



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

computer architecture book by william stallings

English

User Query

IEEE Xplore Digital Library
Non-stalling counterflow architecture
Source: IEEE Xplore Digital Library
Author: Miller, M.F. | Janik, K.J. | Shih-Lien Lu
Research | Reading | UG and PG
ENG

Abstract: The counterflow pipeline concept was originated by Sproull et al (1994) to demonstrate the concept of asynchronous circuits. This architecture relies on distributed decision making and localized clock. [View more](#)

IEEE Xplore Digital Library
Book reviews - Computer network architectures
Source: IEEE Xplore Digital Library
Author: Li, V.
UG and PG
ENG

More William
A Busy Day
Source: Librivox
Author: Crompton, Richmal
ENG

Abstract: More William is the second William collection in the much acclaimed Just William series by Richmal Crompton. It is a sequel to the book Just William. The book was first published in 1922. (Summary by [View more](#))

More William
The Revenge
Source: Librivox
Author: Crompton, Richmal
ENG

Abstract: More William is the second William collection in the much acclaimed Just William series by Richmal Crompton. It is a sequel to the book Just William. The book was first published in 1922. (Summary by [View more](#))

More William
William and the Smuggler
Source: Librivox
Author: Crompton, Richmal
ENG

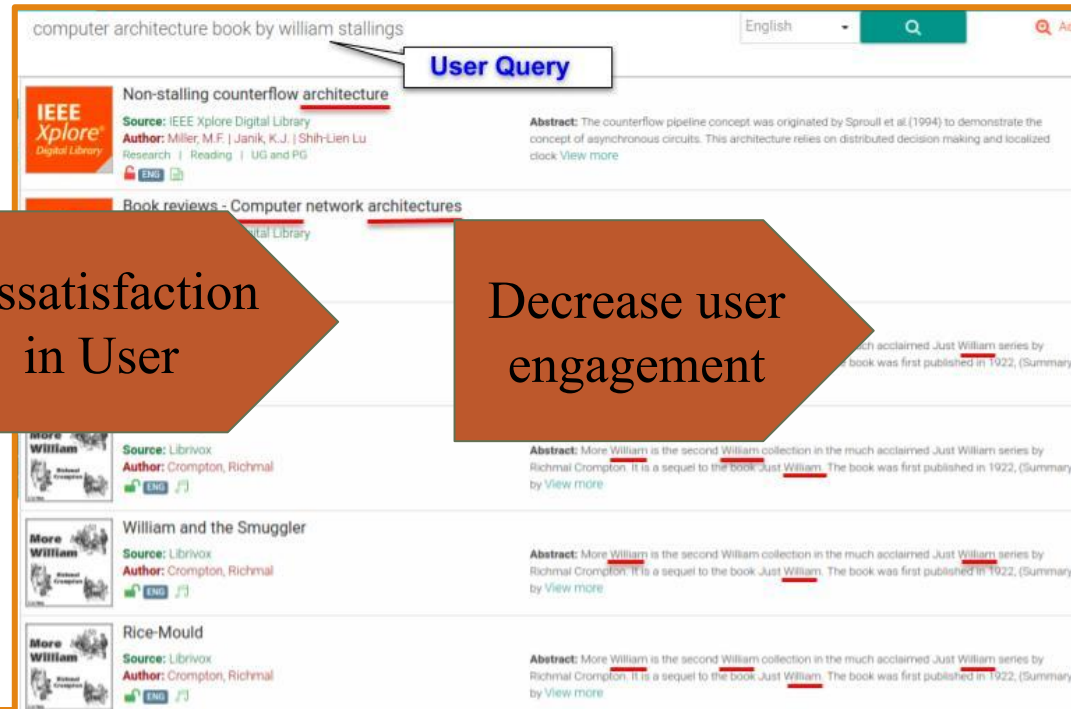
Abstract: More William is the second William collection in the much acclaimed Just William series by Richmal Crompton. It is a sequel to the book Just William. The book was first published in 1922. (Summary by [View more](#))

