Automatic Segmentation and Semantic Annotation of Verbose Queries in Digital Library

Authored By

Susmita Sadhu, Plaban Kumar Bhowmick

Presented By

Susmita Sadhu

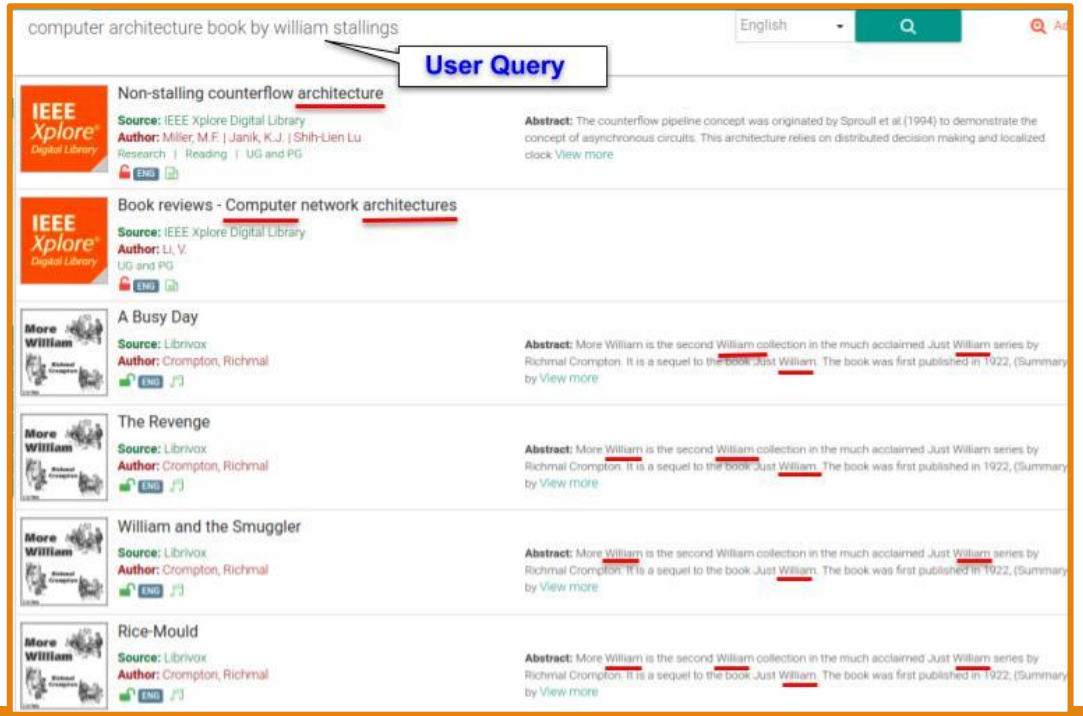
Introduction | What is NDLI and Why this work

- National Digital Library of India(NDLI), single portal, integrate several digital resources.
- NDLI relies only on the metadata for searching and browsing.
- Discoverability of appropriate resources in the portal is very disappointing in certain scenario.
- Need to improve retrieval system.

Motivation | Search Problem in Digital Library

- Query string:
  - ***'computer architecture book by william stallings'***

- Results:
  - Not any book of "computer architecture"
  - Most of them are "audio"
  - Author name not recognized
  - Result according to keyword match
  - No relevant item in result list



