Computational Linguistics

Ling 200 Guest Presentation
Michael Tepper
July 16th 2007

Acknowledgements: Adapted from a presentation given by D. Goss Grubbs, L. Paulson, M. Scanlon and myself in Spring 2006,
as well as a presentation given by Dr. E. Bender in Spring 2007

Overview

- Introduction
- Main Approaches
  - Rule-based:Terminology
  - Rule-based: Applications: Spell-checker
  - Statistical:Terminology
  - Statistical: Applications
    - Supervised: Speech-recognition
    - Unsupervised: Morphological induction
- So, you want to be a computational linguist...

Introduction

- Computational Linguistics
  - processing of human language by computers to facilitate linguistic research
- Natural Language Processing (NLP)
  - development of computer--natural language interface applications
- The line between applications and linguistic research is blurry, so Computational Linguistics may be used as a cover-term for both.

CompLing is multidisciplinary

- Linguistics
  - e.g. grammar engineering
- Electrical Engineering
  - e.g. speech recognition
- Computer Science
  - e.g. machine translation
- Psychology/Cognitive Science
  - e.g. cognitive modeling

Overview

- Introduction
- Main Approaches
  - Rule-based:Terminology
  - Rule-based: Applications: Spell-checker
  - Statistical:Terminology
  - Statistical: Applications
    - Supervised: Speech-recognition
    - Unsupervised: Morphological induction
- So, you want to be a computational linguist...

Terminology for Approaches Used

- Rule-based: Propose a system of hand-coded rules for analyzing and/or interpreting linguistic phenomena.

Knowledge-driven/Rationalist Approach

- Supervised Learning: Machine learning proceeds from data with target answers (labels, analyses, etc) provided.
- Unsupervised Learning: Machine learning proceeds from data with no target answers (labels, analyses, etc) provided.

Machine-Learning/Empiricist Approach

Overview

- Introduction
- Main Approaches
  - Rule-based:Terminology
  - Rule-based: Applications: Spell-checker
  - Statistical:Terminology
  - Statistical: Applications
    - Supervised: Speech-recognition
    - Unsupervised: Morphological induction
- So, you want to be a computational linguist...

How does a spell checker work?

- Main **rule-based** component:
  - Input text is compared to a dictionary (+ morphological analyzer) to detect non-words.
- Other useful components:
  - Runs error types in reverse (insertion, deletion, transposition, substitution) to come up with candidate corrections.
  - Ranks candidate corrections (according to frequency of word in context, severity of fix required to generate the correction)
  - What about spelling or usage mistakes that result in other actual words? (e.g. three/there, their/there, fare/fair)

Overview

- Introduction
- Main Approaches
  - Rule-based:Terminology
  - Rule-based: Applications: Spell-checker
  - Statistical:Terminology
  - Statistical: Applications
    - Supervised: Speech-recognition
    - Unsupervised: Morphological induction
- So, you want to be a computational linguist...

Background Terms: Machine Learning

- Statistical **machine learning**: subfield of artificial intelligence concerned with algorithms enabling computers to "learn". Involves:
  - Designing a probability model to suit the task
  - Estimate the probabilities for the model from some data (training)
  - Use the model to predict analyses, labels, values, etc for some new data (testing)
- Training: the process of extracting observations from data, counting them, and using their frequency to estimate a probability model.
- Testing: the process of using a model extracted in the above fashion to generate an analysis, label, value, etc, for some new linguistic input.

Terminology for Approaches Used

- Rule-based: Propose hand-coded rules for analyzing and/or interpreting linguistic phenomena.

Knowledge-driven/Rationalist Approach

- Supervised Learning: Machine learning proceeds from data that has been ***annotated*** with the learning target (labels, analyses, etc).
- Unsupervised Learning: Machine learning proceeds from data with no **annotation** (labels, analyses, etc) provided.

Machine-Learning/Empiricist Approach

Overview

- Introduction
- Main Approaches
  - Rule-based:Terminology
  - Rule-based: Applications: Spell-checker
  - Statistical:Terminology
  - Statistical: Applications
    - Supervised: Speech-recognition
    - Unsupervised: Morphological induction
- So, you want to be a computational linguist...