More William
Rice-Mould
Source: Librivox
Author: Crompton, Richmal
ENG

Abstract: More William is the second William collection in the much acclaimed Just William series by Richmal Crompton. It is a sequel to the book Just William. The book was first published in 1922. (Summary by [View more](#))

Motivation | Search Problem in Digital Library

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



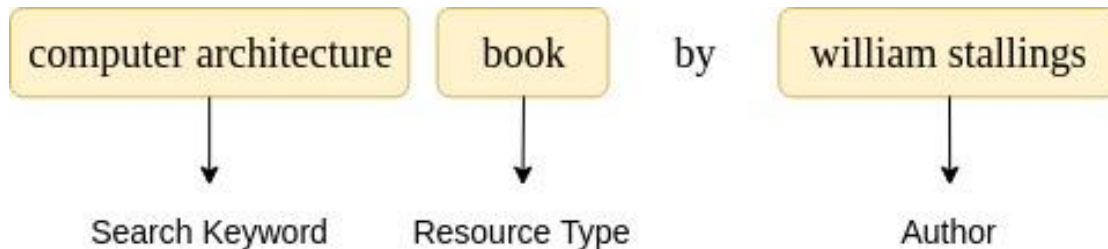
- ❖ System performance depression
 - Result according to keyword match
 - No relevant item in result list