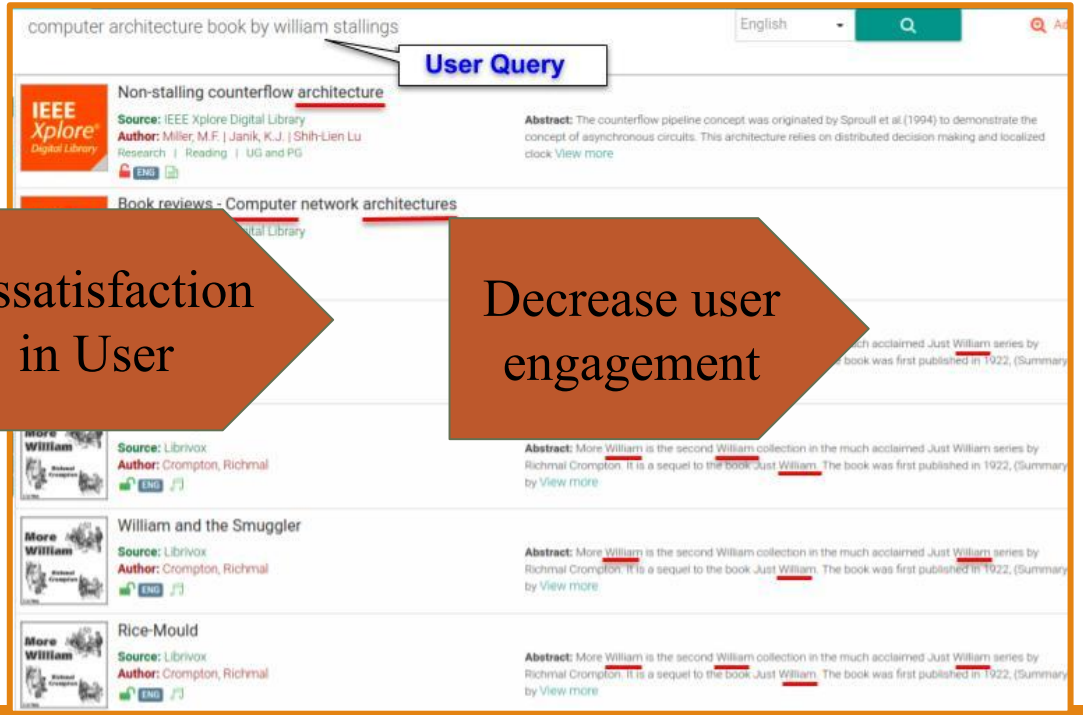
Motivation | Search Problem in Digital Library

- Query string:
  - ***'computer architecture book by william stallings'***
- F
  - puter
  - udio"
  - recognized
  - Result according to keyword match
  - No relevant item in result list

