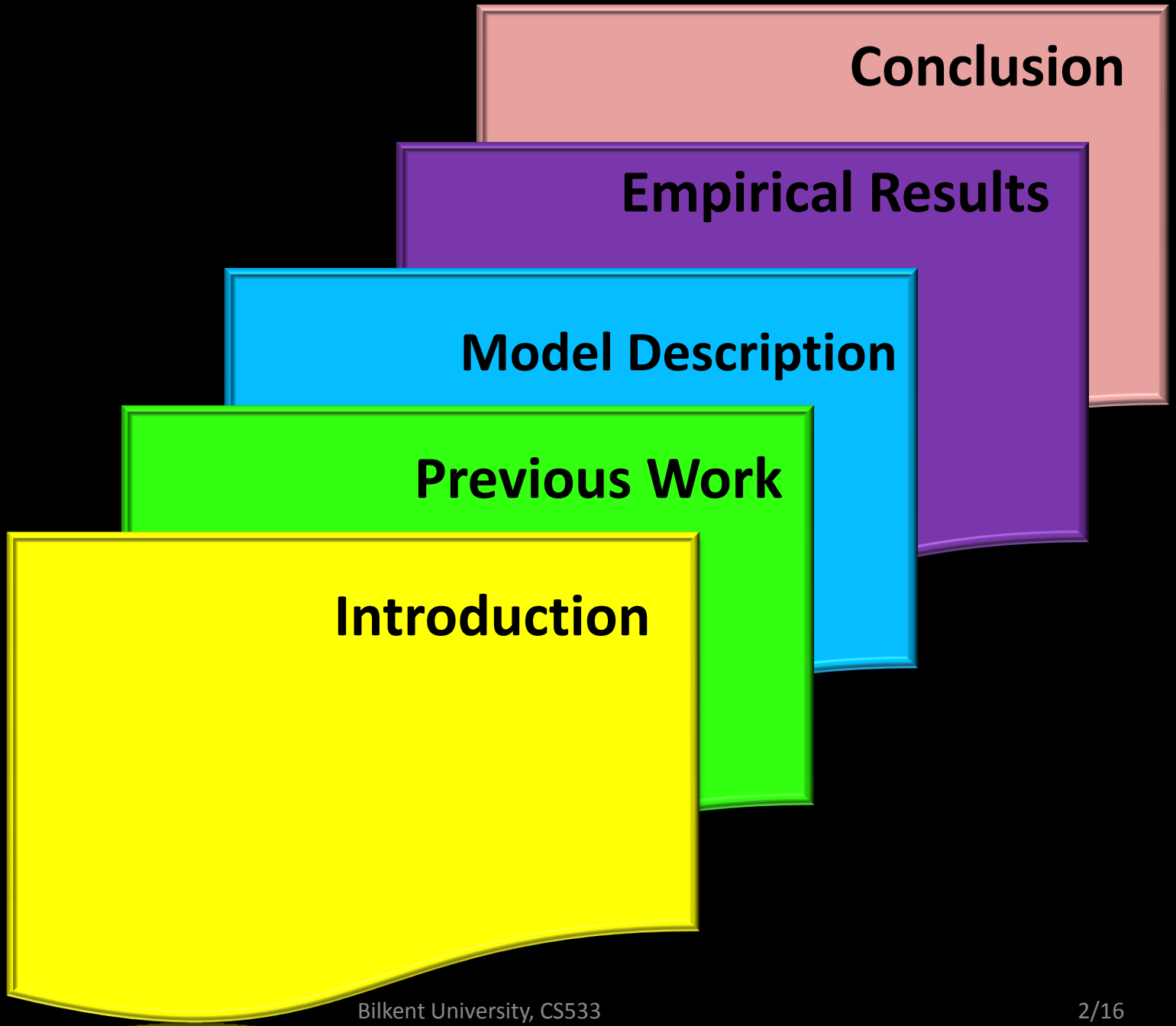
CS533-Information Retrieval Systems

A Language Modeling Approach to Information Retrieval

Ponte & Croft

Mehmet Ali ABBASOĞLU

Gülsüm Ece BIÇAKCI


Conclusion
Empirical Results
Model Description
Previous Work
Introduction

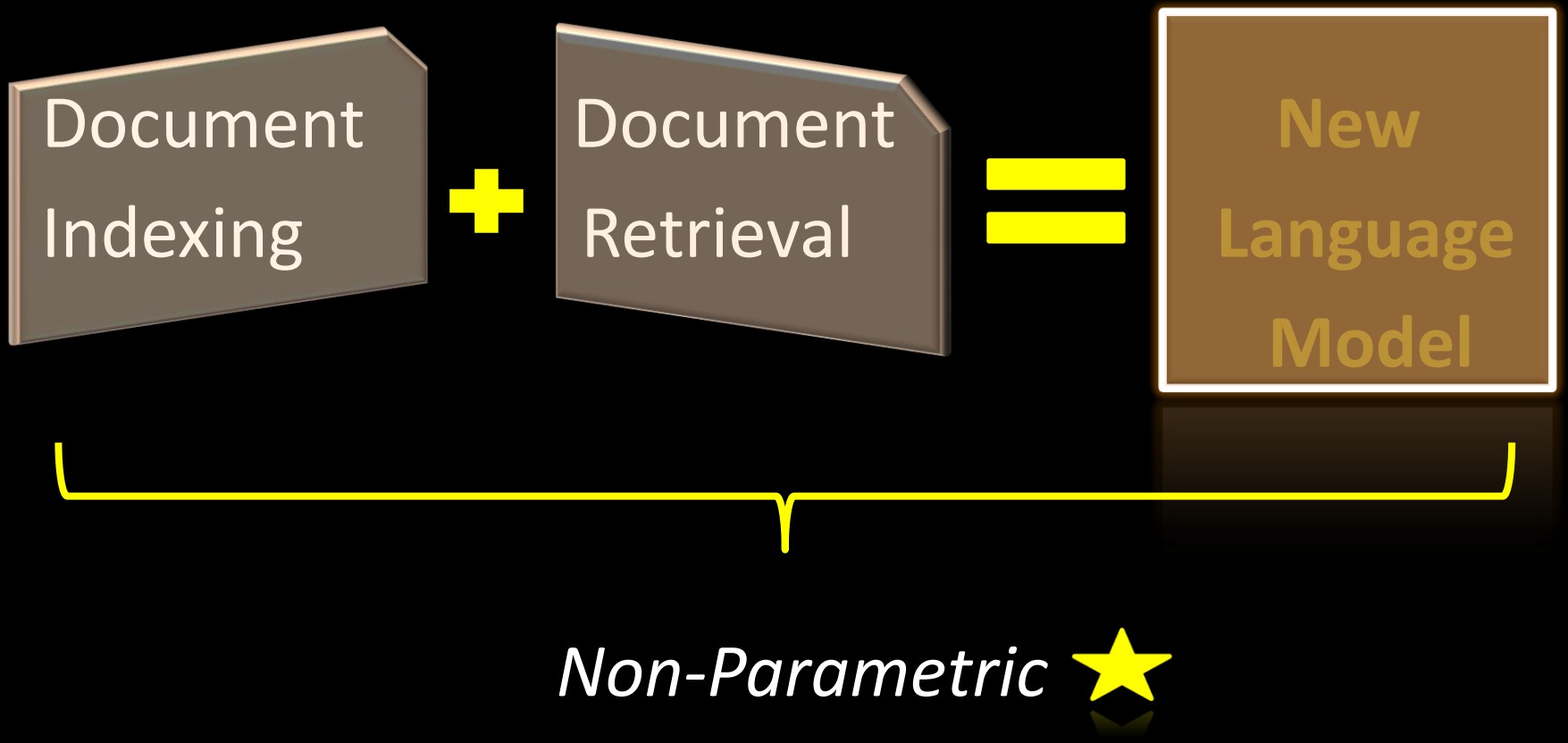
Introduction

Document indexing & retrieval

Difficult

Reason: lack of indexing model

Introduction



Introduction

abstraction of the retrieval task itself

probabilistic sense,
an explanatory model of the data

Predict the probability of the next word in an ordered sequence

Introduction

Approach to Retrieval

Estimate the probability of generating query

Rank the documents to these probabilities

Introduction

- Term frequency
- Document frequency
- Document length statistics

integral part of the model,

but not used heuristically

Previous Work

2- Poisson

- a subset of terms occuring in a document

Robertson & Sparck Jones – Croft & Harper

- estimate the probability of the relevance of each document to the query

Previous Work

Fuhr

- integration of indexing and retrieval models

INQUERY inference by Turtle & Croft

- making inferences of concepts from features

Model Description

- Aim is to estimate **$p(Q \mid M_d)$,** probability of the query given the language model of document d.
- They define maximum likelihood estimate of the probability of term t

$$\hat{p}_{ml}(t|M_d) = \frac{tf_{(t,d)}}{dl_d}$$

Model Description

- For non-occuring terms instead of taking $P_{ml}=0$, they take c_{ft} / c_s
