litsift: Automated Text Categorization in Bibliographic Search

Lukas C. Faulstich — Humboldt Universität zu Berlin, Germany

Peter F. Stadler — Universität Leipzig, Germany

Caroline Thurner — Universität Wien, Austria

Christina Witwer — Universität Wien, Austria

ECML/PKDD 2003 Workshop on

Data Mining and Text Mining for Bioinformatics

Sept 22, 2003

Contents

1 Motivation

Goal: comparison of computational results from bioinformatics with experimental results from life sciences

Task: find relevant literature containing information on *conserved RNA secondary structures in viral genomes* for a fixed virus group

Complications:

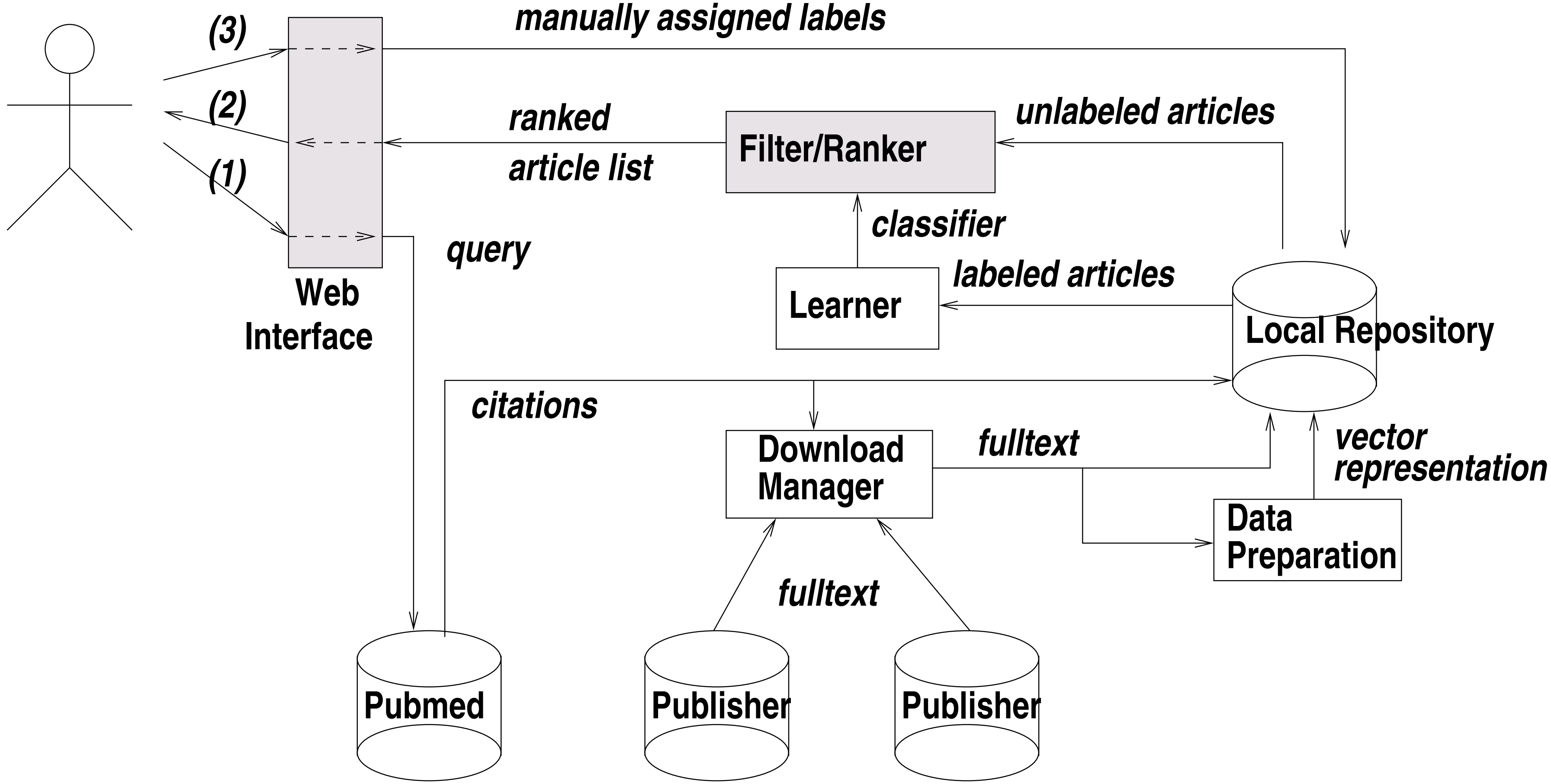
- relevant results may be hidden in articles with differing main topics
- key words may be omitted because context is clear or may be overloaded (e.g. secondary structure)
- no established nomenclature of RNA features in viruses

