Report on the CONALD Workshop on *Learning from Text and the Web*

Jaime Carbonell, Mark Craven, Steve Fienberg, Tom Mitchell and Yiming Yang

An increasing fraction of the world's information and data is now represented in textual form. For example, the World Wide Web, online news feeds, and other Internet sources contain a tremendous volume of textual information. The goal of the CONALD workshop on *Learning from Text and the Web* was to explore computer methods for automatically extracting, clustering and classifying information from text and hypertext sources.

The workshop included ten oral paper presentations, an organized discussion by a panel of distinguished researchers, and a handful of other contributed papers. The workshop provided a good survey of the state of the art in machine learning methods applied to text processing tasks. The presented work involved a wide array of learning approaches, including finite-state-machine induction [HD, MMK], neural networks that can accept *advice* from users [SER], relational learning methods [Moo, SC], statistical clustering algorithms [GS, Hof, LV, YPC], boosting methods [ADW], algorithms for learning with hierarchical classes [Hof, MG], and active learning methods [LT, NM]. A principal limitation of many of these approaches is that they do not directly reflect attempts to develop formal models of the text phenomenon of interest.

The research presented at the workshop also spanned a broad range of application tasks, including: information extraction [HF, HD, LS, Moo, MMK], information finding [SER], information integration from Web sources [MMK], automatic citation indexing [BLG, KP], event detection in text streams [YPC], document routing [ADW] and classification [GWI, Moo], organization and presentation of documents in information retrieval systems [GS, Hof], collaborative filtering [dVN], lexicon learning [GBGH], query reformulation [KK], text generation [Rad] and analysis of the statistical properties of text [MA]. In short, the state of the art in learning from text and the web is that a broad range of methods are currently being applied to many important and interesting tasks. There remain numerous open research questions, however.

Broadly, the goals of the work presented at the workshop fall into two overlapping categories: (i) making textual information available in a structured format so that it can be used for complex queries and problem solving, and (ii) assisting users in finding, organizing and managing information represented in text sources. As an example of research aimed at the former goal, Muslea, Minton and Knoblock [MMK] have developed an approach to learning *wrappers* for semi-structured Web sources, such as restaurant directories. Their method is able to induce extraction rules from small numbers of labeled examples. These learned extraction rules are then applied so that Web pages can be treated like structured databases. As an example of work geared toward the latter goal, Shavlik and Eliassi-Rad [SER] have developed an approach to increasing the communication bandwidth between users and learning agents that perform tasks such as home-page finding. Their approach enables a user to give advice to a learning agent at any time during the agent's lifetime. The advice is incorporated into the agent's learned model where it may be subsequently refined by reinforcement-learning methods.

The likely future impact of research in text learning is twofold. First, much of the information that is currently available only in text form will automatically be mapped into a structured format. This ability would mean that the Web queries would not be limited to keyword searches for individual documents relevant to the query. Instead, we could directly get answers to queries whose answers are distributed across multiple Web sources. Moreover, this ability would mean that text sources, such as the Web, could be used for planning and problem solving (e.g. an agent that would use information on the Web to make travel plans for IJCAI-99). The second likely impact of continued work on learning from text and the Web is greatly improved methods for finding, organizing, and presenting information in free-text data sources including Web pages, emails, transcribed radio/TV broadcasts and newswire stories.

The workshop brought to light numerous promising research topics and raised many open research questions for further exploration:

- *Learning from structure in hypertext.* There are many opportunities to go beyond bag-of-words representations documents by exploiting HTML formatting and patterns of connectivity in hypertext.

- *Developing statistical models that represent hypertext structure.* In addition to developing algorithms that are able to exploit document structure, it is also important to develop to develop well-founded statistical models for such tasks. For example, we might develop stochastic models for graphs of linked objects by exploiting ideas from the literature on *social networks*.

- *Exploiting NLP techniques more in text learning.* Much of the current work in text learning concentrates on word occurrence statistics and ignores other linguistic structure. There is much work to be done in understanding how natural language processing methods can be applied to gain more accurate learners.

- *Learning from temporal patterns in text.* Some text processing tasks, such as topic tracking and event detection, involve a stream of text data over time. New methods are needed to represent, detect and exploit content (and Web structure) that changes over time.

- *Learning from available domain knowledge, advice and data.* In addition to learning from training data, text-learning systems should be able to take advantage of background knowledge and on-line advice offered by users.

