Rare Events and Closed Domains

SyRG Discussion 4/29/04

Brian Langner

Common issues in synthesis

- How to treat large number of rare events
  - Found in several contexts for synthesis
    - Text analysis, duration modeling, database design
- Use of “large” units
  - Only usable for limited/closed domains
  - Recombination rules have tended to be ad hoc
    - Requires significant changes for anything other than original voice and domain

LNRE

- Though individual events are unlikely to occur, there are so many of them that encountering one is extremely likely
- Thus accepting poor models for rare or unseen events degrades quality despite the events being unlikely to be encountered

LNRE in text analysis

- Productivity
  - Generally morphological, but can also refer to production of novel words
- Syllabification
  - Phoneme pronunciation often depends on position in syllable

Productivity

- General TTS is likely to contain words that do not appear in the lexicon
- These words often are formed by a productive morphological process
  - Such a process is able to produce an unlimited vocabulary
  - However, the process is regular and so can be modeled with rules (and a smaller number of explicit patterns)

Productivity

- Growth curve is dependent on sample size
  - Difficult or impossible to compare when sample sizes are not the same (or just similar)
  - Even a corpus with several million words will still encounter unknown word types
    - Such a corpus is still “in the LNRE zone”

Productivity

- Productive patterns have an indefinite growth curve, unproductive patterns have a finite vocabulary

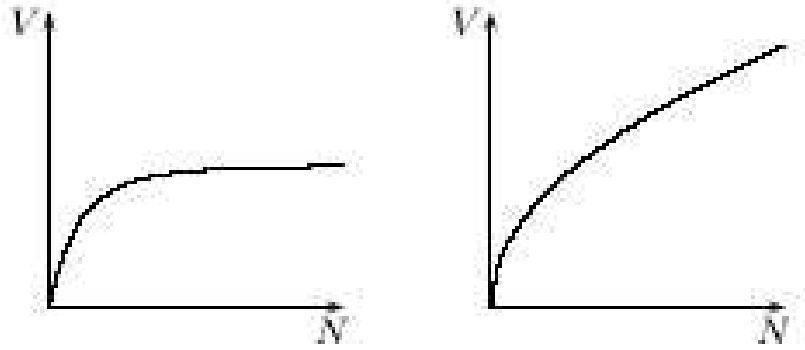

Figure 1: Typical shapes of vocabulary growth curves (V=types, N=tokens): the curve pertaining to an unproductive pattern will flatten out (left panel), whereas the vocabulary of a productive pattern will continue to grow indefinitely (right panel). Adapted from [9].

Productivity

- To provide sufficient coverage, TTS systems should use a morphological analysis component
  - Should provide capability of morphologically decomposing unknown words to get annotation similar to that of words listed in the lexicon

Syllabification

- Phoneme pronunciation often depends on position in syllable
- Syllable structure shows typical LNRE patterns
  - Only about 5% of syllable types are used regularly or systematically
  - Vast majority are encountered infrequently

Syllabification

- Possible solution to LNRE problem uses unsupervised training with unannotated data
  - Multidimensional EM clustering
  - Onset, nucleus, coda (, stress, position)
- Approach beneficial because it will assign probabilities even to syllable types not covered in training

Duration models

- Supposed to predict temporal structure of speech given symbolic input
- Can use automatic method (CART trees, neural networks) or manual construction to cover feature vectors
  - Automatic methods don’t exhaustively cover all vectors
  - Manual methods aren’t feasible for anything but small databases

Duration models

- Majority of feature vectors are infrequently observed, so this too is LNRE
- Cannot ignore or use poor models for rare vectors because any given sentence is likely to have at least one

Duration models

- Model needs to predict durations of vectors not represented in training data (likely by extrapolation)
- CART trees are bad for this because they don't work well with sparse data, and can't extrapolate
- Sums-of-products (Van Santen) far better than CART
  - Needs less training data (will perform adequately with sparse data)
  - Asymptotic performance is better
  - Performs better when training and test data are different
  - Adding more training data will improve performance

Database design

- >15% of diphones did not occur in a Beutnagel & Conkie database designed for unit selection
  - These units were only included when carefully constructed sentences were added to the database, they were not expected to naturally be in the recorded speech
  - In other databases, the number of required units can approach infinity when synthesizing general text, with most units used infrequently
- Unit selection algorithm preferred to select the rare diphones over concatenated demiphones, implying a higher quality result when using them

Database design

- No clear solution for LNRE here
  - 75-80% coverage is “good”
- Can record more (ARCTIC sized and bigger)
  - Will need very cooperative voice talent
- “Carefully define linguistic and phonetic criteria the database should meet”
  - …

Closed domains and larger units

- Improve performance for closed domains by using units larger than phone/diphone
  - Syllable, word, etc.
- Examples: Verbmobil (words), weather (words + syllables)

Weather example (Lewis & Tatham)

- Uses 2000 mono- and poly-syllabic words
- Monosyllable words are recorded in a fixed carrier phrase
  - Often inappropriate to recombine
- Recombination rules seem to be ad hoc
  - Not extensible for other voices, recording, etc.
- Problems make it unlikely this approach can be used for unrestricted TTS or even larger domains

References

- See paper