Report on the Workshop on Operational Text Classification Systems (OTC-02)

Susan T. Dumais
Microsoft Research
Redmond, US
sdumais@microsoft.com

David D. Lewis
Independent Consultant
Chicago, US
dave@daviddlewis.com

Fabrizio Sebastiani
CNR
Pisa, IT
fabrizio@iei.pi.cnr.it

The *2nd Workshop on Operational Text Classification Systems* was held on 15 August 2002 in Tampere, Finland, in conjunction with SIGIR 2002. The workshop was chaired by Fabrizio Sebastiani and the organizing committee was composed by Susan Dumais, David D. Lewis, Tom Montgomery, and Isabelle Moulinier. 42 people attended the workshop.

The research side of text classification has been widely discussed in conferences and journals. In contrast, operational text classification has been covered in the popular media, but less so in technical forums. Issues other than effectiveness, such as engineering and workflow issues, have not been widely discussed in published research. The goal of this workshop was to expose researchers and practitioners to the challenges encountered in building and fielding operational text classification systems.

Eight talks were presented, spanning a wide range of text classification applications. Discussions followed each talk. An extended discussion period in which key themes and directions were identified concluded the workshop.

Presentation Summaries:

1. **Jean-Ronan Vigouroux** presented preliminary experiments by Thomson Multimedia and Singingfish on assigning content categories to multimedia streams on Web sites. Categories are used to aid end user navigation of search results for this material. Singingfish currently categorizes this material by taking advantage of the fact that collections of files in a given directory of a web site often all belong to the same category or categories. Rules, often written specially for each site, classify files based on their URL. The experiments reported in the talk tested the use of a Naive Bayes classifier to further automate this process by using the (typically brief) textual metadata associated with each file.

2. **Frank Smadja** of Elron Software reviewed spam detection and blocking techniques used in commercial systems. He then discussed Elron Software's approach, which runs on a client's mail server and inserts a tag into the subject line of suspected junk messages. Elron periodically trains a spam classifier using a Rocchio-style method and ships it to clients. Traditional word-level features are augmented with collocations, numeric thresholds, ratios, and regular expressions to define patterns. Smadja emphasized that manual ranking, selection, and engineering of predictor features was necessary to achieve and maintain sufficient accuracy over time. He also emphasized the need for substantial experiments with real data, warning that accuracy in the field is inevitably lower than in the lab. He noted that error costs are not the same (classifying good mail as spam is more costly than classifying spam as good mail) so the classifier must be tuned accordingly. Periodic retraining, using additional data supplied by customers, is necessary to

keep up with new types of spam (and new types of nonspam with spam-like characteristics). However, this is difficult, since because of privacy concerns customers are not much willing to provide their own nonspam messages to be used as training examples.

3. **Kees Koster** described collaborative work by University of Nijmegen, Edmond Research, and Fiscaal up to Date in the EU-supported Peking project. The goal of the project is to map a large collection of fiscal law documents from one version of a large taxonomy to a new version. Koster proposed taking a comprehensive approach including tests of classification on the existing category system, discarding or relabeling of difficult-to-classify documents, and revising the category system for both usefulness and ease of assignment.

He reported experiments on the first stage of this process, which involved testing the accuracy of Nijmegen's LCS system on a subset of the original categories. In addition to evaluation against the existing manual categorization, a sample of documents and their automated and manual categorizations was judged blindly by a panel of experts. The experts usually preferred the manually assigned category when there was a disagreement between manual and automated categories, but not always, and the original manual category was not always highly rated.

4. **Sung Hyon Myaeng** described his work with colleagues at ETRI on using text classification to support a question answering system. Their strategy is to classify documents with respect to a large set of financial concepts. They further qualify concepts with attributes describing aspects of the concepts, much as in a faceted indexing vocabulary. The goal is to map user questions to (concept, attribute) pairs, and retrieve documents categorized into the same or similar pairs.

Myaeng's talk focused on the task of assigning (concept, attribute) pairs to documents. There are more than 100,000 such pairs, so training a separate classifier for each would be both inefficient and require labeling too much training data. Defining classifiers separately for concepts and for attributes would greatly reduce the number of classifiers, but would ignore the fact that the character of an attribute is crucially tied up with the concept it modifies or refines. Myaeng and colleagues therefore adopted an intermediate approach by partitioning the concept hierarchy into groups of "alpha-related" concepts. Within an alpha-related group, the same classifier can be used for an attribute.

The concept network is also used to guide choice of documents to label for training data. The demands on training data are further reduced by manual selection of predictor features and by combining learned and manually constructed rules. Another novelty was an attempt to remove predictor features associated only with a single concept from a classifier for an attribute meant to be associated with many concepts. Myaeng presented results showing effectiveness improvements from combining learned and engineered rules, and from taking into account learned associations between attributes and groups of alpha-related concepts.

5. **Marc Krellenstein** described the use of text categorization to support Northern Light Technology's (NLT) search engine. Category assignments are used to group the documents returned by a free text search, and to improve the quality of their ranking. Multiple taxonomies of categories were used including a 16,000 node (9-level) subject taxonomy, a 150 node (3-level) document type or genre taxonomy, and several others. Taxnomies were constructed by librarians, drawing on a variety of existing taxonomies both conventional and unconventional (e.g., labels on grocery store aisles).

