On feature distributional clustering for text categorization

Ron Bekkerman
CS Department
The Technion
Haifa 32000 Israel
ronb@cs.technion.ac.il

Ran El-Yaniv
CS Department
The Technion
Haifa 32000 Israel
rani@cs.technion.ac.il

Naftali Tishby
School of CS and Engineering
and Center for Neural
Computation,
The Hebrew University
Jerusalem 91904 Israel
tishby@cs.huji.ac.il

Yoad Winter
CS Department
The Technion
Haifa 32000 Israel
winter@cs.technion.ac.il

Abstract

We describe a new powerful text categorization method that is based on a combination of distributional features with a support vector machine (SVM) classifier. Our feature selection approach uses distributional clustering of words via the recently introduced information bottleneck method, which generates a more efficient representation of the documents. When combined with the classification power of Support Vector Machines we produce the best known multilabel categorization results on the 20 Newsgroups dataset. Bottleneck (IB) clustering framework. Specifically, in this approach IB clustering is used for representing a document in a feature cluster space (instead of feature space), where each cluster is a distribution over document classes. As we show, this relatively new distributional representation, combined with a Support Vector Machine (SVM) classifier, allows for the best reported result for a multi-class categorization of another well-known 20 Newsgroups dataset.