ABSTRACT

Title of Thesis: UNSUPERVISED THEMATIC PHRASE EXTRACTION FROM SINGLE TEXT ARTIFACTS

Akshay Peshave, Ph.D. - Computer Science, 2021

Thesis directed by: Dr. Tim Oates, Professor
Department of Computer Science and
Electrical Engineering

Automated knowledge discovery is central to augmenting knowledge acquisition and elicitation by humans from vast amounts of content. Precise and concise representations, both structured and semi-structured, of knowledge contained in textual content have the potential to boost human productivity. Further, they can reduce, if not eliminate, human error and bias in knowledge retrieval and curation by humans from vast collections of content to use for their subsequent knowledge-based tasks.

Conventionally, knowledge discovery in text (KDT) approaches and paradigms have been designed to build domain knowledge by processing large collections of text documents and applying them to process individual text documents using this acquired domain knowledge for guidance. Consequently, these approaches are blind to the finer topical features of the individual document because these features are abstracted by topic models that infer topicality in the context of the whole corpus.

We need an unsupervised method to extract topical or thematic phrases from a single text document without the need to access entire collections of texts or background domain or language dictionaries and thesauri. Further, the method should not abstract fine-grained thematic phrases contained in the document, thus, enabling its application for hierarchical knowledge representation and downstream document level text analytics tasks.

This work describes ThemaPhrase (ThP), a novel framework for unsupervised extraction of thematic phrases from single text artifacts. The framework operates without the need for corpus wide statistics and external domain knowledge which makes it domain agnostic. ThP configurations are more robust than competing methods to topic-to-partitions ratio and varying average token occurrence frequencies in a document. Different configurations of ThemaPhrase are identified that outperform competing methods in extracting thematic phrases that represent the topicality of a document at varied granularities.

Further, this work shows that sentence pre-filtering based on thematic phrases and thematic words helps improve extractive summarization for texts, such as patents, that have relatively higher occurrence frequencies of tokens where the baseline TextRank summarizer underperforms. ThemaPhrase configurations that outperform competing thematic phrase extraction methods in extractive summarization using sentence pre-filtering are discussed.

UNSUPERVISED THEMATIC PHRASE EXTRACTION

FROM SINGLE TEXT ARTIFACTS

by

AKSHAY PESHAVE

PhD Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computer Science
2021

Mom and Dad, This One's for You...

Look, I Made Computer Scientist!

ACKNOWLEDGMENTS

I would like to express immense gratitude to my adviser and mentor, Dr. Tim Oates, for being patient and supportive as I indulged in a multitude of tangential explorations throughout my PhD research. This body of work wouldn't have been possible without his recurrent guidance and advice.

I would like to thank my doctoral dissertation committee for their constructive feedback on my dissertation as well as future extensions of this work. I also cannot be grateful enough to my peers in the Cognition, Robotics and Learning (CoRaL) lab at UMBC who have been and continue to be sounding boards for ideas and results of my research.

Lastly, but most importantly, I'd like to thank my parents, extended family and close friends for their patience and tough love that has been a driving force and motivator for me throughout my PhD journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Automated knowledge discovery is defined as the extraction of implicit knowledge that is useful to a particular task or collection of tasks from raw data [1]. It is useful for augmenting knowledge acquisition and elicitation by humans from vast amounts of content in several industries and domains. These domains include academic research, law, public policy, market research and journalism. Text retrieval, curation, collation and summarization are some of the tasks that operate on vast amounts of textual content and are sought to be automated effectively and efficiently using automated knowledge discovery in text (KDT).

Quality knowledge discovered from content should be grounded, comprehensible and actionable by humans [2]. Knowledge representation [3] encompasses ways of organizing and presenting information contained in raw data for efficient assimilation by computer algorithms and humans. Efficiency of assimilation is, thus, a key consideration for any use case and actors involved when formulating representations. Often representations need to be specifically tailored for different domains and use cases, since a single, generalized

knowledge representation may be impractical for effective utility across all use cases. Knowledge representations should serve as accurate, granular substitutes for the verbose, raw information they represent, thus enabling efficient assimilation and further inference by humans as well as other representation systems [3].

Specifically, knowledge contained in textual data is made explicit in various ways through: (a) human understandable representations such as concepts, topics and taxonomies, and (b) machine readable representations such as frequency distributions and co-occurrences of words, phrases and topics. Such representational constructs transform raw, unstructured text collections into semi-structured or structured representations that make the key knowledge contained in the text explicit in a concise and precise form.

A more complex form of knowledge representation, that may be derived from the aforementioned forms, is a hierarchy of knowledge contained in the content. Several knowledge representations that are hierarchical are also inter-operable and integrable, such as, Semantic Networks [4, 5], Ontologies [6] and Concept Maps [7, 8, 9]. Mind Maps [10] are another form of knowledge representation that are relatively more directly assimilable and actionable by humans. The former representations, on the other hand, are befitting to be utilized by algorithms to automate more complex downstream text analyses and knowledge mining tasks such as semantic, sentiment, discourse and temporal analyses of text collections.

All the above representations are graphical in nature and are implicitly hierarchical by design. They may be used as meta-representations too that are particularly relevant in the context of text mining. Graphical forms of knowledge representation are also powerful

because a plethora of efficient graph-theoretic algorithms exist to further mine granular knowledge as well as derive complex inferences. Additionally, hierarchical organization implicitly allows abstraction at any level of the hierarchy. Further, all these characteristics of graphical representations facilitate knowledge visualization for humans that is easily understandable and actionable.

It is intuitive from the above discussion that structured and semi-structured knowledge representations that are concise and precise have the potential to boost human productivity. Further, they can reduce, if not eliminate, human error and bias in knowledge retrieval and curation when dealing with vast collections of textual content. This is achieved by reducing the quantity of raw text that humans need to consider to fulfill the knowledge needs for their tasks by helping them focus on sufficiently concise and precise information. Topic models and text summaries are two such representations of knowledge contained in text collections. They are representations of different granularities and structures, but both aim to capture the core themes addressed in the text.

Topic models [11, 12, 13] are a collection of algorithms to identify topics that can be attributed to documents, either individually or as a group, within a large text collection. Topics can take the form of a set of words and/or phrases that are closely associated with the same topic. These algorithms perform statistical computations to calculate the mutual associativity between words and phrases in the text collection as a measure of their ability to collectively represent one or more topics. The final set of topics is produced based on their collective ability to discriminate between documents belonging to different topics.

These topics can then be associated with individual documents to perform document tagging or clustering for human analyses or for use by information retrieval systems. The topical representations generated by topic models find application in improving downstream text analytics tasks such as automated text summarization, text segmentation, document labeling and discourse analyses.

Text summaries are good examples of how knowledge contained in a text document can be represented in a concise, semi-structured form to help improve human efficiency [14, 15]. Text summarization is the process of extracting key information from a collection of texts and arranging this information in a comprehensible and concise form. In other words, the goal of summarization is to abstract the discourse associated with the most important topics in the content to a suitable granularity. Note-taking or note-making are also a form of summarization of raw content and often take on relatively more structured forms such as bullet or numbered lists, pictorial depictions and mind maps [10].

In addition to providing a concise representation of the core ideas discussed in the text, summaries help improve retention and recall of the elicited knowledge by humans. Conciseness and precision are important for effectiveness of notes [16]. Empirical evidence also shows that hierarchical organization of knowledge improves recall [17]. Further, recall has been shown to more than double when information is presented as a conceptual or associative hierarchy. Thus, good summaries can help humans understand, retain and recall the core ideas contained in a text document efficiently without the need to read entire documents.

Recent research on automated text summarization [14, 15] comprises of methods and approaches for machines to perform text summarization in a semi-supervised or unsupervised manner. One category of approaches selects sentences from the original text to form the summary. The other category identifies core themes or concepts contained in the text and synthesizes sentences to form the summary. Both categories of automated summarization approached depend on accurate theme or concept selection to generate quality summaries. Thus, the effectiveness of automated summarization methods is assessed based on their coverage of concepts that adequately represent the main themes addressed in the content [18, 19, 20, 21].

Conventionally, topic modeling methods have been designed to extract topics from large collections of text documents. KDT approaches that work on individual text documents may then use these corpus-based topics for guidance. Consequently, approaches that use this workflow are blind to the finer topical features of individual documents because these features are abstracted by topic models that infer topicality in the context of entire text collections or corpora.

Automated text summarization and text segmentation are examples of tasks that work on individual text artifacts to summarize and collate their content respectively. They can benefit from topical phrase extraction methods that operate on individual text documents and will not need to depend on domain knowledge built by processing entire text collections that the documents belong. This will ensure efficiency and effectiveness.

Some topical phrase extraction methods that use an external source of quality words

and phrases to guide themselves may be applied to individual text documents. AutoPhrase [22] is an example of one such method. It uses wiki-phrases in the standard version of its implementation to guide the topical phrase extraction process. Wiki-phrases may be replaced or augmented by domain specific lists of phrases too. This requirement of background phrase knowledge for such semi-supervised approaches is an added burden.

We need an unsupervised method to extract topical or thematic phrases from a single text document without the need to access entire collections of texts or background domain or language dictionaries and thesauri. The method should not abstract fine-grained thematic phrases contained in the document, thus, enabling its application for hierarchical knowledge representations. Such a method would find wide application across domains and industries. Further, it can be used to guide myriad automated text analytics tasks that are performed at the level of individual documents as well as at the level of large document collections.

Core Contributions

This work describes ThemaPhrase (ThP), a novel framework for unsupervised extraction of thematic phrases from single text artifacts. The framework operates without the need for corpus wide statistics and external domain knowledge which makes it domain agnostic. It is also more robust than competing methods to topic-to-partitions ratio and varying average word occurrence frequencies in a document. Different configurations of ThemaPhrase are identified that outperform competing methods in extracting thematic phrases at different thematic granularities.

Further, this work describes two datasets created for quantitatively evaluating thematic phrases and automated extractive summarization. The datasets have distinct average word occurrence frequency distributions in a document and have author generated titles and abstracts that can be used as reference gold standards for the both these tasks.

Further, this work shows that sentence pre-filtering based on quality thematic phrases helps improve extractive summarization for texts, such as patents, that have relatively high average word occurrence frequencies where the baseline TextRank summarizer underperforms.

Dissertation Outline

Chapter 2 describes the background and related works for topic modelling and text summarization. Chapter 3 describes the ThemaPhrase framework along with all its phrase heuristics. Chapter 4 discusses the characteristics of datasets and the software libraries/packages used for implementations and evaluations. Chapter 5 provides a thorough quantitative evaluation of the thematic phrases extracted by ThemaPhrase configurations and competing methods. Chapter 6 describes sentence pre-filtering for extractive text summarization using thematic phrases and discusses its effects on summarization quality. Lastly, Chapter 7 concludes the dissertation with key findings, potential extensions of this work and avenues of future research. Detailed tables of all quantitative metrics used for the quantitative evaluations in Chapters 5 and 6 are provided in Appendices B to K

Chapter 2

BACKGROUND

This work addresses the problem of identifying thematic or topical phrases from a single text document in an unsupervised manner. This research problem is closely related to the area of topic models. ThemaPhrase, the unsupervised thematic phrases extraction method described in this work, is compared with some established and popularly used topic modeling methods and their adaptations for extracting thematic phrases from a single text artifact. Further, the quality of automated text summaries generated using sentence pre-filtering based on the extracted thematic phrases is also discussed and evaluated. This chapter describes the state of research in the areas of topic models, topical phrase extraction and automated text summarization.

2.1 Topic Models and Topical Phrases Extraction

Effective representation of knowledge contained in text artifacts requires identification of the topics discussed in the text. These topics form the core themes of the text. Topics are a

conceptual representation of the knowledge contained in a text corpus. These representations take the mathematical form of a probability distribution over a group of tokens (words or phrases). The probabilities quantify the degree to which each token represents a particular topic or theme contained in the corpus, where each topic is collectively represented by its corresponding set of tokens and their in-topic probability distribution. Topic models are algorithmic constructs to infer such topic representations.

The Latent Dirichlet Allocation (LDA) topic model [23] is a probabilistic, generative topic model that models a text collection as a mixture of topics. Each document is modeled as a probability distribution over topics. And, each topic is modeled as a probability distribution over unigram (single word) tokens. This model may be applied to any collections of discrete tokens. The generative story of LDA is to draw a topic from the topic distribution followed by a word from the word distribution of the drawn topic to fill in every word position in the document being generated.

A substantial body of research has extended the LDA unigram model to model topics as a distribution over n-gram (multi-word) tokens i.e. phrases. The Bi-gram Topic Model (BTM) [24] estimates the probability of a word to be generated for a document conditioned on the topics drawn from the topic distribution for that word position as well as the immediately previous word position in the document. The LDA Collocation (LDACOL) model [25] estimates the probability of a word to be generated for a document conditioned on the topic drawn from the topic distribution for that word position and a collocation decision variable that is conditioned on the immediately previous word generated for the

document as well as the current word.

HMM-LDA [26] is a composite model composed of a Hidden Markov Model (HMM) and an LDA topic model. It separates words into syntactic and semantic roles. The model estimates a Markov chain of classes of words as well as topic distributions for the document. The topics and syntactic classes are bags of words and are each represented by a probability distribution over words. The set of classes consists of one semantic class and multiple syntactic classes. The generative story for this model is to draw a class from the class transition distribution. If the class drawn is the semantic class, a word is drawn by choosing a topic from the topic distribution and then by drawing a word from that topic's word distribution. If the class drawn is one of the syntactic classes, a word is drawn from the word distribution of that syntactic class directly. The core motivation of this model is to generate n-grams with a view of generating syntactically correct sentences for a document.

The Hidden Topic Markov Model (HTMM) [27] assumes that all LDA topics, that are bags-of-words, form a Markov chain. This is contrary to the assumption of topic independence in the standard LDA topic model. The motivation of this method stems from the intuition that words in a single sentence may span more than one topic or may belong to the same topic depending on context. This model also focuses on sentence construction, like HMM-LDA, when generating n-grams. The generative story for this model is to first draw a transition decision variable that indicates whether the current word to be drawn should be drawn from a new topic or from the topic of the immediately previous generated word.

The Topical N-grams (TNG) model [28] is a generalization of the BTM [24] and

LDACOL [25] models. In the TNG model, the collocation random variable is conditioned on the topic of the immediately previous word in addition to being conditioned on the previous word, as is the case in LDACOL. As a result, the n-grams generated by the TNG model are based on the context of all the previous word positions (i.e. topics) in the n-gram as well as the previously generated words themselves. This model can approximate n-grams of order n >2 through n-gram concatenation. The resultant topics are bags of n-grams instead of bags of words, as is the case with previously discussed models.

The Pitman-Yor Topic Model (PYTM) [29] is a generalization of the LDA model and Chinese Restaurant Process (CRP). It uses the Pitman-Yor process to guide word generation in the generative story so that word distributions follow Zipf's power law that holds for word distributions in a document corpus as well as individual documents in the corpus. Similar to LDA, the CRP of PYTM assigns topics to each table in the restaurant using the multinomial topic distribution generated using a Dirichlet prior. But in the PYTM, the word for each word position in the document is generated by first determining the topic (table) for the word using a distribution from the Pitman-Yor process and then drawing a word from that topic's word distribution.

PYTM uses the distribution of words over the entire corpus or document collection as the base distribution for the Pitman-Yor process per document. The contrast between document and corpus level word distributions enables the PYTM to model topics based on the document's context. PYTM cannot be applied to a single document, by partitioning the document and providing it as a text corpus input, since the contrast between the document

and corpus level word distributions in this case may be minimal for many partitions that are coherent in topicality that is largely similar to the topicality of the whole document.

The Hierarchical Pitman-Yor Topic Model (HPYTM) [29] assumes a power-law word distribution at both the document and topic level. Several variations of a novel Gibbs sampler have been used with the HPYTM to generate better Bayesian topic language models [30]. Both PYTM and HPYTM aim to reduce the perplexity of their resultant topic language generative models in comparison to LDA. Further, the topics are bags of phrases that don't differentiate between word types or their part-of-speech. As a result, phrases such as "according to" and "in high dimensional" are high probability phrases in certain topics depending on the domain and nature of the text corpus (refer qualitative examples in [30]).

More recent advances in topic modeling are applications of a variety of neural network architectures for supervised and unsupervised topic extraction. The neural topic model (NTM) [31] applies deep learning to generate topics. NTM relies on aggregation of word embeddings pre-trained using large corpora for generation of n-gram topics. The Neural Variational Document Model (NVDM) [32] generates topics as bag-of-words representations using variational inference over a document corpus. Several neural topic models that use variational inference and variational auto-encoders have been proposed that apply different prior distributions and their reparameterization functions [33].

Sequential models such as TopicRNN [34] and TAN-NTM [35] have been proposed that use attention mechanisms to learn bag-of-words topics and employ them for language generation. TAN-NTM is shown to be of utility for keyphrase generation in a supervised

setting in [35].

Adversarial network architectures have also been proposed for topic modeling. The Adversarial-neural Topic Model (ATM) [36] adapts generative adversarial network (GAN) architecture to generate bag-of-words topics with high topic coherence for large corpora. The Bidirectional Adversarial Topic (BAT) model [37] achieved higher topic coherence than ATM. BAT employs bidirectional adversarial training allowing its encoder to infer a topic distribution for an input document unlike ATM.

Neural topic models learn topics as bag-of-words. N-gram topics either require external knowledge, as is the case for NTM, or a downstream supervised learning step for n-gram extraction, as is the case with TAN-NTM. Further, neural topic models need to be trained over a sufficiently large corpus to learn topics that allow them to classify or cluster documents effectively.

The TopMine framework [38] models topics that are bags of phrases with the objective of keeping the complexity of the topic model low. The framework consists of frequent phrase mining and document segmentation to extract quality candidate phrases in the document. The topics are then inferred using a variation of the LDA model wherein words of a phrase are assumed to belong to the same topic and are generated together for each phrase position in the generative story. Therefore, all phrases formed by words belonging to two or more topics are treated as less significant and are excluded from the topics. In the context of applying this method to extract topical phrases from a single document, this assumption may lead to quality topical phrases being rejected. This may happen because words in a

topical phrase may belong to different topics when the framework is run on partitions of a single document that have one or more coherent themes.

AutoPhrase [22] is a framework for automated phrase mining from text corpora. The framework extracts quality phrases by using external knowledge bases, such as Wikipedia, to guide its phrase extraction. It also depends on part-of-speech taggers for the language of the text corpus to extract quality topical phrases. The framework trains an ensemble classifier, guided by the general knowledge base, to classify phrases as quality phrases or otherwise. The phrases that are classified as quality phrases are then processed through a part-of-speech tagger and undergo phrase segmentation to extract the final set of topical phrases.

The dependence of AutoPhrase on external knowledge bases may not serve well for domains that are not sufficiently addressed by the contents of these knowledge bases. The method is also limited in performance by the size of the external knowledge base in addition to the size of the text corpus being processed. A truly unsupervised method should function without the need of such knowledge bases. This is increasingly important for extracting thematic phrases from a single document when the entire corpus that the document belongs to may be unavailable for modeling corpus level topics.

The CQMine framework [39] mines quality, topical phrases from text corpora. The framework utilizes the heuristics of phrase frequency, phraseness, completeness and appropriateness for phrase construction and segmentation. Further, the CPhrLDA topic model utilizes these segmented topical phrases to model topics as bags of phrases. The CPhrLDA topic is more flexible in that it does away with the TopMine's assumption that the topics

for words of a phrase are identical. The CPhrLDA model estimates a distribution over all topic-word pairs for all words that form topical phrases.

In the context of extracting topical or thematic phrases from a single document, the CQMine framework may pose some challenges. Firstly, the formulation of its completeness and appropriateness criteria serves well for large text corpora with sufficient heterogeneity in topicality. For a single text document partitioned into chunks, the heterogeneity may be insufficient as has been explained earlier in this section. Secondly, the completeness criterion seeks to merge two short phrases into a single larger phrase based on satisfaction of a statistical threshold of co-occurrence of the two phrases. The completeness and appropriateness criteria seek to ensure selected phrases are not proper sub-sequences and are disjoint from other selected phrases respectively.

For example, the latter two criteria will end up rejecting one or more of the phrases “data visualization”, “multivariate data” and “multivariate data visualization” that occur in a survey paper on data visualization techniques. The phrases will satisfy the frequency criterion and the phraseness criterion but will not pass the latter two criteria even though they convey separate, finer grained semantics relevant to the thematic basis of the text under consideration. Such fine-grained phrases are termed ”locally frequent phrases” in the CQMine framework. The framework clusters documents into different domains and each cluster is used as the new input for another iteration of searching for domain-specific phrases. This iterative workflow needs additional parameter optimization which is unwieldy for extracting topical phrases from single text documents at scale.

Fig. 2.1 visually summarizes the broader approaches to model and represent topicality of texts using unigram and n-gram representations. It shows example topic modeling methods from groupings of topic models based on their broader approach for topic inference and representations and reflects the discussion above.

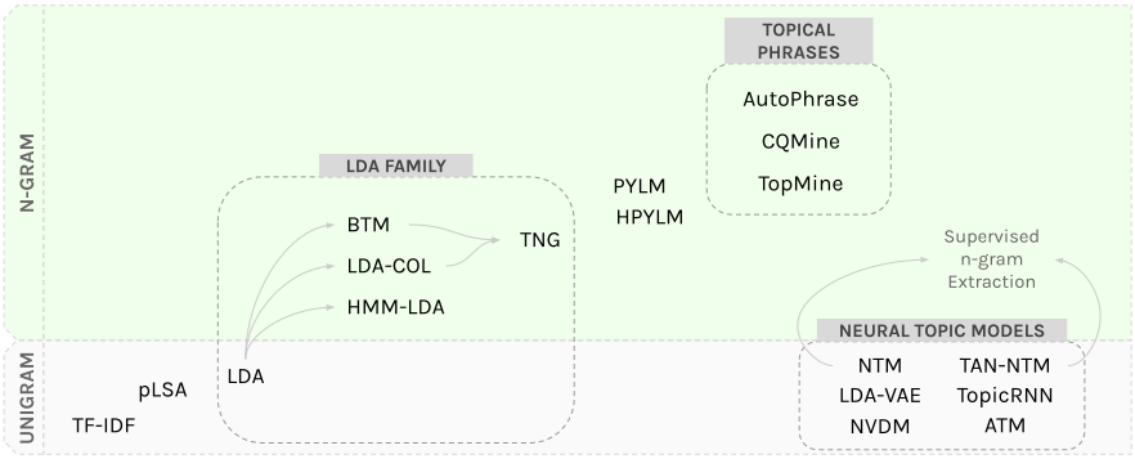


FIG. 2.1: Visual Map of Broad Approaches to Topic Modeling with Unigram and N-Gram Topic Representations

The ThemaPhrase framework, described in this work, is a novel, unsupervised framework that mines thematic phrases from a single text document that represent the thematic basis or topicality of the content contained in the document. The framework uses the LDA unigram topic model over partitions of the document and phrase filtering using structural and semantic heuristics to obtain a ranked list of thematic phrases. The framework is able to consider topicality as well as structural and semantic features of thematic phrases without the need for an external knowledge base or access to entire corpora of texts in the document's domain. Further, the framework doesn't need topic count optimization or multiple

iterations of thematic phrase set refinement.

2.2 Text Summarization

Text summarization is the process of extracting important information from text and arranging it in a comprehensible and concise form, typically as sentences. The goal of summarization is to abstract the discourse that is associated with the most important topics contained in the text to a required level of granularity. Effective text summaries are concise as well as precise. The precision requirement, in addition to conciseness, further reinforces the need for summarizers to know the core themes or topicalities of the text being summarized as discussed in Chapter 1.

Automated text summarization refers to an algorithmic approach to paraphrase textual content using computers. Automatic summarization [14, 15] can be broadly classified into two categories based on their approach to extract summaries:

1. Extractive Summarization: Selects sentences it deems relevant to the summary from the input text and collates them to form the summary.

2. Abstractive Summarization: Extracts representative themes or core topics from the input text and synthesizes sentences using the representative themes and corresponding discourse from the input text to form the summary.

Extractive summarization utilizes the sentences from the input text in their original form in its summaries. Whereas, abstractive summarization involves a language generation component,

either to stitch together segments of sentences from the original text or synthesize new sentences using characteristics of the discourse from the original text. The latter approach to automatic summarization is more complex due to the language generation requirement. Nevertheless, accurate topical representation of content is critical in both cases. The discussion in this section will focus on extractive summarization as it used for experiments and evaluation in this work.

Extractive summarization approaches optimize relevance metrics or heuristics over sentences in the original text in order to choose the best sentences for the text summary. Early works on automatic extractive summarization scored the importance of sentences in the input text based on the lexical frequency of words [40, 41], Term Frequency-Inverse Document Frequency (TF-IDF) [42, 43] and log-likelihood measures [44, 45]. Subsequent advances in extractive summarization applied similarity metrics at different granularities (corpus, document, segment, sentence and phrase) within the input text to rank and choose sentences for generating optimal summaries. These approaches utilized lexical chaining and language processing [46, 47, 48], Singular Value Decomposition (SVD) and Latent Semantic Analysis (LSA) [49, 50, 51], and discourse analysis [52, 53] to extract summaries.

Graph-based extractive summarization methods followed that allow flexibility of sentence selection using metrics and heuristics at word, phrase and sentence granularities. These methods utilize a combination of word frequency and sentence similarity features to extract summaries which makes them useful and effective for single document as well as multi-document summarization. Availability of a variety of graph-based heuristics and

the possibility of incorporating any combination of quantitative and semantic measures as edge weights makes these approaches versatile. LexRank [43] is a graph-based extractive summarizer that uses graph centrality with cosine similarity of sentence-level TF-IDF vectors as edge weights for sentence extraction. TextRank [54] uses graph centrality with word overlap count for sentence pairs as edge weights for sentence extraction.

Accurate topic representation of content at a suitable granularity improves extractive summarization [15]. In this research work, the effectiveness of topical or thematic phrase extraction methods are evaluated by using them to augment the TextRank automatic summarizer. For the augmentation, sentences from the original text are filtered to only retain sentences which contain any of the extracted thematic phrases in whole or in part. The filtered sentences are then provided as input to the TextRank summarizer for summary generation. The effect of augmentation using ThemaPhrase and other competing methods on summarization quality is evaluated using the ROUGE [55, 56, 57] metrics.

Chapter 3

THEMAPHRASE : AN UNSUPERVISED, THEMATIC PHRASE EXTRACTION FRAMEWORK

ThemaPhrase (ThP) is a novel, unsupervised framework to extract thematic phrases from a single text document. The objective is to extract phrases that are representative of the key themes that are addressed in the body of the text. In this chapter, Sec. 3.1 describes notations used in this work along with definitions of terminology and other functions used to describe the framework. This is followed by the problem statement definition for the task that the framework solves in Sec. 3.2.

A detailed description of the ThemaPhrase framework along with its components is provided in the subsequent sections. Sec. 3.3 discusses how the LDA topic model is used along with nounphrase detection to provide an initial set of candidate thematic phrases. Then we describe and discuss the three phrase level heuristics that are used to filter candidate thematic phrases by the ThemaPhrase framework, namely, word sequence heuristic (WSEQ), word position heuristic (WPOS) and word association heuristic (WASS) in Sec. 3.4, 3.5

and 3.6 respectively. Lastly, Sec. 3.7 briefly describes how the ThP framework can be configured to use any combination of its heuristics.

3.1 Definitions & Notations

(a) **Set Notation:** Set names are denoted by bold text. Elements of a set are denoted by regular text using the set name along with a subscript that denotes the index of the element. For example, $\boldsymbol{S}$ is the set and its elements are denoted by S_i.

(b) **Set Membership ($\in$ and $\notin$):** The expression "*LHS* $\in$ *RHS*" indicates a *"belongs to"* or *"is observed in"* relationship of *LHS* with *RHS*. Conversely, the expression "*LHS* $\notin$ *RHS*" indicates a *"does not belong to"* or *"is not observed in"* relationship. Both these operators are utilized in the following ways:

(i) *LHS* and *RHS* can be singleton logical entities such that *LHS* is a part or component of *RHS*. For example, $word \in phrase$ indicates that the word is contained in the phrase and $word \in \boldsymbol{D}$ indicates that the word is contained in the document $\boldsymbol{D}$.

(ii) *LHS* can be a single entity and *RHS* can be a homegeneous or heterogeneous set. For example, $word \in \boldsymbol{W}$ indicates that the word belongs to the set $\boldsymbol{W}$ and $word \in \boldsymbol{Phrases}$ indicates that the set of phrases (***Phrases***) contains $word$ as a unigram phrase.

(iii) *LHS* and *RHS* can be homogeneous or heterogeneous sets. When *LHS* is a set,

the $\in$ and $\notin$ operators indicate their corresponding relationships for all members of the *LHS* set with *RHS*. For example:

$$\{word_1, word_2\} \in \{words\}$$

indicates that both $word_1$ and $word_2$ belong to the set $words$, and,

$$\{word_1, phrase_1\} \in \boldsymbol{D}$$

indicates that both $word_1$ and $phrase_1$ are observed in the document $\boldsymbol{D}$.

(c) **Set Item Count:** The item count of an argument (A) is denoted by $|A|$. If A is a phrase then $|A|$ denotes the number of words in the phrase. If $\boldsymbol{A}$ is a set then $|\boldsymbol{A}|$ denotes the number of items in the set.

(d) **Sequence** : A non-empty sequence of homogeneous or heterogeneous tokens, as opposed to an unordered set of tokens, is indicated using the notation $\langle\ t_1, t_2, ..., t_m\ \rangle$. The generalized shorthand notation used for a sequence is $\langle\ t_i\ \rangle$.

(e) ***D*:** The input document from which thematic phrases need to be extracted.

(f) **Word Set ($\boldsymbol{W}$)** : A word set is denoted by $\boldsymbol{W}$. Further, a word set can be denoted by $\boldsymbol{W^X}$ and means $\boldsymbol{W^X} = \{\ w_i : w_i \in \boldsymbol{X}\ \}$. A shorthand way of writing this word set definition that is used in this work is $\boldsymbol{W^X} = \{\ w_i^X\ \}$ where $w_i^X \equiv w_i \in \boldsymbol{X}$. Therefore, $\boldsymbol{W^D}$ denotes the set of words that occur in document $\boldsymbol{D}$.

(g) **Phrase** : A phrase, ph_i^D, is a sequence of words and is defined as:

$$ph_i^D = \langle\ w_j : w_j \in \boldsymbol{W^D}\ \rangle \text{ where, } ph_i^D \in \boldsymbol{D}$$

For example, the phrase *"multivariate data visualization"* can be written using the ordered sequence notation as $\langle$ *"multivariate", "data", "visualization"* $\rangle$

(h) **Nounphrase**: is defined as *"a word or group of words that functions in a sentence as subject, object, or prepositional object"* [1]. It is a phrase (refer (g)) with additional natural language constraints based on part-of-speech tags. For example, consider the sentence *"Isosurface extraction is an important technique for visualizing large scale three-dimensional scalar fields."*. This sentence has the following nounphrases:

$$\boldsymbol{noun\text{-}phrases} = \left\{ \begin{array}{l} \langle\ \textit{"isosurface", "extraction"}\ \rangle, \\ \langle\ \textit{"an", "important", "technique"}\ \rangle, \\ \langle\ \textit{"large", "scale", "three-dimensional","scalar", "fields"}\ \rangle \end{array} \right\}$$

Throughout this work the terms *"nounphrase"*, *"noun chunk"* and *"phrase"* are used interchangeably. Thus, *"phrase"* should be interpreted to mean a nounphrase present in a document, i.e. it will not contain verbs, adverbs or prepositions.

(i) **Phrase Set (*ph*)** : A phrase set is denoted by $\boldsymbol{ph}$. Further, a phrase set denoted by $\boldsymbol{ph^X}$

[1] Definition of "nounphrase" taken from the Oxford English Dictionary

means $\boldsymbol{ph^X} = \{\, ph_i : ph_i \in \boldsymbol{X} \,\}$. A shorthand way of writing this phrase set definition that is used in this work is $\boldsymbol{ph^X} = \left\{\, ph_i^X \,\right\}$ where $ph_i^X \equiv ph_i \in \boldsymbol{X}$. Therefore, $\boldsymbol{ph^D}$ denotes the set of phrases that occur in document $\boldsymbol{D}$.

(j) **Permutations of phrase** : The set of permutations $\boldsymbol{\widetilde{ph}_i^D}$ of a phrase ph_i^D is defined as,

$$\boldsymbol{\widetilde{ph}_i^D} = \left\{\, \widetilde{ph}_{im}^D : 1 \leq m \leq \left|ph_i^D\right|! - 1 \,\right\}$$

where, all $\widetilde{ph}_{im}^D$ are permutations of ph_i^D and must satisfy the conditions that $\forall \widetilde{ph}_{im}^D \in \boldsymbol{\widetilde{ph}_i^D}$:

(i) $\left|\widetilde{ph}_{im}^D\right| = \left|ph_i^D\right|$: the word length of the permutation $\widetilde{ph}_{im}^D$ and the original phrase ph_i^D is the same

(ii) $\widetilde{ph}_{im}^D \in \boldsymbol{D}$: the permutation occurs in the document $\boldsymbol{D}$

(iii) $\widetilde{ph}_{im}^D = \left\langle\, w_k : w_k \in ph_i^D \,\right\rangle$: the permutation phrase is only formed by words contained in the original phrase

(iv) $\widetilde{ph}_{im}^D \neq ph_i^D$: the permutation phrase is not the same as the original phrase

For example, in the case of a text document that contains both the phrases - *"multiple visualization attributes"* and *"multiple attributes visualization"* - the following are true:

(i) For $ph_i^D = \langle$ *"multiple", "visualization", "attributes"* $\rangle$,

$$\widetilde{\boldsymbol{ph_i}}^{\boldsymbol{D}} = \left\{ \widetilde{ph}_{i1}^{D} = \langle \text{ "multiple", "attributes", "visualization" } \rangle \right\}$$

(ii) For $ph_j^D = \langle$ *"multiple", "attributes", "visualization"* $\rangle$,

$$\widetilde{\boldsymbol{ph_j}}^{\boldsymbol{D}} = \left\{ \widetilde{ph}_{j1}^{D} = \langle \text{ "multiple", "visualization", "attributes" } \rangle \right\}$$

Other permutations, such as, *"visualization multiple attributes"* and *"attributes visualization multiple"* are not part of the permutations set because these phrases $\notin \boldsymbol{D}$, even though all the words $\in \boldsymbol{W^D}$

(k) **Sub-phrases of phrase** : The set of subphrases $\underline{\boldsymbol{ph}}_{\boldsymbol{i}}^{\boldsymbol{D}}$ of a phrase ph_i^D is defined as,

$$\underline{\boldsymbol{ph}}_{\boldsymbol{i}}^{\boldsymbol{D}} = \left\{ \underline{ph}_{in}^{D} : \underline{ph}_{in}^{D} \in \boldsymbol{D}, \underline{ph}_{in}^{D} | ph_i^D, \underline{ph}_{in}^{D} \neq ph_i^D \right\}$$

where the operator | indicates that the LHS is a proper subsequence of the RHS. For example, in the case of a text document that contains the phrase *"multivariate data*

visualization", for $ph_i^D = \langle$ *"multivariate", "data", "visualization"* $\rangle$:

$$\underline{\boldsymbol{ph}}_{\boldsymbol{i}}^{\boldsymbol{D}} = \left\{ \begin{array}{lll} \underline{ph}_{i1}^{D} & = & \langle \text{ "multivariate", "data" } \rangle, \\ \underline{ph}_{i2}^{D} & = & \langle \text{ "data", "visualization" } \rangle, \\ \underline{ph}_{i3}^{D} & = & \langle \text{ "multivariate" } \rangle, \\ \underline{ph}_{i4}^{D} & = & \langle \text{ "data" } \rangle, \\ \underline{ph}_{i5}^{D} & = & \langle \text{ "visualization" } \rangle \end{array} \right\}$$

(l) **Extensions of phrase** : The set of extensions $\widehat{\boldsymbol{ph}}_{\boldsymbol{i}}^{\boldsymbol{D}}$ of a phrase ph_i^D is defined as,

$$\widehat{\boldsymbol{ph}}_{\boldsymbol{i}}^{\boldsymbol{D}} = \left\{ \widehat{ph}_{ic}^{D} : \widehat{ph}_{ic}^{D} \in \boldsymbol{D}, \widehat{ph}_{ic}^{D} \neq ph_i^D \right\}$$

where, all $\widehat{ph}_{ic}^{D}$ are extensions of ph_i^D and must satisfy the following conditions $\forall \widehat{ph}_{ic}^{D} \in \widehat{\boldsymbol{ph}}_{\boldsymbol{i}}^{\boldsymbol{D}}$:

(i) $\left|\widehat{ph}_{ic}^{D}\right| > \left|ph_i^D\right|$: the word length of the extension $\widehat{ph}_{ic}^{D}$ is longer than that of the original phrase ph_i^D

(ii) $\widehat{ph}_{ic}^{D} \in \boldsymbol{D}$: the extension occurs in the document $\boldsymbol{D}$

(iii) The extension can be in either direction of the original phrase or in both directions. That is, $\widehat{ph}_{ic}^{D} = \left\langle \underset{\longleftarrow}{\widehat{ph}_{ic}^{D}},\ ph_i^D,\ \underset{\longrightarrow}{\widehat{ph}_{ic}^{D}} \right\rangle$ where,

$$\underset{\longleftarrow}{\widehat{ph}_{ic}^{D}} = \left\langle\ w_p : w_p \in \{null\} \cup (\boldsymbol{W^D} - ph_i^D)\ \right\rangle$$

$$\overrightarrow{\widehat{ph_{ic}}}^{D} = \left\langle\ w_s : w_s \in \{null\} \cup (\boldsymbol{W^D} - ph_i^D - \overleftarrow{\widehat{ph_{ic}}}^{D})\ \right\rangle$$

The *"−"* operator is used in (v) and (vi) above on two heterogeneous operands, a set of words and a phrase. This should be interpreted as the set difference between the set of words observed in the document ($\boldsymbol{W^D}$) and the set of words that form the phrase operand.

As an example of phrase extensions, let's take the case of a text document that contains the phrases *"multivariate data visualization"*, *"data visualization"* and *"data visualization benefits"*. The following phrase extension sets are possible:

(i) For $ph_i^D = \langle$ *"data", "visualization"* $\rangle$

$$\boldsymbol{\widehat{ph_i}^D} = \left\{ \begin{array}{l} \widehat{ph}_{i1}^{D} = \langle \text{ "multivariate", "data", "visualization" } \rangle, \\ \widehat{ph}_{i2}^{D} = \langle \text{ "data", "visualization", "benefits" } \rangle \end{array} \right\}$$

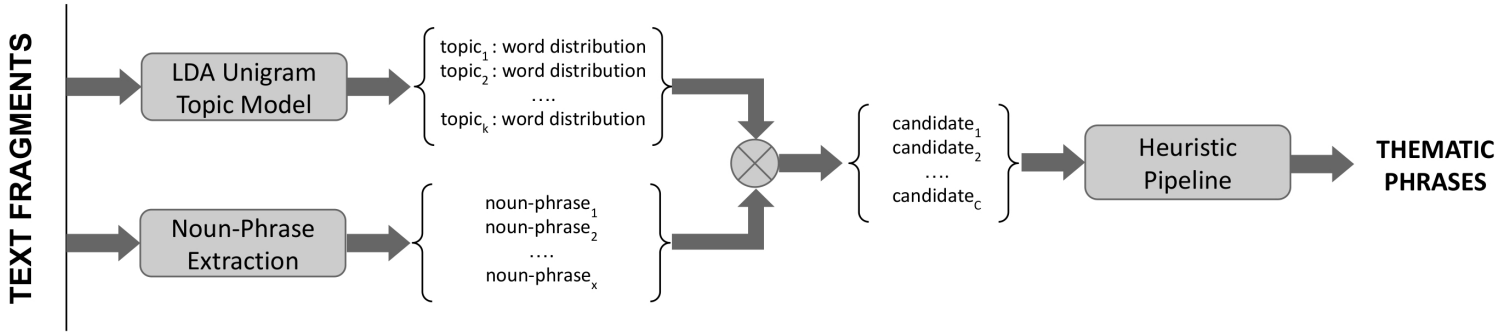
(ii) For $ph_i^D = \langle$ *"multivariate", "data", "visualization"* $\rangle$, $\boldsymbol{\widehat{ph_i}^D} = \varnothing$

(iii) For $ph_i^D = \langle$ *"data", "visualization", "benefits"* $\rangle$, $\boldsymbol{\widehat{ph_i}^D} = \varnothing$

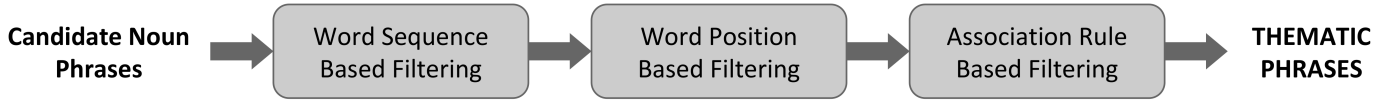
3.2 Problem Statement

The thematic basis or topicality of a text document can be represented by a set of phrases from the document, $\boldsymbol{ph^{tD}}$ such that

$$\boldsymbol{ph^{tD}} = \{\ ph_i \in \boldsymbol{D}\ \} \text{ where, } \boldsymbol{ph^{tD}} \approx \textit{Document Theme} \tag{3.1}$$



(a) Thematic Phrases Extraction Pipeline



(b) Thematic Phrase Filtering Heuristics in the "Heuristic Filtering" Stage of the Extraction Pipeline

FIG. 3.1: ThemaPhrase Framework

The aim is, thus, to find the thematic basis or topicality of an input document represented by a set of thematic phrases extracted from the document. The formal problem statement is as follows. Given a document $\boldsymbol{D}$, extract a set of thematic phrases $\boldsymbol{ph^{tD}}$ that is representative of the document's thematic basis or topicality. The set $\boldsymbol{ph^{tD}}$ is defined as

$$\boldsymbol{ph^{tD}} = \mathcal{M}(\boldsymbol{D}, \lambda, k) = \left\{ ph_x^{tD} : ph_x^{tD} \in \boldsymbol{D}, \left| ph_x^{tD} \right| \leq \lambda \right\}_{x=1}^{k} \tag{3.2}$$

where, k is the number of thematic phrases to be extracted and λ is the maximum allowable word length for each thematic phrase ph_x^{tD}. $\mathcal{M}$ is the thematic extraction method that takes a document $\boldsymbol{D}$ as input along with the parameters λ and k. The method $\mathcal{M}$ deems some phrases more thematic than others, that is:

$$\forall ph_x^D \in \boldsymbol{ph^{tD}}\ ,\ Pr\left(ph_x^D \in \boldsymbol{ph^{tD}} \mid \boldsymbol{D} \right) \geq \max_{ph_y^D \notin \boldsymbol{ph^{tD}}} \left(Pr\left(ph_y^D \in \boldsymbol{ph^{tD}} \mid \boldsymbol{D} \right) \right) \tag{3.3}$$

For readability, we will use "$tdPr\left(ph_x^D \mid \boldsymbol{D} \right)$" to denote $Pr\left(ph_x^D \in \boldsymbol{ph^{tD}} \mid \boldsymbol{D} \right)$ going forward.

The "*Document Theme*" is an abstract notion that is usually represented by human generated meta-elements of text documents such as their titles and abstracts. Since the thematic basis $\boldsymbol{ph^{tD}}$ should be representative of or approximate the document's theme, $\boldsymbol{ph^{tD}} \cap \boldsymbol{ph^{title}}$ and/or $\boldsymbol{ph^{tD}} \cap \boldsymbol{ph^{abstract}}$ can be used to evaluate $\boldsymbol{ph^{tD}}$. This evaluation can also be conducted at the granularity of words i.e. $\boldsymbol{W^{tD}} \cap \boldsymbol{W^{title}}$ and/or $\boldsymbol{W^{tD}} \cap \boldsymbol{W^{abstract}}$

Phrases are composed of words. The semantic meaning that phrases bear in the context of the overarching theme or topicality of a document is not just a consequence of the words

that make up the phrase but the order in which the words are placed to form the phrase. The probability of a phrase being representative of the theme of a document is:

(a) proportional to the probability of its sub-phrases occurring in ***D*** being representative of the theme

(b) inversely proportional to the probability of its permutations occurring in ***D*** being representative of the theme

(c) inversely proportional to the probability of its extensions occurring in ***D*** being representative of the theme

That is, given a function $\mathcal{F} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ that aggregates $tdPr$ over all members of a phrase set (specified as its superscript), the following proportionalities hold for $tdPr\left(ph_x^D \mid \boldsymbol{D} \right)$:

$$tdPr\left(ph_x^D \mid \boldsymbol{D} \right) \propto \overset{\underline{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{D}}}{\mathcal{F}} \left(tdPr\left(\underline{ph}_{xi}^D \mid \boldsymbol{D} \right) \right) \tag{3.4a}$$

$$tdPr\left(ph_x^D \mid \boldsymbol{D} \right) \propto \left[\overset{\widetilde{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{D}}}{\mathcal{F}} \left(tdPr\left(\widetilde{ph}_{xi}^D \mid \boldsymbol{D} \right) \right) \right]^{-1} \tag{3.4b}$$

$$tdPr\left(ph_x^D \mid \boldsymbol{D} \right) \propto \left[\overset{\widehat{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{D}}}{\mathcal{F}} \left(tdPr\left(\widehat{ph}_{xi}^D \mid \boldsymbol{D} \right) \right) \right]^{-1} \tag{3.4c}$$

Given the proportionalities in Eq. (3.4), the condition in Eq. (3.3) can be separated into the

following three inequalities for a thematic phrase ph_x^{tD} :

$$\forall ph_y^D \in \widetilde{ph}_x^{tD} \ , \ tdPr\left(ph_x^{tD} \mid \boldsymbol{D} \right) \geq tdPr\left(ph_y^D \mid \boldsymbol{D} \right) \tag{3.5a}$$

$$\forall ph_y^D \in \underline{ph}_x^{tD} - \left\{ ph_{1 \leq k < x}^{tD} \right\} \ , \ tdPr\left(ph_x^{tD} \mid \boldsymbol{D} \right) \geq tdPr\left(ph_y^D \mid \boldsymbol{D} \right) \tag{3.5b}$$

$$\forall ph_y^D \in \widehat{ph}_x^{tD} - \left\{ ph_{1 \leq k < x}^{tD} \right\} \ , \ tdPr\left(ph_x^{tD} \mid \boldsymbol{D} \right) \geq tdPr\left(ph_y^D \mid \boldsymbol{D} \right) \tag{3.5c}$$

The ThemaPhrase framework utilizes the LDA topic model along with nounphrase detection and three phrase-level heuristics to mine thematic phrases from a text document. Each phrase-level heuristic filters candidate thematic phrases in a manner that addresses one or more of the criteria in Eq. (3.5). The outline of the framework is shown in Fig. 3.1a. The following sections describe all the components of the framework in detail and discuss how each component contributes to the extraction of thematic phrases in alignment with the criteria discussed above for the problem statement.

3.3 LDA with Nounphrase Mapping

Theme identification for a document begins by splitting the document into parts or fragments of near uniform size in terms of sentence counts. These fragments serve as individual documents provided as input to the LDA topic model. The LDA topic model learns topics, each of which is a distribution over unigrams. We use the trained topic model to assign topic probabilities to each fragment. nounphrases are also extracted for each

fragment using part-of-speech tagging.

This stage of the framework utilizes $\boldsymbol{W^D}$ and $\boldsymbol{ph^D}$ as its inputs. Let, $j = \left|\boldsymbol{W^D}\right|$ is the size of the wordset and $\boldsymbol{fr^D} = \left\{ fr_p^{\boldsymbol{D}} \right\}$ be the set of p fragments or partitions of document $\boldsymbol{D}$. Let $\boldsymbol{\mathcal{T}^D} = \left\{ \mathcal{T}_z^{\boldsymbol{D}} \right\}$ be the set of z topics modeled by LDA for the document partitions. At this stage the framework has the following vectors available to it:

Topic Word Distribution ($\overrightarrow{\boldsymbol{TW}}$): This vector is available from the LDA topic model. It contains the word distributions for each topic and has the following form:

$$\overrightarrow{\boldsymbol{TW}} = \begin{pmatrix} \overrightarrow{tw_1} \\ \overrightarrow{tw_2} \\ \vdots \\ \overrightarrow{tw_z} \end{pmatrix} = \begin{pmatrix} Pr\left(w_1^{\boldsymbol{D}} \mid \mathcal{T}_1^{\boldsymbol{D}} \right) & Pr\left(w_2^{\boldsymbol{D}} \mid \mathcal{T}_1^{\boldsymbol{D}} \right) & \cdots & Pr\left(w_j^{\boldsymbol{D}} \mid \mathcal{T}_1^{\boldsymbol{D}} \right) \\ Pr\left(w_1^{\boldsymbol{D}} \mid \mathcal{T}_2^{\boldsymbol{D}} \right) & Pr\left(w_2^{\boldsymbol{D}} \mid \mathcal{T}_2^{\boldsymbol{D}} \right) & \cdots & Pr\left(w_j^{\boldsymbol{D}} \mid \mathcal{T}_2^{\boldsymbol{D}} \right) \\ \vdots & \vdots & \ddots & \vdots \\ Pr\left(w_1^{\boldsymbol{D}} \mid \mathcal{T}_z^{\boldsymbol{D}} \right) & Pr\left(w_2^{\boldsymbol{D}} \mid \mathcal{T}_z^{\boldsymbol{D}} \right) & \cdots & Pr\left(w_j^{\boldsymbol{D}} \mid \mathcal{T}_z^{\boldsymbol{D}} \right) \end{pmatrix}$$

Fragment Noun-phrases ($\boldsymbol{FNP}$): The set of nounphrases for each fragment can be extracted using any off-the-shelf part-of-speech tagging library. The set has the following form:

$$\boldsymbol{FNP} = \begin{pmatrix} \boldsymbol{fnp_1} = \left\{ ph_x^{\boldsymbol{D}} : ph_x^{\boldsymbol{D}} \in fr_1^{\boldsymbol{D}} \right\} \\ \boldsymbol{fnp_2} = \left\{ ph_x^{\boldsymbol{D}} : ph_x^{\boldsymbol{D}} \in fr_2^{\boldsymbol{D}} \right\} \\ \vdots \\ \boldsymbol{fnp_p} = \left\{ ph_x^{\boldsymbol{D}} : ph_x^{\boldsymbol{D}} \in fr_p^{\boldsymbol{D}} \right\} \end{pmatrix}$$

Candidate thematic phrases are selected from $\bigcup \overrightarrow{\boldsymbol{FNP}}$ using $\overrightarrow{\boldsymbol{TW}}$ as follows:

Step 1 - Select Top-*m* Topic Words : A set of top m words per topic are chosen from $\overrightarrow{\boldsymbol{TW}}$ to use for nounphrase filtration as follows:

$$\boldsymbol{mTW_k} = \underset{\boldsymbol{W'} \subseteq \boldsymbol{W^D}, |\boldsymbol{W'}|=m}{arg\,max} \sum^{w_i \in \boldsymbol{W'}} Pr\left(w_i \mid \mathcal{T}_k^{\boldsymbol{D}} \right)$$

$$\boldsymbol{mTW} = \bigcup \{ \boldsymbol{mTW_k} \}$$

Step 2 - Filter FNP Based on Topic Words : The phrases from ***FNP*** are chosen as candidate thematic phrases if they contain one or more words $\in$ ***mTW***. Therefore, every phrase in the set of candidate thematic phrases $\boldsymbol{ph^C} = \left\{ ph_x^C \right\}$ will satisfy the following: (a) $ph_x^C \in \bigcup \boldsymbol{FNP}$ and (b) $ph_x^C \cap \boldsymbol{mTW} \neq \varnothing$

The set of phrases $\boldsymbol{ph^C}$ are the candidate thematic phrases extracted by this stage of the framework. The intuition is that the cumulative probability of words across the words distributions for all topics will be higher for words that occur in thematic phrases that describe the theme of the document as a whole. Also, latent semantic associations between words that determine topicality are also manifested in the bag-of-words distributions and should have a bearing on the candidate noun-phrases selection. This approach of filtering nounphrases using the LDA topic word distributions as criteria ensures quality phrases of any granularity (i.e n-grams, $n \geq 1$) pertinent to the document's themes are considered as the initial set of candidate phrases.

These candidate nounphrases are used by the ThemaPhrase framework as input for heuristics based phrase filtering. The phrase filtering pipeline is shown in Fig. 3.1b. Three heuristics are used to filter phrases: (i) Word Sequence Heuristic (ii) Word Position Heuristic and (iii) Word Association Rules. These heuristics are described in detail in the sections that follow.

3.4 Word Sequence Heuristic (WSEQ)

A phrase is a sequence of words. The word subsequences in the phrase manifest the semantics of the phrase. The commonality of subsequences across multiple phrases hints at the thematic basis collectively represented by the set of those phrases. The word sequence heuristic (*WSEQ*) uses this word sequence information to filter phrases that are representative of the thematic basis of a document. The *WSEQ* heuristic is composed of three computational steps that are described in the subsections below.

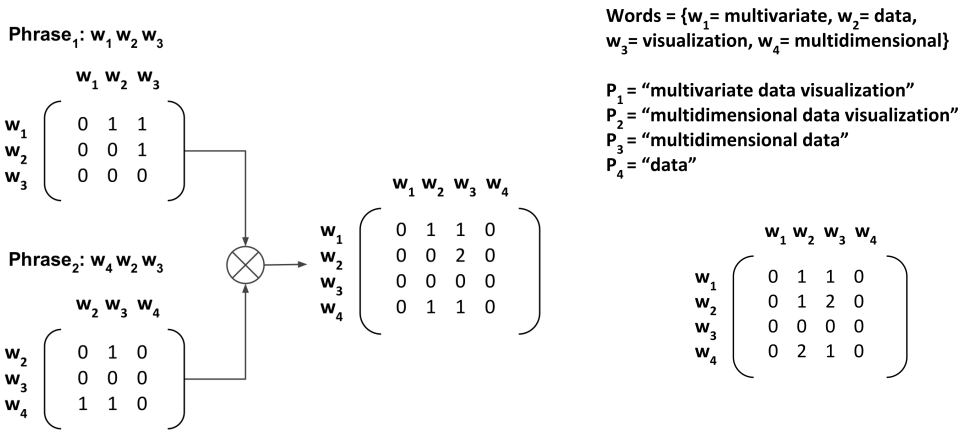
3.4.1 Word Sequence Vectors Generation

The candidate phrase set, $\boldsymbol{ph^C}$, obtained from the LDA nounphrase extraction stage, are used to construct vectors ($\overrightarrow{ws_i}$) for each word, $w_i \in \boldsymbol{W^C}$, to form the vector $\overrightarrow{\boldsymbol{WS}} =$

$\left(\overrightarrow{ws_1},\ \overrightarrow{ws_2},\ \ldots,\ \overrightarrow{ws_j} \right)$ where $j = \left| \boldsymbol{W^C} \right|$. The vector $\overrightarrow{\boldsymbol{WS}}$ takes the matrix form:

$$\overrightarrow{\boldsymbol{WS}} = \begin{pmatrix} \overrightarrow{ws_1} \\ \overrightarrow{ws_2} \\ \vdots \\ \overrightarrow{ws_j} \end{pmatrix} = \begin{pmatrix} fr\left(\langle\, w_1, \varnothing \,\rangle\right) & fr\left(\langle\, w_1 \,..\, w_2 \,\rangle\right) & \cdots & fr\left(\langle\, w_1 \,..\, w_j \,\rangle\right) \\ fr\left(\langle\, w_2 \,..\, w_1 \,\rangle\right) & fr\left(\langle\, w_2, \varnothing \,\rangle\right) & \cdots & fr\left(\langle\, w_2 \,..\, w_j \,\rangle\right) \\ \vdots & \vdots & \ddots & \vdots \\ fr\left(\langle\, w_j \,..\, w_1 \,\rangle\right) & fr\left(\langle\, w_j \,..\, w_2 \,\rangle\right) & \cdots & fr\left(\langle\, w_j, \varnothing \,\rangle\right) \end{pmatrix}$$

Every element $fr\left(\langle\, w_i \,..\, w_j \,\rangle\right)$ in $\overrightarrow{\boldsymbol{WS}}$ represents the frequency of the in-sequence skip-bigram $\langle\, w_i \,..\, w_j \,\rangle$ in the candidate phrases list. Elements of the form $fr\left(\langle\, w_i, \varnothing \,\rangle\right)$ represent the number of times w_i is observed as a single-word (or unigram) phrase in the candidate phrases list. Fig. 3.2a visually describes the process of word sequence vector construction and an example of $\overrightarrow{\boldsymbol{WS}}$ using four candidate phrases is provided in Fig. 3.2b.



(a) Computation (b) Example

FIG. 3.2: Word Sequence Matrix Generation for Phrases

3.4.2 Dimensionality Reduction

The vector space of $\overrightarrow{\boldsymbol{WS}}$ has high dimensionality with a theoretical ceiling value of $\left|\boldsymbol{W^D}\right|$. The curse of dimensionality [58] poses a challenge when working with such a high dimensional space. Dimensionality reduction of $\overrightarrow{\boldsymbol{WS}}$ is useful to alleviate this challenge. The vector $\overrightarrow{\boldsymbol{WS}}$ is projected into a δ-dimensional embedding space, where $\delta \ll j$.

The ThemaPhrase framework uses t-Stochastic Neighbor Embedding (tSNE) [59] for *WSEQ* dimensionality reduction. tSNE uses a parameterized stochastic neighbor embedding space to represent clusters of data points with high dimensionality in a low dimensional embedding space. Due to the high dimensionality and sparsity of $\overrightarrow{\boldsymbol{WS}}$, the distances between the $\overrightarrow{ws_i}$ vector pairs are less discriminative in the average case. Dimensionality reduction using tSNE helps abstract out minor variations and emphasizes relatively larger variations in skip-bigram frequencies between the word sequence vectors. This is critical to ensure effective detection of outlier words based on their *WSEQ* vectors in the next stage of the heuristic.

3.4.3 Outlier Words Detection

The final step of the *WSEQ* heuristic identifies words that are outliers based on their word sequence vectors. This is achieved using density-based clustering of data points in the δ-dimensional embedding space that represent words, and their relative distances, in the original word sequence vector space. Density based clustering uses density-defining parameters as necessary requirements for data points to form clusters. These parameters are

the minimum number of cluster neighbors and the threshold for pairwise distances between cluster neighbors. Outliers are those datapoints (which represent words) that do not satisfy both these requirements. The ThemaPhrase framework uses the DBSCAN [60] clustering algorithm for outlier detection in the *WSEQ* heuristic.

Words form candidate phrases are deemed outliers when their word sequence vector are not sufficiently similar to those of a sufficient number of other words. Outlier words are of two types:

(a) Words that form skip-bigrams with a subset of words $\in \boldsymbol{W^C}$ with unexpectedly lower frequency than all other words $\in \boldsymbol{W^C}$ that form skip-bigrams with the same subset.

(b) Words that form skip-bigrams with a subset of words $\in \boldsymbol{W^C}$ with unexpectedly higher frequency than all other words $\in \boldsymbol{W^C}$ that form skip-bigrams with the same subset.

The presence of the first type of outliers in phrases is indicative of the phrases being either: (i) excessively fine-grained and related to nuances of the core theme of the document, (ii) references to other loosely related concepts that do not represent the core theme of the document. The minimum neighbor count requirement helps identify words that occur with high frequency in only one or two phrases and the phrases can be rejected as stop phrases in the specific domain. Candidate phrases containing one or more outlier words are rejected and the remaining phrases pass through as candidate phrases for the next stage of the framework.

