Phrase Mining from Massive Text and Its Applications

Synthesis Lectures on Data Mining and Knowledge Discovery

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, and Robert Grossman. The series publishes 50- to 150-page publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. Potential topics include: data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Probabilistic Approaches to Recommendations
Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco
2014

Outlier Detection for Temporal Data
Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
2014

Provenance Data in Social Media
Geoffrey Barbier, Zhuo Feng, Pritam Gundecha, and Huan Liu
2013

Graph Mining: Laws, Tools, and Case Studies
D. Chakrabarti and C. Faloutsos
2012

Mining Heterogeneous Information Networks: Principles and Methodologies
Yizhou Sun and Jiawei Han
2012

Privacy in Social Networks
Elena Zheleva, Evimaria Terzi, and Lise Getoor
2012

Community Detection and Mining in Social Media
Lei Tang and Huan Liu
2010

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Phrase Mining from Massive Text and Its Applications

Jialu Liu, Jingbo Shang, and Jiawei Han

www.morganclaypool.com

ISBN: 9781627058988 paperback
ISBN: 9781627059183 ebook

DOI 10.2200/S00759ED1V01Y201702DMK013

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #13
Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of California, Santa Cruz*
Wei Wang, *University of California, Los Angeles*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Print 2151-0067 Electronic 2151-0075

Phrase Mining from Massive Text and Its Applications

Jialu Liu
Google

Jingbo Shang
University of Illinois at Urbana-Champaign

Jiawei Han
University of Illinois at Urbana-Champaign

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #13

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

A lot of digital ink has been spilled on "big data" over the past few years. Most of this surge owes its origin to the various types of unstructured data in the wild, among which the proliferation of text-heavy data is particularly overwhelming, attributed to the daily use of web documents, business reviews, news, social posts, etc., by so many people worldwide. A core challenge presents itself: How can one efficiently and effectively turn massive, unstructured text into structured representation so as to further lay the foundation for many other downstream text mining applications?

In this book, we investigated one promising paradigm for representing unstructured text, that is, through automatically identifying high-quality phrases from innumerable documents. In contrast to a list of frequent *n*-grams without proper filtering, users are often more interested in results based on variable-length phrases with certain semantics such as scientific concepts, organizations, slogans, and so on. We propose new principles and powerful methodologies to achieve this goal, from the scenario where a user can provide meaningful guidance to a fully automated setting through distant learning. This book also introduces applications enabled by the mined phrases and points out some promising research directions.

KEYWORDS

phrase mining, phrase quality, phrasal segmentation, distant supervision, text mining, real-world applications, efficient and scalable algorithms

Contents

Acknowledgments

The authors would like to acknowledge Xiang Ren, Fangbo Tao, and Huan Gui for their contribution to Chapter 4.

The research was supported in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617, and IIS 16-18481, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. The views and conclusions contained in our research publications are those of the authors and should not be interpreted as representing any funding agencies.

Jialu Liu, Jingbo Shang, and Jiawei Han
February 2017

CHAPTER 1

Introduction

1.1 MOTIVATION

The past decade has witnessed the surge of interest in data mining which is broadly construed to discover knowledge from all kinds of data, be it in academia, industry, or daily life. The information explosion brings the "big data" era to the light of the stage. This overwhelming tide of information is largely composed of *unstructured data* such as images, speeches, and videos. It is easy to distinguish them from typical *structured data* (e.g., relational data) in that the latter can be readily stored in the fielded form in databases. Among the various unstructured data, a particularly prominent category comes in the form of text. Examples include news articles, social media messages, as well as web pages and query logs.

In the literature of text mining, during the process of analyzing text, one fundamental problem is how to effectively represent text and model its topic, not only from the perspective of algorithm performance, but also for analysts to better interpret and present the results. A common approach is to use n-gram, i.e., a contiguous sequence of n unigrams, as the basic units. Figure 1.1 shows an example sequence with the corresponding 1-gram, 2-gram, 3-gram and consolidated representation. However, such representation raises concerns of exponential growth of the dictionary as well as the lack of interpretability. One can reasonably expect an intelligent method that only uses a compact subset of n-grams but generates explainable representation given a document.

