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 #14—i.e., up to but not including how search engines work)

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
   • 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
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

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
<table>
<thead>
<tr>
<th>Run-on error</th>
<th>Split error</th>
<th>Phonetic error</th>
<th>Homophone</th>
<th>Insertion, deletion, substitution, transposition</th>
<th>Minimum edit distance</th>
<th>Acyclicity</th>
<th>Topological ordering</th>
<th>Dynamic programming</th>
<th>Noisy channel model</th>
<th>Bayes’ Rule</th>
<th>Confusion matrix</th>
<th>Context-dependent word correction</th>
<th>Grammar checker</th>
<th>Local syntactic error</th>
<th>Long-distance syntactic error</th>
<th>Semantic error</th>
<th>Error pattern</th>
<th>Syntax</th>
<th>Linear order</th>
<th>Constituent</th>
<th>Lexical &amp; phrasal categories</th>
<th>Phrase structure rule</th>
<th>Structural ambiguity</th>
<th>Recursion</th>
<th>Parsing</th>
<th>Top-down &amp; bottom-up parsing</th>
</tr>
</thead>
</table>

### 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
- Search engine
- Meta tag
- Stemming
- Index
- Term-by-document matrix
- Inverted index
- Relevance
- Click-through measurement