Heuristic Efficacy for Thematic Phrases Mining

The *WSEQ* heuristic workflow constructs the word sequence vectors matrix $\overrightarrow{\boldsymbol{WS}}$ in a $\left|\boldsymbol{W^C}\right|$-dimensional vector space. It then projects $\overrightarrow{\boldsymbol{WS}}$ into a δ-dimensional embedding space using tSNE to obtain an embedding vector $\overrightarrow{\boldsymbol{eWS}}$ with reduced dimensionality, where $\delta \ll \left|\boldsymbol{W^C}\right|$. The original *WSEQ* vector and the embedding vector (for $j = \left|\boldsymbol{W^C}\right|$) are as follows:

$$\overrightarrow{\boldsymbol{WS}} \qquad\qquad \overrightarrow{\boldsymbol{eWS}}$$

$$\begin{pmatrix} \mathit{fr}\left(\langle\, w_1, \varnothing \,\rangle\right) & \mathit{fr}\left(\langle\, w_1 \,..\, w_2 \,\rangle\right) & \cdots & \mathit{fr}\left(\langle\, w_1 \,..\, w_j \,\rangle\right) \\ \mathit{fr}\left(\langle\, w_2 \,..\, w_1 \,\rangle\right) & \mathit{fr}\left(\langle\, w_2, \varnothing \,\rangle\right) & \cdots & \mathit{fr}\left(\langle\, w_2 \,..\, w_j \,\rangle\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathit{fr}\left(\langle\, w_j \,..\, w_1 \,\rangle\right) & \mathit{fr}\left(\langle\, w_j \,..\, w_2 \,\rangle\right) & \cdots & \mathit{fr}\left(\langle\, w_j, \varnothing \,\rangle\right) \end{pmatrix} \xrightarrow{\boldsymbol{tSNE}} \begin{pmatrix} e_{11} & e_{12} & .. & e_{1\delta} \\ e_{21} & e_{22} & .. & e_{2\delta} \\ \vdots & \vdots & \ddots & \vdots \\ e_{j1} & e_{j2} & .. & e_{j\delta} \end{pmatrix}$$

The last step of the heuristic workflow identifies outlier words by performing density-based clustering on the $\overrightarrow{eWS_i}$ vectors $\in \overrightarrow{\boldsymbol{eWS}}$ that correspond to each word $w_i \in \boldsymbol{W^C}$. Each cluster is a set of words, $\boldsymbol{Cl_v} = \left\{\, w_a : w_a \in \boldsymbol{W^C} \,\right\}$, such that $\forall w_a \in \boldsymbol{Cl_v}$ there exists one or more sub-clusters $\boldsymbol{Cl_{vm}} \subseteq \boldsymbol{Cl_v}$ that satisfy all the following conditions:

(i) $\boldsymbol{Cl_{vm}} \neq \varnothing$: the sub-cluster is not an empty set

(ii) $w_a \notin \boldsymbol{Cl_{vm}}$: the word does not belong to the sub-cluster

(iii) $|\boldsymbol{Cl_{vm}}| \geq \phi$: the sub-cluster size is at least as large as ϕ, the cluster size threshold

(iv) $\forall\ w_b \in \boldsymbol{Cl_{vm}}, \left|\left|\overrightarrow{eWS_a} - \overrightarrow{eWS_b}\right|\right| \leq \psi$: the distance between the word w_a and each word w_b belonging to the sub-cluster is at most ψ, the distance threshold

Therefore, the outliers are a set of words, $\boldsymbol{\Theta} = \{\ w_o : w_o \in \boldsymbol{W^C}\ \}$, such that, $\forall w_o \in \boldsymbol{\Theta}$ there does not exist a sub-cluster of the form $\boldsymbol{W^{C'}} \subseteq \boldsymbol{W^C}$ that satisfies all the conditions (i)-(iv) described above and expressed in terms of $\boldsymbol{W^{C'}}$ as below:

(i) $\boldsymbol{W^{C'}} \neq \varnothing$

(ii) $w_o \notin \boldsymbol{W^{C'}}$

(iii) $\left|\boldsymbol{W^{C'}}\right| \geq \phi$

(iv) $\forall\ w_b \in \boldsymbol{W^{C'}}, \left|\left|\overrightarrow{eWS_o} - \overrightarrow{eWS_b}\right|\right| \leq \psi$

The *WSEQ* heuristic selects candidate phrases that do not contain any words $\in \boldsymbol{\Theta}$ and makes them available for the next stage of the ThemaPhrase framework. The selected phrases are

$$\boldsymbol{ph^{WS}} = \{\ ph_i^C\ :\ \forall w_o \in \boldsymbol{\Theta}, w_o \notin ph_i^C\ \}$$

The outliers are words that exhibit large deviations in their skip-bigram frequency patterns. This is indicative of the phrases containing these words being either fine-grained, nuanced topical phrases or stop phrases in the document's domain as explained in Sec. 3.4.3. Candidate phrases that are indeed stop phrases need to be rejected from the thematic phrases set; this follows from common intuition that is widely accepted in the information retrieval domain. The rejection of very fine-grained topical phrases is needed because the ThemePhrase framework's objective is to mine thematic phrases that are representative of

the broader thematic basis of the text document.

Consider a candidate phrase $ph_x^C = \langle\, w_i, w_o, w_j \,\rangle$, to understand how the *WSEQ* rejection criterion helps mine $\boldsymbol{ph^{tD}}$. Phrases in a document are formed using sequences of words that are associated with the broader thematic basis of the document and sub-topicality of segments of the document. Thus, if we treat w_i and w_j as random variables keeping w_o constant, then these random variables are:

(a) interdependent and identically distributed as w_o in the context of phrases of the form of ph_x^C for all $w_i, w_j \in \boldsymbol{W^C}$.

(b) interdependent but may or may not be identically distributed in the larger context of the whole document.

If $w_o \in \boldsymbol{\Theta}$, it means $\overrightarrow{\boldsymbol{eWS_o}}$ deviates from $\overrightarrow{\boldsymbol{eWS_x}}, \forall w_x \notin \boldsymbol{\Theta}$ more than the deviations observed among other embeddings. In the context of all phrases of the form ph_x^C in which w_o is fixed, the following is true because of interdependence and distribution characteristics stated in (a) and (b) above:

$$\boldsymbol{fr}\left(\langle\, w_i \,..\, w_j \,\rangle\right) \propto \boldsymbol{fr}\left(\langle\, w_i, w_o \,\rangle\right) \propto \boldsymbol{fr}\left(\langle\, w_o, w_j \,\rangle\right) \tag{3.6}$$

Consider all phrases of the form of ph_x^C and all phrases of the form $\langle\, w_i, w_z, w_j \,\rangle, \forall w_z \notin \Theta$ from the candidate phrases set. When the word sequence vector for w_o indicates that phrases it composes are fine-grained (i.e. low frequency relative to the coarse-grained nounphrases),

then it follows from Eq. (3.6) that:

$$\textit{fr}\left(\langle\, w_i, w_o \,\rangle\right) + \textit{fr}\left(\langle\, w_o, w_j \,\rangle\right) \leq \textit{fr}\left(\langle\, w_i, w_z \,\rangle\right) + \textit{fr}\left(\langle\, w_z, w_j \,\rangle\right) \tag{3.7}$$

The term $\textit{fr}\left(\langle\, w_i \,..\, w_j \,\rangle\right)$ from Eq. (3.6) does not appear in Eq. (3.7) because it is present on either side of the inequality and can be treated as a constant.

Given Eq. (3.6) and (3.7), this inference can be generalized and extended to phrases of any length $\in \boldsymbol{D}$ because presence of $w_o \in \boldsymbol{\Theta}$ in any phrase is consequently indicative of the phrase being a stop phrase or fine grained phrase in the context of $\boldsymbol{tD}$. Therefore, using Eq. (3.4) and (3.5b) with $\sum$ as the function $\mathcal{F}$ it follows that, $\forall ph_x^D = \langle\, w_i \;:\; \nexists w_i \in \boldsymbol{\Theta} \,\rangle$ and $\forall ph_y^D = \langle\, w_j \;:\; \exists w_j \in \boldsymbol{\Theta} \,\rangle$:

$$\sum^{\underline{ph}_x^D} tdPr\left(\, \underline{ph}_{xa}^D \mid \boldsymbol{D} \,\right) \geq \sum^{\underline{ph}_y^D} tdPr\left(\, \underline{ph}_{yb}^D \mid \boldsymbol{D} \,\right) \tag{3.8}$$

$$\text{Hence,} \quad tdPr\left(\, ph_x^D \mid \boldsymbol{D} \,\right) \geq tdPr\left(\, ph_y^D \mid \boldsymbol{D} \,\right) \tag{3.9}$$

Further, all permutations and extensions of ph_y^D will contain the outlier word. Hence, their probability of being part of the set of thematic phrases will be lower than ph_x^D too and will be rejected by the *WSEQ* heuristic.

3.5 Word Position Heuristic (WPOS)

The role every word plays in defining the thematic basis represented by a set of phrases is also indicated by the position at which the word occurs in its corresponding phrases. The word position heuristic (*WPOS*) uses this intra-phrase word position information to filter phrases that are more representative of the thematic basis of a document than other phrases.

The set of candidate phrases, $\boldsymbol{ph^C}$, provided as input to this heuristic is used to construct word position vectors ($\overrightarrow{wp_i}$) for each word, $w_i \in \boldsymbol{W^C}$. Let ρ_d be members of a set of functions such that:

$$\rho_d(ph_x, w_i) = \begin{cases} 1, \text{if } w_i \text{ occurs in } ph_x \text{ at position } d \\ \\ 0, \text{otherwise} \end{cases}$$

where, $0 \leq d \leq \lambda$ and λ is the maximum word length of a thematic phrase as defined in Sec. 3.2. The function $\rho_0 \left(ph_x, w_i \right)$ tests whether w_i occurs at position 0 in ph_x, i.e. whether the phrase is a unigram phrase consisting solely of w_i.

Further, let $\rho(ph_x, w_i) = \widehat{d}$ be a function that finds the word position, $\widehat{d}$, of w_i in ph_x. Let $\rho_d^i = \sum^{\boldsymbol{ph^C}} \rho_d \left(ph_x^C, w_i \right)$ represent the number of times w_i is observed to occur at position d in the phrases in $\boldsymbol{ph^C}$. The word position vector, $\overrightarrow{\boldsymbol{WP}}$, for the set of words $\boldsymbol{W^C}$ takes the form $\overrightarrow{\boldsymbol{WP}} = \left(\overrightarrow{wp_1}, \overrightarrow{wp_2}, \ldots, \overrightarrow{wp_j} \right)$ where $j = \left| \boldsymbol{W^C} \right|$. The vector $\overrightarrow{\boldsymbol{WP}}$ with its individual

word position vectors expanded is as follows:

$$\overrightarrow{\boldsymbol{WP}} = \begin{pmatrix} \overrightarrow{wp_1} \\ \overrightarrow{wp_2} \\ \vdots \\ \overrightarrow{wp_j} \end{pmatrix} = \begin{pmatrix} \rho_0^1 & \rho_1^1 & \cdots & \rho_\lambda^1 \\ \rho_0^2 & \rho_1^2 & \cdots & \rho_\lambda^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_0^j & \rho_1^j & \cdots & \rho_\lambda^j \end{pmatrix}$$

Expanding the vector further by expanding each ρ_d^i entry in $\overrightarrow{\boldsymbol{WP}}$, the complete vector is as follows :

$$\overrightarrow{\boldsymbol{WP}} = \begin{pmatrix} \overrightarrow{wp_1} \\ \overrightarrow{wp_2} \\ \vdots \\ \overrightarrow{wp_j} \end{pmatrix} = \begin{pmatrix} \sum^{\boldsymbol{ph^C}} \rho_0 \left(ph_x^C, w_1 \right) & \sum^{\boldsymbol{ph^C}} \rho_1 \left(ph_x^C, w_1 \right) & \cdots & \sum^{\boldsymbol{ph^C}} \rho_\lambda \left(ph_x^C, w_1 \right) \\ \sum^{\boldsymbol{ph^C}} \rho_0 \left(ph_x^C, w_2 \right) & \sum^{\boldsymbol{ph^C}} \rho_1 \left(ph_x^C, w_2 \right) & \cdots & \sum^{\boldsymbol{ph^C}} \rho_\lambda \left(ph_x^C, w_2 \right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum^{\boldsymbol{ph^C}} \rho_0 \left(ph_x^C, w_j \right) & \sum^{\boldsymbol{ph^C}} \rho_1 \left(ph_x^C, w_j \right) & \cdots & \sum^{\boldsymbol{ph^C}} \rho_\lambda \left(ph_x^C, w_j \right) \end{pmatrix}$$

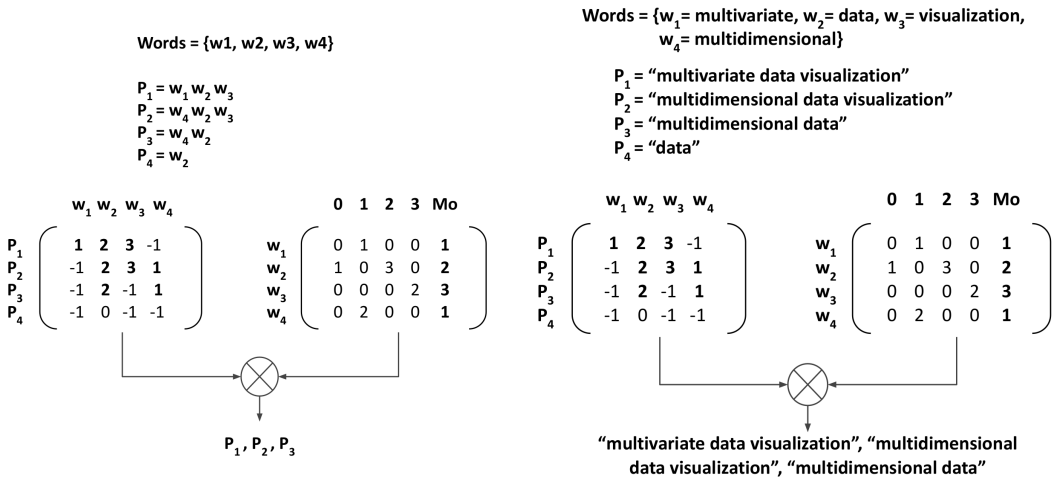
The *WPOS* heuristic uses the most frequently observed position of every word $\in \boldsymbol{W^C}$ as the filtering criteria to select candidate thematic phrases from $\boldsymbol{ph^C}$. The modes of word positions are computed and all ph_x^C containing words that don't occur at their corresponding

mode positions are rejected. The word position mode vector is computed as follows:

$$\overrightarrow{\boldsymbol{MoWP}} = \begin{pmatrix} Mo_1 \\ Mo_2 \\ \vdots \\ Mo_j \end{pmatrix} = \begin{pmatrix} \underset{0 \le d \le \lambda}{arg\,max}\ \rho_d^1 \\ \underset{0 \le d \le \lambda}{arg\,max}\ \rho_d^2 \\ \vdots \\ \underset{0 \le d \le \lambda}{arg\,max}\ \rho_d^j \end{pmatrix}$$

The filtered phrases are:

$$\boldsymbol{ph^{WP}} = \left\{\ ph_x^C\ :\ \forall w_j \in ph_x^C,\ \rho\left(\ ph_x^C, w_j\ \right) = Mo_j\ \right\}$$



(a) Computation (b) Example

FIG. 3.3: Word Position Based Phrase Filtering

Fig. 3.3a visually describes the process of word position vector construction. Fig. 3.3b

shows the same using using four example candidate phrases to show the filtration result. For clarity, both subfigures in Fig. 3.3 shows two matrices to explain the *WPOS* heuristic filtration process as follows:

On the left: Phrase-word position matrix that indicates the position of words $\in \boldsymbol{W^C}$ in each phrase $\in \boldsymbol{ph^C}$. Each entry in the matrix can take on values $-1 \leq d \leq \lambda$. A value of -1 indicates the corresponding word does not occur in the phrase. A value of 0 indicates the phrase is made up of only that word, i.e., it is a unigram phrase. A value ≥ 1 indicates the position at which the word occurs in the phrase.

On the right: This matrix is combination of the two vectors $\overrightarrow{\boldsymbol{WP}}$ and $\overrightarrow{\boldsymbol{MoWP}}$. The last column of the matrix labeled *'Mo'* represents $\overrightarrow{\boldsymbol{MoWP}}$ while all the prior columns represent $\overrightarrow{\boldsymbol{WP}}$.

The word position for each word appearing in a phrase is typeset in bold in the left matrix when the word position is equal to the mode position observed for the word in the candidate phrases. Phrases that have one or more words whose positions in the phrase are not equal to their mode positions are rejected. For example, phrase P_4 in Fig. 3.3a has word w_4 in position 0 while the mode position for w_4 is 1. Hence, P_4 is rejected and phrases P_1, P_2 and P_3 are selected.

Heuristic Efficacy for Thematic Phrases Mining

Consider a candidate phrase $ph_x^C = \langle\, w_i, w_o, w_j \,\rangle$. Phrases in a document are formed

using sequences of words that depend on the thematic basis of the document and sub-topicality of segments of the document. Thus, if we treat w_i and w_j as random variables keeping w_o constant, then these random variables are:

(a) interdependent and identically distributed as w_o in the context of phrases of the form of ph_x^C for all $w_i, w_j \in \boldsymbol{W^C}$.

(b) interdependent but may or may not be identically distributed in the larger context of the whole document.

Consider two cases for the word positions in a candidate phrase of the form of ph_x^C:

Case 1: $\rho\left(\, ph_x^C, w_o \,\right) \neq \boldsymbol{Mo_o}$ In the context of all phrases of the form ph_x^C in which w_o is fixed, the following is true because of interdependence and identical distribution characteristics stated in point (a) above:

$$\boldsymbol{fr}\left(\langle\, w_i, w_o \,\rangle\right) \propto \boldsymbol{fr}\left(\langle\, w_o, w_j \,\rangle\right) \tag{3.10}$$

Further, since w_o is not in its observed mode word position, either w_i or w_j or both must not be in their observed mode word positions. We are not considering the case where only w_o is out of position, since such a case implies that another word $\notin w_i, w_o, w_j$ can take its place to form a phrase with all words in their mode word positions. Such cases are not considered here because this discussion focuses on relative thematic representativeness between phrases that must contain w_o. Also, if

no other candidate phrase exists that is a permutation, subphrase or extension of ph_x^C containing w_o, then it is trivial to note that w_o cannot be out of its observed mode word position in ph_x^C. Thus, such candidate phrases must exist and they are considered in this treatment.

It follows from Eq. (3.10) that for all phrases of the form $ph_x^C = \langle\, w_i, w_o, w_j \,\rangle$ and for $\boldsymbol{ph_y^C} = \left\{ \widetilde{ph}_{xi}^C : \widetilde{ph}_{xi}^C \in \boldsymbol{ph^C} \;,\; \rho(\widetilde{ph}_{xi}^C, w_o) = \boldsymbol{Mo}_o \right\}$

$$\begin{aligned} \boldsymbol{fr}\left(\langle\, w_i, w_o \,\rangle\right) + \boldsymbol{fr}\left(\langle\, w_o, w_j \,\rangle\right) &\leq \sum^{\boldsymbol{ph_y^C}} \boldsymbol{fr}\left(\underline{ph}_{yi}^C\right) \\ \sum^{\boldsymbol{ph_x^C}} \boldsymbol{fr}\left(\underline{ph}_{xi}^C\right) &\leq \sum^{\boldsymbol{ph_y^C}} \boldsymbol{fr}\left(\underline{ph}_{yi}^C\right) \\ tdPr\left(\, ph_x^C \mid \boldsymbol{D} \,\right) &\leq tdPr\left(\, ph_y^C \mid \boldsymbol{D} \,\right) \end{aligned}$$

This case applies to any word in ph_x^C that is at a word position in the phrase that is not equal to the most frequently observed position for that word in phrases $\in \boldsymbol{ph^C}$

Case 2: $\rho\left(\, ph_x^C, w_i \,\right) = \boldsymbol{Mo}_i$, $\rho\left(\, ph_x^C, w_o \,\right) = \boldsymbol{Mo}_o$, $\rho\left(\, ph_x^C, w_j \,\right) = \boldsymbol{Mo}_j$

In the context of all phrases of the form ph_x^C in which w_o is fixed:

(a) All permutations of ph_x^C will displace one or more words in the phrase from their most frequently observed word positions. Consequently, two of more words will occupy positions in the permuted phrase which are not equal to their

corresponding most frequently observed word positions. Hence,

$$\forall ph_y^C \in \widetilde{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{C}}\ , \quad \sum^{\underline{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{C}}} tdPr\left(\underline{ph}_{xa}^{C} \mid \boldsymbol{D} \right) > \sum^{\underline{\boldsymbol{ph}}_{\boldsymbol{y}}^{\boldsymbol{C}}} tdPr\left(\underline{ph}_{yb}^{C} \mid \boldsymbol{D} \right)$$

(b) All subphrases of ph_x^C, considered individually as candidates for thematic phrases, will displace all the words in the subphrases to words positions that are not equal to their respective most frequently observed word positions. Hence,

$$\forall ph_y^C \in \underline{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{C}}\ , \quad \sum^{\underline{\boldsymbol{ph}}_{\boldsymbol{x}}^{\boldsymbol{C}}} tdPr\left(\underline{ph}_{xa}^{C} \mid \boldsymbol{D} \right) > \sum^{\underline{\boldsymbol{ph}}_{\boldsymbol{y}}^{\boldsymbol{C}}} tdPr\left(\underline{ph}_{yb}^{C} \mid \boldsymbol{D} \right)$$

Both the cases above can be trivially generalized and extended to phrases of any length $\in \boldsymbol{D}$. Therefore, it follows that, $\forall ph_x^D = \left\langle\ w_i\ :\ w_i \in ph_y^D, \forall w_i, \rho\left(\ ph_x^D, w_i\ \right) = \boldsymbol{Mo}_i\ \right\rangle$ and $\forall ph_y^D = \left\langle\ w_j\ :\ \exists w_j, \rho\left(\ ph_y^D, w_j\ \right) \neq \boldsymbol{Mo}_j\ \right\rangle$ the following are satisfied:

$$\sum^{\underline{ph}_x^D} tdPr\left(\underline{ph}_{xa}^{D} \mid \boldsymbol{D} \right) > \sum^{\underline{ph}_y^D} tdPr\left(\underline{ph}_{yb}^{D} \mid \boldsymbol{D} \right)$$

$$\text{Hence,} \quad tdPr\left(\underline{ph}_{x}^{D} \mid \boldsymbol{D} \right) > tdPr\left(\underline{ph}_{y}^{D} \mid \boldsymbol{D} \right)$$

Further, the permutations, subphrases and extensions of ph_x^D, as described in *"Case 2"* above, will have lower probability to be part of the set thematic phrases than ph_x^D and will be rejected by the *WPOS* heuristic if they are part of $\boldsymbol{ph}^{\boldsymbol{C}}$.

3.6 Word Association Heuristic (WASS)

Filtration based on the above two heuristics help mine thematic phrases based on structural properties that hint at their representativeness of the document's thematic basis. The *WASS* heuristic induces association rules using a sparse data association rules inducer [61]. This heuristic helps to filter phrases based on co-occurrences of frequent word sets. The intuition here is that phrases that share frequent word-sets with high-confidence associations will collectively represent the thematic basis of the document better than other phrases. The filtration effect of this heuristic can be tuned using the support and rule confidence parameters of association rule mining to alter the granularity or resolution of thematic phrases.

Consider the set of candidate phrases $\boldsymbol{ph^C}$ provided as input to the *WASS* heuristic. The association rule mining algorithm considers each phrase ph_x^C as an itemset of words and induces a set of association rules, $\boldsymbol{R} = \{\ r_g : g \geq 1\ \}$. Each rule, r_g, is an association rule between two itemsets, i.e. disjoint set of words $\in \boldsymbol{W^C}$, and takes the form $r_g = (\ \mathcal{A}_g \rightarrow \mathcal{B}_g, \mathcal{S}_g, \mathcal{C}_g, \mathcal{L}_z\ )$. The definitions and descriptions of all the components of the association rule and related concepts are as follows:

Contains Function : This function receives two sets, $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$, as arguments and checks whether the second set is a subset of the first. It is defined as:

$$\textit{contains}\ (\ \boldsymbol{s_1}, \boldsymbol{s_2}\ ) = \begin{cases} 1 & \text{iff } \boldsymbol{s_2} \subseteq \boldsymbol{s_1} \\ 0 & \text{otherwise} \end{cases}$$

Antecedent ($\mathcal{A}$) and Consequent ($\mathcal{B}$) : $\mathcal{A}_g = \{ w_i \}$ is the antecedent of the rule, $\mathcal{B}_g = \{ w_j \}$ is the consequent of rule. They both satisfy all the following conditions: $\mathcal{A}_g \cap \mathcal{B}_g = \varnothing$, $\mathcal{A}_g \subseteq \boldsymbol{W^C}$ and $\mathcal{B}_g \subseteq \boldsymbol{W^C}$

Support ($\mathcal{S}$) : Support of the rule is the proportion of phrases that contain both the antecedent and consequent set of words of the association rule.

$$\mathcal{S}_g = \frac{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_x, \mathcal{A}_g \cup \mathcal{B}_g \right)}{\left|\boldsymbol{ph^C}\right|}$$

Confidence ($\mathcal{C}$) : Confidence of a rule is the ratio of the support for the association rule $\mathcal{A}_g \rightarrow \mathcal{B}_g$ to the support of the antecedent $\mathcal{A}_g$. It can be trivially reduced to the ratio of the number of phrases that contain both the antecedent and consequent set of words of the association rule to the number of phrases that contain the antecedent as shown below.

$$\begin{aligned}\mathcal{C}_g &= \frac{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_x, \mathcal{A}_g \cup \mathcal{B}_g \right)}{\left|\boldsymbol{ph^C}\right|} \div \frac{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_y, \mathcal{A}_g \right)}{\left|\boldsymbol{ph^C}\right|} \\ &= \frac{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_x, \mathcal{A}_g \cup \mathcal{B}_g \right)}{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_y, \mathcal{A}_g \right)}\end{aligned}$$

Lift ($\mathcal{L}$) : Lift of a rule is the ratio of the confidence of the association rule $\mathcal{A}_g \rightarrow \mathcal{B}_g$ to the

support of the consequent $\mathcal{B}_g$.

$$\mathcal{L}_g = \frac{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_x, \mathcal{A}_g \cup \mathcal{B}_g\right)}{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_x, \mathcal{A}_g\right)} \div \frac{\sum^{\boldsymbol{ph^C}} \textit{contains}\left(\underline{ph^C}_x, \mathcal{B}_g\right)}{\left|\boldsymbol{ph^C}\right|}$$

The association rules inducer takes two parameters as input: (i) the minimum support threshold ($\mathcal{S}_o$) and (ii) the minimum confidence threshold ($\mathcal{C}_o$). The inducer uses these parameters to filter the induced association rules and only generates rules that satisfy the following condition: $\mathcal{S}_g \geq \mathcal{S}_o \wedge \mathcal{C}_g \geq \mathcal{C}_o$

Given the set of association rules $\boldsymbol{R}$, the *WASS* heuristic uses the rules to score and rank the candidate thematic phrases $\boldsymbol{ph^C}$ and creates a final rank list of thematic phrases $\boldsymbol{ph^{WA}}$. The ThemaPhrase framework can then select thematic phrases either using a top-k strategy to select the *k* top phrases ranked by their scores or use a threshold parameter γ to select thematic phrases with scores $\geq \gamma$.

The score, wa_x, for each phrase is computed based on the confidence and lift values of association rules that the phrase satisfies. The score for a phrase and association rule pair is defined as

$$score(ph_x^C, r_g) = \begin{cases} \mathcal{C}_g * \mathcal{L}_g & \text{if } \textit{contains}\left(ph_x^C, \mathcal{A}_g \cup \mathcal{B}_g\right) \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

Using the definitions of confidence ($\mathcal{C}_g$) and lift ($\mathcal{L}_g$) earlier in the chapter, the non-zero part

of the score function in Eq. (3.11) can be expanded as shown below:

$$
\begin{aligned}
score\left(ph_x^C, r_g \right) &= \mathcal{C}_g * \mathcal{L}_g \\
&= \frac{Pr\left(\mathcal{A}_g \wedge \mathcal{B}_g \right)}{Pr\left(\mathcal{A}_g \right)} * \frac{Pr\left(\mathcal{A}_g \wedge \mathcal{B}_g \right)}{Pr\left(\mathcal{A}_g \right)} * \frac{1}{Pr\left(\mathcal{B}_g \right)} \\
&= \frac{Pr\left(\mathcal{A}_g \wedge \mathcal{B}_g \right)}{Pr\left(\mathcal{A}_g \right)} * \frac{Pr\left(\mathcal{A}_g \wedge \mathcal{B}_g \right)}{Pr\left(\mathcal{B}_g \right)} * \frac{1}{Pr\left(\mathcal{A}_g \right)} \\
&= Pr\left(\mathcal{B}_g | \mathcal{A}_g \right) * Pr\left(\mathcal{A}_g | \mathcal{B}_g \right) * \frac{1}{Pr\left(\mathcal{A}_g \right)} \\
&= \frac{Pr\left(\mathcal{B}_g | \mathcal{A}_g \right) * Pr\left(\mathcal{A}_g | \mathcal{B}_g \right)}{Pr\left(\mathcal{A}_g \right)}
\end{aligned}
\tag{3.12}
$$

Eq. (3.12) shows that the score of a rule is rewarded for higher frequency of co-occurrence of its antecedents and consequents conditioned on one another in the set of candidate thematic phrases and is biased towards high frequency consequents. On the other hand, the score is penalized if consequents co-occur with antecedents other than those of the rule more frequently. The net score, wa_x, for a phrase ph_x^C is the cumulative score of all the association rules $\boldsymbol{R}$ that apply to it as follows:

$$
\begin{aligned}
wa_x &= \sum^{r_g \in \boldsymbol{R}} score\left(ph_x^C, r_g \right) \\
&= \sum^{r_g \in \boldsymbol{R}} \frac{Pr\left(\mathcal{B}_g | \mathcal{A}_g \right) * Pr\left(\mathcal{A}_g | \mathcal{B}_g \right)}{Pr\left(\mathcal{A}_g \right)}
\end{aligned}
\tag{3.13}
$$

Heuristic Efficacy for Thematic Phrases Mining

Given Eq. (3.13), phrases that score higher have two broad contributing factors:

(a) Applicable association rules wherein cooccurrence of the rules' respective antecedents and consequents is much higher than the cooccurrence of their consequents with any other antecedents.

(b) Count of applicable of rules.

Since association rules deal with itemsets, the seqeunce or ordering of the items, that are words in the thematic phrases, is inconsequential. Hence, discussing this heuristic in the context of permutations of a candidate phrase is immaterial. Since, wa_x is a sum of scores of all applicable rules, it may appear that candidate phrases that are extensions of other candidate phrases will be invariably scored higher than their corresponding subphrases. But, this is not the case because if the extensions are formed using words $\in \boldsymbol{W^C}$ that are infrequent (i.e form fine-grained thematic phrases with low occurrence frequencies) then their corresponding association rules may not meet the S_0 and/or C_0 thresholds. Consequently, those rules will not contribute to the *WASS* scores for the extension phrases. This implies that the extensions and subphrases will have the same score. The *WASS* heuristic is intended to generate an appropriate rank order for the thematic phrases and not filter out any. The filtration is done using *WSEQ* and *WPOS* heuristics.

Thus, phrases in $\boldsymbol{ph^C}$ ranked by their *WASS* scores (defined in Eq. (3.13)) are analogous to ranking the phrases by the probability of the phrases belonging to the set of thematic phrases $\boldsymbol{ph^{tD}}$ of size $z = \left|\boldsymbol{ph^C}\right|$. Thus, applying a score or count threshold for selecting a $\boldsymbol{ph^{tD}}$ of size $z < \left|\boldsymbol{ph^C}\right|$ still provides a set of thematic phrases where Eq. (3.3) holds.

3.7 ThemaPhrase Framework Configurations

The ThemaPhrase framework can be configured to use any combination of the three phase-level heuristics discussed above. The LDA topic model and nounphrase extraction component is a constant in order to prepare the initial set of candidate phrases. Also, the *WASS* heuristic is required if we need a ranked list of thematic phrases. This is the case for all experiments conducted in this work and, thus, *WASS* will be used in all ThP configurations. The evaluation of the thematic phrases extracted by the framework using all its heuristic combination configurations is discussed in the next chapter.

Chapter 4

DATASETS AND PACKAGES

This chapter describes the datasets and software packages used for experimentation and evaluation in this work. Sec. 4.1 describes the datasets used for evaluating various thematic phrases extraction approaches including a comparison of their distinct characteristics and the preprocessing workflow for the text documents. Sec. 4.2, it describes various software and programming packages used for implementation of the ThemaPhrase framework, modeling using competing methods and evaluation of the various thematic phrases extraction methods.

4.1 Datasets

The different thematic phrases extraction methods are evaluated using two datasets that are corpora of text documents with diverse length and verbosity. These two aspects of a text document are important to analyze the consistency in quality of thematic phrases extracted by different approaches, their sensitivity to text length and word repetition as well

as their robustness in application to diverse textual corpora. Sec. 4.1.1 and 4.1.2 describe the two datasets used in this work and Sec. 4.1.3 compares the two datasets based on various characteristics.

4.1.1 PubMed PMC Original Research Contributions

This dataset consists of approximately 15,000 research publications selected from the PubMed Open Access Subset (OAS)[1]. The OAS archive publishes the fulltext and metadata of the archived research articles as XML files. This dataset is a collection of original research contributions that have titles beginning with the letters *'A'* or *'B'* that are extracted from the non-commercial use PubMed OA archive[2], .

Original research contributions are indicated by the *'subject'* field in the publication XML files. This dataset consists of all publications with the following *'subject'* field values: *'Original Article', 'Original Articles', 'Research Paper', 'Research Papers', 'Original Paper', 'Research Article', 'Research Articles', 'Research-Article', 'Original Research'*.

The title, abstract and body of the publications are extracted by parsing the publications' XML files made available by PubMed. Tab. 4.1 provides the XML parse paths for the various metadata and fulltext body of the research publications discussed above.

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[2] https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/ *(Retrieved June 2019)*

Element	XML Parse Path
Subject	./front/article-meta/article-categories/subj-group/subject
Title	./front/article-meta/title-group/article-title
Abstract	./front/article-meta/abstract
Body	./body

Table 4.1: Pubmed PMC Original Research Contributions: XML Parse Paths for Metadata and Fulltext Body

4.1.2 USPTO Granted Patents

This dataset consists of approximately 12,500 patent applications randomly selected from a larger corpus of patent applications that have been granted in the United States. The United States Patent and Trademark Office (USPTO) made the fulltext and metadata of all granted patents available to the public in collaboration with ReedTech[3]. This dataset consists of patent applications randomly sampled from those granted from January 2019 through June 2019.

The documents are made available in XML form by USPTO. The title, abstract, list of claims and body of the patents are extracted by parsing the patents' XML files. Tab. 4.2 provides the XML parse paths for the various metadata and fulltext body of the patents discussed above.

Element	XML Parse Path
Title	./us-bibliographic-data-grant/invention-title
Abstract	./abstract
Claims	./claims
Body	./description

Table 4.2: USPTO Granted Patents: XML Parse Paths for Metadata and Fulltext Body

[3]https://www.patents.reedtech.com *(Retrieved September 2019)*

4.1.3 Dataset Characteristics Comparison

A text document can be considered to be a sequence of words, phrases or sentences depending on the granularity that a use case requires. The length of the document can be similarly expressed in terms of word count, phrase count and sentence counts. In most statistical methods, including topic models and thematic phrase extraction methods, the frequency of occurrence of tokens (unique words) plays a key role in model estimations. Thus, the evaluation of thematic phrase extraction methods must consider text documents that have varying document lengths as well as varying token frequency distributions.

Text documents in the two datasets exhibit diversity in document lengths as well as token frequency distributions. The following metrics are considered to assess these diversities:

(a) **Word count (WC):** total number of words present in a document

(b) **Sentence count (SC):** total number of sentences in a document

(c) **Token count (TC):** total number of unique words used in a document

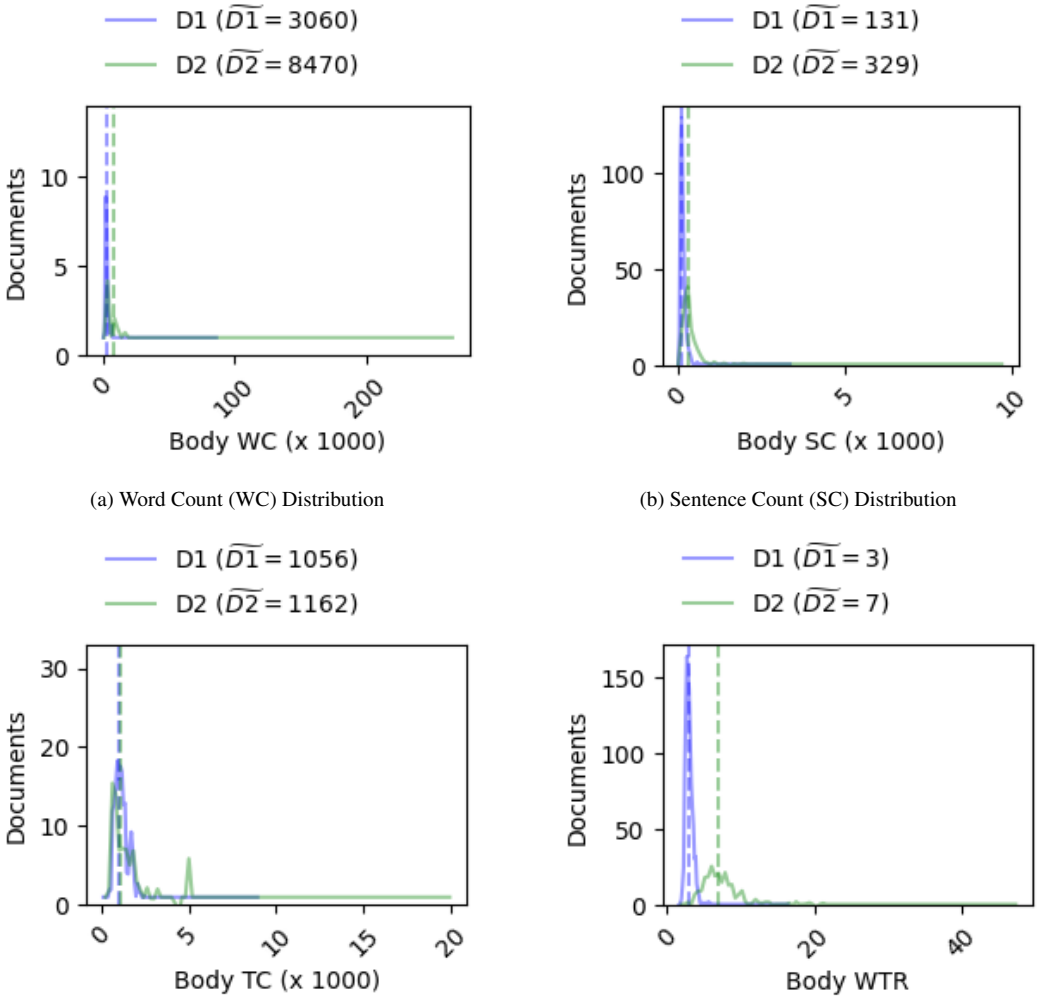
(d) **Word-token count ratio (WTR):** the ratio of word count to token count as a measure of the average repetition in token usage in a document

The word and sentence counts provide a measure of length of documents at two different granularities. Further, these two counts considered collectively provide a measure of sentence-word densities that may affect POS taggers and topic models sensitive to document segment word density. The token counts provide a measure of the dictionary sizes across

documents and the word-token count ratio provides a view of the varying token frequency distributions as a measure of token repetition (or redundancy specific to certain domains such as patents) across documents in the two datasets. All plots in this section will use "D1" to refer to the PubMed dataset and "D2" to refer to the USPTO dataset.

All thematic phrase extraction methods in this work consume text from the body of documents as input. We consider the spread of the four metrics described above for documents bodies. The distribution plots for WC, SC, TC and WTR of document bodies are shown in Fig. 4.1. Both datasets have documents with diverse values across all metrics. The average patent is almost three times the length of the average research publication. The median values of WC (refer Fig. 4.1a) and SC (refer Fig. 4.1b) for patents (D1) and research publications (D2) reflect this. Further, higher document lengths of patents are primarily because of a proportionally higher repetitive use of words relative to that of research publications. This is evident from similar TC for an average patent and research publication (refer Fig. 4.1c) but a WTR for an average patent that is more than twice that of a research publication.

The thematic phrase extraction methods discussed in this work consume individual text documents partitioned into uniformly sized segments as input. The uniform size is computed in terms of number of sentences per segment. We consider different segment counts per document for a thorough evaluation and assessment of the robustness of thematic phrase extraction methods to word densities in document segments that vary for different segment counts. A detailed evaluation and discussion on this is provided in Sec. 5.5.5. It is



(a) Word Count (WC) Distribution

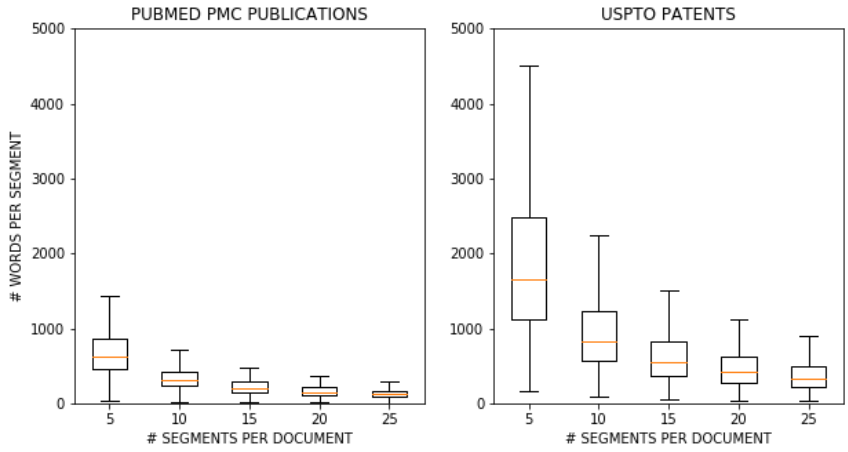
(b) Sentence Count (SC) Distribution

(c) Token Count (TC) Distribution

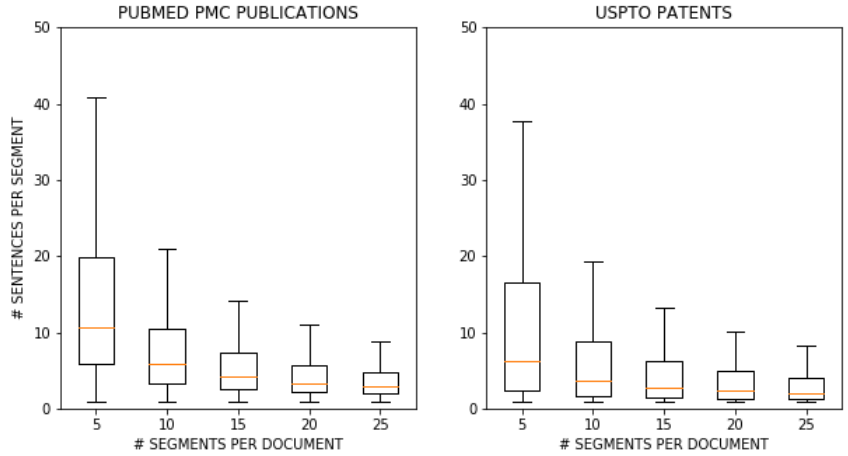
(d) Word-Token Count Ratio (WTR) Distribution

FIG. 4.1: Statistics for Document Bodies

important that there is diversity in word and sentence counts per segment for documents from both datasets. The diversity in word and sentence counts per document segment is shown in Fig. 4.2 as box plots for both datasets. It is trivial to note that the words and sentences per segment decrease as the number of segments per document increase. It is important to note, however, that the spread of both word and sentence counts per segment



(a) Word Count (WC) Distribution



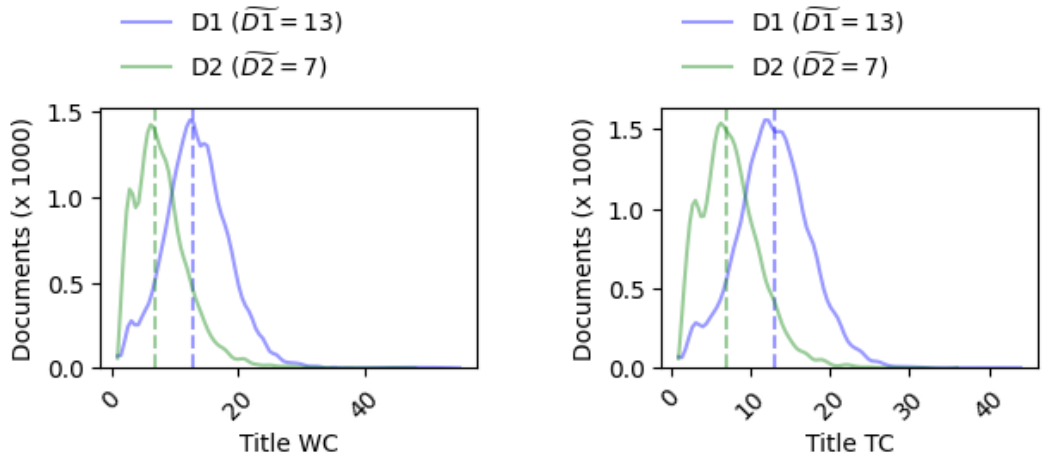
(b) Sentence Count (SC) Distribution

FIG. 4.2: Statistics for Document Body Segments

have different distributions for the two datasets. This ensures a robust evaluation of the thematic phrase extraction methods.

Titles and Abstracts as Gold Standards Patents as well as research publications have titles and abstracts associated with them. These are human synthesised summaries at two largely different granularities and can be considered good representations of the core

themes of their respective documents. It is important to assess the characteristics of these two parts of the documents if they are to be used as reference gold standards for evaluating the quality of extracted thematic phrases and document summaries.



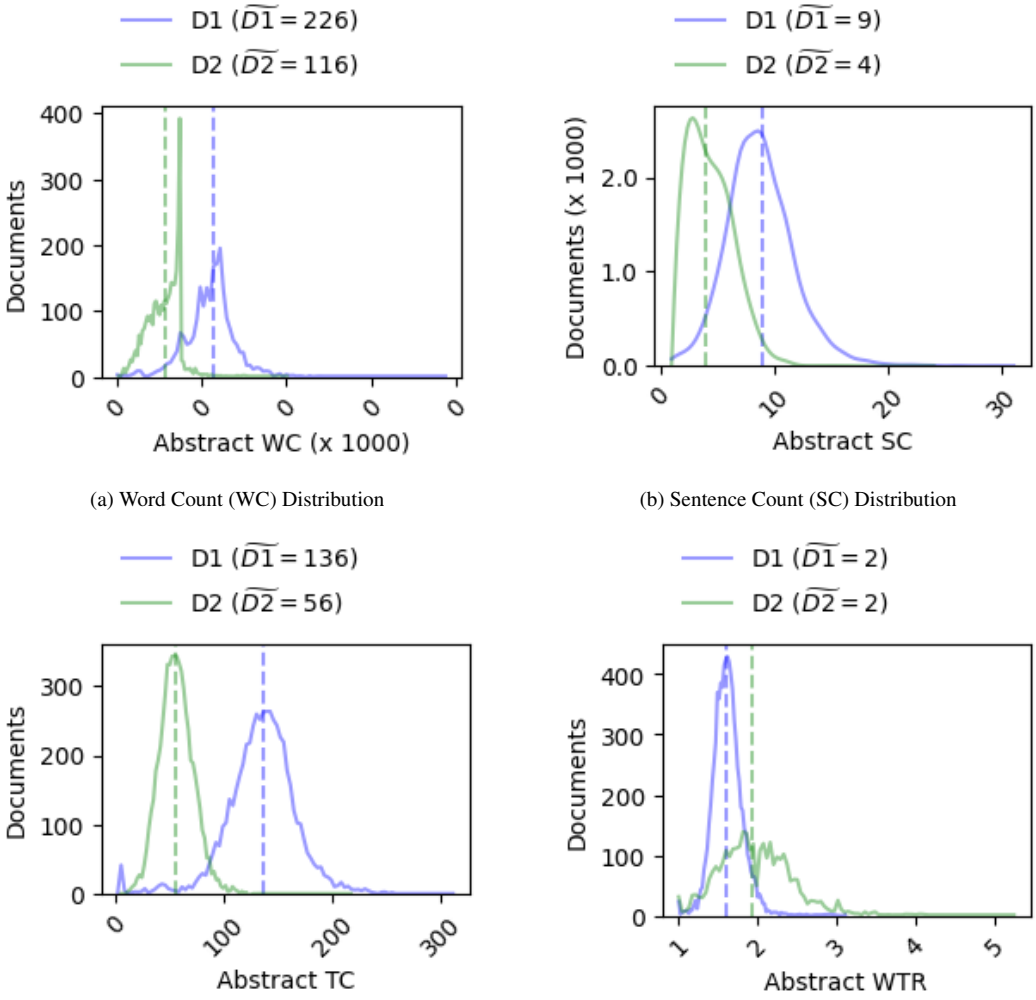
(a) Word Count (WC) Distribution

(b) Token Count (TC) Distribution

FIG. 4.3: Statistics for Document Titles

Fig. 4.3 shows the WC and TC for document titles in the two datasets. Since it is unusual to have multi-sentence titles, the SC metric is not plotted. The diversity in WC and TC across the two datasets as well as across documents in each dataset is shown in the plots. Patents have shorter titles than research publications. The WTR of titles in both datasets is approximately 1.

Fig. 4.4 shows the WC, SC, TC and WTR for document abstracts in the two datasets. The abstracts of research publications are twice as long as those of patents on average in terms of WC and SC (refer Fig. 4.4a and 4.4b respectively). Fig. 4.4d shows that the WTR distribution for patents has a wider spread than that for research publications while the



(a) Word Count (WC) Distribution

(b) Sentence Count (SC) Distribution

(c) Token Count (TC) Distribution

(d) Word-Token Count Ratio (WTR) Distribution

FIG. 4.4: Statistics for Document Abstracts

median WTR for both datasets is comparable. In general, the abstracts of patents, although shorter, have the same token repetition as that in the abstracts of research publications. This indicates that the longer research publication abstracts use a diverse set of tokens and their nounphrases are, therefore, representative of multiple themes and/or sub-themes of the core theme.

Noun phrases extracted from document titles and abstracts are used as the reference gold standard for all thematic phrase evaluations in this work. A manual inspection of the nounphrases extracted from a random sample of 10 documents each from the PubMed and USPTO datasets shows that 66.3%($\pm 3.34\%$) and 84.1%($\pm 3.02\%$) nounphrases respectively are extracted without punctuation errors and are thematically relevant to the corresponding documents. The remainder of nounphrases are mostly generic nounphrases, atleast in the context of the documents' themes, such as *"method"*, *"high rate"*, *"original study"* and *"present invention"*. Using the samples as estimators for the dataset-wide statistics, the percentage of nounphrases extracted from abstracts that can be expected to be thematically relevant lie in the following confidence intervals: 95% CI [59.73%,72.87%] for the PubMed dataset and 95% CI [78.19%, 90.01%] for the USPTO dataset.

Using abstract nounphrases as the gold standard for evaluation of extracted thematic phrases will allow us to understand their representativeness of core themes as well as any important sub-themes or other related themes based on the above discussion. It should also be noted, that patents usually have a narrow focus on a particular invention and its description. Research publications, on the other hand, deal with multiple concepts that are central to one or more research contributions discussed in the document. This qualitative difference in the abstracts and basic nature of documents in the two datasets enables a more robust and holistic evaluation of the thematic phrases extraction methods as well as extractive summarization.

4.2 Libraries and packages

spaCy *(Version: 2.x.x ; Website: https://spacy.io/)*

spaCy a natural language processing toolkit with a python implementation. This work uses spaCy to parse English language sentences within text documents and extract nounphrases from them.

Gensim *(Version: 3.x.x ; Website: https://radimrehurek.com/gensim/)*

Gensim is a topic modelling toolkit implemented as a python package. It includes a modified implementation of LDA [23] with online parameter estimation [62] and can be run on multi-core and cluster hardware configurations. This work uses Gensim's LDA implementation for unigram topic modeling and topical nounphrase estimation in the first, pre-heuristic stage of the ThemaPhrase framework. This work also uses Gensim's implementation of an improved version [63] of the TextRank [54] summarizer for extractive summarization of text documents discussed in Chapter 6.

Mallet *(Version: 2.x.x ; Website: http://mallet.cs.umass.edu/)*

MALLET [64] is a Java-based package that provides natural language processing and several text-based machine learning utilities. This work uses MALLET's Topical-Ngrams (TNG) [28] implementation and adapts it for single document topical phrase extraction as a competing method to ThemaPhrase configurations.

Autophrase *(Version: 3.x.x ; Website: https://github.com/shangjingbo1226/AutoPhrase)*

This is an implementation of AutoPhrase [22] with the standard, english wiki_phrases knowledge base. This work uses the package for thematic phrase extraction using AutoPhrase as a competing method to ThemaPhrase configurations.

ROUGE 2.0 *(Version: 1.x.x ; Website: https://rxnlp.com/rouge-2-0/)*

ROUGE 2.0 is an evaluation toolkit for computer generated summaries. The toolkit is a Java implementation of ROUGE scores and is used to quantitatively evaluate summaries generated using extractive summarization in this work.

Chapter 5

THEMATIC PHRASES EXTRACTION: QUANTITATIVE ANALYSIS

This chapter discusses the evaluation of thematic phrases extracted from a single text document by different configurations of the ThemaPhrase (ThP) framework and two other competing methods, one from the family of topical phrase extraction methods and the other from the LDA family of topic models. The first competing method is AutoPhrase (AP). AutoPhrase extracts topical phrases from a text collection and provides them in the form of a ranked list in its first stage. This output from its first stage suffices for the task of thematic phrase extraction and we do not consider the second stage of AutoPhrase that deals with bag-of-phrases topic induction using the extracted phrases. AutoPhrase is a relevant candidate as a competing method for the task of thematic phrase extraction from a single text document by using segments of a document as its input text collection.

The second competing method is the Topical N-Grams (TNG) topic model that can be adapted for the task of thematic phrase extraction from a single document. TNG is a robust

and widely used n-gram topic model for large text corpora and it is adapted for the task of topical phrase extraction from single documents as described in Sec. 5.1.

These two competing methods are widely cited and used for n-gram topic modeling and topical phrase extraction respectively. Further, though the task of extracting thematic phrases from a single text artifact is different from the common task of achieving this for a corpus, these two methods are applicable since the single text artifact is partitioned and can serve as an input text collection for these methods.

This chapter is organized into the following sections. Sec. 5.1 describes the adaptation of the TNG model for thematic phrase extraction from single text documents. Sec. 5.2 describes the experiment setup for the evaluation of thematic phrase extraction methods and includes various parameter values used to run each competing method for thematic phrase extraction. Sec. 5.3 discusses nuances of the thematic phrases extracted from two example documents using TNG, AP and one ThP configuration. Sec. 5.4 details different quantitative metrics that are used to evaluate the thematic phrases extracted by each competing method. Sec. 5.5 analyzes and compares the quality of the thematic phrases extracted by each method using the quantitative metrics discussed in Sec. 5.4. This section utilizes various combinations of quantitative metrics to compare the thematic phrase extraction methods. It also visualizes these combinations of metrics for ease of readability. The complete tabular representation of all the values of the quantitative metrics for all experiments are provided in Appendices B to J.

5.1 Adapting Topical N-Grams (TNG) for Single Document Thematic Phrases Extraction

A document is a collection of words and phrases associated with the core themes of the document that are arranged in a coherent manner across paragraphs and sections. Hence, partitions of a document are bound to have overlapping words and phrases that are representative of the document's core themes. As discussed in Chapter 4, the input to all thematic phrase extraction methods in this work is a set of partitions of a single text document that have uniform sentence counts. These partitions are treated as a text document corpus/collection by components in each method that require a text collection as input.

Topical N-Grams (TNG) models topics as a bag of n-grams distributed over a collection of documents. Given an evenly partitioned text document as the input corpus for TNG, the topics will have overlapping n-grams due to coherency in the document discourse based on its core themes. TNG can be adapted for our task by consolidating the TNG topics, that are bags-of-phrases, into a single ranked list of topical phrases that are representative of the theme of the document.

TNG infers the likelihood that an n-gram in a topic will be generated at a position in the document given that topic. Consider a set $\{ ph_i^D \}$ of n-grams in the document $\boldsymbol{D}$ where the $L\left(ph_i^D, T_z \right)$ is the likelihood of ph_i^D being generated at a n-gram position in the document given the topic T_z. The set of topics is finite and $\sum_{1}^{z} \boldsymbol{Pr}\left(T_z \right) = 1$ for generating any n-gram in the document. Hence, the likelihood of an n-gram ph_i^D being generated at a

position in the document and, consequently, the frequency of that n-gram being generated in the whole document relative to all other n-grams is proportional to $\sum_{1}^{z} L\left(ph_i^D, T_z \right)$.

Thus, to consolidate all the n-grams across all the topics induced by TNG for the input document we add the likelihood values of each n-gram across all topics and then sort the consolidated list to generate a ranked list of topical n-grams or phrases. We can then slice the ranked list to select the top-k thematic phrases.

5.2 Experiment Setup

Experiments to evaluate the quality of thematic phrases extracted from documents by the competing methods are conducted using the two datasets described in Chapter 4, namely the PubMed research publications dataset and the USPTO patent dataset. Each method receives uniformly sized partitions of each document in the two datasets as input. The methods then perform their respective thematic phrase mining workflows on the document partitions and extract a ranked list of thematic phrases. The evaluation considers three top-k slices of the ranked list of thematic phrases generated by each method: top-5, top-10 and top-20.

Each document in the dataset is partitioned into five different segment counts $\in \{ 5, 10, 15, 20, 25 \}$ as described in Chapter 4. This is done to assess the effects of the number of partitions as well as the size of the partitions (in terms of sentence/phrase/word counts) on each method. Sec. 5.5.5 discusses the effects of segment counts on the efficacy of the thematic phrases extracted by each method.

The following subsections describe the parameters for each thematic phrase extraction method and the values used for each of them for the experiments conducted in this work. This provides clarity about the experiment setup as well as sufficient information for reproducing the experiments.

5.2.1 ThemaPhrase Framework (ThP)

The ThemaPhrase (ThP) framework utilizes the unigram LDA topic model as its first stage followed by a pre-specified combination of the three phrase filtration heuristics. The LDA topic model requires two parameters to be specified, namely, the number of topics (z) and the number of top words per topic (m) to be induced for the input text. Refer to Sec. 3.3 for details on these parameters.

The word sequence heuristic (WSEQ) in ThP performs density-based clustering and requires two input parameters that define density thresholds for cluster generation. They are the maximum allowable distance between neighboring points (ψ) and the minimum number of neighboring points (ϕ). Refer to Sec. 3.4.3 for details on these parameters. The word association heuristic (WASS) in ThP utilizes frequent itemset mining to generate association rules. This heuristic requires two parameters to be specified that serve as thresholds for association rules to be considered acceptable. They are the minimum support (S_0) and the minimum confidence (C_0) required for each association rules. Refer to Sec. 3.6 for details on these parameters.

The values used for each of these parameters to conduct experimentats in this work

are specified in Tab. 5.1. The number of topics and top words per topic are consistent across all ThP configurations and TNG. This is to ensure comparability of performance of these methods. The density thresholds as well as association rule thresholds for WASS in ThP are kept sufficiently low in order to reject only extreme outliers and relying on the final scoring and ranking of the candidate phrases by the WASS heuristic to generate the top-k thematic phrases.

Parameter		**Value**
LDA - Number of Topics		4
LDA - Top words per Topic		10
WSEQ Density Thresholds	Max Distance	10%tile of distances from centroid
	Min Neighbors	5
WASS Thresholds	Min Support	1%
	Min Confidence	0.01

Table 5.1: ThemaPhrase (ThP) Framework Parameters for Experiments

The ThP configurations for experiments conducted in this work are broadly separated into two groups. The first group does not consider the phrase occurrence frequencies within the input document as a factor in its computation. That is, the candidate phrases are analyzed purely based on the structure and semantics of unique candidate phrases relative to others without taking into account the actual frequency with which each candidate phrase occurs in the document. The motivation for this setup of the ThP configurations is to assess if the set of candidate phrases can solely provide cues to the structure, both syntactic as well as semantic, of phrases that represent the thematic basis of a document that ThP heuristics can identify effectively. Thus, these ThP configurations rely on the pattern of words and subphrases within the set of candidate phrases to extract quality thematic phrases.

The second group of ThP configurations considers phrase occurrence frequencies within the input document. That is, the candidate phrases are analyzed based on both the structure and semantics of the candidate phrases as well as the actual frequency with which each candidate phrase occurs in the document. The motivation for this setup of ThP configurations is to assess the importance of phrase occurrence frequencies, which consequently breakdown into word and subphrase occurrence frequencies, to bias ThP heuristics to extract quality thematic phrases. Both these groups of ThP configurations use identical set of parameters and their corresponding values listed in Tab. 5.1 across all experiments.

5.2.2 Topical N-Grams (TNG)

The Topical N-Grams topic model (TNG) that is adapted for the task of single document thematic phrase extraction requires two parameters to be specified. They are the number of topics to be generated and the number of top n-grams per topic that will be used to generate the consolidated list of thematic phrases. The values of each of these parameters are specified in Tab. 5.2 and is consistent with the number of topics and number of top words specified for the LDA modeling stage of the ThP configurations. The only distinction between the two is that of semantics in that LDA requires top words per topic while TNG requires top n-grams per topic to be specified.

Parameter	Value
Number of Topics	4
Top N-grams per Topic	10

Table 5.2: Topical N-Grams (TNG) Parameters for Experiments

5.2.3 AutoPhrase (AP)

The AutoPhrase (AP) method requires two parameters to be specified. The first parameter indicates whether part-of-speech (POS) tagging is enabled or disabled. Enabling POS tagging helps AP perform better and also makes it comparable with ThP configurations since they use LDA topics based nounphrases, extracted by spaCy using POS-based sentence parsing, as their initial candidate phrases. The second parameter is the minimum support threshold that individual phrases should meet for them to be considered candidates for topical phrases. The values of these parameters are specified in Tab. 5.3.

Parameter	Value
POS Tagging	Enabled
Minimum Support	2

Table 5.3: AutoPhrase (AP) Parameters for Experiments

AutoPhrase utilizes an external knowledge base consisting of quality words and phrases in the language and domain of interest to guide topical phrase extraction. The external knowledge base utilized for the experimentation in this work is the standard English language wiki-phrases corpus that is included with the AutoPhrase implementation utilized in this work (refer Sec. 4.2). AP outputs the final ranked list of extracted topical phrases in the file "final_quality_salient.txt". This file includes a consolidated ranked list of n-grams

that AP mines as a holistic representation of the topicality of the document.

5.2.4 Notation for Experiment Codes

The experiments conducted in this work are designated by abbreviated experiment codes that are used for brevity in all figures and tables in the chapters as well as appendices. The standard format followed for experiment codes is "*[Method_Code]-[k]*", where *k* indicates the top-*k* thematic phrases selected from the ranked list of thematic phrases extracted by the method indicated by the "*Method_Code*". The comprehensive list of all method codes is given in Tab. 5.4. For example, the experiment code "*ThP-123-5*" indicates that the top-5 thematic phrases extracted by the method "*ThP-123*" are being considered or discussed.

