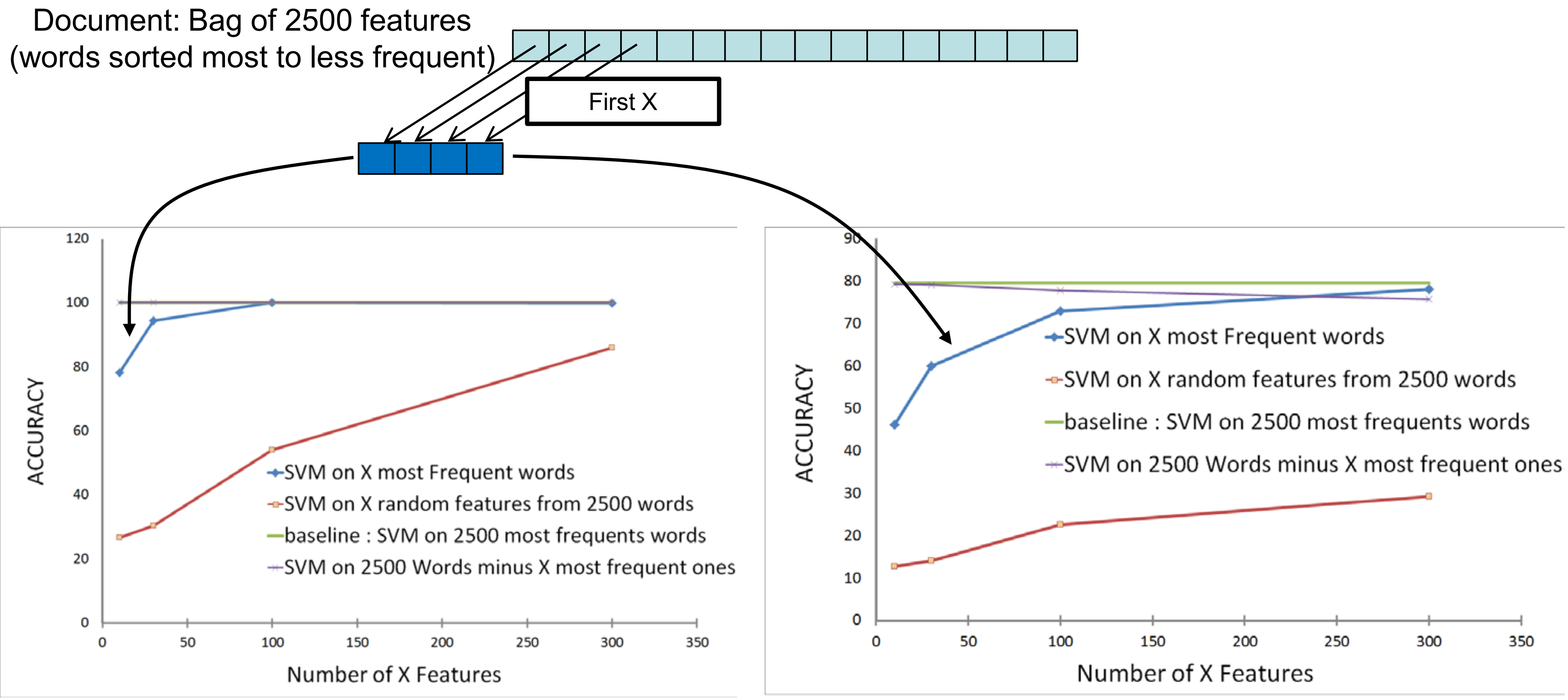
- Red subset of feature alone allows perfect training set discrimination
- Blue subset of feature alone allows either
- Green subset is useless

