Ling/CSE 472: Introduction to Computational Linguistics

4/27/15

Text-to-Speech

Overview

- TTS demo (oddcast)
- Evaluation of TTS systems
- TTS system high-level overview
- Intermediate representation
- Sub-components
- Reading questions

TTS demo

- http://www.oddcast.com/home/demos/tts/tts_example.php?sitepal

Evaluation of TTS

- How can we evaluate TTS systems?
- What are the dimensions on which it should be evaluated?
- What questions would we ask humans about TTS output in order to do that evaluation?

Evaluation of TTS

- Intelligibility:
    - Diagnostic Rhyme Test/Modified Rhyme Test: Present words out of context in bland carrier phrases and ask speakers which it was (sets of 2 or more)
        - Now we will say <word> again
- Semantically unpredictable sentences test:
    - The unsure steaks closed the fish. (D A N V D N)

Evaluation of TTS

- Quality:
    - Mean opinion score:
        - Ask multiple listeners to rate sentences from a system on a scale of 1-5
        - Compare MOS on the same sentences for different systems
- AB score:
    - Same sentences from two different systems
    - Ask listeners to choose which one is better in each case

Evaluation of TTS components

- Measures of intelligibility and quality:
    - *intrinsic* evaluation of system
    - *extrinsic* evaluation of system components

- For each component, we can also think of *intrinsic* evaluation metrics
    - In many cases, these will be more easily automated

High-level overview

- What's the input?
- What's the output?
- Is this an analysis or generation task?

Intermediate representation

- TTS is facilitated by positing an intermediate level of representation
- Effectively breaks the TTS process into two major steps:

Text: standard(-ish) orthography

↓

Intermediate representation

↓

Wave form

Intermediate representation

- What information do we need to specify in the intermediate representation?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Sentence and word segmentation
“Text normalization”

- What’s the input?
- What’s the output?
- What other sources of information can we use?
- In what cases is this task difficult?

Sentence and word segmentation
“Text normalization”

- What’s the input? -- Text in standard(-ish) orthography
- What’s the output?
    - Sentences, each of which consist of
    - Words, each of which is
    - Spelled out (e.g., in case of “non-standard” words)
- What other sources of information can we use?
    - Lists of sentence-ending punctuation & abbreviations
    - Training data: Sentence-segmented text
    - Lists of non-standard words
    - Lists of non-standard word spell-out rules (can be context-dependent)
    - POS tagger
- In what cases is this task difficult?
    - Non-standard words
    - Double-duty punctuation (haplology)

Evaluation: Text normalization

- How do we evaluate text normalization components?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Phones

- What's the input?
- What's the output?
- What other sources of information can we use?
- In what cases is this task difficult?

Phones

- What's the input? -- Sentence and word segmented text
- What's the output? -- Same text with phones aligned to each character (many-to-many alignment)
- What other sources of information can we use?
    - Pronunciation dictionary
    - grapheme-to-phoneme rules
    - Training data with transcriptions (and alignments)
- In what cases is this task difficult?
    - Unknown words: names and non-names
    - Homophones

Evaluation: Phones

- How do we evaluate mapping to phone sequences?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Prosodic structure

- What's the input?
- What's the output?
- What other sources of information can we use?
- In what cases is this task difficult?

Prosodic structure

- What's the input? -- Sentence
- What's the output? -- Sentence with prosodic boundaries marked
- What other sources of information can we use?
  - Marked up training data
  - Features: length of phrases, neighboring POS and punctuation, syntactic parse features
- In what cases is this task difficult?
- How do we evaluate the output of this component?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Prosodic prominence

- What's the input?
- What's the output?
- What other information can we use?
- In what cases is this difficult?

Prosodic prominence

- What's the input? -- Sentence with all mark-up so far
- What's the output? -- Same, plus prominence level (emphatic accent, pitch accent, unaccented, reduced; or two-way system)
- What other information can we use?
    - Word frequencies
    - TF-IDF
    - Stress patterns in sentence
- In what cases is this difficult?
    - When knowledge of information structure is critical

Evaluation: Prosodic prominence

- How do we evaluate predictions of prosodic prominence?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Intonation Contour: Tune

- What's the input?
- What's the output?
- What other sources of information can we use?
- In what cases is this task difficult?

Intonation Contour: Tune

- What's the input? -- Sentence with all mark up so far
- What's the output? -- Same, plus marking of boundary tones
- What other sources of information can we use?
    - ToBI annotation of sample text
    - Rules or learned patterns based on the above
- In what cases is this task difficult?
    - Any time anything other than a basic intonation contour is called for
    - Intonation expressing emotion

Evaluation: Intonation contour

- How do we evaluate predictions of boundary tones?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Phone duration

- What's the input?
- What's the output?
- What other sources of information can we use?
- In what cases is this task difficult?

Phone duration

- What's the input? -- Sequence of phones, plus prosodic structure, plus pitch accent locations plus boundary tones

- What's the output? -- Durations (in ms) for each phone

- What other sources of information can we use? -- Hand-written rules or machine learning features based on generalizations such as:
  - Vowels or syllabic consonants before pauses are longer
  - Vowels bearing an accent are longer

- In what cases is this task difficult?

Evaluation: Phone duration

- How do we evaluate predictions of phone length?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

F0

- What's the input? -- Sentence with prosodic information
- What's the output?
  - F0 target points for each pitch accent
  - Boundary tone, contour connecting those points
    - Exact location w/in accented syllable for each target
- What other sources of information can we use?
  - Pitch range
  - Rules for declination, downstep
- In what cases is this task difficult?

Evaluation: F0

- How do we evaluate predictions of F0?