The prefix "*ThP*" indicates ThemaPhrase configurations that do not consider phrase occurrence frequencies within the input document and only rely on structural and semantic cues within the set of candidate phrases at every heuristic they utilize. The prefix "*ThP-Fr*" indicates ThemaPhrase configurations that consider phrase occurrence frequencies within the input document in addition to the structural and semantic cues within the set of candidate phrases at every heuristic. These ThP configuration categories are described in detail in Sec. 5.2.1 above.

Method Code	Method Description
AP	AutoPhrase
TNG	Topical N-Grams
ThP-123	ThemaPhrase Structural - WSEQ, WPOS and WASS active
ThP-13	ThemaPhrase Structural - WSEQ and WASS active
ThP-23	ThemaPhrase Structural - WPOS and WASS active
ThP-3	ThemaPhrase Structural - WASS active
ThP-Fr-123	ThemaPhrase Structural+Frequency - WSEQ, WPOS and WASS active
ThP-Fr-13	ThemaPhrase Structural+Frequency - WSEQ and WASS active
ThP-Fr-23	ThemaPhrase Structural+Frequency - WPOS and WASS active
ThP-Fr-3	ThemaPhrase Structural+Frequency - WASS active

Table 5.4: Experiment Method Codes

5.3 Examples of Extracted Thematic Phrases and Qualitative Discussion

This section considers two example documents, one from each dataset, and their respective top-10 thematic phrases extracted by three example methods: ThP-123, TNG and AP. The examples allow visual comparison between the thematic phrases extracted by the three methods relative to the two gold standards, namely the title and the abstract of each document. The discussion highlights nuances of thematic phrases and motivates the need for a diverse set of quantitative metrics to evaluate the thematic phrases extracted by the various methods.

Each example document and its extracted thematic phrases are visually presented in two separate figures, one for each gold standard. Fig. 5.1 and 5.2 show the thematic phrases extracted from a PubMed research publication and compares them with the publication's title and abstract respectively. Fig. 5.3 and 5.4 show the thematic phrases extracted from

a USPTO patent and compares them with the patent's title and abstract respectively. The following visual formatting is used in these figures:

(a) Words in a thematic phrase that occur in the gold standard are indicated by green text.

(b) An entire thematic phrase that occurs as a whole nounphrase in the gold standard is indicated by a green underline.

(c) An entire thematic phrase that occurs as a subphrase of a nounphrase in the gold standard is indicated by a blue underline.

(d) A subphrase of a thematic phrase that occurs as a whole nounphrase in the gold standard is also indicated by a blue underline.

(e) Words in a thematic phrase indicated by **orange text** are domain specific abbreviations of a nounphrase in the gold standard.

The thematic phrases extracted by ThP-123 and TNG for the PubMed publication are considerably shorter than those extracted by AP. Further, the methods extract certain thematic phrases such as *"usual inflation technique"* and *"accurate manometer measurements"* that appear generic relative to the document title in Fig. 5.1. But, these phrases are indeed thematic, as can be seen in Fig. 5.2 that visually represents them in the context of the document's abstract. Hence, the quantitative evaluation of thematic phrases is done using both the title and abstract that are gold standards of different granularity and, as such, allow the assessment of the granularity of thematic phrases extracted by the competing methods.

Fig. 5.1 and 5.2 show that certain thematic phrases occur as subphrases of nounphrases in the gold standard. An example is the thematic phrase “cuff pressures” detected by TNG that is a subphrase of “endotracheal tube cuff pressure assessment”, a nounphrase in the title (refer Fig. 5.1). Fig. 5.2 provides several more examples of such cases which are indicated by entire thematic phrases underlined in blue.

Fig. 5.2 also shows cases in which whole nounphrases from the gold standard occur as

TITLE: Endotracheal Tube Cuff Pressure Assessment: Education May Improve but not Guarantee the Safety of Palpation Technique

ThP-123-10	**TNG-10**
in vitro survey	anesthesia personnel
most common technique	tracheal model
usual inflation technique	**ett** cuff
pilot balloon palpation	intubated patients
validated manometer measurements	palpation technique
accurate manometer measurements	cuff pressures
different pressure levels	anesthesia staff
safe pressure limits	safe inflation
different pressure values	pilot balloon
actual **etcps**	safe pressure

AP-10

the safe inflation of **ett** cuff
balloon palpation with validated manometer measurements
fingers by in vitro pilot balloon
accurate measurement of **etcp** with manometer
years of experience in anesthesia
25 cm h 2 o
palpation with validated manometer measurements of endotracheal tube cuff pressure
fingers by in vitro pilot
30 cm h 2 o

FIG. 5.1: Thematic Phrases Example: PubMed Publication With Title as Gold Standard

ABSTRACT: Background:Endotracheal Tube Cuff Pressure (ETCP) should be kept in the range of 20 - 30 cm H2O. Earlier studies suggested that ETCP assessment by palpation of pilot balloon results in overinflation or underinflation and subsequent complications such as tracheal wall damage and aspiration.Objectives:The current study aimed to evaluate the effect of an in vitro educational program on the ability of anesthesia personnel to inflate Endotracheal Tube Cuffs (ETT) within safe pressure limits.Patients and Methods:The survey included two series of blinded ETCP measurements in intubated patients before and two weeks after an in vitro educational intervention. The in vitro educational program included two separate trials. The anesthesia personnel were asked to inflate an ETT cuff inserted in a tracheal model using their usual inflation technique. In the same session six ETTs at different pressure levels were examined by the participants and their estimation of ETCP was recorded. After the in vitro assessment the participants were informed about the actual pressure of the in vitro ETCPs and were allowed to train their fingers by in vitro pilot balloon palpation with validated manometer measurements.Results:The mean ETCP after the in vitro survey was significantly lower than the mean ETCP before the intervention (45 13 vs. 51 15 cm H2O, P = 0.002). The rate of measurements within the safe pressure limits significantly improved after the in vitro education (24.2% vs. 39.7%, P = 0.002).Conclusions:Implementing educational programs with the introduction of estimation techniques besides the use of manometer as a standard intraoperative monitoring will improve the safety of the practice.

ThP-123-10

in vitro survey
most common technique
usual inflation technique
pilot balloon palpation
validated manometer measurements
accurate manometer measurements
different pressure levels
safe pressure limits
different pressure values
actual etcps

TNG-10

anesthesia personnel
tracheal model
ett cuff
intubated patients
palpation technique
cuff pressures
anesthesia staff
safe inflation
pilot balloon
safe pressure

AP-10

the safe inflation of ett cuff
balloon palpation with validated manometer measurements
fingers by in vitro pilot balloon
accurate measurement of etcp with manometer
years of experience in anesthesia
25 cm h 2 o
palpation with validated manometer measurements
of endotracheal tube cuff pressure
fingers by in vitro pilot
30 cm h 2 o

FIG. 5.2: Thematic Phrases Example: PubMed Publication With Abstract as Gold Standard

subphrases of extracted thematic phrases. An example is the nounphrase “ett cuff” from the abstract that occurs as a subphrase in the first thematic phrase extracted by AP. Several more examples of this nature are seen for thematic phrases extracted by AP in Fig. 5.2 and are indicated by blue underlined parts of the thematic phrases. Such cases also occur, although in fewer number, in the case of the example patent seen in Fig. 5.3 and 5.4.

The considerable variations in thematic phrase lengths, as is the case with AP, as well as the nuances discussed above warrant consideration of metrics for quantitative analyses that will help compare thematic phrase sets with the two gold standards at phrase, subphrase and word granularity. Such metrics at varied granularity will allow assessment of which methods extract finer-grained thematic phrases versus more generic, coarser-grained ones. Further, the rank order of the thematic phrases extracted by each method also needs to be considered when assessing the quality of thematic phrases extracted.

Another nuance that needs highlighting is domain specific abbreviations that are used frequently throughout a document but do not occur in the gold standards. Examples of this are observed in the case of the PubMed publication. In Fig. 5.1, the words represented as orange text are abbreviations that are frequently used in the body of the publication but occur in the gold standards in their expanded form. For example, “etcp” is the abbreviation for “endotracheal tube cuff pressure”. And the phrase “ett cuff” is a semi-abbreviated form of the phrase “endotracheal tube cuff”. Fig. 5.1 shows that all three thematic phrase extraction methods extract phrases that contain such abbreviations and these phrases will not contribute to any quantitative metric that compares them with the title as the gold standard.

Lastly, Fig. 5.3 and 5.4 show that a low percentage of words that form thematic phrases and the thematic phrases themselves occur in the title and abstract for the patent document. This is because patents are typically more verbose and contain repetitive boilerplate verbiage. The verbosity and presence of redundant boilerplate language results in frequently occurring finer-grained phrases in the document to be extracted as thematic phrases. Both the figures show that all three methods are affected by this. Hence, the absolute values of quantitative

TITLE: Lighting system for medical appointment progress tracking by wireless detection

ThP-123-10	**AP-10**	**TNG-10**
different virtual networks	ip address	exam room
different physical networks	light bulb	lighting system
present embodiment	ip multicast	user device
present embodiment database	light bulbs	wireless signal
example lighting element	touch screen	fi transceiver
filament lighting element	lighting structures	medical office
mobile device router	signal strength	network signal
mobile device	remote processor	lighting element
mobile device requests	opposite walls	fi extender
long wait times	powerline ethernet adapter	location sensor

FIG. 5.3: Thematic Phrases Example: USPTO Patent With Title as Gold Standard

ABSTRACT: Provided are mechanisms and processes for a lighting system for medical schedule management. According to various examples, an apparatus is provided which comprises a lighting interface configured to connect to a lighting element for illuminating a medical examination room. The apparatus further comprises a power interface coupled to a power source. The apparatus further comprises a Wi-Fi transceiver configured to transmit a wireless signal to connect to a device corresponding to a physician. The wireless signal corresponds to a local area network. The duration of the connection is used to track the presence of the physician in the medical examination room. The Wi-Fi transceiver is tuned to transmit a signal strength corresponding to the size and characteristics of the medical examination room. The apparatus is located in a lighting fixture in the medical examination room. The lighting fixture may be centrally located in the medical examination room.

ThP-123-10	AP-10	TNG-10
different virtual networks	ip address	exam room
different physical networks	light bulb	lighting system
present embodiment	ip multicast	user device
present embodiment database	light bulbs	wireless signal
example lighting element	touch screen	fi transceiver
filament lighting element	lighting structures	medical office
mobile device router	signal strength	network signal
mobile device	remote processor	lighting element
mobile device requests	opposite walls	fi extender
long wait times	powerline ethernet adapter	location sensor

FIG. 5.4: Thematic Phrases Example: USPTO Patent With Abstract as Gold Standard

metrics for the USPTO dataset are expected to be lower than those for the PubMed dataset. This will be evident throughout the quantitative evaluation as well as from the tabular results in appendices for all quantitative metrics across experiments.

The next section describes different quantitative metrics that are used to evaluate the quality of thematic phrases extracted by various methods and their rank order. Each metric is considered based on the discussion of various facets and nuances of the quality of thematic phrases discussed above.

5.4 Quantitative Evaluation Metrics

The thematic phrases extracted by various methods are quantitatively evaluated using several metrics that help assess the thematic phrases in terms of their utility for different usecases and downstream text analytics tasks. These metrics provide different perspectives into the quality of thematic phrases themselves as well as their rank order. Some metrics are computed at the phrase, partial phrase and word granularities while others are computed at only some of those granularities. This is stated clearly when defining and discussing each metric in the following subsections.

The titles and abstracts of the text documents are used as the gold standard when computing the metrics. Both gold standards are generally denoted by 'GS' in all notations hereafter. The following definitions and notations will be used for all the metrics defined in this section:

***GS*:** the gold standard $\in$ { *title, abstract* } used as the reference point for comparison

$\langle\ w_j^{GS}\ \rangle$**:** the sequence of words that form the gold standard

$\langle\ ph_j^{GS}\ \rangle$**:** the sequence of nounphrases that form the gold standard

$\boldsymbol{ph^{GS}}$**:** the set of nounphrases contained in the gold standard

$\boldsymbol{W^{GS}}$**:** the set of words contained in the gold standard

$\boldsymbol{ph^{tD}}$**:** the set of thematic phrases extracted from the document

$\boldsymbol{W^{tD}}$**:** the set of words that form the thematic phrases extracted from the document

Further, each metric is defined at one or more granularities and the granularities are indicated by attaching the following prefixes to the metric names:

***ph*:** Indicates the metric is computed at the phrase granularity. This means the metric looks for exact matches between the thematic phrases and nounphrases in the gold standard.

***wd*:** Indicates the metric is computed at the word granularity. This means the metric looks for matches between the words that form the thematic phrases and words that form the nounphrases in the gold standard.

***ext*:** Indicates the metric is computed at the extended phrase granularity. This means the metric looks for exact matches between the entire thematic phrases or their subsequences and the nounphrases in the gold standard. Performance on this metric

implies the thematic phrases are either exact matches to nounphrases in the gold standard or are extensions of nounphrases in the gold standard, i.e. they are effectively finer-grained than the nounphrases in the gold standard.

***sub*:** Indicates the metric is computed at the subphrase granularity. This means the metric looks for exact matches between entire thematic phrases and entire nounphrases in the gold standard or their sub-sequences. Performance on this metric implies the thematic phrases are either exact matches to nounphrases in the gold standard or are subphrases of nounphrases in the gold standard, i.e. they are effectively coarser-grained than the nounphrases in the gold standard.

The subsections that follow define and describe each quantitative metric used for quantitative evaluation of thematic phrases in detail along with its motivation and utility.

5.4.1 Coverage

Coverage measures the amount of overlap between the gold standard and the thematic phrases. This metric takes into consideration the entire sequence of nounphrases (or words) that forms the gold standard as opposed to the set of the nounphrases (or words) without repetition. The intuition is that frequent nounphrases and words in the author generated gold standards are indicative of their relatively greater relevance to the thematic basis of the

document. The coverage at different granularities is calculated as follows:

$$\textit{ph-COV} = \frac{\left|\left\langle\, ph_j^{GS} \in \boldsymbol{ph^{tD}} \,\right\rangle\right|}{\left|\left\langle\, ph_j^{GS} \,\right\rangle\right|} \qquad \textit{wd-COV} = \frac{\left|\left\langle\, w_j^{GS} \in \boldsymbol{W^{tD}} \,\right\rangle\right|}{\left|\left\langle\, w_j^{GS} \,\right\rangle\right|}$$

$$\textit{ext-COV} = \frac{\left|\left\langle\, ph_j^{GS} : \exists ph_i^{tD}, ph_j^{GS} \in ph_i^{tD} \,\right\rangle\right|}{\left|\left\langle\, ph_j^{GS} \,\right\rangle\right|}$$

$$\textit{sub-COV} = \frac{\left|\left\langle\, ph_j^{GS} : \exists ph_i^{tD}, ph_i^{tD} \in ph_j^{GS} \,\right\rangle\right|}{\left|\left\langle\, ph_j^{GS} \,\right\rangle\right|}$$

5.4.2 Recall

Recall measures the amount of overlap between the set of phrases (or words) in the gold standard and the thematic phrases. The difference between coverage and recall is that recall uses the set of phrases (or words) as the comparison point i.e. the phrases (or words) do not repeat in the comparison group.

For example, consider the gold standard sentence *"This metric takes into consideration the entire sequence of nounphrases that forms the gold standard and not the set of the nounphrases without repetition"*. For word granularity coverage, the gold standard comparison point is the entire sequence of words (i.e. with word repetitions) that composes the sentence. Whereas, for word granularity recall, the gold standard comparison point is the

set of words (i.e. without word repetition) that compose the sentence as shown below:

$$\left\{ \begin{array}{l} \textit{this, metric, takes, into, consideration, the, entire, sequence, of, noun-} \\ \textit{phrases, that, forms, gold, standard, and, not, set, without, repetition} \end{array} \right\}$$

The intuition here is that the set of nounphrases and words in the author generated gold standards indicate the correct tags that manifest the thematic basis of the document without consideration of their relative importance indicated by their frequencies of occurrence. The recall at different granularities is calculated as follows:

$$\textit{ph-REC} = \frac{\left|\boldsymbol{ph^{GS}} \cap \boldsymbol{ph^{tD}}\right|}{\left|\boldsymbol{ph^{GS}}\right|} \qquad \textit{wd-REC} = \frac{\left|\boldsymbol{W^{GS}} \cap \boldsymbol{W^{tD}}\right|}{\left|\boldsymbol{W^{GS}}\right|}$$

$$\textit{ext-REC} = \frac{\left|\left\{\ ph_j^{GS} : \exists ph_i^{tD}, ph_j^{GS} \in ph_i^{tD}\ \right\}\right|}{\left|\boldsymbol{ph^{GS}}\right|}$$

$$\textit{sub-REC} = \frac{\left|\left\{\ ph_j^{GS} : \exists ph_i^{tD}, ph_i^{tD} \in ph_j^{GS}\ \right\}\right|}{\left|\boldsymbol{ph^{GS}}\right|}$$

5.4.3 Precision

Precision measures the proportion of the thematic phrases that overlap with the set of phrases (or words) in the gold standard. Precision also uses the set of phrases (or words) as the comparison point i.e. the phrases (or words) do not repeat in the comparison

group. Precision measures how accurately the thematic phrases represent the topicality of the document indicated by the gold standard. The precision at different granularities is calculated as follows:

$$\textit{ph-PRE} = \frac{\left|\boldsymbol{ph^{GS}} \cap \boldsymbol{ph^{tD}}\right|}{\left|\boldsymbol{ph^{tD}}\right|} \qquad \textit{wd-PRE} = \frac{\left|\boldsymbol{W^{GS}} \cap \boldsymbol{W^{tD}}\right|}{\left|\boldsymbol{W^{tD}}\right|}$$

$$\textit{ext-PRE} = \frac{\left|\left\{\ ph_j^{GS} : \exists ph_i^{tD}, ph_j^{GS} \in ph_i^{tD}\ \right\}\right|}{\left|\boldsymbol{ph^{tD}}\right|}$$

$$\textit{sub-PRE} = \frac{\left|\left\{\ ph_j^{GS} : \exists ph_i^{tD}, ph_i^{tD} \in ph_j^{GS}\ \right\}\right|}{\left|\boldsymbol{ph^{tD}}\right|}$$

5.4.4 Fowlkes-Mallows Index (FMI)

The Fowlkes-Mallows Index is the geometric mean of the precision and recall and is usually used to measure the similarity between two clusters or the similarity of a cluster and ground-truth classification. FMI is commonly used for unsupervised tasks while F-Score (harmonic mean) is commonly used for supervised tasks. In the case of thematic phrases, the FMI is computed at the phrase, partial phrase and word granularities. It provides a composite metric that combines the precision and recall of a thematic phrase set relative to the gold standard at the respective granularities. The FMI at the different granularities is

calculated as follows :

$$\textit{ph-FMI} = \sqrt{\textit{ph-PRE} * \textit{ph-REC}} \qquad \textit{ext-FMI} = \sqrt{\textit{ext-PRE} * \textit{ext-REC}}$$

$$\textit{wd-FMI} = \sqrt{\textit{wd-PRE} * \textit{wd-REC}} \qquad \textit{sub-FMI} = \sqrt{\textit{sub-PRE} * \textit{sub-REC}}$$

Subsequent sections that discuss quantitative performance of the different thematic phrase extraction methods use FMI as a composite recall-precision measure to compare the methods.

5.4.5 Jaccard Similarity

The Jaccard Similarity is a similarity measure between two sets and is calculated as the ratio of the intersection over the union of the two sets. For the purposes of measuring the quality of thematic phrases, this similarity metric is computed for the sets of words that compose the gold standard and the thematic phrases set. The metric does not consider the frequency of occurrence of the words in either sets. It is calculated as follows:

$$\textit{wd-JCI} = \frac{\left|\left\{\ w_i^{GS}\ \right\} \cap \boldsymbol{W^{tD}}\right|}{\left|\left\{\ w_i^{GS}\ \right\} \cup \boldsymbol{W^{tD}}\right|}$$

The Jaccard Similarity is also known as the Critical Success Index (CSI) in the context of classification tasks where it is defined as

$$CSI = \frac{TP}{TP + FP + FN}$$

5.4.6 Cosine Similarity

The Cosine Similarity is a similarity measure between two token lists and is calculated as the cosine of the angle between vectors representing the two lists. The vectorization of the two lists is domain and task dependent. For the purposes of measuring the quality of thematic phrases, the thematic phrases set and the gold standard are count vectorized with their words as the dictionary. Further, the vectors are normalized using their respective L-2 norms to avoid unusually long thematic phrases causing bias that benefits the corresponding method. It is calculated as follows:

$$COS = \frac{A \,.\, B}{||A|| * ||B||}$$

5.4.7 Discounted Cumulative Gain (DCG)

This metric assesses the quality of the rank order of thematic phrases generated by various methods. This metric is particularly important when considering the thematic phrases for tasks such as automated document tagging or for the use of the thematic phrases as an index for retrieval of these documents either independently or as an augmentation on

top of other full text indexes. Both these applications benefit from a ranked list of thematic phrases ordered by relevance or importance. DCG is assessed using abstracts as the only gold standard since titles are short and contain only a few words and phrases that are the most thematically relevant.

DCG_p is the discounted cumulative gain at rank p and considers all members of the rank list from rank 1 through p. It is defined as

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

The relevance score rel_i of the i^{th} member of the ranked list is discounted by the inverse logarithm of its rank. Thus, members ranked lower contribute less to the cumulative gain than those ranked higher. The relevance score for a thematic phrase is calculated in three different ways to assess the quality of thematic phrases:

(a) **Phrase Hits (*phHIT*):** This relevance measure takes into account the intersection of the thematic phrase set with the nounphrase set in the gold standard. That is, it looks for exact matches of thematic phrases with nounphrases in the gold standard. It is defined as

$$rel_i = phHIT_i = \begin{cases} 1 & \text{if } ph_i^{tD} \in GS \\ 0 & \text{otherwise} \end{cases}$$

(b) **Word Hits (*wdHIT*):** This relevance measure takes into account the intersection of words that form thematic phrases with words that form the nounphrases in the gold standard. That is, it looks for phrases that are be formed by thematically relevant words. It is defined as

$$rel_i = wdHIT_i = \frac{\left|ph_i^{tD} \cap \boldsymbol{W^{GS}}\right|}{\left|ph_i^{tD}\right|}$$

(c) **Word Coverage (*wdCOV*):** This relevance measure augments *wdHIT* by also taking into account the frequency with which words that form thematic phrases occur in the gold standard. That is, it scores phrases that are composed of more frequently occurring thematic words from the gold standard higher than those composed of less frequently occurring words. It is defined as

$$rel_i = wdCOV_i = \frac{\left|\left\langle\ w_j^{GS} \in ph_i^{tD}\ \right\rangle\right|}{\left|ph_i^{tD}\right|}$$

The *wdHIT* and *wdCOV* relevance measure for each thematic phrase is normalized using its phrase length $\left|ph_i^{tD}\right|$ to avoid bias in favor of unduly long phrases.

5.5 Quantitative Analysis Discussion

This section discusses the quantitative evaluation of the thematic phrases extracted by AP, TNG and various configurations of ThP using the metrics described in Sec. 5.4.

A brief summary of key findings from the quantitative analyses is provided in Sec. 5.5.1 . The subsequent subsections discuss the quality of thematic phrases using one or more quantitative metrics as follows:

(a) Sec. 5.5.2 discusses thematic phrase quality based on COV and FMI metrics at different granularities using document abstracts as the gold standard

(b) Sec. 5.5.3 discusses thematic phrase quality based on COV and FMI metrics at different granularities using document titles as the gold standard

(c) Sec. 5.5.4 discusses thematic phrase quality based on DCG metrics at different granularities using document abstracts as the gold standard

(d) Sec. 5.5.5 discusses thematic phrases variance as a measure of the robustness of thematic phrase extraction method to varying segment counts

Each subsection contains individual plots of each of the quantitative metrics that the subsection discusses as well as consolidated plots of those metrics. The former allow for detailed visual exploration of the relative performance of each method on the corresponding metric. The latter plots allow for visual comparison of the performance of a few methods of interest on all the metrics being discussed in that subsection. The complete listing of all the values for the quantitative metrics across experiments and gold standards is provided in tabular form in Appendices B to J.

5.5.1 Summary of Results

The performance of TNG across all metrics discussed in the following subsections improves as the segment count increases and it begins to plateau after segment count =15. The performance of all other methods across all the metrics varies relatively lower, if at all, as segment count increases. Hence, when comparing the various thematic phrase extraction methods, conclusions are drawn based on their performance on all the metrics at segment count =25 since TNG's performance is best at that segment count.

The results are summarized based on the different methods' collective performance on COV and FMI metrics at different thematic granularities and k values for both datasets. The thematic granularity is based on : (a) the evaluation gold standards: title (coarse-grained, broad thematic representation) and abstracts (coarse- and fine-grained thematic representation) (b) granularity of the evaluation metrics: ph-*, sub-*, ext-* and wd-* The following are the key findings about the competing thematic phrase extraction methods. They are stated in terms of the granularity at which document themes are covered and represented by the extracted thematic phrases.

(1) **Thematic Representation at the Granularity of Titles and Abstracts:** ThP-Fr-13 and ThP-Fr-3 are the best methods to extract thematic phrases at these granularities for both datasets at k=20. TNG is the best for lower values of k. The detailed discussion on ph-COV and ph-FMI with abstracts as the gold standard is provided in Sec. 5.5.2 and with titles as the gold standard is provided in Sec. 5.5.3.

(2) **Thematic Representation That is Coarser-grained Than Titles and Abstracts:** ThP-Fr-123 and ThP-Fr-23 are the best methods to extract thematic phrases that represent thematic treatment coarser-grained than abstracts and titles for the USPTO dataset. TNG is the best method for the PubMed dataset, especially when recall/coverage is more important. ThP-Fr-123 and ThP-Fr-23 are good for the PubMed dataset at higher k values and when higher FMI is desired in addition to coverage i.e precision is also important to the usecase. The detailed discussion on sub-COV and sub-FMI with abstracts as the gold standard is provided in Sec. 5.5.2 and with titles as the gold standard is provided in Sec. 5.5.3.

(3) **Thematic Representation That is Finer-grained Than Abstracts:** ThP-Fr-13 and ThP-Fr-3 are the best methods to extract thematic phrases that represent thematic treatment finer-grained than abstract nounphrases for the USPTO dataset at $k \in \{10, 20\}$ while TNG is the best method for the PubMed dataset. The detailed discussion on ext-COV and ext-FMI for abstracts is provided in Sec. 5.5.2.

(4) **Thematic Representation That is Finer-grained Than Titles:** TNG is the best method to extract thematic phrases that are finer-grained than the themes represented in document titles across k values for both datasets. The detailed discussion on ext-COV and ext-FMI for titles is provided in Sec. 5.5.3.

(5) **Thematic Phrases Formed by Abstract and Title Words:** TNG thematic phrases have better thematic word recall-precision balance at $k \in \{5, 10\}$. ThP-Fr-13 and ThP-

Fr-3 are comparable to TNG at k=20 for the USPTO dataset. The detailed discussion on wd-COV and wd-FMI of abstracts is provided in Sec. 5.5.2 and of titles is provided in Sec. 5.5.3. The differences in these methods lies mainly in the phrases they extract that are formed by the thematic words. This is evident from the previous results highlighted in (1), (2), (3) and (4).

(6) **Robustness to Topic-Segment Count Ratio:** ThP-Fr-13 and ThP-Fr-3 are the most robust to topics-to-segments ratio as is indicated by their low thematic phrase variance when the segment counts vary. The detailed discussion on effects of segment counts is provided in Sec. 5.5.5.

(7) **Thematic Phrases Ranking:** Ranked list of thematic phrases extracted by TNG have the best DCG performance while ThP configurations consistently lag behind it. Improvements to ThP's phrase scoring is left for future work. The detailed discussion on DCG performance is provided in Sec. 5.5.4.

5.5.2 Quantitative Analyses of Thematic Phrases With Document Abstracts as the Gold Standard

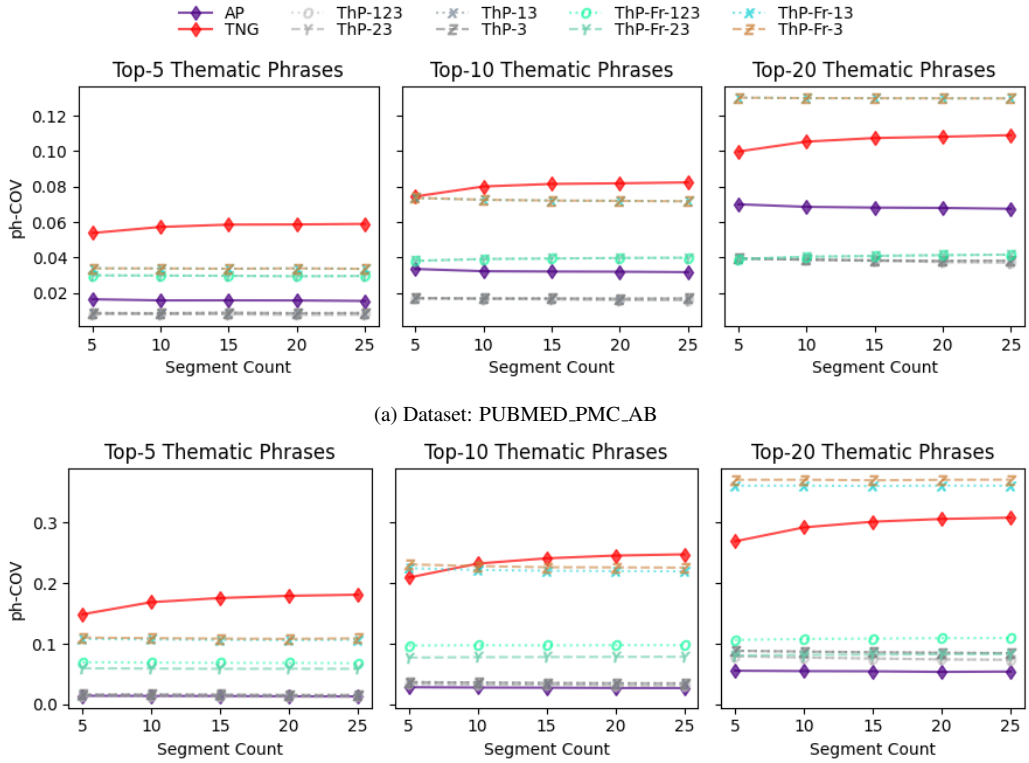
The first gold standard used for evaluating the thematic phrase extraction methods is teh set of abstracts of the documents. Abstracts are representative of their corresponding documents thematic basis that is manifested in the words and phrases that occur in the abstracts. They cover the core themes of the document at a coarse granularity and may

contain finer-grained treatment of one or more concepts that are relevant to the themes and that are addressed in the document.

This subsection discusses the quality of thematic phrases extracted by the thematic phrase extraction methods relative to document abstracts using the COV, FMI, JCI and COS metrics at different granularities. The discussion is divided into three parts under separate headers as follows: analyses of COV at phrase and partial phrase granularities; analyses of FMI at phrase and partial phrase granularities; and analyses of COV, FMI, JCI and COS at word granularity. The discussion under each header will state observations about the performance of the methods on the corresponding metrics followed by key conclusions that are drawn based on those observations. The critical statistical significance level used to report conclusions in this section is α_{C1}=$1.39E{-}06$. Refer Appendix A for details on the Bonferroni corrected α_{C1} critical significance level.

Coverage at Phrase and Partial Phrase Granularities: Coverage at phrase and partial phrase granularities are measured using the ph-COV, sub-COV and ext-COV metrics. These are plotted relative to abstracts as the gold standard in Fig. 5.5, 5.6 and 5.7 respectively. Coverage metrics at all three granularities improve by varying degrees for all methods across both datasets as k increases.

TNG is the top performer on ph-COV at k=5 but is outperformed at k=20 for both datasets. ThP-Fr-13 and ThP-Fr-3 have comparable ph-COV at all k values for the PubMed dataset. At k=20, ThP-Fr-13 and ThP-Fr-3 outperform all other methods on ph-COV for



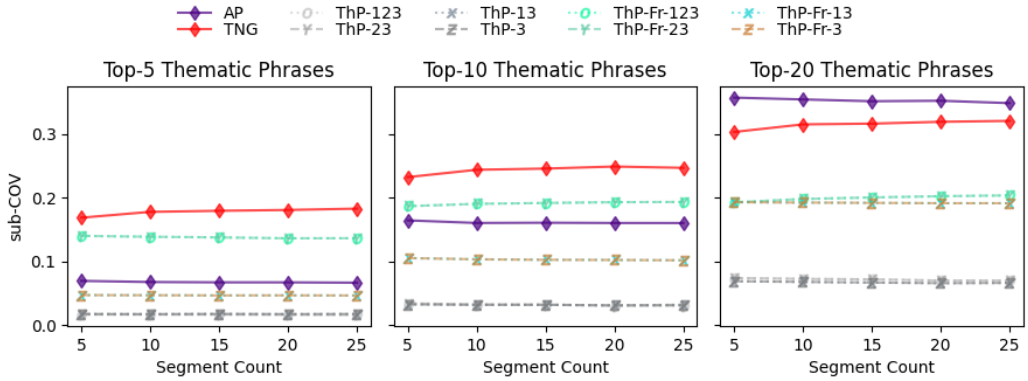
(a) Dataset: PUBMED_PMC_AB

(b) Dataset: PATENT_RAND15K

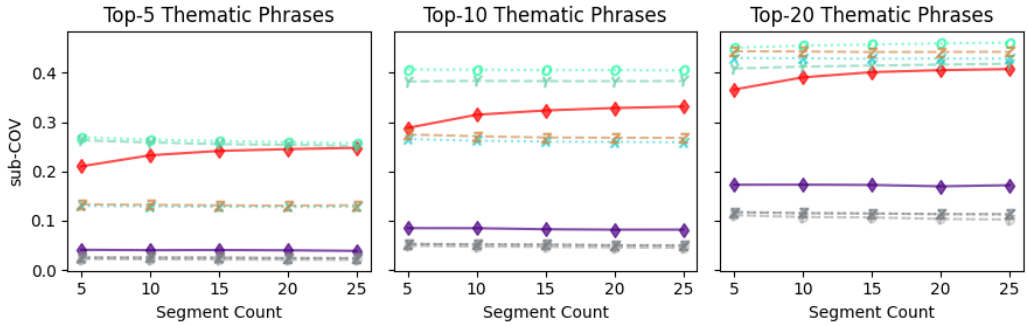
FIG. 5.5: Thematic Phrases ph-COV Comparison With Abstracts as Gold Standard

both datasets ($P{\leq}\alpha_{C1}$). Further, the latter outperforms the former for the USPTO dataset ($P{\leq}\alpha_{C1}$). The ThP-* configurations that do not consider phrase occurrence frequencies have the lowest ph-COV of all the ThP configurations.

TNG is the top performer on sub-COV at k=5 but is outperformed by AP at k=20 for the PubMed dataset. ThP-Fr-123 and ThP-Fr-23 have comparable sub-COV on the PubMed dataset and outperform all other ThP configurations at k=20 ($P{\leq}\alpha_{C1}$). ThP-Fr-123, ThP-Fr-23, ThP-Fr-13 and ThP-Fr-3 outperform TNG and AP on sub-COV at k=20 for the USPTO dataset with ThP-Fr-123 obtaining the highest sub-COV ($P{\leq}\alpha_{C1}$). The ThP-*



(a) Dataset: PUBMED_PMC_AB



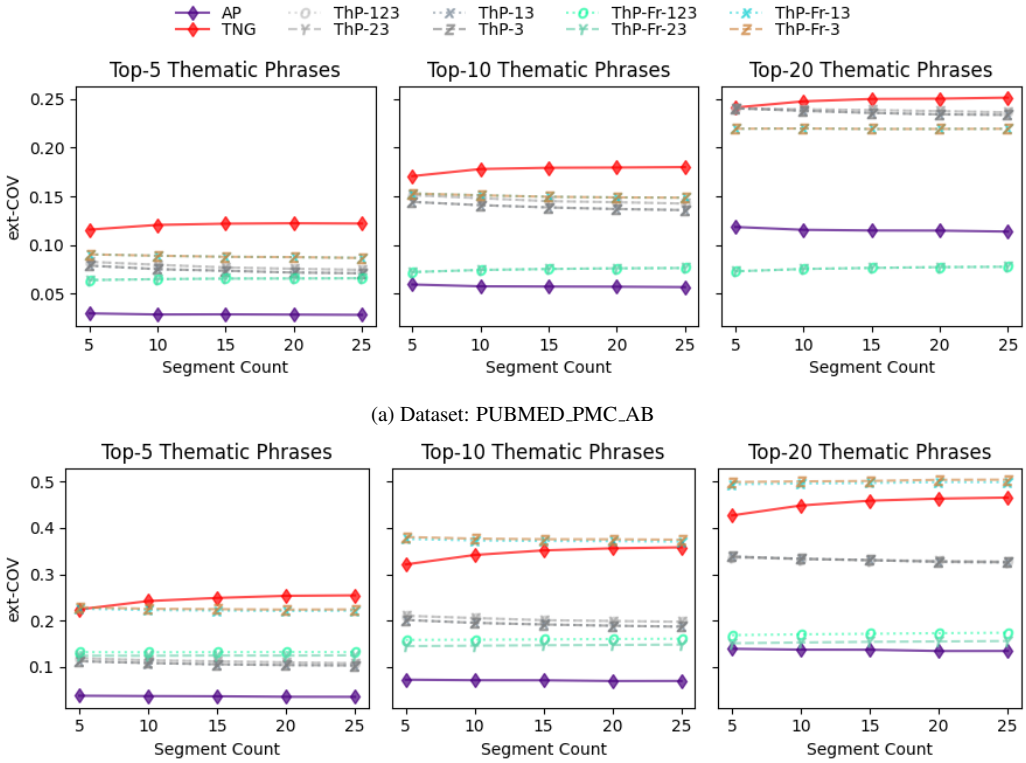
(b) Dataset: PATENT_RAND15K

FIG. 5.6: Thematic Phrases sub-COV Comparison With Abstracts as Gold Standard

configurations that do not consider phrase occurrence frequencies underperform all other methods on sub-COV for both datasets.

TNG outperforms all methods on ext-COV at all k values for the PubMed dataset. AP underperforms most methods on ext-COV for the PubMed dataset and all methods for the USPTO dataset. ThP-* configurations outperform ThP-Fr-* configurations on ext-COV for the PubMed dataset ($P \leq \alpha_{C1}$). ThP-Fr-13 and ThP-Fr-3 outperform all other methods on ext-COV at k=20 for the USPTO dataset ($P \leq \alpha_{C1}$).

We can conclude from the observations on coverage metrics that ThP-Fr-13 and ThP-



(a) Dataset: PUBMED_PMC_AB

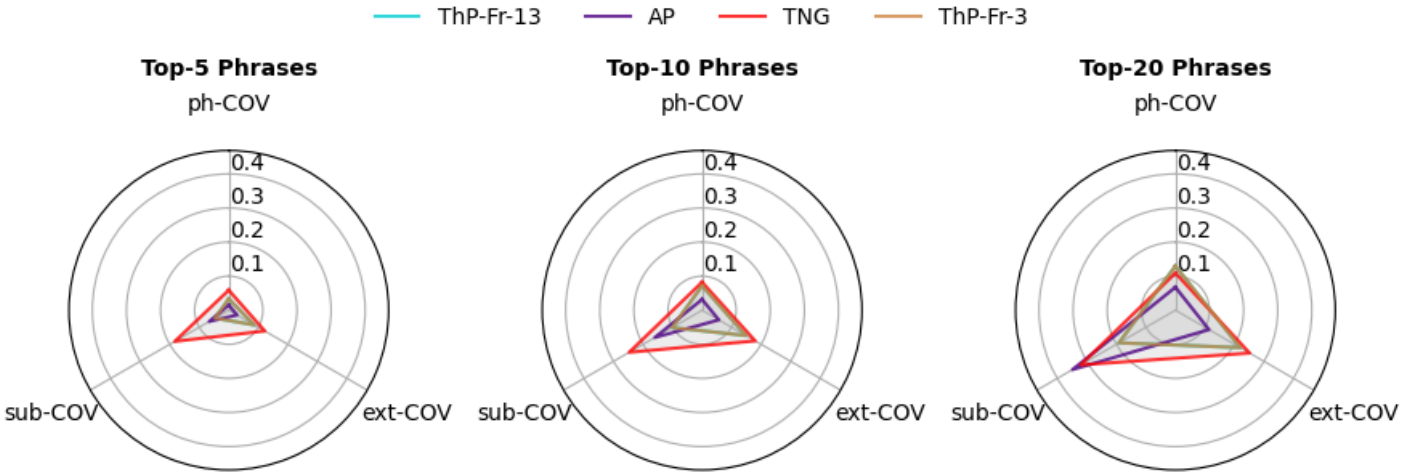
(b) Dataset: PATENT_RAND15K

FIG. 5.7: Thematic Phrases ext-COV Comparison With Abstracts as Gold Standard

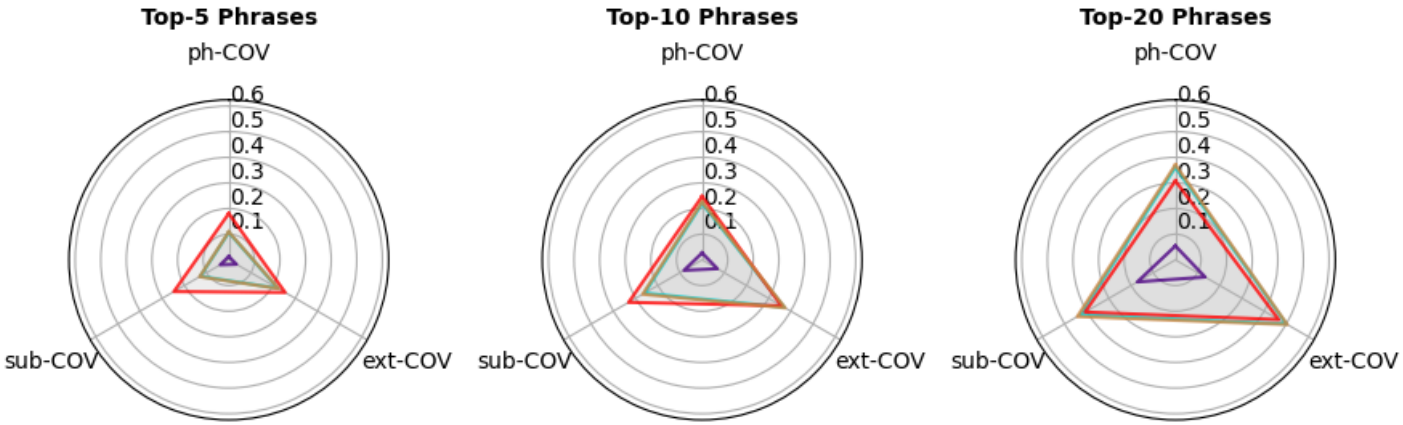
Fr-3 are better at extracting thematic phrases at the granularity at which the themes are represented in document abstracts. AP is better at extracting thematic phrases that are at a coarser granularity than that at which themes are represented in document abstracts for the PubMed dataset whereas ThP-Fr-123 is better for the USPTO dataset. TNG is better at extracting thematic phrases that are at a finer granularity than that at which themes are represented in document abstracts for the PubMed dataset whereas ThP-Fr-13 and ThP-Fr-3 are better for the USPTO dataset. One or more ThP configurations are most averse to effects of high average word occurrence frequencies in documents present in the USPTO dataset as

is reflected by their consistent outperformance of other methods on all coverage metrics for the USPTO dataset.

Fig. 5.8 shows consolidated radar plots that allows for visual comparison of the performance of AP, TNG, ThP-Fr-13 and ThP-Fr-3 on coverage metrics at the three granularities discussed above with the abstract as the gold standard.



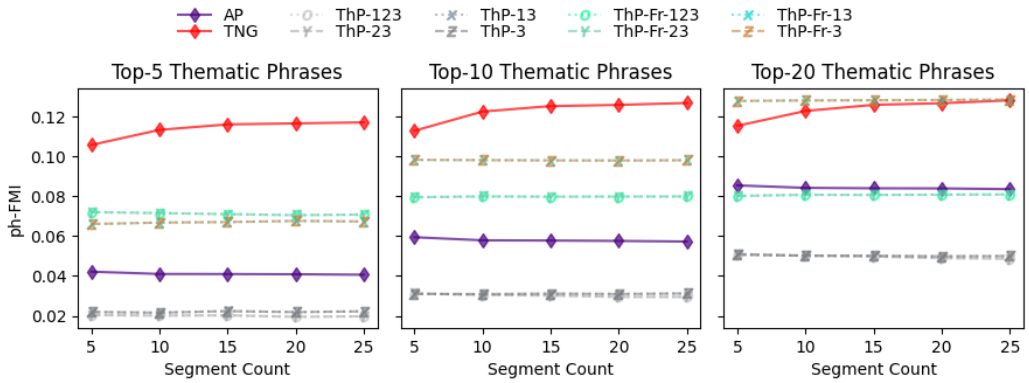
(a) Dataset: PUBMED_PMC_AB



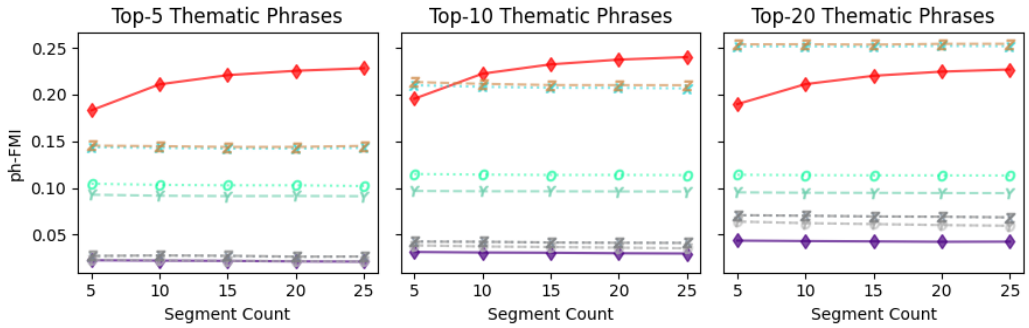
(b) Dataset: PATENT_RAND15K

FIG. 5.8: Consolidated Radar Plot: Thematic Phrases ph-COV, sub-COV and ext-COV With Abstracts as Gold Standard

FMI at Phrase and Partial Phrase Granularities: Fig. 5.9, 5.10 and 5.11 show plots for ph-FMI, sub-FMI and ext-FMI metrics respectively. TNG has the highest ph-



(a) Dataset: PUBMED_PMC_AB



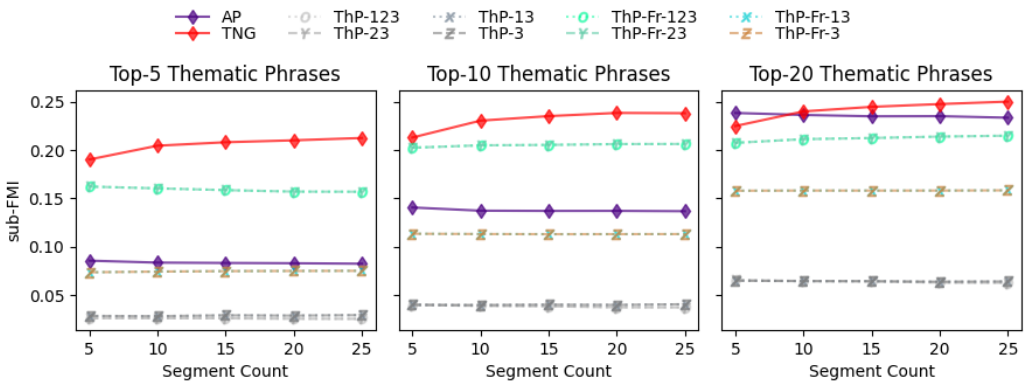
(b) Dataset: PATENT_RAND15K

FIG. 5.9: Thematic Phrases ph-FMI Comparison With Abstracts as Gold Standard

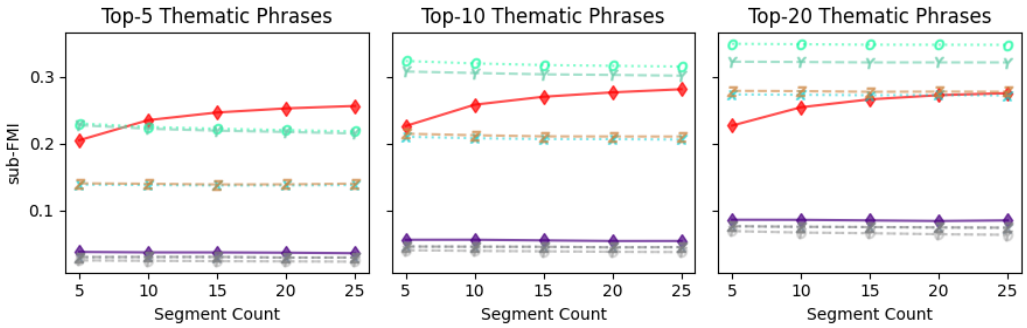
FMI at $k \in \{5, 10\}$ for both datasets. At k=20, ThP-Fr-13 and ThP-Fr-3 have comparable ph-FMI with TNG for the PubMed dataset and they outperform TNG for the USPTO dataset ($P \leq \alpha_{C1}$). Further, the increase in ph-FMI of TNG as k increases is relatively lower than ThP-Fr-13 and ThP-Fr3.

TNG outperforms all other methods on sub-FMI at all values of k for the PubMed dataset. TNG also outperforms all other methods on sub-FMI at k=5 for the USPTO dataset.

ThP-Fr-123 and ThP-Fr-23 outperform on sub-FMI at $k\in\{10, 20\}$ for the USPTO dataset ($P\leq\alpha_{C1}$). ThP-Fr-123 has the highest sub-FMI of the latter two at $k=20$ for the USPTO dataset ($P\leq\alpha_{C1}$). ThP-Fr-13 and ThP-Fr-3 have comparable sub-FMI across k values on both datasets and both are comparable to TNG at $k=20$ for the USPTO dataset.



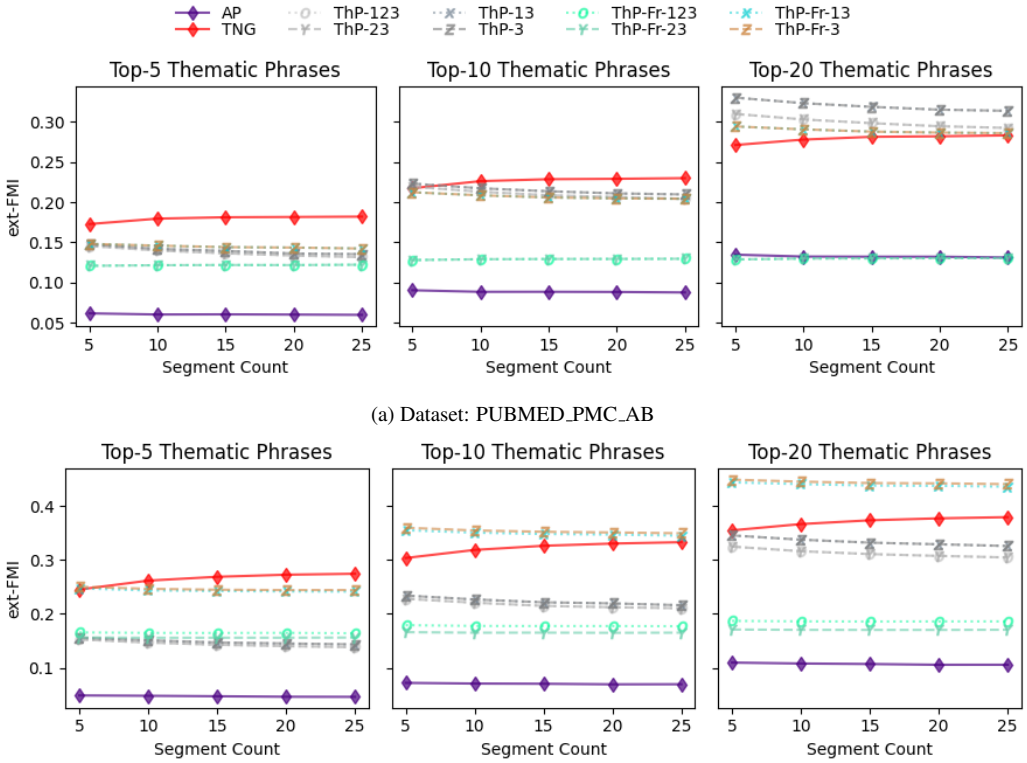
(a) Dataset: PUBMED_PMC_AB



(b) Dataset: PATENT_RAND15K

FIG. 5.10: Thematic Phrases sub-FMI Comparison With Abstracts as Gold Standard

TNG has the highest ext-FMI at $k=5$ for both datasets. At $k=20$, ThP-13 and ThP-3 have comparable ext-FMI and outperform all other methods for the PubMed dataset ($P\leq\alpha_{C1}$). ThP-123 and ThP-23 also outperform TNG on ext-FMI for the PubMed dataset ($P\leq\alpha_{C1}$) while ThP-Fr-13 and ThP-Fr-3 have ext-FMI comparable to TNG. In the case of



(a) Dataset: PUBMED_PMC_AB

(b) Dataset: PATENT_RAND15K

FIG. 5.11: Thematic Phrases ext-FMI Comparison With Abstracts as Gold Standard

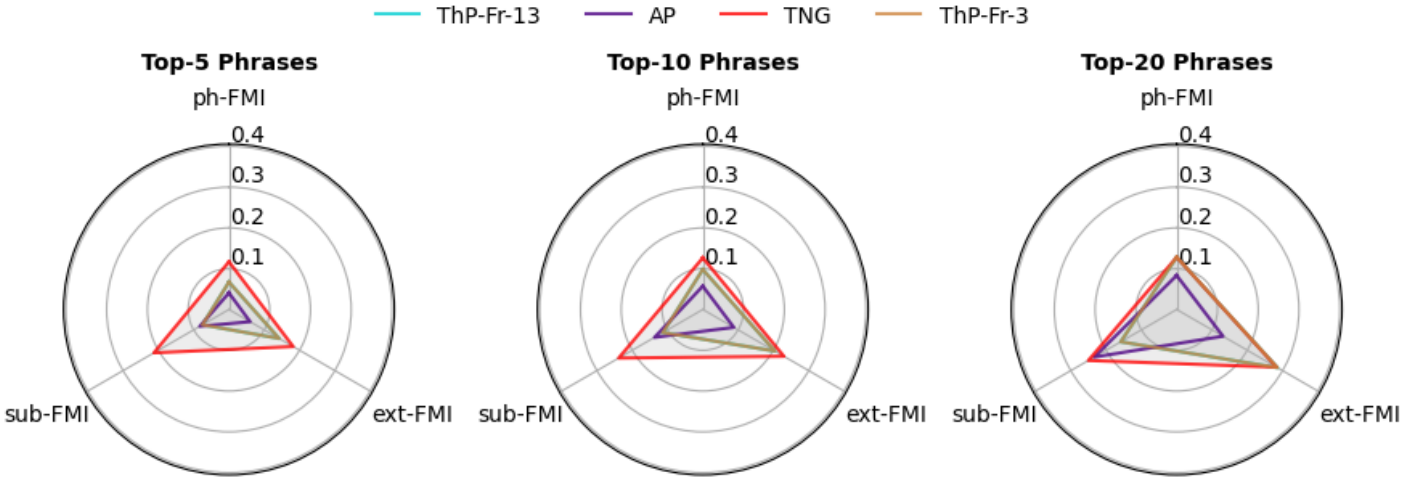
the USPTO dataset, ThP-Fr-13 and ThP-Fr-3 outperform TNG at k=20 ($P \leq \alpha_{C1}$) while the ThP-* configurations have lower ext-FMI than TNG. AP underperforms on ext-FMI across k values for both datasets.

We can draw the following conclusions about the thematic phrase extraction methods from observations of their respective performances on coverage and FMI metrics collectively. TNG is better at extracting thematic phrases at the granularity at which themes are represented in document abstracts at lower k values whereas ThP-Fr-13 and ThP-Fr-3 are better for higher k values.

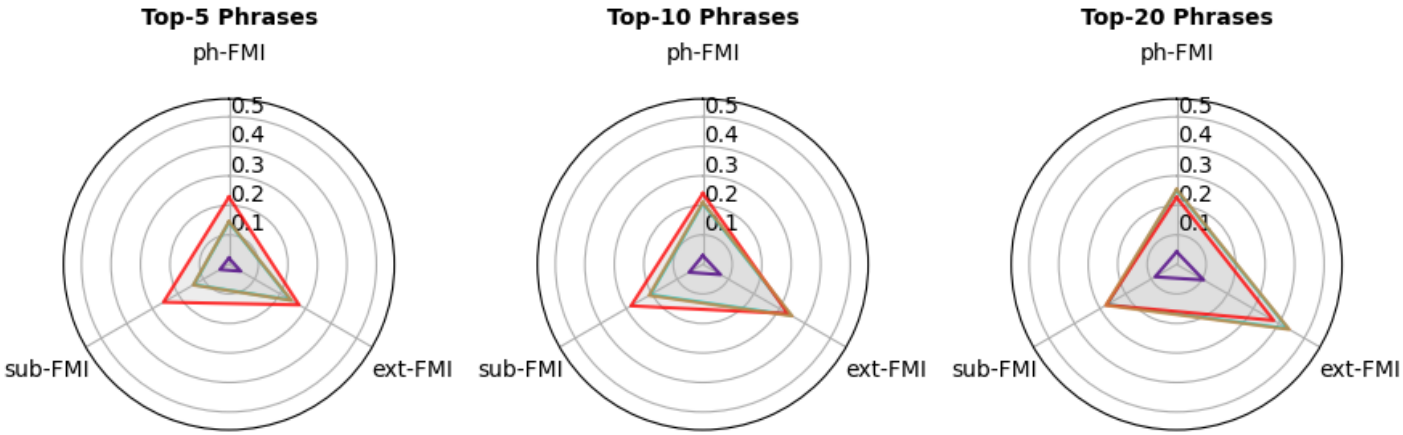
TNG is better at extracting thematic phrases that are at a coarser granularity than that at which themes are represented in document abstracts for datasets like PubMed that contain documents with relatively lower average word occurrence frequencies. ThP-Fr-123 is the better choice for datasets like USPTO that contain documents with relatively higher average word occurrence frequencies.

TNG is better at extracting thematic phrases that are at a finer granularity than that at which themes are represented in document abstracts at lower k values. ThP-Fr-13 and ThP-Fr-3 are the most averse to effects of relatively higher average word occurrence frequencies in documents in the USPTO dataset and outperform on both ext-COV and ext-FMI at higher values of k for this dataset.

Fig. 5.12 shows the consolidated radar plots that allow for visual comparison of the performances of AP, TNG, ThP-Fr-13 and ThP-Fr-3 on the 3 FMI metrics discussed above.

