Relevance Based Language Models

Yusuf Aytaş - Fethi Burak Sazoğlu

Introduction

- New Trend - Language Modeling
- Random Sampling instead of classical models
- Lack of training data

Related Work

- Probability ranking principle
- Depends on the used model
- Shift to estimation of sampling probabilities

A Formal Relevance Model

- No training data
- Assumptions for relevance
- Independent and Identical Sampling(Method1)
- Conditional Sampling(Method 2)
- 2nd is better

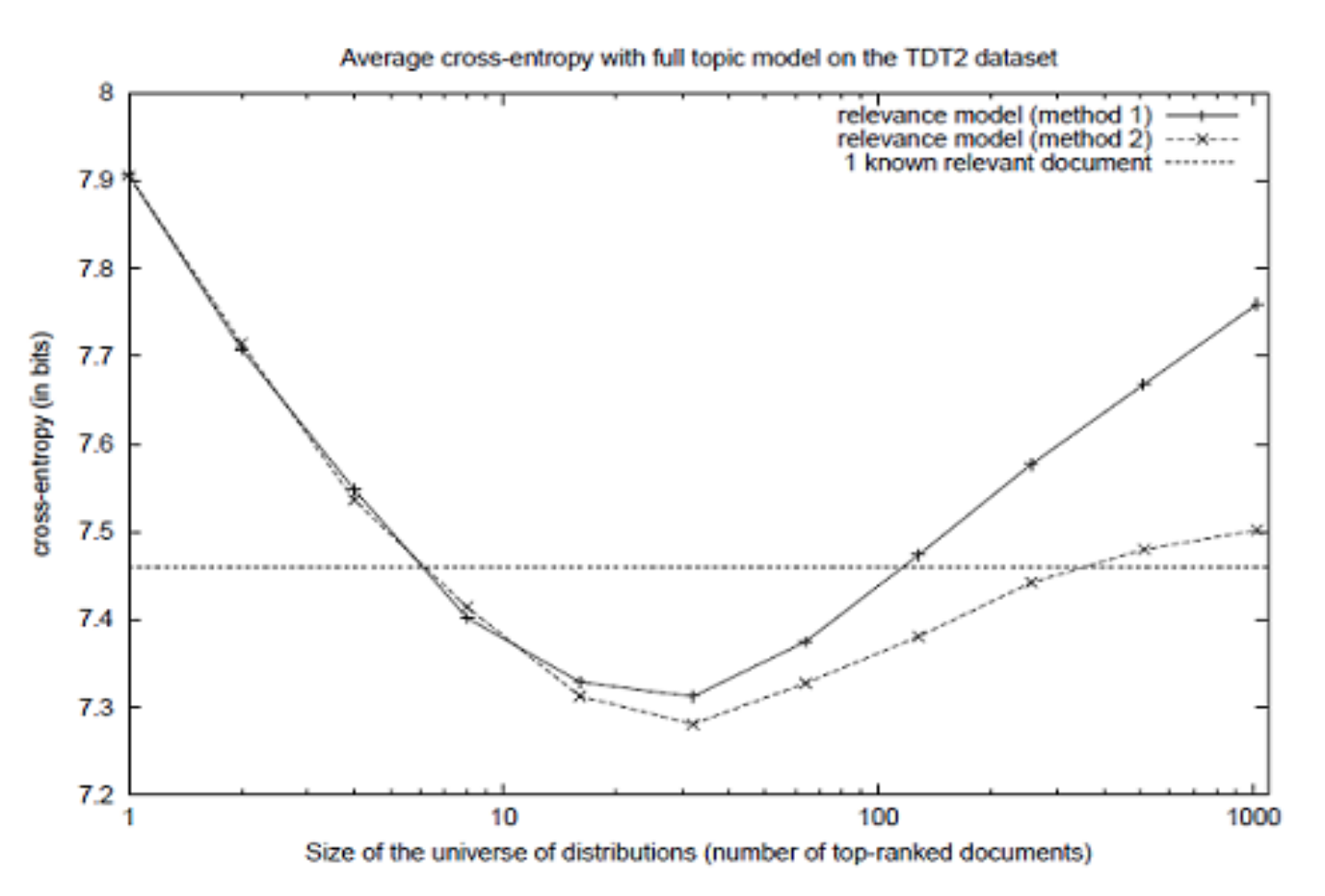
A Formal Relevance Model

- No test data
- Assumptions for relevance
- Independent and Identical Sampling
- Conditional Sampling
- 2nd is better

