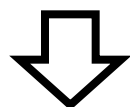


Query-Chain Focused Summarization

1. Introduction

Purpose

- This paper relates to text summarization.
 - Existing task:
 - (i) Generic Multi Document Summarization
 - (ii) Update Summarization
 - (iii) Query-Focused Summarization etc.



In contrast...

- This paper proposes new task, **Query-Chain Focused Summarization** to improve exploratory search

Purpose

- This paper focuses on text summarization.
 - Existing task:
 - (i) Generic Multi Document Summarization
 - (ii) Update Summarization
 - (iii) Query-Focused Summarization etc.

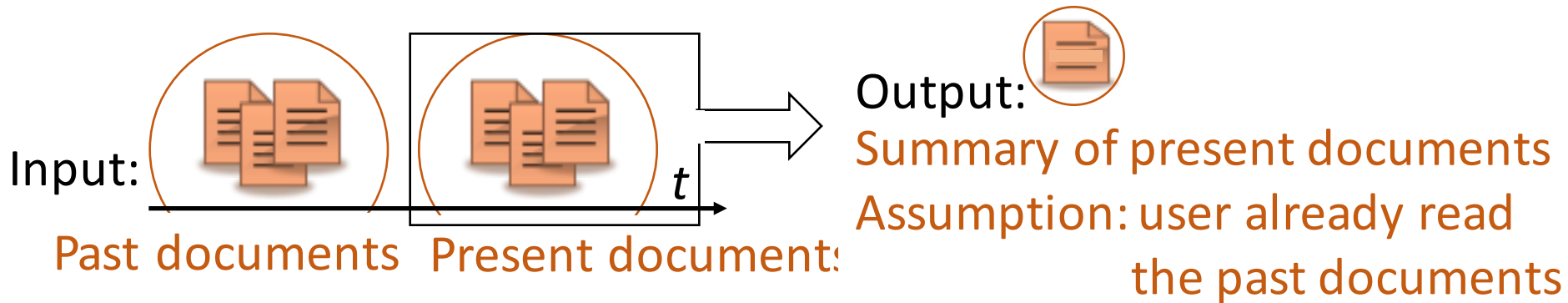


- This paper proposes new task, Query-Chain Focused Summarization to improve exploratory search

(i) Generic Multi Document Summarization



(ii) Update Summarization



(iii) Query-Focused Summarization



Purpose

- This paper focuses on text summarization.
 - Existing task:
 - (i) Generic Multi Document Summarization
 - (ii) Update Summarization
 - (iii) Query-Focused Summarization etc.

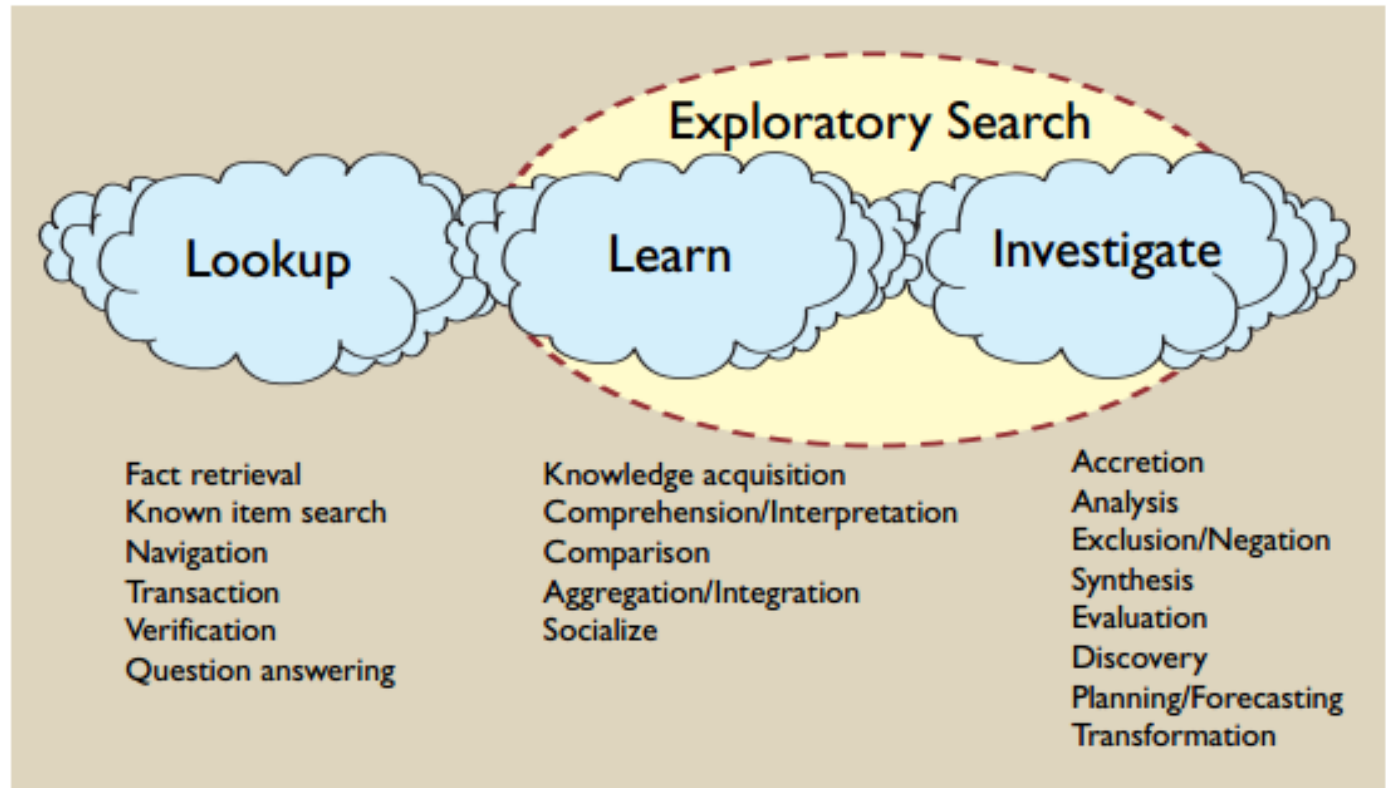


- This paper proposes new task, Query-Chain Focused Summarization to improve **exploratory search**

Preliminary

— Exploratory Search —

- Search to learn or investigate a new topic

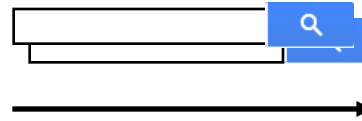


Preliminary

— Exploratory Search: example —



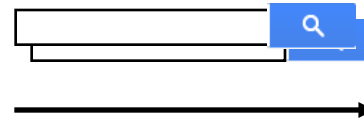
I want to go on a trip to France.
put up at a resort villa
in the countryside.



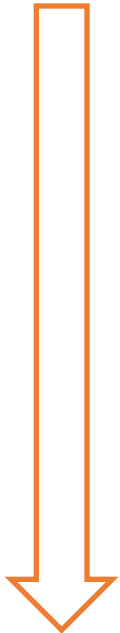
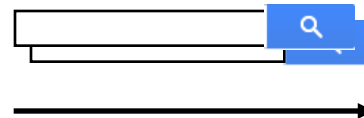
There are several resort villa
in Provence....



Provence is nice!
I plan to put up at a resort villa
in Provence!



No!! The hotel bill is too expensive!
Search another town...

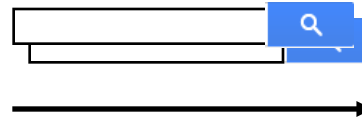


Preliminary

— Exploratory Search: example —



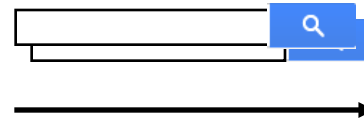
I want to go on a trip to France.
put up at a resort villa
in the countryside.