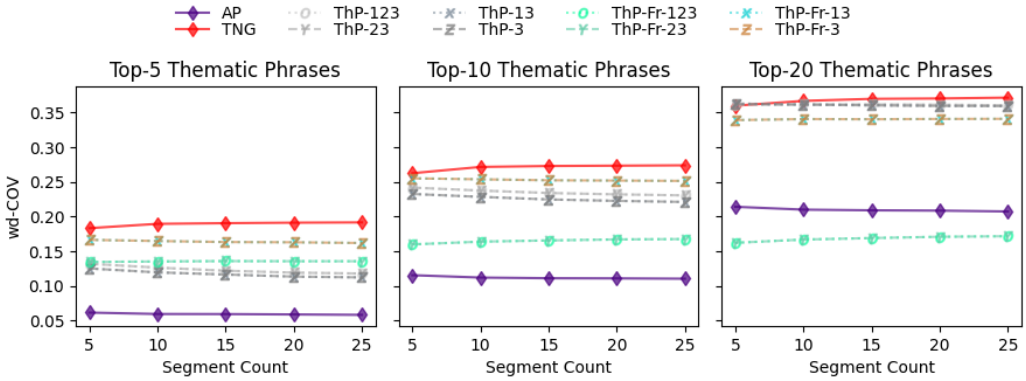


(a) Dataset: PUBMED_PMC_AB

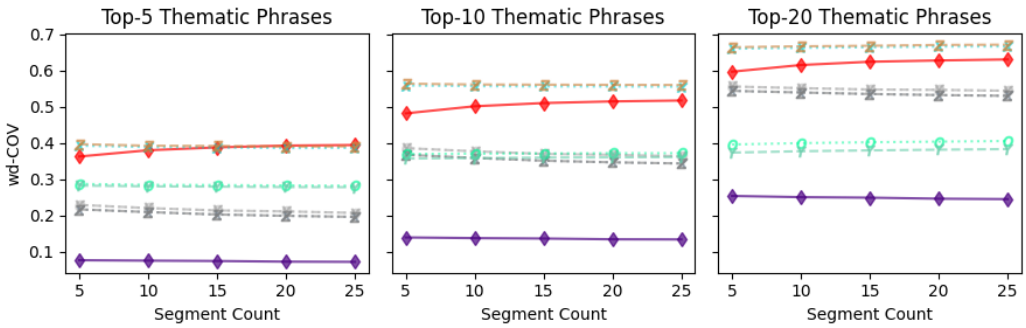


(b) Dataset: PATENT_RAND15K

FIG. 5.12: Consolidated Radar Plot: Thematic Phrases ph-FMI, sub-FMI and ext-FMI With Abstracts as Gold Standard



(a) Dataset: PUBMED_PMC_AB

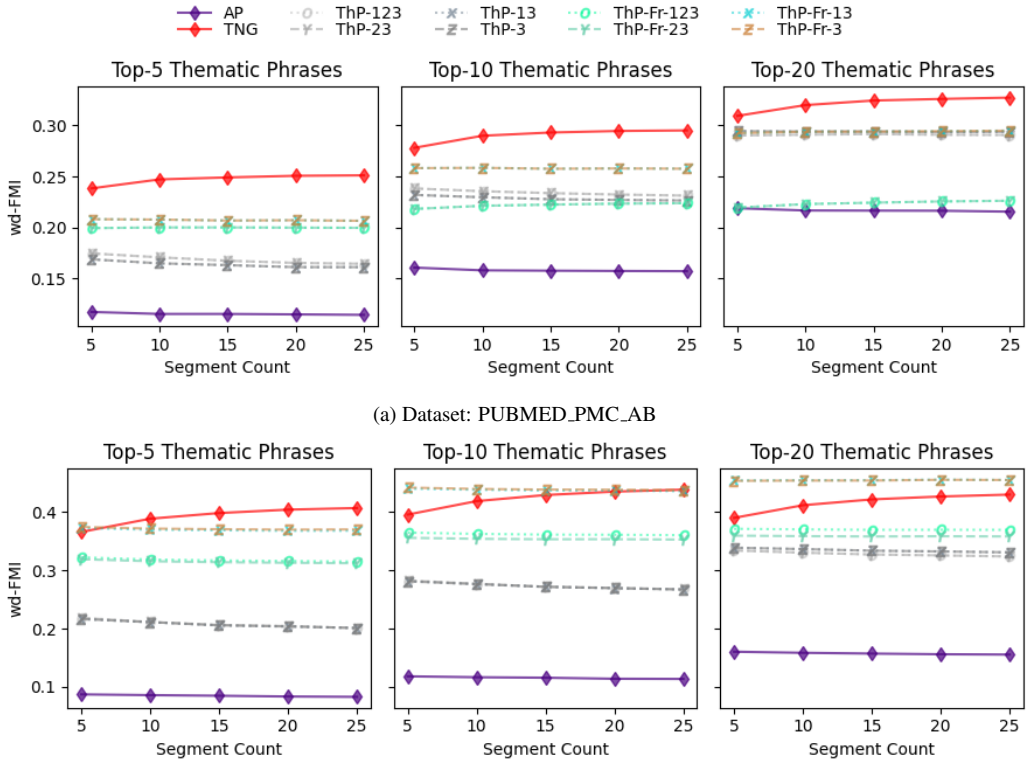


(b) Dataset: PATENT_RAND15K

FIG. 5.13: Thematic Phrases wd-COV Comparison With Abstracts as Gold Standard

Coverage, FMI and Similarity at Word Granularity: Assessment and analysis of the thematic phrases extracted by the various methods at the granularity of words that form the thematic phrases is important to add more context to the observations and conclusions drawn based on the coverage and FMI metrics at the phrase and partial phrase granularities.

Fig. 5.13 and 5.14 show performance of the thematic phrase extraction methods on wd-COV and wd-FMI respectively. All thematic phrase extraction methods have varying degrees of improvement in wd-COV as k increases. At k=20, TNG has the highest wd-COV on the PubMed dataset while ThP-Fr-13 and ThP-Fr-3 have the highest wd-COV for



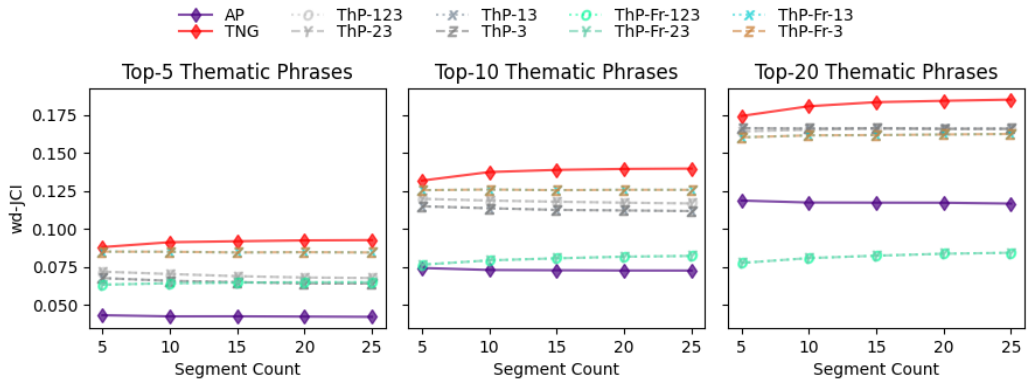
(a) Dataset: PUBMED_PMC_AB

(b) Dataset: PATENT_RAND15K

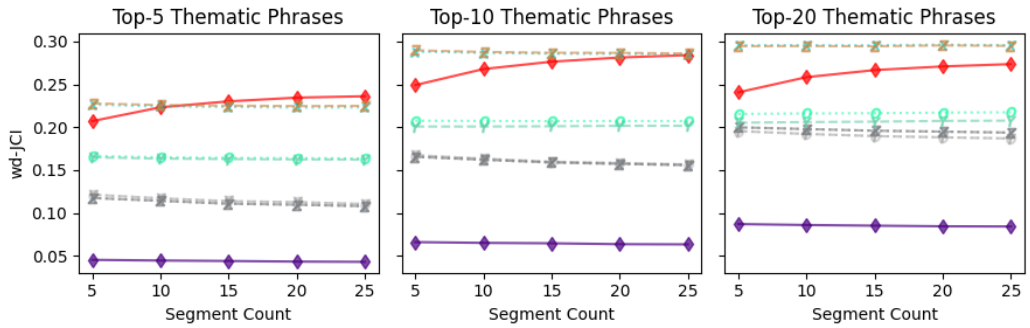
FIG. 5.14: Thematic Phrases wd-FMI Comparison With Abstracts as Gold Standard

the USPTO dataset ($P \leq \alpha_{C1}$). Further, the ThP-* configurations also outperform AP on wd-COV across k values for both datasets.

The performance of thematic phrase extraction methods on wd-FMI provides information on their word-level recall-precision performance. TNG outperforms all other methods on wd-FMI across k values for the PubMed dataset ($P \leq \alpha_{C1}$). At k=20, ThP-*, ThP-Fr-13 and ThP-Fr-3 configurations underperform TNG but outperform AP ($P \leq \alpha_{C1}$). In the case of the USPTO dataset, ThP-Fr-13 and ThP-Fr-3 have comparable wd-FMI and outperform TNG at k=20 ($P \leq \alpha_{C1}$).



(a) Dataset: PUBMED_PMC_AB



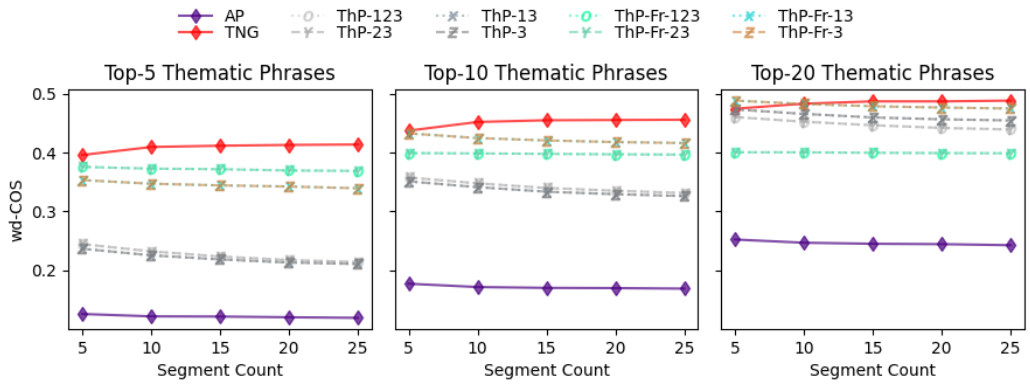
(b) Dataset: PATENT_RAND15K

FIG. 5.15: Thematic Phrases wd-JCI Comparison With Abstracts as Gold Standard

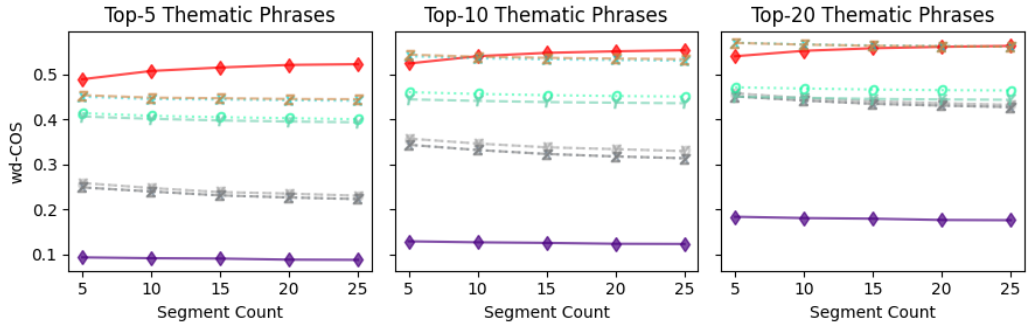
Fig. 5.15 and 5.16 show performance of the thematic phrase extraction methods on wd-JCI and wd-COS respectively. These help assess the word similarity between extracted thematic phrases and the gold standard by considering them as word sets and count-vectorized word vectors respectively. At k=20, TNG outperforms all other methods on wd-JCI for the PubMed dataset whereas ThP-Fr-13 and ThP-Fr-3 outperform for the USPTO dataset. At k=20, ThP-Fr-13, ThP-Fr-3 and TNG have comparable wd-COS for the USPTO dataset whereas TNG outperforms all other methods on wd-COS for the PubMed dataset.

The observations for wd-FMI show that ThP configurations and TNG extract thematic

phrases that are formed by more thematic relevant words than AP across k values for both datasets. ThP-Fr-13, ThP-Fr-3 and TNG extract thematic phrases that score high on word granularity metrics as well as phrase and partial-phrase granularity metrics relative to other methods in most cases. Thus, these three methods extract phrases formed by thematically relevant words that are also structurally aligned with thematic phrases present in abstracts. Fig. 5.17 is a consolidated radar plot of the performance of AP, TNG, ThP-Fr-13 and ThP-Fr-3 on the four word granularity metrics discussed above.

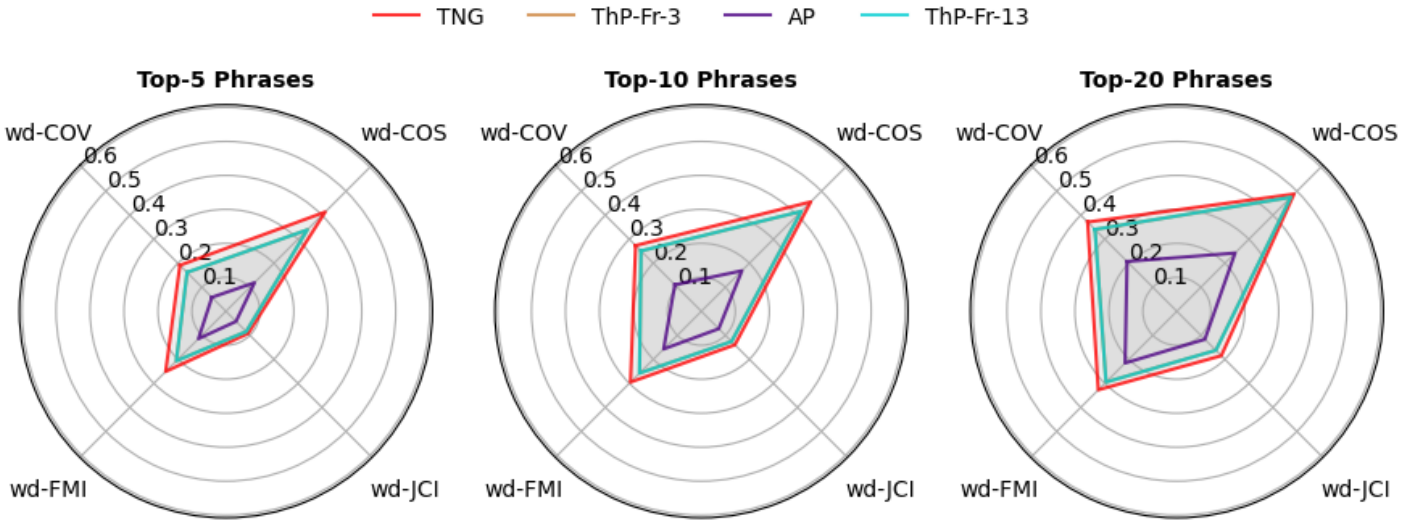


(a) Dataset: PUBMED_PMC_AB

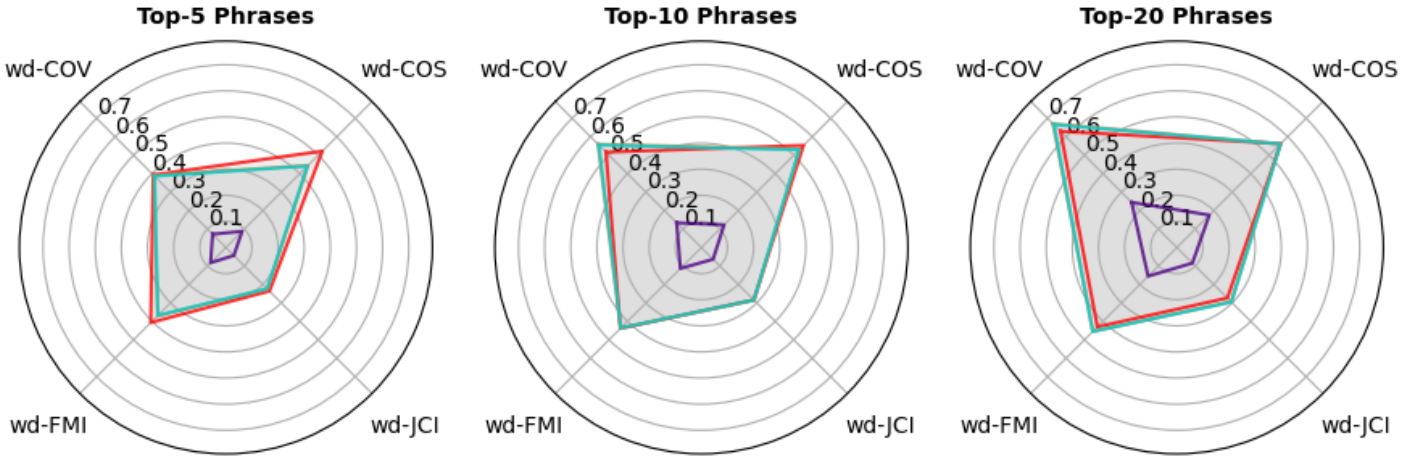


(b) Dataset: PATENT_RAND15K

FIG. 5.16: Thematic Phrases wd-COS Comparison With Abstracts as Gold Standard



(a) Dataset: PUBMED_PMC_AB



(b) Dataset: PATENT_RAND15K

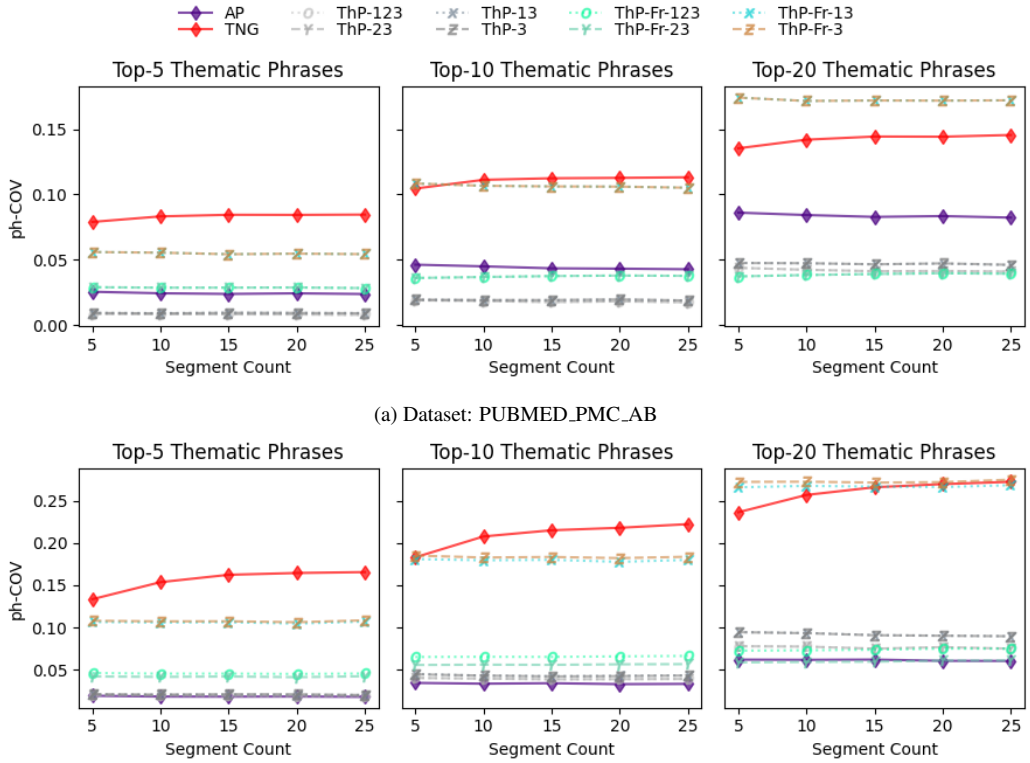
FIG. 5.17: Consolidated Radar Plot: Thematic Phrases wd-COV, wd-FMI, wd-JCI and wd-COS With Abstracts as Gold Standard

5.5.3 Quantitative Analyses of Thematic Phrases With Document Titles as the Gold Standard

The second gold standard used to evaluate the thematic phrase extraction methods is the set of titles of the documents. Titles are a more compact representation of the thematic basis of a document than abstracts. They cover the most important theme of their documents at a relatively coarser granularity while their abstracts and bodies cover the finer grained details. The comparison of the extracted thematic phrases with titles allows assessment of the phrases for precise representation of the overarching theme of the corresponding document.

This subsection discusses the quality of thematic phrases extracted by all the thematic phrase extraction methods relative to the document titles as the gold standard using the COV, FMI, JCI and COS metrics at different granularities. The discussion is divided into three parts under separate headers as follows: analyses of COV at phrase and partial phrase granularities; analyses of FMI at phrase and partial phrase granularities; and analyses of COV, FMI, JCI and COS at word granularity. The discussion under each header will state observations about the performance of the competing methods on the corresponding metrics followed by key conclusions drawn based on those observations. The critical statistical significance level used to report conclusions in this section is α_{C1}=$1.39E{-}06$. Refer Appendix A for details on the Bonferroni corrected α_{C1} critical significance level.

Coverage at Phrase and Partial Phrase Granularities: Fig. 5.18, 5.19 and 5.20 show plots of ph-COV, sub-COV and ext-COV respectively for the thematic phrase extraction



(a) Dataset: PUBMED_PMC_AB

(b) Dataset: PATENT_RAND15K

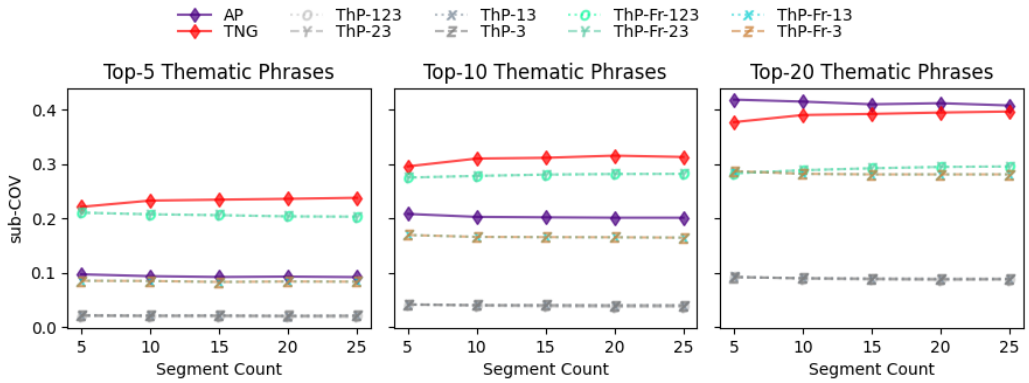
FIG. 5.18: Thematic Phrases ph-COV Comparison With Titles as Gold Standard

methods with document titles as the gold standard. Coverage metrics at all three granularities improve by varying degrees for all methods across both datasets as k increases.

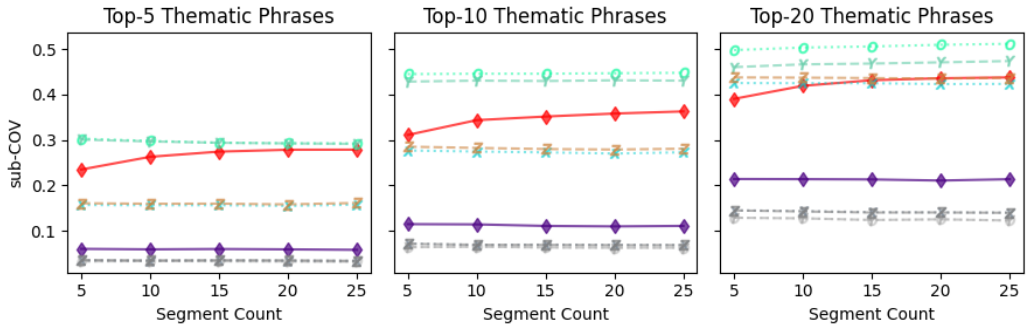
TNG has the highest ph-COV among all the methods at k=5 for both datasets. ThP-Fr-13 and ThP-Fr-3 have comparable ph-COV at all k values for both datasets. They outperform all methods on ph-COV at k=20 for the PubMed dataset ($P \leq \alpha_{C1}$) and are comparable to TNG for the USPTO dataset. All other ThP configurations and AP underperform on ph-COV consistently for both datasets.

TNG outperforms all other methods on sub-COV at $k \in \{5, 10\}$ for the PubMed dataset.

At k=20, TNG and AP have comparable sub-COV and outperform all other methods for that dataset ($P \leq \alpha_{C1}$). In the case of the USPTO dataset, ThP-Fr-123 and ThP-Fr-23 outperform all other methods on sub-COV at k=20 ($P \leq \alpha_{C1}$) with the former having the best sub-COV. ThP-Fr-13, ThP-Fr-3 and TNG have comparable sub-COV at k=20 for that dataset.



(a) Dataset: PUBMED_PMC_AB

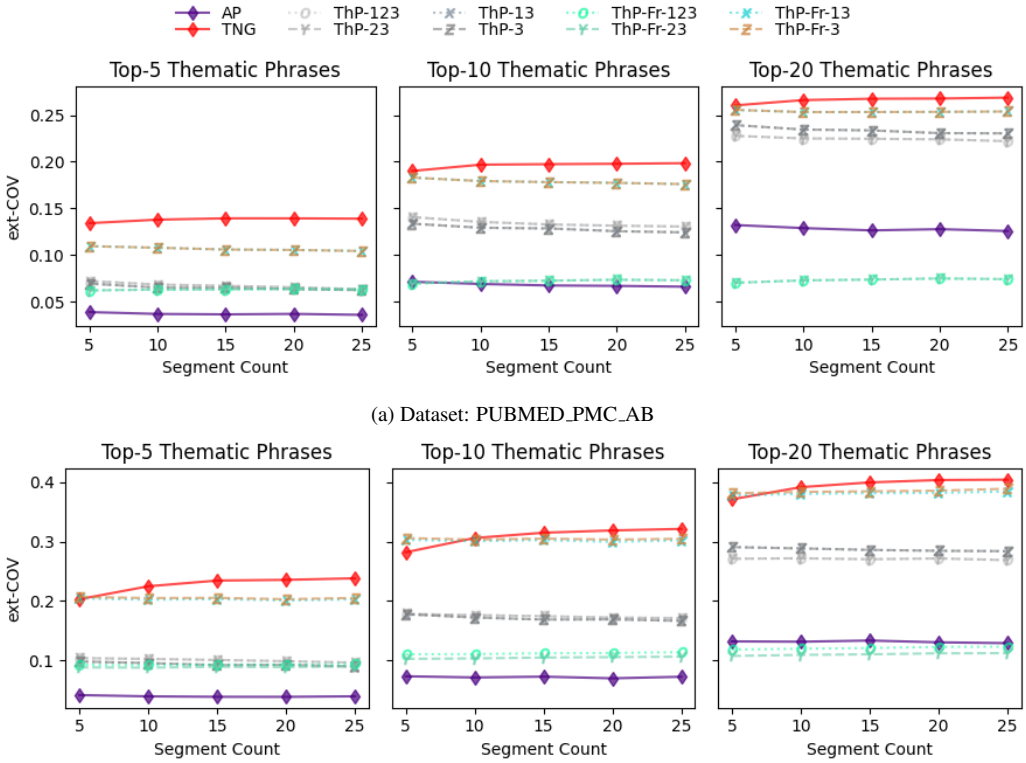


(b) Dataset: PATENT_RAND15K

FIG. 5.19: Thematic Phrases sub-COV Comparison With Titles as Gold Standard

TNG outperforms all methods on ext-COV at all k values for both datasets ($P \leq \alpha_{C1}$). ThP-Fr-13 and ThP-Fr-3 have comparable ext-COV and are the next best method on this metric for both datasets.

We can conclude the following from the observations on coverage metrics relative to



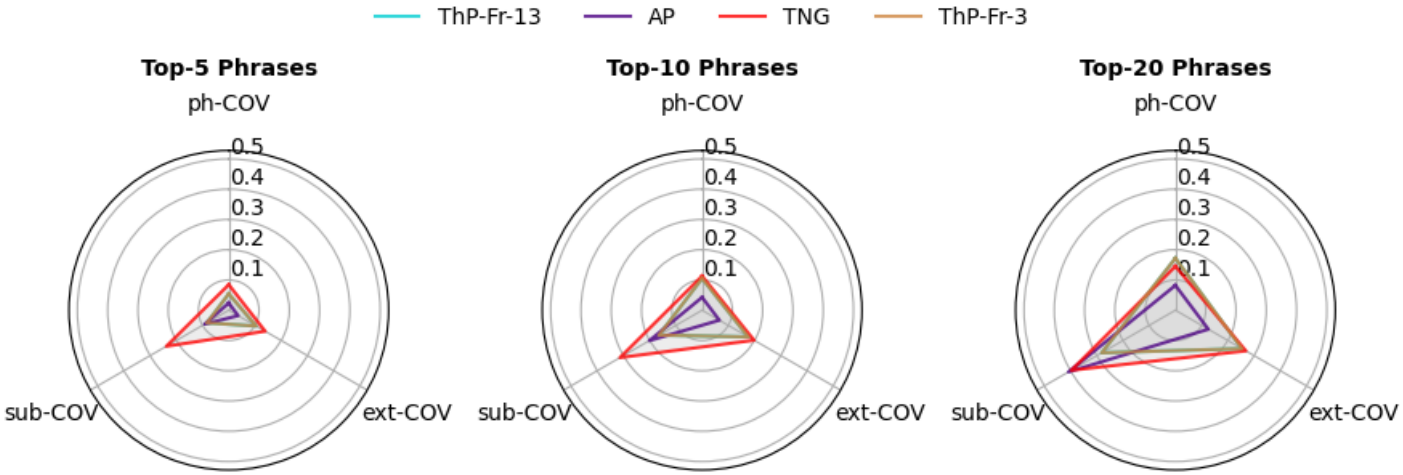
(a) Dataset: PUBMED_PMC_AB

(b) Dataset: PATENT_RAND15K

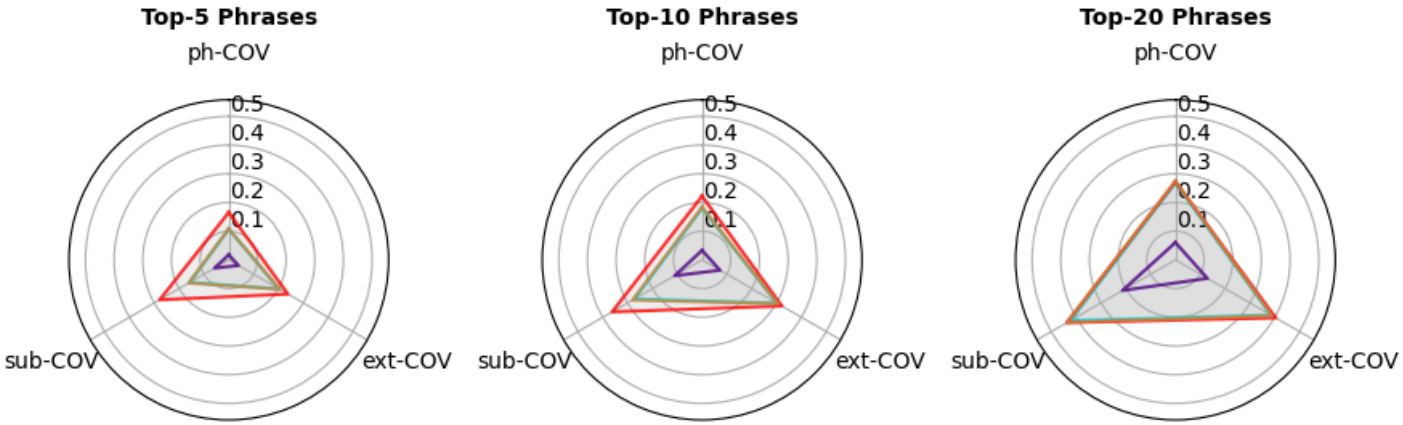
FIG. 5.20: Thematic Phrases ext-COV Comparison With Titles as Gold Standard

documents titles. ThP-Fr-13 and ThP-Fr-3 are the better methods for extracting thematic phrases at the granularity at which themes are represented in document titles at higher values of k; TNG is the better method at low values of k. TNG is better at extracting thematic phrases that are at a coarser granularity than that at which themes are represented in document titles of documents with relatively lower average word occurrence frequencies like the PubMed dataset; ThP-Fr-123 is the better method for datasets with relatively higher average word occurrence frequencies like the USPTO dataset. Lastly, TNG is the better method to extract thematic phrases that are at a finer granularity than that at which themes

are represented in document titles. Fig. 5.21 is a consolidated radar plot that allows for visual comparison of the performance of AP, TNG, ThP-Fr-13 and ThP-Fr-3 on the three COV metrics with titles as the gold standard.



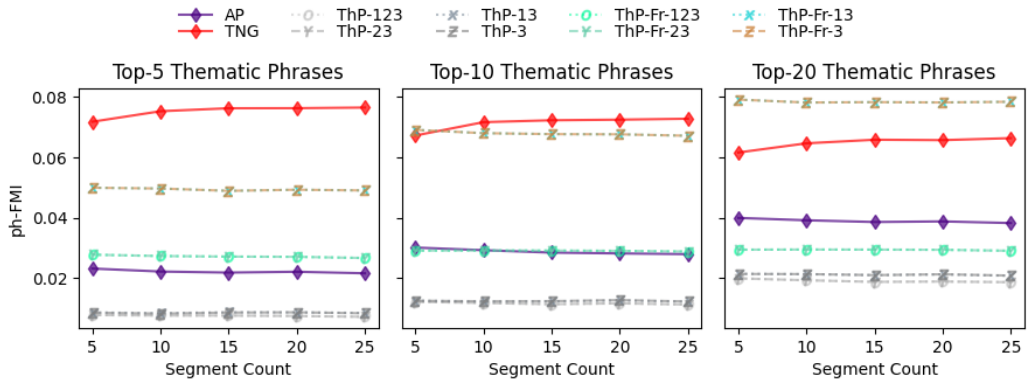
(a) Dataset: PUBMED_PMC_AB



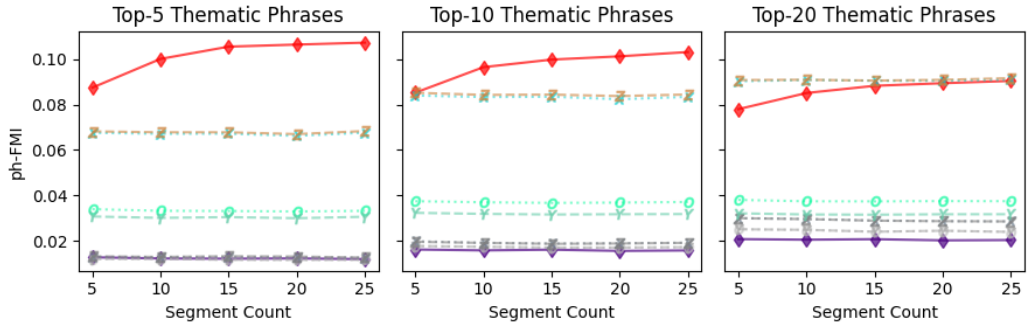
(b) Dataset: PATENT_RAND15K

FIG. 5.21: Consolidated Radar Plot: Thematic Phrases ph-COV, sub-COV and ext-COV With Titles as Gold Standard

FMI at Phrase and Partial Phrase Granularities: Fig. 5.22, 5.23 and 5.24 show plots for ph-FMI, sub-FMI and ext-FMI metrics respectively with titles as the gold standard. Unlike other methods, TNG shows a decline in ph-FMI and sub-FMI as k increases. TNG outperforms all methods on ph-FMI at $k \in \{5, 10\}$ for both datasets. At k=20, ThP-Fr-13 and ThP-Fr-3 outperform TNG on ph-FMI for the PubMed dataset ($P \leq \alpha_{C1}$) and they are comparable to TNG for the USPTO dataset. ThP-* configurations underperform at all k values for both datasets.



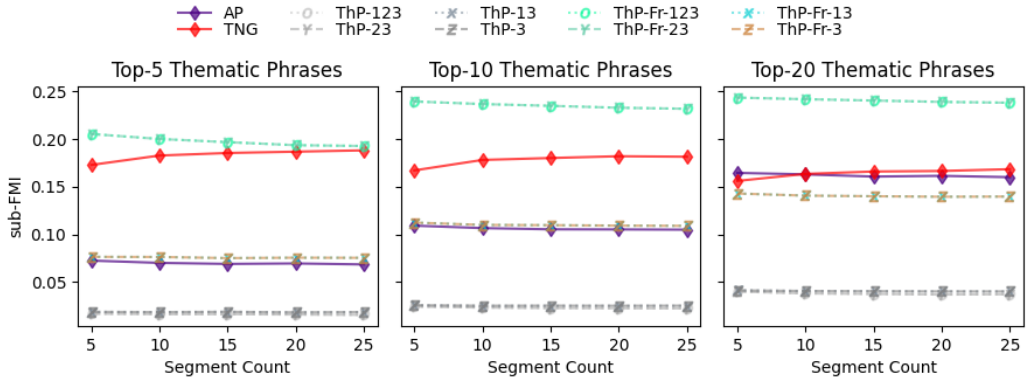
(a) Dataset: PUBMED_PMC_AB



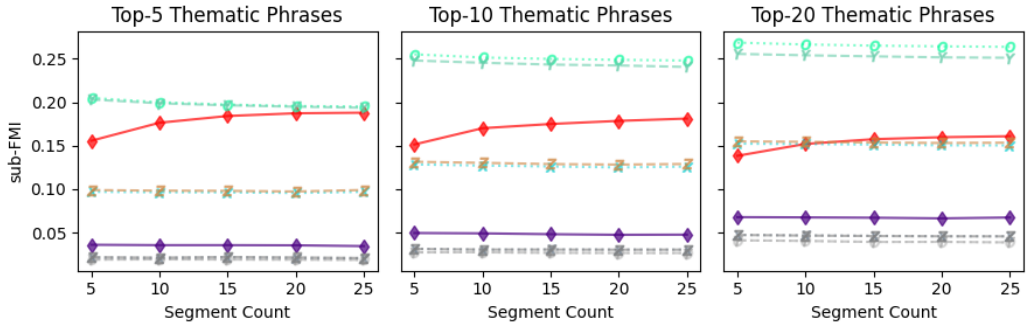
(b) Dataset: PATENT_RAND15K

FIG. 5.22: Thematic Phrases ph-FMI Comparison With Titles as Gold Standard

ThP-Fr-123 and ThP-Fr-23 are comparable with TNG on sub-FMI at k=5 and out-



(a) Dataset: PUBMED_PMC_AB



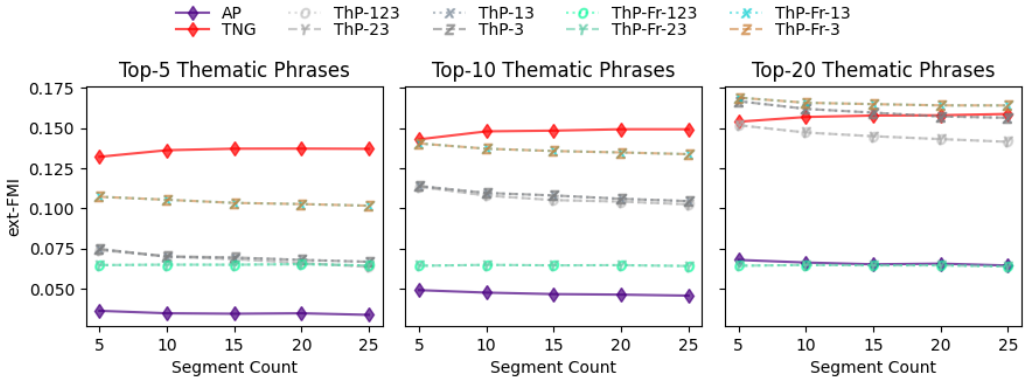
(b) Dataset: PATENT_RAND15K

FIG. 5.23: Thematic Phrases sub-FMI Comparison With Titles as Gold Standard

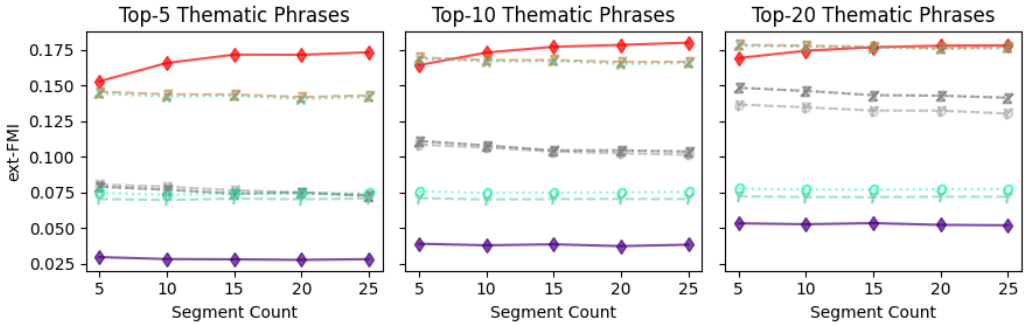
perform all other methods on sub-FMI at $k \in \{10, 20\}$ for both datasets ($P \leq \alpha_{C1}$). Further, ThP-Fr-123 has the highest sub-FMI at $k=20$ for the USPTO dataset. TNG outperforms ThP-Fr-13 and ThP-Fr-3 on sub-FMI across k values for both datasets.

TNG outperforms all other methods on ext-FMI at $k \in \{5, 10\}$ for both datasets. At $k=20$, ThP-Fr-13, ThP-Fr-3 and TNG have comparable ext-FMI for both datasets.

We can draw the following conclusions about the thematic phrase extraction methods from observations of their respective performances on coverage and FMI metrics collectively with titles as the gold standard. TNG is better at extracting thematic phrases at the granularity



(a) Dataset: PUBMED_PMC_AB



(b) Dataset: PATENT_RAND15K

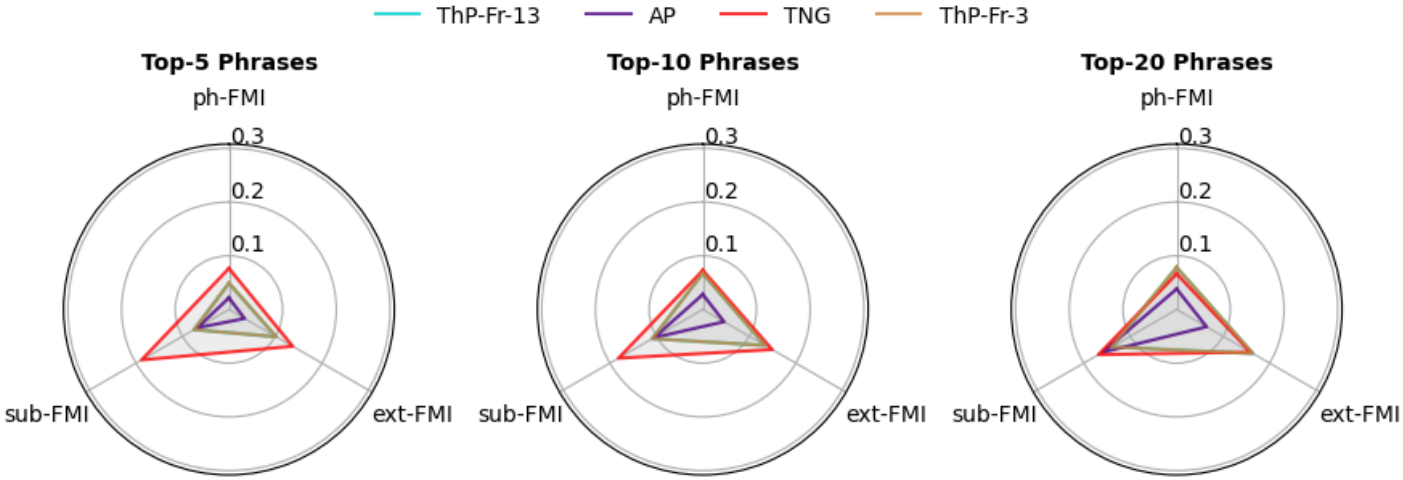
FIG. 5.24: Thematic Phrases ext-FMI Comparison With Titles as Gold Standard

at which themes are represented in document titles for both datasets at lower values of k. ThP-Fr-13 and ThP-Fr-3 are better at this granualrity for both datasets at higher values of k.

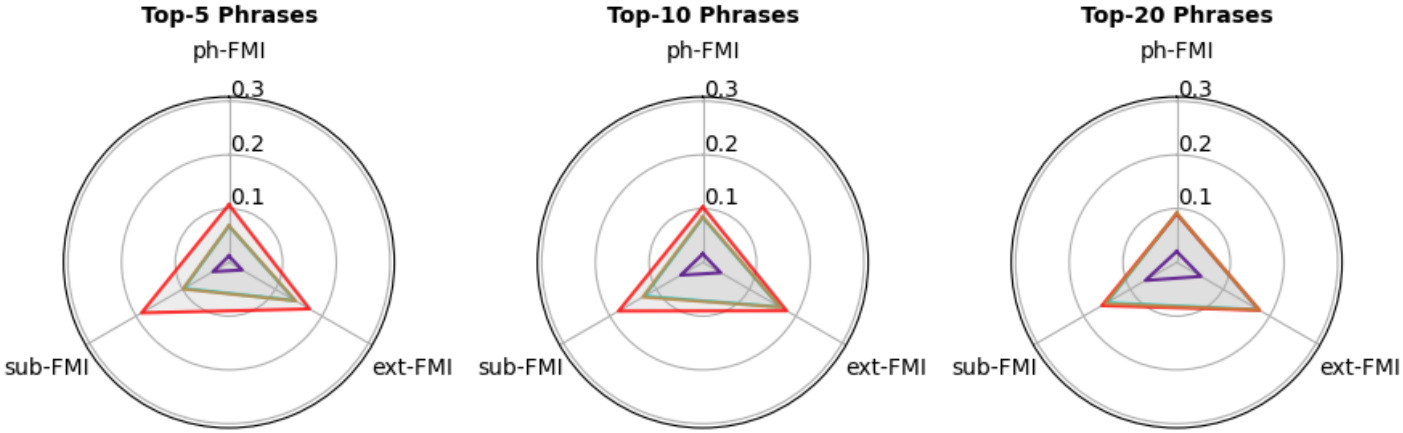
TNG is better at extracting thematic phrases that are at a coarser granularity than that at which themes are represented in document titles for datasets like PubMed with relatively lower word occurrence frequencies across k values when coverage is important for the usecases. ThP-Fr-123 and ThP-Fr-23 are better choices when a balance of recall and precision is desired as is reflected by their sub-FMI performance. ThP-Fr-123 and ThP-Fr-23 are also the better choices for extracting thematic phrases at this granularity across k values

for datasets like USPTO where average word occurrence frequencies are relatively higher.

TNG is better at extracting thematic phrases that are at a finer granularity than that at which themes are represented in document titles across k values for both datasets. Fig. 5.25 is the consolidated radar plot that allows for visual comparison of the performances of AP, TNG, ThP-Fr-13 and ThP-Fr-3 on the three FMI metrics at phrase and partial phrase granularities.

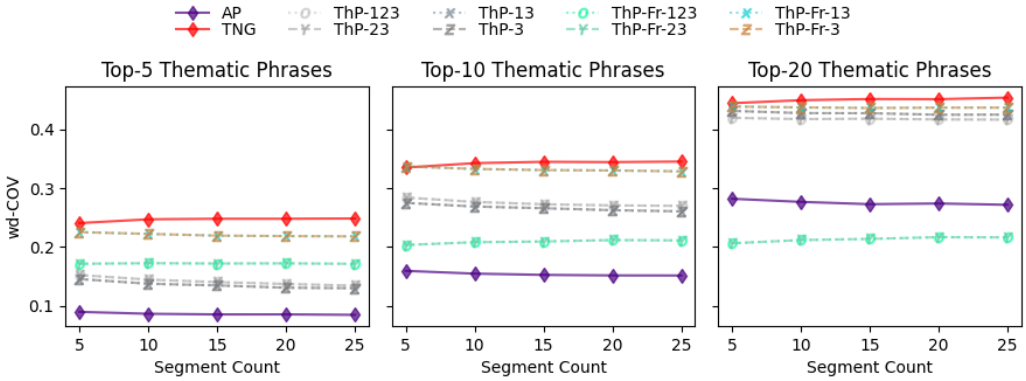


(a) Dataset: PUBMED_PMC_AB

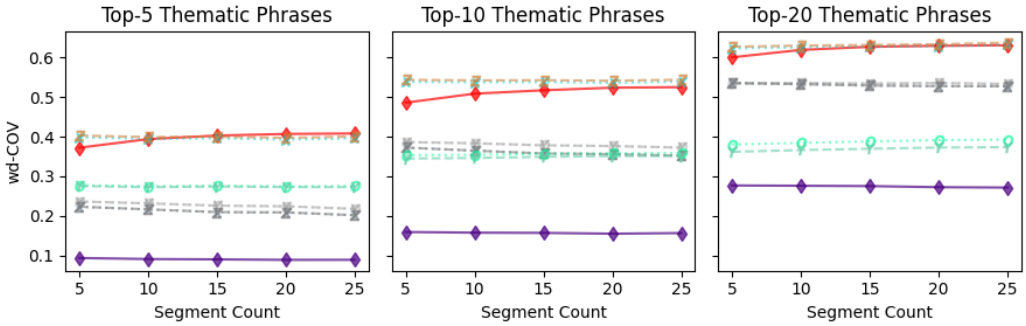


(b) Dataset: PATENT_RAND15K

FIG. 5.25: Consolidated Radar Plot: Thematic Phrases ph-FMI, sub-FMI and ext-FMI With Titles as Gold Standard



(a) Dataset: PUBMED_PMC_AB



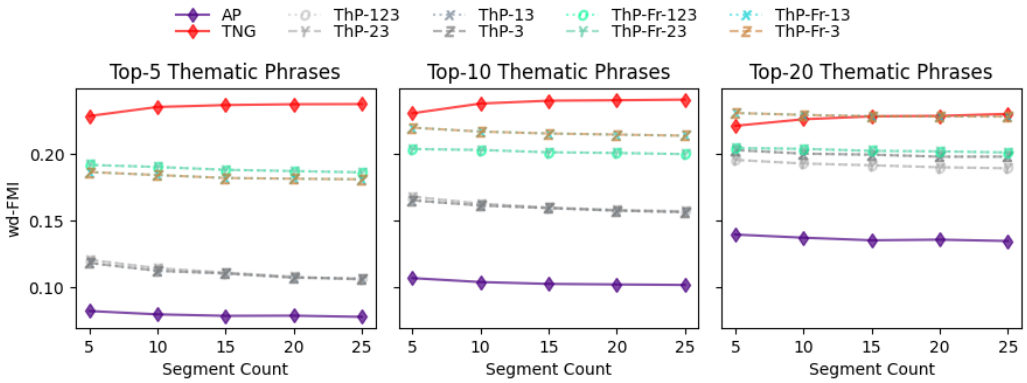
(b) Dataset: PATENT_RAND15K

FIG. 5.26: Thematic Phrases wd-COV Comparison With Titles as Gold Standard

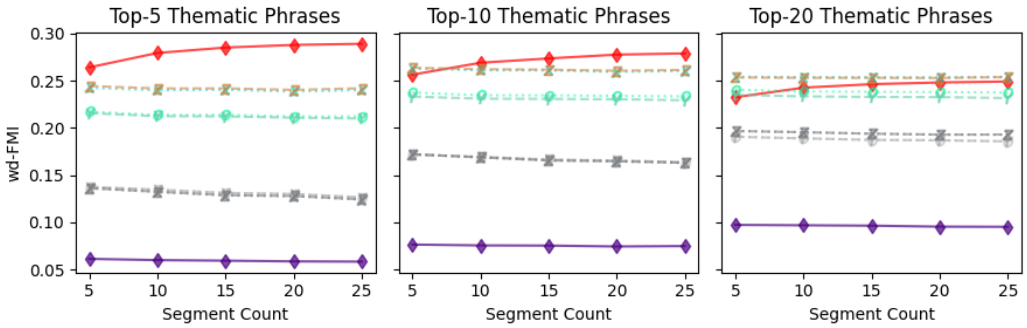
Coverage, FMI and Similarity at Word Granularity: Assessment and analysis of the thematic phrases extracted by the various methods at the granularity of words that form the thematic phrases is important to add more context to the observations and conclusions drawn based on the coverage and FMI metrics at the phrase and partial phrase granularities.

Fig. 5.26 and 5.27 show performance of the thematic phrase extraction methods on wd-COV and wd-FMI respectively with titles as the gold standard. All thematic phrase extraction methods have varying degrees of improvement in wd-COV as k increases. TNG has the highest wd-COV across k values for the PubMed dataset. In the case of the USPTO

dataset, ThP-Fr-13, ThP-Fr-3 and TNG have comparable wd-COV at k=20. Further, the ThP-* configurations outperform AP on wd-COV across k values for both datasets.



(a) Dataset: PUBMED_PMC_AB

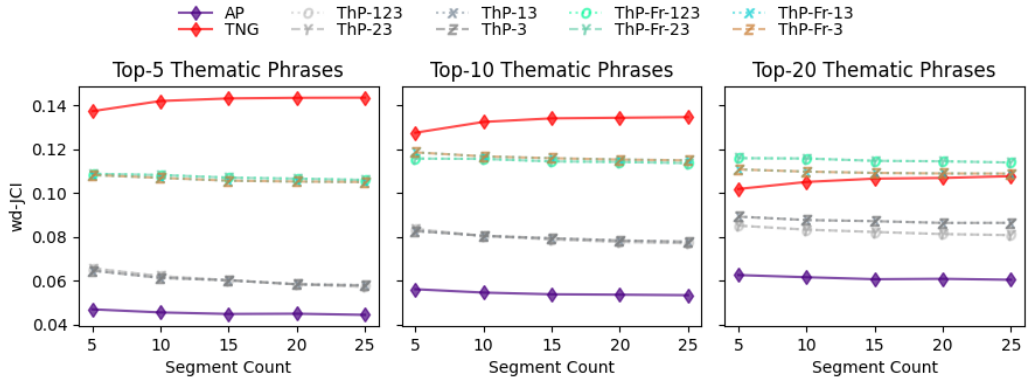


(b) Dataset: PATENT_RAND15K

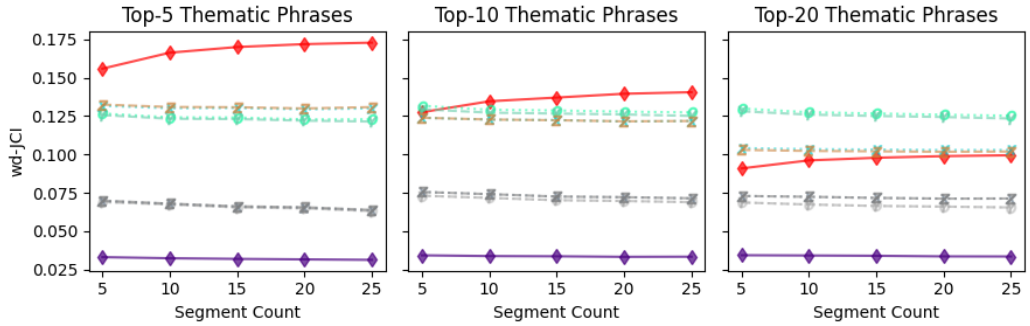
FIG. 5.27: Thematic Phrases wd-FMI Comparison With Titles as Gold Standard

The performance of thematic phrase extraction methods on wd-FMI provides information on their word-level recall-precision performance. TNG outperforms all other methods on wd-FMI at $k \in \{5, 10\}$ for both datasets. At k=20, ThP-Fr-13 and ThP-Fr-3 have comparable wd-FMI to TNG for the PubMed dataset while outperforming TNG for the USPTO dataset ($P \leq \alpha_{C1}$).

Fig. 5.28 and 5.29 show performance of the thematic phrase extraction methods on



(a) Dataset: PUBMED_PMC_AB

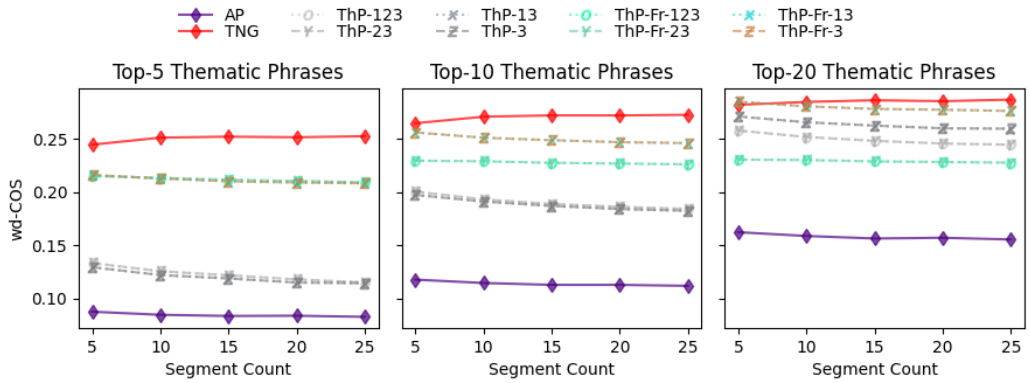


(b) Dataset: PATENT_RAND15K

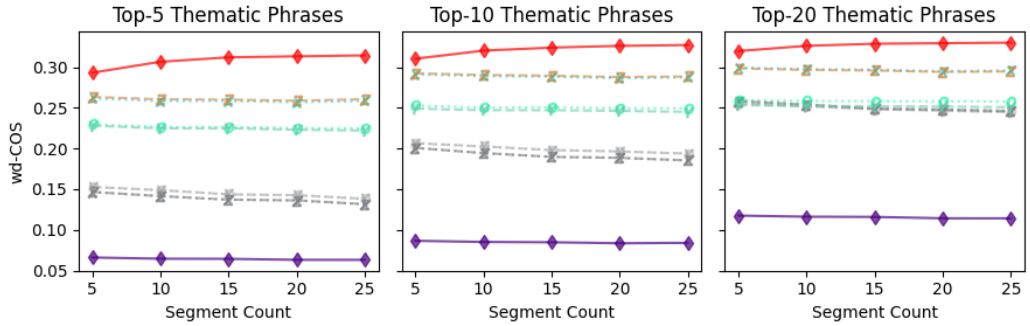
FIG. 5.28: Thematic Phrases wd-JCI Comparison With Titles as Gold Standard

wd-JCI and wd-COS respectively. These help assess the word similarity between extracted thematic phrases and the titles by considering them as word sets and count-vectorized word vectors respectively. At $k \in \{5, 10\}$, TNG outperforms all other methods on wd-JCI for both datasets whereas ThP-Fr-123 and ThP-Fr-23 outperform TNG at $k=20$. Further, TNG outperforms all other methods on wd-COS across all k values for both datasets. AP underperforms all other methods on both wd-JCI and wd-COS across k values for both datasets.

The observations for metrics at thematic word granularity show that ThP configurations



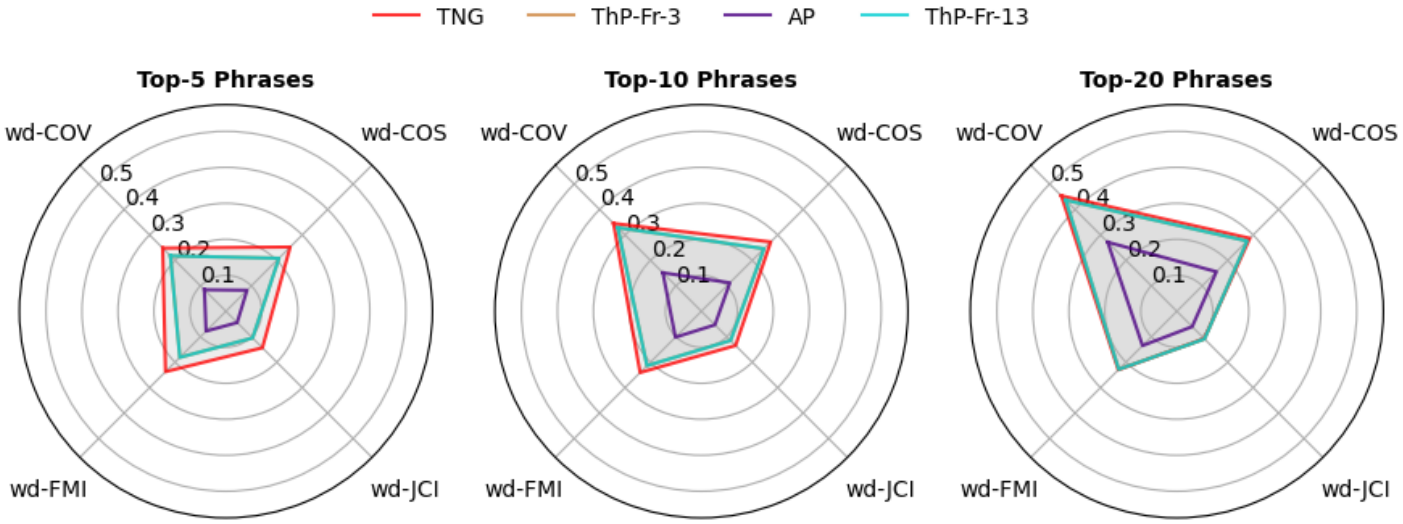
(a) Dataset: PUBMED_PMC_AB



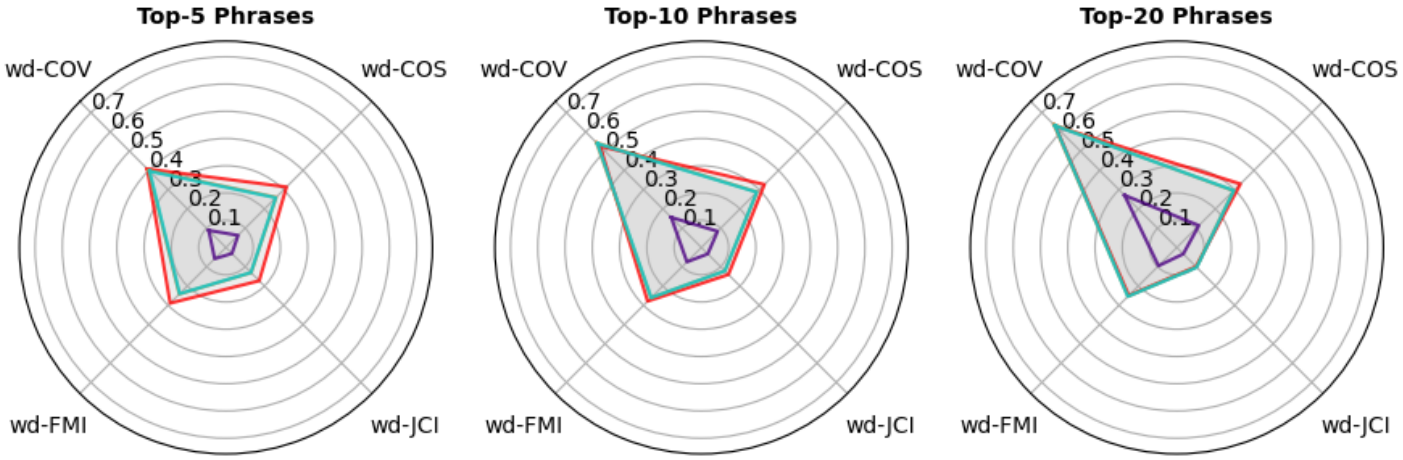
(b) Dataset: PATENT_RAND15K

FIG. 5.29: Thematic Phrases wd-COS Comparison With Titles as Gold Standard

and TNG extract thematic phrases that are formed by more thematic relevant words than AP in general. ThP-Fr-13, ThP-Fr-3 and TNG extract thematic phrases score high on word granularity metrics as well as phrase and partial-phrase granularity metrics relative to other methods in most cases. Thus, these three methods extract phrases formed by thematically relevant words that are also structurally aligned with thematic phrases present in document titles. Fig. 5.30 is a consolidated radar plot of the performance of AP, TNG, ThP-Fr-13 and ThP-Fr-3 on the four word granularity metrics discussed above.



