Summary

LING 570

Fei Xia

Week 11: 12/09/09

Outline

- Main units
- Main techniques
- What’s next?

Main units

Unit #0: introduction and summary

- 2 weeks
- Hw1

- Course overview
- Tokenization
- Introduction to probability theory

- IE
- Summary

Unit #1: finite-state machine

- 2.5 weeks
- Hw2 – Hw4

- Formal language
- Formal grammar
- Regular expression
- FSA
- Regular relation and FST
- Morphological analyzer

Unit #2: LM, HMM, and n-gram tagger

- 3.5 weeks
- Hw5 -- Hw7

- LM: n-gram models
- Smoothing
- HMM
- N-gram model

Unit #3: Classification

- 3 weeks
- Hw8 – Hw10

- Classification problem
- Example tasks:
  - MaxEnt POS tagger
  - NE tagging
  - Chunking

- Sequence labeling and beam search
- clustering

Main techniques

Main techniques (1)

- Probability theory:
  - The chain rule
  - The Bayes' rule
  - The (conditional) independence assumption
  - …

Main techniques (2)

- Regexp, regular language and regular grammar:

- FSA, FST, morphological analysis:
  - Combining simple FSTs in a pipeline can be very powerful

Main Techniques (3)

- LM and Smoothing
- N-gram model
- HMM
  - The Markov assumption
  - Viterbi algorithm

Main techniques (4)

- Classification and sequence labeling problems:
  - Representing an instance as a feature vector
  - Selecting features is very important
  - Many problems can be treated as classification or sequence labeling problems
  - Beam search

Tools created

- English Tokenizer with RegEx: Hw1
- Morphological analyzer with FST: Hw4
- LM and smoothing: Hw5
- Taggers:
  - Unigram model: Hw3
  - N-gram tagger: Hw6-Hw7
  - MaxEnt tagger: Hw9
  - Clustering: Hw10
- Using existing packages:
  - Carmel (Hw2)
  - Mallet (Hw8)

What’s next?

What's next?

- Other tasks ➔ NLP 571 (winter)
  - Ex: parsing, semantics, discourse, …
- Supervised learning ➔ NLP 572 (winter)
  - Ex: MaxEnt, Naïve Bayes, SVM, …
- Information extraction ➔ NLP 575A (winter)
- Emily's ling567 and ling575

Tentative plan for LING 572 (subject to change)

- Unit #0: Introduction
  - 0.5 week
  - Features, training/testing, …
  - Classification algorithms

- Unit #1: Simple algorithms
  - 2 weeks
  - kNN
  - Decision tree
  - Naïve Bayes

LING 572 (cont)

- Unit #2: More sophisticated algorithms
  - 2.5 weeks
  - MaxEnt (*)
  - SVM (**)
- Unit #3: sequence labeling problem
  - 2 weeks
  - TBL (if time permits)
  - CRF (**)
- Other topics: 2 weeks

A head start with LING 572?

- Textbook: none
- Last year's schedule:

http://courses.washington.edu/ling572/winter09/teaching_slides/new_syllabus.pdf

- More math in ling572:
  - Information theory: entropy, mutual information
  - Calculus, derivative of f(x), lagrange multipliers

LING 575

- Theme: Information extraction in the medical domain

- Student presentation
- In-class discussion
- System building

- The workload is about the same as ling570: about 20 hours/week excluding class time

LING 575: prerequisites

- Strong programming skills
- Take LING 572 concurrently
- Team work
- Participate in class discussion

Online option?

- If you cannot attend class live
  - it won't work well

- If you can attend class live,
  - be able to present remotely, and
  - be able to work with a teammate after class

  - You need let me know by 12/25/09

Between now and Jan

- I will turn in 570 grades on 12/19
  - If you have questions about your grades for 570 assignments, please let me know by 12/17.

- I will be traveling on 12/10-12/22, 12/24-12/26, and 1/6-1/10, and will be slow in replying to emails and GoPost.

- No class on Jan 7 for ling572

Course evaluation

- In-class students, fill out the paper forms

- For online students, fill out the form at

https://depts.washington.edu/oeaias/webq/survey.cgi?user=UWDL&survey=1055