Motivation

- Undertraining phenomenon



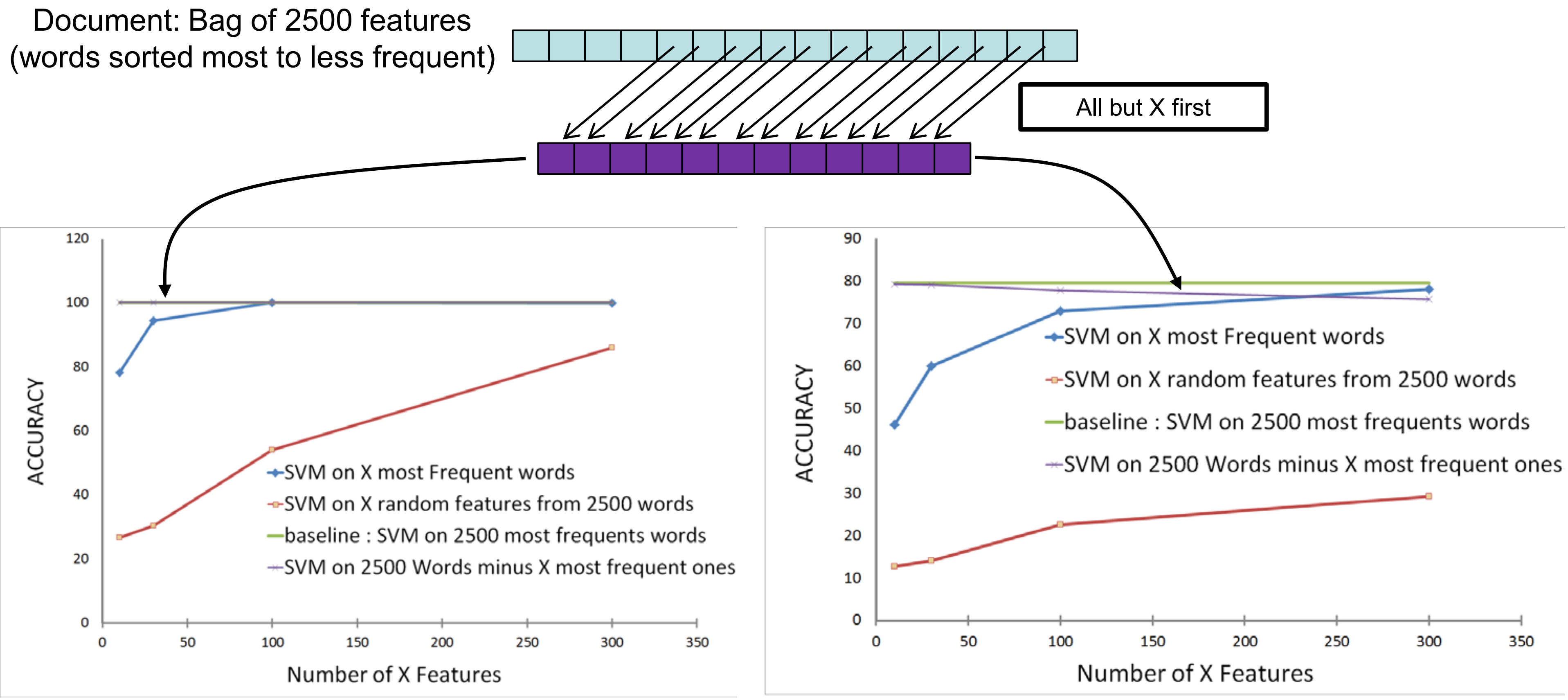
Undertraining investigation



Training accuracy

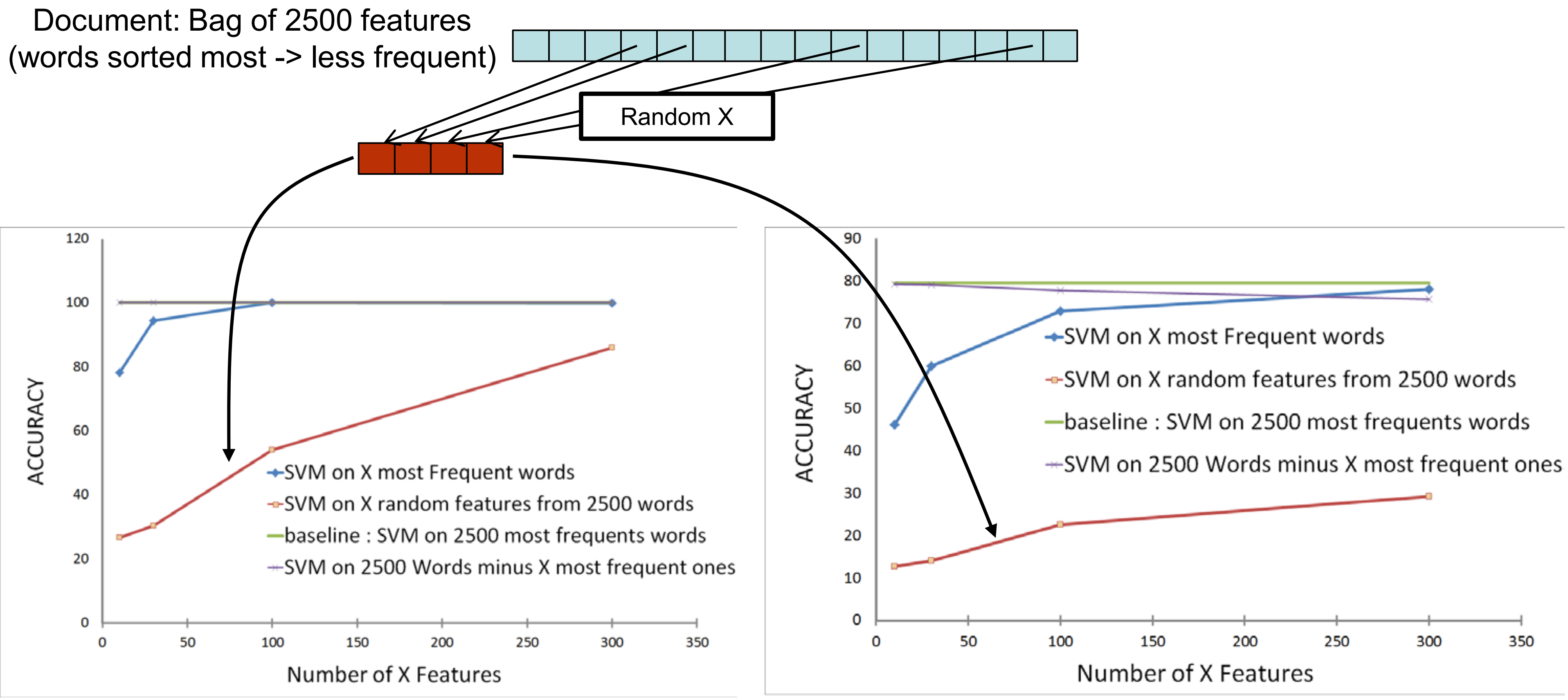
Validation accuracy

Undertraining investigation



Training accuracy Validation accuracy

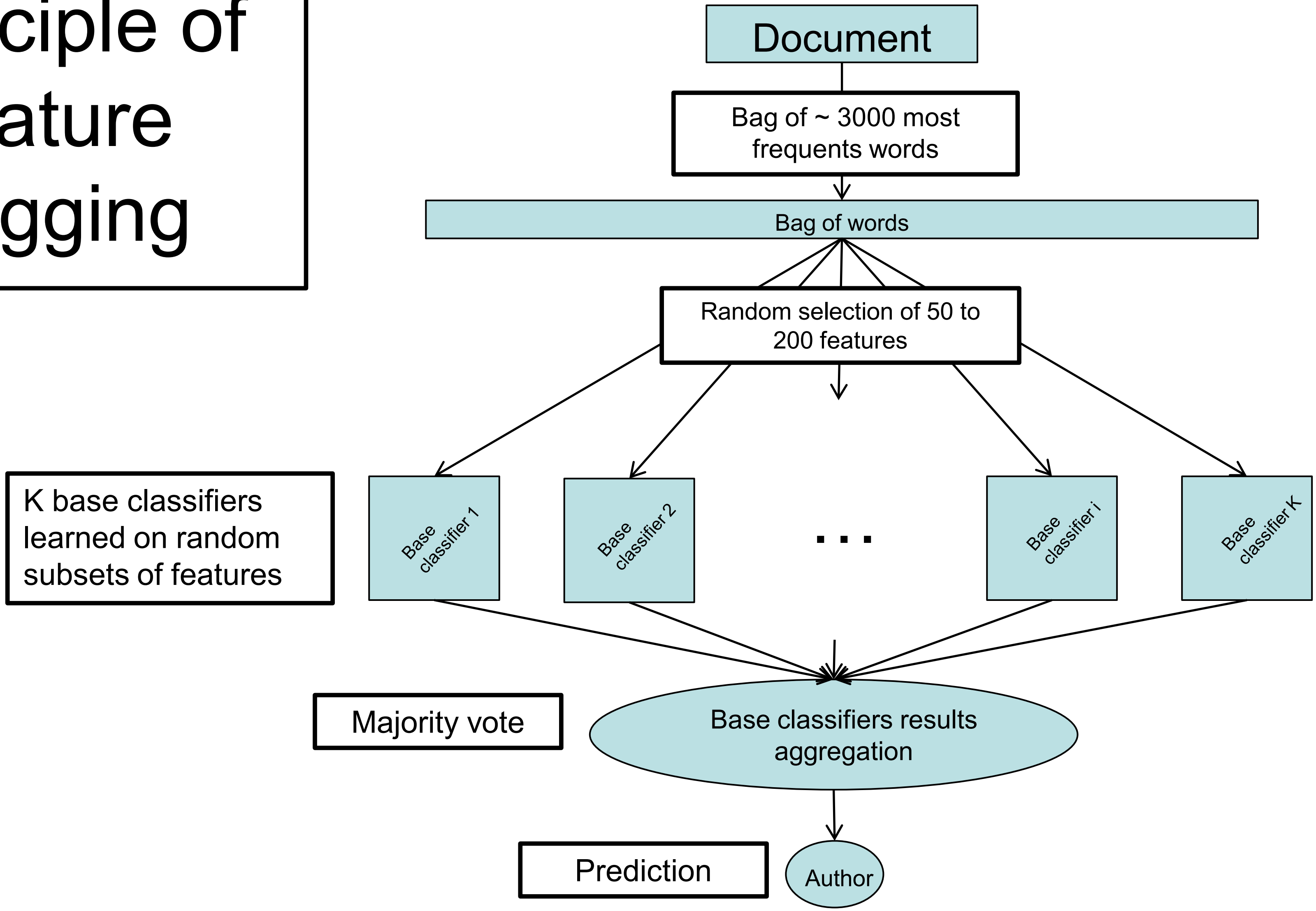
Undertraining investigation