(a) Dataset: PUBMED_PMC_AB



(b) Dataset: PATENT_RAND15K

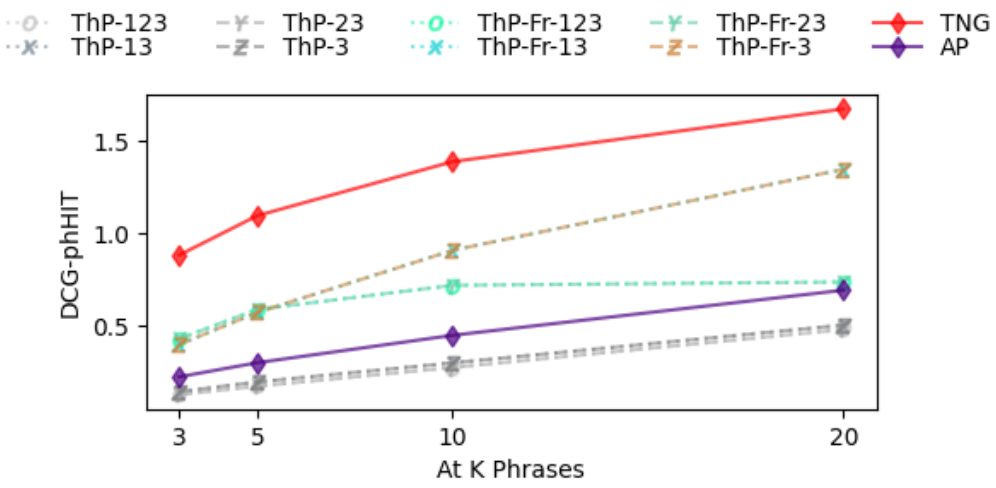
FIG. 5.30: Consolidated Radar Plot: Thematic Phrases wd-COV, wd-FMI, wd-JCI and wd-COS With Titles as Gold Standard

5.5.4 Discounted Cumulative Gain (DCG)

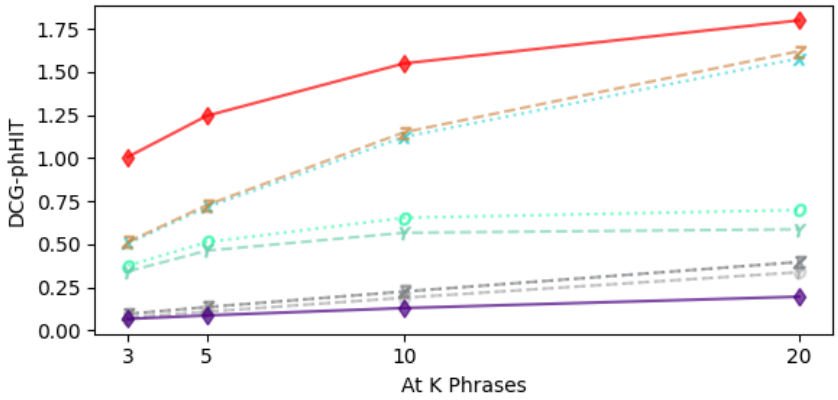
Sec. 5.5.2 and 5.5.3 discussed the quality of thematic phrases extracted by the various methods using abstracts and titles as gold standards. The observations and conclusions in those sections are for the set of thematic phrase as a whole without considering the rank order of the thematic phrases generated by the respective methods. This section focuses on analyzing rank order of the thematic phrases.

Discounted Cumulative Gain (DCG) is used to assess the rank order of the thematic phrases provided by the thematic phrase extraction methods. In the DCG metric, higher ranked relevant phrases contribute more to the DCG than lower ranked relevant ones. In this evaluation, DCG is calculated at ranks (k) 3, 5, 10 and 20 for the thematic phrases extracted by all the methods. Further, the relevance scores for the thematic phrases are calculated in three different ways as described in Sec. 5.4. DCG is calculated using abstracts as the gold standard and take into account phrase recall (in DCG-phHIT), word recall (in DCG-wdHIT) and word coverage (in DCG-wdCOV) as the three phrase relevance metrics.

DCG-phHIT is calculated using thematic phrase relevance based on phrase hits (phHIT) using the nounphrases in document abstracts as the gold standard. Fig. 5.31 shows plots comparing the DCG-phHIT for all the thematic phrase extraction methods for both the datasets. TNG, ThP-Fr-13, ThP-Fr-3 have steadily increasing DCG-phHIT as the k value increases from 3 to 20 and are the top three performers on this metric. TNG outperforms the latter two across all k values. The margin of separation between TNG and the two ThP



(a) Dataset: PUBMED_PMC_AB

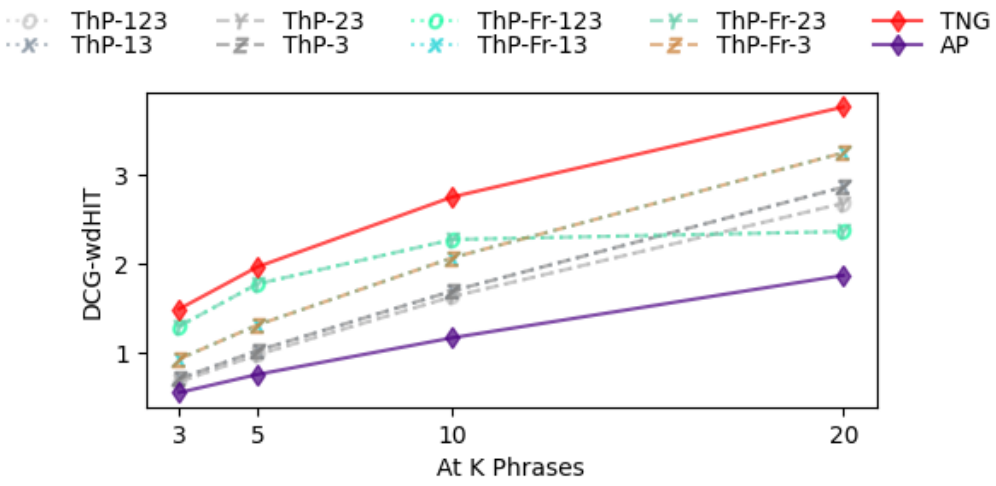


(b) Dataset: PATENT_RAND15K

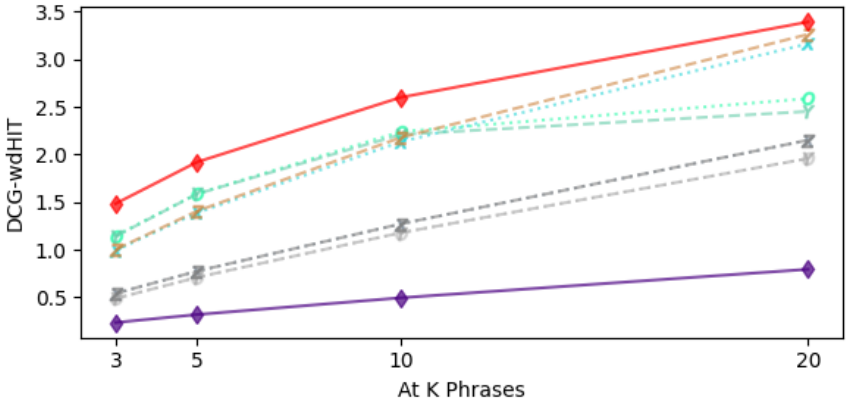
FIG. 5.31: Thematic Phrases DCG-phHIT Comparison With Abstracts as Gold Standard

configurations is relatively less for the USPTO dataset versus the PubMed dataset.

DCG-wdHIT is calculated using thematic phrase relevance based on word hits (wdHIT) using the words in the document abstracts as the gold standard. This relevance measure is based on the intuition that thematic phrases that are formed partially by thematically relevant words are relevant albeit to a lesser degree than thematic phrases that are exact matches to nounphrases in the abstracts. Fig. 5.32 shows plots comparing the DCG-wdHIT for all



(a) Dataset: PUBMED_PMC_AB

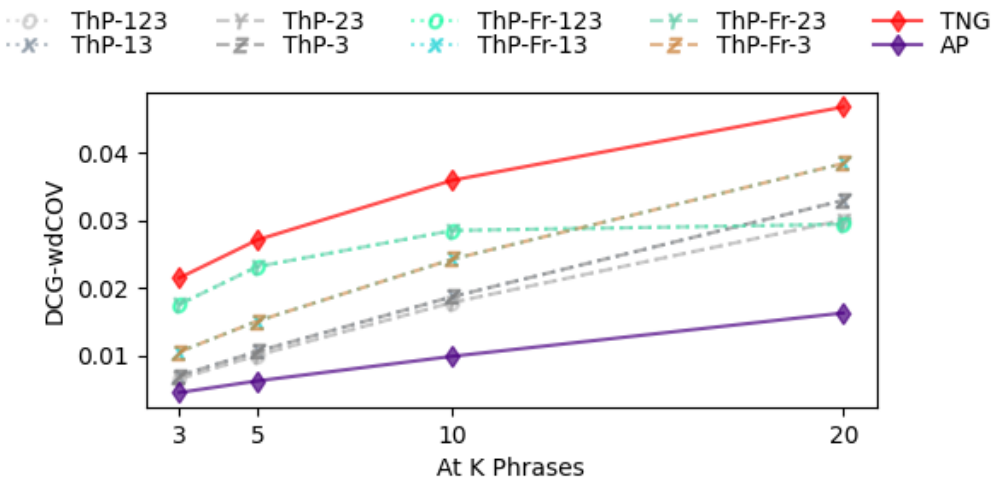


(b) Dataset: PATENT_RAND15K

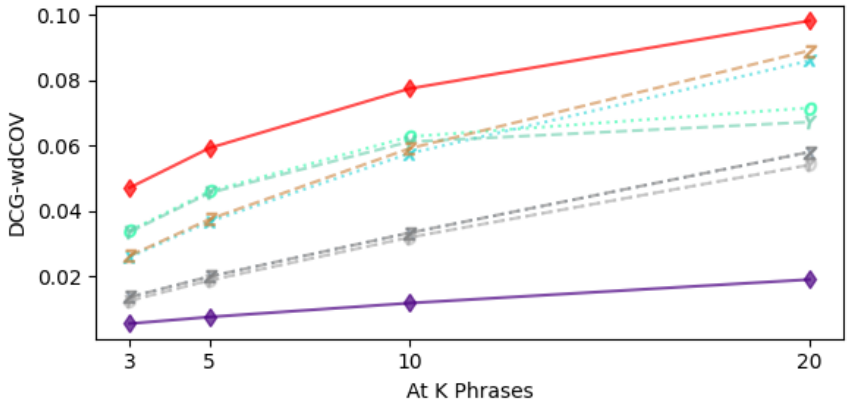
FIG. 5.32: Thematic Phrases DCG-wdHIT Comparison With Abstracts as Gold Standard

the thematic phrase extraction methods for both datasets. TNG, ThP-Fr-23 and ThP-Fr-3 are the top performing methods and show steadily increasing performance as k increases. TNG outperforms the latter two for both datasets. The margin separating TNG from the latter two methods is narrower for the USPTO dataset than for the PubMed dataset.

DCG-wdCOV is calculated using thematic phrase relevance based on word coverage (wdCOV) using the words of the document abstracts as the gold standard. This relevance



(a) Dataset: PUBMED_PMC_AB



(b) Dataset: PATENT_RAND15K

FIG. 5.33: Thematic Phrases DCG-wdCOV Comparison With Abstracts as Gold Standard

measure is based on the intuition that thematic phrases formed by thematically relevant words and occur with higher frequencies in the abstract are more thematically relevant than those formed by thematically relevant words and occur with lower frequencies. This measure adds information of word frequencies to the previous phrase relevance measure. Fig. 5.33 shows plots comparing the DCG-wdCOV of all the thematic phrase extraction methods for both datasets. TNG, ThP-Fr-13, ThP-Fr-3 are the top performing methods on

this metric. TNG outperforms the latter two for both datasets.

A common observation across all three DCG metrics is that TNG, ThP-Fr-13 and ThP-Fr-3 are top performers at k=20 on all three DCG metrics for both datasets. The margin of separation between TNG and the latter two ThP configurations reduces as k increases. This is consistent with observations in previous subsections that show that TNG's FMI performance plateaus or decreases as k increases. The conclusion drawn that TNG adds more thematically irrelevant phrases than relevant ones as k increases, contrary to the ThP configurations, explains the decreasing margin of separation between them on the DCG metrics. Further, AP underperforms on all DCG metrics, which is consistent with it's underperformance on the COV, FMI and similarity metrics discussed in previous subsections.

The ThP configurations that do not take phrase occurrence frequencies into consideration underperform on DCG-phHIT but perform better on DCG-wdHIT and DCG-wdCOV. This indicates that the thematic phrases extracted by these configurations are formed by thematically relevant words but the phrases themselves lack in semantic and syntactic construction relative to the nounphrases in the abstracts. This is also indicated by their underperformance on phrase and partial phrase granularity COV and FMI metrics discussed in Sec. 5.5.2.

Another observation to highlight is the performance of ThP-Fr-123 and ThP-Fr-23. Both these configurations show a steady increase in all three DCG metrics as k increases and plateau at k equal to 10. This is because the WPOS heuristic, that is active in these two configurations, aggressively restricts the candidate phrases that are passed through

to downstream heuristics. Hence, the number of thematic phrases extracted by these configurations across all documents on average is less than 20. Thus, the DCG for these configurations does not increase as k increases for most documents and the mean DCG begins to plateau as a result.

5.5.5 Effects of Segment Count

Chapter 4 described five different segment counts that each input document is partitioned into for evaluating the thematic phrase extraction methods. The methods consider each document's partitions as a corpus of documents. The ThP configurations and TNG perform topic modeling as part of their pipeline while AP computes cross-document n-gram heuristics for thematic phrase extraction. Thus, all methods perform cross-document (cross-partition in our case) computations to extract thematic phrases and can be affected by the nature of the partitions.

The variation in segment counts of the input documents has two effects on these methods. Firstly, the number of segments determines the input document corpus size for each of these methods. Secondly, in the case of ThP and TNG, the number of segments determines the topic to document count ratio for their respective topic modeling components given a constant topic count. The effects of varying segment counts on the thematic phrase extraction methods is assessed using the following measures of thematic phrase variance:

Thematic Phrase Set Variance: This is the variation in the thematic phrases extracted from a document by a method for different segment counts. It is measured using the

mean of the pairwise Jaccard Index (JCI) computed for all pairs of thematic phrase sets extracted for the different segment counts.

A higher JCI indicates higher similarity between the thematic phrase sets extracted across all segment counts, i.e. lower variance in thematic phrases set membership. Hence, a higher JCI is desired.

Thematic Phrase Set and Rank Order Variance: This is the difference in the thematic phrase sets as well as the rank order of the thematic phrases generated by a method from a document for different segment counts. This is measured using the mean of the pairwise Damerau-Levenshtein distances computed for all pairs of ranked order thematic phrases extracted across different segment counts.

The Damerau-Levenshtein distance (DLD) [1] is a variation of the commonly used Levenshtein distance and allows for the transposition edit operation in addition to the insertion, deletion and substitution edit operations considered by the standard Levenshtein distance. DLD is suitable for measuring rank variations in otherwise identical thematic phrase sets. This variance is a consolidated measure of difference in thematic phrase set membership and the rank order of the thematic phrases extracted by a method for different segment counts.

A lower DLD indicates higher similarity between the thematic phrase sets and their rank order across segment counts, i.e. lower variance in thematic phrases set member-

[1] https://en.wikipedia.org/wiki/Damerau-Levenshtein_distance

ship and rank order. Hence, a lower DLD is desired.

Tab. 5.5, 5.6 and 5.7 contain the JCI and DLD values for the top-5, top-10 and top-20 thematic phrases respectively that are extracted by the competing methods for both the datasets. Fig. 5.34 and 5.35 show plots of JCI and DLD respectively for all the methods at these k values for both the datasets. All thematic phrase extraction methods are affected by the segment counts. The effects are of varied degrees for different methods and at different values of k for the same method.

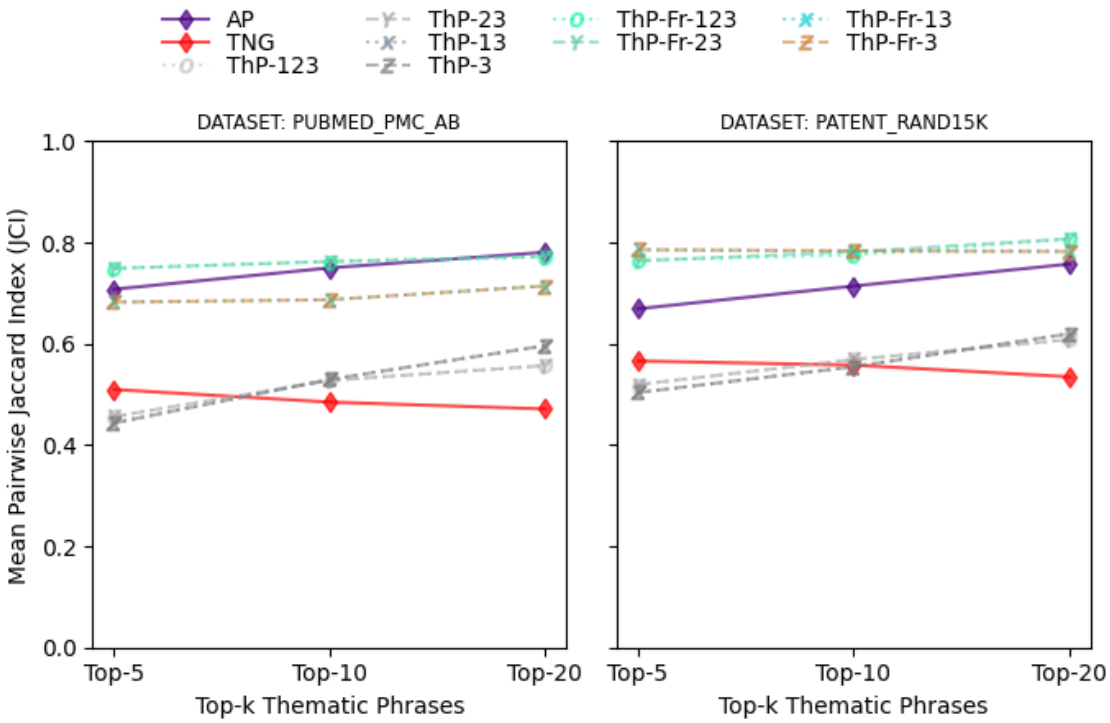


FIG. 5.34: Thematic Phrases Variance Comparison Across Segment Counts Using JCI

Fig. 5.34 shows that ThP configurations that use phrase occurrence frequencies (method prefix ThP-Fr-, referred to here as ThP-Fr-*) and AP have the lowest thematic

phrase variance in terms of set membership indicated by their JCI values. ThP configurations that do not use phrase occurrence frequencies (method prefix ThP-, referred to here as ThP-*) have relatively larger JCI compared to configurations that use phrase occurrence frequencies. Fig. 5.35 shows that ThP-Fr-* configurations have the lowest thematic phrase variance in terms of membership and rank order indicated by their respective DLD values. Further, AP underperforms the ThP-Fr-* configurations as well as ThP-* configurations when rank order of the thematic phrases is considered.

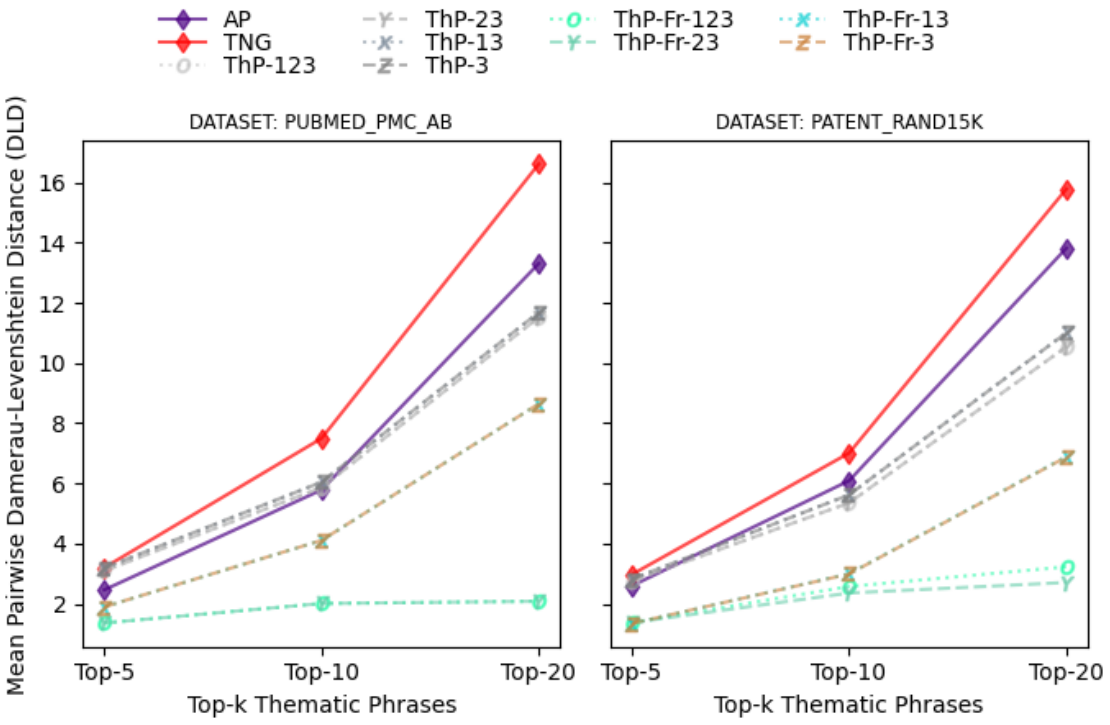


FIG. 5.35: Thematic Phrases Variance Comparison Across Segment Counts Using DLD

TNG has the highest thematic phrase variance across segment counts when both thematic phrase set membership and rank order is considered as can be seen in Fig. 5.35.

Fig. 5.34 shows that TNG has higher JCI at k=5 and comparable JCI at k=10 for both datasets relative to the ThP-* configurations. But, Fig. 5.35 shows that TNG has higher DLD at both those k values relative to the ThP-* configurations. This indicates that the rank order of thematic phrases extracted by TNG at k=5 and k=10 vary as the segment counts change.

Method	**Dataset = PUBMED_PMC_AB**		**Dataset = PATENT_RAND15K**	
	JCI	**DLD**	**JCI**	**DLD**
AP	0.7076	2.4617	0.669	2.5954
TNG	0.5096	3.1994	0.5659	2.9726
ThP-123	0.4561	3.0944	0.5194	2.7554
ThP-13	0.4435	3.1883	0.5036	2.8204
ThP-23	0.4561	3.0944	0.5194	2.7554
ThP-3	0.4434	3.1898	0.5036	2.8215
ThP-Fr-123	0.7488	1.3621	0.7654	1.378
ThP-Fr-13	0.6822	1.894	0.7858	1.3398
ThP-Fr-23	0.7488	1.3621	0.7635	1.3718
ThP-Fr-3	0.6819	1.8951	0.7856	1.3434

Table 5.5: Thematic Phrases Variance Comparison Across Segment Counts for Top-5 Phrases

Method	**Dataset = PUBMED_PMC_AB**		**Dataset = PATENT_RAND15K**	
	JCI	**DLD**	**JCI**	**DLD**
AP	0.7497	5.7831	0.714	6.0888
TNG	0.4847	7.4943	0.5577	7.008
ThP-123	0.5275	5.8617	0.5687	5.3523
ThP-13	0.5295	6.0413	0.5545	5.6215
ThP-23	0.5275	5.8617	0.5687	5.3523
ThP-3	0.5294	6.0424	0.5547	5.6187
ThP-Fr-123	0.7624	2.0093	0.7753	2.5741
ThP-Fr-13	0.6871	4.1044	0.7827	3.0024
ThP-Fr-23	0.7624	2.0093	0.7818	2.3573
ThP-Fr-3	0.6868	4.1063	0.7837	2.9899

Table 5.6: Thematic Phrases Variance Comparison Across Segment Counts for Top-10 Phrases

Sec. 5.5.2 and 5.5.3 discussed that the WPOS heuristic in ThP aggressively filters

Method	Dataset = PUBMED_PMC_AB		Dataset = PATENT_RAND15K	
	JCI	DLD	JCI	DLD
AP	0.7807	13.3046	0.7574	13.8153
TNG	0.4711	16.611	0.5347	15.7797
ThP-123	0.5567	11.5138	0.6081	10.5393
ThP-13	0.5962	11.677	0.6195	11.0107
ThP-23	0.5567	11.5138	0.6081	10.5393
ThP-3	0.5962	11.6784	0.6196	11.0037
ThP-Fr-123	0.7719	2.0833	0.807	3.2282
ThP-Fr-13	0.7141	8.6421	0.7827	6.8934
ThP-Fr-23	0.7719	2.0833	0.807	2.7073
ThP-Fr-3	0.714	8.6438	0.782	6.8763

Table 5.7: Thematic Phrases Variance Comparison Across Segment Counts for Top-20 Phrases

candidate phrases and, as a result, ThP-Fr-123 and ThP-Fr-23 are unable to generate more than 10 thematic phrases for many documents. This explains the plateauing of the DLD between k=10 and k=20 for these two ThP configurations in Fig. 5.35.

The observations above show that ThP configurations that use phrase occurrence frequencies (i.e ThP-Fr-*) are more robust and deterministic in their thematic phrase extraction independent of segment counts than other methods. Further, Sec. 5.5.2 and 5.5.3 discussed that the performance of TNG on all quantitative metrics improves as the segment count increases from 5 to 25. Thus, the thematic phrase variance observed for TNG contributes to quality improvements of its thematic phrases as segment count increases. This quality improvement, though, begins to plateau at segment count = 15. Hence, the thematic phrase variance for TNG for segment counts between 10 and 25 do not help to improve its thematic phrases proportional to the variance. Further, we can conclude that TNG will require parameter optimization of segment counts for a collection of documents in order to extract quality

thematic phrases. This is not the case with the other methods.

Chapter 6

IMPROVING EXTRACTIVE TEXT SUMMARIZATION USING THEMATIC PHRASES BASED SENTENCE PRE-FILTERING

Automatic extractive text summarization is a group of methods that create a gist or summary of an input text. They achieve this by extracting sentences from the text that they deem are most representative of the discourse contained in the text at a desired granularity. Chapters 1 and 2 discussed how inferred topicality or thematic basis of a document can help augment automatic text summarization. In this chapter, the thematic phrases extracted by various approaches evaluated in Chapter 5 are used to bias the TextRank [54] extractive text summarization method. The intuition is that biasing a text summarizer to consider only sentences related to relevant thematic phrases will lead to summaries of better quality since sentences that are irrelevant to the theme of the text are not considered and hence will not be part of the summaries.

Sec. 6.1 briefly describes the TextRank summarization method. Sec. 6.2 describes how input document sentence pre-filtration using thematic phrases is utilized to bias the summarizer. Sec. 6.3 details the experiment setup used to evaluate summarization quality. This is followed by a description of the ROUGE summary evaluation framework and its metrics used for evaluation in Sec. 6.4. Sec. 6.5 provides an analyses of summarization quality of the various methods along with key conclusions. The complete tabular representation of the values of ROUGE metrics for all experiments conducted in this chapter are provided in Appendix K.

6.1 TextRank Summarizer

TextRank [54] is an automatic text summarization method that performs unsupervised, extractive text summarization. It adapts PageRank [65], a graph-based ranking approach, to rank sentences by relevance for an input document. It then chooses top-m sentences to create a summary of size m.

PageRank approximates a webpage's importance in a network of linked webpages. The PageRank of a webpage in such a network is the measure of that webpage's mutual importance among other linked webapges. Computation of PageRank is analogous to eigenvalue decomposition of the adjacency matrix of a directed graph in which the nodes represent webpages, the edges represent links that exist between the webpages, and the edge weights are initialized to *transition-probability* $= \frac{1}{out-degree}$ of each edge's source node.

TextRank constructs a directed graph over a document with sentences as nodes and the

edge weights between sentences (nodes) represent pairwise similarities between sentences. The document level graph is then used to compute PageRank which ranks the nodes, representing sentences, in order of their linkages weighted by sentence similarity scores. This ranking is used by TextRank to extract the top-m sentences in their order of occurrence in the document which is the final extracted summary.

The sentence similarity metric used in the original version of TextRank [54] is the length-normalized unigram intersection between two sentences. Several other sentence similarity metrics [63] have been evaluated for use with TextRank to obtain summarization quality improvements. The BM25 relevance metric [66, 67] is shown to provide the most quality improvement. The modified TextRank implementation with BM25 is utilized for experiments in this work (the implementation package and links are provided in Sec. 4.2).

6.2 TextRank Summarizer With Sentence Pre-filtration

This section discusses how thematic phrases can be used to augment text summarization and improve the quality of the extracted summary. This is done by using thematic phrases to filter document sentences before they are provided as input to TextRank. The sentence pre-filtration uses extracted thematic phrases to filter document sentences in three different ways that correspond to theme consideration at the thematic phrase, thematic sub-phrase, and thematic word granularities as follows:

Thematic phrase granularity: to achieve filtration at this granularity sentences in

the text documents are filtered based on the presence of entire thematic phrases as nounphrases in the sentences.

Thematic sub-phrase granularity: to achieve filtration at this granularity, sentences in the text documents are filtered based on the presence of:

(a) Entire thematic phrases as nounphrases in the sentences.

(b) Sub-phrases of the thematic phrases as nounphrases in the sentences.

(c) Entire thematic phrases as sub-phrases of the nounphrases in the sentences.

Thematic word granularity: to achieve filtration at this granularity, sentences in the text documents are filtered based on the presence of words that form the thematic phrases in the sentences.

The quality of extracted summaries using TextRank and each of these three pre-filtration granularities is evaluated for all the thematic phrase extraction methods.

6.3 Experiment Setup

The evaluation and discussion in Chapter 5 showed that TNG thematic phrase extraction performs best at segment count = 25. On some metrics, AP also performs relatively better at segment count = 25. Further, all ThP configurations have relatively minor variations in performance across all quantitative metrics as segment counts vary. Hence, for the text summarization evaluation and analyses in this chapter we consider thematic phrases

extracted by all the competing methods at segment count = 25 for sentence pre-filtering.

The quantitative evaluation of thematic phrase extraction methods in Chapter 5 also evaluated their respective thematic phrases at k values $\in \{ 5, 10, 20 \}$. At $k = 5$, document sentence pre-filtration using five thematic phrases often results in a filtered sentence count that is lower than the sentence count of the abstracts of the documents being summarized. This skews the text summary quality evaluation. Hence, for the purposes of the evaluation in this chapter only thematic phrases at $k \in \{ 10, 20 \}$ are considered for sentence pre-filtration text summarization.

Text summaries are extracted at $k \in \{ 10, 20 \}$ for all thematic phrase extraction methods using each of the three document sentence filtering strategies discussed in Sec. 6.2. These summaries are compared with the baseline summary extracted by TextRank using the whole text document as input without sentence filtration. The evaluation uses document abstracts as the reference or gold standard summary. The quality of the extracted text summaries is evaluated using the ROUGE automatic summary evaluation framework that is discussed in Sec. 6.4.

6.4 ROUGE Framework

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation [55, 56, 57], is a package for comparing system-generated texts with reference ground-truth texts and evaluating their similarity using metrics at various granularity. It is useful for evaluating automatic summarization and machine translation methods. We evaluate the extractive summaries

generated in the experiments in this chapter using ROUGE metrics that use different word overlap lengths. The following five ROUGE metrics are utilized because they correlate most with human judgments about long and short single document summary quality compared to other ROUGE metrics. [57]:

ROUGE-1: measures similarity based on overlap of unigrams between the system and reference summaries.

ROUGE-2: measures similarity based on overlap of bigrams (two-word phrases) between the system and reference summaries.

ROUGE-3: measures similarity based on overlap of trigrams (three-word phrases) between the system and reference summaries.

ROUGE-L: measures similarity based on the longest common subsequences or longest overlap of in-sequence words per sentence between the reference and system summaries. At the summary level, the union of the longest overlapping in-sequence words per sentence are considered. This differs from ROUGE-1 in that ROUGE-1 considers overlap of words without imposing the requirement of in-sequence occurrence of the words.

ROUGE-SU4: measures similarity based on overlap of unigrams as well as overlap of skip-bigrams with a maximum skip distance of four. That is, the two overlapping words (in the bigrams) need to occur in-sequence but need not be consecutive as is

the case with regular bigrams considered in ROUGE-2.

6.5 ROUGE Evaluation of Extractive Summaries

The text summarization quality evaluation is organized into Sec. 6.5.2, 6.5.3 and 6.5.4 that address thematic phrase granularity pre-filtering, thematic sub-phrase granularity pre-filtering and thematic word granularity pre-filtering respectively. Each sub-section discusses the quality of its respective summaries using the five ROUGE metrics discussed in Sec. 6.4 for both the PubMed and USPTO datasets. The critical statistical significance level used to report conclusions in this section is α_{C2}=$3.64E-06$. Refer Appendix A for details on the Bonferroni corrected α_{C2} critical significance level.

Plots for recall, precision and F-score are provided for each ROUGE metric. The ROUGE scores for baseline summaries extracted by TextRank using the entire text document as input are plotted in yellow color and have a constant value for the two k values of 10 and 20. The ROUGE scores for summaries extracted after sentence pre-filtration will vary with the k value.

6.5.1 Summary of Results

The results are summarized based on the different methods collective performance on all ROUGE metrics for the two datasets for each of the three sentence pre-filtering granularities. The key findings about extractive summarization using thematic phrase based pre-filtering strategies are as follows:

(1) ThP-Fr-13 and ThP-Fr-3 are the best methods to use for thematic phrase based sentence pre-filtering for extractive summarization. They underperform the baseline TextRank the least of all the other methods for the PubMed dataset while they, along with TNG, outperform the baseline and all other methods for the USPTO dataset. Further, the F-scores for these three methods vary marginally for $k \in \{10, 20\}$ relative to AP.

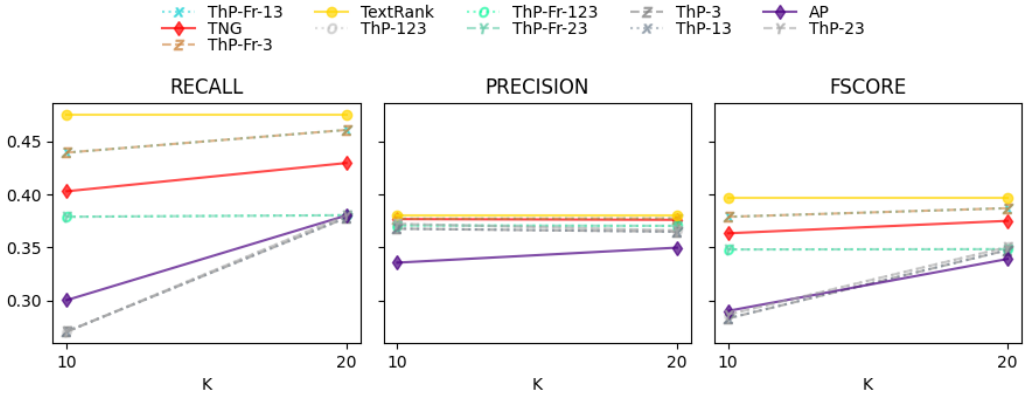
(2) Sentence pre-filtering using thematic sub-phrases and words benefits ThP-Fr-13, ThP-Fr-3 and TNG marginally, if at all. These two granular pre-filtering strategies help lift ROUGE scores for other thematic phrase extraction methods but they fail to achieve the ROUGE scores that the former three methods achieve using sentence pre-filtering based on thematic phrases.

(3) The TextRank baseline underperforms for the USPTO consistently on all ROUGE metrics indicating that summarization guided by thematic phrases is beneficial in the case of datasets that have relatively high average word occurrence frequencies with a coherent topicality for longer documents like patents.

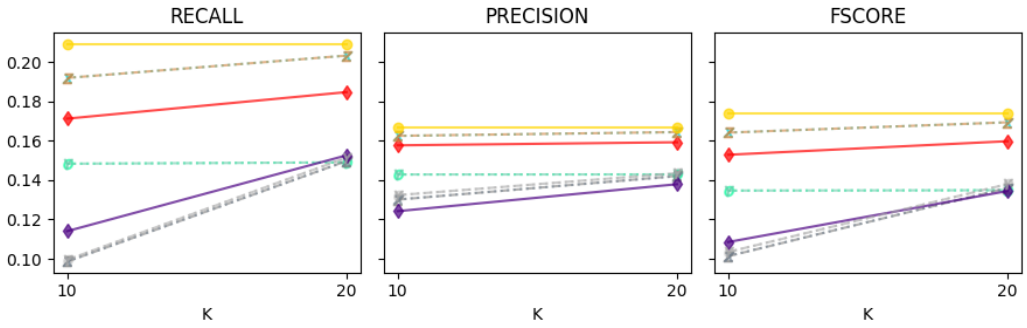
6.5.2 Sentence Pre-filtration using Thematic Phrases

The ROUGE metrics for extractive summaries based on sentence pre-filtration using thematic phrases are shown in Fig. 6.1 and 6.2 for the PubMed dataset and in Fig. 6.3 and 6.4 for the USPTO dataset. Fig. 6.1 and 6.3 show summarization quality measured using ROUGE-1, ROUGE-2 and ROUGE-3 for the respective datasets. Fig. 6.2 and 6.4

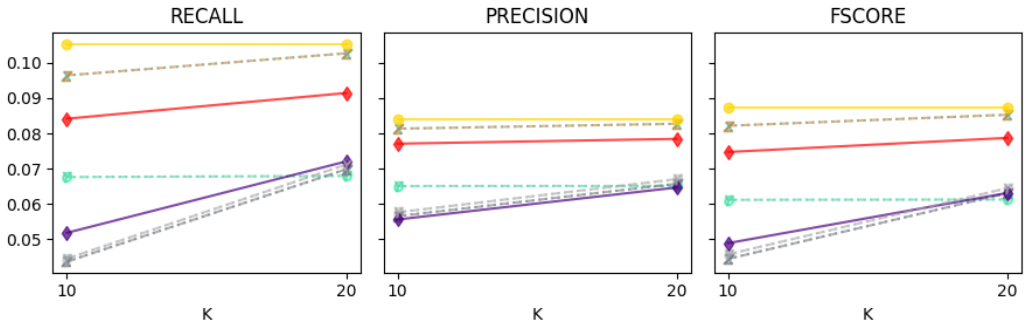
show summarization quality measured using ROUGE-L and ROUGE-SU4 for the respective datasets.



(a) ROUGE-1

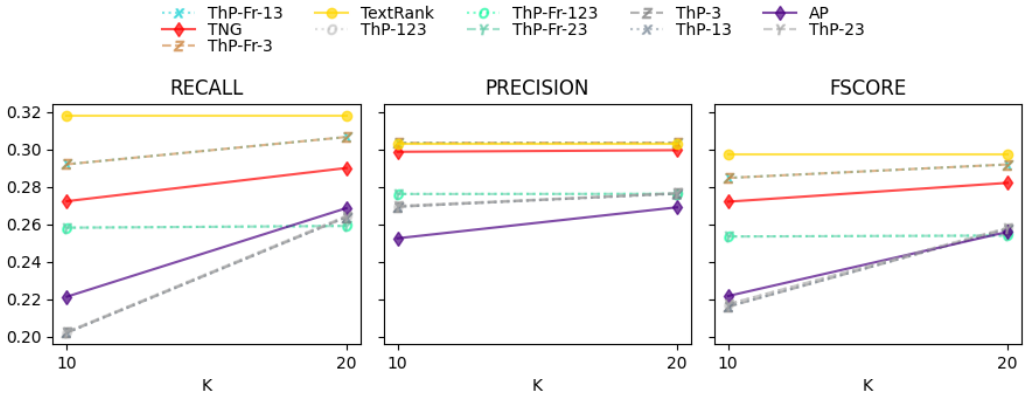


(b) ROUGE-2

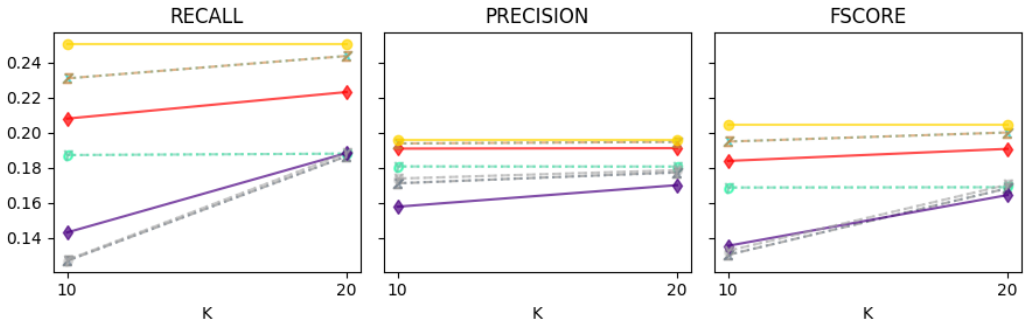


(c) ROUGE-3

FIG. 6.1: ROUGE-1, ROUGE-2 and ROUGE-3 Comparisons for the PUBMED_PMC_AB Dataset with Thematic Phrases Based Sentence Filtration as Input to TextRank Summarizer



(a) ROUGE-L



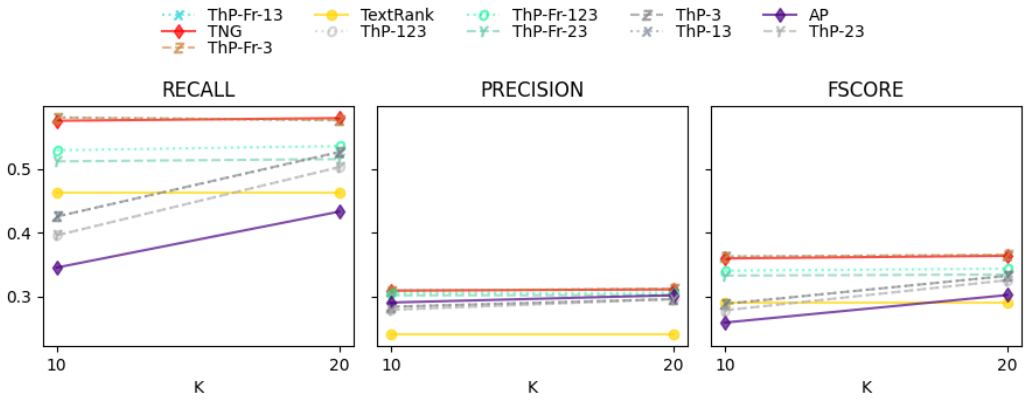
(b) ROUGE-SU4

FIG. 6.2: ROUGE-L and ROUGE-SU4 Comparisons for the PUBMED_PMC_AB Dataset with Thematic Phrases Based Sentence Filtration as Input to TextRank Summarizer

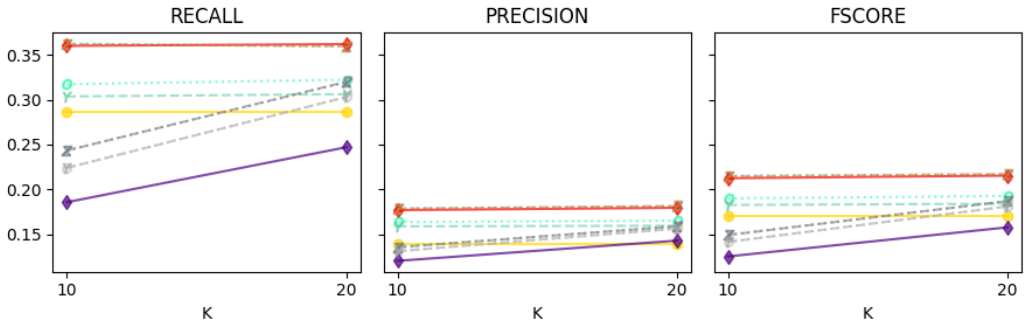
Fig. 6.1 and 6.2 show that the baseline TextRank summary outperforms all summaries extracted after thematic phrase based sentence pre-filtration on all ROUGE metrics for the PubMed dataset. At $k = 20$, ThP-Fr-13 and ThP-Fr-3 phrase pre-filtration summaries score the highest on all five ROUGE metrics of all the thematic phrase extraction methods ($P \leq \alpha_{C2}$).

Fig. 6.3 and 6.4 show that the baseline TextRank summary underperforms several thematic phrase pre-filtration summaries across all ROUGE metrics for the USPTO dataset. At $k = 20$, ThP-Fr-13, ThP-Fr-3 and TNG pre-filtration summaries are comparable and

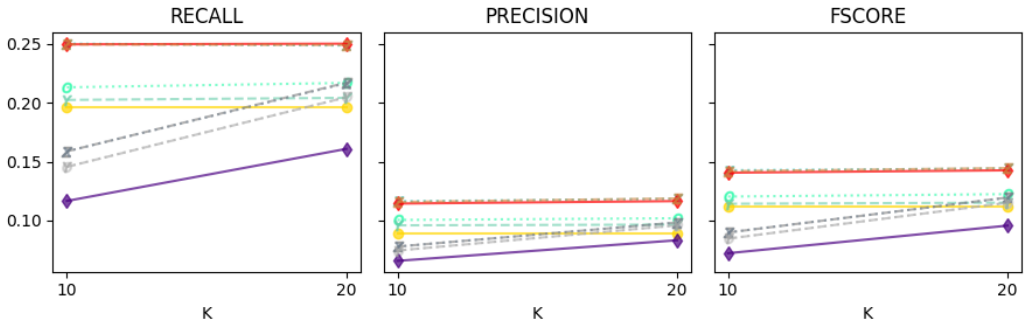
outperform all other thematic phrase pre-filtration summaries on ROUGE-1, -2, -3, -L and -SU4 ($P \leq \alpha_{C2}$).



(a) ROUGE-1

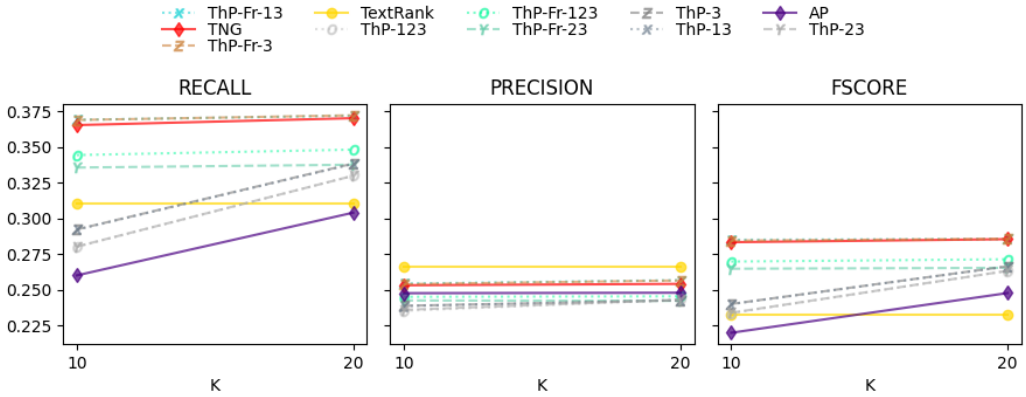


(b) ROUGE-2

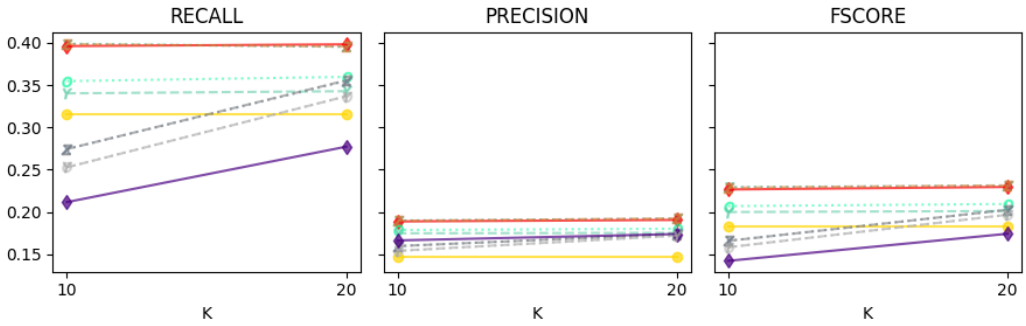


(c) ROUGE-3

FIG. 6.3: ROUGE-1, ROUGE-2 and ROUGE-3 Comparisons for the PATENT_RAND15K Dataset with Thematic Phrases Based Sentence Filtration as Input to TextRank Summarizer



(a) ROUGE-L



(b) ROUGE-SU4

FIG. 6.4: ROUGE-L and ROUGE-SU4 Comparisons for the PATENT_RAND15K Dataset with Thematic Phrases Based Sentence Filtration as Input to TextRank Summarizer

We can conclude that Baseline TextRank summaries outperform other methods for the PubMed dataset in which documents have relatively lower average word occurrence frequencies. It underperforms other methods for the USPTO dataset in which documents have relatively higher average word occurrence frequencies. ThP-Fr-13 and ThP-Fr-3 are the better method to extract thematic phrase pre-filtration summaries since they outperform other methods for the PubMed dataset and are only comparable with TNG for the USPTO dataset.

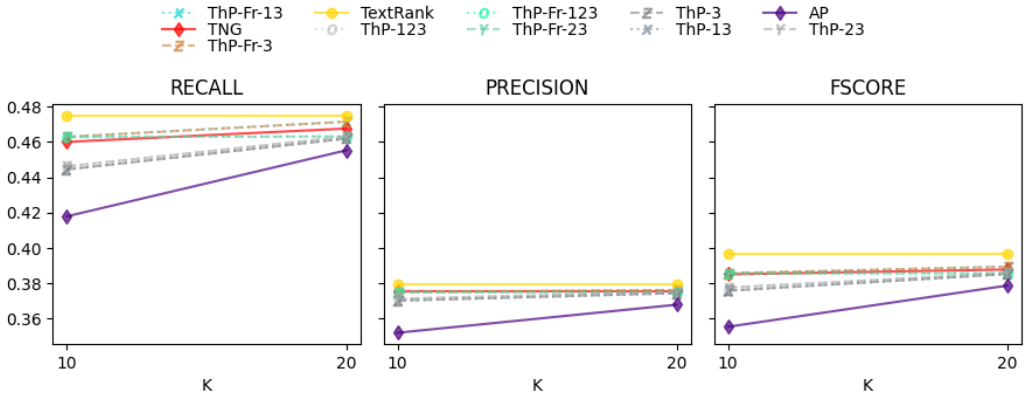
6.5.3 Sentence Pre-filtration Using Thematic Sub-phrases

The ROUGE metrics for extractive summaries based on sentence pre-filtration using thematic sub-phrases are shown in Fig. 6.5 and 6.6 for the PubMed dataset and in Fig. 6.7 and 6.8 for the USPTO dataset. Fig. 6.5 and 6.7 show summarization quality measured using ROUGE-1, ROUGE-2 and ROUGE-3 for the respective datasets. Fig. 6.6 and 6.8 show summarization quality measured using ROUGE-L and ROUGE-SU4 for the respective datasets.

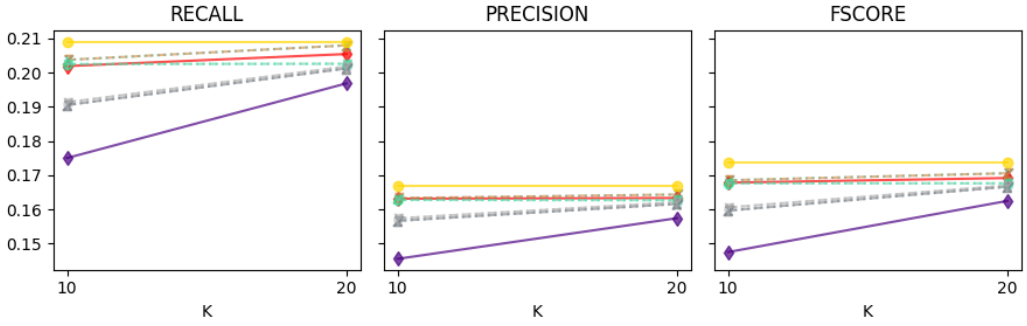
Fig. 6.5 and 6.6 show that the baseline TextRank summary outperforms all summaries extracted after thematic sub-phrases based sentence pre-filtration on all ROUGE metrics for the PubMed dataset. At $k = 20$, ThP-Fr-13, ThP-Fr-3 and TNG thematic sub-phrase pre-filtration summaries are comparable on ROUGE-L and outperform other thematic phrase extraction methods ($P \leq \alpha_{C2}$) whereas the former two outperform all other thematic phrase extraction methods on ROUGE-SU4 ($P \leq \alpha_{C2}$).

Fig. 6.7 and 6.8 show that the baseline TextRank summary underperforms all thematic sub-phrase pre-filtration summaries on all ROUGE metrics for the USPTO dataset. At $k = 20$, all ThP configurations and TNG thematic sub-phrase pre-filtration summaries are comparable on all ROUGE metrics and outperform the baseline TextRank and AP summaries ($P \leq \alpha_{C2}$).

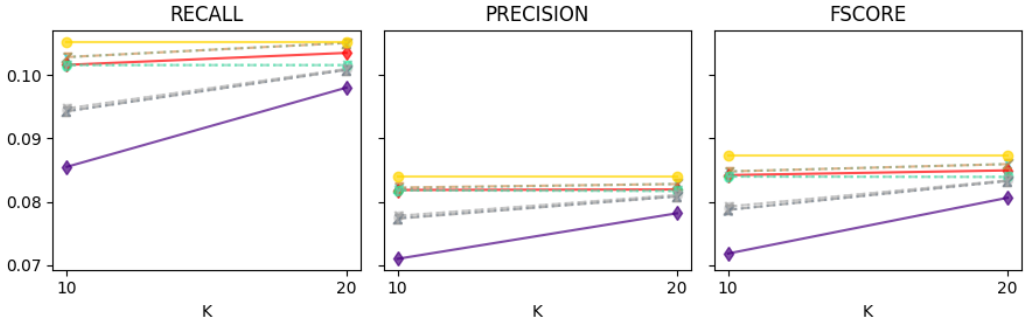
We can conclude that thematic sub-phrase based pre-filtration does not help extractive summaries outperform the baseline TextRank summaries for the PubMed dataset with



(a) ROUGE-1



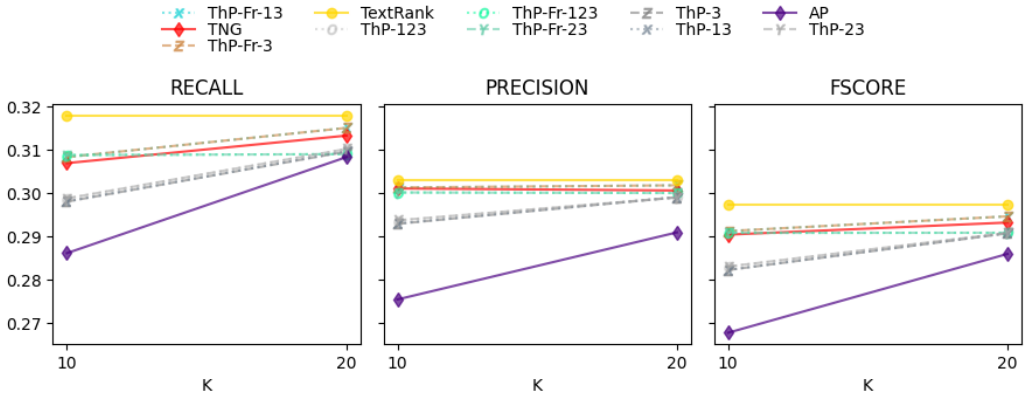
(b) ROUGE-2



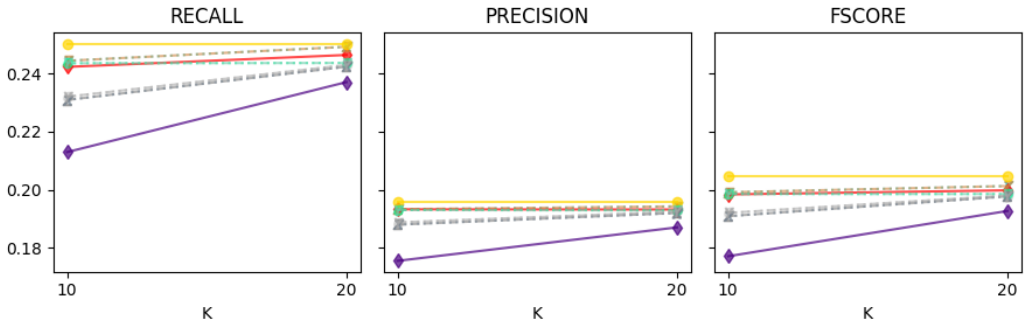
(c) ROUGE-3

FIG. 6.5: ROUGE-1, ROUGE-2 and ROUGE-3 Comparisons for the PUBMED_PMC_AB Dataset with Thematic Sub-phrases Based Sentence Filtration as Input to TextRank Summarizer

documents that have relatively lower average word occurrence frequencies. It does consistently help extractive summaries outperform the baseline for the USPTO dataset across k values for all thematic phrase extraction methods. Sentence pre-filtering using thematic



(a) ROUGE-L



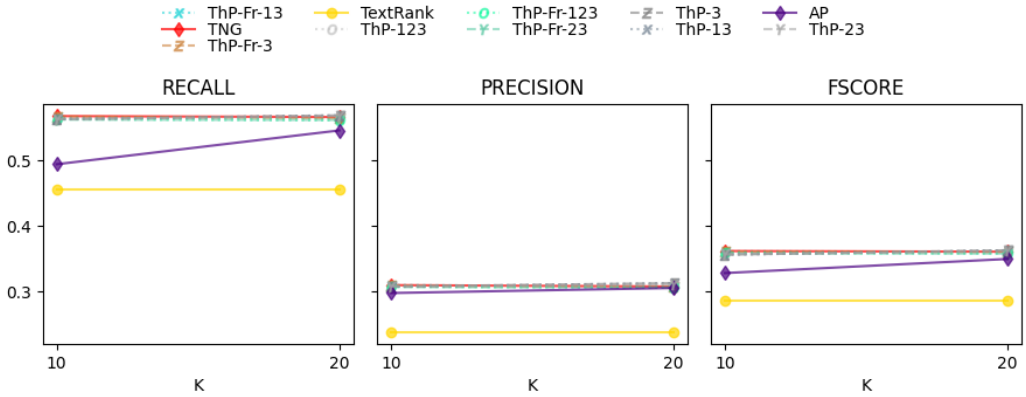
(b) ROUGE-SU4

FIG. 6.6: ROUGE-L and ROUGE-SU4 Comparisons for the PUBMED_PMC_AB Dataset with Thematic Sub-phrases Based Sentence Filtration as Input to TextRank Summarizer

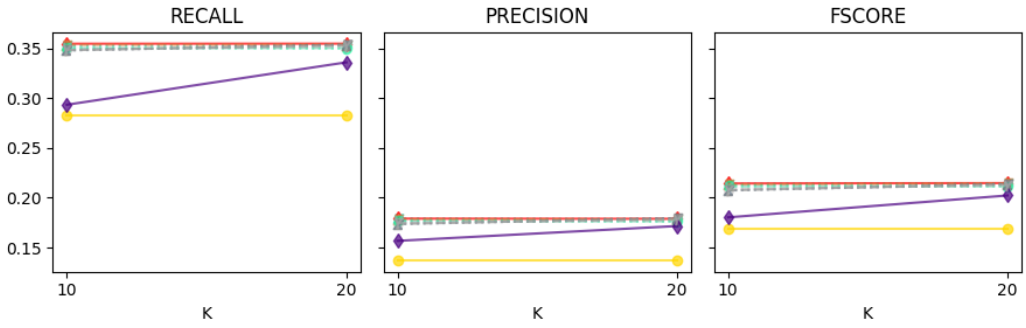
sub-phrases helps lift ROUGE scores marginally for ThP-Fr-13, ThP-Fr-3 and TNG while it benefits the other thematic phrase extraction methods most.

