

Spoken Document Retrieval at CMU

The TREC-7 SDR Track

Matthew Siegler

Adam Berger, Michael Witbrock*, Alex Hauptmann

Carnegie Mellon University

*Justsystem Pittsburgh Research

9 November 1998

Outline

1. System Description
2. LNU+MI Retrieval Equation
3. Word Probabilities
4. Oracle Path
5. LM-based Metric

System Description

Speech Recognition Component

- Sphinx-III Continuous HMM
- Narrow beam search
- Trained on 100-hour BN corpus

Information Retrieval Components

- Usual text processing
- LNU metric + Mutual Information
- "Language Model" Relevance

LNU+Mutual Information

Justification for TF·IDF: ~Mutual Information

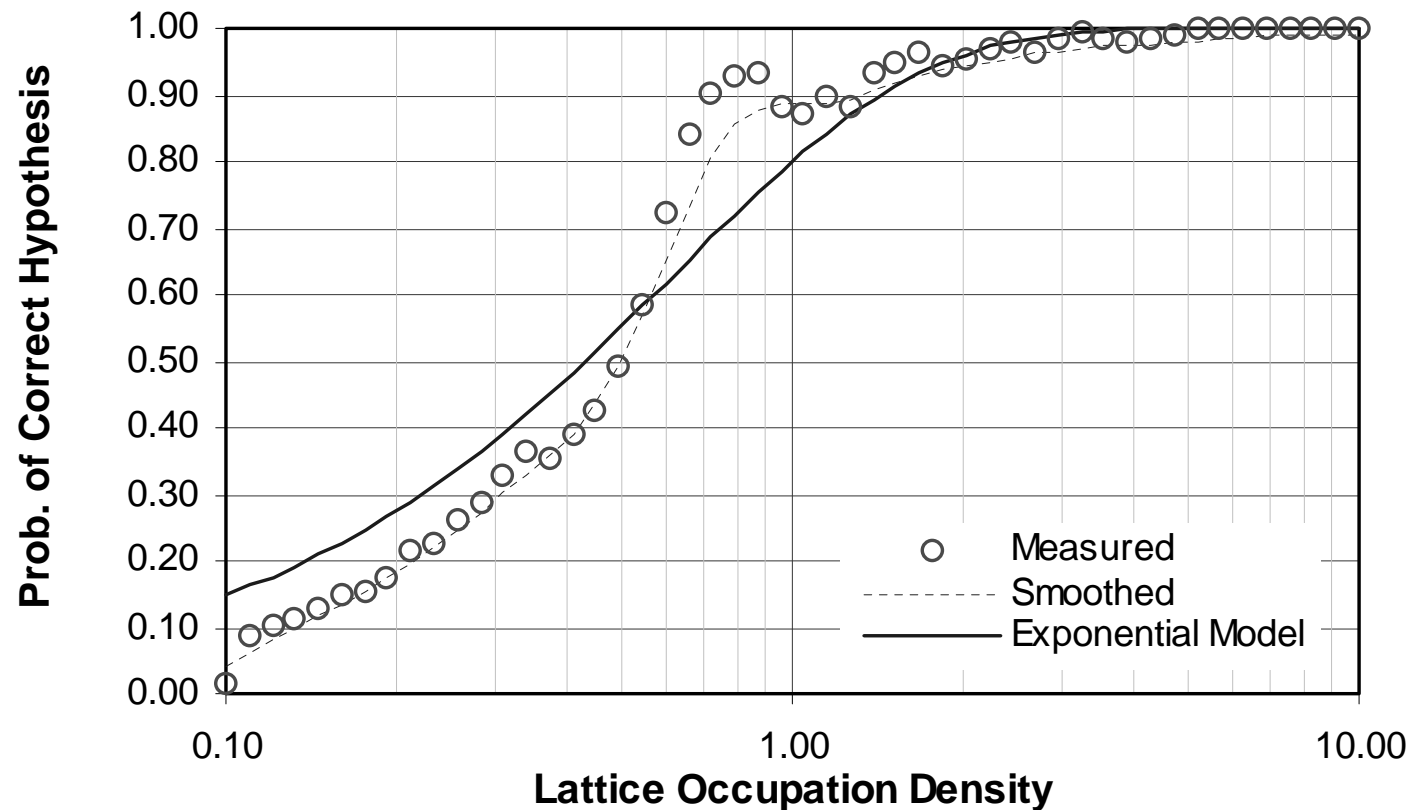
$$\text{Rel}(Q, D_m | \mathbf{D}) = \sum_{i=1}^N E\{I(D; w_i) | Q, D_m\} = \sum_{i=1}^N P(w_i | Q, D_m) I(D; w_i)$$

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

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

Word Probabilities

Use number of choices in lattice to predict error



Oracle Paths

- "Best path" through recognition lattice.
- Chosen to minimize some error criterion.
- Better: minimize errors of **text-processed** path.

	Baseline		Oracle	
	REF	CSR	Pre-Filter	Post-Filter
Training Set (AIR)	0.79	0.75	0.79	0.75
Testing Set (P_{AVG})	0.39	0.36	0.39	0.37

Language Model Relevance Function

Based on work by Ponte/Croft.

- Create unigram model for each document
- Predict likelihood of query for each document
- Competitive with LNU-metric