max planck institut informatik

Linguistic approaches for text classification

Max-Planck Institut für Informatik

AG 5 – Databases and Information Systems Group

Oberseminar-Talk WS 04 / 05

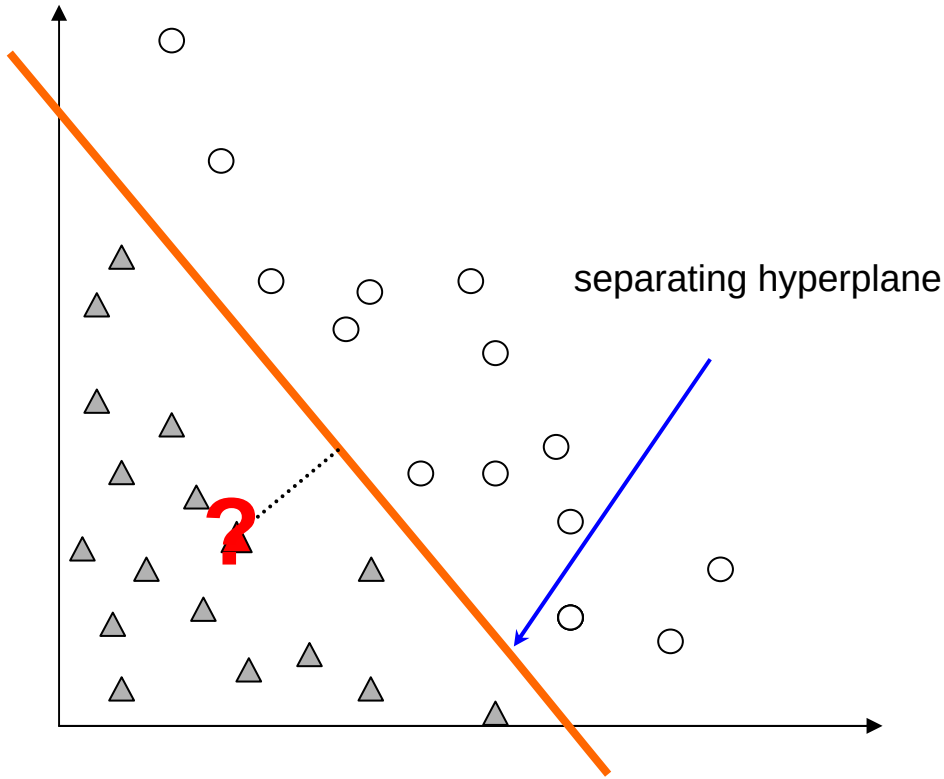
Speaker: Andreas Kaster

Supervisor: Stefan Siersdorfer

Why linguistic approaches?