What is Speech Recognition?

- **Speech recognition** is a multilayered process that converts an acoustic speech signal into text.
  - Open domain: very large vocabulary size, unrestricted recognition of continuous speech
    - e.g. Dictation systems
  - Finite domain: typically recognizes only a small vocabulary; recognition may be restricted isolated words which map directly to actions in the system
    - e.g. Natural language dialogue systems, automated call centers, automatic cell-phone dialing

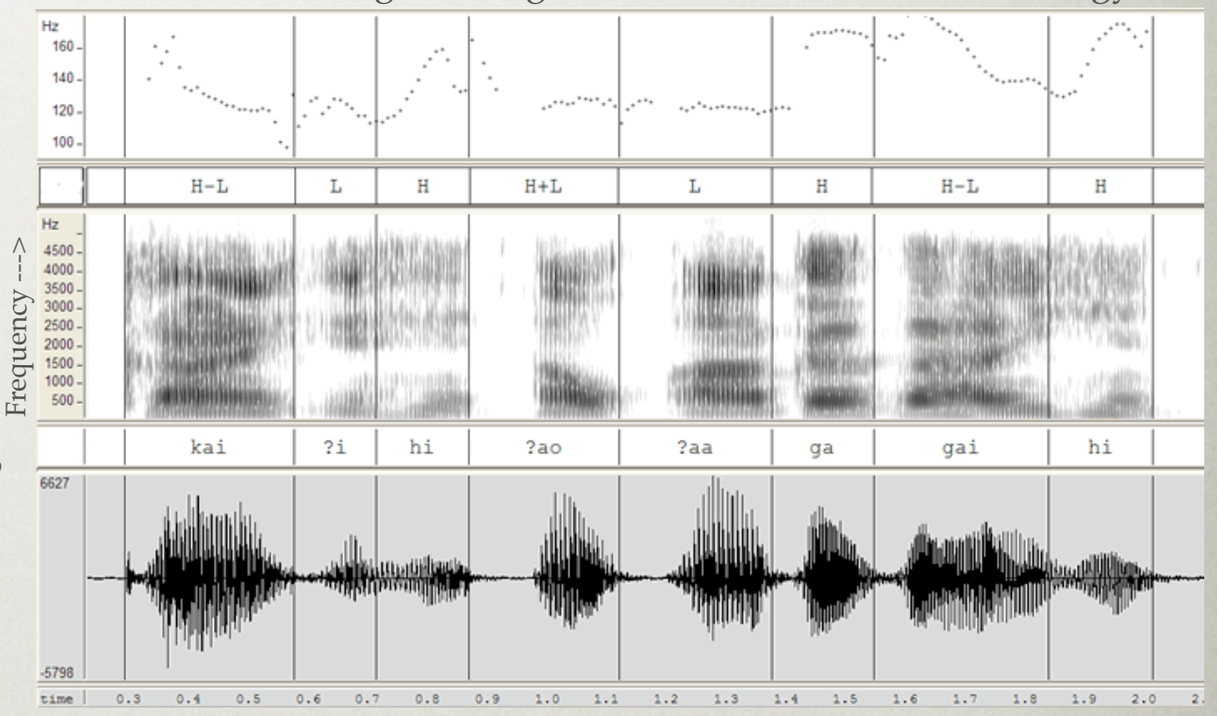
Speech Recognition Layers

- Signal Processing
- Acoustic Modeling
- Pronunciation Modeling
- Language Modeling

Acoustic Layers

- Signal Processing
  - Continuous speech signal transformed into a digital representation of the speech signal. This contains information on how energy in sig. is distributed over various frequencies.
- Acoustic Modeling
  - Computer uses a **supervised** procedure to collect observations on how various sequences of phones map to sequences of energy distributions, using observations extracted from training data.
  - This model is then used to come up with probable sequences of phonemes given an observed signal.