System performance degression

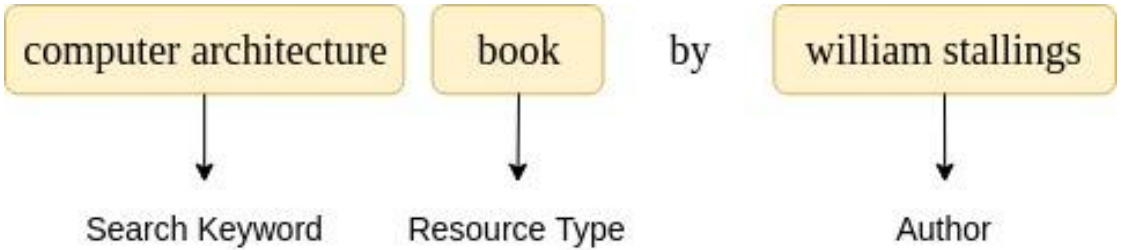
Dissatisfaction in User

Decrease user engagement

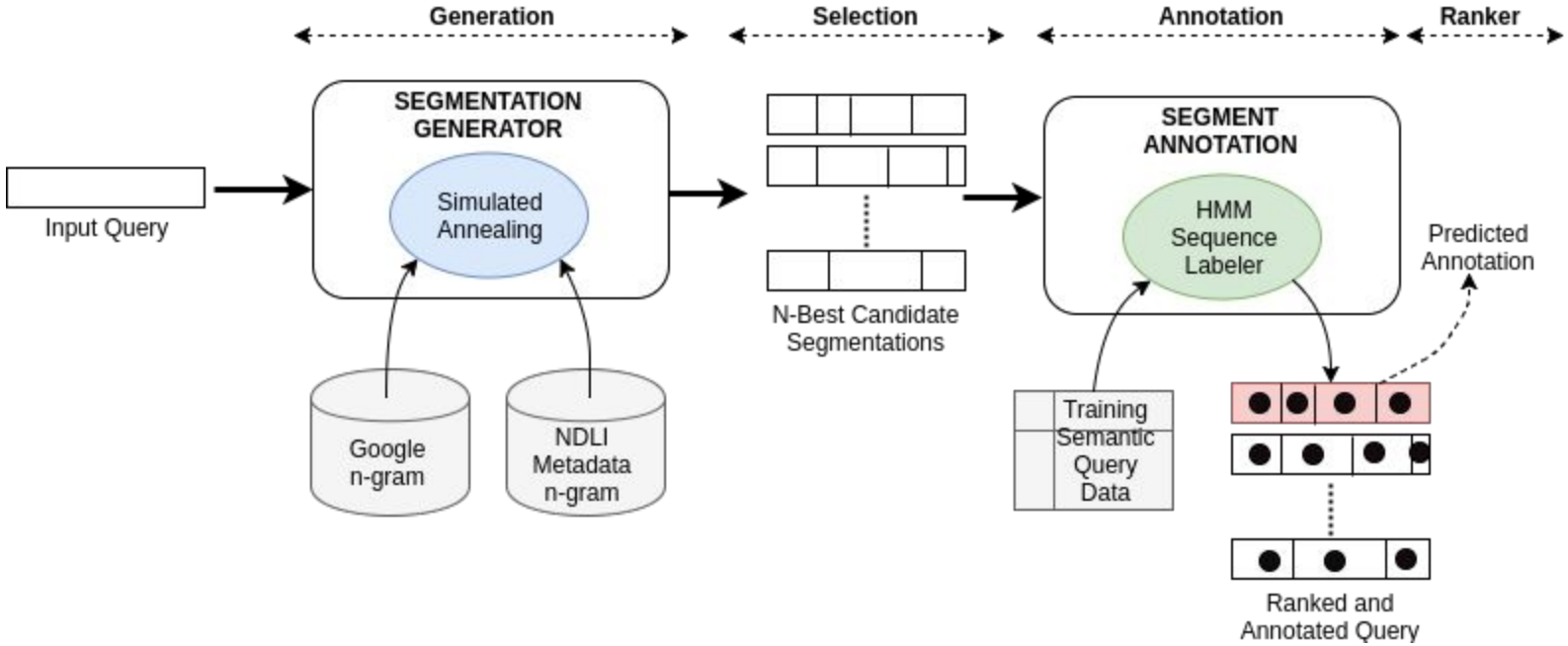


Objective

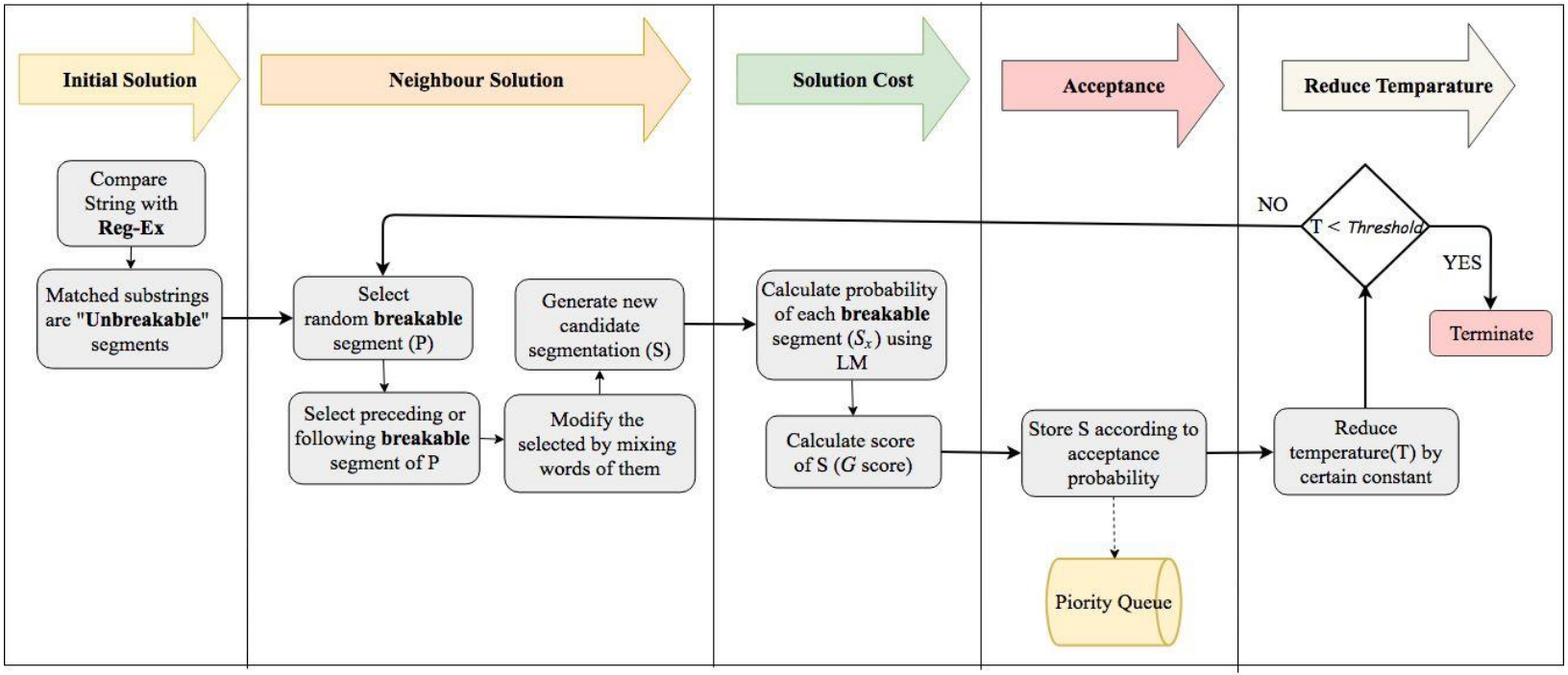
- Semantic analysis of user given natural language query
  - Meaningful segmentation of query string.
  - Annotate each segment with proper metadata.
  - Sample User query:
    “computer architecture book by william stallings”
  - Semantic analysis:



