Introduction to NLP

Craig Martell
Naval Postgraduate School
Aug 09, 2011
(Slides shamelessly stolen)

Acknowledgements

- Slides adapted from those by Pranav Anand, Bonnie Dorr, Jason Eisner, Roger Levy, Mitch Marcus, Rada Mihalcea & Ted Pedersen, and Massimo Poesio.

Overview

- Computational Linguistics vs Natural Language Processing
- Some informative history
- Going Distributional (Part of Speech Tagging)
- Distributional Syntax (Language Modeling)
- Vector Space Modeling and Information Retrieval
- Word Sense Disambiguation

Linguistics vs. Computational Linguistics vs. Natural Language

Linguistics

- The scientific study of language
- Goal is discovery of universal generalizations that apply to all languages, or sets of languages, or whatever the natural grouping is.
- Discovery of what the natural kinds of language are (like whether or not there are natural groupings, etc.)

Computational Linguistics

- Applying computational tools to the study of language
- Corpus linguistics: using data sets of language to formulate and test theories of language
- Using the results from computational complexity to limit the types of theories of language we develop.
- Stays true to the natural phenomenon of language

Natural Language Processing

- Text processing with a goal
- Dealing with spoken words is called “Speech Recognition”
  - Often the output of an ASR system is the input to an NLP system
- Sets of algorithms that solve problems that people want solved.
  - Topic detection
  - Authorship detection
  - Even semantic understanding (not here yet ☺, although on the way)
- Can use any technique that works
- Doesn’t pretend to stay true to cognitive

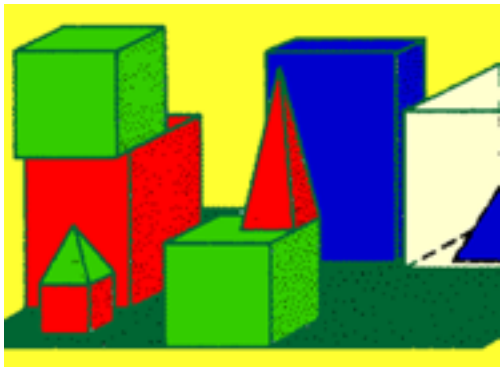
Some informative history

Early NLP successes

- Goals: systems that behaved comparably to humans on given linguistic tasks
- Models often build on models of human reasoning, e.g., reasoning by analogy, logic, etc.
- SHRDLU (Winograd, 1971): interaction with simulation of tabletop
- ELIZA (Weizenbaum, 1966): does it work? Weizenbaum thought no, secretary thought yes.
- Worked because:
  - Restricted domain plus **LOTS of person-hours on given domain**
  - Interactive nature allowed for easy correction
- Very brittle

SHRDLU

Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.



Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.

Eliza

- http://www.manifestation.com/neurotoys/eliza.php3

Another way: statistical

- Church (1985): statistical etymology classifier (important for stress in speech synthesis [calculus vs. spaghetti)
  - Classified a word based on probabilities over 3 letter segments (tri-grams)

    $P(xyz|l) = C_l(xyz)/C_l$ (*)

    Choose the l that maximizes the probability
- Worked!

Angelletti	100%	Italian
Asahara	100%	Japanese
Umeda	96%	Japanese
Anagnostopoulos	100%	Greek
Demetriadis	100%	Greek
Dukakis	99%	Russian
Annette	75%	French
Deneuve	54%	French

Modeling with words

- The founder of Pakistan's nuclear program, Abdul Qadeer Khan, has admitted he transferred nuclear technology to Iran, Libya and North Korea, a Pakistani government official said Monday.

  Khan made the confession in a written statement submitted "a couple of days ago" to investigators probing allegations of nuclear proliferation by Pakistan, the official told The Associated Press on condition on anonymity.

  The transfers were made during the late 1980s and in the early and mid 1990s, and were motivated by "personal greed and ambition," the official said.

  The official said the transfers were not authorized by the government.