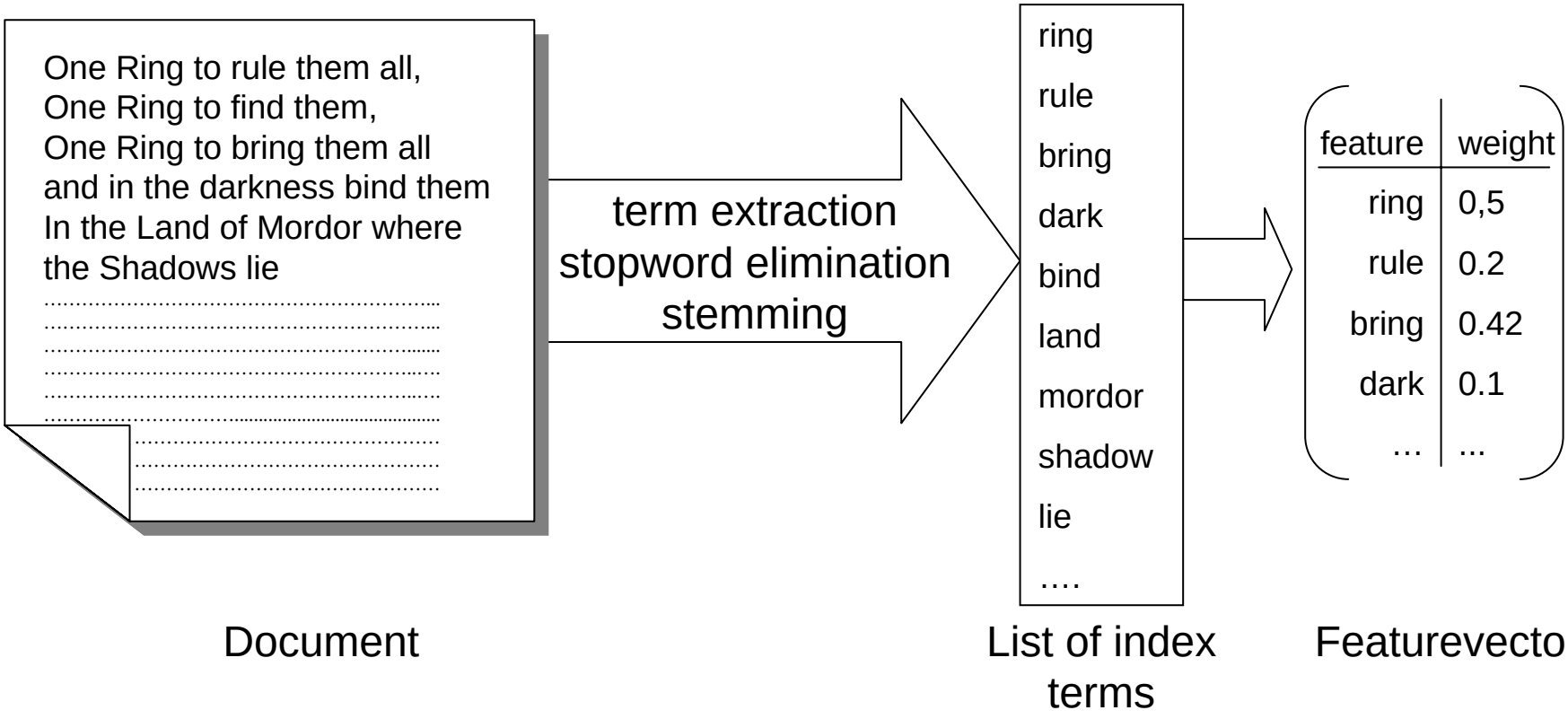
- Automatic text classification important for many different applications (web-portals, focused crawling, spam recognition…)
- Usual task: *topic based* (e.g. Politics vs. Sports)
- State of the art: Bag-of-Words-features
- But what about author recognition, sentiment classification?
- Goal: improve classification by alternative features

SVM - Support Vector Machine



How to map documents?

