Refurbishing the cmudict database

Alex Rudnicky
(air@cs.cmu.edu)

Speech Lunch Presentation
2 March 2006
An unencumbered (public domain) pronunciation dictionary for American English, for use in speech recognition, originally assembled by Bob Weide at Carnegie Mellon University.

Currently available on the web for download:
- [http://www.speech.cs.cmu.edu/cgi-bin/cmudict](http://www.speech.cs.cmu.edu/cgi-bin/cmudict)

Many, many users over time
- “about 10,600 hits” says Google

Some confusion as to which version is the “true” dictionary
- Several seemingly similar versions in concurrent use
- Synthesis people swear that a 1996 version is the “purest”
- No clear metrics for quality
- Currently no formal process for vetting and including new entries
• Initial development was informal
• First “official” release contained ~100k words
• Some observations:
  – Sustained growth in 1994-1998; but not maintained since 2000
  – The proportion of variants started at about 3.5% and grew to 7.7%; most of the variant expansion occurred in versions 0.5 and 0.6 (~1997)
# cmudict through the years

<table>
<thead>
<tr>
<th>Dictionary name</th>
<th>Year</th>
<th>Entries Total</th>
<th>Variant Count</th>
<th>Variant Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>cmudict.0.1</td>
<td>3/1993</td>
<td>99,279</td>
<td>3,498</td>
<td>3.52%</td>
</tr>
<tr>
<td>cmudict.0.2</td>
<td>3/1994</td>
<td>101,767</td>
<td>3,730</td>
<td>3.67%</td>
</tr>
<tr>
<td>cmudict.0.3</td>
<td>9/1994</td>
<td>110,996</td>
<td>3,932</td>
<td>3.54%</td>
</tr>
<tr>
<td>cmudict.0.4</td>
<td>2/1996</td>
<td>116,292</td>
<td>4,885</td>
<td>4.20%</td>
</tr>
<tr>
<td>cmudict.0.5</td>
<td>2/1997</td>
<td>123,885</td>
<td>7,523</td>
<td>6.07%</td>
</tr>
<tr>
<td>cmudict.0.5.a</td>
<td>3/1997</td>
<td>124,134</td>
<td>7,736</td>
<td>6.23%</td>
</tr>
<tr>
<td>cmudict.0.6</td>
<td>8/1997</td>
<td>126,867</td>
<td>7,668</td>
<td>6.04%</td>
</tr>
<tr>
<td>cmudict.0.6d</td>
<td>11/1998</td>
<td>129,482</td>
<td>10,034</td>
<td>7.75%</td>
</tr>
<tr>
<td>cmudict.0.6e</td>
<td>3/2005</td>
<td>129,627</td>
<td>10,090</td>
<td>7.78%</td>
</tr>
<tr>
<td>cmudict.0.7</td>
<td>12/1999</td>
<td>133,724</td>
<td>10,371</td>
<td>7.76%</td>
</tr>
<tr>
<td>cmudict.0.7a</td>
<td>8/2005</td>
<td>133,739</td>
<td>10,326</td>
<td>7.72%</td>
</tr>
</tbody>
</table>
What’s in there anyway?

• Various precursor datasets
  – RM, AN, ATIS, ??
• WSJ 20k dictionary (hand crafted)
• 50k surnames + ~6k names
  – Based on top-n names list from AT&T
  – Pronunciations generated using letter-to-sound (LTS) system
• DARPA evaluation vocabularies
  – Ad hoc additions per domain
• Words needed for local applications
  – Dictation, Communicator(?), *lines, etc.)
• Ad hoc additions and modifications …
What’s wrong

• Multiple sources: lack of consistency
  – Different authors (and different levels of expertise)
  – LTS rules used for some entries

• Confusion between normalization and pronunciation (i.e., correct expansion must be determined in a pre-process)
  – DR     D R AY1 V
  – DR(2)  D AA1 K T ER0
  – DR.    D R AY1 V
  – DR.(2) D AA1 K T ER0
What’s wrong

- Careless inclusion of automatically-derived pronunciations
  - LTS rules fail on unexpected input
  - FS F S [since version 0.1!!]
  - FS (2) EH1 F EH1 S
  - FSI F S IY1
- Ill-considered entries
  - ZZZZ Z IY0 Z
  - ZZZZ (2) Z Z [since version 0.6]
- And probably other oddities …
What is to be done?

• Clean up current dictionary, to the extent possible
  – Partition into sets that behave differently
    • Core vocabulary
    • Special-rules words (e.g., names, places)
    • Acronyms, initialisms (NASA, CIA)
    • Abbreviations, contractions (blvd., they’ll)
    • Alphanumerics (e.g., J3)
    • Foreign terms (e.g. C’EST)
    • Vocalisms (e.g., hmm)
  – Remove junk entries (e.g., zzzz)

• Rationalize maintenance process
  – Audit trails: author, project, …
  – Richer annotation (e.g., POS, word class, etc)
  – Stored as a database, with periodic controlled releases (by date?, by growth/change?)

• Develop techniques for (semi-)automatically adding new words
• Train LTS models based on homogeneous word sets
Why bother fixing things?

• In practice, people now use a combination of look-up and LTS rules to generate domain-specific dictionaries
  – For example, web-based services such as http://www.speech.cs.cmu.edu/tools/lmtool.html

• But current LTS systems are fragile
  – Rule-based approaches depend on quality of authoring (and ongoing maintenance)
  – Learning-based approaches depend on an internally consistent corpus (“garbage in, garbage out”)

• Improvements in dictionaries would in either case lead to better quality generated pronunciations
Questions / Discussion