

# Improving The Suitability Of Imperfect Transcriptions For Information Retrieval From Spoken Documents

Matthew A. Siegler  
Michael J. Witbrock\*

Carnegie Mellon University  
\*Justysystems Pittsburgh Research Center  
(now Lycos Inc.)

1999 March 18 -- SP-16.4  
ICASSP 1999

# Motivation

- Indexing speech databases.
- Transcriptions can have errors.
- Automatic transcriptions: **more** errors.

Relevance between a query and a document should largely depend on the expected error probability of the "relevant" terms.

# Motivation (cont.)

Could develop purely probabilistic engine...

Goal: Improve performance of IR system without changing the relevance engine.

- Modify term weighting scheme only.
- Use existing structure from the SR.

# Mutual Information Term Weights

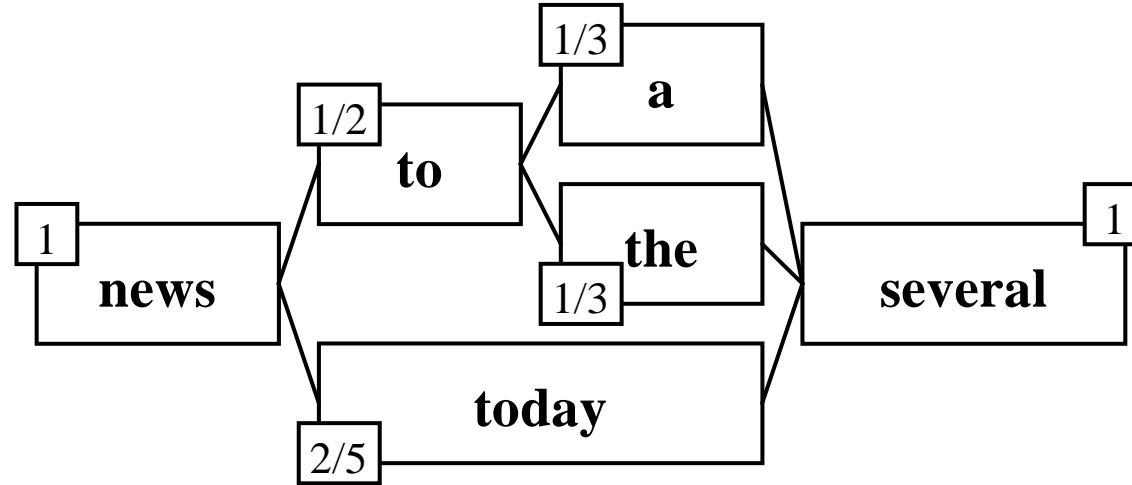
$$IDF_i \equiv -\log \left( \frac{|\{\forall m \text{ s.t. } w_i \in D_m\}|}{M} \right)$$

$$I(\mathbf{D}; w_i) = -\sum_{m=1}^M P(D_m) \log_2 P(D_m) + \sum_{m=1}^M P(D_m | w_i) \log_2 P(D_m | w_i)$$

$$\text{Rel}(Q, D_m | \mathbf{D}) = \sum_{i=1}^N P(w_i | Q) P(w_i | D_m) I(\mathbf{D}; w_i)$$