Proposed Architecture



Query Segmentation | Simulated Annealing

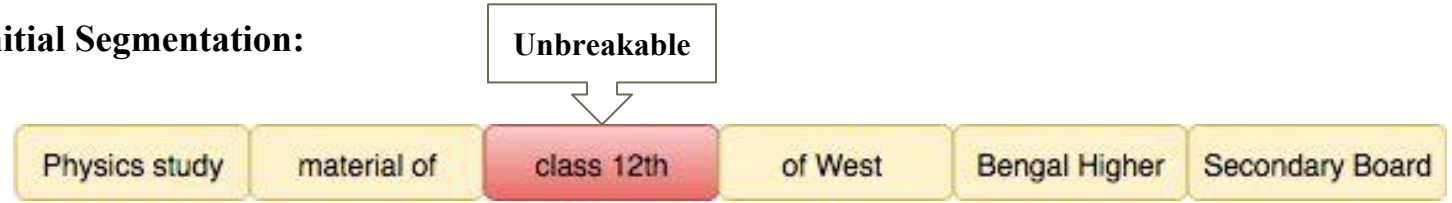


Query Segmentation | Example

Query String:

Physics study material of class 12th of West Bengal Higher Secondary Board

Initial Segmentation:



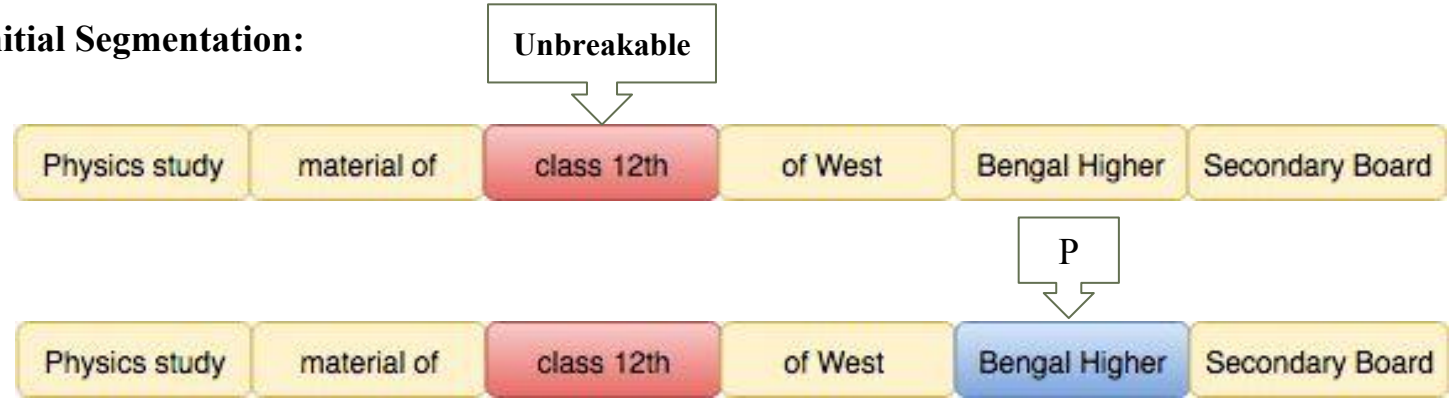
	Reg-Ex Example
1	((class | standard | std)<space>([1-12] | [I-XII])<space>(st | nd | rd | th))
2	((ug | pg | school | kids | children)<space>(students)?)

