A New Method for Bootstrapping an Automatic Text Classification System Utilizing Public Library Resources

Arash Joorabchi and *Abdulhussain E. Mahdi*
Department of Electronic and Computer Engineering, University of Limerick, Ireland
{Arash.Joorabchi, Hussain.Mahdi}@ul.ie

RATIONALE

- Similar to physical libraries, Large-scale digital libraries could contain hundreds of thousands of items and therefore require deploying flexible querying and information retrieval techniques that allow users to easily find the items they are looking for.
- Classification of materials in a digital library based on a pre-defined scheme improves the accuracy of information retrieval significantly and allows users to browse the collection by subject.
- Manual classification is a tedious and expensive job requiring an expert cataloguer in each knowledge domain represented in the collection, and therefore deemed unfeasible in many cases.
- Automated Text Classification/Categorization (ATC) - the automatic assignment of natural language text documents to one or more predefined categories or classes according to their contents - has become one of the key techniques to enhance information retrieval and knowledge management of large digital collections.
- Machine Learning-based ATC methods have shown great potential in this regard as demonstrated by experimental results reported in the literature. however, the performance of these methods in real-world operational projects is considerably below that demonstrated in the experimental setting. This problem has been mainly associated with the lack of high-quantity and/or -quality labeled datasets for training the ML algorithms. This issue is known in literature as the bootstrapping problem.
- In this work we describe a new bootstrapping method for ML-based ATC systems. The proposed method is based on utilising public library resources (i.e., standard library classification schemes, and online public access library catalogues) and book description information retrieved from online book sellers' websites.
- The proposed bootstrapping method could be deployed in variety of ATC systems where there is no labelled text corpus available for training the system. It also has specific applications in developing ATC systems for organising digital libraries, where E-documents are needed to be classified based on library classification schemes such as Library of Congress Classification (LCC) and Dewey Decimal Classification (LCC).

Dewey Decimal Classification Scheme

- The most common library classification scheme, used in about 80% of libraries around the world.
- have undergone numerous revisions and updates since it was developed in 1876.
- Fully hierarchical with ≈100,000 classes.
- We use DDC as a pool of categories/classes that can be selectively adopted by the users to create their own classification scheme according to the requirements of their classification task.

The Library of Congress

- The largest library in the world by shelf space and holds the largest number of books.
- Its collections include more than 32 million catalogued books and other print materials.
- For each DDC Class in the user's classification scheme a search query is submitted to the Library of Congress Online Public Access Catalogue for a list of all the holdings classified in that class by expert cataloguers.
- For each class a list of ISBN numbers of the book items which belong to that class is compiled from the search results. These ISBNs are used to retrieve the description of the books they represent from online booksellers' website.

