Computation in Computational Linguistics

Intel
August 24th, 2007

Emily M. Bender
Assistant Professor & Faculty Director
Professional Master's in Computational Linguistics
University of Washington
ebender@u.washington.edu

Goals of this talk

- Overview of the field of computational linguistics

- Examples of computationally intensive algorithms

Overview

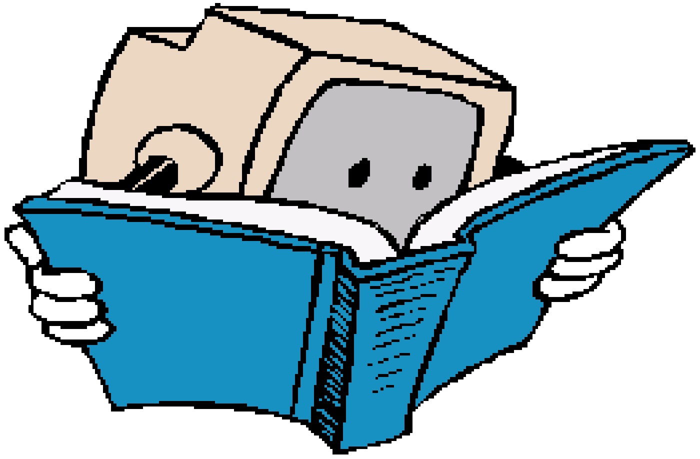
- What is NLP and what it is good for?
- Killer apps & example computationally intensive tasks
- Wrap up

Overview

- What is NLP and what it is good for?
- Killer apps & example computationally intensive tasks
- Wrap up

What is NLP?

- NLP: The processing of natural language text by computers
  - for practical applications
  - ... or linguistic research
- NLU: NLP with the goal of extracting meaning from the text for further machine processing



Human Language Understanding

- Relies on a wealth of intricate grammatical knowledge
- Is supported by an even greater wealth of world knowledge
- This means that information stored in natural language text requires a complex set of keys

Levels of linguistic structure

- Phonetics: Speech sounds, how we make them, how we perceive them
- Phonology: The grammatical structure of sounds and sound systems
- Morphology: How meaningful sub-word units combine to make words
- Syntax: How words combine to make sentences
- Semantics (lexical, propositional): What words mean and how those meanings combine to make sentence meanings
- Pragmatics: How sentence meanings are used to convey communicative intent
- ...

Pervasive ambiguity

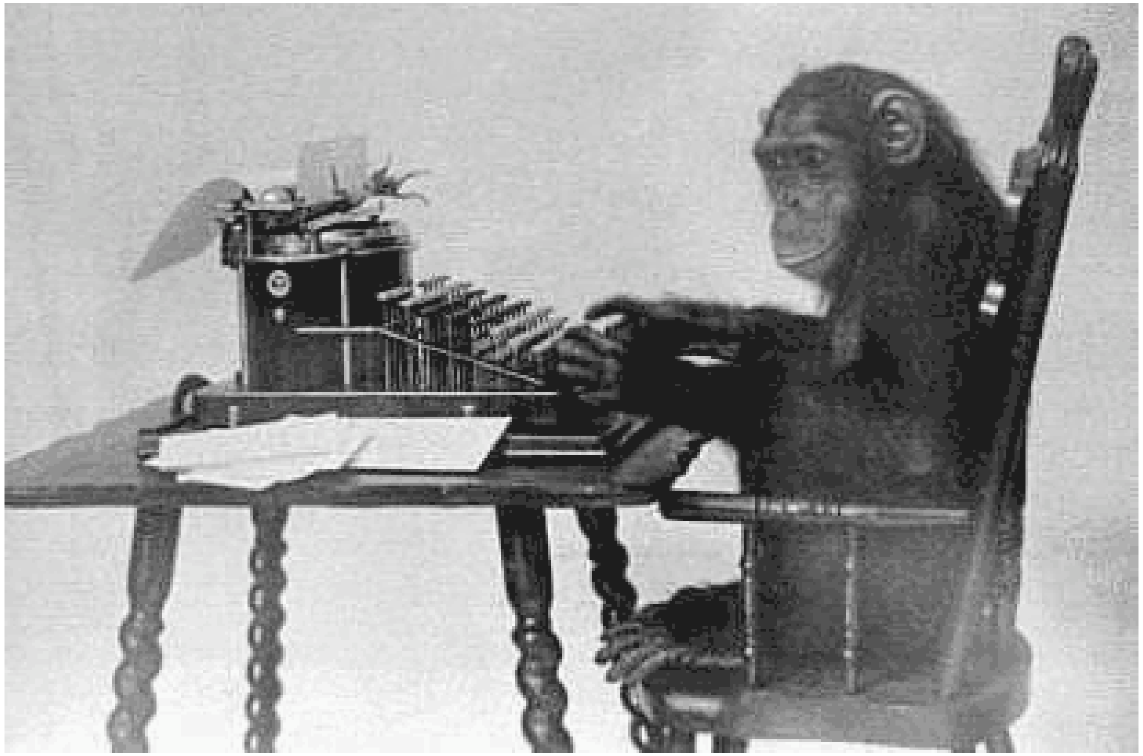
- Phonetic: *It's hard to wreck a nice beach.*
- Morphological: *This choice is undoable.*
- Syntactic: *Time flies like an arrow.*
- Semantic: *Every person read some book.*
- Pragmatic: *You should take those penguins to the zoo!*