Bag of Words

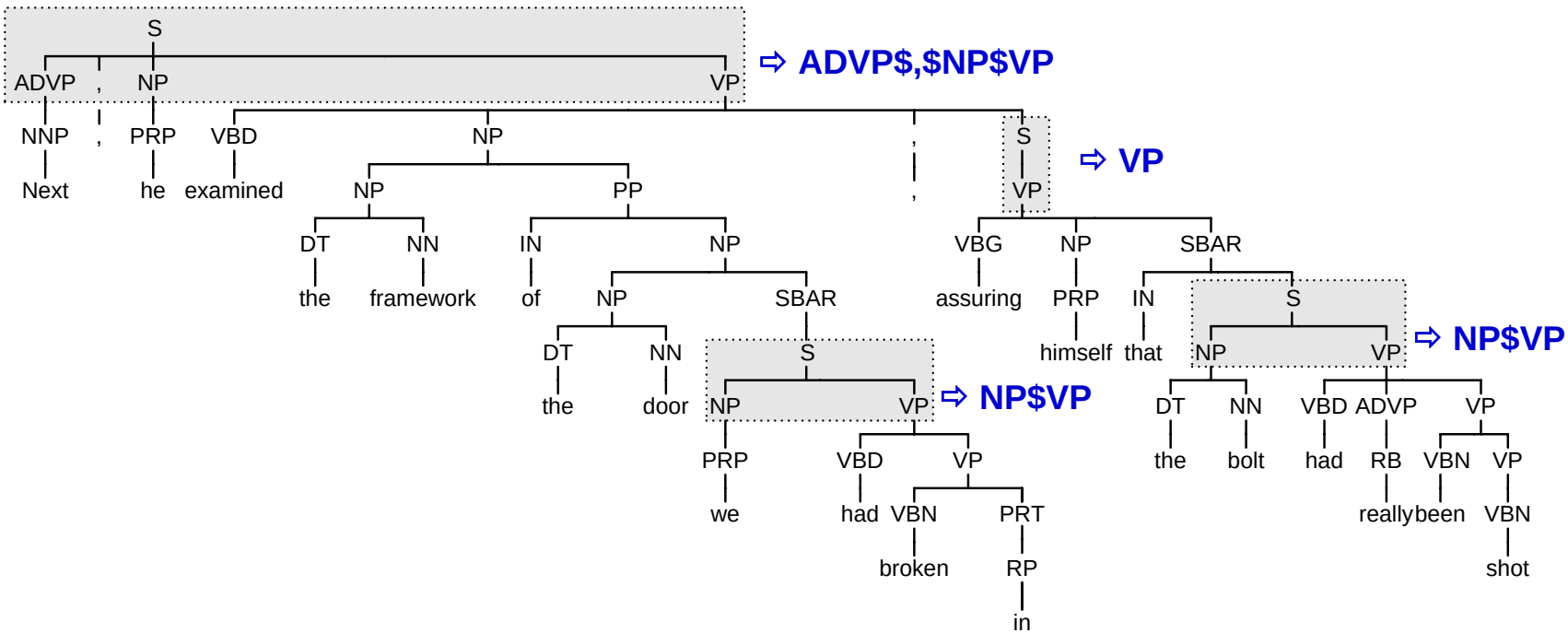

One Ring to rule them all,
One Ring to find them,
One Ring to bring them all
and in the darkness bind them
In the Land of Mordor where
the Shadows lie
term extraction
stopword elimination
stemming
ring
rule
bring
dark
bind
land
mordor
shadow
lie
….
feature
weight
ring
0,5
rule
0.2
bring
0.42
dark
0.1
…
...
Document
List of index terms
Featurevector

Alternative approaches

- word based features
  - ⇨Part of Speech tagging / -extraction (e.g. nouns, verbs…)
  - ⇨Considering word order (e.g. linguistic constituents)
- structure based features
  - ⇨Writingstyle
  - ⇨Functional Dependencies (i.e. subject, predicate, objects)
  - ⇨Sentence Length distributions
- combination techniques
  - ⇨Hybrid vectors
  - ⇨Meta classification

Different authors – different styles

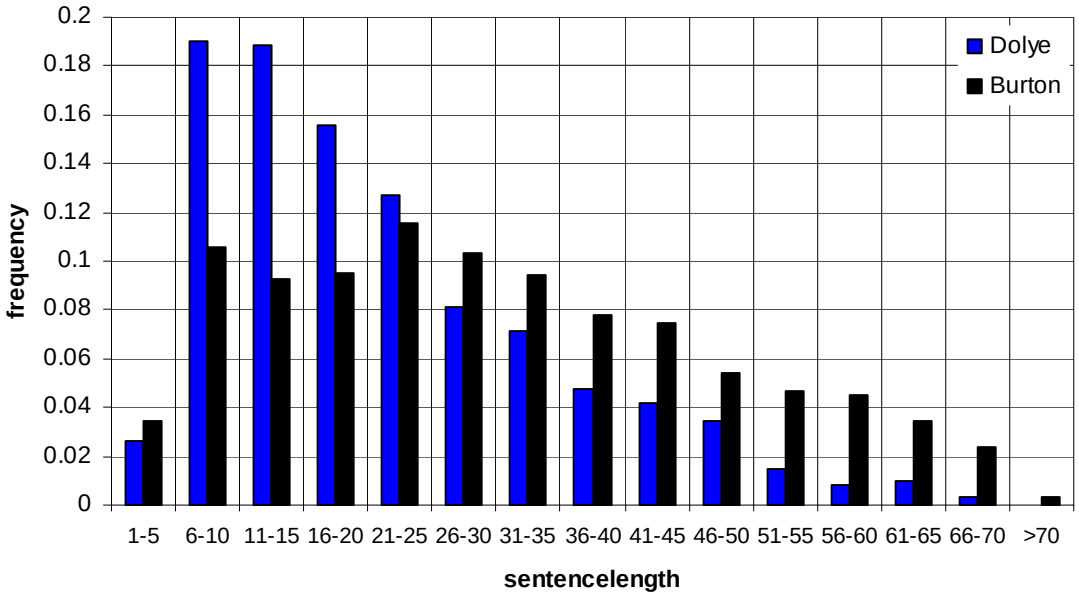
- Don’t consider what an author writes but how.
- Analyse PCFG-trees, use subtrees as features