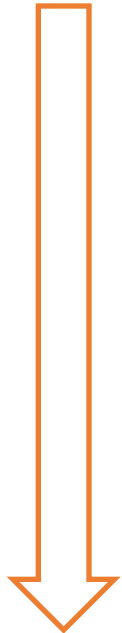
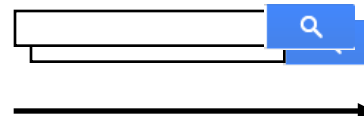
There are several resort villa
in Provence....



Provence is nice!
I plan to put up at a resort villa
in Provence!



No!! The hotel bill is too expensive!
Search another town...



Purpose

- This paper focuses on text summarization.
 - Existing task:
 - (i) Generic Multi Document Summarization
 - (ii) Update Summarization
 - (iii) Query-Focused Summarization etc.



- This paper proposes **new task, Query-Chain Focused Summarization** to improve exploratory search

Query-Chain Focused Summarization: Definition

- Define:

For each query in an exploratory search session, we aim to extract a summary that answers the information need of the user, while not repeating information already provided in previous steps.

- is similar to Query-Focused Summarization.
- is similar to Update Summarization.

Contribution

- The definition of a new summarization task: QCFS
- Construct novel Dataset for QCFS:
 - Novel dataset of Query-Sets with matching manual summarizations in the consumer health domain
- Adapt well-known multi-document algorithms to the task
- Propose a new algorithms to address the task of QCFS, based on a new LDA topic model variant

2. Query-Chain Focused Summarization

Query-Chain Focused Summarization: Formalization

- Formalization:

Given:

ordered chain of queries Q

a set of documents D

For each query $q_i \in Q$, a summary S_i is generated from D answering q_i

under the assumption that the user has already read the summaries $S_{0:i-1}$ for queries $q_{0:i-1}$

- This paper focuses on the *zoom in* aspect of the exploratory search process.

Query-Chain Focused Summarization: example

- A typical example of query chain in the consumer health domain:
 - query chain:
 - causes of asthma*
 - *asthma and allergy*
 - *asthma and mold allergy*
 - Reference set D :
 - documents relevant to the domain of Asthma
 - Task:
 - Generating one summary of D as an answer to each query
 - The successive answers do not repeat information already provided in a previous answer

4. Dataset Collection

Dataset Collection

- Build in Consumer Health domain
 - providing medical information ranging from layman and up to expert information
- Dataset is composed of
 - Query Chains
 - Manually selected from PubMed query logs
 - Document Set
 - Manually selected from various sites to contain relevant information about the queries
 - Manual Summaries
 - Created for each query some were created within the context of the query chain and some weren't

Dataset Collection

- Build in Consumer Health domain
 - providing medical information ranging from layman and up to expert information
- Dataset is composed of
 - Query Chains
 - Manually selected from PubMed query logs
 - Document Set
 - Manually selected from various sites to contain relevant information about the queries
 - Manual Summaries
 - Created for each query some were created within the context of the query chain and some weren't

Dataset Collections

— Query Chains —

- Using PubMed query logs
- Extract laymen queries relating to four topics:

- | | |
|----------------|------------------------|
| 1. Asthma | 3. Obesity |
| 2. Lung Cancer | 4. Alzheimer's disease |

Procedure

1. Extract a single day query log
2. Extract sessions which contained the terms from the query log
3. Sessions containing search tags (such as “[Author]”) were removed
4. The sessions were then manually examined



Dataset Collection

— Example of Query Chain —

- Only *zoom in* query chains of length 3 at most

Asthma:

Asthma causes → asthma allergy → asthma mold allergy;

Asthma treatment → asthma medication → corticosteroids;

Exercise induced asthma → exercise for asthmatic;

Atopic dermatitis → atopic dermatitis medications → atopic dermatitis side effects;

Atopic dermatitis → atopic dermatitis children → atopic dermatitis treatment;

Atopic dermatitis → atopic dermatitis exercise activity → atopic dermatitis treatment;

Figure 1: Queries Used to Construct Dataset

Dataset Collection

- Build in Consumer Health domain
 - providing medical information ranging from layman and up to expert information
- Dataset is composed of
 - Query Chains
 - Manually selected from PubMed query logs
 - Document Set
 - Manually selected from various sites to contain relevant information about the queries
 - Manual Summaries
 - Created for each query some were created within the context of the query chain and some weren't

Dataset Collections

— Document Set —

- Using Wikipedia, WebMD, NHS
 - well-known and reliable consumer health websites
- Asked medical experts to construct four document collections about the four topics
 - Each document provide general information relevant to the queries

Dataset Collection

- Build in Consumer Health domain
 - providing medical information ranging from layman and up to expert information
- Dataset is composed of
 - Query Chains
 - Manually selected from PubMed query logs
 - Document Set
 - Manually selected from various sites to contain relevant information about the queries
 - Manual Summaries
 - Created for each query some were created within the context of the query chain and some weren't

Dataset Collections

— Manual Summaries —

- Asked medical students to manually produce summaries of these four document collections for each query-chain

Instruction

1. Construct a text of up to 250 words that provides a good answer to each query in the chain
2. Assume that the person reading the summaries is familiar with the previous summaries in the chain
 - avoid redundancy
3. Not shown the next steps in the chain

Interface for annotators

Related Documents:

< 1 2 3 ... 11 12 >

- [aafa.org allergic-athma.txt](#)
- [aafa.org allergy.txt](#)
- [aafa.org alternative.txt](#)
- [aafa.org control medicines.txt](#)
- [aafa.org eye-allergies.txt](#)
- [aafa.org food-drug-allergies.txt](#)
- [aafa.org indoor-allergies.txt](#)
- [aafa.org insect-allergies.txt](#)
- [aafa.org letx-allergies.txt](#)
- [aafa.org medications general.txt](#)

Selected Document: aafa.org food-drug-allergies.txt

- 1 Food allergies and allergic reactions to certain drugs are serious.
- 2 They are characterized by a broad range of allergic reactions to ingredients in the foods we eat or the medications we take.
- 3 Food allergy is an overreaction of the immune system, different than food intolerance or food sensitivity.
- 4 The U.S. Food Allergy Labeling Consumer Protection Act (FALCPA) now requires food labels to clearly identify all allergen ingredients (even if it's a spice or flavoring), and to discourage labels with may contain' statements.

Copied Text

For more severe cases, your doctor may prescribe oral corticosteroids, such as prednisone, or an intramuscular injection of corticosteroids to reduce inflammation and to control symptoms.

[Remove](#) [Show Source](#) [Copy](#)

These medications are effective, but cant be used long term because of potential serious side effects, which include cataracts, loss of bone mineral (osteoporosis), muscle weakness, decreased resistance to infection, high blood pressure and thinning of the skin.

[Remove](#) [Show Source](#) [Copy](#)

Immunomodulators

[Remove](#) [Show Source](#) [Copy](#)

Due to possible concerns about the effect of these medications on the immune system when used for prolonged periods, the Food and Drug Administration recommends that Elidel and Protopic be used only when other treatments have failed or if someone cant tolerate other treatments.

[Remove](#) [Show Source](#) [Copy](#)

Alternative medicine.

[Remove](#) [Show Source](#) [Copy](#)

Many alternative therapies including chamomile, evening-primrose oil, witch hazel extract and borage seed oil have been touted as possible ways to treat atopic dermatitis (eczema).

[Remove](#) [Show Source](#) [Copy](#)

However, theres no conclusive evidence that any of these alternative therapies are effective.

[Remove](#) [Show Source](#) [Copy](#)

Asthma is a chronic disease that affects your airways. If you have asthma, the inside walls of your airways become sore and swollen. That makes them very sensitive, and they may react strongly to things that you are allergic to or find irritating. When your airways react, they get narrower and your lungs get less air. This can cause wheezing, coughing, chest tightness and trouble breathing. the causes of asthma are unknown. researchers argue if they are genetic or environmental . the the genetic causes are a tendency to develop allergies, called atopy (AT-o-pe), Parents who have asthma, Certain respiratory infections during childhood. Contact with some airborne allergens or exposure to some viral infections in infancy or in early childhood when the immune system is developing if asthma or atopy runs in your family, exposure to irritants (for example, tobacco smoke) might make your airways more reactive to substances in the air. The Hygiene Hypothesis, one theory researchers have for what causes asthma is called the hygiene hypothesis. They believe that our Western lifestyle with its emphasis on hygiene and sanitation has resulted in changes in our living conditions and an overall decline in infections in early childhood. Many young children no longer have the same types of environmental exposures and infections as children did in the past. This affects the way that young children's immune systems develop during very early childhood, and it may increase their risk for atopy and asthma.