Intermediate representation: Components

- Sentence and word segmentation
- Phones
- Syllable boundaries
- Suprasegmental prosodic structure (intonation phrases)
- Location of pitch accent
- Intonation contour
- Length of phones
- F0

Intermediate representation

- TTS is facilitated by positing an intermediate level of representation
- Effectively breaks the TTS process into two major steps:

Text: standard(-ish) orthography

↓

Intermediate representation

↓

Wave form

Making the wave form: Diphone database

- diphone: A recording of the transition between one phone and the next, starting halfway through the first and ending halfway through the second

- Create a diphone inventory, with text for each one

    - pause t aa **b aa** m aa pause

- Recruit a speaker (voice talent)

- Record the speaker saying each diphone

- Segment, label, and pitch-mark the diphones

- Excise the diphones

Making the wave form: Diphone concatenation

- Putting diphones together willy-nilly leaves lots of artifacts
- Must at least:
    - Match pitch epochs
    - Change pitch
    - Lengthen diphones

Making the wave form: Unit selection (Alternative approach)

- Have the voice talent record a much larger database
  - Multiple copies of each diphone, in different environments
  - Larger segments that could be used
- Given the input (with all mark-up added to the intermediate representation), find the best sequence of stored units (Viterbi)
  - Target cost: how well the target specification matches the potential unit
  - Join cost: how well that potential unit joins with its potential neighbor

Overview

- TTS demo (oddcast)
- Evaluation of TTS systems
- TTS system high-level overview
- Intermediate representation
- Sub-components
- Reading questions

Reading questions

- In the sentence "ANLP Corp. chairman Dr. Smith resigned." there are three period, and when we analyze them with feature templates, we get:

- For the first period:
  - PreviousWord = ANLP; NextWord = chairman;
  - Prefix = Corp; Suffix = NULL;
  - PreviousWordAbbreviation = 1; NextWordAbbreviation = 0;
- For the second period:
  - PreviousWord = chairman; NextWord = Smith;
  - Prefix = Dr; Suffix = NULL;
  - PreviousWordAbbreviation = 0; NextWordAbbreviation = 0;
- For the first period:
  - PreviousWord = Smith; NextWord = NULL;
  - Prefix = resigned; Suffix = NULL;
  - PreviousWordAbbreviation = 0; NextWordAbbreviation = 0;

Reading questions

- Those data seem all different, but the machine should process the first period and second period in the same way, while the third period in a different way. So how does the machine make sure if the previous word is an abbreviation or not, why those feature template data are all necessary, and how exactly these feature templates help tokenization?

Reading questions

- When joining diphones, how fine-grained is the operation? Are we working on the scale of phonemes, or some smaller phonological unit?

- The book said that modern commercial synthesizers use unit selection synthesis rather than diphone synthesis; is diphone synthesis used in any modern applications?

- Diphone waveform synthesis vs. Unit selection waveform synthesis: I have a few questions relating these two forms of sound synthesis. Is one better than the other, or do they have strengths and weaknesses that make one better at certain sounds than other? Would you be able to tell from the output which form of synthesis was used (even if you had to take specific examples to exaggerate the differences)? Is there a significant difference in processing power, time, or output? I guess I am just wanting a general comparison of the two forms of synthesis.

Reading questions

- Are voice talents necessary to build databases of diphones and units for all speech synthesis? Can diphones and units be coded/engineered from scratch?

- For concatenating diphones, windowing function is applied to the edge of diphones. After reading the section about windowing, I'm still confused how it works and how we can measure how close to zero amplitude we should get between each two diphone, since it is highly dependent on context of that text.

- Why was concatenative synthesis chosen by the industry over formant synthesis and articulatory synthesis?

Reading questions

- I thought it was very interesting how stress was determined for the sentences; when reading this one of my first thoughts was how we can handle the case of a sentence like "You gave him money?", where stress on any one word (or none) will change the nuance of the sentence, into 4 different readings. I thought it was interesting that they can use analysis to determine that often, it is the least probable word that receives the stress. However as it later says, in cases like this if there is not enough context or information it will chose the most standard pronunciation; I wonder if increasing training data will produce better results, or if a problem of ambiguity like this may never really be able to be solved?

Reading questions

- I wonder how the difficulty of capturing "tune" varies by language; languages that are either tonal or have regular patterns of sentential stress seem to me to have less variability in prosody and tone, but I don't know if this is just because my ears are not trained to hear tune and prosody in these other languages. Does English rely on use of prosody and tune more than average?

- In the reading, Figure 8.9 listed the accent and boundary tones labels from the ToBI transcription system for American English intonation. What is an example of L-H% continuation rise boundary tones and H-L% final level plateau boundary tones?

Reading questions

- The text notes that typical text-to-speech systems sound wooden because they don't account for discourse, but are there any used out in the world that approach the level of robustness talked about in the chapter? For example, is there a speech synthesis system used by non-researchers that somehow tracks discourse in order to supply accurate stress and pitch accent?

- Given the long list of parts that can go into a speech synthesizer (POS tagger, n-gram model(s), pronunciation dictionary, rules for categorizing NSWs, abbreviation dictionaries, grapheme-to-phoneme conversion, etc) it seems like speech synthesizers would be relatively large and slow. How are they kept fast enough to be practical and small enough to be put on less powerful devices?

Reading questions

- The book talks about how part of speech tagging can be used to produce more natural pronunciations of words in text-to-speech. Is the opposite also true? For speech-to-text, is the pronunciation of the word used to infer the tag? What is done differently in part of speech tagging when you're dealing with spoken language instead of written language?

- What are bprob and eprob in the period decision tree (figure 8.3)?

Reading questions

- Are speech to text programs ever used to evaluate speech synthesizers instead of humans?