From Monsters to Machines

A recurring theme — or nightmare — is the creation of the *talking artifact*:

Caliban: Shakespeare's *The Tempest*

(*) The monster: *Frankenstein*

The false Maria: *Metropolis* (1926), the first great science-fiction film?

The HAL 9000 (*2001: A Space Odyssey*), Ash the android (*Alien*), Agents (*The Matrix*), ...

... But also C3PO, Data (*Star Trek*), automated translation systems, grammar checkers, search engines, ...

Language Processing is Challenging

Natural language processing (NLP) is "AI-complete": All the difficult problems in artificial intelligence manifest themselves in NLP (cf. Turing 1950).

1. List all flights on Tuesday.
2. I saw her duck with a telescope.
3. L'avocat vert est sur la table.

   $\longrightarrow$ "The green lawyer is on the table" (Babel Fish (Altavista)/Systran)
4. (Grishman 1986)

   Q: Do you know when the train to Boston leaves?

   A: Yes.

   Q: I want to know when the train to Boston leaves.

   A: I understand.

A Learning Approach

In some ways, Shelley was very prescient.

- Computers can learn about language directly from language samples.
  - ▷ cf. the Monster listening in on the family
- Computers benefit from human guidance about language samples
  - ▷ cf. the Monster listening to the family teaching Safie

Computers can utilize more data than a human family can produce; but they initally "know" far less about language than humans do.

State of the Art

- Continuous speech recognition: ViaVoice and NaturallySpeaking
  - ▷ Key component: probabilistic language models built from millions of words of text
- New-language machine translation in a day (EGYPT)
  - ▷ Key idea: train statistical English→Czech models to adapt to English→Chinese data

While these techniques seem quite different from human learning, Saffran et al.'s child development studies show that infants learn statistical patterns in visual, aural, and speech stimuli

In the future: use sophisticated computational techniques, more data, and better linguistic models.

Segmentation

Japanese, Chinese, Thai, ...: no spaces between words

社長兼業務部長

theyouthevent = ?

Combining simple statistics from unsegmented Japanese newswire yields results rivaling grammar-based approaches.

(Joint work with Rie Kubota Ando)