$\Rightarrow$ **Exploratory Project:** assess the feasibility of supporting broad bibliographic search with automated text categorization techniques (2PM).

2 Approach

1. learn relevant literature using training corpus (dedicated to a specific virus group, e.g. *Picornaviridae, Flaviviridae*)
2. create test corpus (on some other virus group) by searching bibliographic database and downloading referenced articles
3. apply trained classifier to test corpus
4. present articles as ranked list
5. manually relabel some test articles and use for retraining

2.1 Architecture



3 Data Sets

Corpus	Source	Size	Positive
picorna	Pubmed	40	68%
picorna2	Pubmed + Experts	64	58%
flavi	Pubmed	153	8%
flavi2	Pubmed + Experts	187	12%
hepadna	Pubmed	16	69%

4 Methods

4.1 Data Preparation

1. download: Perl wrapper scripts
2. PDF $\rightarrow$ Text conversion: `pdftotext`, `ps2ascii`
3. tokenization and full text index: `ConceptComposer`
4. term relevance measures: SQL script
   - Odds Ratio $OR(t,c) = \dfrac{P(t|c) \cdot (1 - P(t|\overline{c}))}{(1 - P(t|c)) \cdot P(t|\overline{c})}$
   - Mutual Information $MI(t,c) = \log \dfrac{P(t,c)}{P(t) \cdot P(c)}$
5. vector representation: SQL script, using tfidf term weights

(persistent storage: MySQL relational database)

4.2 Automated Text Categorization

Prototype: Java application on top of Weka 3 and MySQL. Supports crossvalidation on training corpus and validation on separate test corpus. External data download, preparation, and labeling.

Parameters for experiments:

- term relevance measure: $\{\mathrm{OR}, \mathrm{MI}\}$
- dimensionality: $\{10, 20, ..., 200\}$
- target recall: $\{80\%\}$
- classifier type $\{\mathrm{SMO}, \mathrm{J48}, \mathrm{N.B.}\}$
  (i.e., SVM, C4.5, Naive Bayes)
- classifier-specific parameters