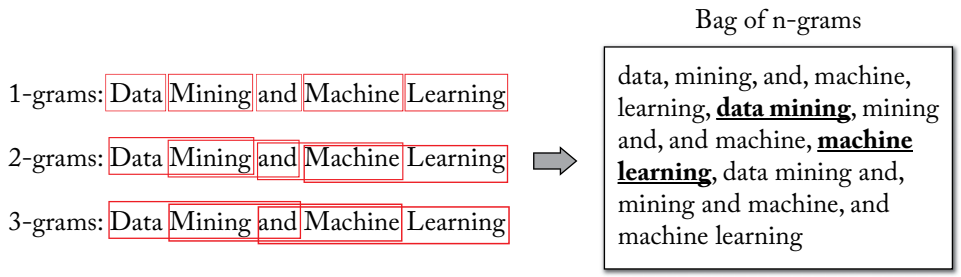

Figure 1.1: Example of n-gram representation.

Along this line of thought, in this book, we formulate such explainable n-gram subset as *quality phrases* (e.g., scientific terms such as "data mining" and "machine learning" outlined in the figure) and *phrase mining* as the corresponding knowledge discovery process.

Phrase mining has been studied in different communities. The natural language processing (NLP) community refers to it as "automatic term recognition" (i.e., extracting technical terms with the use of computers). The information retrieval (IR) community studies this topic to select main concepts in a corpus in an effort to improve search engine. Among existing works published by these two communities, linguistic processors with heuristic rules are primarily used and the most common approach is based on noun phrases. Supervised noun phrase chunking techniques are particularly proposed to leverage annotated documents to learn these rules. Other methods may utilize more sophisticated NLP features, such as dependency parser to further enhance the precision. However, emerging textual data, such as social media messages, can deviate from rigorous language rules. Using various kinds of heavily (pre-)trained linguistic processing makes these approaches difficult to be generalized.

In this regard, we believe that the community would welcome and benefit from a set of *data-driven* algorithms that work for *large-scale* datasets involving irregular textual data in a robust way, while minimizing the human labeling cost. We are also convinced by various study and experiments that our proposed methods embody enough novelty and contribution to add solid building block for various text-related tasks including document indexing, keyphrase extraction, topic modeling, knowledge base construction, and so on.

1.2 WHAT IS PHRASE MINING?

Phrase mining is a text mining technique that discovers semantically meaningful phrases from massive text. By considering the challenge of heterogeneity in the emerging textual data, the principles and methods discussed in this book will not assume particular lexical rules and are primarily compelled by data. Formally, we define the task as follows.

Problem 1.1 Phrase Mining Given a large document corpus $\mathcal{C}$—which can be any textual word sequences with arbitrary lengths such as articles, titles, and queries—phrase mining tries to assign a value between 0 and 1 to indicate the quality of each phrase mentioned in D and discovers a set of quality phrases $K = \{K_1, \cdots, K_M\}$ with their quality scores greater than 0.5. It also seeks to provide a segmenter for locating quality phrase mentions in any unseen text snippet.

Definition 1.2 Quality Phrase. A quality phrase is a sequence of words that appear contiguously in the corpus, and serves as a complete (non-composible) semantic unit in certain context among given documents.

There is no universally accepted definition for phrase quality. However, it is useful to quantify phrase quality based on certain criteria as outlined below:

- **Popularity**: Quality phrases should occur with sufficient frequency in the given document collection.

- **Concordance**: Concordance refers to the collocation of tokens in such a frequency that is significantly higher than what is expected due to chance. A commonly used example of a phraseological-concordance is the two phrases "strong tea" and "powerful tea." One would assume that the two phrases appear in similar frequency, yet in the English language the phrase "strong tea" is considered more proper and appears with much higher frequency. Because a concordant phrase's frequency deviates from what is expected, we consider them as belonging to a whole semantic unit.
- **Informativeness**: A phrase is informative if it is indicative of a specific topic or concept. The phrase "this paper" is popular and concordant, but is not considered to be informative in the bibliographic corpus.
- **Completeness**: Long frequent phrases and their subsequences may both satisfy the above three criteria. But apparently not all of them are qualified. A quality phrase should be interpreted as a complete semantic unit in certain contexts. The phrase "vector machine" is not considered to be complete as it mostly appears with prefix word "support."

Because single-word phrases cannot be decomposed into multiple tokens, the concordance criteria is no longer definable. As an alternative, we propose the independence criteria and will introduce it in more detail in Chapter 3.

