Midterm Review

L245

For the Midterm on Tuesday, March 8, 2016

1 Topics to be covered

1. Text & Speech encoding
2. Writers’ aids
3. Language Tutoring Systems
4. Searching (up through slide #7—i.e., up to but not including unstructured data)

2 Format of the exam

You will have the entire 75 minutes (1:00–2:15pm) should you need or want it.

1. Matching: 5-10 terms (see list below)
2. “Calculations” (relatively closed form questions): 5–10 questions
   a. Binary numbers (different bases), ASCII encoding
   b. Transliteration (converting between writing systems)
   c. N-gram language modeling
   d. Bigram array (positional and non-positional)
   e. Similarity key calculations
   f. Minimum edit distance
   g. Noisy Channel Model
      i. Conditional probabilities
      ii. Bayes’ Law
   h. Confusion matrix (using & representing)
   i. Bigram/Trigram real-word spell checkers (potentially using confusion sets)
   j. Tokenization
   k. Analysis of learner language (e.g., POS evidence)
   l. Boolean expressions
3. Short answer/Essay: something like: “answer 3 out of 5”
   a. 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

3 Some terms/concepts to know

3.1 Text/Speech encoding

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

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