Adding Word Duration Information to Bigram Language Models

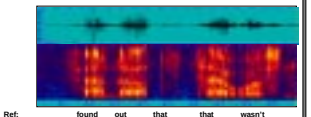
George Doddington

National Institute of Standards and Technology
doddington@nist.gov
http://www.itl.nist.gov

Yufeng Wu, Aravind Ganapathiraju, Joseph Picone

Institute for Signal and Information Processing
Mississippi State University
{wu, ganapath, picone}@isip.msstate.edu
http://www.isip.msstate.edu

Motivation



- Humans follow an internal sense of timing
- Duration is one of the most reliable and accessible prosodic features

Suprasegmental Information

- Word duration represented as a single scalar attribute
- Word duration bigram model ($F \equiv \{w,\tau\}$):

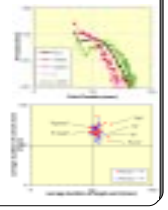
$$Pr(F_i \mid F_{i-1}) = Pr(w_i,\tau_i \mid w_{i-1},\tau_{i-1}) = Pr(\tau_i \mid w_i,w_{i-1},\tau_{i-1})\, Pr(w_i \mid w_{i-1},\tau_{i-1})$$

where W is the word identity and τ is the duration

- Can be implemented in a rescoring paradigm as an additional knowledge source applied to word hypotheses (leads to a feasible implementation)

Duration Analysis For The Word "I"

- Duration distributions for selected bigrams containing the word "I" (WS97 training data)
- Comparison of left context to right context duration for the 750 most common bigrams containing the word "I" (WS97 training data)

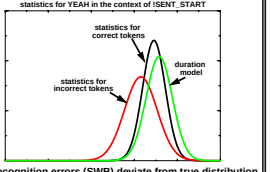


N-best Rescoring Results

- Baseline: 32.4% WER on 637 SWB utterances
- Rescoring of 100-best hypotheses (provided by BBN)
- Oracle WER: 21.2%

	[weight 1d, weight 2d]		
scale	[0.1, 0.1]	[0.1, 0.5]	[0.5, 0.1]
0.01	32.5	32.4	32.3
0.05	32.4	32.3	32.2
0.1	32.3	32.3	32.2

Implicit Duration Models Insufficient



- Recognition errors (SWB) deviate from true distribution
- Word durations preferred over phone durations

Bigram Duration Model

- Duration augmented bigram probability:

$$P(w_i \mid w_{i-1},\tau_{i-1},\tau_i) = P(w_{i-1},\tau_{i-1},w_i,\tau_i) \,/\, P(w_{i-1},\tau_{i-1},\tau_i) = \frac{P(\tau_{i-1},\tau_i \mid w_{i-1},w_i)}{P(\tau_{i-1},\tau_i \mid w_i)} \frac{P(w_{i-1},w_i)}{P(w_{i-1})}$$

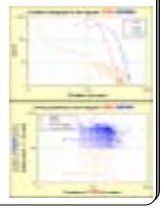
- Begin/end of sentences treated as special cases:

$$P(w_1 \mid S_{beg},\tau_1) = \frac{P(\tau_1 \mid S_{beg},w_1)}{P(\tau_1 \mid S_{beg})} \frac{P(w_1)}{P(S_{beg})}$$

$$P(S_{end} \mid w_{i-1},\tau_{i-1}) = \frac{P(\tau_{i-1} \mid w_{i-1},S_{end})}{P(\tau_{i-1} \mid w_{i-1})} \frac{P(w_{i-1},S_{end})}{P(w_{i-1})}$$

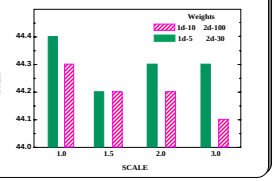
Analysis For The Bigram "You Know"

- Variance of each word in the bigram is low (implies duration is a well-behaved feature)
- Unigram duration of each word in the bigram is not predictable from the other word (warrants the use of higher order n-gram duration models)

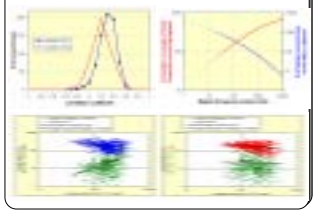


Word Graph Rescoring Results

- Baseline system: WER 44.4% on WS97 test set



Switchboard Data



Back-Off Weighting

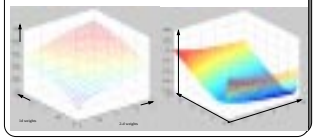
- Many duration bigrams have insufficient training data
- Combine bigram-specific models with word-specific and word-independent models in a back-off framework

$$P_{sm}(\tau_{i-1},\tau_i \mid w_{i-1},w_i) = \frac{\Omega_b P(\tau_{i-1},\tau_i \mid w_{i-1},w_i) + \Omega_w P(\tau_{i-1} \mid w_{i-1})\, P(\tau_i \mid w_i) + \Omega_g P^2(\tau_i)}{\Omega_b+\Omega_w+\Omega_g}$$

- Ω empirically chosen in initial experiments (can be estimated using deleted interpolation or other such smoothing algorithms)

Error Analysis

- Difference between the average duration model score for correct versus incorrect bigrams is crucial to performance (analogous to F-ratio)



Summary

- A consistent statistical modeling framework that exploits word duration models
- Modest improvement on SWB:
  - BBN 100-Best Lists: 0.2% WER absolute
  - ISIP Word Graph Rescoring: 0.3% WER absolute
- Future work:
  - Incorporate duration models into the grammar decoding loop
  - Better models of infrequently occurring bigrams: error analysis indicates greater potential benefits
  - Develop more sophisticated statistical models