

RESEARCH STATEMENT

Christopher M. Collins

Textual data is at the forefront of information management problems today. Thousands of pages of text data, in many languages, are produced daily: emails, news reports, blog posts, product reviews, discussion forums, academic articles, and business reports. Computational linguistics interventions have also increased, as we rely more and more on automated language translation, summarization, enhanced information retrieval, and opinion mining. Managing, exploring, and analysing the flow of linguistic data is becoming both an individual and a societal problem. Data scale issues are not only a challenge for end users of technology, but are equally challenging for natural language engineers as they develop the systems that allow computers to manipulate and analyze text.

My research focus in human-computer interaction uses techniques from information visualization and my background in computational linguistics to create interactive visualizations of language. These visualizations help both computational linguistics researchers and everyday computer users to manage and analyze linguistic data. I approach visual linguistic analysis as a problem of computer-aided cognition — visualizations produced by the high speed and high volume of data analysis achieved by computers complement human linguistic sophistication and decision-making ability. I use a holistic set of methods, including ethnography and user-centered design, as well as drawing on cognitive science findings about human perception and results of laboratory-based studies. My dissertation research focuses on visualizing the processes and outcomes of natural language processing, supporting large scale text analytics, and providing novel information visualization techniques for multi-dimensional data, including linguistic data. I have identified several exciting next steps in each of these areas. Additionally, I see an opportunity to use the synergy between analysis of language and analysis of other types of data to enable new discoveries about language change, language use, and how language affects society. While my dissertation results and mid-term research agenda focus on linguistic data and problems, outcomes of my work have and will continue to include creative and generalizable human-computer interaction and visualization contributions.

DISSERTATION RESEARCH

VISUALIZING NATURAL LANGUAGE PROCESSING

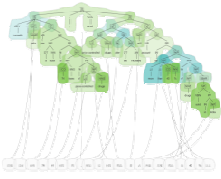


Translation uncertainty lattice visualization with an un-translatable word shown as a photo node.

Traditional natural language processing (NLP) produces a single output, such as a translated sentence. However, these outputs are often associated with well-defined measures of uncertainty, both inherent in the data, and introduced by statistical NLP. To open the black box of NLP, we presented the lattice uncertainty visualization [C.3]¹ at EuroVis 2007. This visualization provides a quick and intuitive visualization of several translation hypotheses and their estimated uncertainty. Embedded in an instant messaging chat client, the visualization enables cross-language conversations, helping readers judge how much to trust the translation accuracy. The technique was also applied to the outputs of the speech recognition system developed in my M.Sc. research [C.1].

Understanding the outcomes of natural language processes is not only important for end users of natural language systems, but also for the researchers who develop them. In yet-unpublished work, I undertook an ethnographic study of natural language engineers at work. During this study, I observed

¹ See CV for self references.



Chinese-to-English translation parse tree with overlaid set relations as concave 2D iso-surfaces.

ad hoc use of visualization (both sketches and computer-based static information graphics) to analyze the quality of their translation system. I discovered that the team of researchers had become overwhelmed with the task of evaluating their system and diagnosing translation quality issues. In response, I created a visualization of their data structures (tree structures overlaid with set relations) which supports annotation and collaboration for analysis and sharing of results. I am in the process of deploying and evaluating this visualization, which is general enough to apply to other sorts of data such as social network graphs and scatter plots with associated set relations.

In a related work, our investigation of word similarity measures led to the contribution of VisLink, a new method for revealing relationships amongst 2D visualizations by linking them in a 3D space. VisLink was applied to reveal patterns of similarity and difference amongst various lexical similarity measures of interest to NLP researchers, but can be used to connect any 2D visualizations. The system, presented in our InfoVis 2007 paper [J.3], has been praised as an early example of the future of integrative visualization, including by Georges Grinstein in his IV 2008 keynote address [1].

