Course Basics

COSI 140 – Natural Language Annotation for Machine Learning

James Pustejovsky

January 15, 2016

Brandeis University

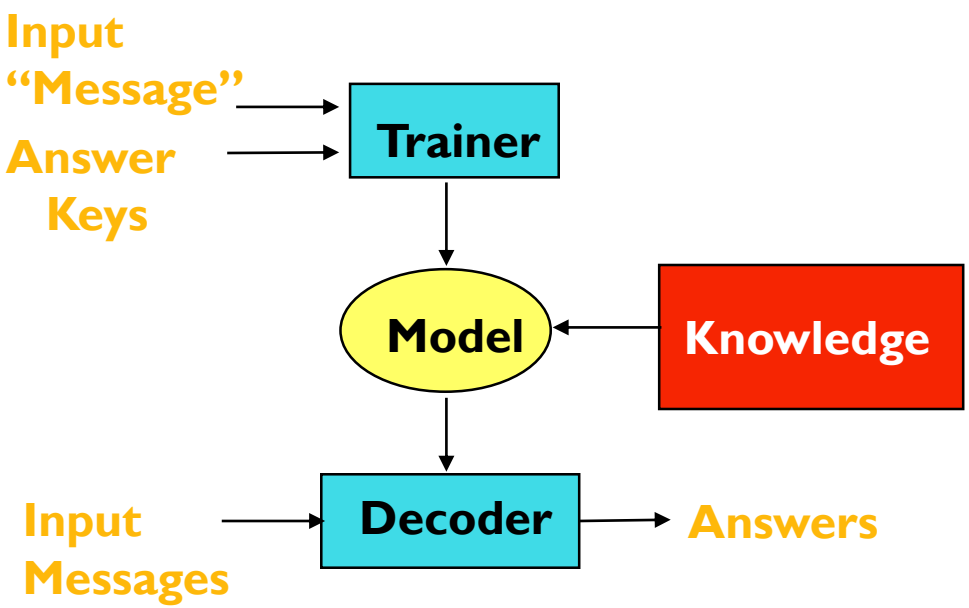
Course overview

- Schedule and assignments
  - www.cs140.org
- Learn by doing
  - Course is centered around group annotation projects
  - We will walk through every step of the process
- Textbook:
  - *Natural Language Annotation for Machine Learning*
  - Pustejovsky & Stubbs. 2012. O'Reilly

Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Variously referred to as the "corpus based," "statistical," or "empirical" approach.
- Statistical learning methods were first applied to speech recognition in the late 1970's and became the dominant approach in the 1980's.
- During the 1990's, the statistical training approach expanded and came to dominate almost all areas of NLP.

Speech and NL Paradigm



- **Requirements:**
  - Annotation of messages with keys
  - Linguistic and Domain Knowledge
  - Statistical Model
  - Training Algorithm
  - Decoding Algorithm
- **Benefits:**
  - Statistical model can combine multiple kinds of information
  - Degrades "softly", finding the most likely answer
  - Learns what information is important to make a decision

Supervised Learning for Language Technologies

Technology	Input	Answers
Speech Recognition	Audio	Transcription
Optical Character Recognition	Image	Characters
Topic classification	Document	Topic labels
Information retrieval	Query	Document
Named entity extraction	Text or speech	Names and categories

Advantages of Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - "The a are of I" is a valid English noun phrase (Abney, 1996)
    - "a" is an adjective for the letter A
    - "are" is a noun for an area of land (as in hectare)
    - "I" is a noun for the letter I
  - "Time flies like an arrow" has 4 parses, including those meaning:
    - Insects of a variety called "time flies" are fond of a particular arrow.
    - A command to record insects' speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
  - "vice president Gore" vs. "dice precedent core"
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

Course Project

- Form Groups
- Annotation goal
- Group Contract
- Task Description and Corpus Selection
- Initial Annotation Spec
- Full Annotation Spec
- Adjudication and precision and recall
- Train Machine Learning Algorithm
- Write-up
- Presentation
- Peer Evaluation