Annotated digitized signal: darker areas have more energy



Speech Recognition Layers

- Signal Processing
- Acoustic Modeling
- Pronunciation Modeling
- Language Modeling

Pronunciation Modeling Layer

- Pronunciation Modeling
  - Models how sequences of phonemes map to words. It usually involves using of a dictionary or database listing how phonetic sequences match to orthographic words.
- Only actual words are accepted at this stage:
  - [rɛkənajspič]
    ‘recognize speech’ ✓
  - [rɛkənajspiš]
    ‘recognize speesh’ !!
- This layer is an example of a rule-based approach.

Speech Recognition Layers

- Signal Processing
- Acoustic Modeling
- Pronunciation Modeling
- Language Modeling

Language Modeling Layer

- Language Modeling: assigns probability to a sequence of words based on observations of word sequences in training data.
- Weights ambiguous word/phrase interpretations according to a mix of factors. This is known as **ambiguity resolution**. Factors include:
    - ... lexical frequency
        - deer occurs 50 times, but dear occurs 215 times
        - for phoneme sequence, e.g. [dir], choose the most frequent: ‘deer’ (50) ‘dear’ (215)
    - ...frequency in in context, i.e. counting pairs, tripples, etc of words.
        - (wreck, a) and (nice, beach) both occur twice
        - (recognize, speech) occurs 10 times
        - “It’s hard to `[rɛkənajspič]`. How do you think the bracketed phoneme sequence would be rendered, given the counts?

Speech Recognition Overview

- Speech recognition has one major supervised machine-learning component built into it. What is it?
  - The acoustic model.
- What are some hard things in speech recognition?
  - Noisy input data.
  - Focus of acoustic model.
    - Don't try to tell [m] from [n] by their features, look to the surrounding vowels.
  - Language model selection.

Overview

- Introduction
- Main Approaches
  - Rule-based:Terminology
  - Rule-based: Applications: Spell-checker
  - Statistical:Terminology
  - Statistical: Applications
    - Supervised: Speech-recognition
    - Unsupervised: Morphological induction
- So, you want to be a computational linguist...

What is Morphological Induction?

- Morphological induction is the machine-learning task that involves learning morphology (morphemes), as well as how to perform a morphological analyses (break words into morphemes).
- Frequently this machine learning task is **unsupervised**.

How does this work?

- How does this typically work? This is **unsupervised** learning; algorithms do not learn by making observations on analyzed, annotated text.
- Algorithms start from a simple list of words and their frequencies:
  - e.g. {...,(27) play, (20) played, (30) seemed, (40) walked, (15) wish, (23) wished, (27) waited, (54) listed, ... }
- A popular technique is to look for the most efficient way to encode the wordlist text where letter-sequences are the prime units instead of characters.
- Think of it as breaking down a Lego building so that it may most efficiently be replicated again, while simultaneously making it most easily storable. Don't take away all the structure, leave some of the big pieces so it's easier to put together next time. But don't leave it all as one piece, because we won't be able to fit it in back in the box.
- In the word list above, what sequences should we treat as one unit?

Demos

- Spell-check
  - Try your own computer.
- Text-to-speech system (Oddcast)
  - http://www.oddcast.com/home/demos/tts/tts_example.php
- Semi-supervised (mostly unsupervised) morphological induction:
  - Demonstration of morphological inducer that uses rewrite rules (demo not on web)
- Speech-recognition Demo Movie (Windows Vista):
  - http://www.istartedsomething.com/20060808/vista-speech-recognition-screencast/ (if we have time)

Many ways to be a computational Linguist

- Most flexible option: training in both computer science and linguistics
- To prepare for UW's Professional Masters Program in Computational Linguistics:
  - Ling 200 ✓
  - CSE 142, 143, 373
  - Stat 391
- To learn more:
  - Ling/CSE 472 (prereq Ling 461 or Ling 200 + CSE 326)
  - http://compling.washington.edu
  - Linguistics colloquia, especially MS/UW symposium

THANKS! SAĞOLUN!
DANKE! !תודה