And that’s only the tip of the iceberg!

- Ambiguities are typically independent, leading to combinatorial explosions.
- *Have that report on my desk by Friday* (32-ways ambiguous)
- Humans are generally bad at detecting ambiguity, a consequence of being so good at *resolving* it.
- In NLP, stochastic models usually stand in for the common sense knowledge people use.

NLP: Spectrum of approaches

- Knowledge engineering
- Stochastic models
  - Supervised v. unsupervised training
  - Incorporation of hand-made resources
- Hybrid approaches



Overview

- What is NLP and what it is good for?
- Killer apps & example computationally intensive tasks
- Wrap up

Killer Apps

- In-car dialogue systems (TellMe, VoiceBox, Bosch, ...)
- Machine translation (Systran, Language Weaver, Microsoft, Google, ...)
- Information extraction (Google, Yahoo!, Microsoft, PowerSet, Cataphora, InQuira, ...)

Killer Apps

- In-car dialogue systems (TellMe, VoiceBox,Bosch, ...)
- Machine translation (Systran, Language Weaver, Microsoft, Google, ...)
- Information extraction (Google, Yahoo!, Microsoft, PowerSet, Cataphora, InQuira, ...)

Dialogue System

- Signal capture
- Speech detection (was that noise or speech?)
- Speech recognition (speech to text)
- Addressee detection (are they talking to me?)
- Utterance segmentation
- Syntactic/semantic processing
- Discourse model
- Reference resolution
- Dialogue management (what to say/ do next?)
- Strategic generation
- Tactical generation
- Speech synthesis

Dialogue System

- Each of those tasks is potentially computationally expensive
- For a dialogue system, need real time performance
- Each level presents ambiguity
  - Potential performance gains by postponing ambiguity resolution
  - Input to each level is a lattice of hypotheses

Dialogue System

- Signal capture
- Speech detection (was that noise or speech?)
- Speech recognition (speech to text)
- Addressee detection (are they talking to me?)
- Utterance segmentation
- Syntactic/semantic processing
- Discourse model
- Reference resolution
- Dialogue management (what to say/ do next?)
- Strategic generation
- Tactical generation
- Speech synthesis

Example 1: Speech-to-text

- Shannon's noisy channel model:
- Underlying signal (intended utterance) sent through a noisy channel (articulatory system/acoustic signal)
- Goal is to estimate the most probable underlying signal given the observed output:

$$\hat{words} = \mathrm{argmax}(p(words|sounds))$$

- Bayes' rule:

$$\hat{words} = \mathrm{argmax}(p(sounds|words)p(words))$$

Example 1: Speech-to-text

$$\hat{words} = \mathrm{argmax}(p(sounds|words)p(words))$$

- Acoustic model: p(sounds|words)
  - Fourrier transform on acoustic signal
  - Machine learning over features of spectrogram
  - Output: lattice of word hypotheses

Example 1: Speech-to-text

$$\hat{words} = \text{argmax}(p(sounds|words)p(words))$$

- Language model: p(words)
  - Which path through the lattice looks the most like English?
  - Most common: n-gram models (HMMs), estimated from counts of word sequences over lots and lots of text
  - Coming into vogue: Structural models based on parsing
- SSLI Lab at UW: 80+ dual core machines, experiments usually run for days

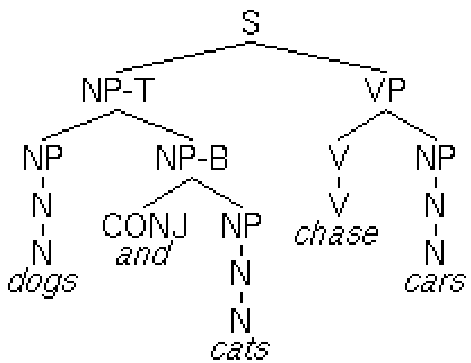
Dialogue System

- Signal capture
- Speech detection (was that noise or speech?)
- Speech recognition (speech to text)
- Addressee detection (are they talking to me?)
- Utterance segmentation
- Syntactic/semantic processing
- Discourse model
- Reference resolution
- Dialogue management (what to say/do next?)
- Strategic generation
- Tactical generation
- Speech synthesis