1.3 OUTLINE OF THE BOOK

The remaining chapters of the book are outlined as follows.

- **Chapter 2: Quality Phrase Mining with User Guidance** In the literature of phrase mining, earlier work focuses on efficiently retrieving recurring word sequences and ranking them according to frequency-based statistics. However, the raw frequency from the data tends to produce misleading quality assessment, and the outcome therefore is unsatisfactory. We attempt to rectify the decisive raw frequency to help discover the true quality of a phrase by examining the context of its mentions. With *limited labeling* effort from the user, the model is able to iteratively segment the corpus into non-overlapped words and phrase sequences such that: (1) the phrase quality estimated in the previous iteration guides the segmentation and (2) segmentation results rectify raw phrase frequency and improve the process of phrase quality estimation. Such an integrated framework benefits from mutual enhancement, and achieves both high quality and high efficiency.
- **Chapter 3: Automated Quality Phrase Mining** Almost all state-of-the-art methods in NLP, IR, and text mining communities require human experts at certain levels. Such reliance on manual efforts from domain experts becomes an impediment for timely analysis of massive, emerging text corpora. Besides this issue, an ideal *automated phrase mining* method is supposed to work smoothly for multiple languages with high performance in terms of precision, recall, and efficiency. We attempt to make the phrase mining automated by utilizing

external knowledge bases to remove human efforts and minimize the language dependency. Modeling single-word phrases at the same time also improves the performance, especially the recall.

Since phrase mining lays the foundation for many other downstream text mining applications, we opt to devote one chapter to discuss its applications during the latest research development.

- **Chapter 4: Phrase Mining Applications** Particularly, we would like to introduce three representative applications using phrase mining results.
- The first is a statistical inference algorithm for detecting latent quality phrases topically relevant to a single document. Previously mentioned phrase mining methods are able to locate any phrase mentions in a document, but they cannot provide the relatedness between the document and the phrase.
- The second application utilizes phrase mining results to systematically analyze large numbers of textual documents from the perspective of topic exploration. We discuss how to group phrases into clusters sharing the same topic, how to summarize commonalities and differences given multiple document collections, and how to incorporate document-associated metadata like authors and tags into the exploration process.
- The last application tries to construct semantically rich knowledge base out of unstructured text. Identifying the phrases in text that constitute entity mentions and assigning types to these spans as well as to the relations between entity mentions are the key to this process.

CHAPTER 2

Quality Phrase Mining with User Guidance

In large, dynamic collections of documents, analysts are often interested in variable-length phrases, including scientific concepts, events, organizations, products, slogans, and so on. Accurate estimation of phrase quality is critical for the extraction of quality phrases and will enable a large body of applications to transform from word granularity to phrase granularity. In this chapter, we study a segmentation-integrated framework to mine multi-word quality phrases with a small set of user-provided binary labels.

2.1 OVERVIEW

Identifying quality phrases has gained increased attention due to its value of handling increasingly massive text datasets. As the origin, the natural language processing (NLP) community has conducted extensive studies mostly known as automatic term recognition [Frantzi et al., 2000, Park et al., 2002, Zhang et al., 2008], referring to the task of extracting technical terms with the use of computers. This topic also attracts attention in the information retrieval (IR) community since appropriate indexing term selection is critical to the improvement of a search engine where the ideal indexing units should represent the main concepts in a corpus, beyond the bag-of-words.

Linguistic processors are commonly used to filter out stop words and restrict candidate terms to noun phrases. With pre-defined part-of-speech (POS) rules, one can generate noun phrases as term candidates to each POS-tagged document. Supervised noun phrase chunking techniques [Chen and Chen, 1994, Punyakanok and Roth, 2001, Xun et al., 2000] leverage annotated documents to automatically learn these rules. Other methods may utilize more sophisticated NLP features such as dependency parser to further enhance the precision [Koo et al., 2008, McDonald et al., 2005]. With candidate terms collected, the next step is to leverage certain statistical measures derived from the corpus to estimate phrase quality. Some methods further resort to reference corpus for the calibration of "termhood" [Zhang et al., 2008]. The various kinds of linguistic processing, domain-dependent language rules, and expensive human labeling make it challenging to apply the phrase mining technique to emerging big and unrestricted corpora which possibly encompass many different domains and topics such as query logs, social media messages, and textual transaction records. Therefore, researchers have sought more general data-driven approaches, primarily based on the frequent pattern mining principle [Ahonen, 1999, Simitsis et al., 2008]. Early work focuses on efficiently retrieving recurring word sequences, but many such se-

quences do not form meaningful phrases. More recent work filters or ranks them according to frequency-based statistics. However, the raw frequency from the data tends to produce misleading quality assessment, and the outcome is unsatisfactory, as the following example demonstrates.