Dissatisfaction in User

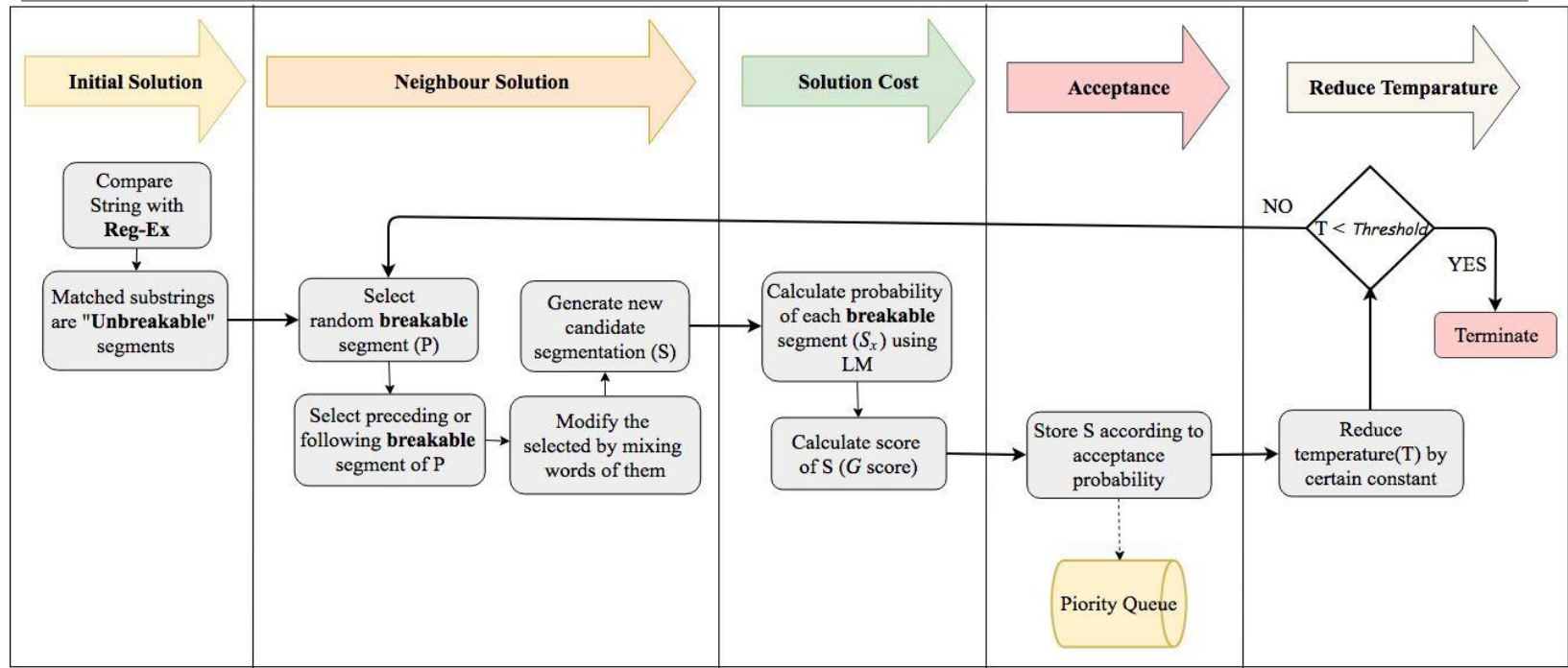
Decrease user engagement

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:



Query Segmentation | Simulated Annealing

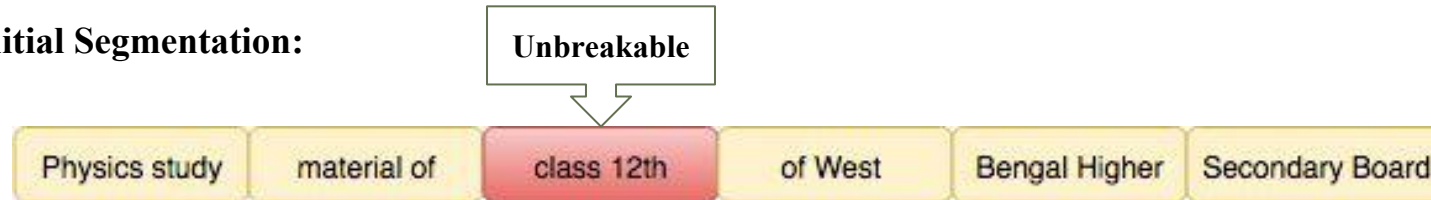


Query Segmentation | Example

Query String:

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

Initial Segmentation:



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

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:

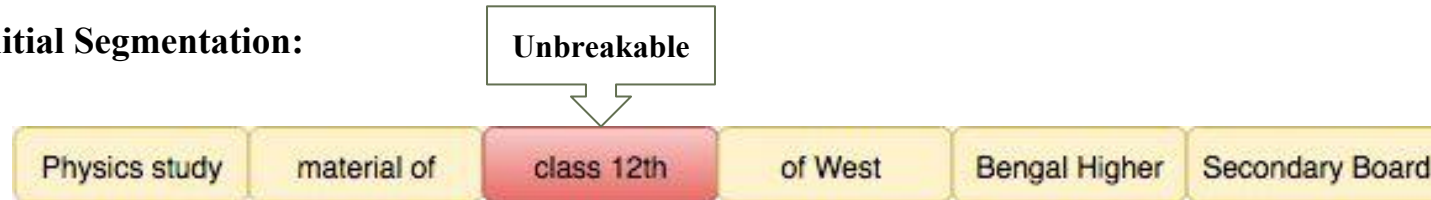


Query Segmentation | Example (cont.)

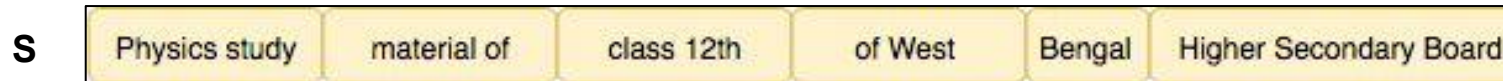
Query String:

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

Initial Segmentation:



New Segmentation:



Query Segmentation | Example (cont.)