Statistics on the collected dataset

Document sets	# Docs	# Sentences	#Tokens / Unique
Asthma	125	1,924	19,662 / 2,284
Lung-Cancer	135	1,450	17,842 / 2,228
Obesity	289	1,615	21,561 / 2,907
Alzheimer's Disease	191	1,163	14,813 / 2,508

Queries	# Sessions	# Sentences	#Tokens / Unique
Asthma	5	15	36 / 14
Lung-Cancer	6	18	71 / 25
Obesity	6	17	45 / 29
Alzheimer's Disease	4	12	33 / 16

Manual Summaries	# Docs	# Sentences	#Tokens / Unique
Asthma	45	543	6,349 / 1,011
Lung-Cancer	54	669	8,287 / 1,130
Obesity	51	538	7,079 / 1,270
Alzheimer's Disease	36	385	5,031 / 966

Verifying the dataset

- The summaries for advanced (second or third) query should contain updated information

Verify:

- Asked additional annotators to create manual summaries for advanced queries **without ever seeing the previous queries**
- Compare the mean ROUGE score
[With-Context]:
each manual summary
 vs. all other summaries about the same query
[Without-Context]:
The mean ROUGE scores of the additionally created summaries

Verifying the dataset

— ROUGE score —

[With-Context]

ROUGE-1 = 0.52, ROUGE-2 = 0.22, ROUGE-SU4 = 0.13

[Without-Context]

ROUGE-1 = 0.40, ROUGE-2 = 0.22, ROUGE-SU4 = 0.01

Except for the ROUGE-2, results showed statistically significant difference with 95% confidence interval

5. Algorithms

Algorithms

- Baselines (adapted existing methods to QCFS)
 - Focused KLSum
 - KL-Chain-Update
 - ChainSum
- New Algorithm for QCFS
 - Adapted LexRank

Algorithms

- Baselines (adapted existing methods to QCFS)
 - Focused KLSum
 - KL-Chain-Update
 - ChainSum
- New Algorithm for QCFS
 - Adapted LexRank

Focused KLSum

— Overview —

- A variation of KLSum which is adapted to Query-Focused Summarization
 - Originally, KLSum is a method for generic multi-document summarization
 - KLSum tries to minimize the KL-divergence between the summary and document set unigram distribution
- Focused KLSum used KLSum on the 10 documents with best TF/IDF matches to the query

Focused KLSum

— Algorithm —

Given a query q ,

1. Select a focused subset of documents from D : $D(q)$
2. Search $S^* = \operatorname{argmin}_{|S| < L} KL(P_{D(q)} || P_S)$,

$$\text{where } KL(P || Q) = \sum_w \log \left(\frac{P(w)}{Q(w)} \right) P(w)$$

- This search is performed by greedy manner
- $D(q)$ is determined by selecting the top-10 documents in D ranked by $TF \times IDF$ scores to the query

Notation

S : candidate summary

$P_{D(q)}$: unigram distribution of $D(q)$

P_S : unigram distribution of candidate summary

L : maximum length of summary

Algorithms

- Baselines (adapted existing methods to QCFS)
 - Focused KLSum
 - KL-Chain-Update
 - ChainSum
- New Algorithm for QCFS
 - Adapted LexRank

KL-Chain-Updates

— Overview —

- A variation of KLSum that answers a query chain $(q_1, q_2, \dots, q_i, q_{i+1})$
- Try to minimize the KL-divergence of the summary and the top 10 TF/IDF retrieved documents for query q_{i+1}
- Select sentences for q_{i+1} assuming the smoothed distribution of the previous summaries (q_1, \dots, q_i) is already part of the summary (eliminates redundancy)

KL-Chain-Updates

— unigram distribution —

Unigram Distribution of Word w	
KLSum	KL-Chain-Updates
$\frac{\text{Count}(w, \text{CurrentSum})}{\text{Length}(\text{CurrentSum})}$	$\frac{\text{Count}(w, \text{CurrentSum}) + \frac{\text{Count}(w, \text{PreviousSum})}{\text{Smoothing Factor}}}{\text{Length}(\text{CurrentSum}) + \frac{\text{Length}(\text{PreviousSum} \cap \text{CurrentSum})}{\text{Smoothing Factor}}}$

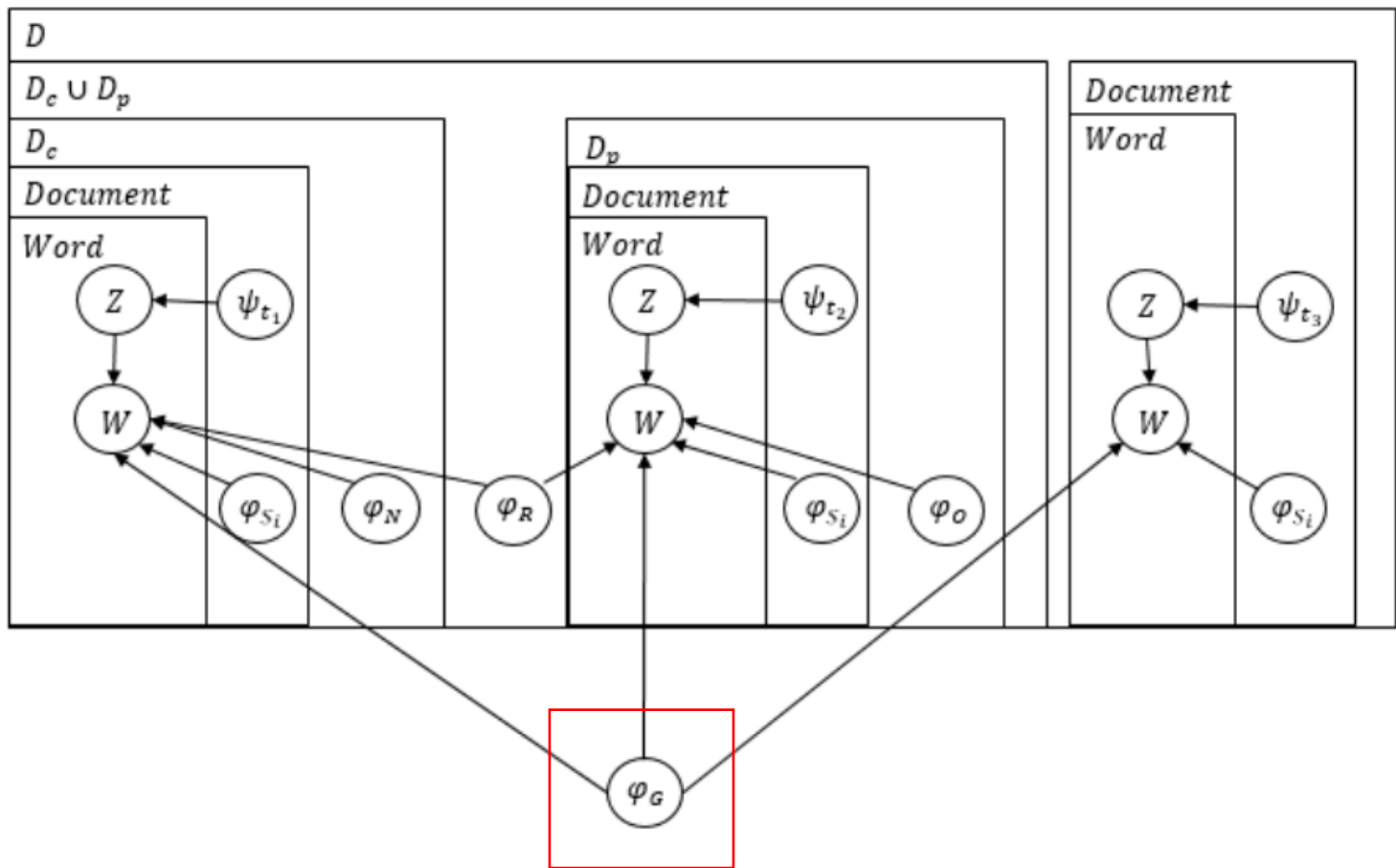
Algorithms

- Baselines (adapted existing methods to QCFS)
 - Focused KLSum
 - KL-Chain-Update
 - ChainSum
- New Algorithm for QCFS
 - Adapted LexRank