⇨Featurevector: (ADVP$,$NP$VP, VP, NP$VP)

Sentence length – histogram approach

- buckets with fixed size for sentence lengths
- $b(i)$:= number of sentences in ith bucket



- Featurevector: $\left(\frac{b(1)}{n},\cdots,\frac{b(m)}{n}\right)$, m: # of buckets, n: # of sentences

- Alternative: $\left(E(X),E(X^2),E(X^3),E([X-E(X)]^2,E([X-E(X)]^3\right)$

Combination Vectors

- Given feature vectors $\vec{v}_1(d),\cdots,\vec{v}_k(d)$ of document d:

$$\vec{v}_i(d) = \left(v_{i1}(d),\cdots,v_{im_i}(d)\right)$$

  m: size of feature space for i[th] method

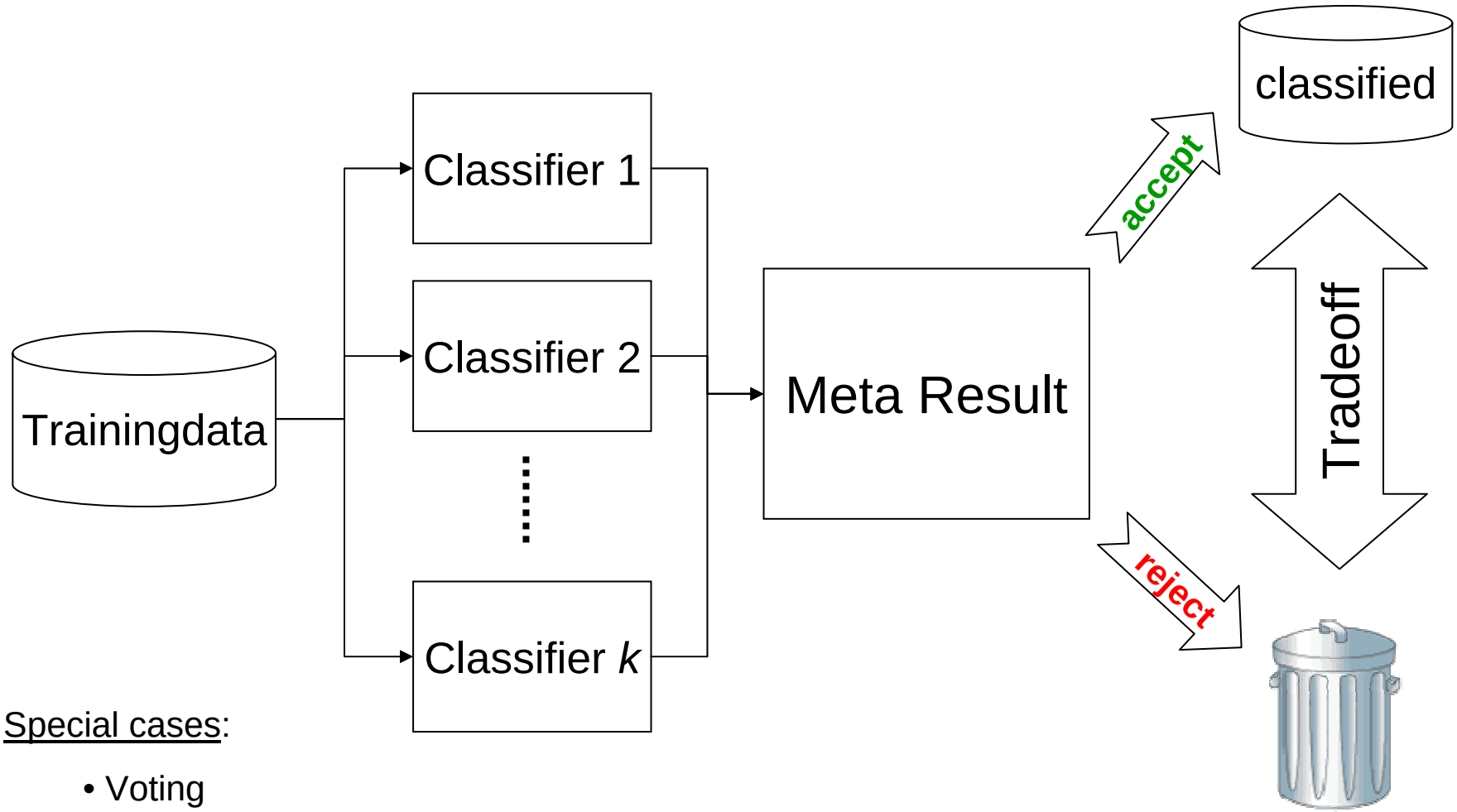
- Combine these vectors to

$$\left(\frac{v_{11}(d)}{c_1},\cdots,\frac{v_{1m_1}(d)}{c_1},\ldots\ldots,\frac{v_{k1}(d)}{c_k},\cdots,\frac{v_{km_k}(d)}{c_k}\right)$$

  with normalization constants $c_1,\ldots,c_k$

- Choose normalization constants that avg. component value is the same for all subvectors.

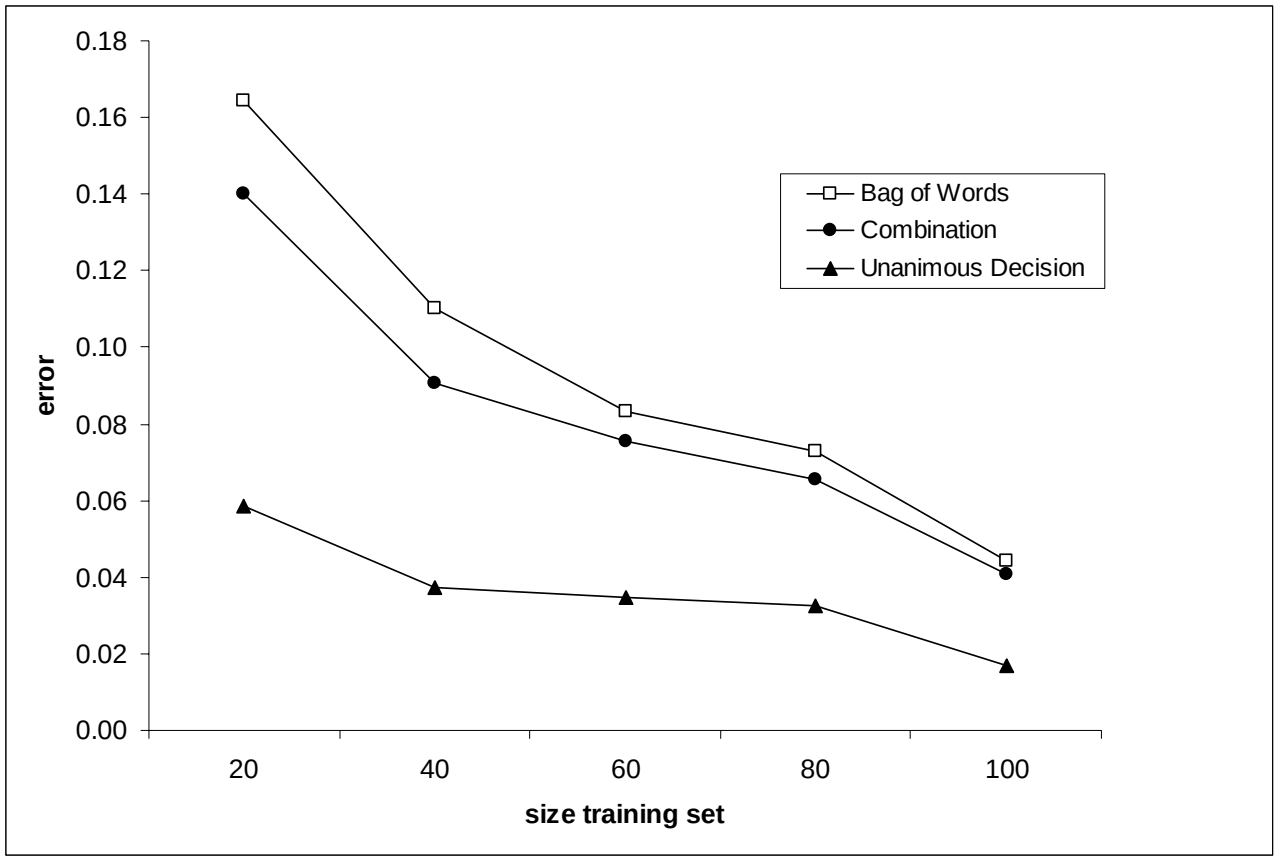
Meta Classification



Special cases:

- Voting
- Unanimous Decision
- Weighted Average

Experiments – Combination techniques



Corpus: Gutenberg, 10 authors ⇨ 45 pairs
Data: 20 – 100 training documents, 500 validation documents, known labels
Combined techniques: Bag of Words, histograms, writingstyle

Conclusion & Future Work

- Conclusion
- Analysis of linguistic structures provides us with interesting new features
- Conventional Bag of Words nearly unbeatable for topic classification task
- Combination techniques improve classification precision

Future Work….

- Linguistic features for query answering
- Anti Aliasing with author recognition
- Duplicate recognition
- Author clustering

Questions

Considering word order - constituents

- Words follow a specific order
- Constituents: syntactic units

> "President George Bush visited Germany at February."

⇨Constituent e.g. [President George Bush]

- Features:
  - ⇨Pairing:
    - bush$president, george$president, bush$george
  - ⇨Constituent features:
    - president$george$bush
- Apply additional techniques to reduce feature space (stopword elimination, stemming…)

TOP 5 Features by MI - Writingstyle

A.C. Doyle		R. Burton	
Feature	**MI**	**Feature**	**MI**
S$,$CCS.	0.23	S$:$S	0.26
PPNPVP$.	0.16	SCCS	0.23
SBAR$,$NPVP.	0.14	XXNP$VP	0.21
SBAR$,$XNPVP$.	0.13	S$:$S$.	0.20
PP$,$NPVP.	0.11	S$:$CC$S	0.18

PCFG – Probabilistic Contextfree Grammar

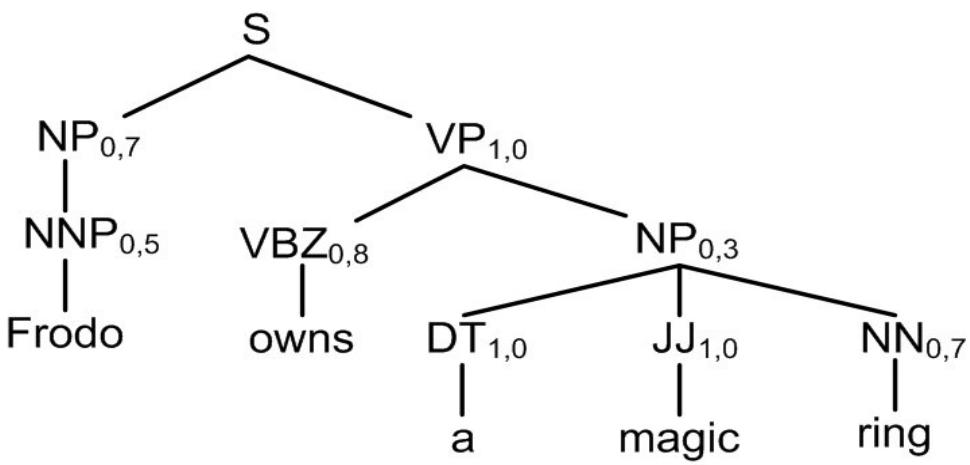
Sei $PCFG = \left(V, \Sigma, \Delta, S\right)$ mit