Query String:

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

New Candidate Segmentation:

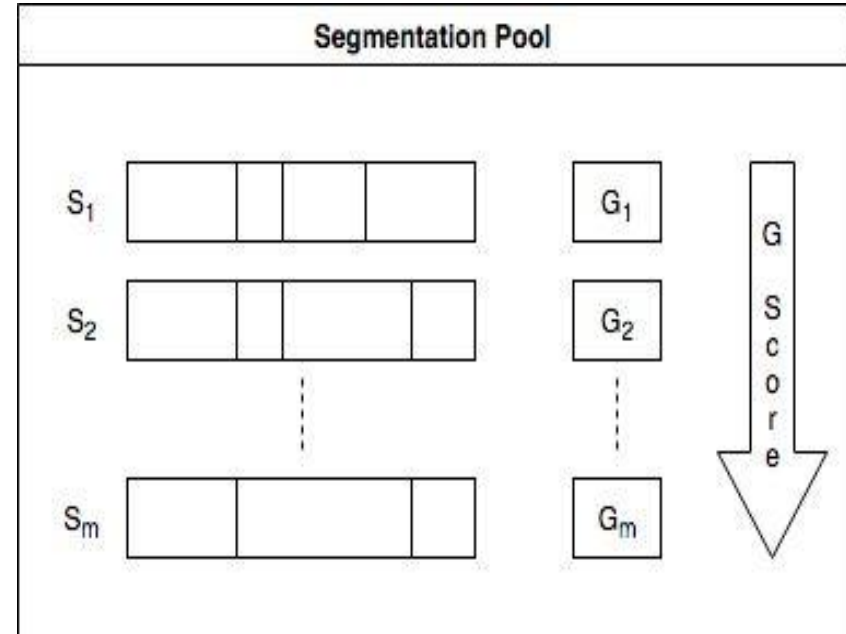
S

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

Equations	
Eq 1	$P(\text{Segment}_x) = \alpha P_{\text{Google}}(\text{Segment}_x) + (1 - \alpha) P_{\text{NDLI}}(\text{Segment}_x), \quad 0 \leq \alpha \leq 1$
Eq 2	$G(S) = \sum_{\text{Segment}_x \in S} P(\text{Segment}_x) \times (1 - \Delta(\text{Segment}_x))$ <p>Where,</p> $\Delta(\text{Segment}_x) = P(\text{End}(\text{Segment}_x), \text{Start}(\text{Segment}_{x+1}))$

