

Midterm Review

L245

For the Midterm on Monday, March 6, 2017

1 Topics to be covered

1. Text & Speech encoding
2. Writers' aids
3. Language Tutoring Systems
4. Searching (~~up through slide #24 i.e., up to but not including semi-structured data up through slide #19 i.e., up to but not including weblinking~~)

2 Format of the exam

You will have the entire 75 minutes (2:30–3:45pm) should you need or want it. FYI: in past semesters, it seems that time has been the biggest challenge.

1. Matching: 5–10 terms (see list below)
2. “Calculations” (relatively closed form questions): 5–10 questions
 - Binary numbers (different bases), ASCII encoding
 - Transliteration (converting between writing systems)
 - N -gram language modeling
 - Bigram array (positional and non-positional)
 - Similarity key calculations
 - Minimum edit distance
 - Noisy Channel Model
 - Conditional probabilities
 - Bayes' Law
 - ~~Confusion matrix (using & representing)~~
 - Bigram/Trigram real-word spell checkers (potentially using confusion sets)
 - Tokenization
 - Analysis of learner language (e.g., POS evidence)
 - Boolean expressions
 - Search engine indexing

- Weblinking & webpage ranking
3. Short answer/Essay: something like: “answer 3 out of 5”
- Types of writing systems, pros & cons
 - Relation of writing systems to languages
 - Types of character encoding systems, e.g., ASCII & Unicode
 - Challenges of ASR & TTS
 - How measurements do & do not correspond to what we hear
 - Types and causes of spelling errors
 - Context-sensitive spelling correction for web queries
 - Error models & language models for spelling correction
 - Designing n -gram grammar correctors
 - Syntactic rules, syntactic trees, parsing, & grammar correction rules
 - Using NLP in CALL (e.g., parsing ill-formed input)
 - Parser-based ICALL (e.g., system design)
 - Learner modeling
 - Authentic-text ICALL
 - Selecting features for ICALL-related machine learning
 - Structured vs. unstructured information: searching in databases vs. on the web
 - How search engines work (indexing, weblinking, etc.)

3 Some terms/concepts to know

3.1 Text/Speech encoding

- | | | |
|------------------------------|------------------------------|--------------------------------------|
| – alphabet | – Unicode | – spectrogram |
| – abjad | – transcription | – Automatic Speech Recognition (ASR) |
| – abugida | – phonetic alphabet | – Text-to-Speech Synthesis (TTS) |
| – syllabary | – coarticulation | – acoustic signal processing |
| – diacritic | – articulatory phonetics | – diphone |
| – logograph | – sampling rate | – n-gram |
| – pictograph | – continuous & discrete data | – word prediction |
| – ideograph | – Hertz | – unigram, bigram, trigram, ... |
| – semantic-phonetic compound | – sound wave | |
| – bit & byte | – amplitude | |
| – ASCII | – frequency | |

3.2 Writers’ aids

- interactive spelling checker
 - automatic spelling corrector
 - non-word error detection / word recognition
 - domain-specificity
 - tokenization (word segmentation)
 - inflection
 - productivity of language
 - (positional or non-positional) bigram array
 - isolated-word error correction
 - run-on error
- split error
 - phonetic error
 - homophone
 - insertion, deletion, substitution, transposition
 - minimum edit distance
 - acyclicity
 - topological ordering
 - dynamic programming
 - noisy channel model
 - Bayes' Rule
 - confusion matrix
 - context-dependent word correction
 - grammar checker
- local syntactic error
 - long-distance syntactic error
 - semantic error
 - error pattern
 - syntax
 - linear order
 - constituent
 - lexical & phrasal categories
 - phrase structure rule
 - (structural) ambiguity
 - recursion
 - parsing
 - top-down & bottom-up parsing

3.3 Language Tutoring Systems

- foreign language teaching (FLT)
 - native speaker
 - language awareness
 - second language acquisition (SLA)
 - cloze (fill-in-the-blank) exercise
 - fallback case (canned text response)
 - frame-based system
 - named entity recognition
 - lexical semantic relations
 - synonymy
 - lemmatization
 - covering & overlapping ambiguity
 - meta-linguistic knowledge
- distribution, morphology, & lexical stem lookup
 - inflectional & derivational suffixes
 - ill-formed input
 - mal-rule
 - modularity
 - demand-driven architecture
 - learner modeling
 - L1-transfer
 - sequencing of teaching material
 - concordance (KWIC)
 - grammatical error detection
 - machine learning/classifiers
 - feature (vector)

3.4 Searching

- database (frontend)
 - stop word
 - querying
 - boolean expression
 - structured data
 - unstructured data
 - information need
- meta tag
 - stemming
 - index
 - term-by-document matrix
 - inverted index
 - relevance
 - click-through measurement