Interactive Visualization for Computational Linguistics

Christopher Collins and Gerald Penn
Department of Computer Science
University of Toronto
10 King's College Road
Toronto, Ontario, Canada
{ccollins,gpenn}@cs.utoronto.ca

Sheelagh Carpendale
Department of Computer Science
University of Calgary
2500 University Dr. NW
Calgary, Canada
sheelagh@ucalgary.ca

Interactive information visualization is an emerging and powerful research technique that can be used to understand models of language and their abstract representations. Much of what computational linguists fall back upon to improve NLP applications and to model language "understanding" is structure that has, at best, only an indirect attestation in observable data. An important part of our research progress thus depends on our ability to fully investigate, explain, and explore these structures, both empirically and relative to accepted linguistic theory. The sheer complexity of these abstract structures, and the observable patterns on which they are based, usually limits their accessibility — often even to the researchers creating or attempting to learn them.

To aid in this understanding, visual 'externalizations' are used for presentation and explanation — traditional statistical graphs and custom-designed illustrations fill the pages of ACL papers. These visualizations provide *post hoc* insight into the representations and algorithms designed by researchers, but visualization can also assist in the process of research itself. There are special statistical methods, falling under the rubric of "exploratory data analysis," and visualization techniques just for this purpose, in fact, but these are not widely used or even known in CL. These techniques offer the potential for revealing structure and detail in data, before anyone else has noticed them.

When observing natural language engineers at work, we also notice that, even without a formal visualization background, they often create sketches to aid in their understanding and communication of complex structures. These are *ad hoc* visualizations, but they, too, can be extended by taking advantage of current information visualization research.

This tutorial will enable members of the ACL community to leverage information visualization theory into exploratory data analysis, algorithm design, and data presentation techniques for their own research. We draw on fundamental studies in cognitive psychology to introduce 'visual variables' — visual dimensions on which data can be encoded. We also discuss the use of interaction and animation to enhance the usability and usefulness of visualizations.

Topics covered in this tutorial include a review of information visualization techniques that are applicable to CL, pointers to existing visualization tools and programming toolkits, and new directions in visualizing CL data and results. We also discuss the challenges of evaluating visualizations, noting differences from the evaluation methods traditionally used in CL, and discuss some heuristic approaches and techniques used for measuring insight. Information visualizations in CL research can also be measured by the impact they have on algorithm and data structure design.

Information visualization is also filled with opportunities to make more creative visualizations that benefit from the CL community's deeper collective understanding of natural language. Given that most visualizations of language are created by researchers with little or no linguistic expertise, we'll cover some open and very ripe possibilities for improving the state of the art in text-based visualizations.

Tutorial Abstracts of ACL-08: HLT, page 6,
Columbus, Ohio, USA, June 2008.