The JanusRTk Switchboard/CallHome System

Pronunciation Modeling

Michael Finke

Interactive Systems Laboratories
Carnegie Mellon University
University of Karlsruhe

May 15, 1997

Interactive Systems Laboratories

Motivation

Do not try to imitate pronunciation; use a dictionary form: "no" will do for "naw," "nah," etc., "oh" for "aw,"; "going to" (not gonna or goin to); "you all" rather than "y'all"; "kind of" instead of "kinda"; etc. Nonstandard words which are not in the dictionary (e.g., kiddo) should be typed normally, i.e. without quotes or other special notation ...

Contractions are allowed, but be conservative. For example, contraction of "is" (it's a boy, running's fun) is common and standard, but there'll (there will) be forms that're (that are) better left uncontracted. It is always permitted to spell out forms in full, even if the pronunciation suggests the contracted form. Thus it is O K to type he is and they are and we would even if it's he's and they're and we'd you heard.

SWITCHBOARD Transcription Manual, Revision 4: 17 March 1992

Multiwords

- Problem: How to handle crossword pronunciation phenomena like contractions and reductions

GOING TO	GONNA
KIND OF	KINDA
DID YOU	DIDJA

- Multiwords: List of 205 word sequences selected by
  - frequency count
  - bigram perplexity reduction
  - hand picked

```
  :            :          :           :             :
DID_HE       DID_YOU   DOES_THAT   DOES_THIS     DON'T_HAVE
DON'T_WANT   DO_IT     DO_WE       DO_YOU        DO_YOU_HAVE
FIND_OUT     FOR_A     FOR_AN      FOR_THE       GET_A
GET_OUT_OF   GIVE_ME   GOING_TO    GOING_TO_BE   GOING_TO_HAVE
  :            :          :           :             :
KIND_OF      LET_ME    LIKE_A      LOOK_AT       LOT_OF
```


Multiwords in the Language Model

- DevTest using unadapted SWB system

Condition	SWB WER	CH WER
LM without Multiwords	34.1%	47.2%
LM with Multiword tokens	34.4%	48.1%

- Decoding: Use Multiwords as acoustic units but split them when it comes to computing the language model probability.

Dictionary Expansion

- Expand the recognition dictionary by applying a set of phonological rules in order to generate a variety of pronunciation variants (about 2 / word)

1	[AX IX] N → (E)N
2	[AX IX] M → (E)M
3	[AX IX] L → (E)L
4	[AX IX] R → AXR
5	[T D] → DX / [+VOWEL] __ [AX IX AXR]
6	[T D] R → DX
7	L → 0 / __ Y [AX IX AXR]
8	IY → Y / __ [AX IX AXR]
9	NG → N
10	HH → 0 / WB __
11	W → 0 / WB __
12	DH → 0 / WB __
13	[T D] → 0 / [+VOWEL] __ [TH DH]
14	[T D] → 0 / [+CONS +CONTINUANT] __ WB
15	R AX → ER / [-WB] __ [-WB]
16	T → 0 / [M N NG] __ [AX IX AXR]
17	BECAUSE → K [AH AO] Z
18	GOING TO → G AH N AX
19	WANT TO → W AH N AX
20	YOU KNOW → Y AX N OW
21	DO YOU → D Y UW

"Transcribing" Switchboard

- Forced Alignment (Viterbi) is used to determine which pronunciation is associated with each word token in the training corpus
- Problem: There are several sources of errors in Switchboard transcriptions
  - Pronunciation
  - Segmentation
  - Incorrect Transcriptions.
- Flexible Transcription Alignment: Utterance HMMs that allow for

1. alternative pronunciations
2. optional multiwords
3. optional beginning and ending words
4. optional silence or breathing models between words,
5. optional noise words to start or end an utterance.

Flexible Transcription Alignment

- Example

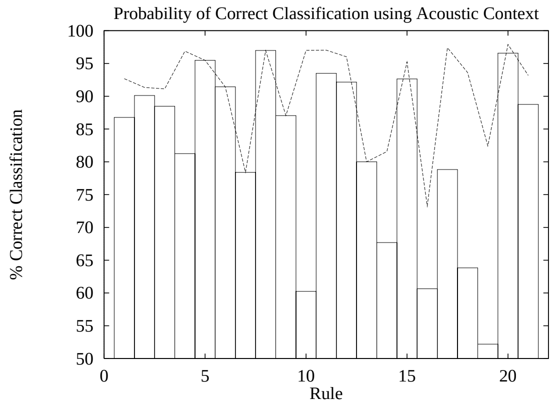
$(<BREATH>) <NOISE>(BREATH) $ AND $(<SBREATH>) I $ YOU_KNOW $ IT'S $ I GUESS IT'S SO NORMAL TO(2) $(<BREATH>) START TO WONDER $ ABOUT THAT EVEN IF(2) SHE DOESN'T(2) NEED THAT BUT $(<SBREATH>) YOU_KNOW SHE'S KIND_OF(KINDA/1) ASKING(1/9/9) QUESTIONS ABOUT WHAT(2) $(<BREATH>) WELL WHAT'S GOING_TO(GONNA/1) HAPPEN THIS CAN'T LAST FOREVER(1/4,18,20/18,20) AND $(<SBREATH>) <NOISE>(THROAT)

Mode dependencies

- Idea: Predict rule probabilities and thus variant probabilities based on
  - word related features like phonetic context and information about the word (unigram count, content vs. function word)
  - speaking style/mode related features
    - speaking rate (word/phone rate etc.),
    - deviation from mean word/phone duration of w,
    - F0, etc.
- Decision Trees were grown to predict $p(rule|word)$ resp. $p(rule|word, mode)$

Decision Tree based Probability Estimates

- Prediction Accuracy: Classification performance of JRTk pronunciation rules using decision tree prediction from word and mode features $p(rule|word, mode)$ (dotted lines), compared to chance performance (bars).



From Rule Probabilities to Variant Probabilities

- Let $\mathbf{r}^+$ be the rules which match with the baseform of w and were applied to derive variant $q_i(w)$
- Let $\mathbf{r}^-$ be the rules which match with the baseform but were not used to get $q_i(w)$
- The probability of pronunciation variant $q_i(w)$ is given by:

$$P(q_i|w) \quad = \quad \frac{\prod_{\mathbf{r}^+} P(\mathbf{r}^+) \prod_{\mathbf{r}^-} (1 - P(\mathbf{r}^-))}{Z}$$

where Z is a normalization constant.

Results

- Flexible Transcription Alignment (unweighted pronunciation variants)

Condition	SWB WER	CH WER
Baseline	32.2%	43.7%
FTA training/test w.basedict	30.7%	41.9%
FTA training/test w.exp.dict	31.1%	42.5%

- Pronunciation Variant Weighting

Condition	SWB WER	CH WER
unweighted	28.7%	38.6%
weighted $p(rule|word)$	27.1%	36.7%
weighted $p(rule|word, mode)$	26.7%	36.1%