Training accuracy

Validation accuracy

Principle of feature bagging



Preliminary results

English public available blog corpus

Statistics on Base classifiers

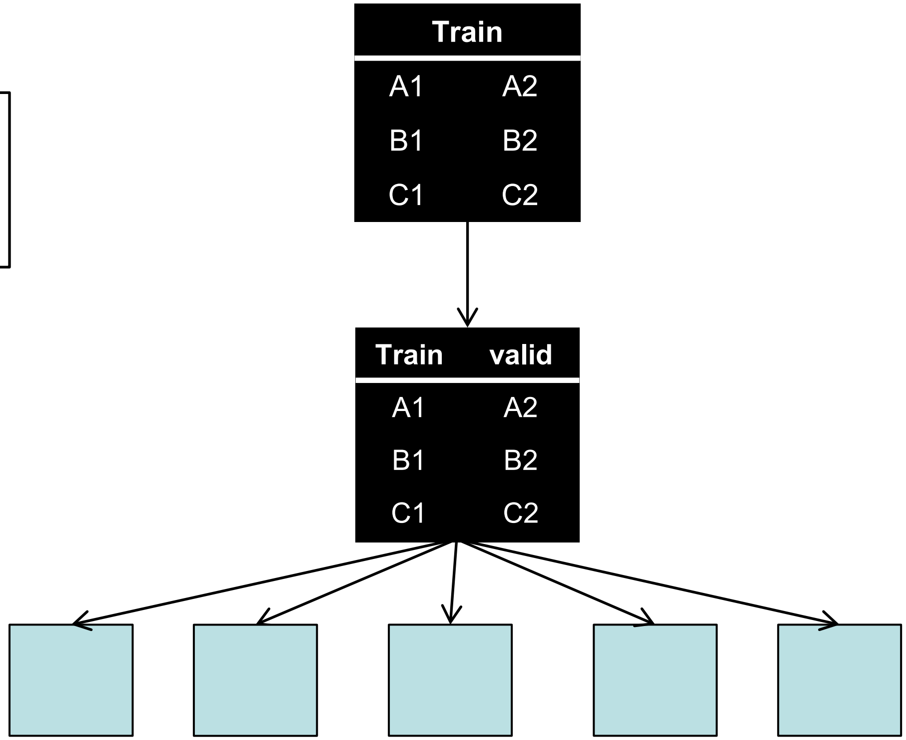
	Minimum			Mean			Maximum		
# features	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
100	99.8	33.2	32.2	99.8	45.6	42.7	100	56.1	53.3
225	99.8	50	46.1	99.9	60.5	55.8	100	69.4	64.4
600	99.8	55	48.9	99.9	65.5	60.1	100	75.5	67.8

Comparison with Baseline

Model	Train	Valid	Test
Bagging (100 features)	99.9	82.2	79.4
Bagging (225 features)	100	83.9	76.7
Bagging (600 features)	100	83.9	76.1
Single SVM with all 3000 features	100	79.4	71.6

