Exam review

COMP90042 Lecture 24

Preprocessing

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

Information retrieval foundations

- "Information need"
- TF*IDF weighting, components
  - Cosine similarity
- Efficient indexing
- Querying algorithm

IR indexing and querying

- Posting list compression
  - Use of gaps between document ids
    - vbyte encoding
    - opt-pfor-delta encoding
- WAND algorithm
- Index construction: static vs incremental
- Phrase search
  - positional index (intersection, extra information etc.)
  - **NOT** suffix array

IR Querying, Evaluation and L2R

- Query completion
  - trie+RMQ algorithm
  - Motivation, Data sources
- Relevance feedback (why, types)
- Evaluation methods
  - precision @ k, (Mean)AveragePrecision, RBP
  - research test collections
- Reranking IR system outputs using learned classifier

Text Classification

- Building a classification system
- Evaluation metrics
- Algorithms
- Text classification tasks
  - including learning-to-rank in IR

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
- Count-based models
  - Basics of singular value decomposition (SVD)
- Predict-based models
  - Skip-gram, CBOW
- Cosine similarity

Part of speech tagging

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

Information extraction

- Named entity recognition
  - Models
  - Tagging formalisms (BIO)
- Relation extraction:
  - How to frame the problem using binary and multi-class classifiers
- Differences between supervised models and OpenIE.

N-gram Language models

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

RNN models

- Basics of neural network structure
- How to frame LM/tagging as a word-by-word classification task
  - feed-forward classifiers vs recurrent neural networks
- Similarity with seq2seq as used in MT
- **not:** mathematical details of formulation

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

Grammars and Languages

- Finite state automata and transducers
  - relationship to n-gram LMs, HMMs
  - Chomsky hierarchy
- Basic syntax of English
  - **not:** detailed nuances of grammar (see Q9 from 2017)
- The context-free grammar formalism
- Parsing
  - CYK 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
- **not**: graph based parsing
- **not:** detailed dependency edge inventory

Question Answering

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

Discourse

- Motivation for modelling larger documents
- Discourse segmentation with TextTiling
- Notion of RST parsing of documents
  (at high level; **not** edge label inventory)
- Anaphora resolution, and the centering algorithm

Machine translation

- Motivation
- Word alignment with IBM model 1
  - **not:** mathematical derivation of alignment posterior
- Phrase based model; stack decoding algorithm
- Sequence to sequence model
  - **not:** mathematical formulation
- Evaluation

Exam Details

Structure and format — closely following last year's

Exam Structure

- Worth 50 marks
- Parts:
  - A: short answer [15-20]
  - B: method questions [15-20]
  - C: algorithm questions [10]
  - D: short essay [8]
- 2 hours in duration

  … 2 minutes 24 seconds / mark
- Download recent years' exams from the University library

Short answer

- 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

- Longer answer
  - larger questions 5-7 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

- Perform algorithmic computations
  - numerical computations for algorithm on some given example data
  - present an outline of an algorithm on your own example
- 2 questions, each worth 4-6 marks.
- You won't be required to simplify maths, i.e., you can leave things as fractions, e.g., log(5/4)

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

- Even coverage of topic from the semester
- Be prepared for concepts that have not yet been assessed by homework / project
- Lecture 23 is *out of scope*
- Prescribed reading is *fair game* for topics mentioned in the lectures and workshops