5 Results

5.1 Feature Selection

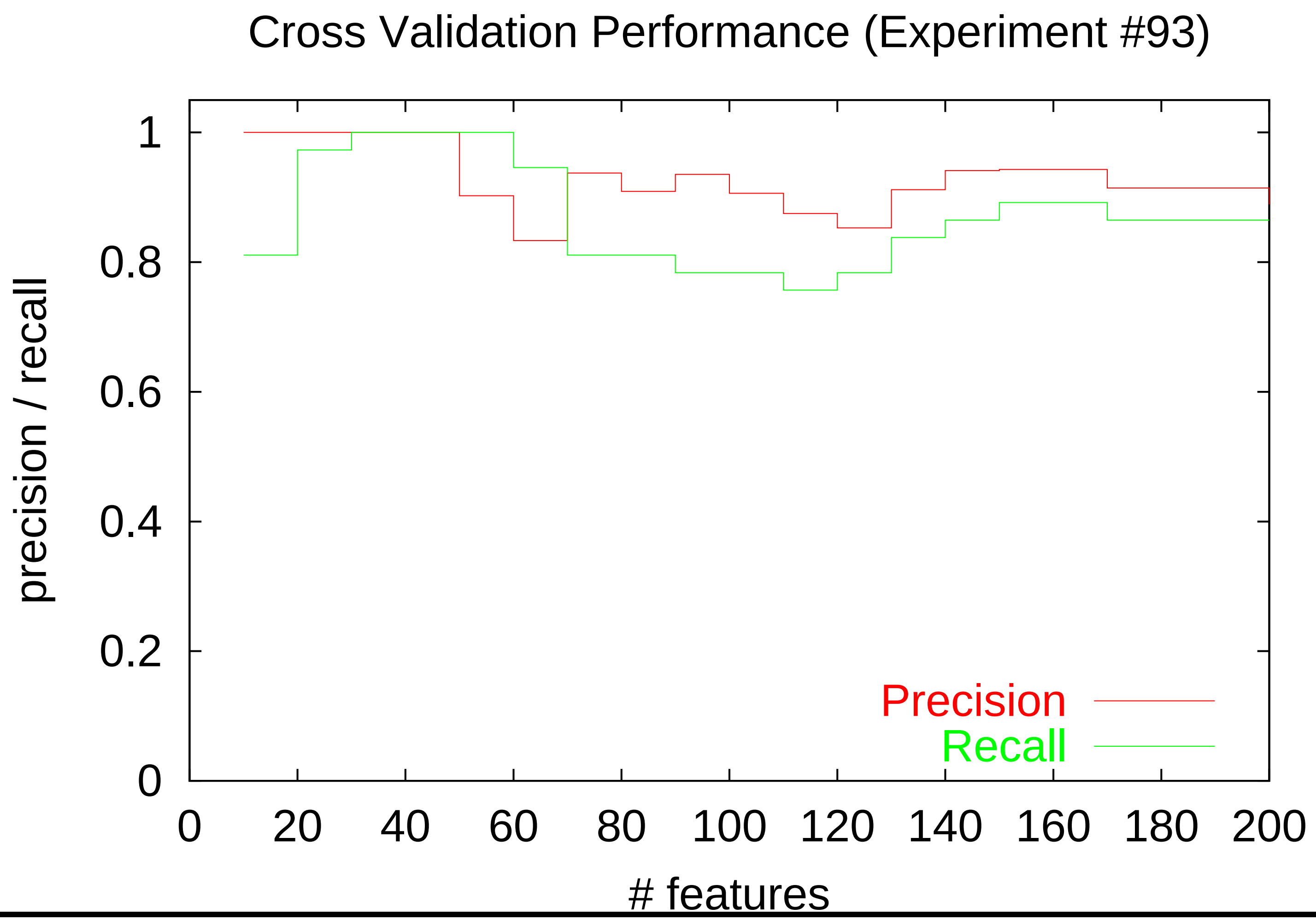
Relevance measure used as classifier. Threshold defined by target recall 100%. Average precision:

p_{avg}	**flavi**	**flavi2**	**picorna**	**picorna2**
OR	7.8%	11.8%	67.6%	58.0%
MI	11.2%	20.2%	76.7%	69.3%

