Ling/CSE 472: Introduction to Computational Linguistics

4/18/17

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

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

- "Thus short, all-capital words(IBM, US) might be LSEQ, longer all-lowercase words with a single-quote (gov't, cap'n) might be EXPN, and all-capital words with multiple vowels (NASA, IKEA) might be more likely to be ASWD." If a program used this rule to distinguish LSEQ from ASWD, without a dictionary, what are the common exceptions that cannot be described by this rule? How about the accuracy?

- Also, has anyone tried to use machine learning to handle expanding abbreviations (EXPN)? The book says that abbreviation dictionaries are usually used, but it seems to me that it might be feasible to recover a full word from its abbreviation, given context (otherwise, the abbreviation probably isn't very good).

Reading questions

- I would be curious to know more about the correlation between prosodic and syntactic structures (section 8.3.1).

- The book gave an example of the difference between intonational and intermediate phrases, but didn't formalize a definition for each of them. How subjective is our understanding of what constitutes an intonational versus a intermediate phrase? I know that intonational phrases are often comprised of one or more intermediate phrases, but what makes "to go" an intermediate phrase rather than "I wanted to" or "I wanted to go"?

Reading questions

- In 8.3 it says that prosodic analysis is the final stage of linguistic analysis, so I don't understand why the phonemic representation in fig. 8.2 comes later in the hourglass metaphor. Isn't the phonemic internal representation already accounted for in the g2p process? How could the underlying forms be represented after all of the analysis on the suprasegmental level has occurred?

- I didn't get the example 8.11 on Page 259.
    - V -> [+ stress] / X __ C* {Vshort CC?|V} {Vshort C*|V}
    - ex: *difficult, oregano*

Reading questions

- Tone doesn't seem to be mentioned in this chapter. Tone can't be treated as totally separate from the individual words, like the suprasegmental features/ prosody are, but it wouldn't work to treat tone the same way as stress. I'm not sure how tone is treated in optimality theory, but I'm familiar with treating tone as an autosegmental feature that is associated with the words but is on a different level than the phonetics, etc, and I was wondering how you would include tone in speech synthesizers or recognizers?

- What are the use of Tilt model? How is its function different from ToBI?

Reading questions

- The book mentions that most synthesis systems use two classifiers to separately handle unknowns for names and other words, so that names can have additional features like language of origin. Wouldn't it make sense to use the same features for other words as well since English adapts words from multiple other languages? (This probably also helps with unknown loanwords.)

Reading questions

- As described on page 275 and illustrated in figure 8.17, to increase the pitch, the pitch-synchronous frames are extracted and moved closer together. Wouldn't it be easier to scale the entire waveform? Does this not sound as natural?

- I'm curious about what the text means by "We meet this goal by measuring the acoustic similarity of the edges of the two units that we will be joining. If the two units have very similar energy, F0, and spectral features, they will probably join well." How do you measure energy?

- This model of concatenating diphones or units seems flawed. Why not synthesize the sounds, instead of recording them? Then you're guaranteed perfect concatenation.