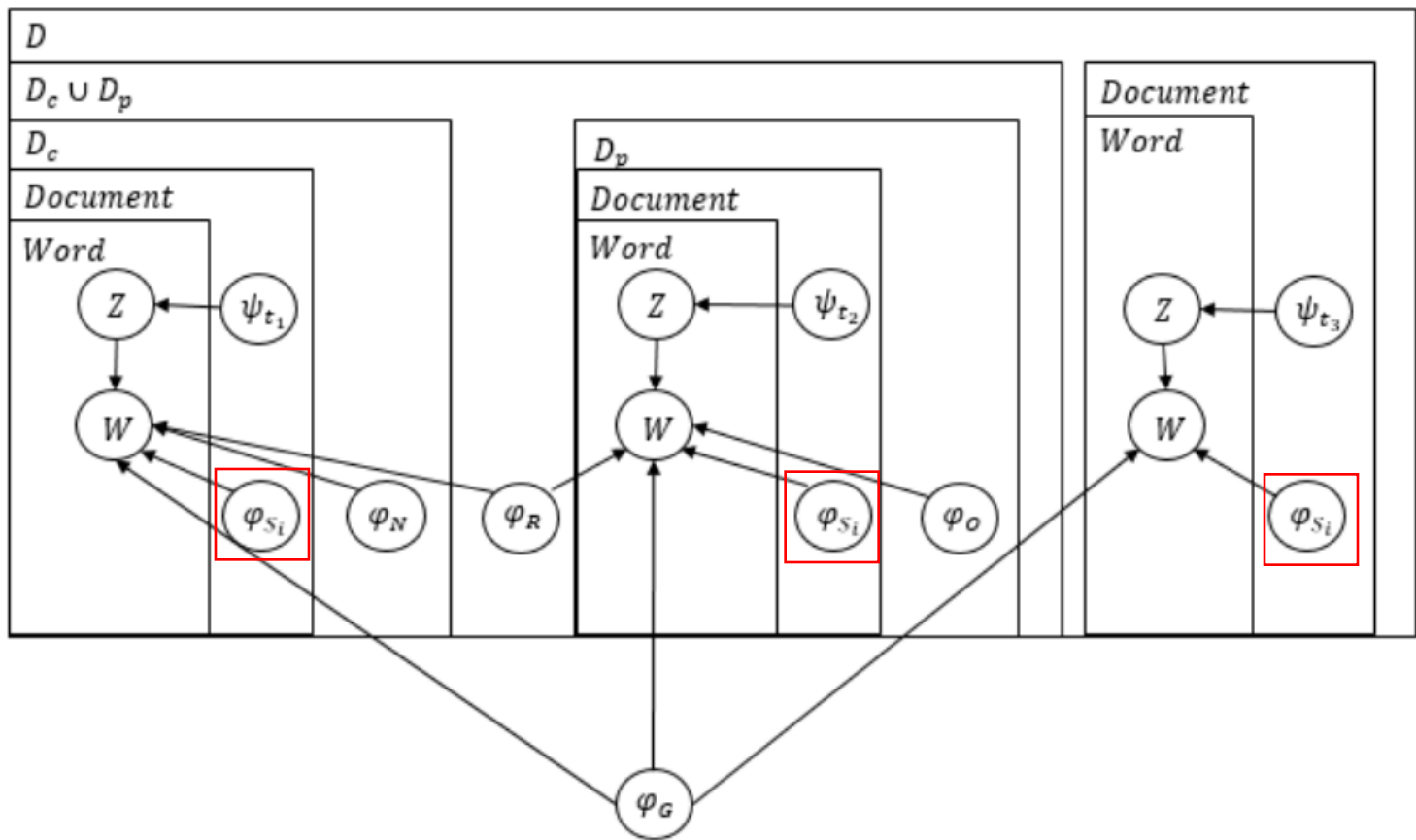
ChainSum

— Overview —

- Adaptation of TopicSum to the QCFS task
- Develop a novel topic model
 - to identify words that are associated to the current query and not shared with the previous queries
- For each query in a chain
 - Consider the documents D_c : “good answers” to the query
 - Also consider the documents D_p : documents used to answer the previous steps of the chain
- Assumption:
 - D_c and D_p are observable
 - But in their implementation, they select these subsets by ranking the documents for the query based on $TF \times IDF$ similarity



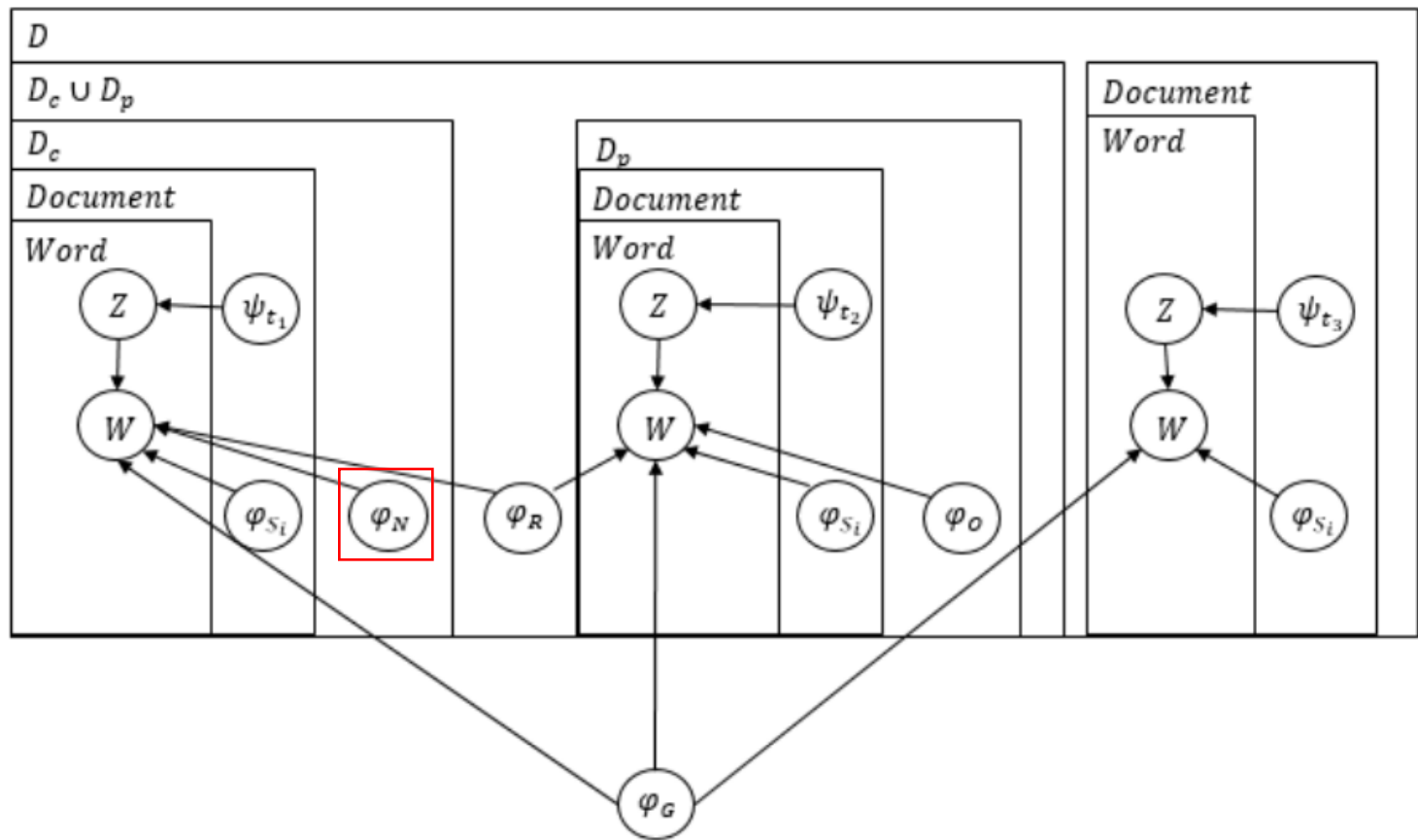
1. G is the *general words* topic,
 - capture stop words, and non-topic specific vocabulary φ_G is drawn for all the documents from $Dirichlet(V, \lambda_G)$