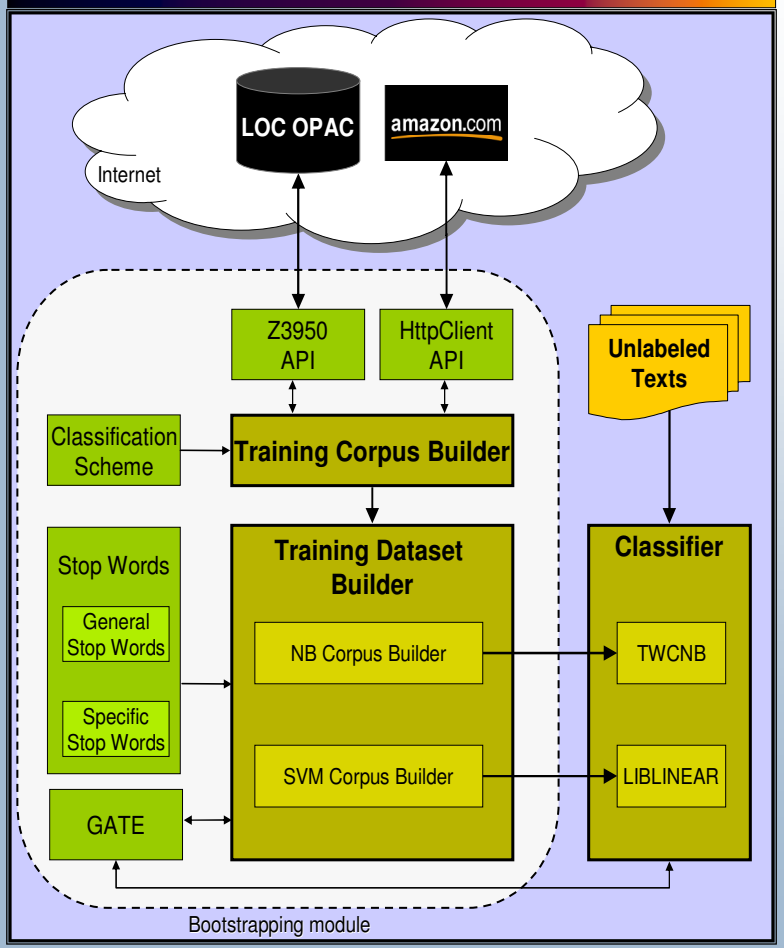
amazon.com

- The biggest online bookseller.
- we use the small parts of books such as cover pages which are publicly available on online book sellers' websites as an alternative to the full text of classified books to be used for training the ML-based classification algorithm. In specific, we use the textual content of *editorial reviews* section of book descriptions provided by Amazon website. This section contains short descriptions of the book such as editors' reviews, topics covered, and the back cover. Although the editorial reviews texts are usually short (less than 500 words) and are not available for all the books, they contain valuable semantically-rich keywords which expose the books' main topics/categories.

Classifier

- The collected training set is used to train an ML-based classification algorithm.
- In the developed experimental ATC system, we measure the performance of the proposed bootstrapping method by using the automatically collected training set to train a Linear SVM and a Naïve Bayes classification algorithm.

PROTOTYPE OVERVIEW



EXPERIMENTAL RESULTS

- An standard benchmarking dataset for text categorization called 20news-18828 collection was used to evaluate the performance of the system. It is a collection of 18,828 Usenet newsgroup articles, partitioned (nearly) evenly across 20 different newsgroups. In order to use this collection as the test dataset, we mapped eight classes in 20-Newsgroups collection to their corresponding classes in Dewey Decimal Classification scheme. Table 1 shows this mapping and also the number of training documents that have been automatically collected for each class
- Table 2 shows the achieved precision in each class by the naive Bayes classifier.
- The Linear SVM classifier achieved an average accuracy of 68%.

Table 1. 20-Newsgroups to DDC mapping

newsgroup	*Dewey Number*	*Dewey Caption*	*No. of training texts collected automatically*
sci.space	520	Astronomy and allied sciences	810
rec.sport.baseball	796.357	Baseball	997
rec.autos	796.7	Driving motor vehicles	587
rec.motorcycles	796.7	Driving motor vehicles	587
soc.religion.christian	230	Christian theology	1043
sci.electronics	537	Electricity and electronics	713
rec.sport.hockey	796.962	Ice hockey	270
sci.med	610	Medicine and health	1653

Table 2. Naïve Bayes classifier results

newsgroup	*Precision%*
sci.space	69.19
rec.sport.baseball	96.78
rec.autos	74.74
rec.motorcycles	71.02
soc.religion.christian	89.36
sci.electronics	69.92
rec.sport.hockey	75.77
sci.med	76.23
	Avg. 77.87

CONCLUSION

- the proposed bootstrapping method has the potential of achieving better accuracies, given richer sources for collecting the training datasets. In order to further investigate and exploit this potential, in future we plan to:
  - Experiment with alternative sources for collecting and building the training dataset. One of these sources is Google Book Search project, which scans and OCRs about one million books a year.
  - Also by increasing the number of library catalogues searched by the system we can retrieve the ISBNs of more classified books in each class and therefore, increase the size of training sets. One possible method to do this is the use of union catalogues such as worldCat which allow us to search the catalogue of many libraries at the same time.