- *Systems that learn to improve the organization and presentation of information.* As an example of this type of system, Perkowitz and Etzioni [PE97] have begun developing adaptive Web sites. Other related issues include learning to recognize certain types of queries, and learning from passive relevance information.

- *Developing statistical models for interaction data.* In addition to developing effective methods for systems such as adaptive Web sites, it is also important to develop statistical models of non-stationary user interaction data.

- *Learning text classifiers when word statistics are sparse.* Current research in this area is exploring such methods as term clustering, feature reduction, document clustering, active learning, and using hierarchical class information.

- *Combining evidence from multiple sources.* This issue is relevant to information extraction, information retrieval, and text classification. In information retrieval, for example, we may want to combine document rankings that consider such factors as document-query similarity, document popularity, and editorial vetting of documents.

Current research in learning from text and the Web is already quite cross-disciplinary. As the above list of promising research topics indicates, however, the hard problems in the area call out for expertise in a wide variety of disciplines including machine learning, statistics, information retrieval, natural language processing, planning, and human-computer interaction.

A challenging question in the inter-disciplinary research is, can we significantly improve the state of the art by introducing more principled formal models about text phenomena of interest? Furthermore, how can we determine the suitability of such models? Empirical evaluation (e.g. cross-validation) **alone** may not be sufficient for analyzing the behavior and limitations of algorithms. On the other hand, comparing models without evaluating their practical impact in specific tasks is clearly inadequate. This workshop presented a promising trend of research that addresses both theoretical and empirical concerns: Hofmann and Lafferty and Venable both proposed new statistical language models for document clustering. Apte presented strong empirical evidence for using boosting to improve the state of the art in text categorization. These exciting research findings encourage further investigation for more satisfactory answers in the future.

References

[ADW] C. Apte, F. Damerau, and S. Weiss. Text mining with decision rules and decision trees.

[BLG] K. D. Bollacker, S. Lawrence, and C. L. Giles. CiteSeer: An autonomous system for processing and organizing scientific literature on the Web.

[dVN] O. de Vel and S. Nesbitt. A collaborative filtering agent system for dynamic virtual communities on the Web.

[GBGH] A. F. Gelbukh, I. A. Bolshakov, and S. N. Galicia-Haro. Automatic learning of a syntactical government patterns dictionary from Web-retrieved texts.

[GS] M. Goldszmidt and M. Sahami. A probabilistic approach to full-text document clustering.

[GWI] B. Gelfand, M. Wulfekuhler, and W. F. Punch III. Automated concept extraction from plain text.

[HD] C.-N. Hsu and M.-T. Dung. Wrapping semistructured Web pages with finite-state transducers.

[HF] Hull and Fluder. Text mining the Web: Extracting chemical compound names.

[Hof] T. Hofmann. Learning and representing topic.

[KK] Y. S. Kwon and N. H. Kim. The effect of relevant input information in ID3's learning performance.

[KP] A. Kehagias and V. Petridis. Automated building of a database of neural network papers.

[LS] Z. Lacroix and A. Sahuguet. Information extraction and heuristics for human-like browsing.

[LT] R. Liere and P. Tadepalli. Active learning with committees: Preliminary results in comparing winnow and perceptron in text categorization.

[LV] J. Lafferty and P. Venable. Simultaneous word and document clustering.

[MA] P. P. Makagonov and M. A. Alexandrov. Tool for measurement of statistical properties of text.

[MG] D. Mladenic and M. Grobelnik. Feature selection for classification based on text hierarchy.

[MMK] I. Muslea, S. Minton, and C. Knoblock. Wrapper induction for semistructured Web-based information sources.

[Moo] R. J. Mooney. Learning for information extraction, querying, and recommending.

[NM] K. Nigam and A. McCallum. Pool-based active learning for text classifcation.

[PE97] M. Perkowitz and O. Etzioni. Adaptive web sites: an ai challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, Nagoya, Japan, 1997. Morgan Kaufmann.

[Rad] D. R. Radev. Learning correlations between linguistic indicators and semantic constrints: Reuse of context-dependent descriptions of entities.

[SC] S. Slattery and M. Craven. Learning to exploit document relationships and structure: The case for relational learning on the Web.

[SER] J. Shavlik and T. Eliassi-Rad. Building intelligent agents for Web-based tasks: A theory-refinement approach.

[YPC] Y. Yang, T. Pierce, and J. Carbonell. Event detection.