www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 07, July 2025, pp : 412-414

e-ISSN : 2583-1062
Impact Factor : 7.001

AUTOMATED PHRASE MINING FROM MASSIVE TEXT

Vijayata Ramteke[*1], Ch. Madhumathi[*2], Ch. Devaji[*3], G. Gowtham Raj[*4]

[*1]Professor, Department Of Computer Science Engineering, Siddhartha Institute Of Technology & Sciences, Hyderabad, India.

[*2,3,4]Department Of Computer Science Engineering, Siddhartha Institute Of Technology & Sciences Hyderabad, India.

DOI: https://www.doi.org/10.58257/IJPREMS42803

ABSTRACT

As one of the fundamental tasks in text analysis, phrase mining aims at extracting quality phrases from a text corpus and has various downstream applications including information extraction/retrieval, taxonomy construction, and topic modeling. Most existing methods rely on complex, trained linguistic analyzers, and thus likely have unsatisfactory performance on text corpora of new domains and genres without extra but expensive adaption.

None of the stateofthe-art models, even data-driven models, bgbghgbgbbbis fully automated because they require human experts for designing rules or labeling phrases. In this paper, we propose a novel framework for automated phrase mining, AutoPhrase, which supports any language as long as a general knowledge base (e.g., Wikipedia) in that language is available, while benefiting from, but not requiring, a POS tagger. Compared to the state-of-the-art methods, AutoPhrase has shown significant improvements in both effectiveness and efficiency on five real-world datasets across different domains and languages. Besides, AutoPhrase can be extend to model single phase.

Keywords: Automatic Phrase Mining, Phrase Mining, Distant Training, Part-Of-Speech Tag, Multiple Language.

1. INTRODUCTION

With the exponential growth of digital content, organizations and researchers are increasingly faced with the challenge of extracting meaningful insights from massive text corpora. Traditional text mining approaches often rely on word-level analysis, which can overlook the rich semantics embedded in multi-word expressions or phrases. Phrases such as "climate change policy," "machine learning algorithm," or "stock market crash" carry far more specific meaning than their constituent words considered separately. This highlights the need for robust automated phrase mining techniques capable of discovering high-quality phrases from large-scale, unstructured text. Automated phrase mining refers to the unsupervised or weakly-supervised process of extracting salient, meaningful phrases from text without requiring human annotation or domain-specific knowledge. Unlike simple n-gram models that extract frequent word sequences based purely on co-occurrence, modern phrase mining algorithms evaluate phrase quality using a combination of linguistic features (e.g., part-of-speech patterns), statistical metrics (e.g., pointwise mutual information, frequency), and contextual informativeness. These techniques are designed to identify non-trivial, domain-relevant phrases that are both statistically significant and semantically coherent. Recent advances in this area include methods like AutoPhrase and SegPhrase, which integrate knowledge bases and distant supervision with efficient phrase segmentation models. Such tools have demonstrated scalability to datasets with billions of tokens, making them highly applicable to industrial and academic settings alike. Automated phrase mining has broad applications, including: Enhancing information retrieval and search engines through better keyword indexing. Supporting topic modeling and document clustering with higher-quality features. Building taxonomies and knowledge graphs by discovering entities and concepts. Enabling summarization and content recommendation systems.

2. LITERATURE SURVEY

Automated phrase mining from massive text corpora is a fundamental task in text analysis that involves extracting high-quality phrases from large amounts of text data. This task has various applications, including: Information Extraction/Retrieval: Phrase mining helps in extracting relevant information from text data, making it easier to retrieve and analyze. Taxonomy Construction: Automated phrase mining enables the construction of taxonomies by identifying key phrases and their relationships. Topic Modeling: Phrase mining is essential in topic modeling, where it helps identify underlying themes and patterns in text data. Approaches to Automated Phrase Mining: Traditional Methods: These methods rely on complex, trained linguistic analyzers, which can have unsatisfactory performance on new domains and genres without additional adaptation. Data-Driven to design rules or label phrases.Methods: Recent developments in data-driven methods have shown promise in extracting phrases from domain-specific text. However, these.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 07, July 2025, pp : 412-414

