

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
- Viterbi algorithm

Unit #3: Classification

- Hw8 - Hw9
- Classification problem
- Sequence labeling and beam search
- Example tasks:
 - POS tagger
 - NE tagging
 - Chunking
- Other topics: IE

Unit #4: NN

- Hw10 – Hw11
- Word embedding
- Activation functions and backpropagation
- word2vec

Main techniques

Main techniques (1)

- Probability theory:

- The chain rule:

$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2|X_1) \dots P(X_n|X_1, \dots, X_{n-1})$$

- The Bayes' rule: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

- The (conditional) independence assumption

$$P(X_n|X_1, \dots, X_{n-1}) = P(X_n|X_{n-k}, \dots, X_{n-1})$$

- ...

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

Main techniques (5)

- Neural network:
 - Representing a word as a vector
 - Hidden layers
 - Non-linear activation functions
 - Iterative training
 - Backpropagation

Tools created

- English Tokenizer with RegEx: Hw1
- FSA and FST acceptors: Hw2 and Hw3
- Morphological analyzer with FST: Hw4
- LM, HMM, and Viterbi: Hw5-Hw7
- Form feat vectors and MaxEnt tagger: Hw8-H
- Word similarity task: Hw10
- Using existing packages:
 - Carmel: Hw2 -- Hw3
 - Mallet: Hw8 -- Hw9
 - Word2vec: Hw11

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, NN, ...
- 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 weeks
 - MaxEnt (*)
 - SVM (**)
- Unit #3: sequence labeling problem
 - 2 weeks
 - TBL (if time permits)
 - CRF (**)
- Unit #4: NN: 1-2 weeks

LING 572 (cont)

- Textbook: none
- More math in ling572:
 - Information theory:
 - Entropy, mutual information
 - Chapter 2 in (Manning & Schutze,1999)
 - Calculus, derivative of $f(x)$, gradient, 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
 - ...

Course evaluation

- The urls are at the syllabus page:
 - Fei: <https://uw.iasystem.org/survey/182074>
 - Ajda: you should have received an email
- Deadline for completing the evaluation:
Dec 14

LING570 due date

- Hw10: today
- Hw11: next Thurs (12/14)