Natural Semantics

Fernando Pereira
Google and University of Pennsylvania

The Delusion of Classification

- Common assumptions
  - there is a systematic mapping from data to "semantic" labels
  - labels govern useful computation
- Evident in:
  - Semantic Web
  - supervised machine learning for data interpretation

Effects of the Delusion

- Slow progress in natural language interpretation, functional annotation of genetic material,...
- Constantly postponed “semantic” search
- Large resources used to manually annotate data for supervised machine learning

Meaning in Natural Activity

- Meaning governs:
  - How texts are translated
  - How people transcribe speech
  - Which Web search results are clicked on
  - How genomes evolve (meaning=function)
- Can machines learn from these processes?

Some Successes

- Statistical machine translation: exploit parallel texts in multiple languages
- Large-vocabulary speech recognition: exploit large bodies of transcribed speech
- Comparative genomics: recognize functional elements through homologies between related genomes

Other Promising Cases

- Web queries and corresponding clicks
- Images and their surrounding text
- Movies and their screenplays
- Documents and their summaries

A General Principle

- Semantics consists of systematic associations produced by human and natural processes
  - Phrases and their translations
  - Conserved genomic elements
  - ...
- Semantic labels, if needed, can be propagated through those associations

Research Opportunities

- Richer mathematical models of semantic associations
- Algorithms for learning associations from large scale data
- Integrate multiple sources of partial semantic evidence:
  - Dream: submit newly sequenced genome and have it functionally annotated overnight using associations and evidence propagation