6.5.4 Sentence Pre-filtration Using Thematic Words

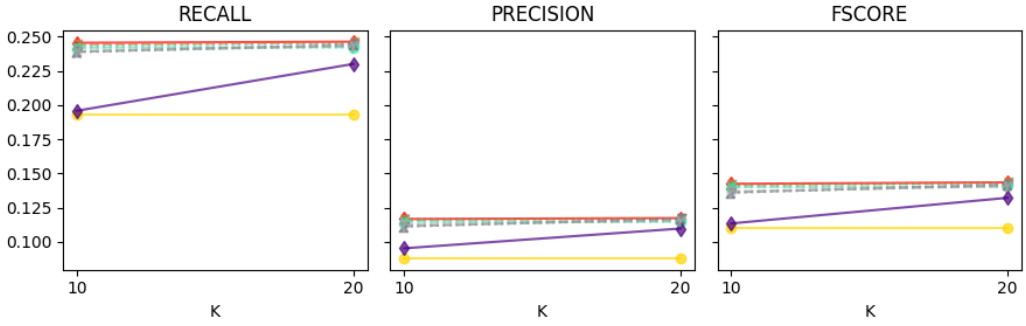
The ROUGE metrics for extractive summaries based on sentence pre-filtration using thematic words are shown in Fig. 6.9 and 6.10 for the PubMed dataset and Fig. 6.11 and 6.12 for the USPTO dataset. Fig. 6.9 and 6.11 show summarization quality measured using ROUGE-1, ROUGE-2 and ROUGE-3 for the respective datasets. Fig. 6.10 and 6.12



(a) ROUGE-1



(b) ROUGE-2

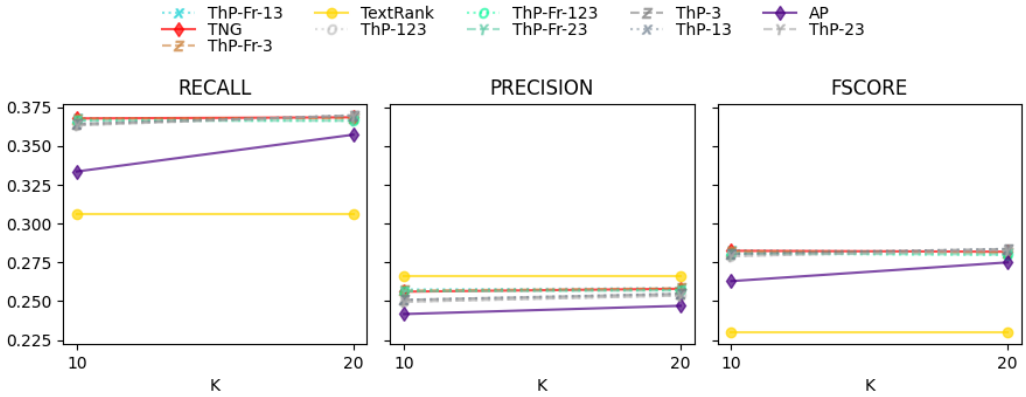


(c) ROUGE-3

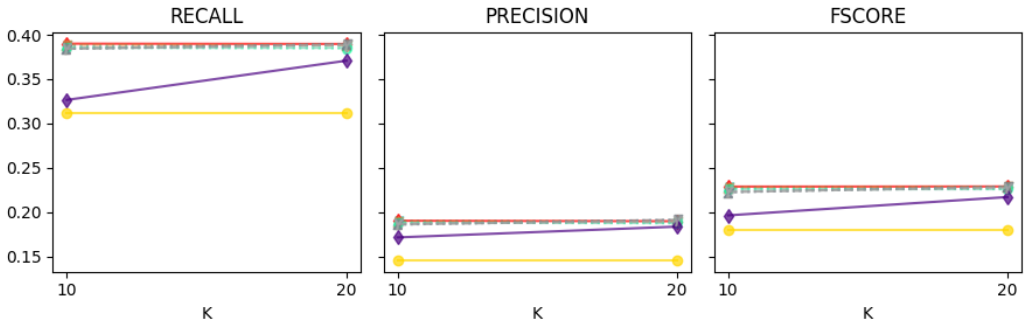
FIG. 6.7: ROUGE-1, ROUGE-2 and ROUGE-3 Comparisons for the PATENT_RAND15K Dataset with Thematic Sub-phrases Based Sentence Filtration as Input to TextRank Summarizer

show summarization quality measured using ROUGE-L and ROUGE-SU4 for the respective datasets.

Fig. 6.9 and 6.10 show that the baseline TextRank summary outperforms all sum-



(a) ROUGE-L

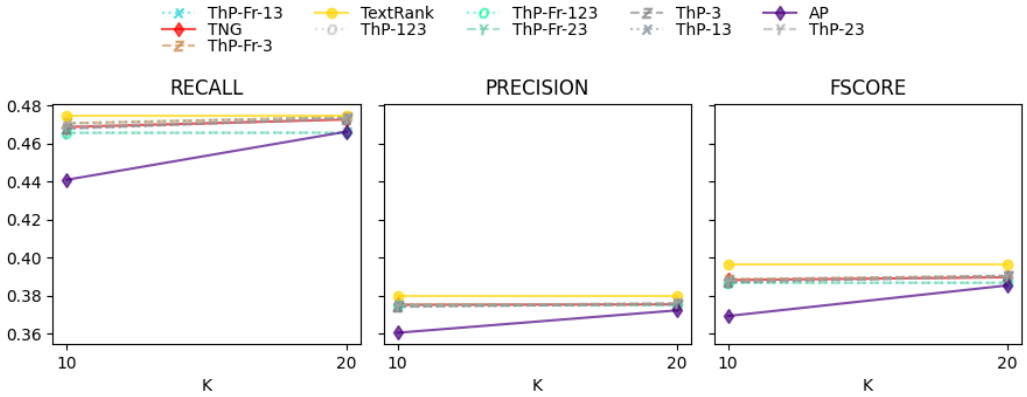


(b) ROUGE-SU4

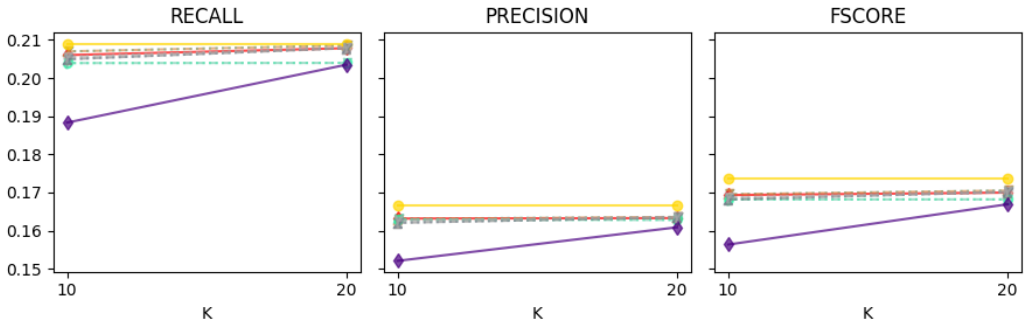
FIG. 6.8: ROUGE-L and ROUGE-SU4 Comparisons for the PATENT_RAND15K Dataset with Thematic Sub-phrases Based Sentence Filtration as Input to TextRank Summarizer

maries extracted after thematic word based sentence pre-filtration on all ROUGE metrics for the PubMed dataset. At $k = 20$, thematic word pre-filtration summaries for ThP-* configurations, ThP-Fr-13, ThP-Fr-3 and TNG have comparable ROUGE scores and outperform other thematic phrase extraction methods.

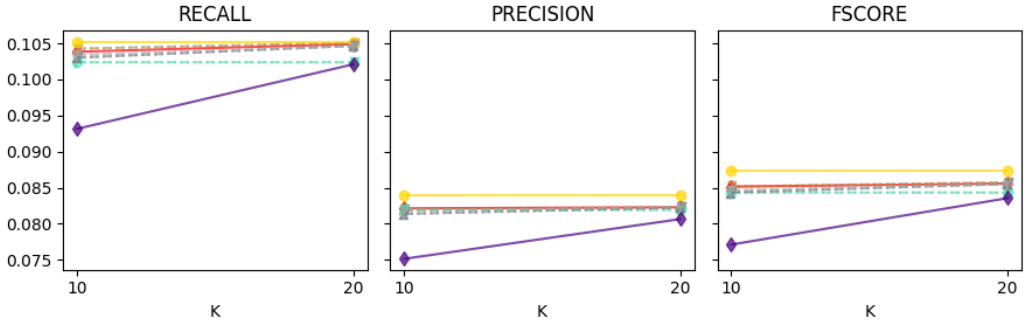
Fig. 6.11 and 6.12 show that the baseline TextRank summary underperforms all thematic word pre-filtration summaries on all ROUGE metrics for the USPTO dataset. At $k = 20$, thematic word pre-filtration summaries for all thematic phrase extraction methods are comparable on all ROUGE metrics.



(a) ROUGE-1



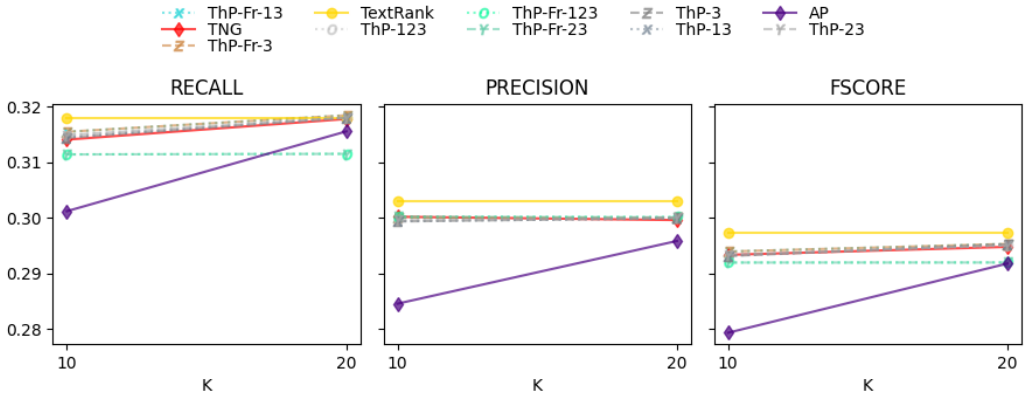
(b) ROUGE-2



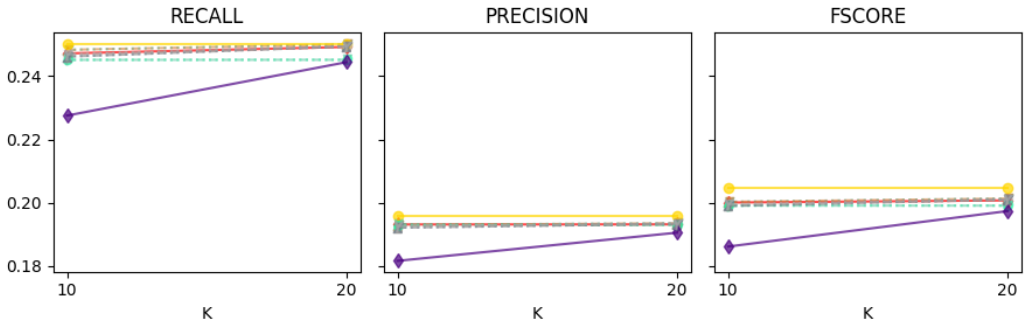
(c) ROUGE-3

FIG. 6.9: ROUGE-1, ROUGE-2 and ROUGE-3 Comparisons for the PUBMED_PMC_AB Dataset with Thematic Words Based Sentence Filtration as Input to TextRank Summarizer

We can conclude that thematic word based sentence pre-filtration helps extractive summaries perform relatively closer to the baseline TextRank summaries for the PubMed dataset. It consistently helps extractive summaries outperform the baseline for the USPTO



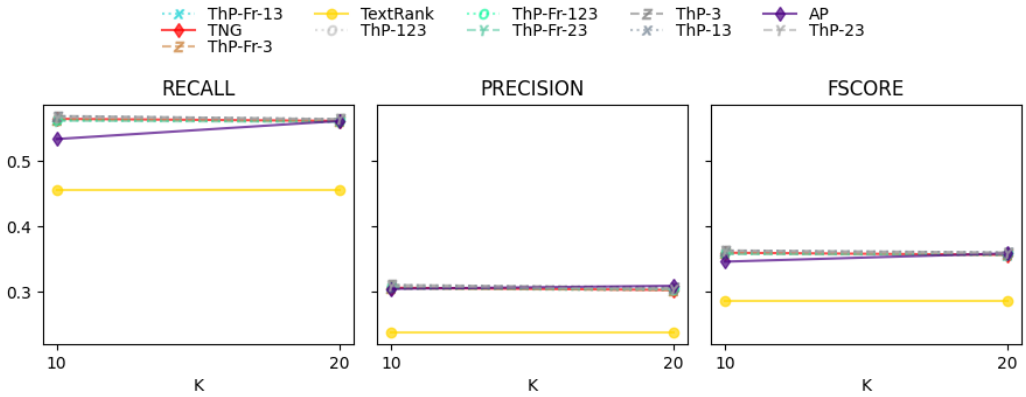
(a) ROUGE-L



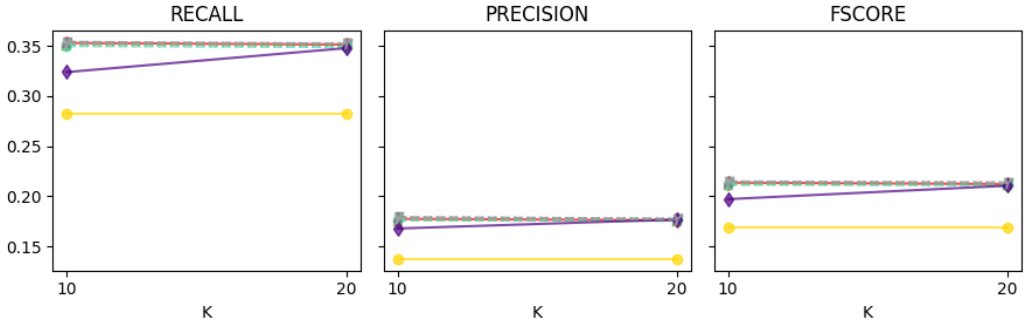
(b) ROUGE-SU4

FIG. 6.10: ROUGE-L and ROUGE-SU4 Comparisons for the PUBMED_PMC_AB Dataset with Thematic Words Based Sentence Filtration as Input to TextRank Summarizer

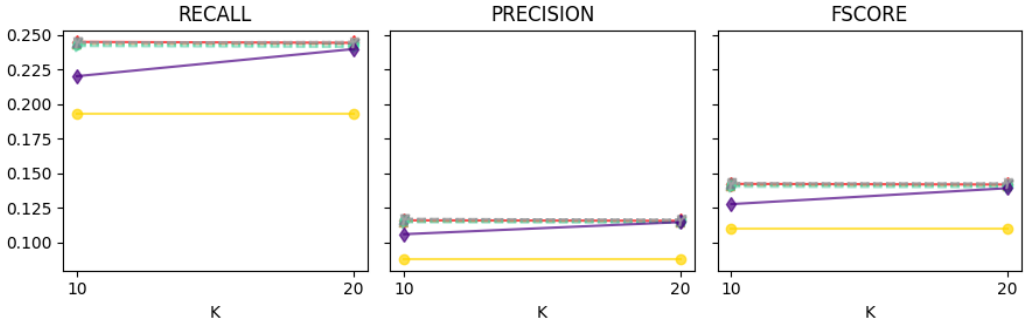
dataset at all k values across all thematic phrase extraction methods. Sentence pre-filtering using thematic words helps most, if not all, thematic phrase extraction methods perform comparably on all ROUGE metrics. Their absolute ROUGE scores for the USPTO dataset, however, do not achieve levels observed for thematic phrases based sentence pre-filtering.



(a) ROUGE-1

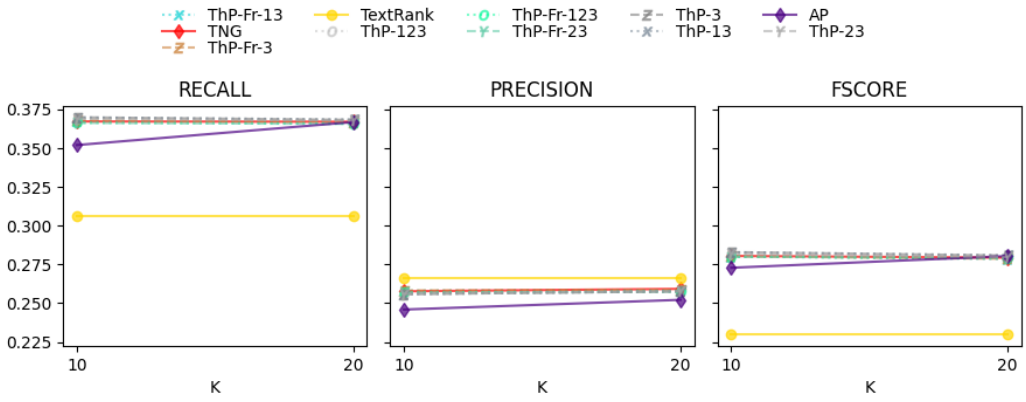


(b) ROUGE-2

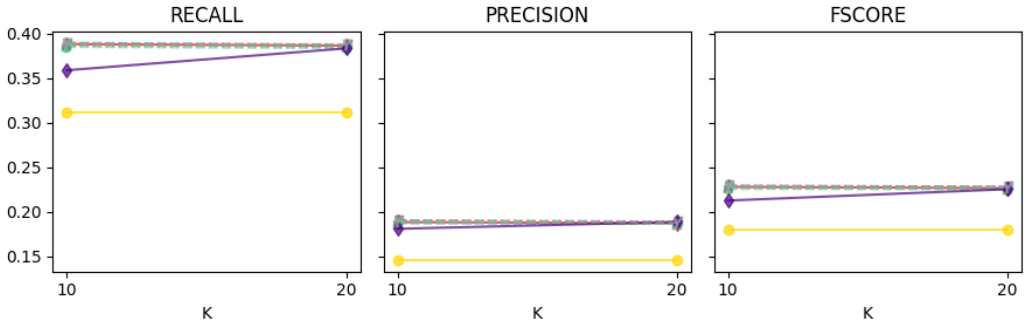


(c) ROUGE-3

FIG. 6.11: ROUGE-1, ROUGE-2 and ROUGE-3 Comparisons for the PATENT_RAND15K Dataset with Thematic Words Based Sentence Filtration as Input to TextRank Summarizer



(a) ROUGE-L



(b) ROUGE-SU4

FIG. 6.12: ROUGE-L and ROUGE-SU4 Comparisons for the PATENT_RAND15K Dataset with Thematic Words Based Sentence Filtration as Input to TextRank Summarizer

Chapter 7

CONCLUSION

This work described ThemaPhrase, a novel framework for the task of unsupervised extraction of thematic phrases from single text artifacts. It is evaluated alongside TNG and AutoPhrase for this task. Two datasets, a subset of PubMed publications and a subset of USPTO patents, with varying token frequency distributions have been created to quantitatively evaluate thematic phrase extraction methods. The titles and abstracts of documents in these datasets have been assessed to serve as reference gold standards for thematic phrases while the latter can serve as reference summaries to evaluate extractive summarization.

ThemaPhrase configurations are more robust to varying ratios of topics to document partitions and average token occurrence frequencies of a document than competing methods. ThP-Fr-13 and ThP-Fr-3 outperform all other methods in extracting thematic phrases that represent themes at the granularity at which they are represented in both document abstracts and titles at $k = 20$ for both the datasets. Further, they are also the best methods in extracting thematic phrases that are finer-grained representations of themes in document abstracts for

the USPTO dataset at $k \in \{10, 20\}$. ThP-Fr-123 and ThP-Fr-23 outperform all other methods in extracting thematic phrases that are coarser-grained representations of themes in both document abstracts and titles across k values for the USPTO dataset.

The DCG analyses show that ThP-Fr-13 and ThP-Fr-3 add more thematically relevant phrases than TNG as k increases. The analyses further shows that the ThP scoring method produces more false positives at lower k relative to TNG and needs to be refined.

Sentence pre-filtering based on thematic phrases at various granularities helps improve extractive summarization for texts, such as patents, that have relatively higher average token occurrence frequencies where the baseline TextRank summarizer underperforms. ThP-Fr-13 and ThP-Fr-13 outperform all other thematic phrase extraction methods and provide the most lift in summary quality in terms of ROUGE metrics for the USPTO dataset using thematic phrase based sentence pre-filtration. Further, these two configurations underperform TextRank by the least amount of all other methods for the PubMed dataset. This makes these two the better choices for extractive summarization that is agnostic of the verbosity and presence of repetitive verbiage in text that needs to be summarized.

Future Work

Improving the WSEQ heuristic with a probability distribution over relative word positions in candidate phrases may help alleviate the overly aggressive rejection of candidate phrases by this heuristic. Combining WSEQ and WPOS into a single filter and integrating them to augment thematic phrase scoring will help rank the extracted thematic phrases better.

The motivation to refine phrase scoring is based on observations of more relevant thematic phrases being observed at higher k values. The DCG analyses made this ranking behavior evident.

The ThemaPhrase framework can be used to build an evolving thematic hierarchy using per-document thematic phrases. This will find application in evolving document tagging, taxonomy generation and granular discourse analysis. Further, it will enable a bottom up paradigm to build thematic hierarchies that can allow inference of cross-domain thematic associations.

Lastly, utilizing extracted thematic phrases as thematic representations for improving text segmentation will be helpful in areas such as topic change detection, discourse analysis and story chaining.

Appendix A

STATISTICAL SIGNIFICANCE TESTS

All statistical significance levels reported in this work are done using p-values of the Student's t-test [68]. The statistical significance of a comparison is reported only when the p-values of both a Student's t-test and a Wilcoxon signed-rank test [69] are in agreement by being less than or equal to a critical significance level. Both tests are conducted as two-sided, paired tests. The Wilcoxon signed-rank tests are conducted by including zero-differences, if any, in the ranking and then splitting the rank of the zero-differences between the two samples.

The number of comparisons for different evaluations are large due to several methods being compared at any given k value on multiple combinations of quantitative metrics and gold standards. Since multiple hypothesis tests are being performed, the critical significance level is obtained by applying the BonFerroni correction [70, 71] to a pre-specified significance level of α=0.001 using the formula α_C=$\alpha \div$ *# Comparisons*.

The critical significance levels for the different evaluations in this work are as provided

in Tab. A.1. There are two distinct critical significance values that are obtained after Bonferronin correction. The first is α_{C1}=$1.39E{-}06$ that applies to statistical significance levels reported for the quantitative evaluation of thematic phrase extraction methods in 5.5.2 and 5.5.3. The second is α_{C2}=$3.64E{-}06$ that applies to statistical significance levels reported for the quantitative evaluation of extracted summaries in 6.5.2, 6.5.3 and 6.5.4.

Evaluation	# Comparisons	Critical $\boldsymbol{\alpha}$
Thematic Phrases Quantitative Analysis w/ Abstracts as Gold Standard (Applies to Sec. 5.5.2)	720	1.39E-06
Thematic Phrases Quantitative Analysis w/ Titles as Gold Standard (Applies to Sec. 5.5.3)	720	1.39E-06
Thematic Phrase Pre-filtered Summarization ROUGE Analysis (Applies to Sec. 6.5.2)	275	3.64E-06
Thematic Subphrase Pre-filtered Summarization ROUGE Analysis (Applies to Sec. 6.5.3)	275	3.64E-06
Thematic Word Pre-filtered Summarization ROUGE Analysis (Applies to Sec. 6.5.4)	275	3.64E-06

Table A.1: Critical Significance Levels for Evaluations After Applying Bonferroni Correction on $\alpha = 0.001$

Appendix B

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : ABSTRACT PH-COV AND PH-FMI

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0421	0.0164	0.0227	0.0144
TNG-5	0.1057	0.0538	0.1832	0.1488
ThP-123-5	0.0204	0.0080	0.0233	0.0141
ThP-13-5	0.0219	0.0086	0.0271	0.0162
ThP-23-5	0.0204	0.0080	0.0233	0.0141
ThP-3-5	0.0219	0.0087	0.0274	0.0164
ThP-Fr-123-5	0.0720	0.0299	0.1046	0.0698
ThP-Fr-13-5	0.0660	0.0339	0.1437	0.1084
ThP-Fr-23-5	0.0720	0.0299	0.0929	0.0598
ThP-Fr-3-5	0.0660	0.0339	0.1454	0.1104
AP-10	0.0594	0.0335	0.0316	0.0285
TNG-10	0.1128	0.0745	0.1956	0.2094
ThP-123-10	0.0312	0.0171	0.0387	0.0339
ThP-13-10	0.0310	0.0170	0.0429	0.0371
ThP-23-10	0.0312	0.0171	0.0387	0.0339
ThP-3-10	0.0310	0.0170	0.0426	0.0369
ThP-Fr-123-10	0.0795	0.0381	0.1150	0.0973
ThP-Fr-13-10	0.0982	0.0736	0.2100	0.2249
ThP-Fr-23-10	0.0795	0.0381	0.0969	0.0774
ThP-Fr-3-10	0.0982	0.0736	0.2134	0.2312
AP-20	0.0854	0.0701	0.0437	0.0558
TNG-20	0.1153	0.0999	0.1898	0.2692
ThP-123-20	0.0510	0.0396	0.0642	0.0796
ThP-13-20	0.0507	0.0393	0.0709	0.0886
ThP-23-20	0.0510	0.0396	0.0642	0.0796
ThP-3-20	0.0507	0.0394	0.0710	0.0887
ThP-Fr-123-20	0.0801	0.0391	0.1142	0.1065
ThP-Fr-13-20	0.1278	0.1303	0.2515	0.3610
ThP-Fr-23-20	0.0801	0.0391	0.0954	0.0808
ThP-Fr-3-20	0.1278	0.1303	0.2538	0.3707

Table B.1: Methods Evaluation : phCOV and phFMI With Abstract as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0410	0.0158	0.0223	0.0142
TNG-5	0.1133	0.0573	0.2110	0.1688
ThP-123-5	0.0201	0.0079	0.0227	0.0136
ThP-13-5	0.0216	0.0086	0.0279	0.0165
ThP-23-5	0.0201	0.0079	0.0227	0.0136
ThP-3-5	0.0216	0.0086	0.0279	0.0166
ThP-Fr-123-5	0.0715	0.0298	0.1034	0.0692
ThP-Fr-13-5	0.0668	0.0338	0.1430	0.1074
ThP-Fr-23-5	0.0715	0.0298	0.0918	0.0594
ThP-Fr-3-5	0.0668	0.0338	0.1448	0.1096
AP-10	0.0579	0.0323	0.0310	0.0280
TNG-10	0.1225	0.0801	0.2224	0.2325
ThP-123-10	0.0304	0.0166	0.0375	0.0326
ThP-13-10	0.0308	0.0168	0.0425	0.0364
ThP-23-10	0.0304	0.0166	0.0375	0.0326
ThP-3-10	0.0309	0.0169	0.0425	0.0364
ThP-Fr-123-10	0.0799	0.0391	0.1144	0.0978
ThP-Fr-13-10	0.0981	0.0726	0.2084	0.2220
ThP-Fr-23-10	0.0799	0.0391	0.0966	0.0781
ThP-Fr-3-10	0.0981	0.0726	0.2115	0.2284
AP-20	0.0842	0.0686	0.0433	0.0551
TNG-20	0.1228	0.1054	0.2112	0.2922
ThP-123-20	0.0502	0.0387	0.0624	0.0775
ThP-13-20	0.0501	0.0388	0.0704	0.0873
ThP-23-20	0.0502	0.0387	0.0624	0.0775
ThP-3-20	0.0502	0.0389	0.0704	0.0873
ThP-Fr-123-20	0.0807	0.0404	0.1138	0.1081
ThP-Fr-13-20	0.1280	0.1300	0.2516	0.3611
ThP-Fr-23-20	0.0807	0.0404	0.0949	0.0819
ThP-Fr-3-20	0.1280	0.1300	0.2539	0.3708

Table B.2: Methods Evaluation : phCOV and phFMI With Abstract as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0410	0.0158	0.0220	0.0140
TNG-5	0.1160	0.0586	0.2208	0.1757
ThP-123-5	0.0202	0.0080	0.0221	0.0132
ThP-13-5	0.0224	0.0088	0.0272	0.0161
ThP-23-5	0.0202	0.0080	0.0221	0.0132
ThP-3-5	0.0224	0.0088	0.0276	0.0164
ThP-Fr-123-5	0.0710	0.0297	0.1029	0.0688
ThP-Fr-13-5	0.0670	0.0337	0.1422	0.1067
ThP-Fr-23-5	0.0710	0.0297	0.0914	0.0589
ThP-Fr-3-5	0.0670	0.0337	0.1440	0.1088
AP-10	0.0578	0.0321	0.0307	0.0278
TNG-10	0.1252	0.0815	0.2323	0.2412
ThP-123-10	0.0302	0.0164	0.0368	0.0322
ThP-13-10	0.0311	0.0169	0.0417	0.0356
ThP-23-10	0.0302	0.0164	0.0368	0.0322
ThP-3-10	0.0311	0.0169	0.0417	0.0356
ThP-Fr-123-10	0.0798	0.0394	0.1140	0.0976
ThP-Fr-13-10	0.0980	0.0722	0.2073	0.2207
ThP-Fr-23-10	0.0798	0.0394	0.0965	0.0784
ThP-Fr-3-10	0.0980	0.0722	0.2102	0.2265
AP-20	0.0840	0.0682	0.0429	0.0549
TNG-20	0.1258	0.1075	0.2201	0.3016
ThP-123-20	0.0497	0.0381	0.0614	0.0759
ThP-13-20	0.0501	0.0384	0.0696	0.0864
ThP-23-20	0.0497	0.0381	0.0614	0.0759
ThP-3-20	0.0501	0.0384	0.0696	0.0862
ThP-Fr-123-20	0.0806	0.0409	0.1136	0.1087
ThP-Fr-13-20	0.1281	0.1300	0.2512	0.3604
ThP-Fr-23-20	0.0806	0.0409	0.0948	0.0825
ThP-Fr-3-20	0.1281	0.1300	0.2535	0.3699

Table B.3: Methods Evaluation : phCOV and phFMI With Abstract as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0409	0.0157	0.0215	0.0135
TNG-5	0.1165	0.0587	0.2255	0.1794
ThP-123-5	0.0195	0.0076	0.0218	0.0130
ThP-13-5	0.0218	0.0086	0.0264	0.0156
ThP-23-5	0.0195	0.0076	0.0218	0.0130
ThP-3-5	0.0219	0.0086	0.0266	0.0158
ThP-Fr-123-5	0.0705	0.0294	0.1030	0.0687
ThP-Fr-13-5	0.0676	0.0339	0.1423	0.1062
ThP-Fr-23-5	0.0705	0.0294	0.0915	0.0591
ThP-Fr-3-5	0.0676	0.0339	0.1440	0.1083
AP-10	0.0576	0.0320	0.0302	0.0272
TNG-10	0.1258	0.0819	0.2374	0.2457
ThP-123-10	0.0295	0.0159	0.0362	0.0311
ThP-13-10	0.0310	0.0168	0.0414	0.0352
ThP-23-10	0.0295	0.0159	0.0362	0.0311
ThP-3-10	0.0309	0.0168	0.0414	0.0351
ThP-Fr-123-10	0.0798	0.0397	0.1140	0.0980
ThP-Fr-13-10	0.0979	0.0720	0.2072	0.2202
ThP-Fr-23-10	0.0798	0.0397	0.0964	0.0786
ThP-Fr-3-10	0.0979	0.0720	0.2102	0.2262
AP-20	0.0839	0.0680	0.0425	0.0538
TNG-20	0.1266	0.1082	0.2246	0.3061
ThP-123-20	0.0489	0.0374	0.0604	0.0745
ThP-13-20	0.0498	0.0381	0.0692	0.0853
ThP-23-20	0.0489	0.0374	0.0604	0.0745
ThP-3-20	0.0499	0.0382	0.0693	0.0854
ThP-Fr-123-20	0.0808	0.0413	0.1135	0.1095
ThP-Fr-13-20	0.1283	0.1299	0.2521	0.3612
ThP-Fr-23-20	0.0808	0.0413	0.0947	0.0830
ThP-Fr-3-20	0.1283	0.1299	0.2543	0.3707

Table B.4: Methods Evaluation : phCOV and phFMI With Abstract as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0406	0.0155	0.0212	0.0134
TNG-5	0.1170	0.0590	0.2282	0.1812
ThP-123-5	0.0197	0.0077	0.0213	0.0127
ThP-13-5	0.0223	0.0087	0.0267	0.0158
ThP-23-5	0.0197	0.0077	0.0213	0.0127
ThP-3-5	0.0222	0.0087	0.0268	0.0159
ThP-Fr-123-5	0.0708	0.0296	0.1021	0.0682
ThP-Fr-13-5	0.0673	0.0336	0.1431	0.1069
ThP-Fr-23-5	0.0708	0.0296	0.0912	0.0589
ThP-Fr-3-5	0.0673	0.0336	0.1451	0.1092
AP-10	0.0573	0.0317	0.0299	0.0271
TNG-10	0.1268	0.0824	0.2402	0.2478
ThP-123-10	0.0295	0.0158	0.0356	0.0306
ThP-13-10	0.0312	0.0169	0.0414	0.0351
ThP-23-10	0.0295	0.0158	0.0356	0.0306
ThP-3-10	0.0312	0.0169	0.0413	0.0349
ThP-Fr-123-10	0.0799	0.0399	0.1138	0.0979
ThP-Fr-13-10	0.0980	0.0718	0.2067	0.2197
ThP-Fr-23-10	0.0799	0.0399	0.0963	0.0787
ThP-Fr-3-10	0.0980	0.0718	0.2101	0.2258
AP-20	0.0835	0.0675	0.0426	0.0544
TNG-20	0.1281	0.1091	0.2269	0.3083
ThP-123-20	0.0487	0.0371	0.0596	0.0738
ThP-13-20	0.0500	0.0382	0.0685	0.0847
ThP-23-20	0.0487	0.0371	0.0596	0.0738
ThP-3-20	0.0500	0.0383	0.0686	0.0848
ThP-Fr-123-20	0.0809	0.0416	0.1134	0.1097
ThP-Fr-13-20	0.1284	0.1299	0.2518	0.3610
ThP-Fr-23-20	0.0809	0.0416	0.0946	0.0832
ThP-Fr-3-20	0.1284	0.1299	0.2544	0.3709

Table B.5: Methods Evaluation : phCOV and phFMI With Abstract as Gold Standard (Segment Count = 25)

Appendix C

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : ABSTRACT SUB-COV AND SUB-FMI

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0857	0.0693	0.0374	0.0409
TNG-5	0.1904	0.1687	0.2050	0.2102
ThP-123-5	0.0262	0.0165	0.0248	0.0226
ThP-13-5	0.0285	0.0174	0.0294	0.0250
ThP-23-5	0.0262	0.0165	0.0248	0.0226
ThP-3-5	0.0285	0.0174	0.0297	0.0253
ThP-Fr-123-5	0.1623	0.1401	0.2301	0.2689
ThP-Fr-13-5	0.0737	0.0469	0.1385	0.1299
ThP-Fr-23-5	0.1623	0.1401	0.2275	0.2630
ThP-Fr-3-5	0.0737	0.0469	0.1403	0.1330
AP-10	0.1407	0.1645	0.0556	0.0849
TNG-10	0.2131	0.2325	0.2263	0.2881
ThP-123-10	0.0399	0.0339	0.0400	0.0492
ThP-13-10	0.0400	0.0323	0.0457	0.0529
ThP-23-10	0.0399	0.0339	0.0400	0.0492
ThP-3-10	0.0400	0.0323	0.0455	0.0527
ThP-Fr-123-10	0.2025	0.1868	0.3238	0.4066
ThP-Fr-13-10	0.1135	0.1052	0.2102	0.2655
ThP-Fr-23-10	0.2025	0.1868	0.3080	0.3823
ThP-Fr-3-10	0.1135	0.1052	0.2147	0.2747
AP-20	0.2385	0.3575	0.0857	0.1728
TNG-20	0.2250	0.3034	0.2269	0.3660
ThP-123-20	0.0655	0.0738	0.0687	0.1112
ThP-13-20	0.0650	0.0688	0.0759	0.1173
ThP-23-20	0.0655	0.0738	0.0687	0.1112
ThP-3-20	0.0650	0.0687	0.0760	0.1173
ThP-Fr-123-20	0.2076	0.1928	0.3496	0.4508
ThP-Fr-13-20	0.1580	0.1931	0.2736	0.4299
ThP-Fr-23-20	0.2076	0.1928	0.3228	0.4085
ThP-Fr-3-20	0.1580	0.1931	0.2790	0.4436

Table C.1: Methods Evaluation : subCOV and subFMI With Abstract as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0837	0.0675	0.0367	0.0401
TNG-5	0.2046	0.1778	0.2351	0.2326
ThP-123-5	0.0261	0.0163	0.0241	0.0217
ThP-13-5	0.0282	0.0171	0.0298	0.0252
ThP-23-5	0.0261	0.0163	0.0241	0.0217
ThP-3-5	0.0282	0.0171	0.0298	0.0252
ThP-Fr-123-5	0.1605	0.1387	0.2249	0.2643
ThP-Fr-13-5	0.0744	0.0467	0.1378	0.1289
ThP-Fr-23-5	0.1605	0.1387	0.2222	0.2587
ThP-Fr-3-5	0.0744	0.0467	0.1396	0.1321
AP-10	0.1374	0.1604	0.0557	0.0847
TNG-10	0.2306	0.2440	0.2582	0.3148
ThP-123-10	0.0389	0.0323	0.0390	0.0475
ThP-13-10	0.0398	0.0316	0.0453	0.0521
ThP-23-10	0.0389	0.0323	0.0390	0.0475
ThP-3-10	0.0398	0.0316	0.0453	0.0519
ThP-Fr-123-10	0.2050	0.1904	0.3202	0.4063
ThP-Fr-13-10	0.1132	0.1032	0.2079	0.2622
ThP-Fr-23-10	0.2050	0.1904	0.3058	0.3835
ThP-Fr-3-10	0.1132	0.1032	0.2124	0.2713
AP-20	0.2365	0.3546	0.0855	0.1730
TNG-20	0.2401	0.3154	0.2544	0.3907
ThP-123-20	0.0646	0.0724	0.0664	0.1073
ThP-13-20	0.0645	0.0676	0.0752	0.1152
ThP-23-20	0.0646	0.0724	0.0664	0.1073
ThP-3-20	0.0645	0.0677	0.0751	0.1152
ThP-Fr-123-20	0.2114	0.1981	0.3490	0.4551
ThP-Fr-13-20	0.1582	0.1924	0.2731	0.4293
ThP-Fr-23-20	0.2114	0.1981	0.3223	0.4124
ThP-Fr-3-20	0.1582	0.1924	0.2786	0.4431

Table C.2: Methods Evaluation : subCOV and subFMI With Abstract as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0833	0.0670	0.0367	0.0405
TNG-5	0.2082	0.1795	0.2466	0.2415
ThP-123-5	0.0261	0.0164	0.0236	0.0212
ThP-13-5	0.0292	0.0175	0.0297	0.0250
ThP-23-5	0.0261	0.0164	0.0236	0.0212
ThP-3-5	0.0291	0.0175	0.0299	0.0252
ThP-Fr-123-5	0.1587	0.1377	0.2221	0.2614
ThP-Fr-13-5	0.0749	0.0464	0.1371	0.1283
ThP-Fr-23-5	0.1587	0.1377	0.2193	0.2552
ThP-Fr-3-5	0.0749	0.0464	0.1386	0.1310
AP-10	0.1372	0.1607	0.0547	0.0826
TNG-10	0.2352	0.2459	0.2702	0.3235
ThP-123-10	0.0385	0.0319	0.0383	0.0470
ThP-13-10	0.0402	0.0316	0.0451	0.0515
ThP-23-10	0.0385	0.0319	0.0383	0.0470
ThP-3-10	0.0402	0.0317	0.0451	0.0514
ThP-Fr-123-10	0.2055	0.1918	0.3177	0.4054
ThP-Fr-13-10	0.1131	0.1024	0.2067	0.2603
ThP-Fr-23-10	0.2055	0.1918	0.3038	0.3830
ThP-Fr-3-10	0.1131	0.1024	0.2108	0.2686
AP-20	0.2351	0.3518	0.0847	0.1724
TNG-20	0.2448	0.3166	0.2663	0.4012
ThP-123-20	0.0642	0.0717	0.0655	0.1065
ThP-13-20	0.0645	0.0667	0.0747	0.1147
ThP-23-20	0.0642	0.0717	0.0655	0.1065
ThP-3-20	0.0645	0.0667	0.0746	0.1145
ThP-Fr-123-20	0.2126	0.2005	0.3483	0.4574
ThP-Fr-13-20	0.1582	0.1919	0.2724	0.4285
ThP-Fr-23-20	0.2126	0.2005	0.3217	0.4146
ThP-Fr-3-20	0.1582	0.1919	0.2776	0.4420

Table C.3: Methods Evaluation : subCOV and subFMI With Abstract as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0830	0.0670	0.0363	0.0399
TNG-5	0.2102	0.1808	0.2527	0.2451
ThP-123-5	0.0256	0.0160	0.0233	0.0209
ThP-13-5	0.0288	0.0171	0.0288	0.0241
ThP-23-5	0.0256	0.0160	0.0233	0.0209
ThP-3-5	0.0288	0.0171	0.0290	0.0243
ThP-Fr-123-5	0.1571	0.1362	0.2201	0.2595
ThP-Fr-13-5	0.0750	0.0465	0.1372	0.1276
ThP-Fr-23-5	0.1571	0.1362	0.2174	0.2539
ThP-Fr-3-5	0.0750	0.0465	0.1388	0.1304
AP-10	0.1372	0.1603	0.0538	0.0817
TNG-10	0.2385	0.2491	0.2767	0.3285
ThP-123-10	0.0374	0.0303	0.0376	0.0457
ThP-13-10	0.0399	0.0310	0.0443	0.0503
ThP-23-10	0.0374	0.0303	0.0376	0.0457
ThP-3-10	0.0399	0.0309	0.0442	0.0501
ThP-Fr-123-10	0.2063	0.1930	0.3166	0.4053
ThP-Fr-13-10	0.1131	0.1022	0.2066	0.2597
ThP-Fr-23-10	0.2063	0.1930	0.3029	0.3829
ThP-Fr-3-10	0.1131	0.1022	0.2106	0.2682
AP-20	0.2352	0.3525	0.0838	0.1696
TNG-20	0.2477	0.3194	0.2725	0.4052
ThP-123-20	0.0630	0.0698	0.0639	0.1035
ThP-13-20	0.0639	0.0656	0.0741	0.1132
ThP-23-20	0.0630	0.0698	0.0639	0.1035
ThP-3-20	0.0640	0.0657	0.0741	0.1133
ThP-Fr-123-20	0.2141	0.2025	0.3482	0.4595
ThP-Fr-13-20	0.1582	0.1914	0.2727	0.4289
ThP-Fr-23-20	0.2141	0.2025	0.3217	0.4165
ThP-Fr-3-20	0.1582	0.1914	0.2779	0.4422

Table C.4: Methods Evaluation : subCOV and subFMI With Abstract as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0825	0.0665	0.0355	0.0387
TNG-5	0.2126	0.1829	0.2564	0.2477
ThP-123-5	0.0255	0.0159	0.0228	0.0205
ThP-13-5	0.0294	0.0173	0.0290	0.0241
ThP-23-5	0.0255	0.0159	0.0228	0.0205
ThP-3-5	0.0293	0.0173	0.0290	0.0241
ThP-Fr-123-5	0.1570	0.1364	0.2182	0.2574
ThP-Fr-13-5	0.0752	0.0463	0.1378	0.1281
ThP-Fr-23-5	0.1570	0.1364	0.2151	0.2517
ThP-Fr-3-5	0.0752	0.0463	0.1399	0.1314
AP-10	0.1369	0.1603	0.0538	0.0818
TNG-10	0.2383	0.2471	0.2814	0.3314
ThP-123-10	0.0374	0.0303	0.0372	0.0451
ThP-13-10	0.0406	0.0316	0.0446	0.0500
ThP-23-10	0.0374	0.0303	0.0372	0.0451
ThP-3-10	0.0406	0.0316	0.0445	0.0499
ThP-Fr-123-10	0.2064	0.1934	0.3155	0.4048
ThP-Fr-13-10	0.1132	0.1017	0.2059	0.2590
ThP-Fr-23-10	0.2064	0.1934	0.3018	0.3832
ThP-Fr-3-10	0.1132	0.1017	0.2105	0.2680
AP-20	0.2336	0.3488	0.0847	0.1719
TNG-20	0.2502	0.3209	0.2756	0.4075
ThP-123-20	0.0628	0.0698	0.0631	0.1023
ThP-13-20	0.0642	0.0660	0.0736	0.1128
ThP-23-20	0.0628	0.0698	0.0631	0.1023
ThP-3-20	0.0642	0.0661	0.0738	0.1131
ThP-Fr-123-20	0.2150	0.2037	0.3480	0.4610
ThP-Fr-13-20	0.1585	0.1914	0.2721	0.4285
ThP-Fr-23-20	0.2150	0.2037	0.3216	0.4183
ThP-Fr-3-20	0.1585	0.1914	0.2777	0.4425

Table C.5: Methods Evaluation : subCOV and subFMI With Abstract as Gold Standard (Segment Count = 25)

Appendix D

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : ABSTRACT EXT-COV AND EXT-FMI

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0617	0.0296	0.0487	0.0373
TNG-5	0.1727	0.1157	0.2453	0.2240
ThP-123-5	0.1454	0.0826	0.1526	0.1184
ThP-13-5	0.1475	0.0787	0.1557	0.1119
ThP-23-5	0.1454	0.0826	0.1526	0.1184
ThP-3-5	0.1476	0.0787	0.1559	0.1120
ThP-Fr-123-5	0.1208	0.0639	0.1656	0.1315
ThP-Fr-13-5	0.1481	0.0902	0.2474	0.2254
ThP-Fr-23-5	0.1208	0.0639	0.1564	0.1239
ThP-Fr-3-5	0.1481	0.0902	0.2504	0.2286
AP-10	0.0904	0.0593	0.0720	0.0719
TNG-10	0.2173	0.1708	0.3038	0.3212
ThP-123-10	0.2191	0.1513	0.2281	0.2102
ThP-13-10	0.2230	0.1442	0.2341	0.2012
ThP-23-10	0.2191	0.1513	0.2281	0.2102
ThP-3-10	0.2230	0.1442	0.2340	0.2010
ThP-Fr-123-10	0.1280	0.0721	0.1791	0.1579
ThP-Fr-13-10	0.2122	0.1529	0.3554	0.3762
ThP-Fr-23-10	0.1280	0.0721	0.1661	0.1445
ThP-Fr-3-10	0.2122	0.1529	0.3600	0.3805
AP-20	0.1347	0.1184	0.1097	0.1388
TNG-20	0.2709	0.2413	0.3551	0.4272
ThP-123-20	0.3096	0.2406	0.3249	0.3365
ThP-13-20	0.3299	0.2404	0.3456	0.3384
ThP-23-20	0.3096	0.2406	0.3249	0.3365
ThP-3-20	0.3298	0.2404	0.3457	0.3383
ThP-Fr-123-20	0.1287	0.0729	0.1870	0.1684
ThP-Fr-13-20	0.2940	0.2194	0.4442	0.4946
ThP-Fr-23-20	0.1287	0.0729	0.1710	0.1508
ThP-Fr-3-20	0.2940	0.2194	0.4492	0.4993

Table D.1: Methods Evaluation : extCOV and extFMI With Abstract as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0603	0.0285	0.0481	0.0366
TNG-5	0.1795	0.1206	0.2618	0.2420
ThP-123-5	0.1399	0.0796	0.1468	0.1139
ThP-13-5	0.1419	0.0750	0.1510	0.1079
ThP-23-5	0.1399	0.0796	0.1468	0.1139
ThP-3-5	0.1420	0.0751	0.1512	0.1080
ThP-Fr-123-5	0.1216	0.0648	0.1647	0.1313
ThP-Fr-13-5	0.1459	0.0889	0.2436	0.2225
ThP-Fr-23-5	0.1216	0.0648	0.1558	0.1237
ThP-Fr-3-5	0.1459	0.0889	0.2466	0.2253
AP-10	0.0885	0.0574	0.0708	0.0709
TNG-10	0.2261	0.1780	0.3188	0.3416
ThP-123-10	0.2129	0.1480	0.2208	0.2052
ThP-13-10	0.2171	0.1408	0.2268	0.1949
ThP-23-10	0.2129	0.1480	0.2208	0.2052
ThP-3-10	0.2172	0.1409	0.2268	0.1949
ThP-Fr-123-10	0.1290	0.0743	0.1777	0.1584
ThP-Fr-13-10	0.2084	0.1510	0.3508	0.3730
ThP-Fr-23-10	0.1290	0.0743	0.1651	0.1454
ThP-Fr-3-10	0.2084	0.1510	0.3549	0.3770
AP-20	0.1323	0.1155	0.1080	0.1369
TNG-20	0.2777	0.2476	0.3668	0.4487
ThP-123-20	0.3027	0.2394	0.3164	0.3328
ThP-13-20	0.3229	0.2379	0.3378	0.3334
ThP-23-20	0.3027	0.2394	0.3164	0.3328
ThP-3-20	0.3229	0.2379	0.3378	0.3335
ThP-Fr-123-20	0.1299	0.0753	0.1862	0.1701
ThP-Fr-13-20	0.2904	0.2197	0.4409	0.4963
ThP-Fr-23-20	0.1299	0.0753	0.1704	0.1523
ThP-Fr-3-20	0.2904	0.2197	0.4456	0.5008

Table D.2: Methods Evaluation : extCOV and extFMI With Abstract as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0605	0.0286	0.0472	0.0361
TNG-5	0.1812	0.1218	0.2689	0.2488
ThP-123-5	0.1364	0.0768	0.1427	0.1116
ThP-13-5	0.1393	0.0733	0.1466	0.1050
ThP-23-5	0.1364	0.0768	0.1427	0.1116
ThP-3-5	0.1393	0.0734	0.1471	0.1053
ThP-Fr-123-5	0.1219	0.0654	0.1646	0.1315
ThP-Fr-13-5	0.1440	0.0878	0.2425	0.2215
ThP-Fr-23-5	0.1219	0.0654	0.1559	0.1243
ThP-Fr-3-5	0.1440	0.0878	0.2449	0.2245
AP-10	0.0885	0.0572	0.0704	0.0707
TNG-10	0.2285	0.1793	0.3265	0.3515
ThP-123-10	0.2085	0.1450	0.2149	0.2006
ThP-13-10	0.2135	0.1386	0.2215	0.1914
ThP-23-10	0.2085	0.1450	0.2149	0.2006
ThP-3-10	0.2135	0.1385	0.2216	0.1913
ThP-Fr-123-10	0.1293	0.0753	0.1775	0.1590
ThP-Fr-13-10	0.2058	0.1495	0.3484	0.3720
ThP-Fr-23-10	0.1293	0.0753	0.1652	0.1465
ThP-Fr-3-10	0.2058	0.1495	0.3525	0.3757
AP-20	0.1322	0.1148	0.1070	0.1367
TNG-20	0.2811	0.2501	0.3737	0.4590
ThP-123-20	0.2981	0.2386	0.3110	0.3302
ThP-13-20	0.3184	0.2358	0.3323	0.3306
ThP-23-20	0.2981	0.2386	0.3110	0.3302
ThP-3-20	0.3184	0.2357	0.3323	0.3305
ThP-Fr-123-20	0.1301	0.0764	0.1859	0.1712
ThP-Fr-13-20	0.2876	0.2192	0.4382	0.4973
ThP-Fr-23-20	0.1301	0.0764	0.1703	0.1539
ThP-Fr-3-20	0.2876	0.2192	0.4429	0.5017

Table D.3: Methods Evaluation : extCOV and extFMI With Abstract as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0602	0.0283	0.0463	0.0350
TNG-5	0.1815	0.1223	0.2728	0.2533
ThP-123-5	0.1336	0.0754	0.1402	0.1094
ThP-13-5	0.1361	0.0716	0.1449	0.1035
ThP-23-5	0.1336	0.0754	0.1402	0.1094
ThP-3-5	0.1361	0.0716	0.1452	0.1036
ThP-Fr-123-5	0.1217	0.0655	0.1648	0.1319
ThP-Fr-13-5	0.1435	0.0874	0.2418	0.2206
ThP-Fr-23-5	0.1217	0.0655	0.1559	0.1241
ThP-Fr-3-5	0.1435	0.0874	0.2443	0.2235
AP-10	0.0883	0.0570	0.0693	0.0691
TNG-10	0.2289	0.1796	0.3306	0.3560
ThP-123-10	0.2063	0.1439	0.2124	0.1989
ThP-13-10	0.2108	0.1370	0.2195	0.1887
ThP-23-10	0.2063	0.1439	0.2124	0.1989
ThP-3-10	0.2108	0.1369	0.2195	0.1886
ThP-Fr-123-10	0.1294	0.0759	0.1775	0.1599
ThP-Fr-13-10	0.2046	0.1488	0.3473	0.3716
ThP-Fr-23-10	0.1294	0.0759	0.1649	0.1470
ThP-Fr-3-10	0.2046	0.1488	0.3512	0.3756
AP-20	0.1321	0.1147	0.1056	0.1340
TNG-20	0.2818	0.2504	0.3774	0.4635
ThP-123-20	0.2943	0.2376	0.3077	0.3291
ThP-13-20	0.3148	0.2343	0.3292	0.3265
ThP-23-20	0.2943	0.2376	0.3077	0.3291
ThP-3-20	0.3149	0.2342	0.3293	0.3266
ThP-Fr-123-20	0.1303	0.0771	0.1861	0.1726
ThP-Fr-13-20	0.2864	0.2193	0.4377	0.4993
ThP-Fr-23-20	0.1303	0.0771	0.1703	0.1546
ThP-Fr-3-20	0.2864	0.2193	0.4423	0.5041

Table D.4: Methods Evaluation : extCOV and extFMI With Abstract as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0599	0.0281	0.0461	0.0349
TNG-5	0.1819	0.1220	0.2745	0.2543
ThP-123-5	0.1316	0.0743	0.1381	0.1076
ThP-13-5	0.1353	0.0709	0.1433	0.1026
ThP-23-5	0.1316	0.0743	0.1381	0.1076
ThP-3-5	0.1353	0.0708	0.1435	0.1028
ThP-Fr-123-5	0.1222	0.0657	0.1641	0.1316
ThP-Fr-13-5	0.1423	0.0866	0.2414	0.2213
ThP-Fr-23-5	0.1222	0.0657	0.1562	0.1247
ThP-Fr-3-5	0.1423	0.0866	0.2443	0.2240
AP-10	0.0877	0.0566	0.0695	0.0692
TNG-10	0.2298	0.1800	0.3331	0.3581
ThP-123-10	0.2040	0.1425	0.2106	0.1972
ThP-13-10	0.2095	0.1358	0.2162	0.1867
ThP-23-10	0.2040	0.1425	0.2106	0.1972
ThP-3-10	0.2095	0.1358	0.2161	0.1866
ThP-Fr-123-10	0.1296	0.0763	0.1772	0.1605
ThP-Fr-13-10	0.2042	0.1486	0.3452	0.3706
ThP-Fr-23-10	0.1296	0.0763	0.1652	0.1479
ThP-Fr-3-10	0.2042	0.1486	0.3497	0.3746
AP-20	0.1312	0.1137	0.1056	0.1340
TNG-20	0.2829	0.2514	0.3796	0.4658
ThP-123-20	0.2922	0.2363	0.3048	0.3272
ThP-13-20	0.3135	0.2337	0.3260	0.3255
ThP-23-20	0.2922	0.2363	0.3048	0.3272
ThP-3-20	0.3135	0.2337	0.3261	0.3258
ThP-Fr-123-20	0.1305	0.0775	0.1861	0.1735
ThP-Fr-13-20	0.2857	0.2194	0.4362	0.4992
ThP-Fr-23-20	0.1305	0.0775	0.1705	0.1556
ThP-Fr-3-20	0.2857	0.2194	0.4409	0.5045

Table D.5: Methods Evaluation : extCOV and extFMI With Abstract as Gold Standard (Segment Count = 25)

Appendix E

THEMATIC PHRASES QUANTITATIVE METRICS TABLES : ABSTRACT WORD GRANULARITY METRICS

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.1172	0.0615	0.0432	0.1257	0.0873	0.0764	0.0455	0.0933
TNG-5	0.2383	0.1832	0.0881	0.3957	0.3660	0.3623	0.2071	0.4887
ThP-123-5	0.1746	0.1319	0.0719	0.2444	0.2185	0.2289	0.1214	0.2584
ThP-13-5	0.1687	0.1251	0.0676	0.2364	0.2164	0.2165	0.1175	0.2485
ThP-23-5	0.1746	0.1319	0.0719	0.2444	0.2185	0.2289	0.1214	0.2584
ThP-3-5	0.1687	0.1252	0.0677	0.2365	0.2165	0.2166	0.1176	0.2487
ThP-Fr-123-5	0.1994	0.1345	0.0633	0.3755	0.3224	0.2867	0.1666	0.4134
ThP-Fr-13-5	0.2081	0.1664	0.0851	0.3531	0.3728	0.3922	0.2263	0.4500
ThP-Fr-23-5	0.1994	0.1345	0.0633	0.3755	0.3191	0.2821	0.1649	0.4059
ThP-Fr-3-5	0.2081	0.1664	0.0851	0.3531	0.3746	0.3964	0.2278	0.4534
AP-10	0.1606	0.1154	0.0743	0.1770	0.1182	0.1392	0.0662	0.1290
TNG-10	0.2781	0.2625	0.1318	0.4370	0.3958	0.4812	0.2491	0.5236
ThP-123-10	0.2382	0.2417	0.1198	0.3575	0.2827	0.3854	0.1674	0.3569
ThP-13-10	0.2319	0.2325	0.1149	0.3509	0.2813	0.3673	0.1656	0.3432
ThP-23-10	0.2382	0.2417	0.1198	0.3575	0.2827	0.3854	0.1674	0.3569
ThP-3-10	0.2319	0.2325	0.1149	0.3509	0.2812	0.3671	0.1655	0.3432
ThP-Fr-123-10	0.2182	0.1598	0.0765	0.3989	0.3648	0.3693	0.2076	0.4600
ThP-Fr-13-10	0.2582	0.2551	0.1255	0.4324	0.4403	0.5581	0.2881	0.5399
ThP-Fr-23-10	0.2182	0.1598	0.0765	0.3989	0.3559	0.3566	0.2008	0.4441
ThP-Fr-3-10	0.2582	0.2551	0.1255	0.4324	0.4421	0.5631	0.2896	0.5433
AP-20	0.2189	0.2140	0.1187	0.2521	0.1605	0.2536	0.0872	0.1834
TNG-20	0.3094	0.3598	0.1744	0.4743	0.3901	0.5962	0.2406	0.5397
ThP-123-20	0.2902	0.3610	0.1643	0.4602	0.3343	0.5552	0.1955	0.4576
ThP-13-20	0.2946	0.3629	0.1664	0.4732	0.3390	0.5433	0.1998	0.4507
ThP-23-20	0.2902	0.3610	0.1643	0.4602	0.3343	0.5552	0.1955	0.4576
ThP-3-20	0.2945	0.3628	0.1664	0.4731	0.3390	0.5433	0.1998	0.4507
ThP-Fr-123-20	0.2194	0.1621	0.0776	0.4003	0.3714	0.3954	0.2154	0.4708
ThP-Fr-13-20	0.2924	0.3390	0.1603	0.4881	0.4543	0.6594	0.2955	0.5699
ThP-Fr-23-20	0.2194	0.1621	0.0776	0.4003	0.3596	0.3731	0.2052	0.4500
ThP-Fr-3-20	0.2924	0.3390	0.1603	0.4881	0.4537	0.6636	0.2944	0.5693

Table E.1: Methods Evaluation : Word Granularity Metrics With Abstract as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.1152	0.0595	0.0425	0.1216	0.0859	0.0752	0.0447	0.0915
TNG-5	0.2471	0.1895	0.0913	0.4093	0.3887	0.3796	0.2232	0.5071
ThP-123-5	0.1707	0.1263	0.0703	0.2320	0.2115	0.2201	0.1172	0.2472
ThP-13-5	0.1647	0.1193	0.0658	0.2252	0.2110	0.2091	0.1142	0.2394
ThP-23-5	0.1707	0.1263	0.0703	0.2320	0.2115	0.2201	0.1172	0.2472
ThP-3-5	0.1649	0.1195	0.0659	0.2253	0.2112	0.2091	0.1143	0.2395
ThP-Fr-123-5	0.2000	0.1352	0.0642	0.3725	0.3187	0.2843	0.1650	0.4075
ThP-Fr-13-5	0.2077	0.1647	0.0850	0.3469	0.3698	0.3882	0.2242	0.4445
ThP-Fr-23-5	0.2000	0.1352	0.0642	0.3725	0.3157	0.2801	0.1634	0.4009
ThP-Fr-3-5	0.2077	0.1647	0.0850	0.3469	0.3718	0.3922	0.2258	0.4478
AP-10	0.1580	0.1120	0.0730	0.1714	0.1167	0.1374	0.0653	0.1268
TNG-10	0.2899	0.2715	0.1375	0.4518	0.4189	0.5008	0.2678	0.5399
ThP-123-10	0.2355	0.2374	0.1187	0.3476	0.2770	0.3767	0.1633	0.3459
ThP-13-10	0.2296	0.2283	0.1136	0.3411	0.2763	0.3583	0.1620	0.3315
ThP-23-10	0.2355	0.2374	0.1187	0.3476	0.2770	0.3767	0.1633	0.3459
ThP-3-10	0.2296	0.2284	0.1136	0.3411	0.2765	0.3584	0.1621	0.3316
ThP-Fr-123-10	0.2213	0.1638	0.0793	0.3984	0.3627	0.3702	0.2072	0.4561
ThP-Fr-13-10	0.2585	0.2537	0.1260	0.4245	0.4384	0.5563	0.2865	0.5352
ThP-Fr-23-10	0.2213	0.1638	0.0793	0.3984	0.3542	0.3580	0.2008	0.4407
ThP-Fr-3-10	0.2585	0.2537	0.1260	0.4245	0.4398	0.5609	0.2877	0.5387
AP-20	0.2167	0.2099	0.1174	0.2467	0.1586	0.2502	0.0861	0.1805
TNG-20	0.3200	0.3667	0.1808	0.4829	0.4118	0.6146	0.2583	0.5518
ThP-123-20	0.2910	0.3616	0.1653	0.4522	0.3304	0.5508	0.1922	0.4467
ThP-13-20	0.2941	0.3614	0.1662	0.4655	0.3365	0.5387	0.1979	0.4414
ThP-23-20	0.2910	0.3616	0.1653	0.4522	0.3304	0.5508	0.1922	0.4467
ThP-3-20	0.2940	0.3614	0.1661	0.4654	0.3366	0.5388	0.1979	0.4414
ThP-Fr-123-20	0.2229	0.1668	0.0808	0.4002	0.3704	0.3993	0.2161	0.4681
ThP-Fr-13-20	0.2939	0.3405	0.1616	0.4824	0.4546	0.6620	0.2956	0.5662
ThP-Fr-23-20	0.2229	0.1668	0.0808	0.4002	0.3588	0.3769	0.2060	0.4475
ThP-Fr-3-20	0.2939	0.3405	0.1616	0.4824	0.4542	0.6662	0.2945	0.5655