Query Segmentation | Example

Query String:

Physics study material of class 12th of West Bengal Higher Secondary Board

Initial Segmentation:

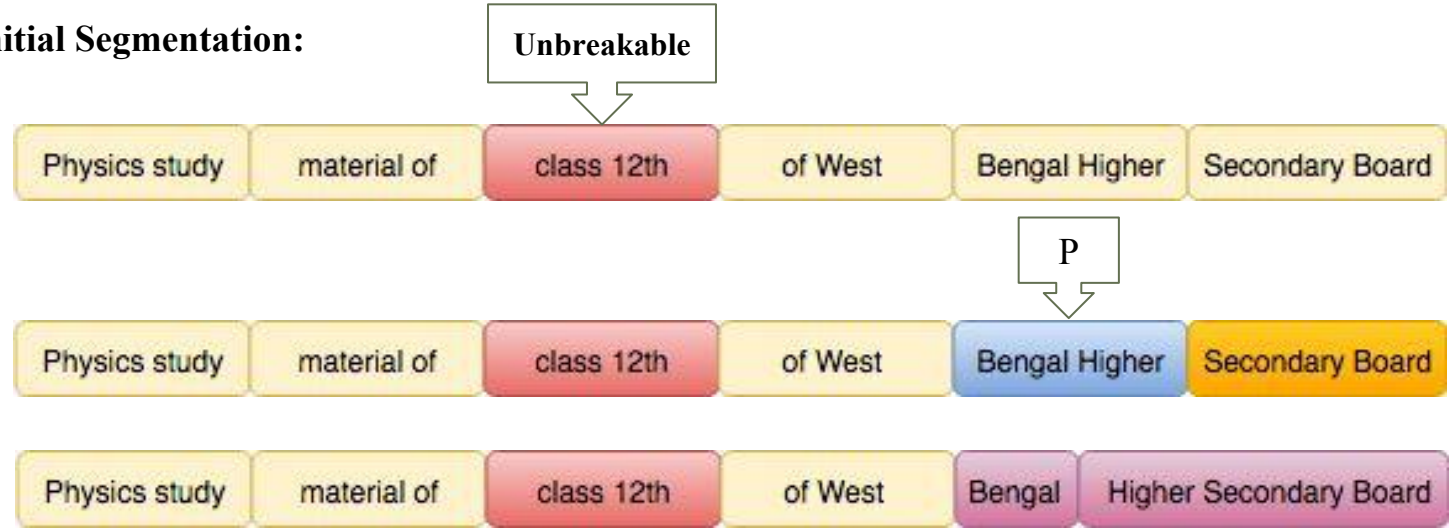


Query Segmentation | Example (cont.)

Query String:

Physics study material of class 12th of West Bengal Higher Secondary Board

Initial Segmentation:

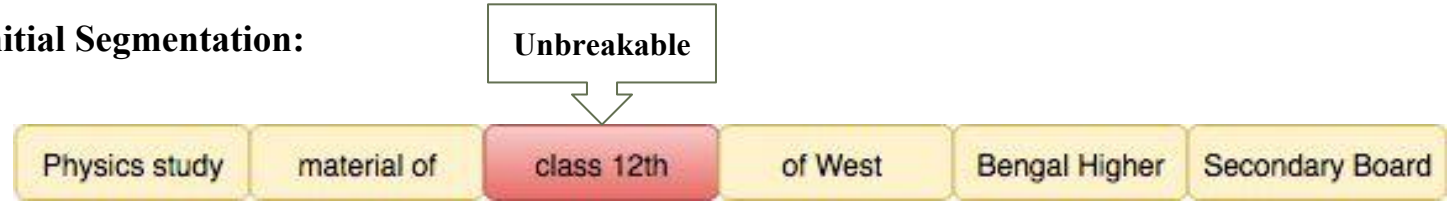


Query Segmentation | Example (cont.)

Query String:

Physics study material of class 12th of West Bengal Higher Secondary Board

Initial Segmentation:



New Segmentation:

S

Query Segmentation | Example (cont.)