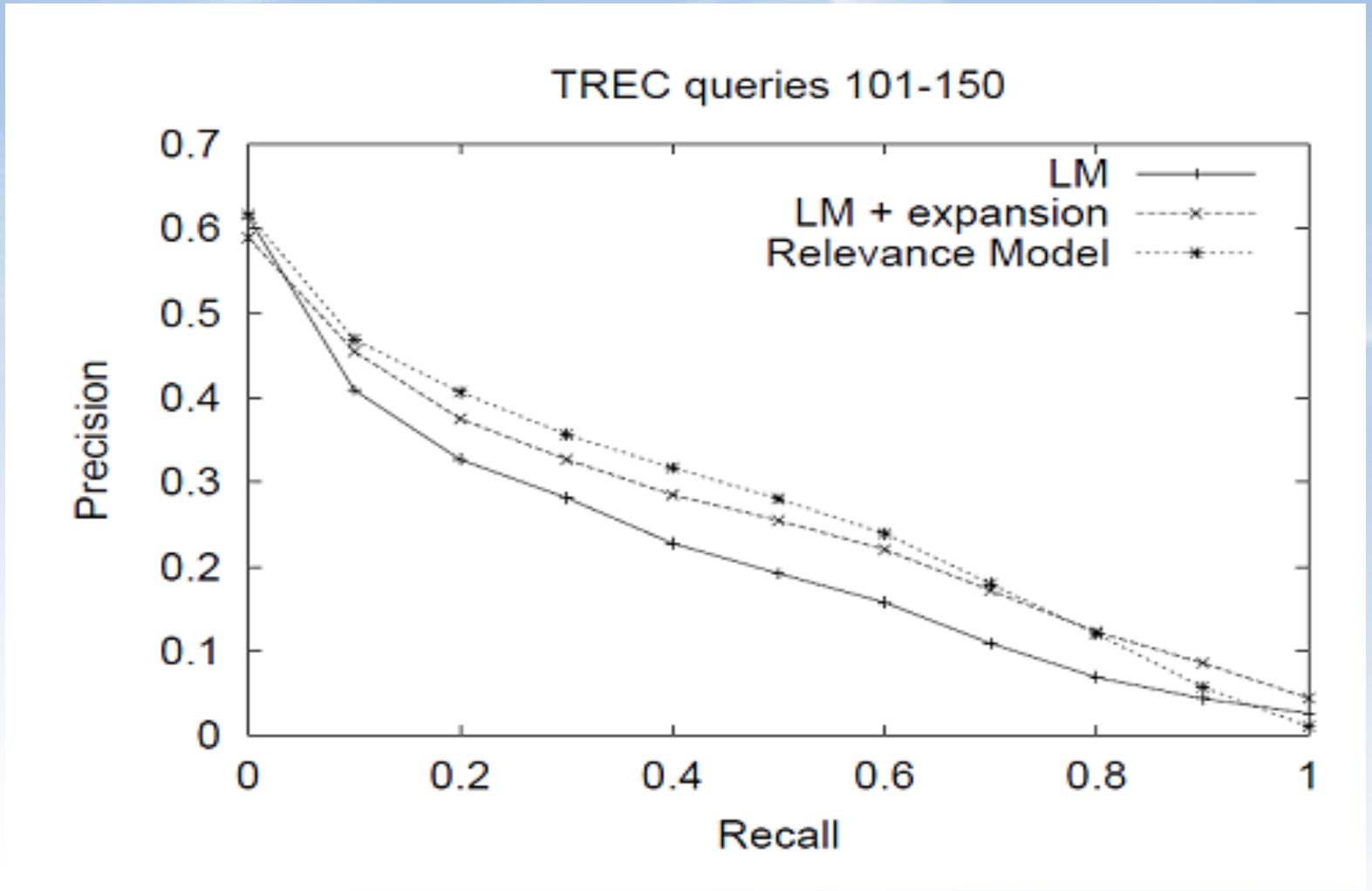
Experiments

- Cross-entropy with true relevance model
- Relevance model vs. Language model on TREC data
- Relevance model vs. TDT training set with 1, 2, 4 elements.

