NLP overview

Lillian Lee
Cornell University
http://www.cs.cornell.edu/home/llee

NSF workshop on Automated Content Analysis and the Law, 2009

Lots of NLP/info. retrieval work ($\approx$ 3000 researchers)

Lots of different interests among this audience

Goal: connect techniques from my community with interests here, and perhaps broaden the landscape of possible collaborations.

Method: mention lots of different directions, and hope some of them catch your eye.

Caveat: in practice, there aren't out-of-the-box solutions. Defining the problems, data, and algorithms demands strong collaboration among all parties.

Some of "our" forays into political/legal text

⋆: techniques perhaps "most different/surprising" to this audience

- Automatic ideology/perspective/bias classification for discussion-board posts, websites, newsfeeds [Efron '04, Grefenstette '04, Lin and Hauptmann '06⋆, Fortuna et al '07⋆, Lin, Xing and Hauptmann '08⋆, Malouf and Mullen '08]
- TREC Legal Track annual competition: e-discovery [http://trec-legal.umiacs.umd.edu/]
- eRulemaking: analyzing or facilitating citizen comments about proposed legislation [Shulman et al. '06, Cardie et al. '06,]
- Election prediction based on individuals' predictions [Kim and Hovy '07]
- Support/opposition classification within Congressional floor debates, or Supreme-Court turn-taking [Thomas, Pang, and Lee '06⋆, Hawes, Lin and Resnik '09⋆]

Some potentially relevant research

No magic bullets, but there are plenty of bullets.

- real-valued text classification using regression, metric labeling$\star$, SVM-separator-distance$\star$, etc. (*ideology classification?)*; labels with hierarchical or other complex relationships$\star$(*topic hierarchies?)*
- document similarity/clustering$\star$(*finding document relationships in a multidimensional space?)*
- burst detection, correlated topic models, and other topic/word detection and tracking$\star$(*changes in word usage over time?)*

- various IR techniques for finding relevant documents (link analysis and other metadata usage$\star$, etc.) (*e-discovery?)*

Intra-document analysis

- Sentiment analysis/opinion [ha] mining: analysis of subjective sentences, documents, or document collections$\star$ (*influence of public opinion, opinion dynamics, etc.?)*
- Textual entailment$\star$: is this text unit implied by that one? (*identifying conflicts/agreement in regulations or arguments?)*
- Citation analysis (a special case of *information extraction*) (*what citations represent support, competing explanations, or background information?)*
- Summarization (*What's the common content in this blog post or this collection of briefs?)*