Example 2.1 Raw Frequency-based Phrase Mining Consider a set of scientific publications and the raw frequency counts of two phrases "relational database system" and "support vector machine" and their subsequences in the *frequency* column of Table 2.1. The numbers are hypothetical but manifest several key observations: (i) the frequency generally decreases with the phrase length; (ii) both good and bad phrases can possess high frequency (e.g., "support vector" and "vector machine"); and (iii) the frequency of one sequence (e.g., "relational database system") and its subsequences can have a similar scale of another sequence (e.g., "support vector machine") and its counterparts.

Table 2.1: A hypothetical example of word sequence raw frequency

Sequence	Raw Frequency	Quality Phrase?	Rectified Frequency
relational database system	100	yes	70
relational database	150	yes	40
database system	160	yes	35
relational	500	N/A	20
database	1000	N/A	200
system	10000	N/A	1000
Sequence	**Raw Frequency**	**Quality Phrase?**	**Rectified Frequency**
support vector machine	100	yes	80
support vector	160	yes	50
vector maching	150	no	6
support	500	N/A	150
vector	1000	N/A	200
machine	10000	N/A	50

Obviously, a method that ranks the word sequences solely according to the frequency will output many false phrases such as "vector machine." In order to address this problem, different heuristics have been proposed based on comparison of a sequence's frequency and its sub- (or super-) sequences, assuming that a good phrase should have high enough (normalized) frequency compared with its sub-sequences and/or super-sequences [Danilevsky et al., 2014, Parameswaran et al., 2010]. However, such heuristics can hardly differentiate the quality of, e.g., "support vector" and "vector machine" because their frequencies are so close. Finally, even if the heuristics can indeed draw a line between "support vector" and "vector machine" by discriminating their fre-

quencies (between 160 and 150), the same separation could fail for another case like "relational database" and "database system."

Using the frequency in Table 2.1, all heuristics will produce identical predictions for "relational database" and "vector machine," guaranteeing one of them to be wrong. This example suggests the intrinsic limitations of using raw frequency counts, especially in judging whether a sequence is too long (longer than a minimum semantic unit), too short (broken and not informative), or right in length. It is a critical bottleneck for all frequency-based quality assessment.

2.2 PHRASAL SEGMENTATION

In this chapter, we discuss how to address this bottleneck through rectifying the decisive raw frequency that hinders discovering the true quality of a phrase. The goal of the *rectification* is to estimate how many times each word sequence should be interpreted in whole as a phrase in its occurrence context. The following example illustrates this idea.

Example 2.2 Rectification Consider the following occurrences of the six multi-word sequences listed in Table 2.1.

1. A ⌈relational database system⌋ for images…
2. ⌈Database system⌋ empowers everyone in your organization…
3. More formally, a ⌈support vector machine⌋ constructs a hyperplane…
4. The ⌈support vector⌋ method is a new general method of ⌈function estimation⌋…
5. A standard ⌈feature vector⌋ ⌈machine learning⌋ setup is used to describe…
6. ⌈Relevance vector machine⌋ has an identical ⌈functional form⌋ to the ⌈support vector machine⌋…
7. The basic goal for ⌈object-oriented relational database⌋ is to ⌈bridge the gap⌋ between…

The first four instances should provide positive counts to these sequences, while the last three instances should not provide positive counts to "vector machine" or "relational database" because they should not be interpreted as a whole phrase (instead, sequences like "feature vector" and "relevance vector machine" can). Suppose one can correctly count true occurrences of the sequences, and collect rectified frequency as shown in the *rectified* column of Table 2.1. The rectified frequency now clearly distinguishes "vector machine" from the other phrases, since "vector machine" rarely occurs as a whole phrase.

The success of this approach relies on reasonably accurate rectification. Simple arithmetics of the raw frequency, such as subtracting one sequence's count with its quality super sequence, are prone to error. First, which super sequences are quality phrases is a question in and of itself.