2. S_i is the *document specific* topic,

— represents words, which are local for a specific document

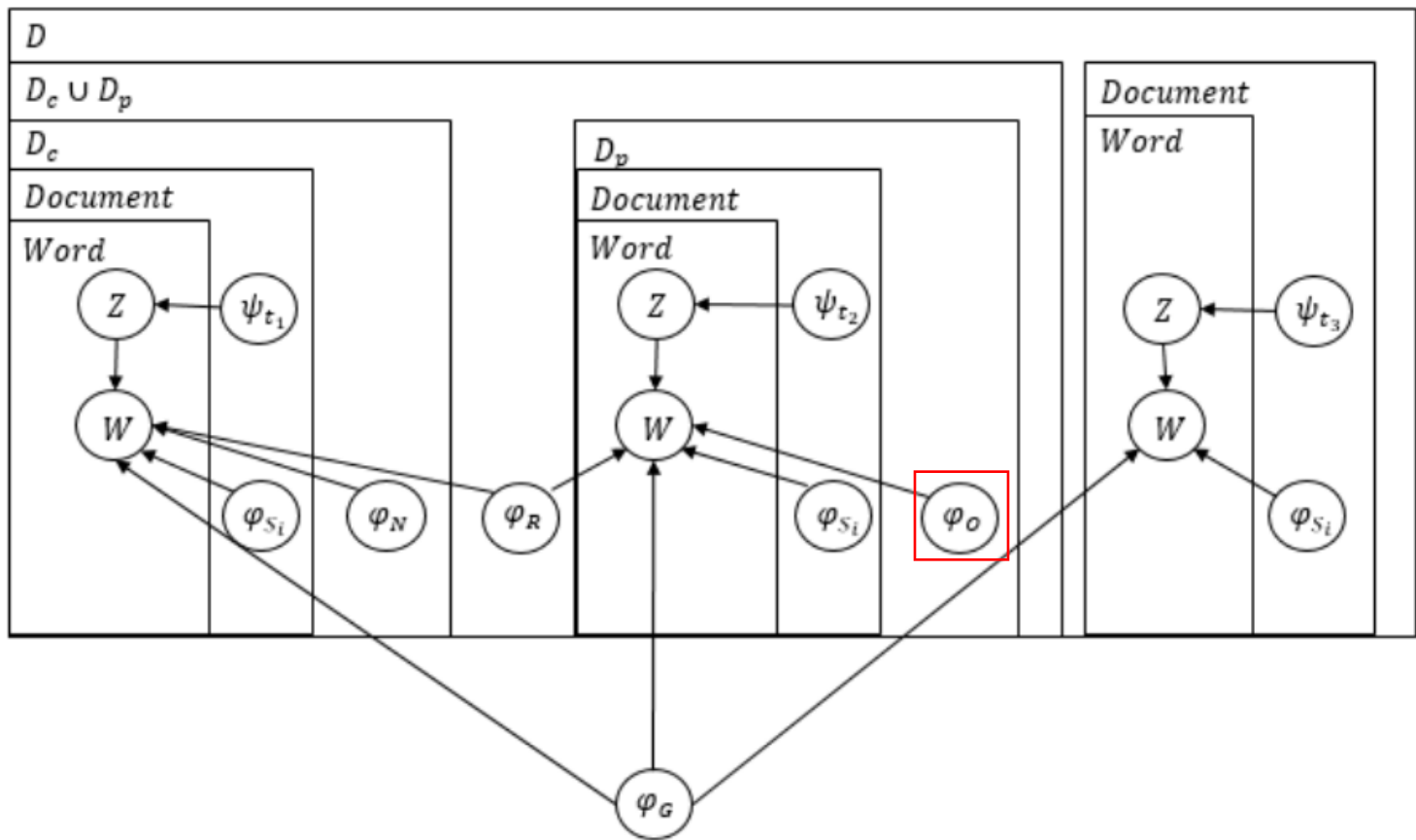
φ_{S_i} is drawn for each document from $Dirichlet(V, \lambda_{S_i})$



