Foundations of Natural Language Processing
Lecture 18
Wrapup, review, and exam information

Alex Lascarides

20 March 2020

School of informatics

WARNING: this isn't the same course it was in 2015 and before

When revising for the exam, past exam papers are useful, but be careful of overfitting.

- Most topics in common with last year
- Changed a lot in 2016
- Different topics; some new approaches/models

Topics in common with previous years

- Corpora, annotation, evaluation
- Ambiguity at all levels
- N-gram models, entropy, smoothing
- Noisy channel framework
- Spelling correction, edit distance
- HMMs, part-of-speech tagging
- Syntax, parsing algorithms, PCFGs, other grammar formalisms
- Lexical semantics: word senses

Eliminated from previous years

You will not be expected to answer questions about these topics.

- corpus markup
- mathematical details of backoff in N-gram models
- details of forward-backward algorithm for HMMs
- feature structure grammars
- crowdsourcing in detail
- implementation details of Good-Turing smoothing
- pronoun resolution
- machine translation

New since 2015

So past papers are not a good guide for these!

- Updated discussion of evaluation
- High-level overview of more modern smoothing methods (K-N)
- More complete example of spelling correction (end-to-end system)
- Generalized discussion of EM (showing application in both spelling correction and HMMs)
- Text classification (tasks and methods)
- Dependency grammar and related algorithms (e.g., SR parsing)
- Semantic roles and distributional semantics

Format of the exam

As in previous years, the exam has two parts:

- Part A: 8 short-answer questions, each worth 3 marks (total of 24 marks).

- Part B: 3 longer questions worth 13 marks each, of which you must answer **two** (total of 26 marks).
  - Be clear which questions you are answering.
  - If you (start to) answer **more** than two, you **must** clearly cross out one answer.
  - Part A and each question for part B in **separate booklets** (i.e., 3 booklets overall).

What counts and what doesn't

Things that **do** matter (not necessarily a complete list):

- Complete answer (double check you've answered everything that was asked!)
- Clear explanations/reasoning where appropriate
- Correct equations, all variables defined
- Legible

What counts and what doesn't

Things that **do not** matter:

- Perfect spelling/grammar/handwriting: as long as it is **clear what you mean**. Do not waste time writing drafts/copying over, but clearly cross out any scratch work that should not be marked. You can lose marks for have both correct and incorrect answers unless one is crossed out.

- Full sentences. If a word or short phrase conveys the meaning, no need for more.

Other ways to prepare

- Lecture summary slides are a good place to start: they don't have all the details, but make sure you understand the details underlying the main points mentioned.
- Do the labs! Make sure you understand the answers you get
- Heed any feedback on your courseworks and talk to your classmates or post on Piazza if you still don't understand.
- Post questions on Piazza. We will not always answer immediately but will try to ensure questions are answered. **Exception:** we will **not** answer any questions asked less than 48 hours before the exam.

What courses follow on next year?

- IAML: if you haven't already taken it, do! ML underlies most of NLP, and fourth year courses assume a strong background.
- Natural Language Understanding, Generation and MT (NLU+): more advanced models and algorithms for processing syntax, semantics, and discourse.
- Text Technologies: focus on search engines, IR, working systems.
- Automatic Speech Recognition: builds on knowledge from this course, but focuses on speech processing.

Other related courses

- Other machine learning courses (MLPR, MLP, PMR): These cover modern statistical approaches and deep learning models that are increasingly popular in NLP.

- Extreme computing: for dealing with huge data sets.

- Computational Cognitive Science, Topics in Cognitive Modeling: include sections on computational models of human language processing.

That's all folks!