Second, it is context-dependent to decide whether a sequence should be deemed a whole phrase. For example, the fifth instance in Example 2.2 prefers "feature vector" and "machine learning" over "vector machine," even though neither "feature vector machine" nor "vector machine learning" is a quality phrase. The context information is lost when we only collect the frequency counts.

In order to recover the true frequency with best effort, we ought to examine the context of every occurrence of each word sequence and decide whether to count it as a phrase. The examination for one occurrence may involve enumeration of alternative possibilities, such as extending the sequence or breaking the sequence, and comparison among them. The test for word sequence occurrences could be expensive, losing the advantage in efficiency of the frequent pattern mining approaches.

Facing the challenge of accuracy and efficiency, we propose a segmentation approach called "phrasal segmentation," and integrate it with the phrase quality assessment in a unified framework with linear complexity (w.r.t the corpus size). First, the segmentation assigns every word occurrence to only one phrase. In the first instance of Example 2.2, "relational database system" are bundled as a single phrase. Therefore, it automatically avoids double counting "relational database" and "database system" within this instance. Similarly, the segmentation of the fifth instance contributes to the count of "feature vector" and "machine learning" instead of "feature," "vector machine," and "learning." This strategy condenses the individual tests for each word sequence and reduces the overall complexity while ensures correctness. Second, although there are an exponential number of possible partitions of the documents, we are concerned with those relevant to the phrase extraction task only. Therefore, we can integrate the segmentation with the phrase quality assessment, such that: (i) only frequent phrases with reasonable quality are taken into consideration when enumerating partitions; and (ii) the phrase quality guides the segmentation, and the segmentation rectifies the phrase quality estimation. Such an integrated framework benefits from mutual enhancement, and achieves both high quality and high efficiency.

A phrasal segmentation defines a partition of a sequence into subsequences, such that every subsequence corresponds to either a single word or a phrase. Example 2.2 shows instances of such partitions, where all phrases with high quality are marked by brackets ⌈⌋. The phrasal segmentation is distinct from word, sentence or topic segmentation tasks in natural language processing. It is also different from the syntactic or semantic parsing which relies on grammar to decompose the sentences with rich structures like parse trees. Phrasal segmentation provides the necessary granularity we need to extract quality phrases. The total count for a phrase to appear in the segmented corpus is called *rectified frequency*.

It is beneficial to acknowledge that a sequence's segmentation may not be unique, for two reasons. First, as we mentioned above, a word sequence may be regarded as a phrase or not, depending on the adoption customs. Some phrases, like "bridge the gap" in the last instance of Example 2.2, are subject to a user's requirement. Therefore, we seek for segmentation that accommodates the phrase quality, which is learned from user-provided examples. Second, a sequence could be ambiguous and have different interpretations. Nevertheless, in most cases, it does not

require perfect segmentation, no matter if such a segmentation exists, to extract quality phrases. In a large document collection, the popularly adopted phrases appear many times in a variety of context. Even with a few mistakes or debatable partitions, a reasonably high quality segmentation (e.g., yielding no partition like "support ⌈vector machine⌋") would retain sufficient support (i.e., rectified frequency) for these quality phrases, albeit not for false phrases with high raw frequency.

With the above discussions, we have the following formalization.

Definition 2.3 Phrasal Segmentation. Given a word sequence $C = w_1 w_2 \dots w_n$ of length n, a segmentation $S = s_1 s_2 \dots s_m$ for C is induced by a boundary index sequence $B = \{b_1, b_2, \dots, b_{m+1}\}$ satisfying $1 = b_1 < b_2 < \cdots < b_{m+1} = n + 1$, where a segment $s_t = w_{b_t} w_{b_t+1} \dots w_{b_t+|s_t|-1}$. Here $|s_t|$ refers to the number of words in segment s_t. Since $b_t + |s_t| = b_{t+1}$, for clearness we use $w_{[b_t, b_{t+1})}$ to denote word sequence $w_{b_t} w_{b_t+1} \cdots w_{b_t+|s_t|-1}$.

Example 2.4 Continuing our previous Example 2.2 and specifically for the first instance, the word sequence and marked segmentation are

$$C = \text{a relational database system for images}$$
$$S = / \text{ a } / \text{ relational database system } / \text{ for } / \text{ images } /$$

with a boundary index sequence $B = \{1, 2, 5, 6, 7\}$ indicating the location of segmentation symbol /.

