Midterm Review

L245

For the Midterm on Thursday, October 17, 2013

1 Topics to be covered

1. Text & Speech encoding
2. Writers’ aids
3. Language Tutoring Systems
4. Searching

2 Format of the exam

You will have the entire 75 minutes (9:30-10:45am) should you need/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
   - Regular expressions
   - Search engine indexing
   - Weblinking & webpage ranking
- Finite-state automata

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
- 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
- Searching in databases vs. on the web vs. in a corpus
- 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
- ideographs / ideogram
- semantic-phonetic compound
- bit & byte
- ASCII
- meta-information
- Unicode
- information theory
- transcription
- phonetic alphabet
- coarticulation
- articulatory phonetics
- sampling rate
- continuous & discrete data
- Hertz
- sound wave
- speech flow
- loudness / amplitude
- frequency
- pitch
- intonation
- 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
- inflection
- productivity of language
- non-word error detection / word recognition
- dictionary construction & lookup
- domain-specificity
- (positional or non-positional) bigram array
- insertion, deletion, substitution, transposition
- confusion matrix
- Noisy Channel Model
- Bayes’ Rule
- minimum edit distance
- topological ordering
- dynamic programming
- context-dependent word correction
- grammar checker
- part of speech
- syntax
- linear order
- constituent
- lexical & phrasal categories
- phrase structure rule
- (structural) ambiguity
- recursion
- context-free grammar
- non-terminal & terminal elements
- pushdown automaton
- top-down & bottom-up parsing

3.3 Language Tutoring Systems

- language awareness
- second language acquisition (SLA)
- concordance
- cloze exercise
- fallback case
- canned text response
- frame-based system
- named entity recognition
- lexical semantic relations
- synonymy
- hyponymy/hypernymy
- tokenization
- 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
- machine learning/classifiers
- grammatical error detection
- machine learning
- 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
– semi-structured data
– unstructured data
– regular expression
– formal language (theory)
– corpus