Automated Phrase Mining and Phrasal Segmentation from Massive Corpora

Uma E. S[1] Saranya U[2]
[1]Assistant Professor [2]PG Scholar
[1,2]Department of Computer Science & Engineering
[1,2]Cochin College of Engineering, India

Abstract— As one of the fundamental tasks in text analysis, phrase mining aims at extracting quality phrases from a text corpus. Compared to the state-of- the-art methods, the new method has shown significant improvements on effectiveness on five real-world datasets in different domains recently, a few data-driven methods have been developed successfully for extraction of phrases from massive domain-specific text. However, none of the state-of-the-art models is 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 can achieve high performance with minimal human effort. In addition, we develop a POS-guided phrasal segmentation model, which incorporates the shallow syntactic information in part-of-speech (POS) tags to further enhance the performance, when a POS tagger is available. Note that, AutoPhrase can support 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.

Key words: Quality Phrases, AutoPhrase, POS Tagger

I. INTRODUCTION

Mining quality phrases refers to automatically extracting salient phrases from a given corpus. It is a fundamental task for text analytics of various domains, such as science, news, social media and enterprise documents. In these large, dynamic collections of documents, analysts are often interested in variable-length phrases, including scientic concepts, events, organizations. This paper deals with quality phrase extraction from a large collection of documents. The input documents can be any textual word sequences with arbitrary lengths, such as. articles, titles, queries, tags. Representing the text with quality phrases instead of n-grams can improve computational models for applications such as information. Therefore, for distant training, we leverage the existing high-quality phrases, as available from general knowledge bases, such as Wikipedia.

Bearing this in mind, we propose a novel automated phrase mining framework AutoPhrase in this paper, going beyond SegPhrase, to further get rid of additional manual labeling effort and enhance the performance, mainly using the following two new techniques.

A. Robust Positive-Only Distant Training

In fact, many high-quality phrases are freely available in general knowledge bases, and they can be easily obtained to a scale that is much larger than that produced by human experts. Domain-specific corpora usually contain some quality phrases also encoded in general knowledge bases, even when there may be no other domain-specific knowledge bases. Therefore, for distant training, we leverage the existing high- quality phrases, as available from general knowledge bases, such as Wikipedia and Freebase, to get rid of additional manual labelling effort. We independently build samples of positive labels from general knowledge bases and negative labels from the given domain corpora, and train a number of base classifiers. We then aggregate the predictions from these classifiers, whose independence helps reduce the noise from negative labels.

B. POS-Guided Phrasal Segmentation

There is a trade off between the performance and domain-independence when incorporating linguistic processors in the phrase mining method. On the domain independence side, the accuracy might be limited without linguistic knowledge. It is difficult to support multiple languages, if the method is completely language-blind. On the accuracy side, relying on complex, trained linguistic analyzers may hurt the domain-independence of the phrase mining method. For example, it is expensive to adapt dependency parsers to special domains like clinical reports. As a compromise, we propose to incorporate a pre-trained part-of-speech (POS) tagger to further enhance the performance, when it is available for the language of the document collection. The POS-guided phrasal segmentation leverages the shallow syntactic information in POS tags to guide the phrasal segmentation model locating the boundaries of phrases more accurately.

As demonstrated in our experiments, AutoPhrase works effectively in multiple domains like scientific papers, business reviews, and Wikipedia articles and also supports multiple languages, such as English, Spanish, and Chinese. Our main contributions are highlighted as follows:

- We study an important problem, automated phrase mining, and analyse its major challenges as above.
- We propose a robust positive-only distant training method for phrase quality estimation to minimize the human effort.
- We develop an novel phrasal segmentation model to leverage POS tags to achieve further improvement, when a POS tagger is available.
- We demonstrate the robustness and accuracy of our method and show improvements over prior methods, with results of experiments conducted on five real-world datasets in different domains (scientific papers, business reviews, and Wikipedia articles) and different languages (English, Spanish, and Chinese).

II. RELATED WORK

The unstructured text data is transforming unstructured text into structured units (semantically meaningful phrase), extracting salient phrases from corpus as science, news, social media and enterprise documents. Applications as topic tracking, document categorization. Human judgement is

required. This method does not depend on external sources, it mainly serves the purpose of rectifying phrase frequencies and checking phrase quality. Mainly four algorithms are; (i) frequent phrase detection (ii) dynamic programing (iii) Viterbi algorithm (iv) Penalty learning. [3]. The main drawback of this method is quality is unsatisfactory.

ATR techniques are mostly based on frequency, C-value also uses the parameter of frequency. This is a domain independent method, semiautomatic extraction of multi-word terms from special language English corpora. Algorithm used here is C-value algorithm. Mainly two methods are using here; (i) C-value, which is used for collocation extraction. (ii) NC-value, the method for extraction of term context words [6].