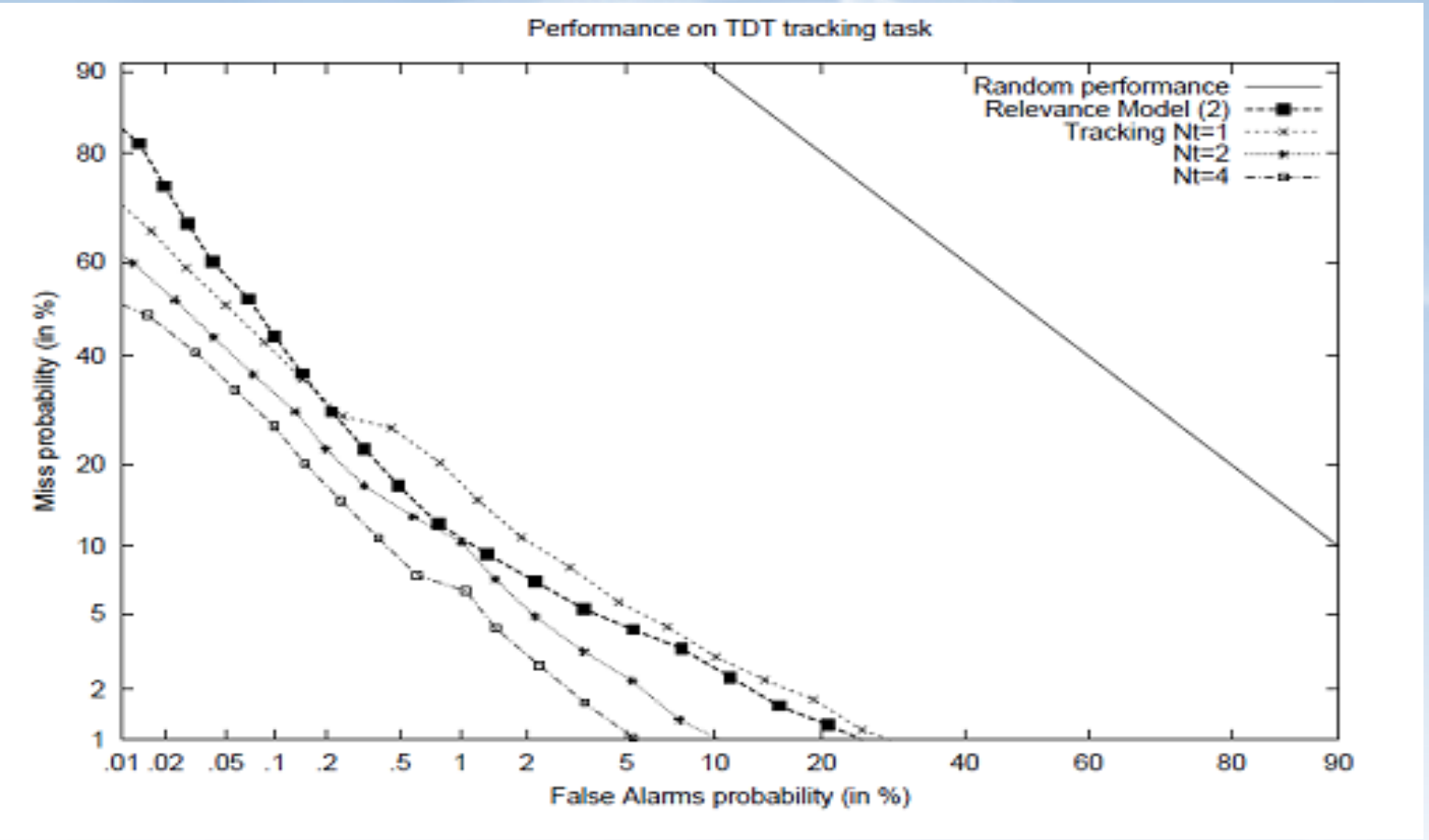
Cross - Entropy



TREC data


TREC queries 101-150
LM
LM + expansion
Relevance Model
Precision
Recall
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1

TDT (topic tracking) data



Conclusion

- Proposed model uses no training data
- Unites classical probabilistic models of relevance with language model approach
- Main contribution: Formal probabilistic approach to estimate a relevance model instead of heuristics (tf * idf)