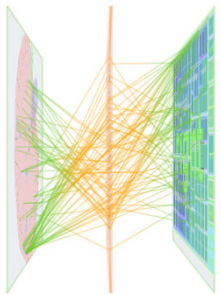
Future Directions

There is a lot of potential for improving the effectiveness of the outputs of NLP systems through application of process visualization. Areas of importance include investigating real-time visualization of automata used in tasks such as dialogue system construction and visualizing the process of chart pruning and beam search for parsing and other applications. At ACL in 2008, we enumerated several of these possibilities as we engaged the research community in a tutorial on the potential for interactive visualization to aid NLP research [TW.1]. The exciting discussion raised several possible future research areas. Interactive visualization is well-suited to exploratory data analysis, a type of task which is increasingly common as NLP moves towards greater reliance on statistical techniques. Exploratory tasks include ensuring the quality and balanced coverage of the corpora used to train NLP systems. Interactive visual exploration of parameter spaces, e.g., “What changes when I adjust this parameter?” could build on our work with machine translation researchers to provide an even tighter feedback loop in NLP research.

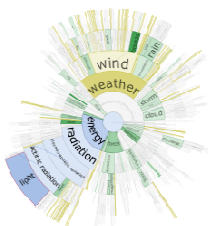
While the VisLink system holds promise for additional linguistic analyses, it also provides a test bed for investigation into long-held visualization challenges such as 3D interaction, data ordering, and visual occlusion. We have been invited to apply VisLink in collaborations by researchers in aerospace (Boeing), chemistry (Dow Chemical), and forensic analysis (Royal Canadian Mounted Police).

VISUAL TEXT ANALYTICS

One of the most fundamental analytic tasks for text is to provide an overview of a long document — the electronic version of thumbing through a book. Existing visualizations, such as tag clouds, provide an unstructured overview, and do not make use of the wealth of expert-created knowledge stored in ontologies such as WordNet. We designed DocuBurst, the first visualization of document content which combines word frequency with the human-created structure in lexical databases, to create a simple visualization that reflects semantic content [TR.1, O.1, P.1-3, *full paper in submission*]. DocuBurst is a radial, space-filling layout of hyponymy (the IS-A relation), overlaid with occurrence counts of words in a document to provide visual summaries at varying levels of granularity. Interactive analysis is supported with geometric and semantic zoom, selection of individual words, and linked access to the source text. Our radial space filling layout and interaction code has been downloaded 75 times and is in active use in several research labs. DocuBurst has been featured in mainstream newspapers, television, and radio. Deployment inquiries include patent databases (Dow Chemical)

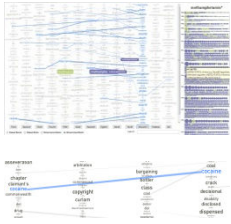


VisLink relating various lexical visualizations (top) and heterogeneous data about US elections (bottom).



DocuBurst of terms under ‘physics’ in a general science textbook.

and governmental repositories (NATO). In addition, many educators have seen its potential as a teaching aid, and are eager to use it in the classroom. I am investigating options for wide distribution.



Visualization of lexical differences amongst 14 US courts (top).

Enlargement of several words with connections (bottom).

Moving beyond a single text, while working with Martin Wattenberg and Fernanda Viégas in the summer of 2008, I developed a system which supports exploratory real-time visual analytics on very large numbers of long documents. Specifically, our visualization enables detection and display of changing themes amongst the 14 upper courts of the USA, over a period of more than 200 years, covering a total of over 600,000 written court decisions. Linked access to the source text allows for drill down by investigators interested in discovering more about a term or time period. Trials with legal scholars resulted in interesting discoveries about the geographically distributed themes (e.g. miner’s lung disease in the 4th Circuit) and fascinating legal jargon and language use differences across states and even judges (e.g. the 7th Circuit uses the “ostrich instruction” to jurors abnormally often).

Future Directions

While advanced computational techniques such as latent semantic analysis are powerful tools for data experts, they produce results that can be unintuitive for the average reader. Simple analysis techniques such as word-counting and lexical relationship mapping require little training to understand, however, we have only scratched the surface of how to present the results of these analyses. A lot of potential remains for better visualization, increased capacity for large and changing datasets, support for different viewing platforms, and large-scale collaboration. The Web offers a programming and visualization platform, exciting data sources, and an eager user base for continued research. Imagine a visualization platform for collaborative visual analytics of government documents to increase transparency, or a method for tracking and understanding how linguistic ‘memes’ propagate online.

The long documents we have worked with are luxury items, each providing a wealth of data. An exciting text analytics challenge is to develop data analysis and visualization techniques for short texts — ‘tweets’ on Twitter, FaceBook status updates, and consumer-submitted product reviews comprise a significant portion of computer-mediated communication, especially amongst younger generations. Visualizations of these data sources generally focus on keyword spotting and word counting. Closer integration of analysis techniques such as sentiment analysis and word likelihood scoring could have a dramatic impact on how we understand the flurry of short texts produced all day, every day, online.