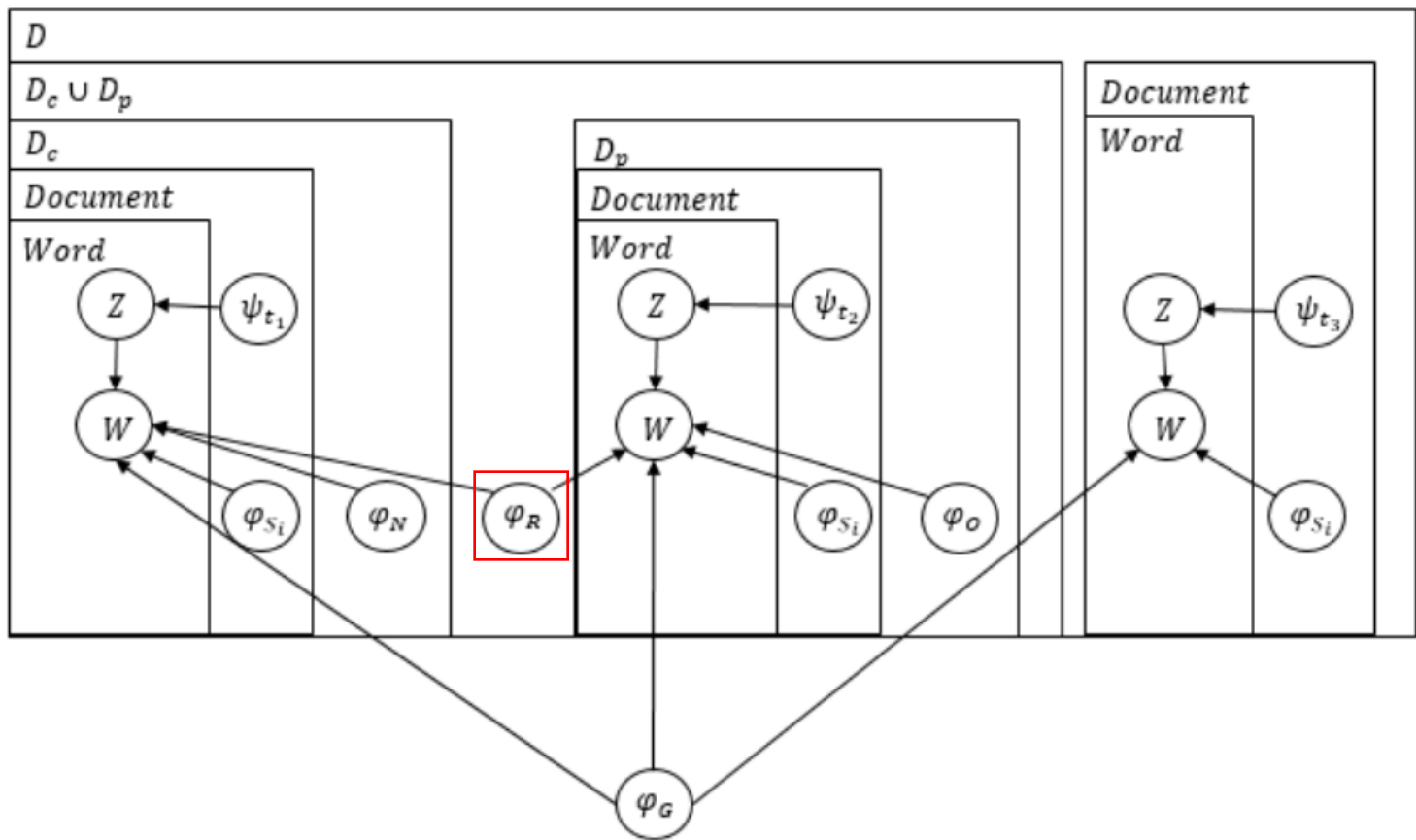
3. N is the *new content* topic,

— capture words that are characteristic for D_c

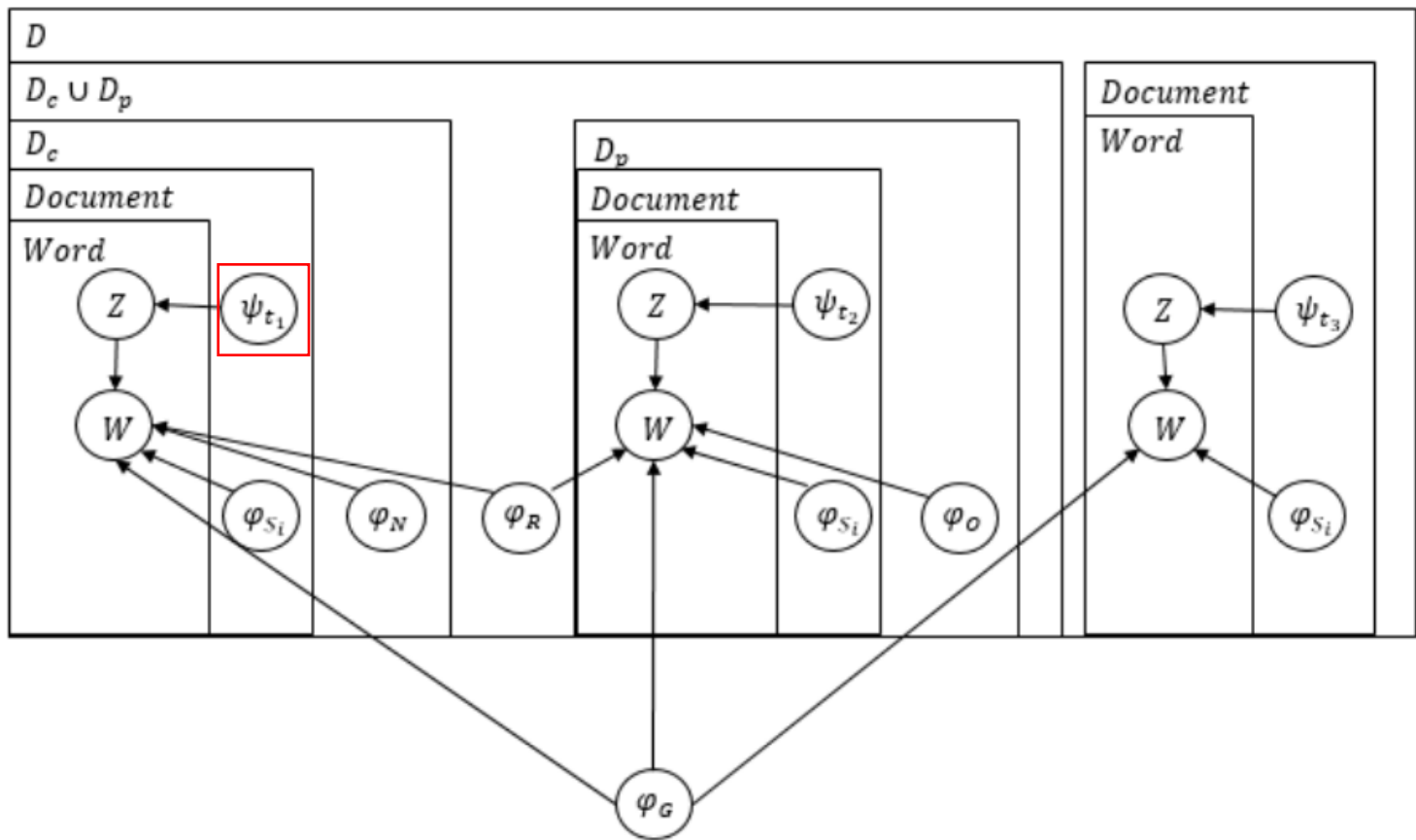
φ_N is drawn for all the documents in D_c from $Dirichlet(V, \lambda_N)$



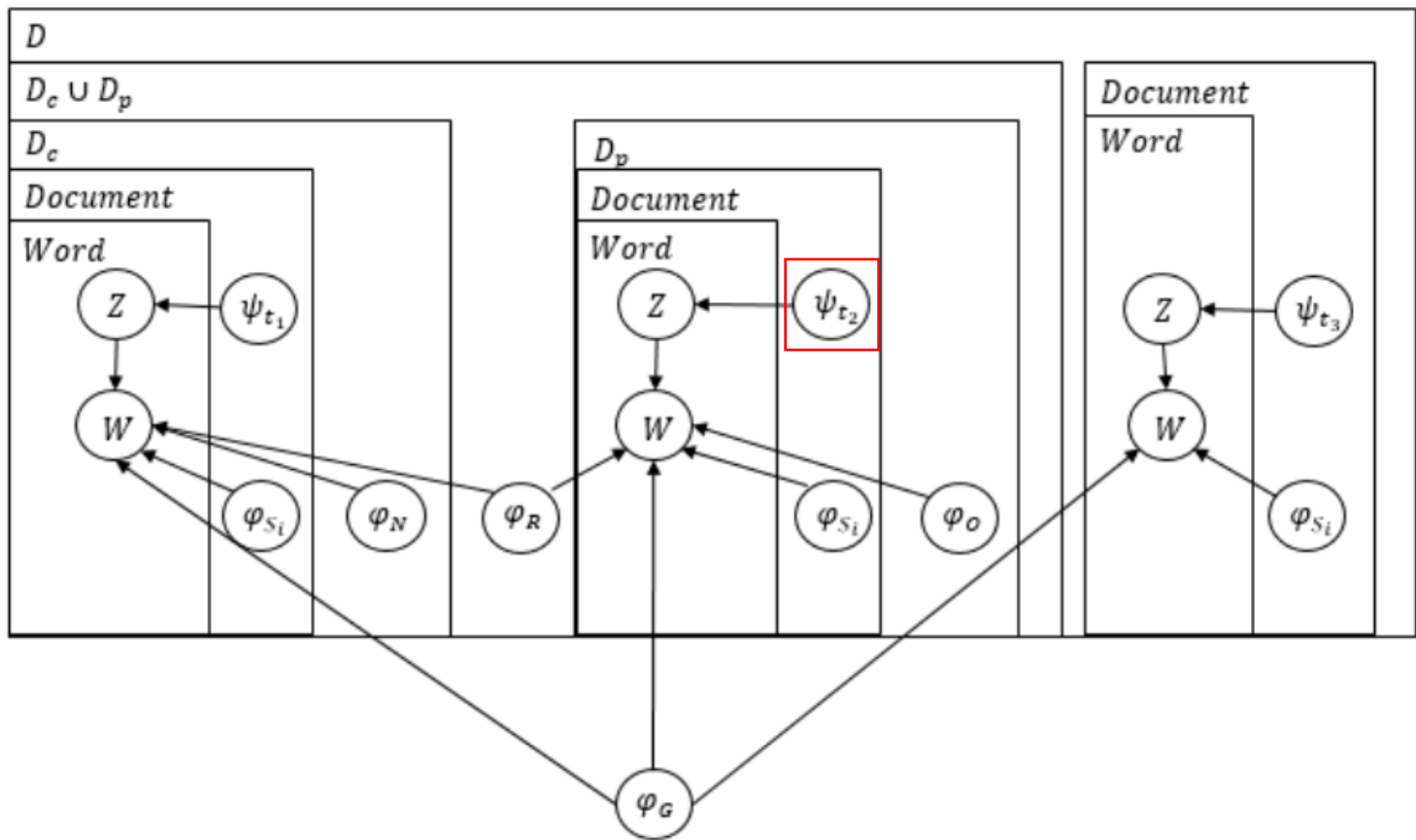
4. O captures old content from D_P ,
 φ_O is drawn for all the documents in D_P from $Dirichlet(V, \lambda_O)$