Table E.2: Methods Evaluation : Word Granularity Metrics With Abstract as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.1152	0.0594	0.0425	0.1213	0.0849	0.0743	0.0442	0.0907
TNG-5	0.2491	0.1904	0.0919	0.4115	0.3984	0.3875	0.2301	0.5148
ThP-123-5	0.1674	0.1219	0.0689	0.2233	0.2063	0.2136	0.1139	0.2382
ThP-13-5	0.1630	0.1164	0.0650	0.2184	0.2055	0.2023	0.1108	0.2309
ThP-23-5	0.1674	0.1219	0.0689	0.2233	0.2063	0.2136	0.1139	0.2382
ThP-3-5	0.1630	0.1164	0.0650	0.2184	0.2058	0.2025	0.1109	0.2309
ThP-Fr-123-5	0.2000	0.1357	0.0647	0.3716	0.3170	0.2836	0.1644	0.4044
ThP-Fr-13-5	0.2068	0.1632	0.0845	0.3442	0.3689	0.3873	0.2236	0.4439
ThP-Fr-23-5	0.2000	0.1357	0.0647	0.3716	0.3141	0.2792	0.1629	0.3972
ThP-Fr-3-5	0.2068	0.1632	0.0845	0.3442	0.3707	0.3915	0.2252	0.4466
AP-10	0.1576	0.1111	0.0728	0.1700	0.1158	0.1363	0.0648	0.1255
TNG-10	0.2931	0.2729	0.1389	0.4547	0.4294	0.5097	0.2763	0.5472
ThP-123-10	0.2337	0.2339	0.1179	0.3397	0.2719	0.3694	0.1598	0.3376
ThP-13-10	0.2277	0.2246	0.1126	0.3334	0.2719	0.3509	0.1589	0.3228
ThP-23-10	0.2337	0.2339	0.1179	0.3397	0.2719	0.3694	0.1598	0.3376
ThP-3-10	0.2277	0.2246	0.1125	0.3334	0.2718	0.3507	0.1589	0.3228
ThP-Fr-123-10	0.2224	0.1655	0.0807	0.3976	0.3614	0.3707	0.2070	0.4533
ThP-Fr-13-10	0.2576	0.2523	0.1255	0.4205	0.4372	0.5555	0.2856	0.5329
ThP-Fr-23-10	0.2224	0.1655	0.0807	0.3976	0.3534	0.3593	0.2011	0.4380
ThP-Fr-3-10	0.2576	0.2523	0.1255	0.4205	0.4387	0.5603	0.2867	0.5363
AP-20	0.2165	0.2090	0.1173	0.2449	0.1573	0.2489	0.0853	0.1792
TNG-20	0.3245	0.3697	0.1835	0.4868	0.4219	0.6235	0.2666	0.5576
ThP-123-20	0.2916	0.3617	0.1661	0.4462	0.3273	0.5470	0.1897	0.4398
ThP-13-20	0.2941	0.3603	0.1663	0.4595	0.3338	0.5345	0.1959	0.4341
ThP-23-20	0.2916	0.3617	0.1661	0.4462	0.3273	0.5470	0.1897	0.4398
ThP-3-20	0.2941	0.3602	0.1663	0.4595	0.3338	0.5346	0.1959	0.4342
ThP-Fr-123-20	0.2243	0.1690	0.0824	0.3995	0.3696	0.4016	0.2164	0.4658
ThP-Fr-13-20	0.2940	0.3403	0.1618	0.4785	0.4548	0.6635	0.2954	0.5637
ThP-Fr-23-20	0.2243	0.1690	0.0824	0.3995	0.3582	0.3792	0.2065	0.4450
ThP-Fr-3-20	0.2940	0.3403	0.1618	0.4785	0.4542	0.6676	0.2943	0.5630

Table E.3: Methods Evaluation : Word Granularity Metrics With Abstract as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.1148	0.0588	0.0423	0.1201	0.0836	0.0724	0.0435	0.0882
TNG-5	0.2507	0.1912	0.0925	0.4128	0.4043	0.3924	0.2345	0.5205
ThP-123-5	0.1653	0.1191	0.0680	0.2172	0.2045	0.2110	0.1127	0.2349
ThP-13-5	0.1610	0.1134	0.0641	0.2131	0.2036	0.1988	0.1097	0.2262
ThP-23-5	0.1653	0.1191	0.0680	0.2172	0.2045	0.2110	0.1127	0.2349
ThP-3-5	0.1611	0.1134	0.0641	0.2131	0.2038	0.1988	0.1097	0.2263
ThP-Fr-123-5	0.1998	0.1355	0.0649	0.3694	0.3161	0.2825	0.1640	0.4023
ThP-Fr-13-5	0.2071	0.1629	0.0847	0.3421	0.3679	0.3855	0.2229	0.4420
ThP-Fr-23-5	0.1998	0.1355	0.0649	0.3694	0.3131	0.2780	0.1625	0.3948
ThP-Fr-3-5	0.2071	0.1629	0.0847	0.3421	0.3700	0.3898	0.2246	0.4450
AP-10	0.1573	0.1108	0.0727	0.1696	0.1140	0.1341	0.0638	0.1235
TNG-10	0.2946	0.2733	0.1395	0.4552	0.4352	0.5142	0.2812	0.5507
ThP-123-10	0.2323	0.2320	0.1173	0.3350	0.2700	0.3671	0.1584	0.3335
ThP-13-10	0.2270	0.2226	0.1122	0.3292	0.2694	0.3462	0.1573	0.3172
ThP-23-10	0.2323	0.2320	0.1173	0.3350	0.2700	0.3671	0.1584	0.3335
ThP-3-10	0.2270	0.2226	0.1122	0.3291	0.2695	0.3463	0.1573	0.3174
ThP-Fr-123-10	0.2234	0.1670	0.0818	0.3969	0.3612	0.3714	0.2072	0.4514
ThP-Fr-13-10	0.2579	0.2518	0.1258	0.4177	0.4371	0.5554	0.2855	0.5314
ThP-Fr-23-10	0.2234	0.1670	0.0818	0.3969	0.3535	0.3601	0.2016	0.4365
ThP-Fr-3-10	0.2579	0.2518	0.1258	0.4177	0.4385	0.5599	0.2866	0.5346
AP-20	0.2164	0.2085	0.1173	0.2443	0.1561	0.2460	0.0846	0.1763
TNG-20	0.3260	0.3702	0.1844	0.4868	0.4268	0.6271	0.2708	0.5606
ThP-123-20	0.2909	0.3612	0.1658	0.4417	0.3257	0.5460	0.1882	0.4355
ThP-13-20	0.2936	0.3594	0.1660	0.4564	0.3326	0.5317	0.1949	0.4301
ThP-23-20	0.2909	0.3612	0.1658	0.4417	0.3257	0.5460	0.1882	0.4355
ThP-3-20	0.2935	0.3593	0.1660	0.4563	0.3325	0.5317	0.1949	0.4302
ThP-Fr-123-20	0.2255	0.1708	0.0837	0.3989	0.3697	0.4034	0.2170	0.4645
ThP-Fr-13-20	0.2944	0.3407	0.1622	0.4764	0.4557	0.6654	0.2962	0.5623
ThP-Fr-23-20	0.2255	0.1708	0.0837	0.3989	0.3586	0.3811	0.2073	0.4439
ThP-Fr-3-20	0.2944	0.3407	0.1622	0.4764	0.4554	0.6699	0.2953	0.5621

Table E.4: Methods Evaluation : Word Granularity Metrics With Abstract as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.1143	0.0582	0.0421	0.1191	0.0830	0.0719	0.0432	0.0878
TNG-5	0.2511	0.1916	0.0927	0.4136	0.4070	0.3940	0.2361	0.5221
ThP-123-5	0.1644	0.1178	0.0677	0.2140	0.2009	0.2069	0.1106	0.2302
ThP-13-5	0.1612	0.1122	0.0641	0.2106	0.2011	0.1958	0.1081	0.2225
ThP-23-5	0.1644	0.1178	0.0677	0.2140	0.2009	0.2069	0.1106	0.2302
ThP-3-5	0.1611	0.1122	0.0640	0.2106	0.2012	0.1959	0.1081	0.2228
ThP-Fr-123-5	0.1996	0.1355	0.0649	0.3687	0.3151	0.2817	0.1636	0.3999
ThP-Fr-13-5	0.2065	0.1618	0.0845	0.3392	0.3681	0.3861	0.2231	0.4411
ThP-Fr-23-5	0.1996	0.1355	0.0649	0.3687	0.3123	0.2778	0.1623	0.3931
ThP-Fr-3-5	0.2065	0.1618	0.0845	0.3392	0.3703	0.3903	0.2249	0.4445
AP-10	0.1571	0.1104	0.0726	0.1687	0.1138	0.1338	0.0636	0.1231
TNG-10	0.2952	0.2738	0.1397	0.4556	0.4386	0.5167	0.2840	0.5531
ThP-123-10	0.2312	0.2305	0.1168	0.3312	0.2674	0.3631	0.1566	0.3292
ThP-13-10	0.2265	0.2210	0.1118	0.3259	0.2671	0.3431	0.1557	0.3133
ThP-23-10	0.2312	0.2305	0.1168	0.3312	0.2674	0.3631	0.1566	0.3292
ThP-3-10	0.2264	0.2210	0.1118	0.3259	0.2672	0.3431	0.1558	0.3133
ThP-Fr-123-10	0.2238	0.1675	0.0822	0.3963	0.3604	0.3715	0.2071	0.4503
ThP-Fr-13-10	0.2576	0.2513	0.1257	0.4161	0.4359	0.5548	0.2846	0.5296
ThP-Fr-23-10	0.2238	0.1675	0.0822	0.3963	0.3529	0.3608	0.2017	0.4354
ThP-Fr-3-10	0.2576	0.2513	0.1257	0.4161	0.4378	0.5595	0.2860	0.5333
AP-20	0.2155	0.2073	0.1167	0.2425	0.1556	0.2449	0.0844	0.1758
TNG-20	0.3273	0.3714	0.1851	0.4881	0.4301	0.6299	0.2735	0.5626
ThP-123-20	0.2907	0.3604	0.1658	0.4390	0.3239	0.5435	0.1869	0.4318
ThP-13-20	0.2938	0.3591	0.1661	0.4544	0.3309	0.5294	0.1937	0.4268
ThP-23-20	0.2907	0.3604	0.1658	0.4390	0.3239	0.5435	0.1869	0.4318
ThP-3-20	0.2938	0.3591	0.1661	0.4544	0.3310	0.5296	0.1938	0.4270
ThP-Fr-123-20	0.2262	0.1716	0.0843	0.3986	0.3696	0.4051	0.2174	0.4638
ThP-Fr-13-20	0.2949	0.3409	0.1625	0.4746	0.4551	0.6658	0.2955	0.5609
ThP-Fr-23-20	0.2262	0.1716	0.0843	0.3986	0.3583	0.3828	0.2077	0.4431
ThP-Fr-3-20	0.2949	0.3409	0.1625	0.4746	0.4551	0.6707	0.2947	0.5610

Table E.5: Methods Evaluation : Word Granularity Metrics With Abstract as Gold Standard (Segment Count = 25)

Appendix F

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : ABSTRACT DCG

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	phHIT	wdHIT	wdCOV	phHIT	wdHIT	wdCOV
AP-3	0.2325	0.5726	0.0048	0.0702	0.2466	0.0059
TNG-3	0.8135	1.4285	0.0206	0.8106	1.3473	0.0440
ThP-123-3	0.1323	0.7361	0.0077	0.0813	0.5452	0.0146
ThP-13-3	0.1457	0.7527	0.0079	0.0978	0.5879	0.0155
ThP-23-3	0.1323	0.7361	0.0077	0.0813	0.5452	0.0146
ThP-3-3	0.1457	0.7527	0.0079	0.0976	0.5877	0.0155
ThP-Fr-123-3	0.4564	1.3393	0.0184	0.3915	1.1920	0.0358
ThP-Fr-13-3	0.3867	0.9401	0.0109	0.5029	1.0055	0.0266
ThP-Fr-23-3	0.4564	1.3393	0.0184	0.3595	1.1857	0.0353
ThP-Fr-3-3	0.3867	0.9401	0.0109	0.5101	1.0162	0.0271
AP-5	0.3113	0.7807	0.0066	0.0917	0.3336	0.0080
TNG-5	0.9991	1.8651	0.0259	0.9959	1.7401	0.0557
ThP-123-5	0.1803	1.0623	0.0117	0.1168	0.7854	0.0215
ThP-13-5	0.1944	1.0781	0.0119	0.1375	0.8367	0.0224
ThP-23-5	0.1803	1.0623	0.0117	0.1168	0.7854	0.0215
ThP-3-5	0.1944	1.0782	0.0119	0.1376	0.8368	0.0224
ThP-Fr-123-5	0.6034	1.7893	0.0238	0.5265	1.6425	0.0482
ThP-Fr-13-5	0.5589	1.3303	0.0158	0.7141	1.4101	0.0379
ThP-Fr-23-5	0.6034	1.7893	0.0238	0.4751	1.6426	0.0476
ThP-Fr-3-5	0.5589	1.3303	0.0158	0.7263	1.4291	0.0386
AP-10	0.4658	1.2051	0.0105	0.1363	0.5155	0.0124
TNG-10	1.2493	2.5914	0.0341	1.2495	2.3764	0.0734
ThP-123-10	0.2906	1.7490	0.0201	0.2043	1.2883	0.0359
ThP-13-10	0.2979	1.7701	0.0205	0.2306	1.3641	0.0371
ThP-23-10	0.2906	1.7490	0.0201	0.2043	1.2883	0.0359
ThP-3-10	0.2979	1.7701	0.0205	0.2309	1.3642	0.0371
ThP-Fr-123-10	0.7064	2.1892	0.0283	0.6559	2.2692	0.0645
ThP-Fr-13-10	0.9022	2.1035	0.0255	1.1335	2.1827	0.0597
ThP-Fr-23-10	0.7064	2.1892	0.0283	0.5667	2.2255	0.0625
ThP-Fr-3-10	0.9022	2.1035	0.0255	1.1606	2.2301	0.0613
AP-20	0.7116	1.9145	0.0170	0.2020	0.8235	0.0199
TNG-20	1.5155	3.5558	0.0447	1.4751	3.1336	0.0942
ThP-123-20	0.5036	2.8359	0.0333	0.3608	2.1295	0.0604
ThP-13-20	0.5085	2.9766	0.0357	0.4072	2.2884	0.0634
ThP-23-20	0.5036	2.8359	0.0333	0.3608	2.1295	0.0604
ThP-3-20	0.5086	2.9763	0.0357	0.4076	2.2888	0.0634
ThP-Fr-123-20	0.7175	2.2445	0.0289	0.6908	2.5609	0.0720
ThP-Fr-13-20	1.3363	3.2972	0.0401	1.5782	3.2387	0.0893
ThP-Fr-23-20	0.7175	2.2445	0.0289	0.5815	2.4197	0.0674
ThP-Fr-3-20	1.3363	3.2972	0.0401	1.6199	3.3455	0.0927

Table F.1: Methods Evaluation Abstract Phrase Metrics (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	phHIT	wdHIT	wdCOV	phHIT	wdHIT	wdCOV
AP-3	0.2282	0.5609	0.0046	0.0700	0.2430	0.0058
TNG-3	0.8660	1.4770	0.0213	0.9372	1.4271	0.0459
ThP-123-3	0.1305	0.7104	0.0072	0.0793	0.5223	0.0137
ThP-13-3	0.1425	0.7302	0.0075	0.1000	0.5723	0.0148
ThP-23-3	0.1305	0.7104	0.0072	0.0793	0.5223	0.0137
ThP-3-3	0.1427	0.7306	0.0075	0.0998	0.5726	0.0148
ThP-Fr-123-3	0.4446	1.3245	0.0180	0.3855	1.1738	0.0351
ThP-Fr-13-3	0.3938	0.9363	0.0107	0.5020	0.9974	0.0261
ThP-Fr-23-3	0.4446	1.3245	0.0180	0.3529	1.1687	0.0347
ThP-Fr-3-3	0.3938	0.9363	0.0107	0.5099	1.0090	0.0266
AP-5	0.3046	0.7664	0.0064	0.0916	0.3295	0.0079
TNG-5	1.0717	1.9380	0.0268	1.1536	1.8417	0.0576
ThP-123-5	0.1777	1.0285	0.0110	0.1147	0.7552	0.0204
ThP-13-5	0.1925	1.0497	0.0113	0.1402	0.8149	0.0215
ThP-23-5	0.1777	1.0285	0.0110	0.1147	0.7552	0.0204
ThP-3-5	0.1925	1.0500	0.0113	0.1403	0.8153	0.0215
ThP-Fr-123-5	0.5986	1.7881	0.0235	0.5210	1.6195	0.0473
ThP-Fr-13-5	0.5665	1.3233	0.0155	0.7127	1.3969	0.0373
ThP-Fr-23-5	0.5986	1.7881	0.0235	0.4698	1.6206	0.0468
ThP-Fr-3-5	0.5665	1.3233	0.0155	0.7254	1.4171	0.0381
AP-10	0.4552	1.1818	0.0101	0.1358	0.5095	0.0122
TNG-10	1.3535	2.7038	0.0355	1.4351	2.4954	0.0754
ThP-123-10	0.2843	1.7028	0.0192	0.1984	1.2431	0.0342
ThP-13-10	0.2960	1.7338	0.0197	0.2310	1.3286	0.0356
ThP-23-10	0.2843	1.7028	0.0192	0.1984	1.2431	0.0342
ThP-3-10	0.2961	1.7338	0.0197	0.2314	1.3294	0.0356
ThP-Fr-123-10	0.7152	2.2340	0.0285	0.6546	2.2566	0.0638
ThP-Fr-13-10	0.9061	2.0875	0.0248	1.1276	2.1607	0.0587
ThP-Fr-23-10	0.7152	2.2340	0.0285	0.5671	2.2216	0.0621
ThP-Fr-3-10	0.9061	2.0875	0.0248	1.1540	2.2067	0.0603
AP-20	0.6997	1.8871	0.0166	0.2011	0.8133	0.0196
TNG-20	1.6211	3.6842	0.0461	1.6705	3.2674	0.0960
ThP-123-20	0.4953	2.7738	0.0320	0.3521	2.0588	0.0576
ThP-13-20	0.5036	2.9217	0.0345	0.4065	2.2360	0.0613
ThP-23-20	0.4953	2.7738	0.0320	0.3521	2.0588	0.0576
ThP-3-20	0.5041	2.9218	0.0345	0.4067	2.2362	0.0613
ThP-Fr-123-20	0.7291	2.3018	0.0292	0.6937	2.5720	0.0718
ThP-Fr-13-20	1.3427	3.2739	0.0393	1.5775	3.2078	0.0878
ThP-Fr-23-20	0.7291	2.3018	0.0292	0.5837	2.4337	0.0674
ThP-Fr-3-20	1.3427	3.2739	0.0393	1.6192	3.3116	0.0911

Table F.2: Methods Evaluation Abstract Phrase Metrics (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	phHIT	wdHIT	wdCOV	phHIT	wdHIT	wdCOV
AP-3	0.2274	0.5617	0.0046	0.0682	0.2402	0.0057
TNG-3	0.8810	1.4894	0.0214	0.9804	1.4585	0.0465
ThP-123-3	0.1316	0.6952	0.0069	0.0759	0.5041	0.0131
ThP-13-3	0.1473	0.7248	0.0073	0.0969	0.5560	0.0142
ThP-23-3	0.1316	0.6952	0.0069	0.0759	0.5041	0.0131
ThP-3-3	0.1472	0.7250	0.0073	0.0967	0.5559	0.0142
ThP-Fr-123-3	0.4353	1.3122	0.0178	0.3788	1.1554	0.0344
ThP-Fr-13-3	0.3992	0.9324	0.0106	0.5012	0.9950	0.0261
ThP-Fr-23-3	0.4353	1.3122	0.0178	0.3481	1.1533	0.0341
ThP-Fr-3-3	0.3992	0.9324	0.0106	0.5093	1.0060	0.0265
AP-5	0.3047	0.7672	0.0064	0.0901	0.3256	0.0078
TNG-5	1.0916	1.9539	0.0269	1.2089	1.8835	0.0584
ThP-123-5	0.1789	1.0073	0.0106	0.1104	0.7315	0.0195
ThP-13-5	0.1980	1.0413	0.0110	0.1377	0.7936	0.0207
ThP-23-5	0.1789	1.0073	0.0106	0.1104	0.7315	0.0195
ThP-3-5	0.1978	1.0412	0.0110	0.1377	0.7937	0.0207
ThP-Fr-123-5	0.5932	1.7827	0.0234	0.5155	1.6010	0.0466
ThP-Fr-13-5	0.5709	1.3171	0.0153	0.7093	1.3915	0.0371
ThP-Fr-23-5	0.5932	1.7827	0.0234	0.4662	1.6029	0.0461
ThP-Fr-3-5	0.5709	1.3171	0.0153	0.7218	1.4090	0.0378
AP-10	0.4545	1.1810	0.0100	0.1325	0.5037	0.0120
TNG-10	1.3796	2.7354	0.0358	1.5020	2.5496	0.0763
ThP-123-10	0.2836	1.6729	0.0186	0.1952	1.2117	0.0331
ThP-13-10	0.3001	1.7164	0.0193	0.2263	1.2975	0.0345
ThP-23-10	0.2836	1.6729	0.0186	0.1952	1.2117	0.0331
ThP-3-10	0.3001	1.7163	0.0193	0.2265	1.2980	0.0345
ThP-Fr-123-10	0.7160	2.2523	0.0285	0.6521	2.2448	0.0633
ThP-Fr-13-10	0.9082	2.0762	0.0246	1.1217	2.1470	0.0583
ThP-Fr-23-10	0.7160	2.2523	0.0285	0.5664	2.2124	0.0616
ThP-Fr-3-10	0.9082	2.0762	0.0246	1.1469	2.1912	0.0598
AP-20	0.6981	1.8838	0.0165	0.1984	0.8050	0.0194
TNG-20	1.6554	3.7316	0.0465	1.7459	3.3335	0.0970
ThP-123-20	0.4926	2.7333	0.0311	0.3461	2.0104	0.0559
ThP-13-20	0.5067	2.8947	0.0337	0.4018	2.1941	0.0597
ThP-23-20	0.4926	2.7333	0.0311	0.3461	2.0104	0.0559
ThP-3-20	0.5064	2.8945	0.0337	0.4016	2.1941	0.0597
ThP-Fr-123-20	0.7315	2.3285	0.0293	0.6935	2.5731	0.0716
ThP-Fr-13-20	1.3460	3.2586	0.0389	1.5736	3.1860	0.0870
ThP-Fr-23-20	0.7315	2.3285	0.0293	0.5844	2.4371	0.0672
ThP-Fr-3-20	1.3460	3.2586	0.0389	1.6145	3.2859	0.0902

Table F.3: Methods Evaluation Abstract Phrase Metrics (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	phHIT	wdHIT	wdCOV	phHIT	wdHIT	wdCOV
AP-3	0.2243	0.5587	0.0046	0.0671	0.2361	0.0055
TNG-3	0.8877	1.4994	0.0215	0.9983	1.4762	0.0470
ThP-123-3	0.1273	0.6832	0.0067	0.0777	0.5001	0.0129
ThP-13-3	0.1443	0.7122	0.0071	0.0929	0.5433	0.0138
ThP-23-3	0.1273	0.6832	0.0067	0.0777	0.5001	0.0129
ThP-3-3	0.1446	0.7125	0.0071	0.0931	0.5437	0.0138
ThP-Fr-123-3	0.4314	1.3046	0.0176	0.3771	1.1521	0.0342
ThP-Fr-13-3	0.3990	0.9337	0.0106	0.5020	0.9936	0.0259
ThP-Fr-23-3	0.4314	1.3046	0.0176	0.3464	1.1494	0.0339
ThP-Fr-3-3	0.3990	0.9337	0.0106	0.5101	1.0041	0.0264
AP-5	0.3012	0.7609	0.0063	0.0872	0.3195	0.0075
TNG-5	1.1003	1.9679	0.0271	1.2324	1.9094	0.0591
ThP-123-5	0.1730	0.9890	0.0102	0.1105	0.7234	0.0192
ThP-13-5	0.1941	1.0255	0.0107	0.1329	0.7811	0.0202
ThP-23-5	0.1730	0.9890	0.0102	0.1105	0.7234	0.0192
ThP-3-5	0.1944	1.0257	0.0107	0.1334	0.7817	0.0203
ThP-Fr-123-5	0.5905	1.7786	0.0232	0.5148	1.5976	0.0464
ThP-Fr-13-5	0.5734	1.3174	0.0152	0.7102	1.3884	0.0369
ThP-Fr-23-5	0.5905	1.7786	0.0232	0.4665	1.5991	0.0459
ThP-Fr-3-5	0.5734	1.3174	0.0152	0.7229	1.4069	0.0375
AP-10	0.4515	1.1757	0.0100	0.1304	0.4952	0.0118
TNG-10	1.3871	2.7493	0.0359	1.5329	2.5828	0.0771
ThP-123-10	0.2759	1.6517	0.0182	0.1926	1.1953	0.0324
ThP-13-10	0.2969	1.7001	0.0189	0.2230	1.2819	0.0337
ThP-23-10	0.2759	1.6517	0.0182	0.1926	1.1953	0.0324
ThP-3-10	0.2971	1.7004	0.0189	0.2234	1.2824	0.0338
ThP-Fr-123-10	0.7198	2.2682	0.0285	0.6540	2.2461	0.0631
ThP-Fr-13-10	0.9080	2.0734	0.0244	1.1222	2.1422	0.0579
ThP-Fr-23-10	0.7198	2.2682	0.0285	0.5683	2.2148	0.0615
ThP-Fr-3-10	0.9080	2.0734	0.0244	1.1479	2.1862	0.0594
AP-20	0.6959	1.8789	0.0164	0.1966	0.7965	0.0190
TNG-20	1.6674	3.7516	0.0466	1.7809	3.3720	0.0977
ThP-123-20	0.4828	2.6995	0.0304	0.3421	1.9828	0.0548
ThP-13-20	0.5024	2.8680	0.0332	0.3973	2.1721	0.0587
ThP-23-20	0.4828	2.6995	0.0304	0.3421	1.9828	0.0548
ThP-3-20	0.5031	2.8680	0.0332	0.3981	2.1725	0.0587
ThP-Fr-123-20	0.7368	2.3511	0.0293	0.6970	2.5852	0.0717
ThP-Fr-13-20	1.3473	3.2508	0.0386	1.5777	3.1773	0.0864
ThP-Fr-23-20	0.7368	2.3511	0.0293	0.5874	2.4498	0.0673
ThP-Fr-3-20	1.3473	3.2508	0.0386	1.6185	3.2765	0.0895

Table F.4: Methods Evaluation Abstract Phrase Metrics (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	phHIT	wdHIT	wdCOV	phHIT	wdHIT	wdCOV
AP-3	0.2243	0.5562	0.0045	0.0662	0.2345	0.0055
TNG-3	0.8841	1.4943	0.0215	1.0070	1.4841	0.0471
ThP-123-3	0.1281	0.6769	0.0066	0.0738	0.4866	0.0125
ThP-13-3	0.1435	0.7110	0.0070	0.0949	0.5404	0.0136
ThP-23-3	0.1281	0.6769	0.0066	0.0738	0.4866	0.0125
ThP-3-3	0.1435	0.7109	0.0070	0.0951	0.5408	0.0136
ThP-Fr-123-3	0.4301	1.3004	0.0176	0.3757	1.1468	0.0340
ThP-Fr-13-3	0.3997	0.9328	0.0105	0.5045	0.9925	0.0259
ThP-Fr-23-3	0.4301	1.3004	0.0176	0.3420	1.1401	0.0335
ThP-Fr-3-3	0.3997	0.9328	0.0105	0.5124	1.0026	0.0263
AP-5	0.3009	0.7595	0.0062	0.0866	0.3184	0.0076
TNG-5	1.0982	1.9664	0.0271	1.2455	1.9213	0.0592
ThP-123-5	0.1735	0.9806	0.0100	0.1082	0.7091	0.0188
ThP-13-5	0.1963	1.0265	0.0106	0.1346	0.7741	0.0199
ThP-23-5	0.1735	0.9806	0.0100	0.1082	0.7091	0.0188
ThP-3-5	0.1962	1.0265	0.0106	0.1350	0.7745	0.0199
ThP-Fr-123-5	0.5899	1.7760	0.0232	0.5119	1.5876	0.0460
ThP-Fr-13-5	0.5732	1.3146	0.0151	0.7143	1.3872	0.0368
ThP-Fr-23-5	0.5899	1.7760	0.0232	0.4627	1.5864	0.0455
ThP-Fr-3-5	0.5732	1.3146	0.0151	0.7275	1.4063	0.0375
AP-10	0.4495	1.1718	0.0099	0.1293	0.4945	0.0118
TNG-10	1.3910	2.7535	0.0359	1.5496	2.6008	0.0774
ThP-123-10	0.2752	1.6345	0.0179	0.1893	1.1781	0.0319
ThP-13-10	0.2998	1.6981	0.0187	0.2250	1.2705	0.0333
ThP-23-10	0.2752	1.6345	0.0179	0.1893	1.1781	0.0319
ThP-3-10	0.2998	1.6982	0.0187	0.2253	1.2712	0.0333
ThP-Fr-123-10	0.7203	2.2742	0.0285	0.6533	2.2403	0.0627
ThP-Fr-13-10	0.9103	2.0701	0.0243	1.1224	2.1336	0.0576
ThP-Fr-23-10	0.7203	2.2742	0.0285	0.5664	2.2101	0.0612
ThP-Fr-3-10	0.9103	2.0701	0.0243	1.1499	2.1793	0.0591
AP-20	0.6940	1.8726	0.0163	0.1953	0.7939	0.0190
TNG-20	1.6766	3.7620	0.0468	1.7990	3.3940	0.0981
ThP-123-20	0.4813	2.6806	0.0300	0.3371	1.9588	0.0540
ThP-13-20	0.5044	2.8628	0.0330	0.3959	2.1513	0.0580
ThP-23-20	0.4813	2.6806	0.0300	0.3371	1.9588	0.0540
ThP-3-20	0.5046	2.8627	0.0330	0.3966	2.1520	0.0580
ThP-Fr-123-20	0.7387	2.3639	0.0294	0.6984	2.5887	0.0715
ThP-Fr-13-20	1.3495	3.2479	0.0384	1.5785	3.1650	0.0859
ThP-Fr-23-20	0.7387	2.3639	0.0294	0.5867	2.4533	0.0672
ThP-Fr-3-20	1.3495	3.2479	0.0384	1.6205	3.2642	0.0890

Table F.5: Methods Evaluation Abstract Phrase Metrics (Segment Count = 25)

Appendix G

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : TITLE PH-COV AND PH-FMI

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0232	0.0255	0.0128	0.0191
TNG-5	0.0718	0.0789	0.0874	0.1334
ThP-123-5	0.0078	0.0085	0.0121	0.0199
ThP-13-5	0.0086	0.0093	0.0129	0.0207
ThP-23-5	0.0078	0.0085	0.0121	0.0199
ThP-3-5	0.0085	0.0093	0.0132	0.0211
ThP-Fr-123-5	0.0278	0.0289	0.0339	0.0461
ThP-Fr-13-5	0.0499	0.0559	0.0676	0.1065
ThP-Fr-23-5	0.0278	0.0289	0.0307	0.0420
ThP-Fr-3-5	0.0499	0.0559	0.0682	0.1079
AP-10	0.0301	0.0462	0.0162	0.0343
TNG-10	0.0673	0.1045	0.0852	0.1830
ThP-123-10	0.0122	0.0190	0.0177	0.0402
ThP-13-10	0.0125	0.0195	0.0199	0.0450
ThP-23-10	0.0122	0.0190	0.0177	0.0402
ThP-3-10	0.0125	0.0194	0.0196	0.0445
ThP-Fr-123-10	0.0291	0.0360	0.0375	0.0651
ThP-Fr-13-10	0.0691	0.1084	0.0840	0.1809
ThP-Fr-23-10	0.0291	0.0360	0.0323	0.0556
ThP-Fr-3-10	0.0691	0.1084	0.0852	0.1850
AP-20	0.0399	0.0860	0.0208	0.0619
TNG-20	0.0616	0.1353	0.0780	0.2364
ThP-123-20	0.0199	0.0438	0.0251	0.0778
ThP-13-20	0.0214	0.0476	0.0300	0.0944
ThP-23-20	0.0199	0.0438	0.0251	0.0778
ThP-3-20	0.0214	0.0474	0.0300	0.0944
ThP-Fr-123-20	0.0294	0.0373	0.0380	0.0725
ThP-Fr-13-20	0.0792	0.1741	0.0902	0.2659
ThP-Fr-23-20	0.0294	0.0373	0.0321	0.0585
ThP-Fr-3-20	0.0792	0.1741	0.0908	0.2723

Table G.1: Methods Evaluation : phCOV and phFMI With Title as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0222	0.0243	0.0123	0.0182
TNG-5	0.0753	0.0832	0.1001	0.1536
ThP-123-5	0.0076	0.0082	0.0121	0.0197
ThP-13-5	0.0084	0.0091	0.0129	0.0205
ThP-23-5	0.0076	0.0082	0.0121	0.0197
ThP-3-5	0.0082	0.0089	0.0129	0.0206
ThP-Fr-123-5	0.0273	0.0286	0.0333	0.0453
ThP-Fr-13-5	0.0497	0.0554	0.0672	0.1058
ThP-Fr-23-5	0.0273	0.0286	0.0301	0.0414
ThP-Fr-3-5	0.0497	0.0554	0.0678	0.1072
AP-10	0.0292	0.0449	0.0158	0.0334
TNG-10	0.0717	0.1112	0.0965	0.2077
ThP-123-10	0.0118	0.0184	0.0174	0.0395
ThP-13-10	0.0122	0.0190	0.0192	0.0429
ThP-23-10	0.0118	0.0184	0.0174	0.0395
ThP-3-10	0.0123	0.0191	0.0191	0.0430
ThP-Fr-123-10	0.0291	0.0368	0.0370	0.0653
ThP-Fr-13-10	0.0680	0.1066	0.0833	0.1795
ThP-Fr-23-10	0.0291	0.0368	0.0319	0.0559
ThP-Fr-3-10	0.0680	0.1066	0.0843	0.1828
AP-20	0.0392	0.0842	0.0205	0.0617
TNG-20	0.0647	0.1419	0.0851	0.2568
ThP-123-20	0.0193	0.0423	0.0249	0.0773
ThP-13-20	0.0213	0.0473	0.0297	0.0935
ThP-23-20	0.0193	0.0423	0.0249	0.0773
ThP-3-20	0.0213	0.0473	0.0296	0.0932
ThP-Fr-123-20	0.0295	0.0383	0.0374	0.0730
ThP-Fr-13-20	0.0782	0.1715	0.0908	0.2676
ThP-Fr-23-20	0.0295	0.0383	0.0316	0.0591
ThP-Fr-3-20	0.0782	0.1715	0.0909	0.2725

Table G.2: Methods Evaluation : phCOV and phFMI With Title as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0218	0.0238	0.0122	0.0181
TNG-5	0.0763	0.0844	0.1055	0.1623
ThP-123-5	0.0076	0.0082	0.0117	0.0192
ThP-13-5	0.0087	0.0095	0.0130	0.0205
ThP-23-5	0.0076	0.0082	0.0117	0.0192
ThP-3-5	0.0086	0.0094	0.0132	0.0210
ThP-Fr-123-5	0.0271	0.0286	0.0332	0.0453
ThP-Fr-13-5	0.0489	0.0542	0.0672	0.1062
ThP-Fr-23-5	0.0271	0.0286	0.0305	0.0419
ThP-Fr-3-5	0.0489	0.0542	0.0678	0.1073
AP-10	0.0285	0.0434	0.0161	0.0340
TNG-10	0.0723	0.1124	0.0998	0.2150
ThP-123-10	0.0113	0.0176	0.0170	0.0386
ThP-13-10	0.0123	0.0190	0.0189	0.0424
ThP-23-10	0.0113	0.0176	0.0170	0.0386
ThP-3-10	0.0124	0.0192	0.0187	0.0420
ThP-Fr-123-10	0.0291	0.0375	0.0367	0.0651
ThP-Fr-13-10	0.0677	0.1061	0.0834	0.1800
ThP-Fr-23-10	0.0291	0.0375	0.0316	0.0557
ThP-Fr-3-10	0.0677	0.1061	0.0844	0.1834
AP-20	0.0386	0.0828	0.0207	0.0619
TNG-20	0.0659	0.1443	0.0883	0.2659
ThP-123-20	0.0188	0.0412	0.0240	0.0749
ThP-13-20	0.0210	0.0465	0.0290	0.0909
ThP-23-20	0.0188	0.0412	0.0240	0.0749
ThP-3-20	0.0210	0.0465	0.0289	0.0908
ThP-Fr-123-20	0.0294	0.0391	0.0374	0.0736
ThP-Fr-13-20	0.0783	0.1719	0.0905	0.2668
ThP-Fr-23-20	0.0294	0.0391	0.0315	0.0595
ThP-Fr-3-20	0.0783	0.1719	0.0905	0.2715

Table G.3: Methods Evaluation : phCOV and phFMI With Title as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0221	0.0243	0.0123	0.0182
TNG-5	0.0763	0.0843	0.1064	0.1644
ThP-123-5	0.0074	0.0081	0.0117	0.0190
ThP-13-5	0.0086	0.0093	0.0127	0.0202
ThP-23-5	0.0074	0.0081	0.0117	0.0190
ThP-3-5	0.0087	0.0093	0.0131	0.0209
ThP-Fr-123-5	0.0270	0.0288	0.0329	0.0450
ThP-Fr-13-5	0.0493	0.0546	0.0662	0.1046
ThP-Fr-23-5	0.0270	0.0288	0.0300	0.0411
ThP-Fr-3-5	0.0493	0.0546	0.0670	0.1062
AP-10	0.0282	0.0431	0.0156	0.0329
TNG-10	0.0725	0.1127	0.1012	0.2180
ThP-123-10	0.0116	0.0181	0.0171	0.0388
ThP-13-10	0.0126	0.0194	0.0190	0.0427
ThP-23-10	0.0116	0.0181	0.0171	0.0388
ThP-3-10	0.0127	0.0196	0.0189	0.0424
ThP-Fr-123-10	0.0290	0.0379	0.0369	0.0657
ThP-Fr-13-10	0.0677	0.1060	0.0824	0.1777
ThP-Fr-23-10	0.0290	0.0379	0.0317	0.0564
ThP-Fr-3-10	0.0677	0.1060	0.0837	0.1821
AP-20	0.0388	0.0834	0.0202	0.0605
TNG-20	0.0657	0.1442	0.0894	0.2697
ThP-123-20	0.0189	0.0414	0.0245	0.0767
ThP-13-20	0.0212	0.0471	0.0287	0.0903
ThP-23-20	0.0189	0.0414	0.0245	0.0767
ThP-3-20	0.0213	0.0472	0.0287	0.0902
ThP-Fr-123-20	0.0294	0.0396	0.0375	0.0749
ThP-Fr-13-20	0.0782	0.1718	0.0904	0.2662
ThP-Fr-23-20	0.0294	0.0396	0.0317	0.0607
ThP-Fr-3-20	0.0782	0.1718	0.0909	0.2720

Table G.4: Methods Evaluation : phCOV and phFMI With Title as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ph-FMI	ph-COV	ph-FMI	ph-COV
AP-5	0.0216	0.0237	0.0120	0.0178
TNG-5	0.0765	0.0846	0.1072	0.1655
ThP-123-5	0.0072	0.0078	0.0117	0.0191
ThP-13-5	0.0085	0.0092	0.0126	0.0200
ThP-23-5	0.0072	0.0078	0.0117	0.0191
ThP-3-5	0.0083	0.0090	0.0127	0.0200
ThP-Fr-123-5	0.0267	0.0283	0.0332	0.0454
ThP-Fr-13-5	0.0491	0.0543	0.0677	0.1070
ThP-Fr-23-5	0.0267	0.0283	0.0306	0.0422
ThP-Fr-3-5	0.0491	0.0543	0.0684	0.1084
AP-10	0.0279	0.0427	0.0158	0.0332
TNG-10	0.0728	0.1131	0.1031	0.2223
ThP-123-10	0.0112	0.0174	0.0172	0.0390
ThP-13-10	0.0122	0.0188	0.0192	0.0432
ThP-23-10	0.0112	0.0174	0.0172	0.0390
ThP-3-10	0.0122	0.0188	0.0191	0.0430
ThP-Fr-123-10	0.0288	0.0377	0.0371	0.0663
ThP-Fr-13-10	0.0672	0.1051	0.0834	0.1800
ThP-Fr-23-10	0.0288	0.0377	0.0318	0.0566
ThP-Fr-3-10	0.0672	0.1051	0.0844	0.1837
AP-20	0.0382	0.0822	0.0204	0.0603
TNG-20	0.0664	0.1455	0.0904	0.2724
ThP-123-20	0.0186	0.0410	0.0240	0.0749
ThP-13-20	0.0208	0.0460	0.0286	0.0896
ThP-23-20	0.0186	0.0410	0.0240	0.0749
ThP-3-20	0.0208	0.0461	0.0286	0.0896
ThP-Fr-123-20	0.0291	0.0393	0.0375	0.0748
ThP-Fr-13-20	0.0785	0.1720	0.0909	0.2684
ThP-Fr-23-20	0.0291	0.0393	0.0317	0.0609
ThP-Fr-3-20	0.0785	0.1720	0.0916	0.2749

Table G.5: Methods Evaluation : phCOV and phFMI With Title as Gold Standard (Segment Count = 25)

Appendix H

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : TITLE SUB-COV AND SUB-FMI

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0724	0.0973	0.0360	0.0600
TNG-5	0.1728	0.2214	0.1558	0.2348
ThP-123-5	0.0168	0.0204	0.0190	0.0325
ThP-13-5	0.0187	0.0221	0.0214	0.0351
ThP-23-5	0.0168	0.0204	0.0190	0.0325
ThP-3-5	0.0187	0.0220	0.0217	0.0356
ThP-Fr-123-5	0.2053	0.2107	0.2051	0.3028
ThP-Fr-13-5	0.0764	0.0854	0.0970	0.1566
ThP-Fr-23-5	0.2053	0.2107	0.2032	0.3006
ThP-Fr-3-5	0.0764	0.0854	0.0990	0.1605
AP-10	0.1091	0.2084	0.0496	0.1144
TNG-10	0.1671	0.2959	0.1515	0.3109
ThP-123-10	0.0243	0.0414	0.0273	0.0636
ThP-13-10	0.0257	0.0419	0.0315	0.0717
ThP-23-10	0.0243	0.0414	0.0273	0.0636
ThP-3-10	0.0257	0.0418	0.0312	0.0710
ThP-Fr-123-10	0.2394	0.2753	0.2552	0.4449
ThP-Fr-13-10	0.1123	0.1696	0.1285	0.2767
ThP-Fr-23-10	0.2394	0.2753	0.2481	0.4284
ThP-Fr-3-10	0.1123	0.1696	0.1316	0.2851
AP-20	0.1645	0.4187	0.0678	0.2138
TNG-20	0.1560	0.3773	0.1384	0.3904
ThP-123-20	0.0400	0.0930	0.0411	0.1285
ThP-13-20	0.0414	0.0922	0.0473	0.1448
ThP-23-20	0.0400	0.0930	0.0411	0.1285
ThP-3-20	0.0412	0.0917	0.0473	0.1446
ThP-Fr-123-20	0.2434	0.2832	0.2686	0.4978
ThP-Fr-13-20	0.1428	0.2872	0.1524	0.4253
ThP-Fr-23-20	0.2434	0.2832	0.2557	0.4603
ThP-Fr-3-20	0.1428	0.2872	0.1550	0.4376

Table H.1: Methods Evaluation : subCOV and subFMI With Title as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0701	0.0939	0.0356	0.0589
TNG-5	0.1827	0.2329	0.1765	0.2628
ThP-123-5	0.0162	0.0197	0.0193	0.0331
ThP-13-5	0.0184	0.0216	0.0210	0.0341
ThP-23-5	0.0162	0.0197	0.0193	0.0331
ThP-3-5	0.0184	0.0216	0.0212	0.0345
ThP-Fr-123-5	0.2000	0.2077	0.2001	0.2977
ThP-Fr-13-5	0.0762	0.0850	0.0965	0.1560
ThP-Fr-23-5	0.2000	0.2077	0.1988	0.2966
ThP-Fr-3-5	0.0762	0.0850	0.0983	0.1593
AP-10	0.1065	0.2030	0.0492	0.1139
TNG-10	0.1780	0.3102	0.1702	0.3435
ThP-123-10	0.0232	0.0395	0.0274	0.0647
ThP-13-10	0.0250	0.0407	0.0304	0.0688
ThP-23-10	0.0232	0.0395	0.0274	0.0647
ThP-3-10	0.0251	0.0408	0.0305	0.0690
ThP-Fr-123-10	0.2366	0.2783	0.2516	0.4458
ThP-Fr-13-10	0.1099	0.1661	0.1272	0.2744
ThP-Fr-23-10	0.2366	0.2783	0.2455	0.4306
ThP-Fr-3-10	0.1099	0.1661	0.1303	0.2824
AP-20	0.1629	0.4150	0.0675	0.2136
TNG-20	0.1634	0.3904	0.1521	0.4192
ThP-123-20	0.0380	0.0892	0.0405	0.1273
ThP-13-20	0.0405	0.0902	0.0469	0.1433
ThP-23-20	0.0380	0.0892	0.0405	0.1273
ThP-3-20	0.0405	0.0904	0.0466	0.1424
ThP-Fr-123-20	0.2417	0.2888	0.2666	0.5034
ThP-Fr-13-20	0.1406	0.2821	0.1519	0.4250
ThP-Fr-23-20	0.2417	0.2888	0.2541	0.4664
ThP-Fr-3-20	0.1406	0.2821	0.1545	0.4371

Table H.2: Methods Evaluation : subCOV and subFMI With Title as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0689	0.0923	0.0356	0.0597
TNG-5	0.1853	0.2346	0.1842	0.2742
ThP-123-5	0.0163	0.0197	0.0191	0.0328
ThP-13-5	0.0188	0.0219	0.0218	0.0349
ThP-23-5	0.0163	0.0197	0.0191	0.0328
ThP-3-5	0.0187	0.0217	0.0219	0.0352
ThP-Fr-123-5	0.1967	0.2061	0.1974	0.2938
ThP-Fr-13-5	0.0751	0.0833	0.0963	0.1561
ThP-Fr-23-5	0.1967	0.2061	0.1965	0.2937
ThP-Fr-3-5	0.0751	0.0833	0.0981	0.1596
AP-10	0.1053	0.2021	0.0482	0.1105
TNG-10	0.1800	0.3117	0.1750	0.3515
ThP-123-10	0.0227	0.0384	0.0268	0.0631
ThP-13-10	0.0250	0.0405	0.0307	0.0693
ThP-23-10	0.0227	0.0384	0.0268	0.0631
ThP-3-10	0.0252	0.0407	0.0305	0.0687
ThP-Fr-123-10	0.2348	0.2808	0.2499	0.4458
ThP-Fr-13-10	0.1096	0.1657	0.1261	0.2725
ThP-Fr-23-10	0.2348	0.2808	0.2434	0.4297
ThP-Fr-3-10	0.1096	0.1657	0.1289	0.2799
AP-20	0.1606	0.4101	0.0672	0.2130
TNG-20	0.1659	0.3924	0.1575	0.4315
ThP-123-20	0.0375	0.0879	0.0393	0.1236
ThP-13-20	0.0404	0.0896	0.0464	0.1410
ThP-23-20	0.0375	0.0879	0.0393	0.1236
ThP-3-20	0.0403	0.0895	0.0461	0.1400
ThP-Fr-123-20	0.2403	0.2924	0.2652	0.5058
ThP-Fr-13-20	0.1400	0.2812	0.1514	0.4249
ThP-Fr-23-20	0.2403	0.2924	0.2529	0.4682
ThP-Fr-3-20	0.1400	0.2812	0.1537	0.4363

Table H.3: Methods Evaluation : subCOV and subFMI With Title as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0695	0.0931	0.0354	0.0589
TNG-5	0.1867	0.2362	0.1874	0.2783
ThP-123-5	0.0159	0.0194	0.0189	0.0323
ThP-13-5	0.0181	0.0210	0.0210	0.0341
ThP-23-5	0.0159	0.0194	0.0189	0.0323
ThP-3-5	0.0182	0.0211	0.0214	0.0348
ThP-Fr-123-5	0.1935	0.2039	0.1957	0.2925
ThP-Fr-13-5	0.0755	0.0841	0.0956	0.1546
ThP-Fr-23-5	0.1935	0.2039	0.1949	0.2924
ThP-Fr-3-5	0.0755	0.0841	0.0974	0.1580
AP-10	0.1051	0.2015	0.0475	0.1095
TNG-10	0.1819	0.3156	0.1786	0.3581
ThP-123-10	0.0225	0.0377	0.0266	0.0626
ThP-13-10	0.0250	0.0402	0.0306	0.0688
ThP-23-10	0.0225	0.0377	0.0266	0.0626
ThP-3-10	0.0251	0.0403	0.0303	0.0682
ThP-Fr-123-10	0.2328	0.2818	0.2489	0.4468
ThP-Fr-13-10	0.1091	0.1654	0.1250	0.2700
ThP-Fr-23-10	0.2328	0.2818	0.2424	0.4313
ThP-Fr-3-10	0.1091	0.1654	0.1283	0.2788
AP-20	0.1613	0.4121	0.0664	0.2105
TNG-20	0.1665	0.3949	0.1598	0.4359
ThP-123-20	0.0369	0.0864	0.0394	0.1250
ThP-13-20	0.0401	0.0890	0.0458	0.1403
ThP-23-20	0.0369	0.0864	0.0394	0.1250
ThP-3-20	0.0401	0.0891	0.0458	0.1402
ThP-Fr-123-20	0.2389	0.2949	0.2644	0.5096
ThP-Fr-13-20	0.1394	0.2812	0.1505	0.4229
ThP-Fr-23-20	0.2389	0.2949	0.2517	0.4708
ThP-Fr-3-20	0.1394	0.2812	0.1532	0.4352

Table H.4: Methods Evaluation : subCOV and subFMI With Title as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	sub-FMI	sub-COV	sub-FMI	sub-COV
AP-5	0.0684	0.0921	0.0345	0.0578
TNG-5	0.1881	0.2381	0.1880	0.2785
ThP-123-5	0.0155	0.0190	0.0187	0.0318
ThP-13-5	0.0184	0.0214	0.0206	0.0336
ThP-23-5	0.0155	0.0190	0.0187	0.0318
ThP-3-5	0.0184	0.0215	0.0207	0.0335
ThP-Fr-123-5	0.1927	0.2034	0.1952	0.2923
ThP-Fr-13-5	0.0754	0.0837	0.0970	0.1574
ThP-Fr-23-5	0.1927	0.2034	0.1937	0.2913
ThP-Fr-3-5	0.0754	0.0837	0.0991	0.1611
AP-10	0.1049	0.2014	0.0477	0.1106
TNG-10	0.1814	0.3132	0.1813	0.3627
ThP-123-10	0.0224	0.0378	0.0266	0.0624
ThP-13-10	0.0250	0.0403	0.0305	0.0684
ThP-23-10	0.0224	0.0378	0.0266	0.0624
ThP-3-10	0.0252	0.0404	0.0304	0.0682
ThP-Fr-123-10	0.2317	0.2821	0.2481	0.4478
ThP-Fr-13-10	0.1089	0.1648	0.1260	0.2727
ThP-Fr-23-10	0.2317	0.2821	0.2407	0.4305
ThP-Fr-3-10	0.1089	0.1648	0.1290	0.2806
AP-20	0.1600	0.4079	0.0674	0.2135
TNG-20	0.1683	0.3968	0.1609	0.4376
ThP-123-20	0.0371	0.0873	0.0388	0.1225
ThP-13-20	0.0400	0.0891	0.0458	0.1399
ThP-23-20	0.0371	0.0873	0.0388	0.1225
ThP-3-20	0.0401	0.0892	0.0457	0.1394
ThP-Fr-123-20	0.2380	0.2957	0.2638	0.5117
ThP-Fr-13-20	0.1396	0.2812	0.1504	0.4233
ThP-Fr-23-20	0.2380	0.2957	0.2512	0.4737
ThP-Fr-3-20	0.1396	0.2812	0.1535	0.4368

Table H.5: Methods Evaluation : subCOV and subFMI With Title as Gold Standard (Segment Count = 25)

Appendix I

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : TITLE EXT-COV AND EXT-FMI

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0362	0.0386	0.0298	0.0414
TNG-5	0.1322	0.1340	0.1531	0.2030
ThP-123-5	0.0741	0.0719	0.0807	0.1036
ThP-13-5	0.0747	0.0692	0.0791	0.0977
ThP-23-5	0.0741	0.0719	0.0807	0.1036
ThP-3-5	0.0747	0.0692	0.0792	0.0981
ThP-Fr-123-5	0.0646	0.0619	0.0748	0.0925
ThP-Fr-13-5	0.1072	0.1093	0.1444	0.2044
ThP-Fr-23-5	0.0646	0.0619	0.0705	0.0881
ThP-Fr-3-5	0.1072	0.1093	0.1459	0.2068
AP-10	0.0490	0.0711	0.0392	0.0731
TNG-10	0.1430	0.1901	0.1644	0.2821
ThP-123-10	0.1135	0.1405	0.1088	0.1782
ThP-13-10	0.1140	0.1335	0.1113	0.1777
ThP-23-10	0.1135	0.1405	0.1088	0.1782
ThP-3-10	0.1139	0.1334	0.1112	0.1776
ThP-Fr-123-10	0.0642	0.0693	0.0758	0.1102
ThP-Fr-13-10	0.1405	0.1829	0.1688	0.3030
ThP-Fr-23-10	0.0642	0.0693	0.0712	0.1025
ThP-Fr-3-10	0.1405	0.1829	0.1698	0.3061
AP-20	0.0679	0.1320	0.0535	0.1319
TNG-20	0.1540	0.2604	0.1694	0.3715
ThP-123-20	0.1518	0.2277	0.1368	0.2712
ThP-13-20	0.1668	0.2392	0.1486	0.2911
ThP-23-20	0.1518	0.2277	0.1368	0.2712
ThP-3-20	0.1668	0.2392	0.1485	0.2908
ThP-Fr-123-20	0.0643	0.0701	0.0779	0.1180
ThP-Fr-13-20	0.1690	0.2556	0.1779	0.3778
ThP-Fr-23-20	0.0643	0.0701	0.0724	0.1074
ThP-Fr-3-20	0.1690	0.2556	0.1788	0.3816

Table I.1: Methods Evaluation : extCOV and extFMI With Title as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0346	0.0366	0.0285	0.0392
TNG-5	0.1363	0.1378	0.1659	0.2248
ThP-123-5	0.0699	0.0681	0.0792	0.1024
ThP-13-5	0.0699	0.0648	0.0771	0.0953
ThP-23-5	0.0699	0.0681	0.0792	0.1024
ThP-3-5	0.0701	0.0649	0.0770	0.0953
ThP-Fr-123-5	0.0649	0.0627	0.0736	0.0916
ThP-Fr-13-5	0.1054	0.1077	0.1424	0.2021
ThP-Fr-23-5	0.0649	0.0627	0.0698	0.0876
ThP-Fr-3-5	0.1054	0.1077	0.1440	0.2050
AP-10	0.0475	0.0688	0.0381	0.0711
TNG-10	0.1480	0.1968	0.1732	0.3062
ThP-123-10	0.1080	0.1354	0.1067	0.1762
ThP-13-10	0.1096	0.1290	0.1081	0.1721
ThP-23-10	0.1080	0.1354	0.1067	0.1762
ThP-3-10	0.1096	0.1292	0.1082	0.1722
ThP-Fr-123-10	0.0648	0.0716	0.0749	0.1105
ThP-Fr-13-10	0.1372	0.1792	0.1671	0.3009
ThP-Fr-23-10	0.0648	0.0716	0.0703	0.1032
ThP-Fr-3-10	0.1372	0.1792	0.1680	0.3035
AP-20	0.0662	0.1287	0.0529	0.1314
TNG-20	0.1570	0.2661	0.1745	0.3920
ThP-123-20	0.1473	0.2249	0.1349	0.2720
ThP-13-20	0.1620	0.2344	0.1463	0.2888
ThP-23-20	0.1473	0.2249	0.1349	0.2720
ThP-3-20	0.1620	0.2345	0.1463	0.2886
ThP-Fr-123-20	0.0647	0.0726	0.0772	0.1196
ThP-Fr-13-20	0.1658	0.2532	0.1774	0.3805
ThP-Fr-23-20	0.0647	0.0726	0.0720	0.1091
ThP-Fr-3-20	0.1658	0.2532	0.1780	0.3839

Table I.2: Methods Evaluation : extCOV and extFMI With Title as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0343	0.0362	0.0283	0.0385
TNG-5	0.1372	0.1392	0.1718	0.2345
ThP-123-5	0.0684	0.0667	0.0766	0.1005
ThP-13-5	0.0692	0.0644	0.0743	0.0917
ThP-23-5	0.0684	0.0667	0.0766	0.1005
ThP-3-5	0.0694	0.0646	0.0744	0.0920
ThP-Fr-123-5	0.0649	0.0630	0.0743	0.0929
ThP-Fr-13-5	0.1034	0.1058	0.1429	0.2033
ThP-Fr-23-5	0.0649	0.0630	0.0708	0.0891
ThP-Fr-3-5	0.1034	0.1058	0.1439	0.2052
AP-10	0.0466	0.0672	0.0388	0.0725
TNG-10	0.1485	0.1973	0.1773	0.3151
ThP-123-10	0.1052	0.1327	0.1038	0.1744
ThP-13-10	0.1080	0.1284	0.1046	0.1686
ThP-23-10	0.1052	0.1327	0.1038	0.1744
ThP-3-10	0.1081	0.1284	0.1046	0.1688
ThP-Fr-123-10	0.0645	0.0723	0.0750	0.1119
ThP-Fr-13-10	0.1358	0.1781	0.1672	0.3027
ThP-Fr-23-10	0.0645	0.0723	0.0705	0.1047
ThP-Fr-3-10	0.1358	0.1781	0.1681	0.3052
AP-20	0.0651	0.1263	0.0536	0.1333
TNG-20	0.1579	0.2675	0.1770	0.3999
ThP-123-20	0.1450	0.2246	0.1326	0.2704
ThP-13-20	0.1597	0.2337	0.1433	0.2858
ThP-23-20	0.1450	0.2246	0.1326	0.2704
ThP-3-20	0.1597	0.2335	0.1434	0.2862
ThP-Fr-123-20	0.0645	0.0735	0.0771	0.1207
ThP-Fr-13-20	0.1650	0.2533	0.1769	0.3819
ThP-Fr-23-20	0.0645	0.0735	0.0719	0.1101
ThP-Fr-3-20	0.1650	0.2533	0.1774	0.3849

Table I.3: Methods Evaluation : extCOV and extFMI With Title as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0346	0.0367	0.0279	0.0384
TNG-5	0.1373	0.1392	0.1717	0.2357
ThP-123-5	0.0660	0.0653	0.0752	0.0985
ThP-13-5	0.0679	0.0634	0.0748	0.0922
ThP-23-5	0.0660	0.0653	0.0752	0.0985
ThP-3-5	0.0679	0.0633	0.0751	0.0925
ThP-Fr-123-5	0.0652	0.0636	0.0746	0.0929
ThP-Fr-13-5	0.1027	0.1053	0.1407	0.2007
ThP-Fr-23-5	0.0652	0.0636	0.0704	0.0885
ThP-Fr-3-5	0.1027	0.1053	0.1420	0.2030
AP-10	0.0462	0.0667	0.0376	0.0696
TNG-10	0.1493	0.1978	0.1786	0.3190
ThP-123-10	0.1042	0.1313	0.1027	0.1721
ThP-13-10	0.1060	0.1255	0.1047	0.1690
ThP-23-10	0.1042	0.1313	0.1027	0.1721
ThP-3-10	0.1060	0.1254	0.1047	0.1690
ThP-Fr-123-10	0.0646	0.0733	0.0752	0.1123
ThP-Fr-13-10	0.1349	0.1773	0.1653	0.2997
ThP-Fr-23-10	0.0646	0.0733	0.0706	0.1054
ThP-Fr-3-10	0.1349	0.1773	0.1666	0.3035
AP-20	0.0656	0.1276	0.0524	0.1303
TNG-20	0.1580	0.2678	0.1781	0.4039
ThP-123-20	0.1432	0.2240	0.1325	0.2717
ThP-13-20	0.1573	0.2305	0.1431	0.2847
ThP-23-20	0.1432	0.2240	0.1325	0.2717
ThP-3-20	0.1573	0.2307	0.1430	0.2846
ThP-Fr-123-20	0.0645	0.0747	0.0775	0.1226
ThP-Fr-13-20	0.1642	0.2535	0.1756	0.3822
ThP-Fr-23-20	0.0645	0.0747	0.0722	0.1118
ThP-Fr-3-20	0.1642	0.2535	0.1763	0.3860