THE NEXT STEP: LEVERAGING LINGUISTIC DATA FOR NON-LINGUISTIC INSIGHT

Languages are always changing — over time new words emerge, spelling and pronunciation shift, fashionable slang explodes and fizzles, and older words lose favour. Sociolinguists and lexical etymologists have been studying how written language is influenced by social structures for many years [e.g., 2]. The massive scale of digital data available on the Internet introduces new computer science research challenges for these types of analyses. How do we analyse ever-changing linguistic style over time? How do we link computational linguistics to the closely intertwined social structures (friends, colleagues, geography) backing the use of language online? And, given two or more heterogeneous and mutually influential types of data that are changing through time, what types of visual metaphors and graphical techniques will best support teasing apart the networks of influence? Can we then use these models to predict linguistic influence? Through developing new linguistic algorithms and visualization techniques to address these computational challenges, we can also gain insight into the sociological aspects of language change in ways not previously possible.

Explicit online connections such as “friending” patterns and bookmark sharing activities have been investigated using methods such as information visualization [3] and ethnography [4]. These

investigations rely on surface relationships of who communicates with whom. However, less is known about how the semantic content and linguistic style within online artefacts relates to social networks.

This type of research must be carried out with care and attention to the sensitive and private nature of social network information. To address this, I propose three phases of investigation, each using publicly available data. First, we will analyze the transcripts of the Canadian Parliament by speaker, party, and over time, to capture the emergence of new topics of debate and changes in style and phrasing. These linguistic phenomena will then be visually related back to the explicit social structures of political parties and positions of influence. Second, we will investigate the written decisions of the Supreme Court. New visualizations combining the social network of co-authorship and linguistic analysis of decisions will answer questions such as when does a particular legal theme or linguistic oddity first appear, and how does it propagate through the set of judges over time. Pushing the data size an order of magnitude larger, the third stage will examine the social networks defined by co-citation (cross-linked blogs or citations within academic papers) and the contents of the artefacts of these networks (the blog posts or papers) to discover patterns of linguistic influence. In the long term, this research can contribute visualization techniques for relating multiple heterogeneous but mutually influencing types of data that are changing over time.

As I begin an academic career, I believe this research agenda is well-suited to attracting quality graduate students. I have witnessed the excitement of visualizing linguistic data while giving many demonstrations of my various research projects over the past years, and in working with undergraduate and Master's students through collaboration with my thesis supervisor. There are many opportunities for students of computer science and design. New and general techniques will need to be developed, such as closer coupling of natural language processing and visualization, appropriate interaction techniques and/or devices, creative and justifiable visual design for multi-dimensional and large scale data, highly efficient data processing and rendering for large amounts of text, solutions for constraints imposed by the need for clear text legibility despite the relatively low resolution of digital displays, and methods to support collaborative analysis.

As with most investigations in information visualization, this research offers the opportunity to learn something new about the data, in this case about language and how we use it. The workplace relevance of linguistic data sources such as blogs, emails, and other computer-mediated communication, as well as the societal relevance of data such as parliamentary and court records will certainly enhance my ability to attract the funding necessary to conduct the research. I personally have been awarded grants based on research merit, including NSERC (the Canadian federal science funding agency) scholarships and the IBM Scholarship, and I have been involved as a reviewer of successful grant applications for my supervisors. My past research has attracted the attention of potential industry, academic, and governmental partners, and I have developed an international network of potential academic collaborators through my involvement in various conferences and visits to research labs. I am confident that I have the motivation, skill, and eagerness to continue to learn that is necessary to carry this out, and that it is the right time to embark on a research program to enhance our ability to interact with and understand language in the digital age.

References

- [1] Grinstein, George. (2008) Important Future Research Areas for Information Visualization. Keynote Lecture, 12th Int. Conf. on Information Visualisation, London, England.
- [2] Mayes, P. (2003) Language, Social Structure, and Culture. John Benjamins Publishing, Amsterdam.
- [3] Heer, J.; Boyd, D. (2005) Vizster: Visualizing Social Networks. In Proc. IEEE InfoVis., pp 5–13.
- [4] Lampe, C.; Ellison, N.; Steinfeld, C. (2006) A face(book) in the crowd. In Proc. CSCW, pp 167–170.