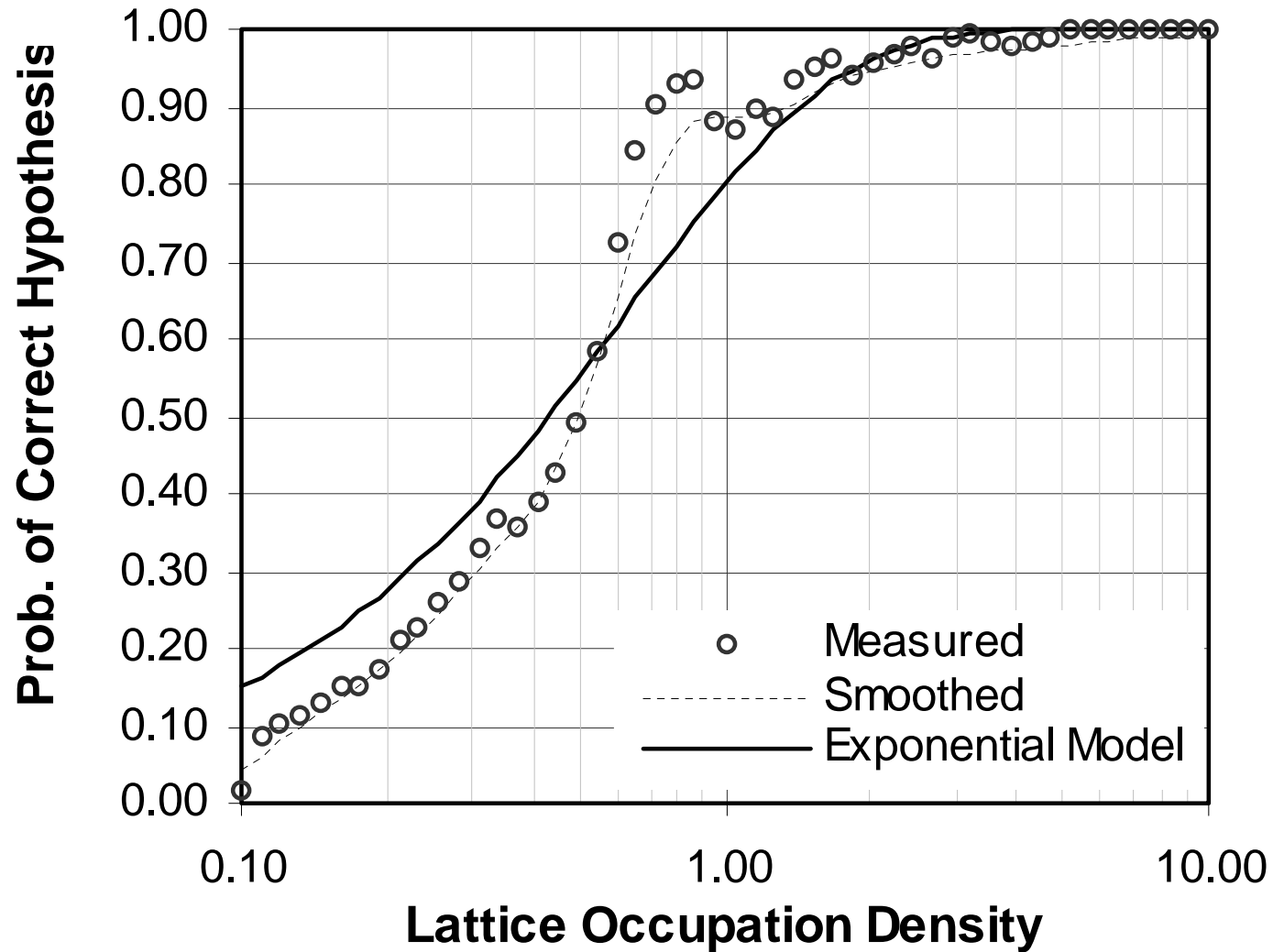
# Estimating Word Probabilities

Proportional to number of competing hypotheses in the SR lattice.

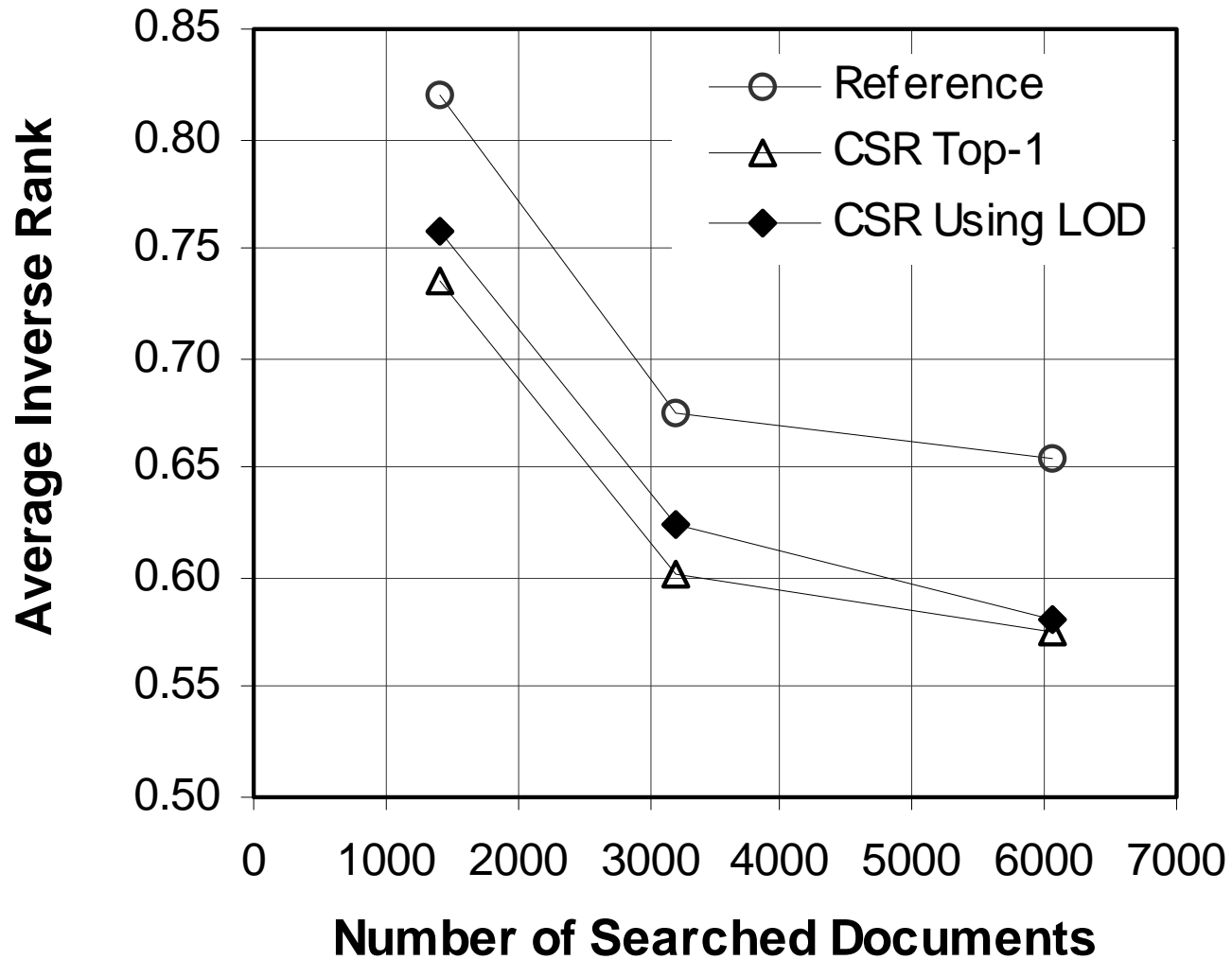


"Lattice Occupation Density" of a word.

# Probability Model



# Experimental Results (TREC-7)



# Summary

- Recovered 13%-38% of loss from SR.
- Re-used existing features.
- Limitation: Use of nonstandard IR metric.

## Work in Progress

- Incremental estimation with new data.
- Better estimates of word probability.