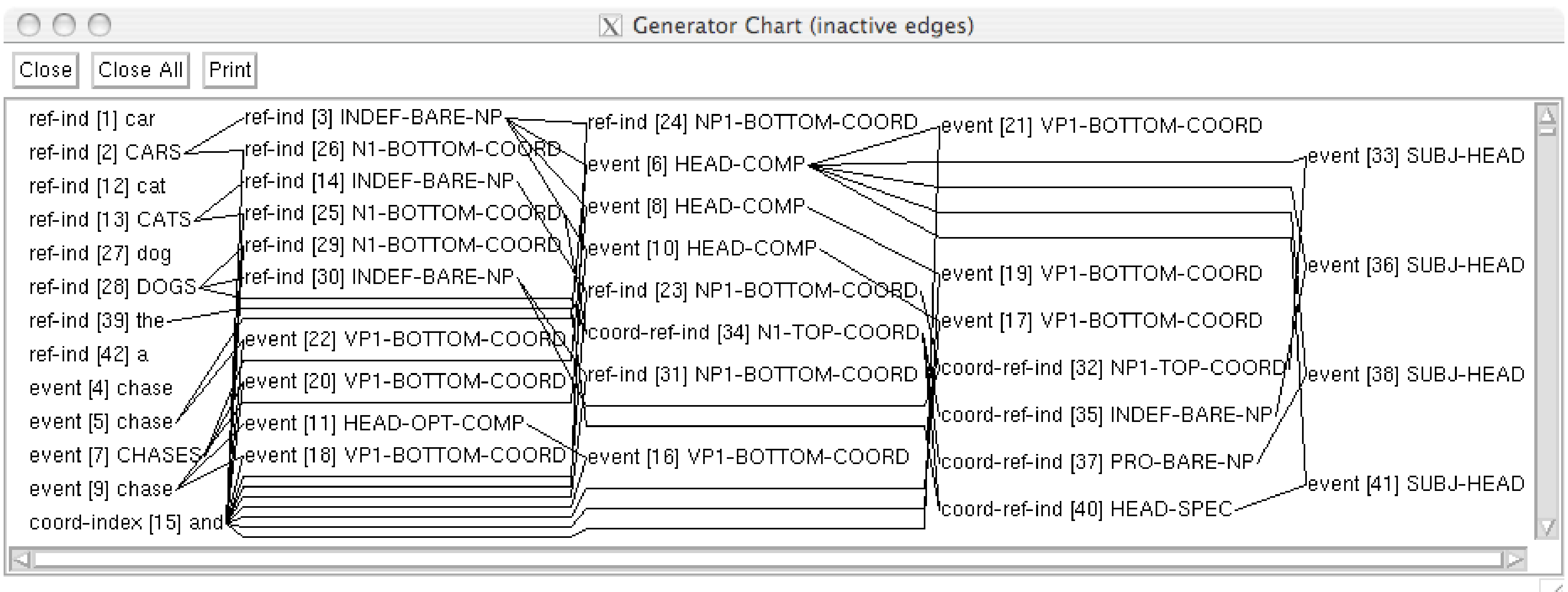
Example 2: Tactical generation

- Input: Semantic representation
- Output: Well-formed string(s) corresponding to input semantics
- Knowledge base:
  - Lexicon: maps semantic relations to words + syntactic constraints
  - Grammar: rules for constructing phrases (and phrase meanings) from words/smaller phrases
- Find all words that match semantic relations in input semantics