The framework is used for topical key phrase generation and ranking, that is shifting from a unigram-centric to a phrase centric approach. Domain here using are as multiple real world document collections. There are basically three steps; (i) Clustering words using topic modelling. (ii) Candidate key phrase generation. (iii) Ranking key phrases for topic representation. And mainly KERT (key phrase extraction and ranking by topic) algorithm is using in this method, com [1]. As a negative point, we can say that human judgement is depending here and it is formally analysing shorter texts.

There is a novel frequent pattern tree (FP-tree) structure for storing compressed, crucial information about frequent pattern. Divide and conquer method is using here on FP-tree, this is like the apriori-like candidate set generation and test approach. Following are the steps of this approach; (i) a novel compact data structure FP-tree is creating. (ii) Mining frequent patterns. (iii) FP-tree based pattern fragment growth method. Two algorithms are used here; (i) FP-tree construction algorithm. (ii) FP-growth: Mining frequent patterns with FP-tree by pattern fragment growth [7]. The FP-tree is not can be used in large industrial databases, so no satisfactory performance will get.

The hierarchical generative probabilistic model of topical phrases, this model simultaneously infers the location, length and topic of the phrases within corpus by using hierarchy of "Pitman-Yor processes". And here using "Markov chain Monte Carlo technique" for approximate inference in the model and perform slice sampling to learn its hyper parameters. This method related to "Bag- of words" assumption, that is word order is ignored. The domain we choose as, human subjects. The archetypal topic model, Latent Dirichlet Allocation (LDA), posits that words within a document are conditionally independent given their topic. Topic phrases found by PDLDA (Phrase- Discovering LDA). Policy iteration algorithm is used here [2]. As a drawback we can say that it is difficult to find topical phrases in large corpora.

For topical phrase mining we present a novel framework CITPM. In this method the corpus as a mixture of cluster's (domain), has similar topical distribution in each cluster. Until a satisfactory final result is obtained, the cluster is updating. There are four stages are here; (i) Pre-processing (ii) Phrase mining (iii) Topic modelling (iv) Clustering. Phrase frequency counting, phraseness checking and DPBK clustering are the algorithms used in this technique. But for the clustering stage and for the topic modelling stage there are two separate algorithms are used, this is a drawback in this paper [4]. In future we can combine the both algorithms.

Extracting domain specific glossaries (a brief dictionary) from large document in an automatic manner, from a large document collection. We have used these methods to build GlossEx, a glossary extraction tool. This technique can be used in automotive engineering and computer help desk domain. Glossary extraction algorithm is used in this paper. In this paper, we will show that, from the identification of single word forms, abbreviations, verbs and salience, the glossary items can be obtained. POS tagging and parsing, induction and application of multiple forms and statistical computations of item distribution etc. are the techniques used in this paper [5]. The negative point of this paper is, only create the list of glossary, no further applications performed here. And we can add the ontology as a future extension.

Mining interesting phrases, specified using features such as keywords, from document sub-collections. Conditional independence assumption (probability of a given word is independent of the other words) is used here. In this paper mainly two algorithms are used; (i) Scoring using disk-resident indexes (NRA) algorithm. (ii) Scoring using phrase ID-ordered lists (SMJ) algorithm [8]. In large size of corpus, it is difficult to mining the interested phrases as keywords.

There is an efficient high-quality phrase mining approach (EQPM). We propose an efficient integrated framework for high quality topical phrase mining, which adopts complete phrase mining to guarantee completeness, and utilizes a novel phrasal segmentation model to handle overlapping phrases. Here we consider both intra-cohesion and inter- isolation in mining phrases, which is able to guarantee appropriateness [10]. This method is not automatic, this is a negative point in this paper.

Presenting topical n-grams, at topic model that discovers topics as well as topical phrases. The probabilistic model generates words in their textual order by, for each word, first sampling a topic, then sampling its status as a unigram or bigram, and then sampling the word from a topic-specific unigram or bigram distribution. There are three N-gram based topic models; (i) Bigram Topic Model (BTM). (ii) LDA Collocation Model (LDACOL). (iii) Topical N-gram Model (TNG) [11]. Accuracy is less sometimes here, and phrase not giving exact meaning all the time.

The technique presented the task of topic labelling, that is the generation and scoring of labels for a given topic. We generate a set of label candidates from the top-ranking topic terms, titles of Wikipedia articles containing the top-ranking topic terms, and also a filtered set of sub-phrases extracted from the Wikipedia article titles. We rank the label candidates using a combination of association measures, lexical features and an Information Retrieval feature [9]. In future we can further improve the method by including segmentations with its algorithms.

