COMP90042

SUBJECT EXAM REVIEW

PREPROCESSING

- Sentence segmentation
- Tokenization
- Word normalization
  - Derivational vs. inflectional morphology
  - Lemmatisation vs. stemming
- Stop words

TEXT CLASSIFICATION

- Building a classification system
- Evaluation metrics
- Algorithms
- Text classification tasks

PART OF SPEECH TAGGING

- English parts-of-speech
- Tagsets
  - **not:** fine-grained tags of any particular tagset
- Approaches

CONTEXT-FREE GRAMMARS

- Basic syntax of English
- The context-free grammar formalism
- Parsing
  - CYK
  - Earley

LEXICAL SEMANTICS

- Lexical relationships (-*nyms*)
- Structure of WordNet
- Similarity metrics
- Approaches to Word Sense Disambiguation

DISTRIBUTIONAL SEMANTICS

- Matrices for distributional semantics
- Association measures
  - Calculating (P)PMI from a co-occurrence matrix
- Dimensionality reduction
  - Basics of singular value decomposition (SVD)
- Cosine similarity

N-GRAM LANGUAGE MODELS

- Derivation
- Smoothing techniques
  - Add-k
  - Interpolation vs. backoff
  - Absolute discounting
  - **not:** continuation counts
- Perplexity

INFORMATION EXTRACTION

- Named entity recognition
  - Models
  - Tagging formalisms (BIO)
- **not:** relation extraction
- **not:** event extraction

QUESTION ANSWERING

- Major approaches
- Information Retrieval QA pipeline
  - Passage retrieval
  - Answer extraction

DISCOURSE

- Discourse segmentation
  - TextTiling algorithm
- Discourse parsing
  - Rhetorical Structure Theory
  - Discourse markers
- Anaphor resolution
  - Antecedent restrictions and preferences
  - **not:** Centering algorithm

SEQUENCE MODELS FOR TAGGING

- Markov Models vs Hidden Markov Model
    - mathematical formulation of HMM, assumptions
- Training on fully observed data, e.g., tagging
- Viterbi algorithm

PROB. CFGS

- Ambiguity in grammars
- Probabilistic context free grammars: rules, generative process, probability of a tree
- PCYK algorithm for parsing
- Comparing to Viterbi and other ‘decoding’ methods

DEPENDENCY GRAMMAR

- Notion of dependency between words
- Dependency grammars and dependency parse trees
- Projectivity vs non-projectivity
- Transition based parsing algorithm

WORD VECTOR LEARNING

- Formulation as term-term matrix
- Models
  - skip-gram
  - CBOW
- Training algorithm (**not:** training tricks like negative sampling)
- Evaluation tasks and general uses elsewhere

INFORMATION RETRIEVAL FOUNDATIONS

- Boolean retrieval
  - Posting list intersection
- TF*IDF weighting, components
  - Cosine similarity
- Efficient indexing
- Querying algorithm
- Evaluation metrics & resources

BM25 AND LMS

- BM25 formulation, components
- Language model formulation
- Smoothing
- Relating BM25 and LMs to other models
  - TF*IDF in IR
  - LMs in NLP

INDEX COMPRESSION

- Motivation for posting list compression
- Use of gaps between document ids
  - vbyte encoding
  - opt-p-for-delta encoding
- **not:** details of WAND beyond high level overview

WEB AS A GRAPH

- Importance of hyperlinks in web retrieval
- Graph properties
- PageRank algorithm
- HITS algorithm

MACHINE TRANSLATION

- Motivation
- Word alignment with IBM model 1
  - **not:** mathematical derivation of alignment posterior
- Phrase based model; stack decoding
  - **not:** mathematical details of sequence to sequence models
- Evaluation
  - manually vs automatically using WER, BLEU
  - learning translation metrics and evaluating metrics
  - task based “quality estimation”

EXAM STRUCTURE

- Worth 50 marks
- Parts:
  - A: short answer [10]
  - B: method questions [14]
  - C: algorithm questions [18]
  - D: short essay [8]
- 2 hours in duration

  … 2 minutes 24 seconds / mark

SHORT ANSWER (10 MARKS)

- Several short questions
  - 1-2 sentence answers for each
  - 1 mark per question
- Often
  - definitional, e.g., *what is X?*
  - conceptual, e.g., *relate X and Y? What is the purpose of Z?*
  - may call for an example illustrating a technique/problem

METHOD QUESTIONS (14 MARKS)

- Longer answer
  - larger questions 5 or 6 marks each
  - broken down into parts
- Focus on analysis and understanding, e.g.,
  - contrast different methods
  - outline or analyze an algorithm
  - motivate a modelling technique
  - explain or derive mathematical equation

ALGORITHMIC QUESTIONS (18 MARKS)

- Perform algorithmic computations
  - numerical computations for algorithm on some given example data
  - present an outline of an algorithm on your own example
- 3 Questions (longer this year than in the past)
- Each question worth 5-7 marks.
- You won’t be required to simplify maths, i.e., you can leave things as fractions; and will be given table of useful numbers

ESSAY QUESTION (8 MARKS)

- Expect to write 1 page
- Several broad topics in WSTA given, you should select **one**
  - no marks given for attempting many
- Provide
  - Definition and motivation
  - Relation to multiple tasks discussed in the class
  - Compare/contrast use across these tasks

WHAT TO EXPECT

- In proportion to lectures, i.e.,
  - 25% information retrieval / web search
  - 75% text analysis
- Greater focus on concepts that have not yet been assessed by homework / project
  - e.g., increased focus on IR components
- Guest lectures are *fair game*
- Prescribed reading is *fair game*