Example 2: Tactical Generation



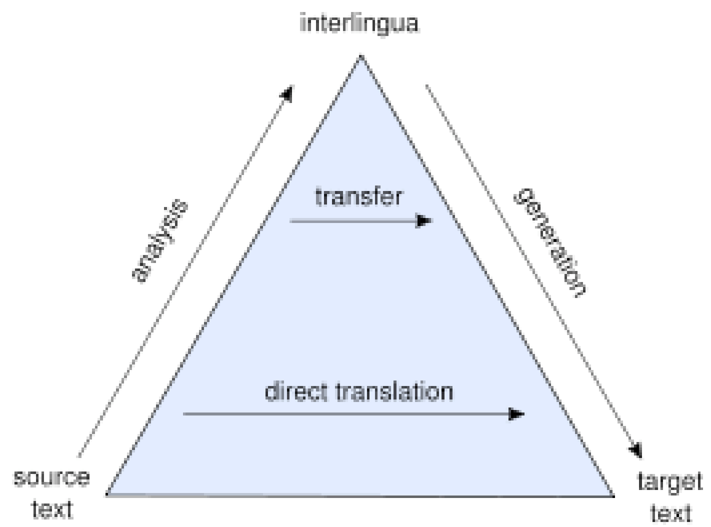
h1 e2{ prop-or-ques }
{ h3:_dog_n_rel(x4{ 3 non-sg type-id })
  h5:exist_q_rel(x4, h6, h7)
  h8:_and_coord_rel(x9, h11, x4, h12, x10{ 3 non-sg type-id })
  h13:_cat_n_rel(x10)
  h14:exist_q_rel(x10, h15, h16)
  h17:exist_q_rel(x9, h8, h18)
  h1:_chase_v_rel(e2, x9, x19{ 3 non-sg type-id })
  h20:_car_n_rel(x19)
  h21:exist_q_rel(x19, h22, h23) }
{ h6 =q h3 h15 =q h13 h22 =q h20 }



Killer Apps

- In-car dialogue systems (TellMe, VoiceBox,Bosch, ...)
- Machine translation (Systran, Language Weaver, Microsoft, Google, ...)
- Information extraction (Google, Yahoo!, Microsoft, PowerSet, Cataphora, InQuira, ...)

Machine Translation: Vauquois Triangle



Statistical Machine Translation

- Noisy channel again:
  - Speaker intended to speak English, but the signal came out in Japanese.

$$\hat{e}\text{-string} = \text{argmax}(p(j\text{-string}|e\text{-string})\,p(e\text{-string}))$$

  - Language models same as with speech-to-text
  - Translation models learned from parallel corpora (bitexts)
  - Step 0: Align sentences
  - Step 1: Align words

Example 3: Word Alignment

- Input: Sentence aligned bitext (the more words the better)
- Output: Probabilistic bilingual dictionary
- Expectation Maximization (EM) algorithm (hill-climbing):
  - Initialize: Align every source word with every target word, with equal probability
  - E step: Count alignments of each source word to each target word, and estimate probabilities
  - M step: Reassign probabilities to alignments based on previous E step
- M step is easily parallelized, E step requires more coordination

Killer Apps

- In-car dialogue systems (TellMe, VoiceBox,Bosch, ...)
- Machine translation (Systran, Language Weaver, Microsoft, Google, ...)
- Information extraction (Google, Yahoo!, Microsoft, PowerSet, Cataphora, InQuira, ...)

Information Extraction

- Miyao et al (2006): Retrieval of relational concepts from massive text databases
- Biomedical domain
- (Domain) expert users
- Availability of resources (e.g., ontologies)

Miyao et al: Problem

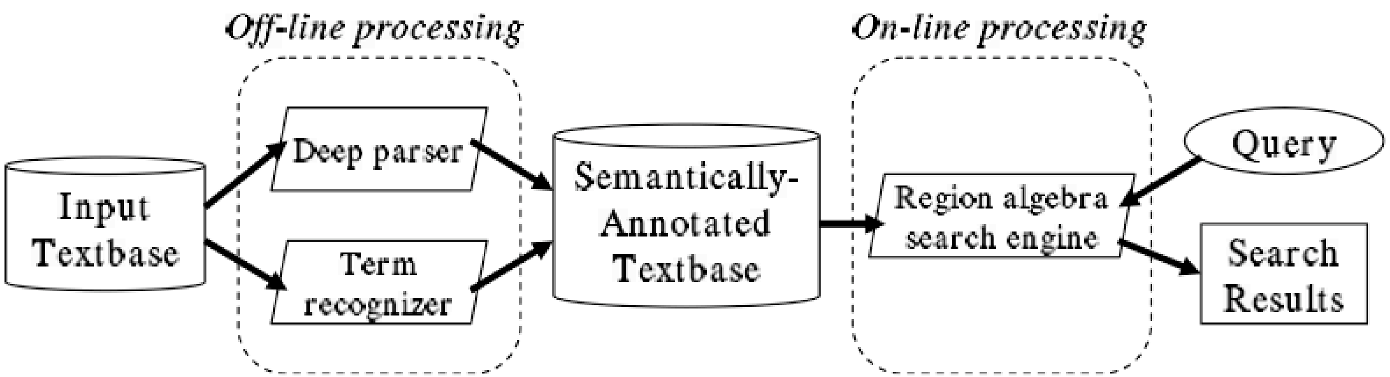
- Biomedical results are reported in natural language text.
- MEDLINE indexes 4500 journals (14,785,094 articles as of 2006).
- Researchers want answers to queries like: “What triggers diabetes?”, “What inhibits ERK2?”
- State-of-the-art: Keyword based searches.
- Can semantic search (using ontologies and parsing for predicate argument structure) do better?
- Big problem: Lots of text, a broad range of concepts
- Also narrow: Queries target simple relations between two entities

