Report on the Workshop on Operational Text Classification Systems (OTC-01)

David D. Lewis
Independent Consultant
Chicago, US
E-mail: <dave@daviddlewis.com>

Fabrizio Sebastiani
Consiglio Nazionale delle Ricerche
Pisa, IT
E-mail: <fabrizio@iei.pi.cnr.it>

1 Goal of the Workshop

The Workshop on Operational Text Classification (OTC-01), occurred September 13, 2001 in New Orleans, Louisiana, US. It was co-located with ACM SIGIR 2002 and brought together researchers, practitioners, and system designers interested in building and fielding operational text classification systems. The workshop organizers were David Lewis (chair), Susan Dumais, Ronen Feldman, and Fabrizio Sebastiani.

Text classification research and practice has exploded in the past decade. This work has been pursued under a variety of headings (text categorization, automated indexing, document clustering, text mining, topic detection and tracking, etc.). Both the automated assignment of textual data to classes, and the automated discovery of such classes (by techniques such as clustering) have been of intense interest. A variety of practical applications have been fielded, in areas such as indexing of documents for retrieval, classification of Web sites under hierarchical directories, alerting, routing of news, creation of customized information products, enforcement of information security, filtering of unsuitable content, help desk automation, knowledge discovery in textual and partially textual databases, and many others.

However, while experiments on text classification data sets have been widely presented in a variety of forums, the details of text classification in actual use have rarely been discussed. The workshop therefore encouraged contributions discussing issues such as cost effectiveness, user needs, personnel and training, resource demands, maintaining effectiveness over time, and integration of classification with existing processes and software.

2 Format of the Workshop

Prospective speakers submitted a total of 18 extended abstracts, from which the program committee selected 9 for presentation. A call for demos was later issued among prospective attendees. Only 1 demo proposal was submitted, and it was accepted for presentation. About 45 participants attended the workshop.

The selected abstracts, along with brief interest paragraphs from attendees, were included in an informal notebook, whose distribution was limited to participants to encourage frank discussion. A subset of authors agreed to have their extended abstracts and slides posted publicly. These appear at `http://www.DavidDLewis.com/events/otc2001`.

3 Presentations

We briefly summarize here the presentations from the workshop, giving our personal interpretations of these detail-rich talks. As requested by the speakers, we do not include certain details, such as personnel costs, that came up in the discussions.

Two talks, by Richard Tong (Tarragon Consulting) and Mark Krellenstein (Northern Light Technology) were canceled due to travel interruptions after the September 11 terrorist attacks in the US. A third talk, by Mark Shewhart (LexisNexis), was presented by his colleague Mark Wasson. The schedule of presentations was compressed to allow for the rapidly changing travel plans of participants, with a single extended discussion session after the presentations.

Mark Wasson (LexisNexis) presented two talks on TTI (Term-based Topic Identification System), a text categorization system which has been in operation at LexisNexis for many years. It classifies tens of thousands of news articles, legal cases, Web sites, company reports and other documents on a daily basis. Over 70,000 categories are used, including both subject categories and named entities. Categories can be included in Boolean searches by end users. They are also used to create "specialized libraries", allowing a user search to run against a virtual collection of material on a specific topic. The set of categories has evolved over time, and customer requests to supply particular categorizations (which may or may not fit well in the existing taxonomy) present ongoing challenges. Demands for accuracy and consistency are high, as the user base includes many information professionals.

Classifiers are built manually in an iterative fashion. Pure supervised machine learning approaches were not felt to be effective across the wide range of source types to be handled. However, supervised learning in the form of chi-square tests and stepwise linear regression is used to aid manual selection and weighting of words and groups of words. Similarly, while glossaries, name variant generation, and other "knowledge-based" approaches are used, the outputs of all of these are reviewed and often overridden based on human expertise. Techniques that would limit the classifier builders' flexibility (punctuation and stop word removal, stemming, etc.) are avoided. Building of a classifier may require anywhere from five minutes (for, say, a simple company name detector) to eight hours (for categories corresponding to complex topics) of staff time.