Table I.4: Methods Evaluation : extCOV and extFMI With Title as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB		DATASET = PATENT_RAND15K	
	ext-FMI	ext-COV	ext-FMI	ext-COV
AP-5	0.0336	0.0356	0.0284	0.0393
TNG-5	0.1371	0.1388	0.1734	0.2383
ThP-123-5	0.0635	0.0631	0.0734	0.0961
ThP-13-5	0.0667	0.0623	0.0725	0.0893
ThP-23-5	0.0635	0.0631	0.0734	0.0961
ThP-3-5	0.0667	0.0622	0.0728	0.0897
ThP-Fr-123-5	0.0644	0.0629	0.0748	0.0938
ThP-Fr-13-5	0.1017	0.1042	0.1421	0.2028
ThP-Fr-23-5	0.0644	0.0629	0.0709	0.0898
ThP-Fr-3-5	0.1017	0.1042	0.1432	0.2052
AP-10	0.0456	0.0659	0.0385	0.0723
TNG-10	0.1493	0.1984	0.1802	0.3215
ThP-123-10	0.1025	0.1304	0.1016	0.1711
ThP-13-10	0.1045	0.1242	0.1037	0.1664
ThP-23-10	0.1025	0.1304	0.1016	0.1711
ThP-3-10	0.1045	0.1241	0.1038	0.1665
ThP-Fr-123-10	0.0641	0.0728	0.0756	0.1137
ThP-Fr-13-10	0.1338	0.1759	0.1658	0.3023
ThP-Fr-23-10	0.0641	0.0728	0.0706	0.1061
ThP-Fr-3-10	0.1338	0.1759	0.1668	0.3049
AP-20	0.0644	0.1256	0.0522	0.1289
TNG-20	0.1589	0.2687	0.1783	0.4046
ThP-123-20	0.1415	0.2220	0.1304	0.2689
ThP-13-20	0.1565	0.2303	0.1416	0.2839
ThP-23-20	0.1415	0.2220	0.1304	0.2689
ThP-3-20	0.1565	0.2304	0.1415	0.2841
ThP-Fr-123-20	0.0638	0.0740	0.0775	0.1230
ThP-Fr-13-20	0.1642	0.2539	0.1762	0.3846
ThP-Fr-23-20	0.0638	0.0740	0.0721	0.1120
ThP-Fr-3-20	0.1642	0.2539	0.1772	0.3893

Table I.5: Methods Evaluation : extCOV and extFMI With Title as Gold Standard (Segment Count = 25)

Appendix J

THEMATIC PHRASES QUANTITATIVE METRICS

TABLES : TITLE WORD GRANULARITY METRICS

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.0824	0.0896	0.0469	0.0876	0.0612	0.0935	0.0330	0.0661
TNG-5	0.2286	0.2407	0.1373	0.2447	0.2642	0.3722	0.1560	0.2931
ThP-123-5	0.1208	0.1523	0.0656	0.1334	0.1376	0.2362	0.0690	0.1527
ThP-13-5	0.1184	0.1455	0.0645	0.1295	0.1360	0.2228	0.0697	0.1464
ThP-23-5	0.1208	0.1523	0.0656	0.1334	0.1376	0.2362	0.0690	0.1527
ThP-3-5	0.1185	0.1454	0.0645	0.1295	0.1362	0.2232	0.0699	0.1466
ThP-Fr-123-5	0.1919	0.1714	0.1087	0.2152	0.2173	0.2775	0.1267	0.2301
ThP-Fr-13-5	0.1866	0.2252	0.1082	0.2162	0.2425	0.3984	0.1317	0.2613
ThP-Fr-23-5	0.1919	0.1714	0.1087	0.2152	0.2157	0.2753	0.1257	0.2280
ThP-Fr-3-5	0.1866	0.2252	0.1082	0.2162	0.2445	0.4030	0.1325	0.2634
AP-10	0.1071	0.1594	0.0560	0.1178	0.0764	0.1592	0.0342	0.0866
TNG-10	0.2306	0.3351	0.1275	0.2648	0.2565	0.4861	0.1275	0.3104
ThP-123-10	0.1682	0.2840	0.0835	0.2005	0.1718	0.3865	0.0731	0.2065
ThP-13-10	0.1654	0.2751	0.0826	0.1975	0.1722	0.3727	0.0756	0.2008
ThP-23-10	0.1682	0.2840	0.0835	0.2005	0.1718	0.3865	0.0731	0.2065
ThP-3-10	0.1654	0.2750	0.0826	0.1974	0.1721	0.3724	0.0755	0.2007
ThP-Fr-123-10	0.2039	0.2034	0.1157	0.2296	0.2375	0.3537	0.1318	0.2528
ThP-Fr-13-10	0.2199	0.3369	0.1185	0.2561	0.2628	0.5386	0.1239	0.2911
ThP-Fr-23-10	0.2039	0.2034	0.1157	0.2296	0.2333	0.3447	0.1295	0.2490
ThP-Fr-3-10	0.2199	0.3369	0.1185	0.2561	0.2638	0.5437	0.1239	0.2923
AP-20	0.1397	0.2820	0.0625	0.1624	0.0972	0.2768	0.0342	0.1175
TNG-20	0.2213	0.4446	0.1018	0.2821	0.2325	0.6006	0.0909	0.3198
ThP-123-20	0.1957	0.4197	0.0851	0.2580	0.1907	0.5349	0.0686	0.2555
ThP-13-20	0.2032	0.4316	0.0892	0.2712	0.1968	0.5356	0.0729	0.2584
ThP-23-20	0.1957	0.4197	0.0851	0.2580	0.1907	0.5349	0.0686	0.2555
ThP-3-20	0.2031	0.4314	0.0892	0.2712	0.1967	0.5355	0.0729	0.2583
ThP-Fr-123-20	0.2046	0.2064	0.1159	0.2306	0.2406	0.3804	0.1299	0.2596
ThP-Fr-13-20	0.2309	0.4388	0.1107	0.2852	0.2542	0.6230	0.1042	0.2995
ThP-Fr-23-20	0.2046	0.2064	0.1159	0.2306	0.2350	0.3618	0.1281	0.2532
ThP-Fr-3-20	0.2309	0.4388	0.1107	0.2852	0.2535	0.6274	0.1030	0.2985

Table J.1: Methods Evaluation : Word Granularity Metrics With Title as Gold Standard (Segment Count = 5)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.0799	0.0864	0.0455	0.0847	0.0599	0.0909	0.0323	0.0646
TNG-5	0.2354	0.2471	0.1420	0.2513	0.2795	0.3940	0.1664	0.3066
ThP-123-5	0.1145	0.1445	0.0620	0.1257	0.1346	0.2315	0.0673	0.1488
ThP-13-5	0.1125	0.1375	0.0612	0.1220	0.1322	0.2162	0.0678	0.1414
ThP-23-5	0.1145	0.1445	0.0620	0.1257	0.1346	0.2315	0.0673	0.1488
ThP-3-5	0.1126	0.1375	0.0612	0.1220	0.1323	0.2164	0.0679	0.1415
ThP-Fr-123-5	0.1904	0.1724	0.1082	0.2134	0.2137	0.2745	0.1242	0.2264
ThP-Fr-13-5	0.1843	0.2222	0.1069	0.2129	0.2399	0.3943	0.1301	0.2583
ThP-Fr-23-5	0.1904	0.1724	0.1082	0.2134	0.2124	0.2728	0.1233	0.2247
ThP-Fr-3-5	0.1843	0.2222	0.1069	0.2129	0.2419	0.3991	0.1310	0.2605
AP-10	0.1041	0.1547	0.0545	0.1146	0.0755	0.1575	0.0337	0.0852
TNG-10	0.2379	0.3426	0.1324	0.2710	0.2691	0.5088	0.1347	0.3203
ThP-123-10	0.1629	0.2766	0.0804	0.1932	0.1694	0.3834	0.0716	0.2025
ThP-13-10	0.1615	0.2688	0.0804	0.1912	0.1688	0.3647	0.0741	0.1943
ThP-23-10	0.1629	0.2766	0.0804	0.1932	0.1694	0.3834	0.0716	0.2025
ThP-3-10	0.1615	0.2688	0.0804	0.1912	0.1688	0.3648	0.0741	0.1944
ThP-Fr-123-10	0.2032	0.2082	0.1155	0.2291	0.2347	0.3541	0.1291	0.2503
ThP-Fr-13-10	0.2169	0.3327	0.1167	0.2512	0.2611	0.5373	0.1227	0.2888
ThP-Fr-23-10	0.2032	0.2082	0.1155	0.2291	0.2310	0.3462	0.1273	0.2470
ThP-Fr-3-10	0.2169	0.3327	0.1167	0.2512	0.2622	0.5424	0.1228	0.2904
AP-20	0.1374	0.2767	0.0615	0.1588	0.0968	0.2760	0.0341	0.1162
TNG-20	0.2262	0.4496	0.1050	0.2849	0.2427	0.6194	0.0961	0.3262
ThP-123-20	0.1930	0.4175	0.0832	0.2519	0.1889	0.5359	0.0673	0.2519
ThP-13-20	0.2004	0.4276	0.0876	0.2657	0.1955	0.5327	0.0724	0.2540
ThP-23-20	0.1930	0.4175	0.0832	0.2519	0.1889	0.5359	0.0673	0.2519
ThP-3-20	0.2004	0.4277	0.0876	0.2657	0.1955	0.5326	0.0724	0.2540
ThP-Fr-123-20	0.2040	0.2117	0.1157	0.2303	0.2387	0.3845	0.1275	0.2585
ThP-Fr-13-20	0.2294	0.4373	0.1097	0.2807	0.2538	0.6259	0.1035	0.2977
ThP-Fr-23-20	0.2040	0.2117	0.1157	0.2303	0.2334	0.3663	0.1260	0.2522
ThP-Fr-3-20	0.2294	0.4373	0.1097	0.2807	0.2531	0.6303	0.1023	0.2966

Table J.2: Methods Evaluation : Word Granularity Metrics With Title as Gold Standard (Segment Count = 10)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.0788	0.0853	0.0448	0.0836	0.0593	0.0901	0.0318	0.0644
TNG-5	0.2368	0.2480	0.1432	0.2522	0.2852	0.4029	0.1701	0.3120
ThP-123-5	0.1112	0.1403	0.0602	0.1219	0.1311	0.2258	0.0654	0.1436
ThP-13-5	0.1105	0.1345	0.0601	0.1189	0.1285	0.2092	0.0660	0.1370
ThP-23-5	0.1112	0.1403	0.0602	0.1219	0.1311	0.2258	0.0654	0.1436
ThP-3-5	0.1105	0.1347	0.0601	0.1189	0.1287	0.2096	0.0661	0.1372
ThP-Fr-123-5	0.1883	0.1719	0.1070	0.2117	0.2139	0.2759	0.1241	0.2265
ThP-Fr-13-5	0.1821	0.2193	0.1056	0.2102	0.2405	0.3961	0.1303	0.2584
ThP-Fr-23-5	0.1883	0.1719	0.1070	0.2117	0.2122	0.2742	0.1231	0.2248
ThP-Fr-3-5	0.1821	0.2193	0.1056	0.2102	0.2419	0.4001	0.1308	0.2600
AP-10	0.1027	0.1525	0.0537	0.1129	0.0753	0.1570	0.0335	0.0848
TNG-10	0.2401	0.3448	0.1341	0.2722	0.2737	0.5173	0.1371	0.3240
ThP-123-10	0.1600	0.2727	0.0788	0.1889	0.1665	0.3782	0.0701	0.1981
ThP-13-10	0.1595	0.2658	0.0793	0.1870	0.1656	0.3576	0.0725	0.1897
ThP-23-10	0.1600	0.2727	0.0788	0.1889	0.1665	0.3782	0.0701	0.1981
ThP-3-10	0.1596	0.2659	0.0794	0.1871	0.1656	0.3576	0.0725	0.1897
ThP-Fr-123-10	0.2014	0.2093	0.1144	0.2276	0.2345	0.3569	0.1287	0.2506
ThP-Fr-13-10	0.2155	0.3308	0.1159	0.2489	0.2609	0.5380	0.1224	0.2882
ThP-Fr-23-10	0.2014	0.2093	0.1144	0.2276	0.2307	0.3492	0.1267	0.2470
ThP-Fr-3-10	0.2155	0.3308	0.1159	0.2489	0.2618	0.5428	0.1223	0.2894
AP-20	0.1354	0.2728	0.0606	0.1565	0.0964	0.2752	0.0339	0.1159
TNG-20	0.2283	0.4517	0.1065	0.2864	0.2464	0.6272	0.0978	0.3287
ThP-123-20	0.1917	0.4182	0.0822	0.2483	0.1873	0.5345	0.0664	0.2483
ThP-13-20	0.1997	0.4275	0.0871	0.2626	0.1939	0.5293	0.0716	0.2490
ThP-23-20	0.1917	0.4182	0.0822	0.2483	0.1873	0.5345	0.0664	0.2483
ThP-3-20	0.1996	0.4275	0.0871	0.2626	0.1939	0.5293	0.0716	0.2489
ThP-Fr-123-20	0.2025	0.2137	0.1147	0.2289	0.2383	0.3881	0.1267	0.2582
ThP-Fr-13-20	0.2285	0.4364	0.1092	0.2783	0.2538	0.6278	0.1032	0.2970
ThP-Fr-23-20	0.2025	0.2137	0.1147	0.2289	0.2328	0.3695	0.1251	0.2521
ThP-Fr-3-20	0.2285	0.4364	0.1092	0.2783	0.2531	0.6320	0.1020	0.2959

Table J.3: Methods Evaluation : Word Granularity Metrics With Title as Gold Standard (Segment Count = 15)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.0789	0.0853	0.0449	0.0839	0.0586	0.0888	0.0315	0.0632
TNG-5	0.2374	0.2480	0.1434	0.2516	0.2880	0.4070	0.1720	0.3132
ThP-123-5	0.1082	0.1371	0.0583	0.1180	0.1300	0.2242	0.0649	0.1426
ThP-13-5	0.1074	0.1307	0.0583	0.1151	0.1277	0.2085	0.0654	0.1361
ThP-23-5	0.1082	0.1371	0.0583	0.1180	0.1300	0.2242	0.0649	0.1426
ThP-3-5	0.1074	0.1307	0.0583	0.1150	0.1278	0.2087	0.0655	0.1363
ThP-Fr-123-5	0.1873	0.1721	0.1065	0.2105	0.2122	0.2744	0.1229	0.2251
ThP-Fr-13-5	0.1816	0.2185	0.1052	0.2091	0.2386	0.3925	0.1294	0.2567
ThP-Fr-23-5	0.1873	0.1721	0.1065	0.2105	0.2108	0.2728	0.1220	0.2230
ThP-Fr-3-5	0.1816	0.2185	0.1052	0.2091	0.2404	0.3972	0.1300	0.2587
AP-10	0.1023	0.1516	0.0536	0.1129	0.0744	0.1549	0.0332	0.0836
TNG-10	0.2404	0.3444	0.1343	0.2722	0.2777	0.5239	0.1396	0.3261
ThP-123-10	0.1582	0.2707	0.0777	0.1863	0.1655	0.3764	0.0696	0.1964
ThP-13-10	0.1576	0.2625	0.0783	0.1841	0.1646	0.3560	0.0721	0.1886
ThP-23-10	0.1582	0.2707	0.0777	0.1863	0.1655	0.3764	0.0696	0.1964
ThP-3-10	0.1576	0.2625	0.0783	0.1842	0.1647	0.3559	0.0721	0.1885
ThP-Fr-123-10	0.2009	0.2117	0.1142	0.2269	0.2340	0.3578	0.1279	0.2495
ThP-Fr-13-10	0.2147	0.3301	0.1153	0.2470	0.2593	0.5359	0.1215	0.2862
ThP-Fr-23-10	0.2009	0.2117	0.1142	0.2269	0.2304	0.3509	0.1261	0.2462
ThP-Fr-3-10	0.2147	0.3301	0.1153	0.2470	0.2606	0.5415	0.1216	0.2878
AP-20	0.1359	0.2740	0.0608	0.1571	0.0954	0.2725	0.0335	0.1142
TNG-20	0.2287	0.4514	0.1069	0.2856	0.2482	0.6298	0.0989	0.3294
ThP-123-20	0.1901	0.4169	0.0812	0.2458	0.1869	0.5358	0.0659	0.2465
ThP-13-20	0.1981	0.4249	0.0863	0.2600	0.1929	0.5279	0.0711	0.2479
ThP-23-20	0.1901	0.4169	0.0812	0.2458	0.1869	0.5358	0.0659	0.2465
ThP-3-20	0.1981	0.4249	0.0863	0.2600	0.1929	0.5279	0.0711	0.2478
ThP-Fr-123-20	0.2021	0.2167	0.1145	0.2284	0.2380	0.3912	0.1259	0.2580
ThP-Fr-13-20	0.2283	0.4369	0.1089	0.2775	0.2536	0.6289	0.1029	0.2950
ThP-Fr-23-20	0.2021	0.2167	0.1145	0.2284	0.2327	0.3726	0.1244	0.2515
ThP-Fr-3-20	0.2283	0.4369	0.1089	0.2775	0.2529	0.6336	0.1017	0.2941

Table J.4: Methods Evaluation : Word Granularity Metrics With Title as Gold Standard (Segment Count = 20)

METHOD CODE	DATASET = PUBMED_PMC_AB				DATASET = PATENT_RAND15K			
	wd-FMI	wd-COV	wd-JCI	wd-COS	wd-FMI	wd-COV	wd-JCI	wd-COS
AP-5	0.0780	0.0846	0.0443	0.0828	0.0583	0.0889	0.0312	0.0632
TNG-5	0.2376	0.2485	0.1435	0.2526	0.2892	0.4081	0.1729	0.3144
ThP-123-5	0.1060	0.1342	0.0571	0.1151	0.1265	0.2177	0.0630	0.1382
ThP-13-5	0.1067	0.1299	0.0580	0.1145	0.1240	0.2015	0.0636	0.1315
ThP-23-5	0.1060	0.1342	0.0571	0.1151	0.1265	0.2177	0.0630	0.1382
ThP-3-5	0.1066	0.1298	0.0579	0.1144	0.1243	0.2018	0.0638	0.1316
ThP-Fr-123-5	0.1863	0.1713	0.1059	0.2093	0.2122	0.2752	0.1229	0.2246
ThP-Fr-13-5	0.1811	0.2180	0.1050	0.2084	0.2401	0.3959	0.1300	0.2584
ThP-Fr-23-5	0.1863	0.1713	0.1059	0.2093	0.2100	0.2727	0.1215	0.2219
ThP-Fr-3-5	0.1811	0.2180	0.1050	0.2084	0.2421	0.4006	0.1309	0.2607
AP-10	0.1020	0.1515	0.0533	0.1120	0.0748	0.1565	0.0333	0.0840
TNG-10	0.2409	0.3453	0.1346	0.2727	0.2790	0.5252	0.1406	0.3271
ThP-123-10	0.1572	0.2699	0.0770	0.1843	0.1637	0.3725	0.0688	0.1937
ThP-13-10	0.1564	0.2605	0.0777	0.1824	0.1629	0.3519	0.0713	0.1852
ThP-23-10	0.1572	0.2699	0.0770	0.1843	0.1637	0.3725	0.0688	0.1937
ThP-3-10	0.1566	0.2606	0.0778	0.1825	0.1630	0.3523	0.0714	0.1854
ThP-Fr-123-10	0.2001	0.2112	0.1136	0.2262	0.2335	0.3590	0.1274	0.2493
ThP-Fr-13-10	0.2139	0.3286	0.1149	0.2461	0.2604	0.5393	0.1218	0.2873
ThP-Fr-23-10	0.2001	0.2112	0.1136	0.2262	0.2293	0.3512	0.1251	0.2450
ThP-Fr-3-10	0.2139	0.3286	0.1149	0.2461	0.2614	0.5441	0.1218	0.2886
AP-20	0.1349	0.2717	0.0604	0.1555	0.0952	0.2715	0.0334	0.1142
TNG-20	0.2302	0.4540	0.1077	0.2870	0.2492	0.6311	0.0995	0.3301
ThP-123-20	0.1895	0.4166	0.0807	0.2446	0.1857	0.5332	0.0654	0.2446
ThP-13-20	0.1981	0.4251	0.0863	0.2597	0.1929	0.5272	0.0712	0.2461
ThP-23-20	0.1895	0.4166	0.0807	0.2446	0.1857	0.5332	0.0654	0.2446
ThP-3-20	0.1982	0.4251	0.0864	0.2597	0.1930	0.5274	0.0712	0.2461
ThP-Fr-123-20	0.2013	0.2164	0.1139	0.2277	0.2375	0.3929	0.1252	0.2577
ThP-Fr-13-20	0.2282	0.4370	0.1088	0.2764	0.2542	0.6322	0.1029	0.2959
ThP-Fr-23-20	0.2013	0.2164	0.1139	0.2277	0.2316	0.3735	0.1234	0.2508
ThP-Fr-3-20	0.2282	0.4370	0.1088	0.2764	0.2539	0.6377	0.1019	0.2947

Table J.5: Methods Evaluation : Word Granularity Metrics With Title as Gold Standard (Segment Count = 25)

Appendix K

ROUGE EVALUATION TABLES

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.3797	0.4752	0.3965	0.2384	0.4584	0.2875
AP-10	0.3357	0.3004	0.2902	0.2907	0.3454	0.2594
TNG-10	0.3768	0.4030	0.3634	0.3094	0.5749	0.3597
ThP-123-10	0.3723	0.2706	0.2868	0.2789	0.3957	0.2785
ThP-13-10	0.3678	0.2708	0.2833	0.2842	0.4250	0.2888
ThP-23-10	0.3723	0.2706	0.2868	0.2796	0.3957	0.2786
ThP-3-10	0.3677	0.2708	0.2831	0.2841	0.4251	0.2885
ThP-Fr-123-10	0.3708	0.3790	0.3482	0.3047	0.5288	0.3405
ThP-Fr-13-10	0.3771	0.4394	0.3789	0.3094	0.5803	0.3630
ThP-Fr-23-10	0.3707	0.3790	0.3482	0.3019	0.5114	0.3329
ThP-Fr-3-10	0.3771	0.4395	0.3789	0.3091	0.5799	0.3624
AP-20	0.3499	0.3802	0.3392	0.3021	0.4329	0.3024
TNG-20	0.3761	0.4297	0.3751	0.3113	0.5790	0.3638
ThP-123-20	0.3659	0.3804	0.3505	0.2961	0.5025	0.3257
ThP-13-20	0.3648	0.3780	0.3476	0.2962	0.5255	0.3324
ThP-23-20	0.3661	0.3804	0.3505	0.2962	0.5025	0.3257
ThP-3-20	0.3650	0.3781	0.3476	0.2960	0.5263	0.3325
ThP-Fr-123-20	0.3704	0.3804	0.3484	0.3060	0.5354	0.3437
ThP-Fr-13-20	0.3774	0.4608	0.3871	0.3119	0.5755	0.3646
ThP-Fr-23-20	0.3705	0.3804	0.3484	0.3022	0.5150	0.3345
ThP-Fr-3-20	0.3774	0.4608	0.3871	0.3124	0.5759	0.3655

Table K.1: Text Summarization Quality: ROUGE-1 Evaluation of Summaries Extracted After Thematic Phrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.1667	0.2091	0.1737	0.1373	0.2840	0.1691
AP-10	0.1242	0.1140	0.1085	0.1201	0.1856	0.1250
TNG-10	0.1577	0.1713	0.1529	0.1768	0.3603	0.2124
ThP-123-10	0.1324	0.0994	0.1035	0.1306	0.2240	0.1412
ThP-13-10	0.1302	0.0986	0.1014	0.1354	0.2434	0.1492
ThP-23-10	0.1324	0.0994	0.1035	0.1309	0.2239	0.1412
ThP-3-10	0.1301	0.0986	0.1013	0.1352	0.2433	0.1489
ThP-Fr-123-10	0.1429	0.1484	0.1347	0.1634	0.3174	0.1898
ThP-Fr-13-10	0.1625	0.1920	0.1642	0.1788	0.3627	0.2152
ThP-Fr-23-10	0.1429	0.1484	0.1347	0.1586	0.3037	0.1825
ThP-Fr-3-10	0.1625	0.1920	0.1642	0.1783	0.3621	0.2144
AP-20	0.1379	0.1526	0.1345	0.1427	0.2470	0.1575
TNG-20	0.1593	0.1848	0.1598	0.1794	0.3622	0.2153
ThP-123-20	0.1434	0.1515	0.1380	0.1561	0.3031	0.1809
ThP-13-20	0.1420	0.1497	0.1360	0.1588	0.3196	0.1868
ThP-23-20	0.1435	0.1515	0.1380	0.1562	0.3031	0.1809
ThP-3-20	0.1421	0.1497	0.1360	0.1586	0.3202	0.1869
ThP-Fr-123-20	0.1429	0.1490	0.1349	0.1650	0.3224	0.1925
ThP-Fr-13-20	0.1644	0.2033	0.1694	0.1809	0.3597	0.2166
ThP-Fr-23-20	0.1429	0.1490	0.1349	0.1593	0.3063	0.1839
ThP-Fr-3-20	0.1644	0.2033	0.1694	0.1814	0.3600	0.2171

Table K.2: Text Summarization Quality: ROUGE-2 Evaluation of Summaries Extracted After Thematic Phrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.0840	0.1053	0.0873	0.0880	0.1942	0.1106
AP-10	0.0556	0.0518	0.0489	0.0657	0.1166	0.0722
TNG-10	0.0771	0.0842	0.0747	0.1144	0.2498	0.1406
ThP-123-10	0.0577	0.0445	0.0457	0.0744	0.1458	0.0847
ThP-13-10	0.0566	0.0437	0.0445	0.0778	0.1588	0.0899
ThP-23-10	0.0576	0.0445	0.0457	0.0745	0.1456	0.0845
ThP-3-10	0.0565	0.0437	0.0444	0.0778	0.1588	0.0898
ThP-Fr-123-10	0.0650	0.0676	0.0612	0.1004	0.2132	0.1203
ThP-Fr-13-10	0.0813	0.0965	0.0822	0.1164	0.2503	0.1428
ThP-Fr-23-10	0.0651	0.0676	0.0612	0.0959	0.2025	0.1142
ThP-Fr-3-10	0.0813	0.0965	0.0822	0.1158	0.2497	0.1419
AP-20	0.0647	0.0721	0.0631	0.0832	0.1609	0.0956
TNG-20	0.0785	0.0915	0.0787	0.1165	0.2503	0.1427
ThP-123-20	0.0669	0.0712	0.0645	0.0959	0.2048	0.1151
ThP-13-20	0.0657	0.0698	0.0631	0.0983	0.2166	0.1194
ThP-23-20	0.0671	0.0712	0.0645	0.0959	0.2048	0.1151
ThP-3-20	0.0658	0.0699	0.0631	0.0983	0.2172	0.1195
ThP-Fr-123-20	0.0651	0.0679	0.0613	0.1019	0.2170	0.1224
ThP-Fr-13-20	0.0827	0.1027	0.0853	0.1183	0.2487	0.1441
ThP-Fr-23-20	0.0651	0.0679	0.0613	0.0965	0.2043	0.1152
ThP-Fr-3-20	0.0828	0.1027	0.0853	0.1188	0.2492	0.1446

Table K.3: Text Summarization Quality: ROUGE-3 Evaluation of Summaries Extracted After Thematic Phrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.3030	0.3181	0.2975	0.2660	0.3077	0.2307
AP-10	0.2525	0.2213	0.2216	0.2475	0.2600	0.2198
TNG-10	0.2988	0.2724	0.2721	0.2531	0.3652	0.2832
ThP-123-10	0.2698	0.2025	0.2173	0.2352	0.2800	0.2336
ThP-13-10	0.2695	0.2020	0.2160	0.2385	0.2920	0.2399
ThP-23-10	0.2698	0.2025	0.2173	0.2359	0.2802	0.2338
ThP-3-10	0.2694	0.2020	0.2159	0.2387	0.2922	0.2400
ThP-Fr-123-10	0.2763	0.2582	0.2535	0.2450	0.3443	0.2697
ThP-Fr-13-10	0.3036	0.2921	0.2848	0.2543	0.3690	0.2849
ThP-Fr-23-10	0.2762	0.2582	0.2535	0.2425	0.3355	0.2647
ThP-Fr-3-10	0.3036	0.2921	0.2848	0.2539	0.3688	0.2844
AP-20	0.2691	0.2687	0.2560	0.2479	0.3039	0.2477
TNG-20	0.2997	0.2901	0.2822	0.2541	0.3701	0.2854
ThP-123-20	0.2764	0.2644	0.2581	0.2429	0.3298	0.2632
ThP-13-20	0.2764	0.2639	0.2576	0.2427	0.3380	0.2663
ThP-23-20	0.2766	0.2644	0.2581	0.2430	0.3298	0.2632
ThP-3-20	0.2765	0.2638	0.2576	0.2426	0.3383	0.2664
ThP-Fr-123-20	0.2763	0.2592	0.2540	0.2454	0.3481	0.2713
ThP-Fr-13-20	0.3036	0.3067	0.2920	0.2566	0.3718	0.2855
ThP-Fr-23-20	0.2763	0.2591	0.2540	0.2422	0.3375	0.2653
ThP-Fr-3-20	0.3037	0.3067	0.2920	0.2565	0.3721	0.2856

Table K.4: Text Summarization Quality: ROUGE-L Evaluation of Summaries Extracted After Thematic Phrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.1957	0.2504	0.2047	0.1457	0.3133	0.1809
AP-10	0.1579	0.1432	0.1356	0.1663	0.2116	0.1421
TNG-10	0.1909	0.2080	0.1839	0.1888	0.3962	0.2266
ThP-123-10	0.1739	0.1275	0.1328	0.1536	0.2529	0.1582
ThP-13-10	0.1712	0.1271	0.1306	0.1596	0.2743	0.1659
ThP-23-10	0.1738	0.1275	0.1328	0.1540	0.2527	0.1581
ThP-3-10	0.1712	0.1271	0.1305	0.1595	0.2744	0.1657
ThP-Fr-123-10	0.1807	0.1872	0.1688	0.1786	0.3547	0.2069
ThP-Fr-13-10	0.1938	0.2308	0.1949	0.1899	0.3991	0.2294
ThP-Fr-23-10	0.1808	0.1872	0.1688	0.1747	0.3403	0.1999
ThP-Fr-3-10	0.1938	0.2309	0.1949	0.1894	0.3985	0.2287
AP-20	0.1701	0.1884	0.1645	0.1738	0.2772	0.1742
TNG-20	0.1912	0.2231	0.1908	0.1905	0.3983	0.2296
ThP-123-20	0.1785	0.1883	0.1704	0.1720	0.3369	0.1967
ThP-13-20	0.1773	0.1865	0.1683	0.1739	0.3550	0.2024
ThP-23-20	0.1786	0.1883	0.1704	0.1720	0.3369	0.1967
ThP-3-20	0.1774	0.1866	0.1683	0.1738	0.3557	0.2025
ThP-Fr-123-20	0.1806	0.1880	0.1689	0.1801	0.3598	0.2094
ThP-Fr-13-20	0.1947	0.2436	0.2001	0.1919	0.3953	0.2308
ThP-Fr-23-20	0.1807	0.1881	0.1690	0.1753	0.3428	0.2011
ThP-Fr-3-20	0.1948	0.2436	0.2001	0.1924	0.3956	0.2314

Table K.5: Text Summarization Quality: ROUGE-SU4 Evaluation of Summaries Extracted After Thematic Phrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.3797	0.4752	0.3965	0.2384	0.4584	0.2875
AP-10	0.3521	0.4179	0.3555	0.2980	0.4946	0.3285
TNG-10	0.3756	0.4600	0.3851	0.3101	0.5680	0.3622
ThP-123-10	0.3713	0.4462	0.3776	0.3082	0.5633	0.3571
ThP-13-10	0.3703	0.4444	0.3758	0.3078	0.5632	0.3565
ThP-23-10	0.3714	0.4462	0.3776	0.3082	0.5633	0.3571
ThP-3-10	0.3703	0.4444	0.3759	0.3079	0.5633	0.3567
ThP-Fr-123-10	0.3751	0.4628	0.3857	0.3077	0.5633	0.3593
ThP-Fr-13-10	0.3750	0.4629	0.3858	0.3103	0.5666	0.3618
ThP-Fr-23-10	0.3751	0.4628	0.3857	0.3071	0.5626	0.3585
ThP-Fr-3-10	0.3751	0.4629	0.3858	0.3093	0.5674	0.3612
AP-20	0.3680	0.4553	0.3788	0.3057	0.5462	0.3500
TNG-20	0.3755	0.4676	0.3879	0.3080	0.5666	0.3611
ThP-123-20	0.3751	0.4630	0.3859	0.3114	0.5684	0.3625
ThP-13-20	0.3744	0.4619	0.3853	0.3131	0.5692	0.3631
ThP-23-20	0.3751	0.4630	0.3859	0.3114	0.5684	0.3625
ThP-3-20	0.3745	0.4619	0.3854	0.3131	0.5692	0.3633
ThP-Fr-123-20	0.3750	0.4631	0.3857	0.3062	0.5616	0.3583
ThP-Fr-13-20	0.3763	0.4714	0.3895	0.3077	0.5651	0.3609
ThP-Fr-23-20	0.3750	0.4631	0.3857	0.3064	0.5621	0.3582
ThP-Fr-3-20	0.3763	0.4714	0.3895	0.3071	0.5647	0.3606

Table K.6: Text Summarization Quality: ROUGE-1 Evaluation of Summaries Extracted After Thematic Subphrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.1667	0.2091	0.1737	0.1373	0.2840	0.1691
AP-10	0.1455	0.1750	0.1475	0.1568	0.2934	0.1804
TNG-10	0.1631	0.2019	0.1678	0.1791	0.3547	0.2144
ThP-123-10	0.1574	0.1913	0.1605	0.1746	0.3484	0.2082
ThP-13-10	0.1566	0.1905	0.1596	0.1736	0.3479	0.2073
ThP-23-10	0.1573	0.1913	0.1605	0.1746	0.3484	0.2082
ThP-3-10	0.1566	0.1905	0.1596	0.1738	0.3482	0.2076
ThP-Fr-123-10	0.1627	0.2025	0.1676	0.1768	0.3505	0.2118
ThP-Fr-13-10	0.1632	0.2037	0.1685	0.1792	0.3527	0.2138
ThP-Fr-23-10	0.1627	0.2025	0.1676	0.1766	0.3503	0.2114
ThP-Fr-3-10	0.1633	0.2037	0.1685	0.1784	0.3535	0.2134
AP-20	0.1574	0.1969	0.1625	0.1717	0.3360	0.2024
TNG-20	0.1633	0.2055	0.1692	0.1789	0.3548	0.2148
ThP-123-20	0.1621	0.2018	0.1670	0.1781	0.3526	0.2128
ThP-13-20	0.1615	0.2013	0.1666	0.1795	0.3538	0.2138
ThP-23-20	0.1620	0.2018	0.1670	0.1781	0.3526	0.2128
ThP-3-20	0.1616	0.2012	0.1666	0.1796	0.3536	0.2139
ThP-Fr-123-20	0.1627	0.2026	0.1676	0.1763	0.3500	0.2117
ThP-Fr-13-20	0.1643	0.2080	0.1706	0.1786	0.3537	0.2145
ThP-Fr-23-20	0.1627	0.2026	0.1676	0.1764	0.3504	0.2116
ThP-Fr-3-20	0.1643	0.2080	0.1706	0.1783	0.3537	0.2144

Table K.7: Text Summarization Quality: ROUGE-2 Evaluation of Summaries Extracted After Thematic Subphrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.0840	0.1053	0.0873	0.0880	0.1942	0.1106
AP-10	0.0710	0.0855	0.0718	0.0953	0.1958	0.1135
TNG-10	0.0818	0.1016	0.0842	0.1167	0.2454	0.1424
ThP-123-10	0.0777	0.0947	0.0792	0.1125	0.2393	0.1369
ThP-13-10	0.0774	0.0943	0.0787	0.1112	0.2385	0.1359
ThP-23-10	0.0778	0.0947	0.0792	0.1125	0.2393	0.1369
ThP-3-10	0.0773	0.0943	0.0787	0.1114	0.2389	0.1362
ThP-Fr-123-10	0.0818	0.1016	0.0840	0.1148	0.2419	0.1403
ThP-Fr-13-10	0.0821	0.1028	0.0848	0.1167	0.2431	0.1417
ThP-Fr-23-10	0.0818	0.1016	0.0840	0.1148	0.2419	0.1401
ThP-Fr-3-10	0.0822	0.1028	0.0848	0.1162	0.2441	0.1416
AP-20	0.0782	0.0980	0.0806	0.1097	0.2300	0.1322
TNG-20	0.0819	0.1035	0.0849	0.1174	0.2463	0.1435
ThP-123-20	0.0810	0.1010	0.0834	0.1156	0.2431	0.1408
ThP-13-20	0.0808	0.1008	0.0833	0.1168	0.2440	0.1417
ThP-23-20	0.0810	0.1010	0.0834	0.1156	0.2431	0.1408
ThP-3-20	0.0808	0.1008	0.0833	0.1168	0.2438	0.1417
ThP-Fr-123-20	0.0817	0.1016	0.0839	0.1148	0.2420	0.1406
ThP-Fr-13-20	0.0828	0.1051	0.0859	0.1170	0.2452	0.1431
ThP-Fr-23-20	0.0817	0.1016	0.0839	0.1148	0.2423	0.1404
ThP-Fr-3-20	0.0828	0.1051	0.0859	0.1169	0.2455	0.1432

Table K.8: Text Summarization Quality: ROUGE-3 Evaluation of Summaries Extracted After Thematic Subphrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.3030	0.3181	0.2975	0.2660	0.3077	0.2307
AP-10	0.2755	0.2862	0.2678	0.2417	0.3336	0.2629
TNG-10	0.3011	0.3070	0.2905	0.2562	0.3678	0.2825
ThP-123-10	0.2937	0.2988	0.2831	0.2496	0.3633	0.2791
ThP-13-10	0.2930	0.2981	0.2823	0.2508	0.3641	0.2804
ThP-23-10	0.2938	0.2988	0.2831	0.2497	0.3633	0.2791
ThP-3-10	0.2930	0.2981	0.2823	0.2509	0.3642	0.2805
ThP-Fr-123-10	0.3002	0.3089	0.2909	0.2566	0.3666	0.2808
ThP-Fr-13-10	0.3013	0.3084	0.2913	0.2575	0.3678	0.2828
ThP-Fr-23-10	0.3002	0.3089	0.2909	0.2564	0.3663	0.2806
ThP-Fr-3-10	0.3014	0.3084	0.2913	0.2568	0.3676	0.2821
AP-20	0.2910	0.3084	0.2860	0.2470	0.3573	0.2750
TNG-20	0.3007	0.3134	0.2933	0.2580	0.3685	0.2819
ThP-123-20	0.2991	0.3103	0.2909	0.2537	0.3698	0.2829
ThP-13-20	0.2991	0.3097	0.2908	0.2548	0.3697	0.2837
ThP-23-20	0.2990	0.3103	0.2909	0.2537	0.3698	0.2829
ThP-3-20	0.2991	0.3097	0.2908	0.2547	0.3696	0.2836
ThP-Fr-123-20	0.3001	0.3091	0.2909	0.2571	0.3663	0.2798
ThP-Fr-13-20	0.3019	0.3151	0.2947	0.2584	0.3681	0.2816
ThP-Fr-23-20	0.3001	0.3091	0.2909	0.2570	0.3663	0.2802
ThP-Fr-3-20	0.3019	0.3151	0.2947	0.2583	0.3680	0.2813

Table K.9: Text Summarization Quality: ROUGE-L Evaluation of Summaries Extracted After Thematic Subphrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.1957	0.2504	0.2047	0.1457	0.3133	0.1809
AP-10	0.1756	0.2130	0.1772	0.1716	0.3266	0.1964
TNG-10	0.1933	0.2423	0.1984	0.1904	0.3900	0.2289
ThP-123-10	0.1888	0.2320	0.1921	0.1867	0.3843	0.2231
ThP-13-10	0.1880	0.2309	0.1909	0.1860	0.3841	0.2225
ThP-23-10	0.1888	0.2320	0.1921	0.1867	0.3843	0.2231
ThP-3-10	0.1880	0.2309	0.1910	0.1861	0.3844	0.2227
ThP-Fr-123-10	0.1931	0.2436	0.1986	0.1882	0.3855	0.2263
ThP-Fr-13-10	0.1933	0.2444	0.1992	0.1903	0.3878	0.2282
ThP-Fr-23-10	0.1931	0.2436	0.1986	0.1880	0.3853	0.2259
ThP-Fr-3-10	0.1934	0.2444	0.1992	0.1896	0.3888	0.2278
AP-20	0.1871	0.2371	0.1927	0.1837	0.3708	0.2171
TNG-20	0.1932	0.2464	0.1998	0.1898	0.3896	0.2290
ThP-123-20	0.1925	0.2430	0.1981	0.1899	0.3886	0.2277
ThP-13-20	0.1920	0.2424	0.1977	0.1915	0.3897	0.2287
ThP-23-20	0.1925	0.2430	0.1981	0.1899	0.3886	0.2277
ThP-3-20	0.1921	0.2424	0.1978	0.1915	0.3895	0.2287
ThP-Fr-123-20	0.1930	0.2437	0.1986	0.1876	0.3848	0.2261
ThP-Fr-13-20	0.1943	0.2492	0.2013	0.1896	0.3883	0.2287
ThP-Fr-23-20	0.1929	0.2437	0.1986	0.1878	0.3853	0.2260
ThP-Fr-3-20	0.1943	0.2492	0.2013	0.1893	0.3884	0.2287

Table K.10: Text Summarization Quality: ROUGE-SU4 Evaluation of Summaries Extracted After Thematic Subphrase Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.3797	0.4752	0.3965	0.2384	0.4584	0.2875
AP-10	0.3603	0.4410	0.3692	0.3049	0.5337	0.3464
TNG-10	0.3751	0.4688	0.3881	0.3065	0.5646	0.3600
ThP-123-10	0.3747	0.4690	0.3878	0.3090	0.5669	0.3623
ThP-13-10	0.3741	0.4680	0.3871	0.3104	0.5684	0.3631
ThP-23-10	0.3747	0.4690	0.3878	0.3090	0.5669	0.3623
ThP-3-10	0.3740	0.4679	0.3870	0.3106	0.5689	0.3635
ThP-Fr-123-10	0.3750	0.4657	0.3867	0.3070	0.5624	0.3594
ThP-Fr-13-10	0.3749	0.4708	0.3885	0.3053	0.5631	0.3589
ThP-Fr-23-10	0.3750	0.4657	0.3867	0.3056	0.5615	0.3579
ThP-Fr-3-10	0.3749	0.4708	0.3885	0.3054	0.5633	0.3590
AP-20	0.3721	0.4664	0.3853	0.3094	0.5613	0.3590
TNG-20	0.3753	0.4728	0.3896	0.3031	0.5615	0.3568
ThP-123-20	0.3758	0.4734	0.3901	0.3047	0.5641	0.3590
ThP-13-20	0.3756	0.4732	0.3901	0.3069	0.5644	0.3605
ThP-23-20	0.3758	0.4734	0.3901	0.3047	0.5641	0.3590
ThP-3-20	0.3756	0.4732	0.3900	0.3067	0.5640	0.3603
ThP-Fr-123-20	0.3749	0.4658	0.3867	0.3048	0.5603	0.3572
ThP-Fr-13-20	0.3756	0.4740	0.3904	0.3026	0.5609	0.3564
ThP-Fr-23-20	0.3749	0.4658	0.3867	0.3045	0.5599	0.3567
ThP-Fr-3-20	0.3756	0.4740	0.3904	0.3030	0.5610	0.3569

Table K.11: Text Summarization Quality: ROUGE-1 Evaluation of Summaries Extracted After Thematic Word Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.1667	0.2091	0.1737	0.1373	0.2840	0.1691
AP-10	0.1521	0.1883	0.1563	0.1678	0.3240	0.1971
TNG-10	0.1632	0.2060	0.1693	0.1775	0.3531	0.2136
ThP-123-10	0.1627	0.2056	0.1687	0.1783	0.3531	0.2142
ThP-13-10	0.1621	0.2049	0.1682	0.1790	0.3539	0.2145
ThP-23-10	0.1627	0.2056	0.1687	0.1783	0.3531	0.2142
ThP-3-10	0.1621	0.2050	0.1682	0.1792	0.3543	0.2148
ThP-Fr-123-10	0.1629	0.2039	0.1682	0.1773	0.3505	0.2126
ThP-Fr-13-10	0.1631	0.2070	0.1695	0.1766	0.3516	0.2126
ThP-Fr-23-10	0.1629	0.2039	0.1682	0.1761	0.3499	0.2114
ThP-Fr-3-10	0.1631	0.2070	0.1695	0.1764	0.3518	0.2126
AP-20	0.1609	0.2035	0.1669	0.1768	0.3479	0.2105
TNG-20	0.1633	0.2078	0.1700	0.1760	0.3515	0.2121
ThP-123-20	0.1635	0.2079	0.1701	0.1768	0.3528	0.2133
ThP-13-20	0.1634	0.2079	0.1701	0.1777	0.3523	0.2138
ThP-23-20	0.1635	0.2079	0.1701	0.1768	0.3528	0.2133
ThP-3-20	0.1635	0.2078	0.1701	0.1775	0.3521	0.2136
ThP-Fr-123-20	0.1629	0.2040	0.1682	0.1759	0.3493	0.2113
ThP-Fr-13-20	0.1636	0.2085	0.1705	0.1756	0.3510	0.2118
ThP-Fr-23-20	0.1628	0.2040	0.1682	0.1755	0.3490	0.2108
ThP-Fr-3-20	0.1636	0.2085	0.1705	0.1760	0.3511	0.2122

Table K.12: Text Summarization Quality: ROUGE-2 Evaluation of Summaries Extracted After Thematic Word Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.0840	0.1053	0.0873	0.0880	0.1942	0.1106
AP-10	0.0751	0.0932	0.0771	0.1060	0.2201	0.1277
TNG-10	0.0821	0.1038	0.0852	0.1160	0.2449	0.1423
ThP-123-10	0.0818	0.1034	0.0846	0.1167	0.2443	0.1426
ThP-13-10	0.0814	0.1030	0.0844	0.1169	0.2445	0.1425
ThP-23-10	0.0818	0.1034	0.0846	0.1167	0.2443	0.1426
ThP-3-10	0.0814	0.1030	0.0844	0.1171	0.2448	0.1428
ThP-Fr-123-10	0.0820	0.1024	0.0843	0.1158	0.2426	0.1414
ThP-Fr-13-10	0.0820	0.1043	0.0852	0.1154	0.2435	0.1415
ThP-Fr-23-10	0.0820	0.1024	0.0843	0.1146	0.2419	0.1402
ThP-Fr-3-10	0.0820	0.1043	0.0852	0.1151	0.2435	0.1413
AP-20	0.0807	0.1021	0.0836	0.1147	0.2397	0.1392
TNG-20	0.0823	0.1049	0.0856	0.1157	0.2442	0.1418
ThP-123-20	0.0823	0.1046	0.0855	0.1162	0.2450	0.1426
ThP-13-20	0.0823	0.1047	0.0855	0.1164	0.2441	0.1425
ThP-23-20	0.0823	0.1046	0.0855	0.1162	0.2449	0.1426
ThP-3-20	0.0823	0.1047	0.0855	0.1163	0.2439	0.1424
ThP-Fr-123-20	0.0819	0.1024	0.0843	0.1148	0.2417	0.1405
ThP-Fr-13-20	0.0823	0.1051	0.0857	0.1152	0.2433	0.1413
ThP-Fr-23-20	0.0819	0.1024	0.0843	0.1143	0.2413	0.1399
ThP-Fr-3-20	0.0823	0.1051	0.0857	0.1154	0.2434	0.1416

Table K.13: Text Summarization Quality: ROUGE-3 Evaluation of Summaries Extracted After Thematic Word Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.3030	0.3181	0.2975	0.2660	0.3077	0.2307
AP-10	0.2846	0.3012	0.2793	0.2459	0.3520	0.2729
TNG-10	0.3002	0.3141	0.2934	0.2577	0.3672	0.2805
ThP-123-10	0.2996	0.3149	0.2934	0.2559	0.3689	0.2818
ThP-13-10	0.2994	0.3145	0.2932	0.2559	0.3698	0.2828
ThP-23-10	0.2996	0.3149	0.2934	0.2559	0.3689	0.2818
ThP-3-10	0.2994	0.3144	0.2931	0.2559	0.3699	0.2830
ThP-Fr-123-10	0.3002	0.3115	0.2920	0.2569	0.3664	0.2800
ThP-Fr-13-10	0.3002	0.3155	0.2940	0.2581	0.3674	0.2801
ThP-Fr-23-10	0.3002	0.3114	0.2920	0.2569	0.3663	0.2800
ThP-Fr-3-10	0.3001	0.3155	0.2940	0.2582	0.3679	0.2803
AP-20	0.2959	0.3156	0.2918	0.2522	0.3668	0.2804
TNG-20	0.2996	0.3178	0.2948	0.2593	0.3670	0.2794
ThP-123-20	0.3000	0.3181	0.2951	0.2577	0.3678	0.2798
ThP-13-20	0.2999	0.3181	0.2952	0.2578	0.3686	0.2810
ThP-23-20	0.3000	0.3181	0.2951	0.2577	0.3678	0.2798
ThP-3-20	0.3000	0.3181	0.2953	0.2576	0.3683	0.2807
ThP-Fr-123-20	0.3002	0.3116	0.2920	0.2575	0.3658	0.2791
ThP-Fr-13-20	0.3000	0.3185	0.2953	0.2585	0.3667	0.2785
ThP-Fr-23-20	0.3001	0.3115	0.2920	0.2575	0.3656	0.2793
ThP-Fr-3-20	0.3000	0.3185	0.2953	0.2586	0.3668	0.2787

Table K.14: Text Summarization Quality: ROUGE-L Evaluation of Summaries Extracted After Thematic Word Based Sentence Pre-Filtration (Segment Count = 25)

METHOD CODE	DATASET = PUBMED_PMC_AB			DATASET = PATENT_RAND15K		
	PRECISION	RECALL	F-SCORE	PRECISION	RECALL	F-SCORE
TextRank	0.1957	0.2504	0.2047	0.1457	0.3133	0.1809
AP-10	0.1817	0.2276	0.1862	0.1808	0.3586	0.2125
TNG-10	0.1932	0.2472	0.2001	0.1884	0.3878	0.2279
ThP-123-10	0.1926	0.2469	0.1995	0.1897	0.3883	0.2288
ThP-13-10	0.1922	0.2462	0.1991	0.1903	0.3890	0.2290
ThP-23-10	0.1926	0.2469	0.1995	0.1897	0.3883	0.2288
ThP-3-10	0.1922	0.2462	0.1990	0.1904	0.3894	0.2292
ThP-Fr-123-10	0.1931	0.2451	0.1991	0.1884	0.3854	0.2269
ThP-Fr-13-10	0.1931	0.2483	0.2003	0.1876	0.3864	0.2269
ThP-Fr-23-10	0.1931	0.2451	0.1991	0.1871	0.3847	0.2257
ThP-Fr-3-10	0.1930	0.2483	0.2003	0.1876	0.3865	0.2270
AP-20	0.1906	0.2444	0.1974	0.1885	0.3832	0.2252
TNG-20	0.1932	0.2492	0.2008	0.1870	0.3863	0.2265
ThP-123-20	0.1934	0.2494	0.2010	0.1879	0.3876	0.2276
ThP-13-20	0.1934	0.2494	0.2010	0.1889	0.3872	0.2282
ThP-23-20	0.1934	0.2494	0.2010	0.1879	0.3876	0.2276
ThP-3-20	0.1934	0.2494	0.2010	0.1888	0.3869	0.2281
ThP-Fr-123-20	0.1930	0.2452	0.1991	0.1870	0.3840	0.2256
ThP-Fr-13-20	0.1935	0.2500	0.2014	0.1865	0.3855	0.2259
ThP-Fr-23-20	0.1930	0.2452	0.1991	0.1865	0.3837	0.2250
ThP-Fr-3-20	0.1935	0.2500	0.2014	0.1869	0.3856	0.2263

Table K.15: Text Summarization Quality: ROUGE-SU4 Evaluation of Summaries Extracted After Thematic Word Based Sentence Pre-Filtration (Segment Count = 25)

Bibliography

[1] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus, "Knowledge Discovery in Databases : An Overview," *AI Magazine*, vol. 13, no. 3, pp. 57–70, 1992.

[2] Y. Kodratoff, "Knowledge discovery in texts: A definition, and applications," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1609, pp. 16–29, 1999.

[3] R. Davis, H. Shrobe, and P. Szolovits, "What Is a Knowledge Representation?," *AI Magazine*, vol. 14, no. 1, p. 17, 1993.

[4] J. F. Sowa, *Semantic Networks*. American Cancer Society, 2006.

[5] J. F. Sowa, *Principles of Semantic Networks : Explorations in the Representation of Knowledge.* Burlington : Elsevier Science, 2014., 2014.

[6] S. Staab and R. Studer, *Handbook on Ontologies*. Springer Publishing Company, Incorporated, 2nd ed., 2009.

[7] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," *Inter-*

national Conference on Knowledge Discovery and Data Mining (KDD), pp. 112–117, 1995.

[8] A. J. Cañas, R. Carff, G. Hill, M. Carvalho, M. Arguedas, T. C. Eskridge, J. Lott, and R. Carvajal, *Concept Maps: Integrating Knowledge and Information Visualization*, pp. 205–219. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[9] J. Novak and A. Caas, "The theory underlying concept maps and how to construct them," Tech. Rep. MSU-CSE-06-2, Institute for Human and Machine Cognition, Pensacola Florida, 2006.

[10] T. Buzan and B. Buzan, *The Mind Map Book*. London: BBC Books, 2 ed., 1995.

[11] V. Jelisavi, B. Furlan, J. Proti, and V. Milutinovi, "Topic models and advanced algorithms for profiling of knowledge in scientific papers," in *2012 Proceedings of the 35th International Convention MIPRO*, pp. 1030–1035, 2012.

[12] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, 2015.

[13] L. Xia, D. Luo, C. Zhang, and Z. Wu, "A survey of topic models in text classification," in *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 244–250, 2019.

[14] A. Nenkova and K. McKeown, "Automatic Summarization," *Foundations and Trends® in Information Retrieval*, vol. 5, no. 3, pp. 235–422, 2011.

[15] A. Nenkova and K. McKeown, *A Survey of Text Summarization Techniques*, pp. 43–76. Boston, MA: Springer US, 2012.

[16] M. J. A. Howe, "Using students' notes to examine the role of the individual learner in acquiring meaningful subject matter," *The Journal of Educational Research*, vol. 64, no. 2, pp. 61–63, 1970.

[17] G. H. Bower, M. C. Clark, A. M. Lesgold, and D. Winzenz, "Hierarchical retrieval schemes in recall of categorized word lists," *Journal of Verbal Learning and Verbal Behavior*, vol. 8, no. 3, pp. 323–343, 1969.

[18] A. L. Brown, J. C. Campione, and J. D. Day, "Learning to Learn: On Training Students to Learn from Texts," *Educational Researcher*, vol. 10, no. 2, 1981.

[19] R. Garner, "Efficient text summarization: Costs and benefits," 1982.

[20] A. L. Brown and J. D. Day, "Macrorules for summarizing texts: The development of expertise," *Journal of Verbal Learning and Verbal Behavior*, vol. 22, no. 1, pp. 1–14, 1983.

[21] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "Summac: A text summarization evaluation," *Nat. Lang. Eng.*, vol. 8, pp. 43–68, Mar. 2002.

[22] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, "Automated phrase mining from massive text corpora," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, pp. 1825–1837, Oct 2018.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[24] H. M. Wallach, "Topic Modeling: Beyond Bag-of-Words Hanna," *Icml2006*, no. 1, pp. 977–984, 2006.

[25] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation.," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[26] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum, "Integrating topics and syntax," *Advances in neural information processing systems*, vol. 17, pp. 537–544, 2005.

[27] A. Gruber, Y. Weiss, and M. Rosen-Zvi, "Hidden topic markov models.," *Journal of Machine Learning Research - Proceedings Track*, vol. 2, pp. 163–170, 01 2007.

[28] X. Wang, A. McCallum, and X. Wei, "Topical N-grams: Phrase and topic discovery, with an application to information retrieval," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 697–702, 2007.

[29] I. Sato and H. Nakagawa, "Topic models with power-law using pitman-yor process," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, (New York, NY, USA), pp. 673–682, ACM, 2010.

[30] H. Noji, "Improvements to the Bayesian Topic N-gram Models," *Emnlp2013*, no. October, pp. 1180–1190, 2013.

[31] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, p. 22102216, AAAI Press, 2015.

[32] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, p. 17271736, JMLR.org, 2016.

[33] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine, "Topic modelling meets deep neural networks: A survey," 2021.

[34] A. B. Dieng, C. Wang, J. Gao, and J. W. Paisley, "Topicrnn: A recurrent neural network with long-range semantic dependency," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[35] M. Panwar, S. Shailabh, M. Aggarwal, and B. Krishnamurthy, "TAN-NTM: topic attention networks for neural topic modeling," *CoRR*, vol. abs/2012.01524, 2020.

[36] R. Wang, D. Zhou, and Y. He, "Atm: Adversarial-neural topic model," *Information Processing & Management*, vol. 56, no. 6, p. 102098, 2019.

[37] R. Wang, X. Hu, D. Zhou, Y. He, Y. Xiong, C. Ye, and H. Xu, "Neural topic modeling with bidirectional adversarial training," in *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, (Online), pp. 340–350, Association for Computational Linguistics, July 2020.

[38] A. El-Kishky, Y. Song, C. Wang, C. Voss, and J. Han, "Scalable topical phrase mining from text corpora," *Proceedings of the VLDB Endowment*, vol. 8, 06 2014.

[39] B. Li, X. Yang, R. Zhou, B. Wang, C. Liu, and Y. Zhang, "An efficient method for high quality and cohesive topical phrase mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, pp. 120–137, Jan 2019.

[40] W.-t. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, (San Francisco, CA, USA), pp. 1776–1782, Morgan Kaufmann Publishers Inc., 2007.

[41] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," *Inf. Process. Manage.*, vol. 43, pp. 1606–1618, Nov. 2007.

[42] E. Hovy and C.-Y. Lin, "Automated text summarization and the summarist system," in *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, (Stroudsburg, PA, USA), pp. 197–214, Association for Computational Linguistics, 1998.

[43] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Int. Res.*, vol. 22, pp. 457–479, Dec. 2004.

[44] C.-Y. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, (Stroudsburg, PA, USA), pp. 495–501, Association for Computational Linguistics, 2000.

[45] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary, "Topic-focused multi-document summarization using an approximate oracle score," in *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, (Stroudsburg, PA, USA), pp. 152–159, Association for Computational Linguistics, 2006.

[46] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *Intelligent Scalable Text Summarization*, 1997.

[47] B. Schiffman, A. Nenkova, and K. McKeown, "Experiments in multidocument summarization," in *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, (San Francisco, CA, USA), pp. 52–58, Morgan Kaufmann Publishers Inc., 2002.

[48] H. G. Silber and K. F. McCoy, "Efficiently computed lexical chains as an intermediate representation for automatic text summarization," *Comput. Linguist.*, vol. 28, pp. 487–496, Dec. 2002.

[49] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, (New York, NY, USA), pp. 19–25, ACM, 2001.

[50] J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Jeek, "Two uses of anaphora resolution in summarization," *Information Processing & Management*, vol. 43, no. 6, pp. 1663 – 1680, 2007. Text Summarization.

[51] J. Steinberger and K. Ježek, "Update summarization based on novel topic distribution," in *Proceedings of the 9th ACM Symposium on Document Engineering*, DocEng '09, (New York, NY, USA), pp. 205–213, ACM, 2009.

[52] D. Marcu, "From discourse structures to text summaries," in *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 82–88, 1997.

[53] D. Marcu, *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press, 2000.

[54] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[55] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Conference of the North Ameri-*

can Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, (USA), p. 7178, Association for Computational Linguistics, 2003.

[56] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, (USA), p. 605es, Association for Computational Linguistics, 2004.

[57] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[58] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory — ICDT 2001* (J. Van den Bussche and V. Vianu, eds.), (Berlin, Heidelberg), pp. 420–434, Springer Berlin Heidelberg, 2001.

[59] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[60] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second Inter-*

national Conference on Knowledge Discovery and Data Mining, KDD'96, p. 226231, AAAI Press, 1996.

[61] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

[62] M. Hoffman, D. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *Nips*, pp. 1–9, 2010.

[63] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *ArXiv*, vol. abs/1602.03606, 2016.

[64] A. K. McCallum, "Mallet: A machine learning for language toolkit." http://mallet.cs.umass.edu, 2002.

[65] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 107–117, 1998. Proceedings of the Seventh International World Wide Web Conference.

[66] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *SIGIR '94* (B. W. Croft and C. J. van Rijsbergen, eds.), (London), pp. 232–241, Springer London, 1994.

[67] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[68] D. Kalpić, N. Hlupić, and M. Lovrić, *Student's t-Tests*, pp. 1559–1563. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[69] D. Rey and M. Neuhäuser, *Wilcoxon-Signed-Rank Test*, pp. 1658–1659. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[70] *Bonferroni Correction*, pp. 227–227. Dordrecht: Springer Netherlands, 2008.

[71] T. Morikawa and T. Yamanaka, *Multiple Comparison*, pp. 888–890. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.