Miyao et al: Resources

- Ontologies: GENA (metadatabase of genes and gene products; Koike & Takagi 2004); UMLS (other biomedical and health concepts; Lindberg et al 1993)
  - Map textual expressions to real-world entities
- Term recognizer: map expressions in the text to ontology entries (Tsuruoka and Tsujii 2004)
- Parsing technology: A probabilistic HPSG parser (Miyao & Tsujii 2005), which extracts predicate argument structure. (97.6% coverage on MEDLINE corpus)
  - *exclude* (ARG1: *CRP*, ARG2: *thrombosis*)
- Treebank: GENIA Treebank (Tateisi et al 2005), contains biomedical domain text

Miyao et al: Methodology

- Parse corpus offline, store predicate-argument structures in a structured database.
- Run term recognizer to annotate sentences with links to ontology



- Convert queries to extended region algebra
- Match queries to semantic annotations to return relevant passages

Miyao et al: Evaluation

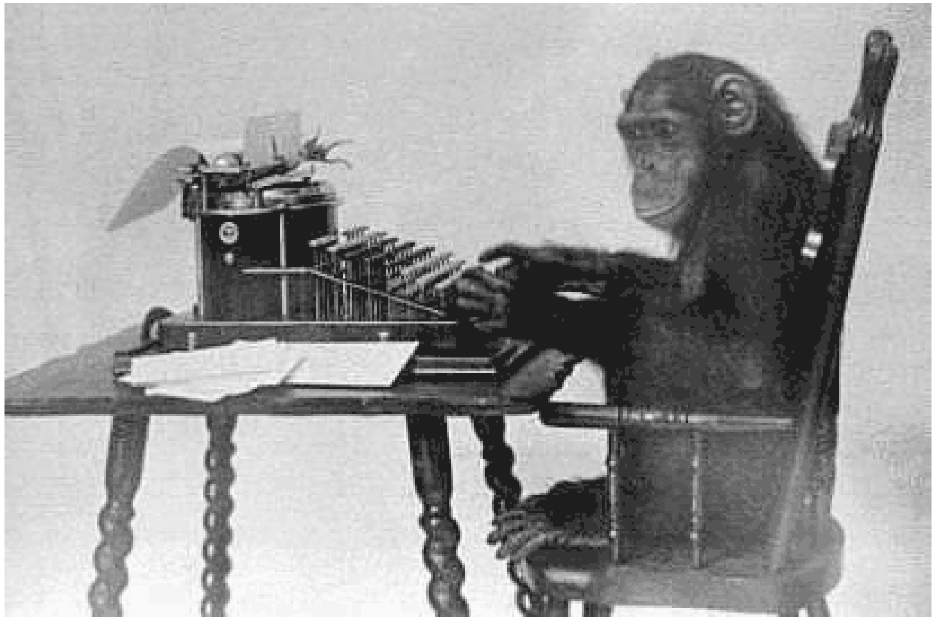
- 8 sample queries
- Max 100 results per query
- With and without ontological mapping; keyword or semantic matching
- Present results from different conditions in random order to biologist for evaluation
- Keyword precision ranges from 0-74%
- Semantic search precision 60-97%
- Effect of ontological mapping also clear
- This task values precision over recall

Overview

- What is NLP and what it is good for?
- Killer apps & example computationally intensive tasks
- Wrap up

Knowledge engineering and machine learning

- After swinging hard towards machine learning, the pendulum is returning to hybrid approaches
- Knowledge engineering contributes precision, depth of analysis
- Machine learning contributes robustness and scalability



The promise of NLP

- The amount of information stored in digitized text is increasing every day
- Long-promised applications seem closer than ever:
  - Dialogue systems (“personal assistants”)
  - Machine translation and other multilingual NLP
  - Automated question answering based on web content
  - NLP for business intelligence
  - ...

To learn more...

- The ACL recently launched a wiki:

  http://aclweb.org/aclwiki

- Papers from top conferences back to 1965 are available online:

  http://acl.ldc.upenn.edu

- Computational linguistics at the University of Washington

  http://www.compling.washington.edu

Thank you!