The diversity of material to be classified is partly dealt with by source specification files. These map from the fields of each document to a canonical fields that are referenced by rules (e.g. mapping *Title* from one source and *Heading* from another to a canonical title field). Even so, properties of specific sources must sometimes be manually accounted for in classifiers.

Mark reported precision and recall rates higher than 90–95% for most topics and sources, with classification speeds on the order of 100,000 characters/sec. Precision, recall, and similar measures are used in quality assurance and management, and considerable resources are devoted to benchmarking on real data. Effectiveness data from the research literature or text classification vendors have not always been found to be good predictors of effectiveness on LexisNexis data.

Max Copperman and **Scott Waterman** (Kanisa) reported on their experience with using text categorization to support interactive, self-service help systems. Documents consist of FAQs, troubleshooting guides, product information, and the like, and need to be classified under a variety of dimensions: this means that multiple taxonomies coexist (up to 150 taxonomies, with up to 2,000 categories per taxonomy), addressing orthogonal concerns (e.g. a Product taxonomy, a Symptoms taxonomy, etc.). User queries are also classified under the same taxonomies, and matching on categories drives the response to each query.

The large number of closely related categories (e.g. Creating Forms vs. Editing Form Fields) and shortage of data on many categories (as well as labeling costs) argued against a pure supervised learning approach. At least one commercial classification tool was found wanting in internal tests. Kanisa's current approach relies heavily on manual definition and choice of features, followed by some use of training data to set weights. As with LexisNexis, the classifier environment allows manual overriding of defaults for issues such as stemming and phrase matching. Also as with LexisNexis, source definitions are used to control which parts of documents are used in classification.

Kanisa found that improving classifier effectiveness did not always lead to better help dialogues, which is the goal of their system. Testing therefore emphasizes dialogue quality as evaluated by customers and by semi-automated full system testing.

Thomas Montgomery (Ford Motor Company) reported progress toward a system for classifying patents to support competitive intelligence. Currently, manual classification of patents into technology categories is done to support detection of trends and levels of activity by competitors. Automating this classification promises to speed up the availability of such information, allow greater coverage, and allow reclassification as categories change to reflect new technologies.

Commercial text classification tools, research software, and internally built tools were all evaluated for this application. Tests were carried out using a sample of 136,063 manually labeled patents produced by the current manual process. A taxonomy of 4,000 categories were used, with documents allowed to belong to multiple categories. Preliminary results on high level categories showed support vectors machines (SVMs) outperforming a nearest neighbor approach (k-NN), and that in turn outperforming a naive Bayes approach. Evaluation is ongoing. Lower level categories were harder to assigned than higher level ones, and some collapsing together of lower level categories with too few training examples was necessary. Tom indicated that the choice of a final approach will depend not just on effectiveness, but on issues such as vendor size and stability, maintenance effort, cost, etc.

Bryan Goodman (Ford Motor Company) reported work to develop a personalized information delivery systems. Here, there was no existing classified data or taxonomy, so the two were created together. A taxonomy of 57 Ford-specific categories (plus "Other") was created, guided by existing taxonomies from other applications as well as self-organizing map (SOM) software. A total of 2600 papers were labeled, with bootstrapping from initially labeled data used to aid the labeling (and correction of labeling) of further documents. As with the project reported by Tom Montgomery, a variety of machine learning approaches were tested. The vast majority of work on the project involved writing software to connect tools together, reformat data for each piece of software, etc. Hope was expressed for standardization as text classification software matures.

Tamas Doszkocs (National Library of Medicine) reported on preliminary work on text classification to improve search of a document base on hazardous substances. Shallow parsing and matching of strings against UMLS (Unified Medical Language System) have been investigated, and a test collection has been built to support future studies of machine learning approaches.