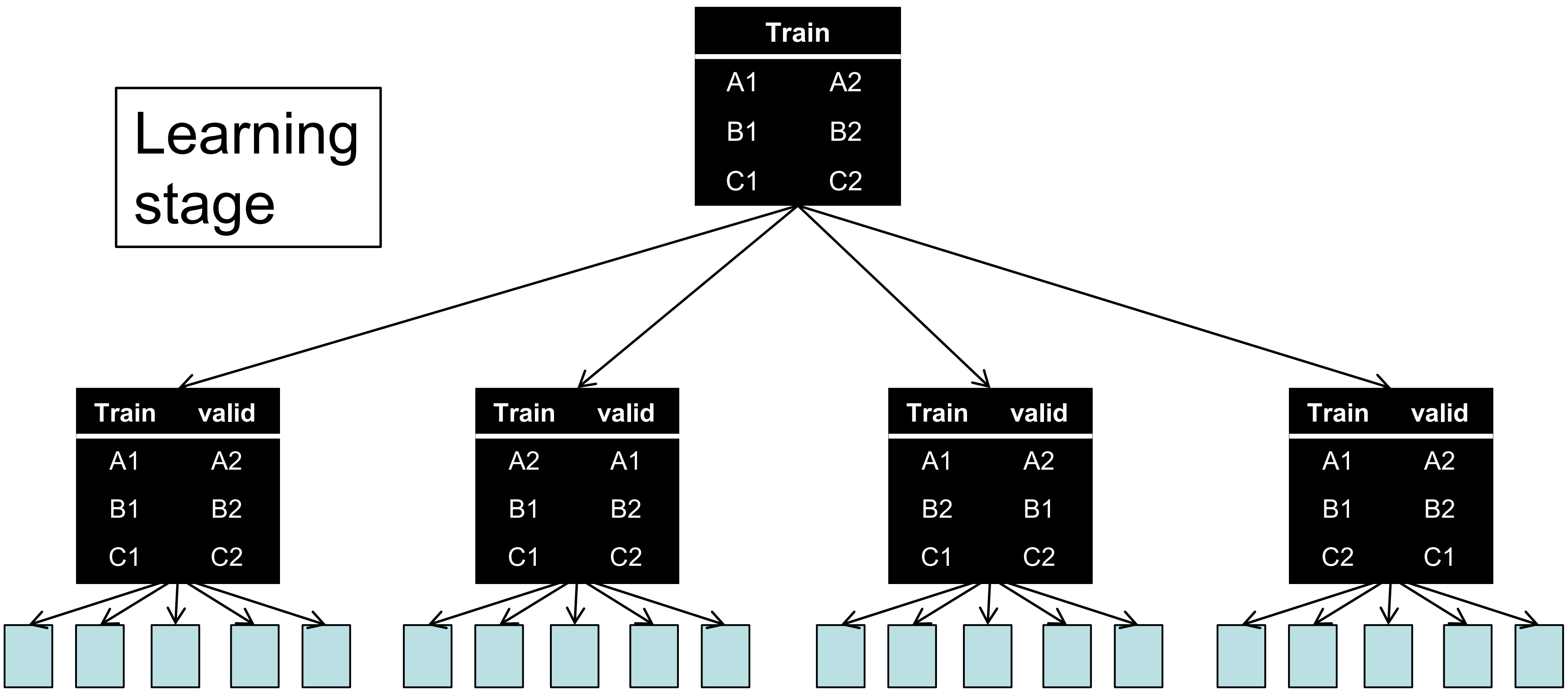
Experimental methodology for PAN

Learning stage



Experimental methodology for PAN

Learning stage

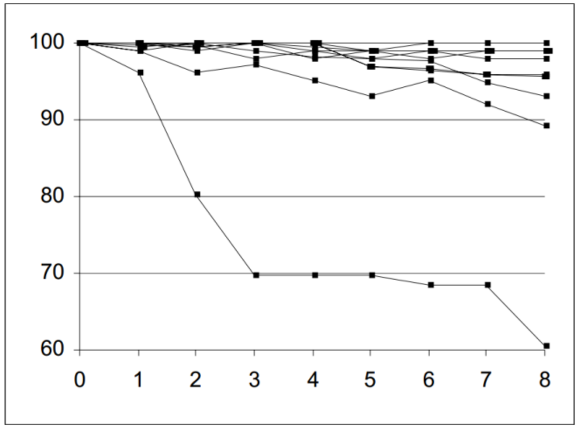


Comments on PAN results

- Less random features works better.
- Better ranks on closed tasks
- Reject method have to be improved
- Interest to use severals training/validation split

Perspective : A two Stage Approach

- Motivation
  - The way the classifier behaves when removing features depends on the author [Koppel 2007]



