Summary

LING 570
Fei Xia
Outline

• Main units

• Main techniques

• What’s next?
Main units
Unit #0: introduction and summary

- Hw1
- Course overview
- Tokenization
- Introduction to probability theory
- Summary
Unit #1: finite-state machine

- Hw2 – Hw4
- Formal language
- Formal grammar
- Regular expression
- FSA
- Regular relation and FST
- Morphological analyzer
Unit #2: LM, HMM, and n-gram tagger

- Hw5 -- Hw7
- LM: n-gram models
- Smoothing
- HMM
- N-gram model
Unit #3: Classification

• Hw8 – Hw10

• Classification problem

• Sequence labeling and beam search

• Example tasks:
  – MaxEnt POS tagger
  – NE tagging
  – Chunking

• Other topics: IE, clustering
Main techniques
Main techniques (1)

• Probability theory:
  – The chain rule:
    \[
    P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1) \ldots P(X_n|X_1, \ldots, X_{n-1})
    \]
  – The Bayes’ rule:
    \[
    P(B|A) = \frac{P(A|B)P(B)}{P(A)}
    \]
  – The (conditional) independence assumption
    \[
    P(X_n|X_1, \ldots, X_{n-1}) = P(X_n|X_{n-k}, \ldots, X_{n-1})
    \]
  – …
Main techniques (2)

• Regexp, regular language and regular grammar, and FSA

• Regular relation and FST

• FSA, FST, and morphological analysis:
  – FSA: DFA, NFA, PFA, WFA
  – 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

• FSA and FST acceptors: Hw2 and Hw3

• Morphological analyzer with FST: Hw4

• LM and smoothing: Hw5

• Taggers:
  – N-gram tagger: Hw6-Hw7
  – MaxEnt tagger: Hw9

• Classifiers: Hw8 and Hw10

• Using existing packages:
  – Carmel (Hw2)
  – Mallet (Hw8)
What’s next?
What’s next?

• Other tasks ➔ LING 571 (winter)
  – Ex: parsing, semantics, discourse, …

• Supervised learning ➔ LING 572 (winter)
  – Ex: MaxEnt, Naïve Bayes, SVM, …

• System ➔ LING 573 (spring)

• LING 575 (winter, spring)

• Thesis or internship (starting ASAP)
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
LING 572 (cont)

- Textbook: none

- Website: http://courses.washington.edu/ling572/

- More math in ling572:
  - Information theory:
    - Entropy, mutual information
    - Chapter 2 in (Manning & Schutze, 1999)
  - Calculus, derivative of f(x), lagrange multipliers

- Remember to register 2 weeks before the quarter starts, especially for the online option.
Beyond Ling572

• Advanced machine learning algorithms:
  – Graphical models
  – Bayesian
  – Joint inference
  – Sampling: e.g., Gibbs sampling
  – ....

• More applications:
  – Speech
  – MT
  – IR
  – QA
  – ...
Fei’s LING 575 in winter

- Website: [http://courses.washington.edu/ling575x/](http://courses.washington.edu/ling575x/)

- Topic: TBA

- Once the website is updated, I will email the CLMS mailing list. If you are not on the mailing list, you can check the website in late Dec.
Course evaluation

• The url is at the syllabus page: https://uw.iasystem.org/survey/150533

• Deadline for completing the evaluation: Dec 12