BBN TECHNOLOGIES
A Verizon Company

Using a Large LM

Nicolae Duta
Richard Schwartz

EARS Technical Workshop
September 5,6 2003
Martigny, Switzerland

“There is no data like more data”

-- Bob Mercer,

Arden House, 1995

Corollary:

More data only helps if you don’t throw it away.

Ngram Pruning

BBN TECHNOLOGIES
A Verizon Company

- When we train an n-gram LM on a large corpus, most of the observed n-grams only occur a small number of times.
- We typically discard these, for two reasons:
  - We assume they are not statistically significant
  - We don't want to use a LM with 700M distinct 4-grams!
- Questions:
  - Does the fact that an n-gram occurred one time provide useful information?
  - Is it practical to use a really large LM?

N-gram Hit Rate

BBN TECHNOLOGIES
A Verizon Company

- **We find the n-gram “hit rate” to be a useful diagnostic.**
- **The hit rate is the percentage of n-gram tokens in the reference transcription that are explicitly in the LM model.**
- **We find that the probability that a word is recognized is affected significantly by whether the corresponding n-gram is in the LM, because if it is not, the LM probability (from backing off) is significantly lower.**

Experimental Results

BBN TECHNOLOGIES
A Verizon Company

- English broadcast news test, (H4Dev03)

LM Order	Cutoffs [4g, 3g]	LM size [4g, 3g]	Hit Rates [4g,3g]	Perplex	WER
3	[inf, 6]	[0, 36M]	[0, 76%]	201	12.6
3	[inf, 0]	[0, 305M]	[0, 84%]	164	12.1
4	[6, 6]	[40M, 36M]	[49%,76%]	208	12.1
4	[0, 0]	[710M,305M]	[61%,84%]	139	11.8

- Cutoff of 6 for trigram loses 0.5% absolute
- 4-gram with cuttoff of 6 gains 0.5%
- 4-gram cutoff of 6 loses 0.3%

Batch Implementation

BBN TECHNOLOGIES
A Verizon Company

- Count ALL ngrams. Store in (4) sorted files.
  - Total size is about 8.5 GB.
- Do first pass recognition using 'normal' LM
  - Produce n-best (or lattice)
  - Make sorted list of all recognized n-grams of all orders in the hypotheses in the test set
- Make one pass through the count file
  - Extract only those counts needed (in the hypotheses)
  - Accumulate total count and number unique transitions for each state
- Create mini-LM
- Apply LM to n-best (or lattice) as re-scoring.

- Total process requires 15 minutes for a 3 hour test.
  - < 0.1 xRT
  - Could be implemented to be fast on a short test file.

Discussion

BBN TECHNOLOGIES
A Verizon Company

- **Even a single observed token of an n-gram tells you that it is *possible.***
  - **It is important to know the difference between n-grams that are unobserved because they are rare and those that are impossible. [If we could really know this, we would have much better results.]**
- **The gain from keeping all n-grams is significant (0.5% for 3-grams, 0.3% for 4-grams).**
- **Small question:**
  - **Would this result hold for other back-off methods?**