Evaluating the proposed methods on a large Twitter data set. Experiments show that these methods are very effective for topical key phrase extraction. In this paper, we studied the novel problem of topical key phrase extraction for summarizing and analysing Twitter content. We proposed the context-sensitive topical Page Rank (cTPR) method for keyword ranking. Experiments showed that cTPR is

consistently better than the original TPR and other base line methods in terms of top keyword and key phrase extraction. For key phrase ranking, we proposed a probabilistic ranking method, which models both relevance and interestingness of key phrases. In our experiments, this method is shown to be very effective to boost the performance of key phrase extraction for different kinds of keyword ranking methods [14].

Providing a new perspective to key phrase extraction: regarding a document and its key phrases as descriptions to the same object written in two languages. There are two methods are here; (i) State of the Art. (ii) Key phrase Extraction by Bridging Vocabulary Gap Using WAM. As a future scope Explore more complicated methods to extract important sentences for constructing translation pairs [16]. And as a negative point we can say that, this paper is not working in many languages and other type of articles.

We present a survey of the state of the art in automatic key phrase extraction, examining the major sources of errors made by existing systems and discussing the challenges ahead. Our analysis revealed that there are at least three major challenges ahead; (i) Incorporating background knowledge. (ii) Handling long documents. (iii) Improving evaluation schemes [12]. This paper is not giving good results comparatively, because only one model is using here.

Introducing a new method for identifying candidate phrasal terms (also known as multiword units) which applies a nonparametric, rank-based heuristic measure. Evaluation of this measure, the mutual rank ratio metric, shows that it produces better results than standard statistical measures when applied to this task. Evaluation indicates that this method may outperform standard lexical association measures, including mutual information, chi-squared, log-likelihood, and the T-score [15]. This method is not automatic type of mining, so we can say this point as a negative.

Proposed a frequent pattern-based data enrichment, a general method for improving topic model performance with frequent pattern mining. Among the variations of patterns, the proposed compressed/closed frequent pattern-based data enrichment consistently outperformed two representative topic models. In addition to its performance improvement, frequent pattern based data enrichment has advantage in its generality, which is shown in our experiment that we applied frequent pattern based data enrichment to both PLSA and LDA topic models. We may apply other sophisticated patterns to improve the performance of our frequent pattern- based data enrichment [13]. Exploration of other types of ungapped patterns would help us to further explore the effect of gaps. This is a drawback of this paper.
Introducing Energy Evaluation and Prediction System (EEPS). Based on the data provided by SEDS, this paper proposed four models to select and aggregate important information. Firstly, Energy Profile Model (EPM) is established to cluster the data. Secondly, based on EPM, time is considered and we get the main energy percentage diagram of each state during 50 years. Thirdly, to determine which of the four states appeared to use clean energy best in 2009, New Energy Profile Model (NEPM) is established. Fourthly, Energy Profile Prediction Model (EPPM) is established to predict the energy profile of 2025 and 2050. Fifthly, EPPM and NEPM are used to predict and evaluate the use condition of energy in 2025 and 2050. From the prediction results, the clean energy used is increasing [19]. This paper is not working automatically and having expensive method.

With the increase of technology and computational storage facilities, usage of Real World Data (RWD) and Big Data Mining (BDM) techniques is proving to be a useful tool for automated data analysis. Entities dealing daily with medical practice such as clinics and hospitals possess databases with a wellspring of information worthwhile being studied to aid clinicians establishing disease patterns identification, future trends, and therapeutic relationships. Aiming at assessing cardiovascular disease (CVD) progression of diabetic patients, a nearly 20 years old private clinic database was studied. Primarily goal, and subject of this paper, was the evaluation of the data-base reliability to continue the study of CVD progression. Manual inspection of the database content revealed missing and misleading fields, inconsistency of inputted instrumental data, temporal and user dependency of fields filling, particularly concerning CV data [17]. We are dependent on a clinician or technician understanding of the database, so it is not automatic, human requirement is needed.

Presented a topical phrase mining framework, Top Mine that discovers arbitrary length topical phrases. Our framework mainly consists of two parts: phrase mining and phrase-constrained topic modelling. In the first part, we use frequent phrase mining to efficiently collect necessary aggregate statistics for our significance score - the objective function that guides our bottom-up phrase construction. Upon termination, our phrase mining step segments each document into a bag of phrases. In the first part, we use frequent phrase mining to efficiently collect necessary aggregate statistics for our significance score - the objective function that guides our bottom-up phrase construction. Upon termination, our phrase mining step segments each document into a bag of phrases [20]. The techniques used here are less accurate, and not automatic.

