Week 13: Understanding natural language--part 2

Difficulty
State of the art today
Some techniques

COMP 378 -- Spring, 2015
Mr. Weisert

Magnitude of the problem
- Understanding natural language is a huge topic.
- In an A.I. curriculum it's a whole course and that's just for an introduction!
- Science-fiction predictions including speech recognition (e.g. Star Trek, conversational HAL in 2001) were far ahead of reality.
  - Will we ever get there?
  - How soon?
  - Would we want to?
  - Do we risk losing control?

Watson (2013)
- IBM's supercomputer equipped with huge, fast memory banks supported massive amounts of stored information with highly efficient and sophisticated indexing.
- Demonstrated as a contestant against two human champions on the TV quiz program Jeopardy:
  - Questions were in text, not spoken, form.
  - It almost always produced a correct or at least understandable answer faster than the competition.
  - But occasionally it would produce a really stupid and irrelevant answer!

Why? Can it get better? In what ways?

Strength and limitations
- Watson is vast in scale and raw speed.
- Much of the memory searching is based simply on key words.
- But Watson, although impressive, lacks world view, context from experience, common sense, recognition of absurdity.
- So automated dialogues from Eliza to Watson and beyond will keep getting better at the illusion of understanding, without ever actually understanding meaning.

Will they ever pass the Turing test?
Introductory material

- The chapter from a UNM book (link from course schedule) shows some of the potential, but doesn't demand deep background in specific tools & techniques.
  - The negative Amazon review by Helzerman recommends three others that we should review and consider.
  - The lecture notes by Pereira & Shieber are particularly well regarded.

Language representation

- Prolog was originally designed mainly to support logic:
  - Predicate calculus, proofs, derivations, etc.
- But it's also well suited to representing natural languages:
  - Propositions (sentences) can be decomposed into graphs.
  - Graphs can be represented naturally in predicate calculus (see examples pp. 108-109 in UNM chapter).

UNM's type of relation

(see diagram p. 110)

- These are a bit oversimplified, but they may help us to get started.
  - **agent**: relation between an action and the animate object causing the action.
  - **experiencer**: relation between a mental state and the animate object experiencing it.
  - **instrument**: links an act with an entity
  - **object**: links an event or state with an entity; represents verb-object relation
  - **part**: relation between whole and part.
- Whether we adopt them exactly or not, they can help to clarify structure and meaning.

Context-free grammar

- The rules: for example
  
  Sentence <-> NounPhrase VerbPhrase
  
  i.e. a sentence (at least one kind of sentence) consists of (is defined as) a noun phrase followed by a verb phrase.
- Compare <-> with BNF :=
- Section 8.3 in the UNM chapter.
- This is the essence of decomposing a proposition. We've seen it before. (see example p. 112)
Special difficulties
- Natural languages weren't designed; they evolved.
  - Influenced heavily by unsophisticated, even illiterate speakers
- Consequently they're loaded with ambiguity and "unnecessary" complexity:
  - of structure
  - of vocabulary
  - of spelling & pronunciation, but we won't worry about those here.

Some languages are more difficult than others
- English, now the world's most popular second language, is particularly irregular and therefore very difficult to process by computer.
  - Many identical words have multiple, very different meanings.
  - Huge vocabulary (the largest of any language! Why?) presents some very subtle choices.
  - Many illogical idioms.
  - Few inflections (e.g. Russian noun and adjective case endings are hard for people to memorize, but they resolve many ambiguities.)

Probabilistic parsing
- Complexity of natural languages leads to multiple legal interpretations of sentences.
  - Of course, some of those interpretations are much less likely than others. But who is to judge and how.
- The UNM chapter contains a lengthy example (section 8.4) that's a bit hard (or tedious) to follow:
  - Don't worry about understanding every line.
  - But note the weighting principle involved.