Author profiles for unmasking method, [Koppel 2007]

- Investigate mixing
  - this result

  with
  - our feature bagging approach

Two Stage Approach

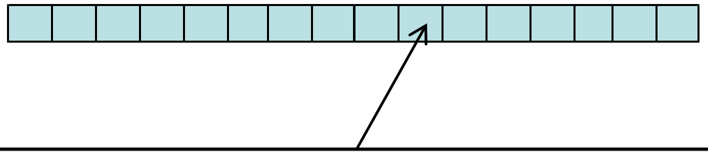
1. Bagging Appproach
   Learn multiple base classifiers erxploiting random selected subsets of features.

2. Building new data (called profile) for each pair (document, author)

3. (Optional) sort all vectors of the new dataset according to.

4. Learn a binary classifier to say if a profile is correct or not

Profile vector for document d and author a



j^{th} component stand for the percentage of base classifier that exploit feature j and predict a

Two Stage Approach

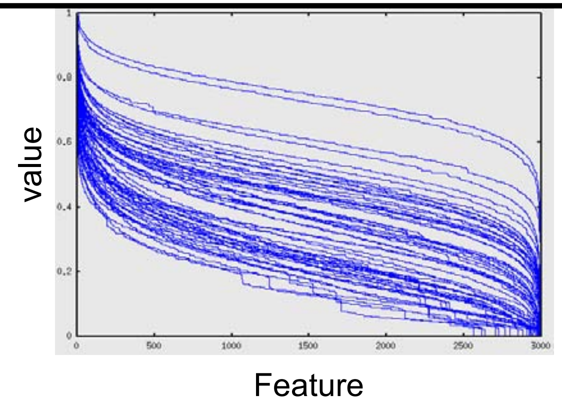
1. Bagging Appproach
   Learn multiple base classifiers erxploiting random selected subsets of features.

2. Building new data (called profile) for each pair (document, author)

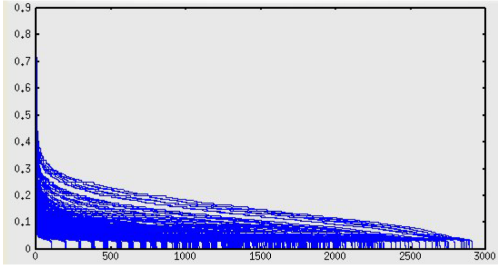
3. (Optional) sort all vectors of the new dataset according to.

4. Learn a binary classifier to say if a profile is correct or not

True author (sorted) profiles



False author profiles



Similar results as Bagging approach

Conclusion and further works

- Feature bagging approach to enforce exploiting all features
  - ⇒ Outperforms the SVM baseline
  - ⇒ Should be improved for handling open problems (cf PAN results)
- Similar results of the second approach
  - While different representation
  - ⇒ Should be combined

ANY QUESTION ?

Additional results on PAN

TASK	Run Name	K (# splits)	N (# Models / split)	# Models overal	Type of feature	# Random features	Open/Closed task	Accurracy
A	Lip6 1	8	100	800	WORDS-1500	200	closed	100
B	Lip6 1	8	100	800	WORDS-1500	200	open	70
A	Lip6 2	8	1	8	3CHAR-3500	3500	closed	100
B	Lip6 2	8	1	8	3CHAR-3500	3500	open	60
A	Lip6 3	8	100	800	WORDS-1500	300	closed	100
B	Lip6 3	8	100	800	WORDS-1500	300	open	70
C	Lip6 1	10	100	1000	WORDS-1500	400	closed	100
D	Lip6 1	10	100	1000	WORDS-1500	400	open	41.18
C	Lip6 2	10	300	3000	3CHAR-3500	1000	closed	75
D	Lip6 2	10	300	3000	3CHAR-3500	1000	open	52.94
C	Lip6 3	10	300	3000	3CHAR-3500	1250	closed	62.5
D	Lip6 3	10	300	3000	3CHAR-3500	1250	open	35.29
I	Lip6 1	12	1	12	WORDS-1500	1500	closed	85.71
J	Lip6 1	12	1	12	WORDS-1500	1500	open	81.25
I	Lip6 2	12	1	12	WORDS-2000	2000	closed	78.57
J	Lip6 2	12	1	12	WORDS-2000	2000	open	68.75
I	Lip6 3	12	1	12	WORDS-2500	2500	closed	78.57
J	Lip6 3	12	1	12	WORDS-2500	2500	open	75