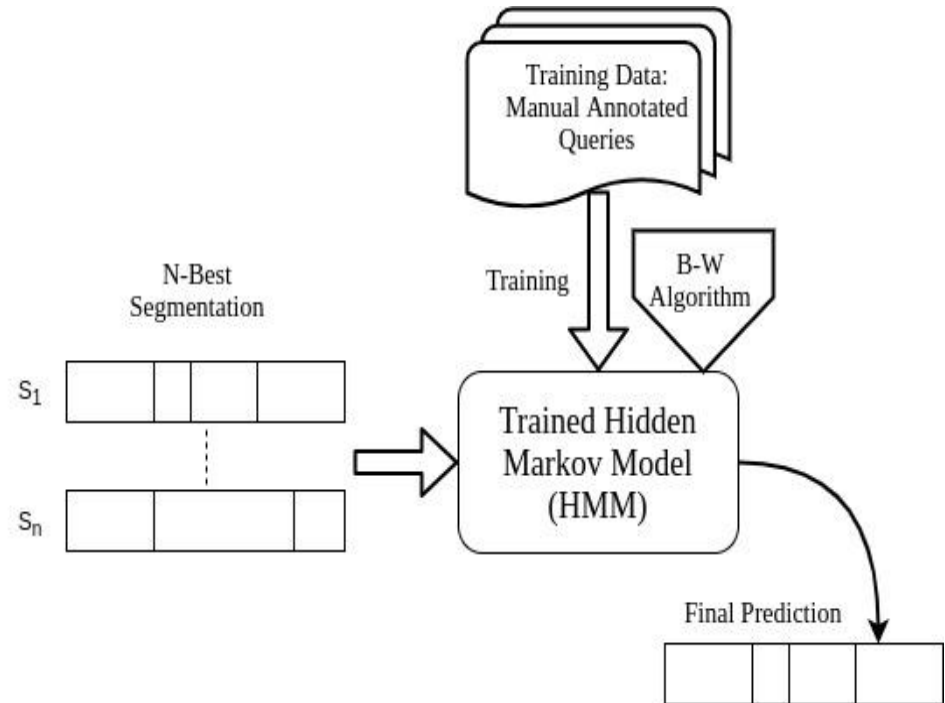
Query Segmentation | Selection

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



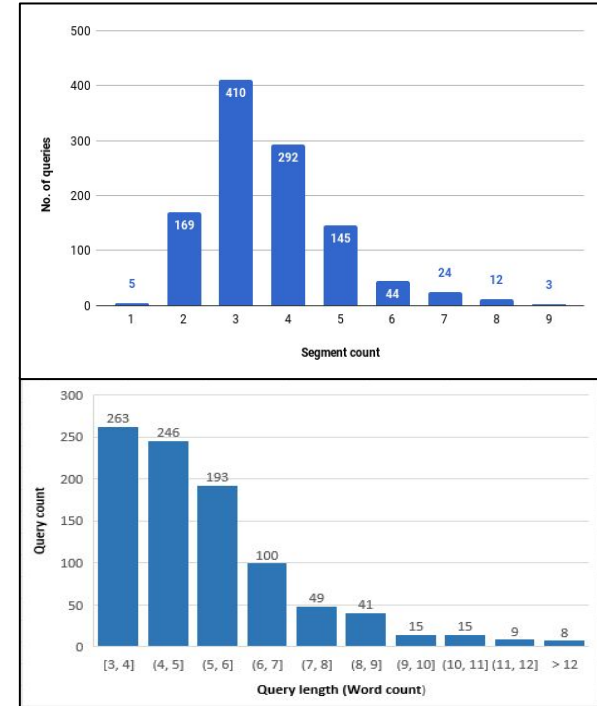
Query Segmentation | Annotation & Ranking

- 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

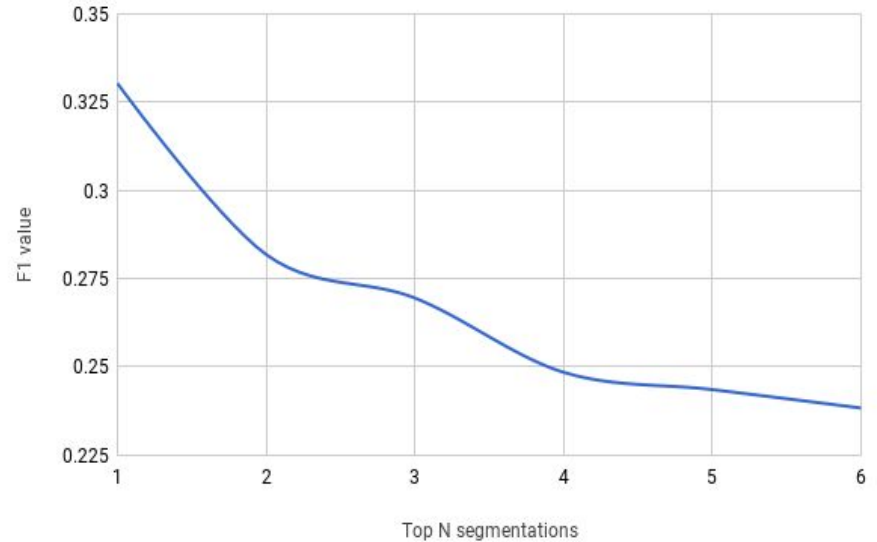
- Segmentation Task
 - Performed with WindowDiff as well as Boundary Similarity-based measures.
 - Achieved 80% accuracy.
- Segment Annotation
 - Performance of the system has been measured with accuracy metric and achieved 56% accuracy.
 - Performed 10-fold cross validation.

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/>

