JUST-IN-TIME LANGUAGE MODELLING

Adam Berger, Robert Miller
School of Computer Science
Carnegie Mellon University
aberger@cs.cmu.edu

May 14, 1998

Why language modelling is difficult

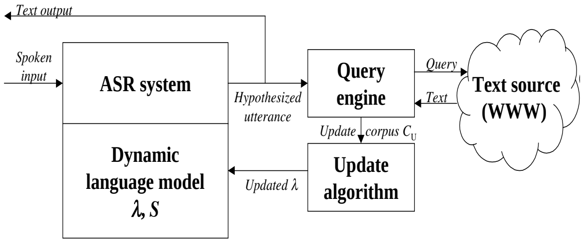
1. English is *non-stationary*

   So adaptive is better than static

2. New topics, not seen in training, are always cropping up
   - Even an adaptive LM becomes stale
   - e.g.: lexical correlations (*Simpson, Bronco*) and (*Lewinski, President*).

A "just-in-time" architecture

A diligent but ignorant transcriptionist:



Interleaved training and application
Query engine stops after fetching $10,000$ words

An exponential model of language

Start with trigram model; adjust marginal probabilities of topical words:

$$p(y \mid x) = \frac{1}{Z(x)} q(y \mid x) e^{\lambda_y}$$

where:

q: trigram model

e^{λ_y}: "boost" to probability of word y

$Z(x)$: normalization

A Bayesian approach

How much should we trust update corpus? Use a Gaussian prior over models:

$$p(\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_y \exp\left(-\frac{\lambda_y^2}{2\sigma^2}\right)$$

- "Smoothness" constraint: large adjustments less likely (Occam's Razor)
- σ^2 adjusts degree of trust in adjustments λ_y

Interleaved training/applying

1. Receive utterance.
2. Produce hypothesis with current model.
3. Fetch update corpus $\mathcal{C}$.
4. Compute MAP-optimal language model:

$$\lambda^\star = \arg\max_\lambda \frac{p(\mathcal{C} \mid \lambda)p(\lambda)}{p(\mathcal{C})} = \arg\max_\lambda p(\mathcal{C} \mid \lambda)p(\lambda)$$

   (using a modified *iterative scaling* algorithm)

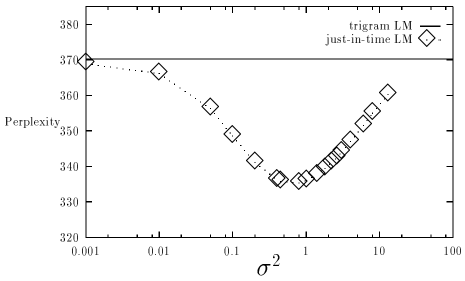
5. Go to 1

Experimental Results

Perplexity of 3 AP newswire stories of 11/1/97:

	IRS	China	Marathon
Static trigram model	313	781	1060
14k trigger BN model	307	750	1552
Dynamic (using AltaVista)	315	781	1038
Dynamic (using News Index)	298	704	1016

Determining σ^2



Final thoughts

- As described, *not* a real-time system, but how about once per n utterances, in the background?
- Dynamic vocabularies
- Complement to, not competing with, standard adaptive techniques.

Plans:

- Interface with CMU experimental speech dictation system
- Only modify marginal probabilities of true on-topic words (use IR techniques to identify these)