e-ISSN : 2583-1062
Impact Factor : 7.001

AUTO PHRASE

A Novel Framework AutoPhrase is a fully automated phrase mining framework that leverages a large amount of highquality phrases from public knowledge bases like Wikipedia. This framework: POS-Guided Phrasal Segmentation: Incorporates shallow syntactic information from partofspeech (POS) tags to enhance performance when a POS tagger is available. Significant Improvements: AutoPhrase has shown significant improvements in effectiveness and efficiency on real-world datasets across different domains and languages. Key Benefits: Fully Automated: AutoPhrase eliminates the need for human experts to design rules or label phrases. Flexibility: Supports various languages and domains. Improved Performance: Achieves better performance compared to state-of-the-art methods

3. METHODOLOGY

In this section, we focus on introducing our two new techniques. First, a novel robust positiveonly distant training method is developed to leverage the quality phrases in public, general knowledge bases. Second, we introduce the part-of-speech tags into the phrasal segmentation process and try to let our model take advantage of these language-dependent information, and thus perform more smoothly in different languages.

Robust Positive-Only Distant Training To estimate the phrase quality score for each phrase candidate, our previous work [23] required domain experts to first carefully select hundreds of varying-quality phrases from millions of candidates, and then annotate them with binary labels. For example, for computer science papers, our domain experts provided hundreds of positive labels (e.g., “spanning tree” and “computer science”) and negative labels (e.g., “paper focuses” and “important form of “). However, creating such a label set is expensive, especially in specialized domains like clinical reports and business reviews, because this approach provides no clues for how to identify the phrase candidates to be labeled. In this paper, we introduce a method that only utilizes existing general knowledge bases without any other human effort.

Label Pools—Public knowledge bases (e.g., Wikipedia) usually encode a considerable number of high-quality phrases in the titles, keywords, and internal links of pages. For example, by analyzing the internal links and synonyms in English Wikipedia, more than a hundred thousand high-quality phrases were discovered. As a result, we place these phrases in a positive pool. Knowledge bases, however, rarely, if ever, identify phrases that fail to meet our criteria, what we call inferior phrases. An important observation is that the number of phrase candidates, based on n-grams (recall leftmost box of Figure 1), is huge and the majority of them are actually of inferior quality (e.g., “Francisco opera and”). In practice, based on our experiments, among millions of phrase candidates, usually, only about 10% are in good quality . Therefore, phrase candidates that are derived from the given corpus but that fail to match any high-quality phrase derived from the given knowledge base, are used to populate a large but noisy negative pool.

Noise Reduction—Directly training a classifier based on the noisy label pools is not a wise choice: some phrases of high quality from the given corpus may have been missed (i.e., inaccurately binned into the negative pool) simply because they were not present in the knowledge base. Instead, we propose to utilize an ensemble classifier that averages the results of T independently trained base classifiers. As shown in Figure 2, for each base classifier, we randomly draw K phrase candidates with replacement from the positive pool and the negative pool respectively (considering a canonical balanced classification scenario). This size-2K subset of the full set of all phrase candidates is called a perturbed training set [5], because the labels of some (δ in the figure) quality phrases are switched from positive to negative. In order for the ensemble classifier to alleviate the effect of such noise, we need to use base classifiers with the lowest possible training errors. We grow an unpruned decision tree to the point of separating all phrases to meet this requirement. In fact, such decision tree will always reach 100% training accuracy when no two positive and negative phrases share identical feature representations in the perturbed training set. In this case, its ideal error is , , which approximately equals to the proportion of switched labels among all phrase candidates (i.e., $2\delta K \approx 10\%$.)

4. EXISTING SYSYTEM

Existing Systems for Automated Phrase Mining Several existing systems and techniques are used for automated phrase mining from massive text. Here are some notable ones:

1 AUTO PHRASE Description: AutoPhrase is a fully automated phrase mining framework that leverages high-quality phrases from public knowledge bases like Wikipedia. Features: Supports multiple languages, POS-guided phrasal segmentation, and significant improvements in effectiveness and efficiency.