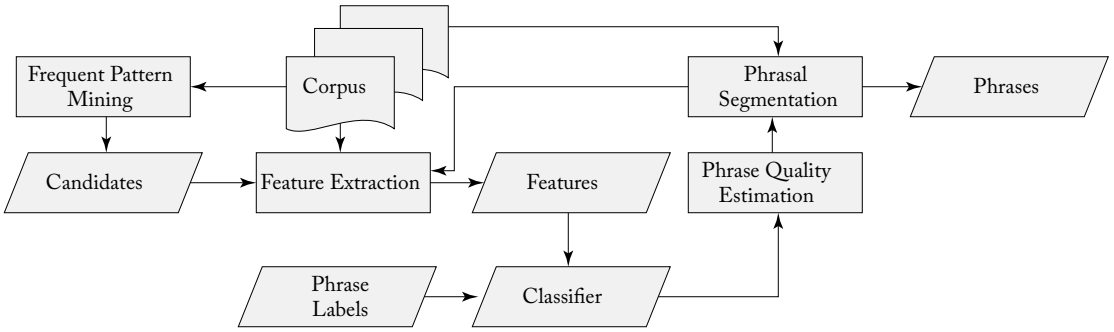

Figure 2.1: The supervised phrase mining framework.

2.3 SUPERVISED PHRASE MINING FRAMEWORK

In this chapter, in addition to the input corpus originally mentioned in Definition 1.1, users are required to provide a small set L of labeled quality phrases and $\bar{L}$ of inferior ones, which serves as the training data to guide the phrasal segmentation. The supervised framework comprises the following five steps and try to mine quality phrases following the quality criteria described in Section 1.2.

1. Generate frequent phrase candidates according to popularity criterion (Section 2.3.1).
2. Estimate phrase quality based on features design for concordance and informativeness criteria (Section 2.3.2).
3. Estimate rectified frequency via phrasal segmentation (Section 2.3.3).
4. Add segmentation-based features derived from rectified frequency into the feature set of phrase quality classifier (Section 2.3.4). Repeat steps 2 and 3.
5. Filter phrases with low rectified frequencies to satisfy the completeness criterion as post-processing step.

An complexity analysis for this framework is given at Section 2.3.5 to show that both of its computation time and required space grow linearly as the corpus size increases.

2.3.1 FREQUENT PHRASE DETECTION

The task of detecting frequent phrases can be defined as collecting aggregate counts for all phrases in a corpus that satisfy a certain minimum support threshold τ, according to the popularity criterion. In practice, one can also set a maximum phrase length ω to restrict the phrase length. Even if no explicit restriction is added, ω is typically a small constant. For efficiently mining these frequent phrases, we draw upon two properties.

1. Downward Closure property: If a phrase is not frequent, then any its super-phrase is guaranteed to be not frequent. Therefore, those longer phrases will be filtered and never expanded.
2. Prefix property: If a phrase is frequent, any of its prefix units should be frequent too. In this way, all the frequent phrases can be generated by expanding their prefixes.

The algorithm for detecting frequent phrases is given in Algorithm 1. We use $\mathcal{C}[\cdot]$ to index a word in the corpus string and $|\mathcal{C}|$ to denote the corpus size. The $\oplus$ operator is for concatecating two words or phrases. Algorithm 1 returns a key-value dictionary f. Its keys are vocabulary $\mathcal{U}$ containing all frequent phrases $\mathcal{P}$, and words $\mathcal{U} \setminus \mathcal{P}$. Its values are their raw frequency.

2.3.2 PHRASE QUALITY ESTIMATION

Estimating phrase quality from only a few training labels is challenging since a huge number of phrase candidates might be generated from the first step and they are messy. Instead of using one or two statistical measures [El-Kishky et al., 2015, Frantzi et al., 2000, Park et al., 2002], we opt to compute multiple features for each candidate in $\mathcal{P}$. A classifier is trained on these features to predict quality Q for all unlabeled phrases. For phrases not in $\mathcal{P}$, their quality is simply 0.

We divide the features into two categories according to concordance and informativeness criteria in the following two subsections. Only representative features are introduced for clearness. We then discuss about the classifier in Section 14.