- For the situations that arbitrary sized sampled data, maximum likelihood estimator could be reasonably confident. To estimate larger amount of data:

$$\hat{p}_{avg}(t) = \frac{\left(\sum_{d_{(t \in d)}} p_{ml}(t|M_d)\right)}{df_t}$$

Model Description

- They define risk as

$$\hat{R}_{t,d} = \left(\frac{1.0}{(1.0+\overline{f}_t)}\right) \times \left(\frac{\overline{f}_t}{(1.0+\overline{f}_t)}\right)^{tf_{t,d}}$$

- They define ***p(Q | M_d)*** as

$$\hat{p}(Q|M_d) = \prod_{t \in Q} \hat{p}(t|M_d) \times \prod_{t \notin Q} 1.0 - \hat{p}(t|M_d)$$

Results

- They performed recall / precision experiments on two data sets
  - TREC topics 202-250 on TREC disks 2 and 3, TREC 4 ad-hoc task
  - TREC topics 51-100 on TREC disk 3 using concept fields.
- Implementation is done using ***Labrador*** information retrieval engine.

Results

- On the eleven point recall/precision:
  - The language modeling approach achieves better precision at all levels of recall.
  - Most of the improvements are statistically significant according to ***Wilcoxon*** test.

Conclusion

- They presented a novel way of text retrieval based on probabilistic ***language modeling***.
- Their language models are accurate representations of the data
- They claim that
  - Users of their methodology can understand their approach to retrieval.
  - Users also will have some sense of term distribution.
- Their language modeling has significantly better recall/precision results than ***INQUERY.***


Thank You