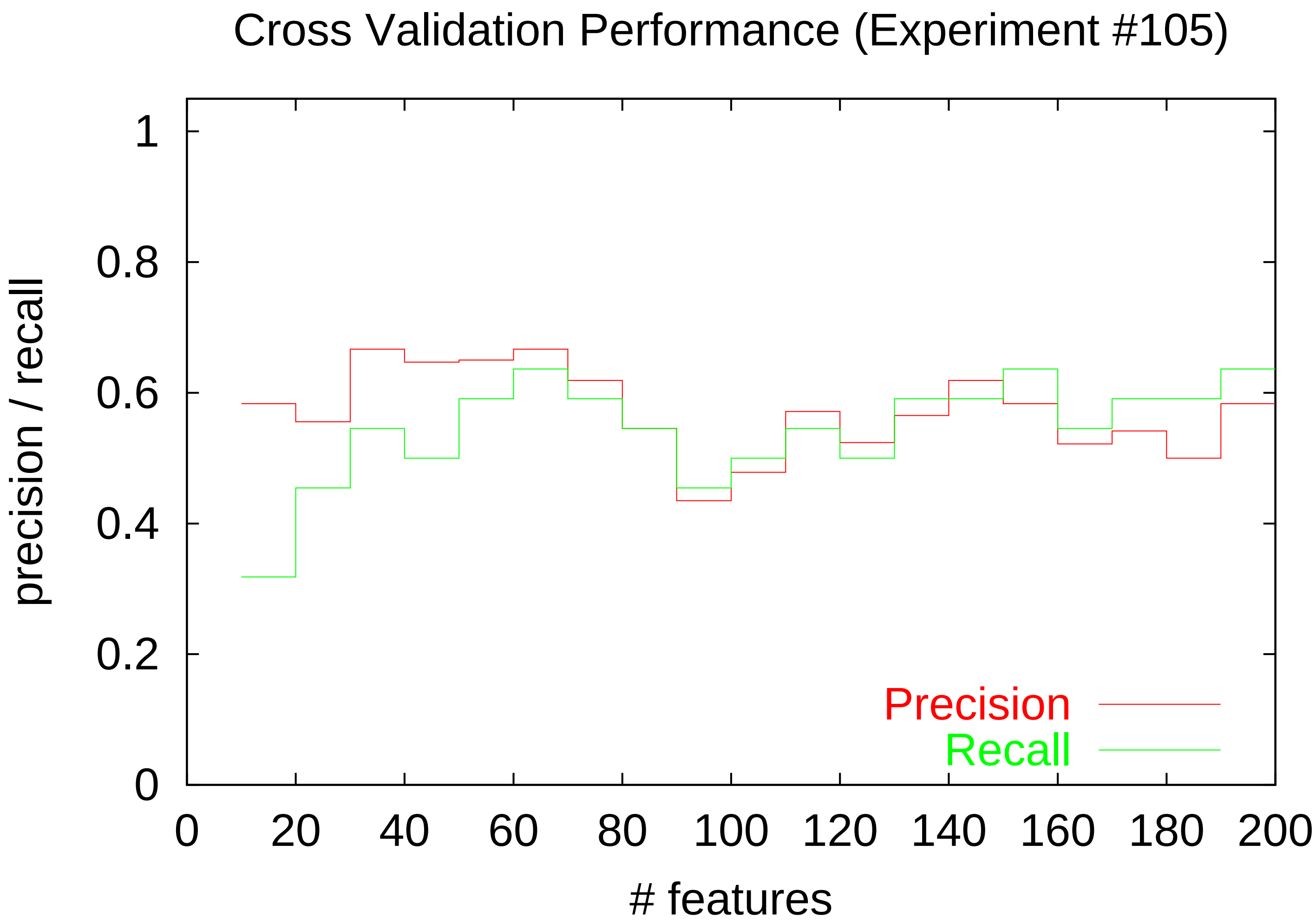
$\Rightarrow$ baseline for cross evaluation.

5.2 Cross Evaluation

Picorna corpora: easy to classify. E.g., SMO with MI on picorna2:



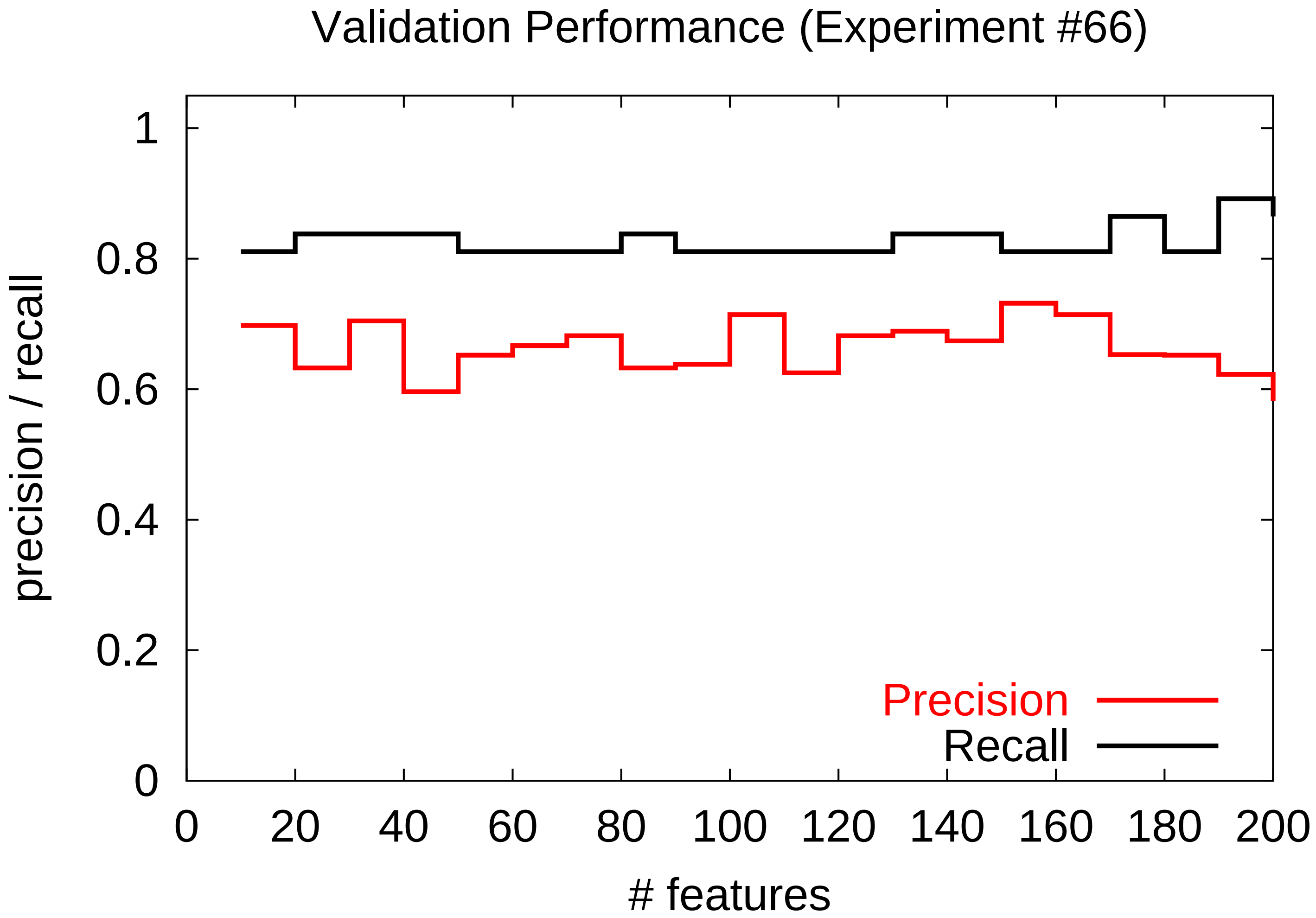
Flavi corpora: harder to classify. E.g., SMO with MI on flavi2:



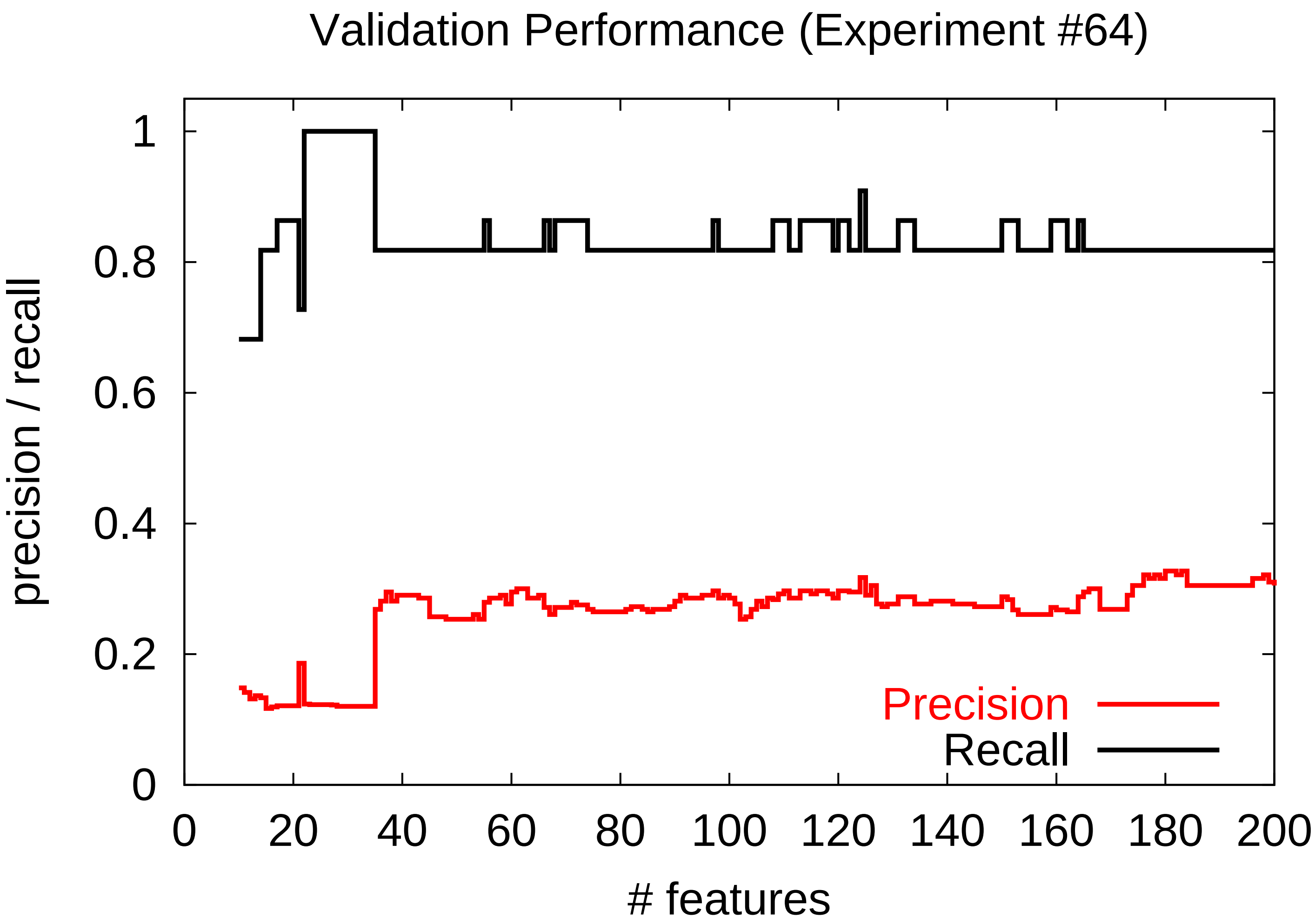
Typically less than 50 features needed for maximum precision.

5.3 Validation on Separate Test Corpus

Classifiers trained on Flavi corpora transfer well to Picorna corpora (e.g., SMO with OR, flavi2 $\longrightarrow$ picorna2)...



... but not vice versa (e.g., SMO with OR, picorna2 $\longrightarrow$ flavi2)



Still, even a low precision may save work...

6 Cost Model

Task: find at least a fraction r of all relevant documents within a bibliographic search result, i.e., target recall is r.

Goal: minimize fraction q of articles to be inspected manually.

Baseline: random selection with probability r requires $q_{\text{rand}} = r$ and yields recall r.

With classifier: classifier with precision p requires
$q_{\text{auto}} = \min(P(c)r/p, 1)$ where
$P(c)$ frequency of relevant documents

Work reduction: $s = (q_{\text{rand}} - q_{\text{auto}})/q_{\text{rand}} = 1 - P(c)/p$ if $P(c) \leq p$

6.1 Work Reduction (Examples)

training	**test**	**class**	**msr**	$P(c)$	p_{max}	r	s
flavi2	picorna2	SMO	MI	58%	83.3%	100.0%	30%
picorna2	flavi2	SMO	OR	12%	32.7%	81.8%	**63%**
flavi2	hepadna	SMO	OR	69%	90.9%	90.9%	25%
picorna2	hepadna	SMO	OR	69%	90.0%	81.8%	25%

7 Conclusions

- classifiers can be transferred among corpora on different virus groups, at the cost of reduced precision
- low precision can still reduce manual work significantly, especially with infrequent classes
- work reduction allows to broaden search queries and to increase overall recall

8 Future Plans

- experiment with classifiers for partially unlabeled data sets
- complete implementation of *litsift* tool:
  - implement Web interface based on `Apache Cocoon`
  - re-implement download manager in Java, based on `Apache xalan` and `JaxME`.