Siddharth Jonathan J.B.

Overview of Research at Stanford

Web Sociologists Workbench:

The Communications Department at Stanford wanted to analyze online media coverage of popular events/happenings like the California Special Election, Hurricane Katrina etc. They wanted to analyze Web crawls of targeted news sites to calculate aggregate statistics to assess the focus of the media coverage of these events. The Web Sociologists go through the crawl manually and classify it into a set of pre-specified categories. As part of the Stanford InfoLab, I worked with Dr. Paepcke and Prof. Hector Garcia Molina on automating this process. My research was in the following areas,

- Identification of relevant (to the topic under study eg. The California special election) and irrelevant documents in the crawl using a suite of Naïve Bayes Classifiers.
- Classifying Web pages into one of 5 pre-specified topic categories using an SVM.
- Identifying near duplicate documents in the Web crawl.

InfoMem:

My research here with Dr. Fruchter of the Project Based Learning Lab was to design and implement a Search Engine for the enterprise. This text based Search Engine was designed to model hierarchical containment relationships and leveraged this information for ranking results. It allows search at multiple granularities and also includes a search visualization module that makes us of the Tree Map metaphor for displaying results. As a beta test, it was deployed to make the data in a live discussion forum searchable.

Yahoo! Search Relevance Group Research - Region/Language Mixing:

In International Web Search, for some queries, it is better to return local region/language results (eg. Weather, commercial queries) while for others there is no need to favor local region/language results (eg. information seeking queries). The way to blend the results appropriately to improve user satisfaction varies by query category and also based on whether the local language is English or not. My Internship research focused on coming up with ways to solve this problem and using this information as a feature for the machine learned ranking. This involved,

- Click log analysis
- Leveraging Yahoo!'s click data from various other Yahoo! properties
- Implementing an SVM based Query language classifier for German/English based on features like bigram characters, relative frequency in the query logs etc.
- Using a Commercial Query Classifier trained on commercial queries
- Incorporating the Region/Language information as features for Machine Learned Ranking for the scoring function.
- Analysis of the correlation between Click through rate and DCG (discounted cumulative gain)

Context Driven Ranking:

I worked on Context Driven Ranking both during my Independent Research Study with Dr.Andreas Paepcke & Prof. Hector Garcia Molina as well as during my Internship with Yahoo! Inc. Here 'context' is defined to be words that describe or are associated with a query word. For eg. For 'motorcycle' the context would be 'two wheeler', 'bike', 'ride', 'oil' etc.
The task of Context Driven Ranking is twofold,

- Automatically generate context for a query
- Weight it appropriately for use in the ranking to improve precision and recall

Some of the techniques that I have worked on to generate and incorporate contextual information for a query are,

- Use the postings list as a training set and view the documents as being generated from a generative model.
- Use chi-square feature selection to come up with context words
- Use selective boosting of query and high scoring chi-square words
- Used word net based context words (courtesy Fuchun Peng, Yahoo! Inc.)
- Used context words derived from the incoming anchor text links. (research at Yahoo! Search Relevance)

SQUINT – An SVM based approach for identifying relevant sections in a Web Page for a Query (CS229 Course Project):

SQUINT works by generating features from the top most relevant results returned in response to a query from a Web Search Engine, to learn more about the query and its context. SQUINT uses a supervised learning model to score sections of a Web page based on these features. These scores can be used for many applications. Some applications include, some form of highlighting of the sections to indicate which section is most likely to be interesting to the user given his query. If the result page has a lot of (possibly diverse) content sections, this could be very useful to the user in terms of reducing his time to get the information he needs. Another advantage of this scheme as compared to simple search term highlighting is that, it would even score sections which do not mention the key word at all. We also think SQUINT could be used to generate better summaries for queries in Web Search. One can also envision SQUINT as being able to create succinct summaries of pages of results, by pulling out the most relevant section in each page and creating a meta summary page of the results. The training set for SQUINT is generated by querying a Web Search Engine and manual labeling of sections.

Statistical Machine Translation System for Turkish (CS224N Course Project):

There has been minimal research in MT systems where Turkish is either the source or target language. This project attempts to fill this gap, if only slightly. As part of my NLP Course Project, my project partner and I designed and implemented a suite of tools intended for aiding the statistical MT creation process by exploiting linguistic features of Turkish and how

these features differ from their counterparts in English. The tools we designed were then coupled with existing software to produce an English-to-Turkish MT system that would perform better than the baseline system constructed without concern for the idiosyncrasies of the specific task at hand.