Unigrams

Word	# in Document
Khan	15
nuclear	14
Pakistan	10
transfers	9
official	8
scientists	5
journalists	5
government	5
Libya	5
officials	4
military	4

Modeling with word pairs

- The founder of Pakistan's nuclear program, Abdul Qadeer Khan, has admitted he transferred nuclear technology to Iran, Libya and North Korea, a Pakistani government official said Monday.

  Khan made the confession in a written statement submitted "a couple of days ago" to investigators probing allegations of nuclear proliferation by Pakistan, the official told The Associated Press on condition on anonymity.

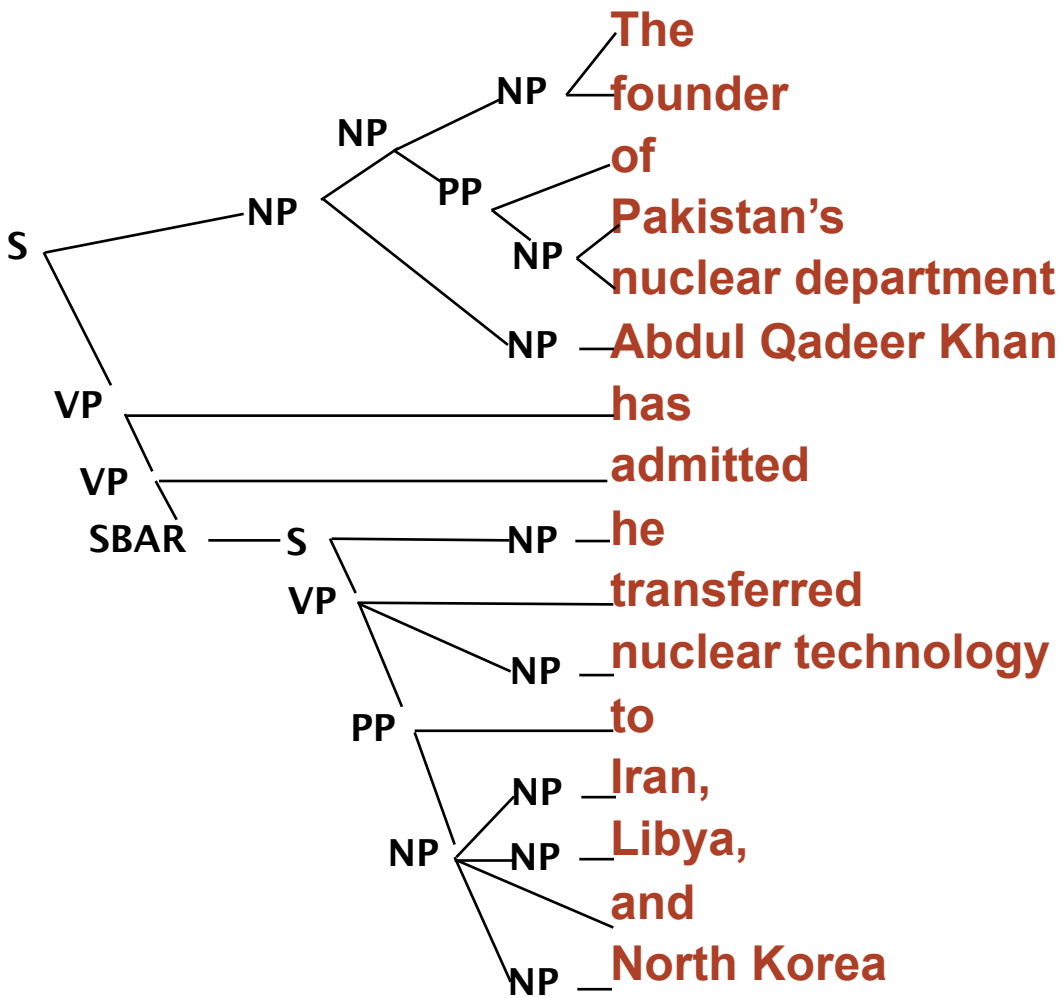
  The transfers were made during the late 1980s and in the early and mid 1990s, and were motivated by "personal greed and ambition," the official said.

