Fragments and Text Categorization

Jan Blaťák, Eva Mráková and Luboš Popelínský
Knowledge Discovery Lab
Faculty of Informatics, Masaryk University
602 00 Brno, Czech Republic
{xblatak,glum,popel}@fi.muni.cz

- two novel methods of text categorization in which documents are split into fragments.
- experiments on English, French and Czech.
- both methods increase the accuracy of text categorization
- for the Naïve Bayes classifier this increase is significant

Methods

`skip-tail`

only the first X sentences of a document are used

`fragments`

splits the documents into fragments which are classified independently of each others

Data

	n	docs	ave_s	sdev_s
20 Newsgroups	138	4040	15.79	5.99
Reuters-21578	4	1022	11.03	2.02
Medline	1	235	12.54	0.22
French cooking	36	1370	9.41	1.24
Czech newspaper	15	2545	22.04	4.22

Experiments

Feature (i.e. significant word) selection

chi, **ig**, **f1** and Probability Ratio (**pr**). **ig** yielded the best results

Three learning algorithms

J48, Naïve Bayes, SVM (SMO).

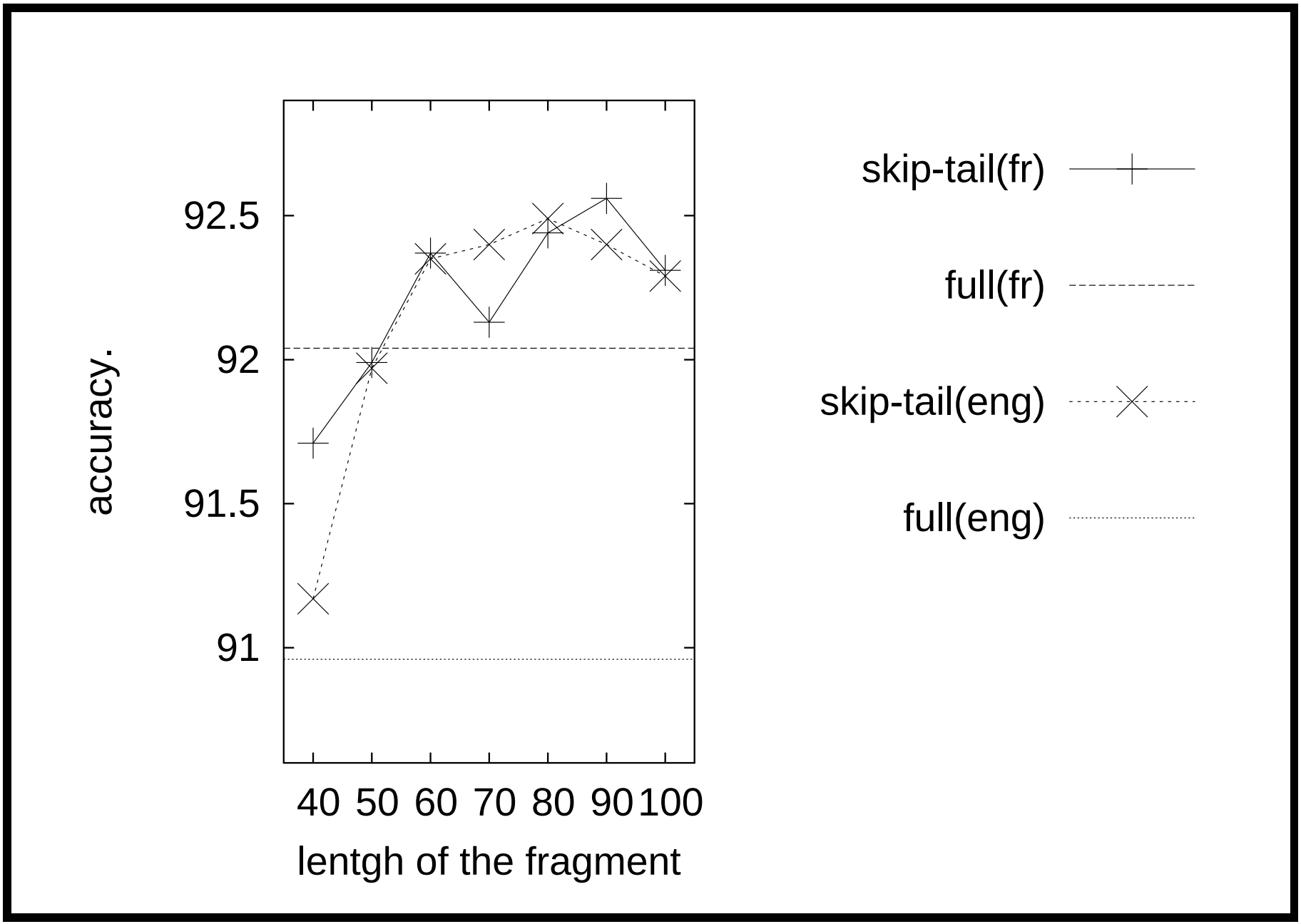
Length of fragments

1–15, 20, 25, 30, and 40 sentences.

Evaluation criterion

accuracy (the percentage of correctly classified documents from the test set)

10-fold cross validation.


accuracy.
92.5
92
91.5
91
40 50 60 70 80 90 100
lentgh of the fragment
skip-tail(fr)
full(fr)
skip-tail(eng)
full(eng)