Neural networks represent an important tool in data classification. It is the only technique that allows generalizations based on a set of data to be analysed. Using this method, a number close to the optimal number of hidden layers of a multi-layer neural network can be obtained. There are three clustering techniques; (i) Optimization Problem. (ii) Proposed method for optimization. (iii) Number of hidden layers using clustering technique [18]. In future we can extend the method with new accurate techniques.

III. CONCLUSION AND FUTURE WORK

In this paper, we present an automated phrase mining framework with two novel techniques: the robust positive only distant training and the POS-guided phrasal segmentation incorporating part-of-speech (POS) tags, for the development of an automated phrase mining framework Auto Phrase. Our extensive experiments show that Auto Phrase is domain independent, outperforms other phrase mining methods, and supports multiple languages (e.g., English, Spanish, and Chinese) effectively, with minimal human effort.

We develop a segmentation-integrated approach for this purpose, which significantly boosts the final quality of

extracted phrases. It is the first work to explore the mutual benefit of phrase extraction and phrasal segmentation. By integrating them, this work addresses a fundamental limitation of phrase mining and empowers widespread applications. Meanwhile, the method is scalable: both computation time and required space grow linearly as corpus size increases.

For future work, it is interesting to (1) refine quality phrases to entity mentions, (2) apply Auto Phrase to more languages, such as Japanese, and (3) for those languages without general knowledge bases, seek an unsupervised method to generate the positive pool from the corpus, even with some noise. Another direction is to replace Viterbi Training by other parameter estimation approaches to further improve the phrasal segmentation.

REFERENCES

[1] M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, and J. Han. Automatic construction and ranking of topical keyphrases on collections of short documents. In SDM, ©2014.
[2] Robert V. Lindsey, Michael J. Stipicevic. A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. @2012 Association for Computational Linguistics.
[3] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, Jiawei Han. Mining Quality Phrases from Massive Text Corpora. In 2015.
[4] B. Li, B. Wang, R. Zhou, X. Yang, and C. Liu. Citpm: A clusterbased iterative topical phrase mining framework. In International Conference on Database Systems for Advanced Applications, pages 197–213. Springer, ©2016.
[5] Roy J Byrd, Youngja Park and Branimir K Boguraev. Automatic Glossary Extraction: Beyond Terminology Identification. IBM Thomas J Watson Research centre, USA.
[6] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. JODL, 3(2):115– 130, 2000.
[7] Jiawei Han, Jian Pei, and Yiwen Yin. Mining Frequent Patterns without Candidate Generation. School of Computing Science Simon Fraser University.
[8] D. P, A. Dey, and D. Majumdar. Fast mining of interesting phrases from subsets of text corpora. In EDBT, ©2014.
[9] JeyHanLau, KarlGrieser, DavidNewman and Timothy Baldwin. Automatic Labelling of Topic Models. @ 2011 Association for Computational Linguistics.
[10] B. Li, X. Yang, B. Wang, and W. Cui. Efficiently mining high quality phrases from texts. In AAAI, pages 3474–3481, ©2017.
[11] Xuerui Wang, Andrew McCallum, Xing Wei. Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. University of Massachusetts, ©2007.
[12] Kazi Saidul Hasan and Vincent Ng. Automatic Key phrase Extraction: A Survey of the State of the Art. Human Language Technology Research Institute University of Texas at Dallas. 2014 Association for Computational Linguistics.
[13] Hyun Duk Kim, Dae Hoon Park, Yue Lu, Cheng Xiang Zhai. Enriching Text Representation with Frequent Pattern Mining for Probabilistic Topic Modeling. University of Illinois at Urbana- Champaign, ©2012.
[14] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, Xiaoming Li. Topical Key phrase Extraction from Twitter. School of Electronics Engineering and Computer Science, Peking University, School of Information Systems, Singapore Management University.
[15] Paul Deane. A Nonparametric Method for Extraction of Candidate Phrasal Terms. Center for Assessment, Design and Scoring Educational Testing Service.
[16] Z. Liu, X. Chen, Y. Zheng, and M. Sun. Automatic keyphrase extraction by bridging vocabulary gap. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 135–144, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
[17] M. G. Ruano, G. P. Almeida, F. Palma, J. F. Raposo, R. T. Ribeiro. Reliability of Medical Databases for the use of Real Word Data and Data Mining Techniques for Cardio vascular Diseases Progression in Diabetic Patients, © 2018 IEEE.
[18] Mohamed Lafif Tej, Stefan Holban. Determining Neural Network Architecture Using Data Mining Techniques, ©2018 IEEE.
[19] Zhaocong Sun, Chi Zhang, Tianyi Shi, Wan Cui. Energy Evaluation and Prediction System Based on Data Mining. ©2018 IEEE.
[20] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora, Department of Computer Science The University of Illinois at Urbana Champaign.