  The official said the transfers were not authorized by the government.

Bigrams

Bigram	# in Document
North Korea	4
nuclear transfers	3
Government official	3
Pakistan's nuclear	3
written statement	2
told investigators	2
other suspects	2
other Muslim	2
nuclear program	2
nuclear powers	2
military officials	2
become nuclear	2

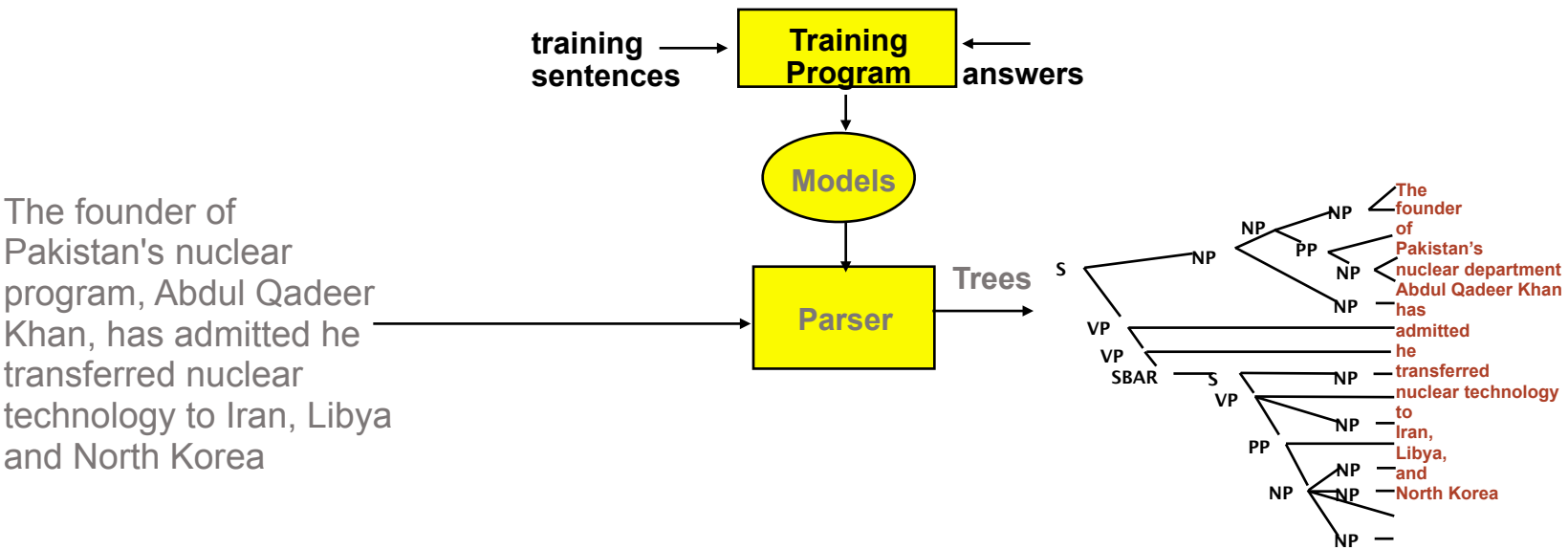
Phrase Structure requires annotation



Machine Learning with Corpora

10^6 words of Treebank Annotation

+ Machine Learning = Robust Parsers



- **1990 Best hand-built parsers: ~40-60% accuracy**
- **1995+ Statistical parsers: ~90% accuracy**

Upshot

- Hand-crafted rules are brittle
- Distributional information can often be a proxy for rich structure
- Large-scale annotations of rich structure can produce better models (with machine learning) than hand-built methods.
- Large-scale annotation is VERY expensive and time consuming