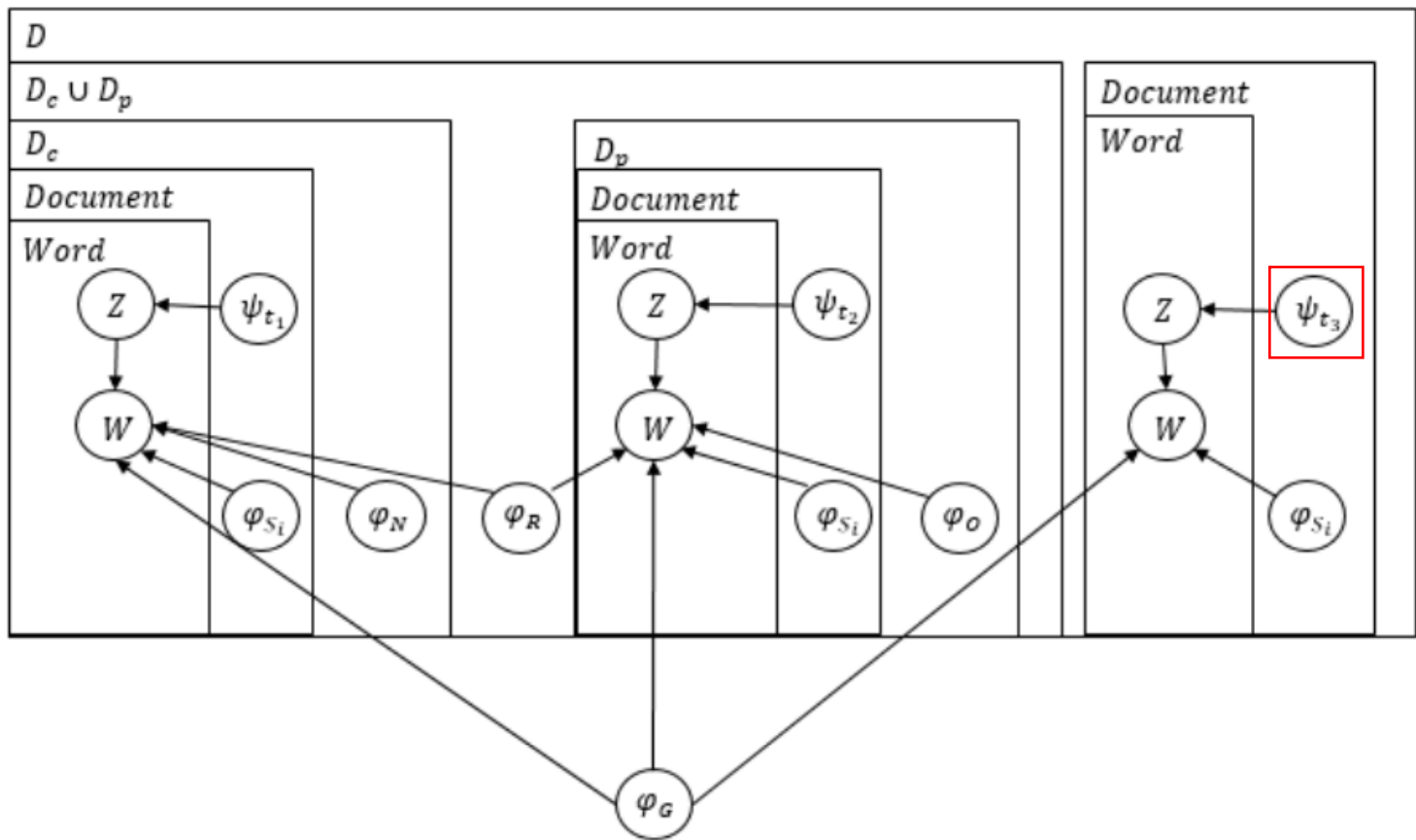
5. R captures *redundant information* between D_c and D_p ,
 φ_R is drawn for all the documents in $D_p \cup D_c$ from $Dirichlet(V, \lambda_R)$



- For documents from D_c draw from the distribution ψ_{t_1} over topics (G, N, R, S_i) from a Dirichlet prior with pseudo-counts (10.0, 15.0, 15.0, 1.0). For each word in the document, we draw a topic Z from ψ_{t_1} , and a word W from the topic indicated by Z .



- For documents from D_P , draw from the distribution ψ_{t_2} over topics (G, O, R, S_i) from a Dirichlet prior with pseudo-counts $(10.0, 15.0, 15.0, 1.0)$.
The words are drawn in the same manner as in t_1 .



8. For documents from $D \setminus (D_C \cup D_P)$, draw from the distribution ψ_{t_3} over topics (G, S_i) from a Dirichlet prior with pseudo-counts (10.0, 1.0). The words are drawn in the same manner as in t_1 .

ChainSum

— Procedure —

- Apply KLSum only on words that are assigned to the *new content* topic after the topic model is applied to the current query

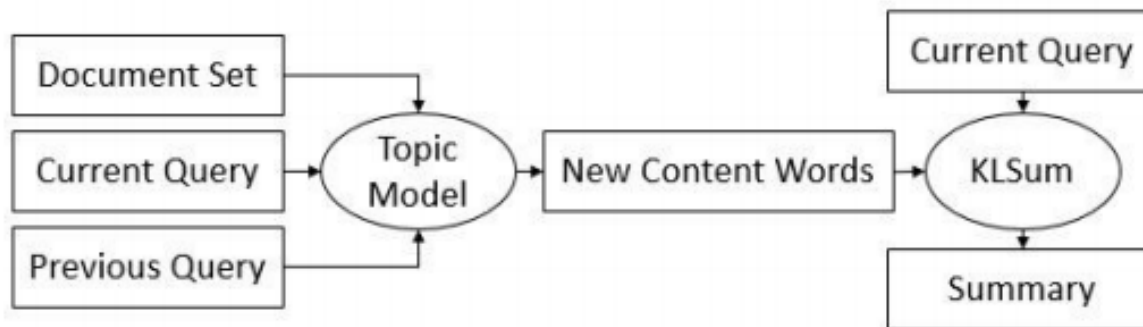


Figure 4 ChainSum Architecture

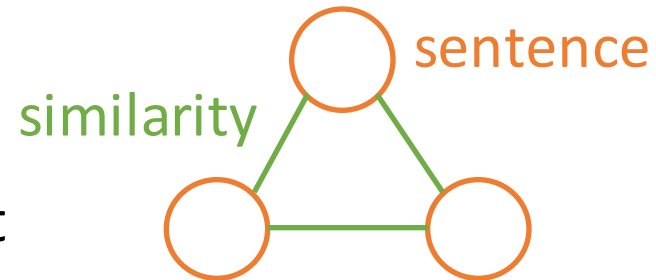
Algorithms

- Baselines (adapted existing methods to QCFS)
 - Focused KLSum
 - KL-Chain-Update
 - ChainSum
- New Algorithm for QCFS
 - Adapted LexRank

LexRank

— Overview —

- A stochastic graph-based method for computing the importance of sentences
 - Original algorithm does not address update or query-focused variants
- A summary is created by extracting top ranking sentences



1. Create a graph where nodes represent the sentences from the text
 - weighted edges represent the cosine similarity of each sentence's $TF \times IDF$ vectors
2. PageRank is run to rank sentences

Adapted LexRank

— Overview —

- A variation of LexRank which is adapted to QCFS
- For the adaptation,
 - Extend the sentence representation scheme to capture semantic information
 - Refine the model of sentence's similarity so that it captures query answering instead of centrality
 - Adapt to update of query