Khalid Al-Kofahi (Thomson Legal & Regulatory) described a "case routing program" (CARP) for legal publishing in use at Thomson, which classifies a weekly feed of case law summaries (CLSs) under a proprietary classification scheme, called American Law Reports (ALR). ALR is actually a set of about 13,800 *articles* ("annotations") that provide in-depth analysis of narrow points of law. The task of CARP is to suggest court cases that should be cited by each ALR article.

CARP has been operational since January 2001. It processes about 12,000 headnotes per week (each case has on the order of five headnotes associated with it), and makes on the order of 1,600 suggestions to link particular headnotes with ALR articles. Roughly 900 of these suggestions are accepted, 170 rejected as inappropriate, and 530 considered on topic but redundant or too general. This compares with the previous purely manual process which produced about 700 new citations per week. Supervised learning was used to train classifiers, which take the form of a weighted committee of individual classifiers. Inputs to classifiers include words, phrases, and previously assigned categories from other taxonomies. (Khalid mentioned that over 200 separate taxonomies are in use for different purposes in Thomson.)

Finally, **Ronen Feldman** (Clearforest) gave a demonstration of the ClearCat categorization system. The system comprises a suite of tools, including different learning (SVM, perceptron, Rocchio, naive Bayes, kNN), and indexing (e.g. bag of words, noun phrases, etc.) methods, with options for manual feature selection.

4 Discussion

A wide-ranging discussion followed the presentations, lasting about four hours. There was a particular emphasis on issues that were common among the talks, but which have received little attention in the research literature.

For instance, while training on labeled data is widely used, it is rarely used in the "data in / classifier out" model common in experimental work. Manual feature construction and feature selection before training, and/or modification of classifiers after the fact, appear common in operational systems. Reasons include both the value provided by human domain knowledge (particularly in constructing features that would not be found by a learning systems), as well as problems of lack of data to label, cost of labeling, and inaccuracies in labeling. The important role of domain knowledge sometimes leads to learning approaches or classifier forms that would not be optimal in a pure supervised learning setting, but which make manual intervention easier. For software vendors, being able to quickly provide manual input during a demo was another argument for this.

Effectiveness measures such as recall and precision are sometimes used. Customers on first learning of them often have unrealistic expectations (e.g. 100% precision and recall). Education about the limits of technology and the subjectiveness of classification, as well as looking into the actual needs and context of an application, helps.

However, these measures do not capture everything one would want. In particular, the notion of degrees of correctness of category assignments came up several times. Some mistakes are worse than others in terms of end user perception. Several attendees mentioned tuning of thresholds to avoid blatant false positives, especially in systems which do not involve human checking of automated output. The question is not merely one of avoiding false positives, however. Assigning a category which is wrong, but closely related to a correct category, is viewed as less bad by users than assigning a completely unrelated one. This raised the question of whether we need new effectiveness measures that take such relationships into account.

Some techniques that have excited interest or controversy in the research community were discussed. Multiword phrases were mentioned several times, but often in the context of manual feature construction, so their role in a purely automated setting is unclear. Using hierarchical relationships among categories was found useful in tests by some groups and useless by others. None of the talks discussed using links between hypertext documents to aid classification, and there was speculation on the difference in linkage structure between the Web and, for instance, intranets.

There was, not surprisingly given the workshop topic, considerable enthusiasm for automated classification. Many participants felt that automated systems, or automation plus human checking, can reduce both costs and increase consistency.

5 Future activities

Participants overwhelmingly indicated that the workshop was successful, and discussions were lively. Many expressed the theme that insights from the operational world will shake up the implicit assumptions of text classification research, with benefits to both.

There was strong support for a follow-up workshop. Not surprisingly given the setting, a straw poll favored SIGIR-02 in Tampere as the location, but KDD-02 in Edmonton was also discussed. David Lewis also announced his intention to edit a book-length collection on operational text classification, with papers to be solicited widely. Look for announcements of these and other activities on the DDLBETA text categorization mailing list and other forums.