

COMPUTATIONAL LINGUISTICS IN INDUSTRY*Interesting things gleaned from the Lecture Series*

This paper summarizes certain characteristics mentioned by industry speakers working either directly or tangentially in computational linguistics. It fulfills the required coursework for LSA 306: Computational Linguistics in Industry (Lecture Series).

The companies / research labs represented (by way of association of the speakers at some point in their career) in the lecture series included Google, Microsoft, Yahoo, PARC, Nuance, SRI, Powerset, NASA Ames, Cataphora, Facebook, etc. The applications and technologies discussed were machine translation, speech recognition and synthesis, natural language search, information retrieval, document analysis, language modeling, multimodal processing, information extraction, grammar checkers, etc. The speakers themselves were a mixed bag of doctorates (some Master's) in linguistics, computer science, and related fields. Their work experience oscillated between hard core research and practical customer-driven product / software development.

One of the most frequently discussed themes was small startups vs. large companies. This aspect of industrial research gains prominence in a multi-faceted and a relatively young field like computational linguistics. The general consensus was that one needs to find a suitable niche while balancing the availability of resources, usefulness / marketability of the application and company priorities / infrastructure. Nevertheless there are factors universal in industrial computational linguistics. Some of these follow below.

Peter Norvig, Director of Research at Google Inc., conceded that increasing the amount of training data in data-driven language technologies can compensate to a certain extent the quality of the algorithmic performance.

Mike Cohen, who previously worked on Voice User Interfaces at Nuance and is now involved in the speech technology division at Google Inc., stated that real data (interspersed with noise) is the key to statistical language processing research. Secondly, data, end users, business models, markets and resources all drive with varying degrees of influence the industrial research projects.

David Bean, Chief Technology Officer of Attensity (text analytics software) disclosed that although a research idea may be theoretically sound, it does not endure in industry unless it works. Consequently, industrial research is goal driven, getting the results is more important than reflecting on its theoretical foundations.

Ron Kaplan, Chief Technology and Science Officer at Powerset (specializing in natural language search) reflected on the competition between small startups and large companies. For example, Powerset, in contrast to larger companies' search technologies (data driven) like Google and Yahoo, endeavors to incorporate natural language understanding and deep linguistic processing to enable information search.

Beth Ann Hockey, who worked at NASA Ames on Clarissa (a fully voice-operated procedure browser) reiterated the novelty of hybrid computational linguistic models, i.e. combining rule-based (grammatical) systems with data-driven (statistical) training sets to boost performance and accuracy scores.

There were several speakers who did not work in traditional computational linguistics application areas but applied similar models and methodologies in their respective areas. This enforces the idea that any field can take inspirations from results in other related fields. This is particularly true for such an inter-disciplinary field. For example, Jared Bernstein from Ordinate Corporation showed an application of computational linguistic modeling to automated language proficiency tests. Also Christopher Cox from Facebook - a social network website, demonstrated use of statistical models (as used in computational linguistics) for several phases of their product developments like spam filtering and nickname recognizers.

Elizabeth Strand from Tellme asserted importance of academic success in graduate schools and interning in companies for a career in industrial research.

Chris Hogan from H5 summarized some of the skills and nature of a computational linguistic job. There are several types of computational linguistic jobs including pure research, lexicography, quality assurance, professional services, and software development.

In a nutshell, the lecture series was a true learning experience. I think that the most important skill in this field is flexibility, the ability to adapt and change one's methodologies as well as accept and incorporate alternate theories into one's work goal.