Adapted LexRank

— Overview —

- A variation of LexRank which is adapted to QCFS
- For the adaptation,

- Extend the sentence representation scheme to capture semantic information
- Refine the model of sentence's similarity so that it captures query answering instead of centrality
- Adapt to update of query

Adapted LexRank

— sentence representation scheme —

- Original LexRank only deal with lexical information for sentence representation
- Extend the sentence representation scheme to capture semantic information
 - ⇒ Using Wikipedia terms and UMLS terms
- How to detect Wikipedia and UMLS terms?
 - ⇒ Wikifier, HealthTermFinder

Adapted LexRank

— sentence representation scheme —

- edges scoring : $Score(U, V)$
 - use the sum of Lexical Semantic Similarity (LSS) functions on:
 - Lexical terms, Wikipedia terms, UMLS terms
 - $Score(U, V) = LSS_{lexical}(U, V) + a * LSS_{wiki}(U, V) + b * LSS_{UMLS}(U, V)$

$$\text{where } LSS(S_1, S_2) = \frac{\sum_i \max_j \left(\frac{Sim(w_i^1, w_j^2)}{Sim(w_i^1, w_j^1)} \right) \times IDF(W_i^1)}{\sum_i IDF(w_i^1)}$$

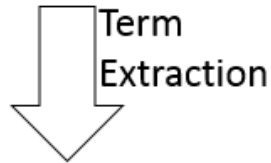
- For lexical terms Sim is identity function
- For Wikipedia terms Sim is calculated by Wikiminer
- For UMLS terms Sim is calculated by Ted Pedersen UMLS similarity function

Adapted LexRank

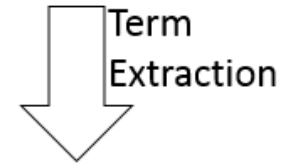
— similarity example —

"Asthma is a common chronic inflammatory disease."

"inhaler commonly used for long-term control."



['Asthma', 'Inflammation']



['Inhaler']

$\text{Sim}(\text{'Asthma'}, \text{'Inhaler'}) = 0.725798946382$

$\text{Sim}(\text{'Inflammation'}, \text{'Inhaler'}) = 0.559059383727$

Adapted LexRank

— Overview —

- A variation of LexRank which is adapted to QCFS
- For the adaptation,

- Extend the sentence representation scheme to capture semantic information
- Refine the model of sentence's similarity so that it captures query answering instead of centrality
- Adapt to update of query

Adapted LexRank

— Refine the Model —

- To handle the query, added to the LexRank graph:
 - a new node representing the query
 - all the required edges
- Change the page rank algorithm to achieve similarity
 - In PageRank, the damping factor jumps to a random node in the graph
 - ⇒ Allowed the damping factor to only jump back to the query node

Simulate the probability of reaching a sentence when starting a random walk at the query

Adapted LexRank

— Overview —

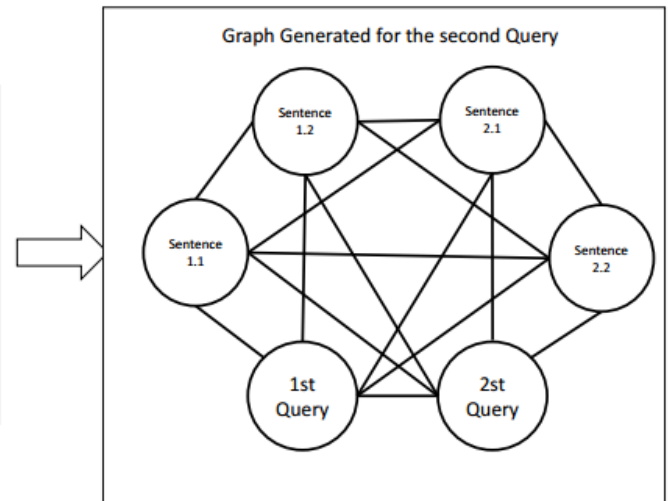
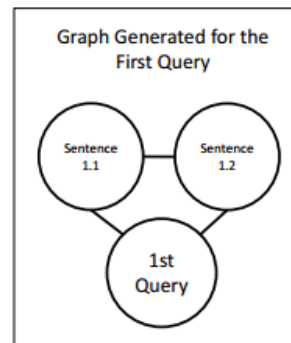
- A variation of LexRank which is adapted to QCFS
- For the adaptation,
 - Extend the sentence representation scheme to capture semantic information
 - Refine the model of sentence's similarity so that it captures query answering instead of centrality
 - Adapt to update of query

Adapted LexRank

— Adapt to update of query —

- To adapt the query updates, the model is incorporated into the following changes:
 - Did not create a new graph
 - Merged the graph from the previous query and with the new query and sentences from the top N documents fetched

⇒ Enable a drill down effect to the summary



Adapte LexRank

— Avoid Redundancy —

- After ranking, select only sentences that are:
 - different from sentences that are selected for the current summary
 - also different from sentences that are selected for the previous summaries in the session

6. Evaluation

Evaluation Dataset

- Using dataset created for QCFS
- Added semantic tags:
 - 10% of the tokens had Wikipedia annotations
 - 33% had a UMLS annotation

Results

— Automatic Evaluation —

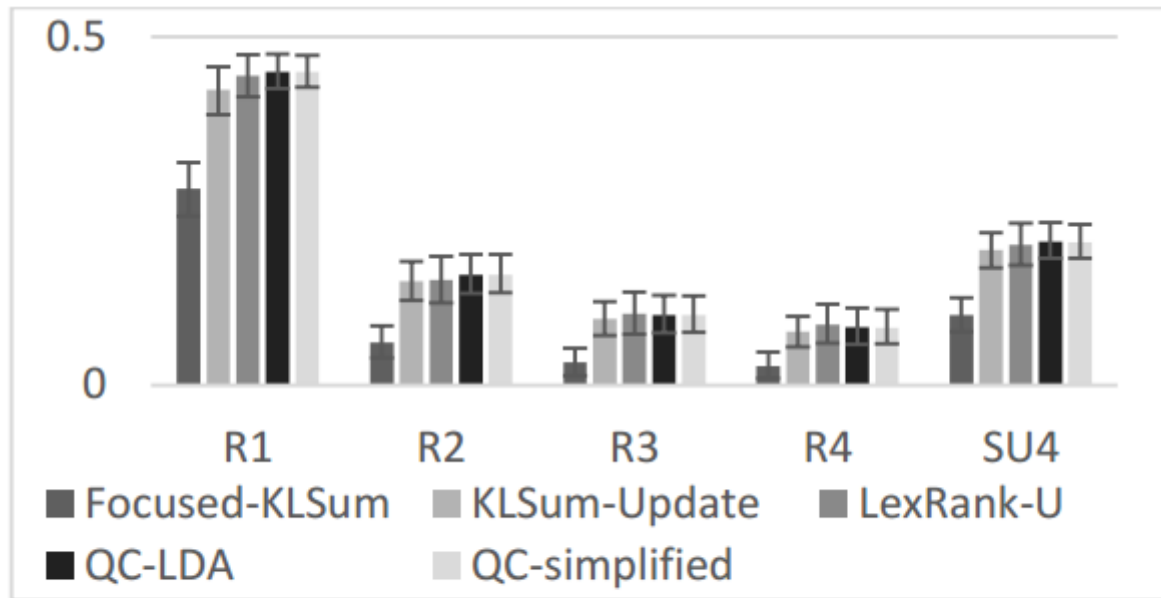


Figure 5: ROUGE Recall Scores (with stemming and stop-words)

Conclusions

- Presented new summarization task: QCFS
- Construct a novel dataset for QCFS containing human summaries
- Four methods were evaluated for the task. The baseline methods based on KLSum show a significant improvement when penalizing redundancy with the previous summarization.

Future Work

- This paper only concentrated on *zoom in* query chains
 - *Zoom out* or *switch topic* were left to future work
- Attempt to derive a task-specific evaluation metric that exploits the structure of the chains better assess relevance, redundancy and contrast