A variety of classification approaches were combined, including:

* linear classifiers trained by supervised learning

* manually engineered rule-based classifiers
* metarules that replaced multiple more specific category assignments with a single ancestor of them all
* mapping of existing category labels to NLT's categories (this was mostly for content NLT obtained from publishers, rather than from the web)
* limited manual categorization

Precision was emphasized over recall, with NLT's studies suggesting 90% precision was necessary for users to trust results. NLT was able, after considerable tuning, to produce classifiers that had 90-95% precision according to user judgments, though only 60-65% precision according to NLT's stricter judges. Recall was about 25% for web documents, but many of the "missed" documents were ones with little text or one serving a purely navigational purpose, where the notion of subject categorization may not be appropriate. Precision and recall were higher on non-web data.

6. **Peter Schäuble** discussed the experiences of Eurospider Information Technology AG with developing text categorization applications for commercial clients. He reviewed 10 criteria, emphasizing how they differed from TREC and test collection style evaluations. On the effectiveness side, he emphasized the need to support narrow categories, subtle distinctions among categories, and time varying categories. Average effectiveness is often less important than avoiding blatant errors, having reasonable behavior on all inputs that will be encountered, and distinguishing degrees of certainty in classification. Classification behavior which is easy for both users (and lawyers) to understand is important for both ongoing maintenance, as is verification that classifiers obey any legal restrictions (e.g. as in anti-money laundering systems). Techniques for reducing the amount of data needed for supervised learning, and leveraging existing knowledge of categories and relationships among categories are important. Finally cost and efficiency is always a concern.

7. **David D. Lewis**, an independent consultant, described a project undertaken with the National Center for Charitable Statistics (NCCS) to classify descriptions of activities of US non-profit organizations. The resulting category labels are widely used for locating and studying non-profit organizations in the U.S. The taxonomy used to classify programs is large and hierarchical. About 20,000 labeled examples were available, yet there were a number of issues with the training data. The quality of the labeled data varied (some labeling was done by interns and some by domain experts), some of the labels were derived from a slightly different version of the category hierarchy resulting in missing data and other inconsistencies, and some of the data was from a geographical sampling of organizations, which introduces some biases. And, even with this large amount of training data, more than 70% of the categories had fewer than 20 examples.

Each program has a short textual description which is the primary input into the text classification algorithm. In addition, there are other sources of potentially useful information to supplement these descriptions. Each organization has a textual description, and is classified into a different but related hierarchy. Category names are long and descriptive. And, programs and organizations have additional data associated with them that might be useful including location, dollar amount of programs, etc. Several different classification algorithms were explored. Classification accuracy remains below human accuracy levels, but is nonetheless useful.

8. **Natalia Loukachevitch** of Moscow State University described their Thesaurus on Sociopolitical Life, which has been used as the basis of eight distinct text categorization systems since 1995. The Thesaurus is a large hierarchical vocabulary of concepts, with extensive information on nominal, verbal, and adjectival constructs through which the concept might be

expressed in text. The Thesaurus is structured using broader term, narrower term, related term, and other relationships.

Categorization systems are then built on top of the Theasurus by manually writing rules to map from Thesaurus concepts to categories from the target hierarchy. The rule language allows making use of the hierarchical structure of the Thesaurus, specifying for instance what kinds of related concepts also provide evidence for a target category. Despite good support for rule writing, it is still a labor-intensive process, taking perhaps 1 person month to write 100 to 200 rules.

Categorization is then a two-stage process, first mapping text to Thesaurus concepts, and then mapping Thesaurus concepts to target categories. Loukachevitch argued that this approach has advantages over supervised learning approaches given the constraints they have faced in fielding systems: little or no training data, low levels of interindexer consistency for the target vocabularies, and low frequencies of some categories. She cited results indicating that the Thesaurus approach gave comparable effectiveness to SVM-based classification on one data set where training and test data were available, and had higher coverage on another data set.

Themes:

There were a number of recurring themes among the talks and in discussion among the workshop participants. A common theme is that the levels of data quality in operational settings are far below those found in experimental work. Shortages of labeled data, unknown biases in choice of data to label, low quality of labeling, and differences between training documents and those encountered when the system is used are all factors.

A related issue is the instability of categorization situations. The kinds of documents to be processed, the set of categories to be assigned, and the definitions of categories, can all change over time. Several speakers discussed issues of mapping from one version of a category set to a newer one, an issue which has received some attention in information science ("vocabulary switching") but very little in machine learning-based text categorization.

As at last year's workshop, many speakers discussed the desirability, and sometimes the absolute necessity of manually engineering features or, more generally, combining learned and engineered classifiers. It appears that relatively few operational systems use pure supervised learning approaches. The main reason for this seems to be to take advantage of human domain knowledge to construct features that would not be found by a learning system.

Another issue that comes up in practice is that of being able to understand a classifier once it is learned. Again there has been little attention to this issue in the research literature, not surprisingly, given that human experts do not enter into the experimental context.

A related area that will be challenging to explore in the research community is the notion, mentioned by at least three of the speakers, that some errors are plausible near-misses, while others are blatantly wrong. Unfortunately, it appears difficult for indexers to include all possible near-misses when labeling data, and even if they could it would greatly increase the expense of such labeling.

In spite of the challenges of deploying operational text classification systems, there is considerable enthusiasm for automated text classification. Most participants felt that automated systems, or automation in combination with human checking, can reduce classification costs in a variety of applications and increase consistency.