Query String:

Physics study material of class 12th of West Bengal Higher Secondary Board

New Candidate Segmentation:

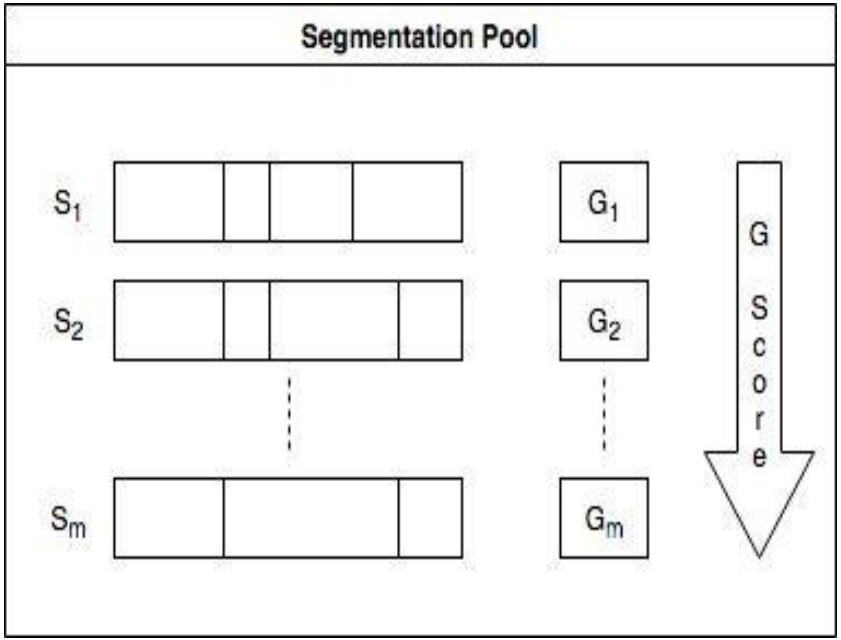
S

Physics study	material of	class 12th	of West	Bengal	Higher Secondary Board

Equations	
Eq 1	$P(Segment_x) = \alpha P_{Google}(Segment_x) + (1-\alpha)P_{NDLI}(Segment_x), \quad 0 \le \alpha \le 1$
Eq 2	$\mathcal{G}(\mathbb{S}) = \sum_{Segment_x \in \mathbb{S}} P(Segment_x) \times (1 - \Delta(Segment_x))$ Where, $\Delta(Segment_x) = P(End(Segment_x).Start(Segment_{x+1}))$

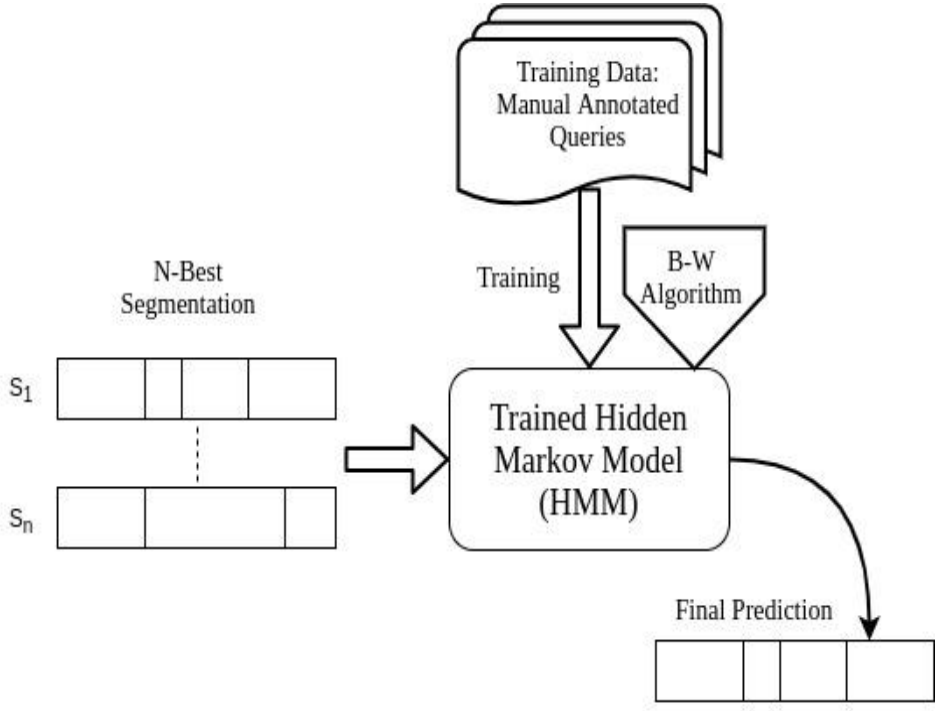
Query Segmentation | Selection

- A priority queue is maintained to store all generated candidate segmentations including G score.
- N-best segmentations are fed to the segmentation annotation stage.