$\Sigma = \{\text{Frodo, Bilbo, owns, sleeps, a, magic, ring, carpet}\}$

$V = \{\text{S, NP, VP, NNP, VBZ, NP, DT, JJ, NN}\}$

Startsymbol: $S \in V$

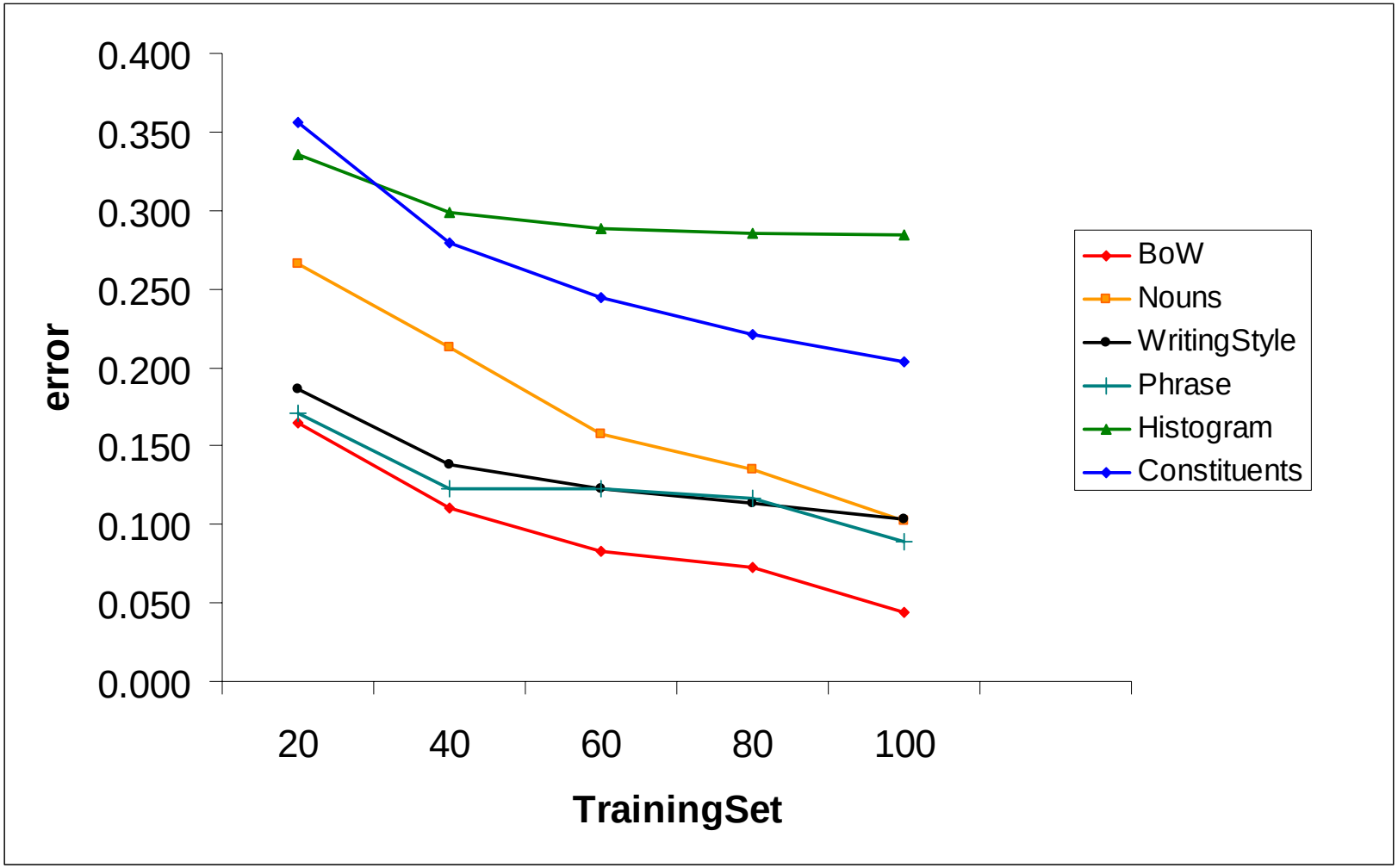
Δ			P[Δ]	Δ			P[Δ]
S	→	NP VP	1,0	VBZ	→	owns	0,8
NP	→	NNP	0,7	VBZ	→	sleeps	0,2
NP	→	DT JJ NN	0,3	DET	→	a	1,0
VP	→	VBZ NP	1,0	JJ	→	magic	1,0
NNP	→	Frodo	0,5	NN	→	ring	0,7
NNP	→	Bilbo	0,5	NN	→	carpet	0,3



Result-Table Gutenberg

# T	BoW	Nouns	WS	FD	Histogram	Constituent	Combi	U.D. err	U.D. loss
20	0.16	0.27	0.19	0.17	0.34	0.36	0.14	0.06	0.48
40	0.11	0.21	0.14	0.12	0.30	0.28	0.09	0.04	0.40
60	0.08	0.16	0.12	0.12	0.29	0.24	0.08	0.03	0.36
80	0.07	0.14	0.11	0.12	0.29	0.22	0.07	0.03	0.35
100	0.04	0.10	0.10	0.09	0.28	0.20	0.04	0.02	0.35

Experiments – Gutenberg



Experiments – 20 Newsgroups