Midterm Review

L245

For the Midterm on Tuesday, March 10, 2015

1 Topics to be covered

1. Text & Speech encoding
2. Writers’ aids
3. Language Tutoring Systems
4. Searching (up through slide #23—i.e., up to but not including semi-structured data)

2 Format of the exam

You will have the entire 75 minutes (2:30–3:45pm) should you need or want it.

1. Matching: 5-10 terms (see list below)
2. “Calculations” (relatively closed form questions): 5–10 questions
   - Binary numbers (different bases), ASCII encoding
   - n-gram language modeling
   - Precision/Recall
   - Similarity key calculations
   - Minimum edit distance
   - Bigram array (positional and non-positional)
   - Confusion matrix
   - Noisy Channel Model
     - Conditional probabilities
     - Bayes’ Law
   - Bigram/Trigram real-word spell checkers (using confusion sets)
   - Tokenization
   - POS evidence for learner language
   - 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
- Context-free grammars
- Syntactic trees & parsing
- Second Language Acquisition (SLA)
- Using NLP in CALL
- Authentic-text ICALL
- Parser-based ICALL
- Learner modeling & CALL exercise types (and motivation for them)
- Selecting features for ICALL-related machine learning
- Structured vs. unstructured information
- How search engines work (indexing, weblinking, etc.)

3 Some terms/concepts to know

3.1 Text/Speech encoding

- alphabet
- abjad
- abugida
- syllabary
- diacritic
- pictograph / pictogram
- ideograph / ideogram
- semantic-phonetic compound
- bit & byte
- ASCII
- Unicode
- information theory
- transcription
- phonetic alphabet
- coarticulation
- articulatory phonetics
- sampling rate
- continuous & discrete data
- Hertz
- sound wave
- speech flow
- loudness / 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
- error detection & error correction
- run-on error
- split error
- phonetic error
– homophone
– tokenization (word segmentation)
– inflection
– productivity of language
– non-word error detection / word recognition
– dictionary construction & lookup
– domain-specificity
– (positional or non-positional) bigram array
– isolated-word error correction
– 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
– syntax
– linear order
– constituent
– lexical & phrasal categories
– phrase structure rule
– (structural) ambiguity
– recursion
– context-free grammar
– top-down & bottom-up parsing

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

3.4 Searching
– database (frontend)
– stop word
– literal string
– querying
– boolean expression
– search engine
– meta tag
– index
– stemming
– term-by-document matrix
– inverted index
– relevance
– click-through measurement
– structured data
– unstructured data