Query Segmentation | Annotation & Ranking

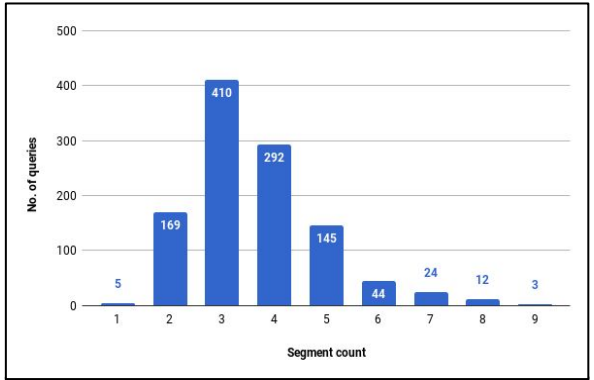
- Model the annotation task as a sequence labeling problem.
- Implemented with **Hidden Markov Model (HMM).**

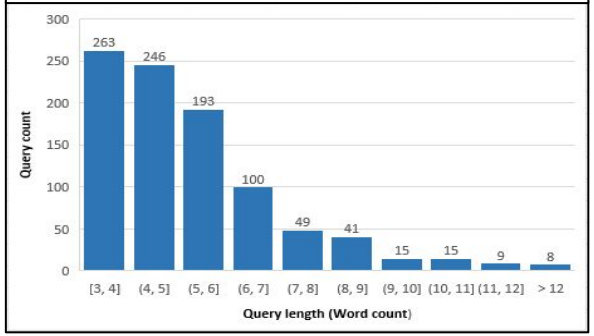


Benchmark Data

- Benchmark corpus contains 1100 user queries
  - Collected from NDLI user query log of certain duration.
  - Having length between 3-15 word.
  - Contains only english queries.
  - Do not contain queries having special characters.

- Segmented & annotated manually.





System Evaluation & Performance

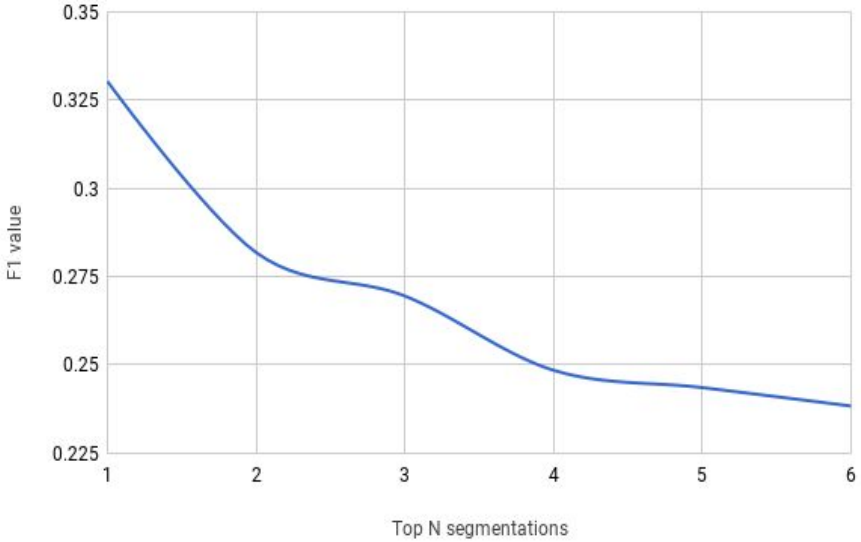
- Segmentation Task
  - Performed with WindowDiff as well as Boundary Similarity-based measures.
  - Achieved 80% accuracy.
- Segment Annotation
  - Performance of the system has been measured with accuracy metric and achieved 56% accuracy.
  - Performed 10-fold cross validation.

Measure	N-Best		Best
	Micro Avg.	Macro Avg.	Average
WindowDiff	0.411	0.420	0.384
B-F1	0.784	0.779	0.802

Individual Performance of Query Segmentation Model

System Evaluation & Performance

- Best performance obtained by taking Top-1 segmentation.
- Achieved 33% F1 score in holistic evaluation.



Questions ?

Thank You

https://ndl.iitkgp.ac.in/