PHRASE MINING TOOLS Description: Various phrase mining tools are available, including those based on statistical methods, machine learning algorithms, and hybrid approaches. Features: These tools often provide features like phrase extraction, ranking, and filtering.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 07, July 2025, pp : 412-414

e-ISSN : 2583-1062
Impact Factor : 7.001

3 NATURAL LANGUAGE PROCESSING (NLP) Libraries: Description: NLP libraries like NLTK, spaCy, and Stanford CoreNLP provide tools and resources for phrase mining and other NLP tasks. Features: These libraries often include features like tokenization, part-of-speech tagging, and named entity recognition.

INFORMATION EXTRACTION SYSTEM: Description: Information extraction systems like OpenIE and Never-Ending Language Learning (NELL) are designed to extract specific information from text. Features: These systems often use machine learning algorithms and knowledge bases to extract relevant information.Comparison of Existing Systems When comparing existing systems for automated phrase mining, consider the following factors: Accuracy: Evaluate the system's ability to extract high-quality phrases. Efficiency: Consider the system's performance in terms of processing large amounts of text. Flexibility: Assess the system's ability to adapt to different domains and languages. Ease of use: Evaluate the system's usability and the level of expertise required. Challenges and Future Directions Despite the advancements in automated phrase mining, Handling ambiguity: Dealing with ambiguous phrases and context-dependent meanings. Domain adaptation: Adapting phrase mining systems to specific domains and industries. Scalability: Processing large amounts of text data efficiently. Future research directions may include: Multilingual phrase mining: Developing systems that can handle multiple languages and cultural nuances. Domain-specific phrase mining: Creating systems that can adapt to specific domains and industries. Explainability and transparency: Developing systems that provide insights into the phrase mining process and results.

5. IMPLEMENTATION

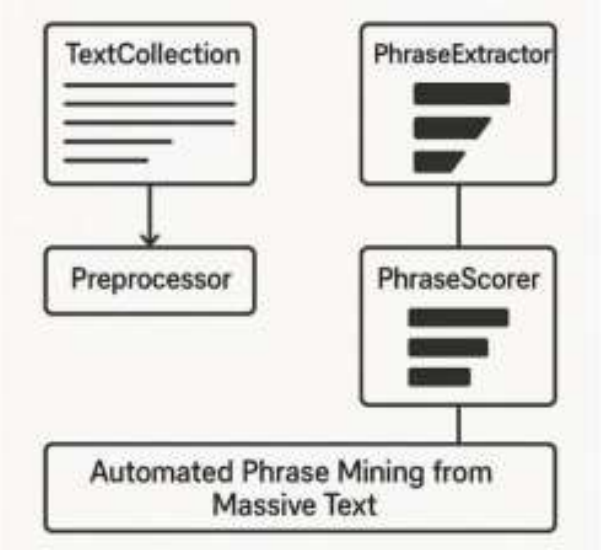

Figure 1: Class Diagram

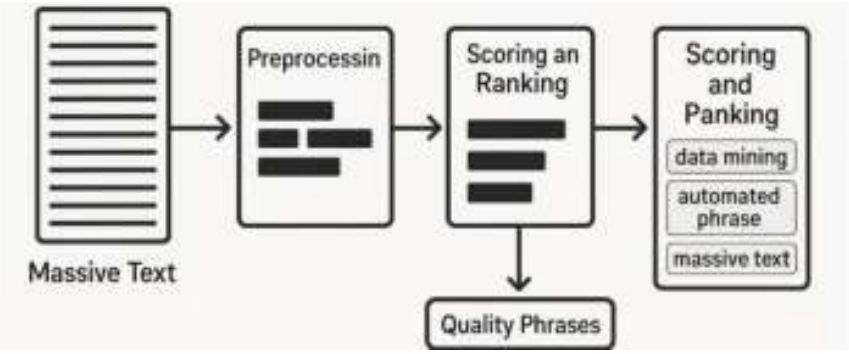

Figure 2: Sequential Diagram

6. CONCLUSION

Automated phrase mining efficiently extracts meaningful phrases from large text corpora with minimal human input. It helps in organizing unstructured data for tasks like summarization, information retrieval, and knowledge discovery. While effective and scalable, challenges remain in handling noisy, domain-specific, or multilingual text. Overall, it's a crucial technique in modern natural language processing.