AUTOMATED PHRASE MINING: SIGNIFICANT IMPROVEMENT PHRASE MINING FROM MASSIVE TEXT

DHARMA VARDHANI.MORA

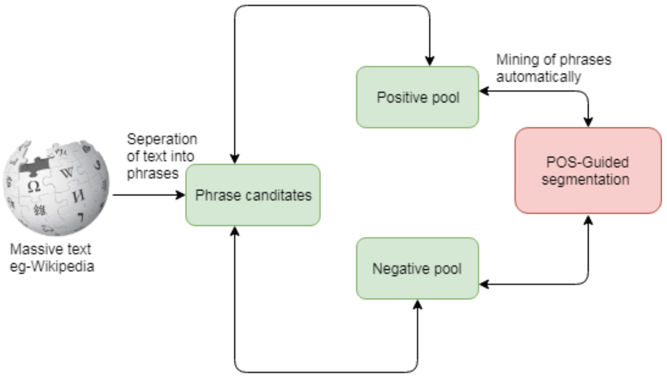
Lecturer, Dept. of Computer Science, Sri Durga Malleswara Siddhratha Mahila kalasala, A.P., India.

ABSTRACT: Phrase mining refers to the process of automatic extraction of high-quality phrases (e.g., scientific terms and general entity names) in a given corpus (e.g., research papers and news). Representing the text with quality phrases instead of n-grams can improve computational models for applications such as 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 state-of-the-art models, even data-driven 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, Auto Phrase, 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, Auto Phrase has shown significant improvements in both effectiveness and efficiency on five real-world datasets across different domains and languages. Besides, Auto Phrase can be extending to model single-word quality phrases.

Keywords: – Automated Phrase Mining, Text.

INTRODUCTION

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. Almost all the state-of-the-art methods, however, require human experts at certain levels. Most existing methods rely on complex, trained linguistic analyzers (e.g., dependency parsers) to locate phrase mentions, and thus may have unsatisfactory performance on text corpora of new domains and genres without extra but expensive adaption. Our latest domain-independent method Rephrase outperforms many other approaches, but still needs domain experts to first carefully select hundreds of varyingquality phrases from millions of candidates, and then annotate them with binary labels. Such reliance on manual efforts by domain and linguistic experts becomes an impediment for timely analysis of massive, emerging text corpora in specific domains. An ideal automated phrase mining method is supposed to be domain-independent, with minimal human effort 1 or reliance on linguistic analyzers. Bearing this in mind, we propose a novel automated phrase mining framework


Mining of phrases automatically
Positive pool
Seperation of text into phrases
Phrase canditates
POS-Guided segmentation
Massive text eg-Wikipedia
Negative pool

PROPOSED SYSTEM

Introducing our two new techniques. First, a novel robust positive-only 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. 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.

Identifying quality phrases efficiently has become ever more central and critical for effective handling of massively increasing-size text datasets. In contrast to key phrase extraction this task goes beyond the scope of single documents and utilizes useful cross-document signals. Interesting phrases can be queried efficiently for ad-hoc subsets of a corpus, while the phrases are based on simple frequent pattern mining methods. The natural language processing (NLP) community has conducted extensive studies typically referred to as automatic term recognition for the computational task of extracting terms (such as technical phrases). This topic also attracts attention in the information retrieval (IR) community since selecting appropriate indexing terms is critical to the improvement of search engines where the ideal indexing units represent the main concepts in a corpus, not just literal bag-of-words. Text indexing algorithms typically filter out stop words and restrict candidate terms to noun phrases. With predefined part-of-speech (POS) rules, one can identify noun phrases as term candidates in POS-tagged documents. Supervised noun phrase chunking techniques exploit such tagged documents to automatically learn rules for identifying noun phrase boundaries. Other methods may utilize more sophisticated NLP technologies such as dependency parsing to further enhance the precision. With candidate terms collected, the next step is to leverage certain statistical measures derived from the corpus to estimate phrase quality. Some methods rely on other reference corpora for the calibration of "term hood". The dependency on these various kinds of linguistic analyzers, domain-dependent language rules, and expensive human labeling, makes it challenging to extend these approaches to emerging, big, and unrestricted corpora, which may include many different domains, topics, and languages. To overcome this limitation, data-driven approaches opt instead to make use of frequency statistics in the corpus to address both candidate generation and quality estimation. They do not rely on complex linguistic feature generation, domain-specific rules or extensive labeling efforts. Instead, they rely on large corpora containing hundreds of thousands of documents to help deliver superior performance.

IMPLEMENTATION

- ✓ Concordance
- ✓ Informativeness
- ✓ Popularity
- ✓ Completeness

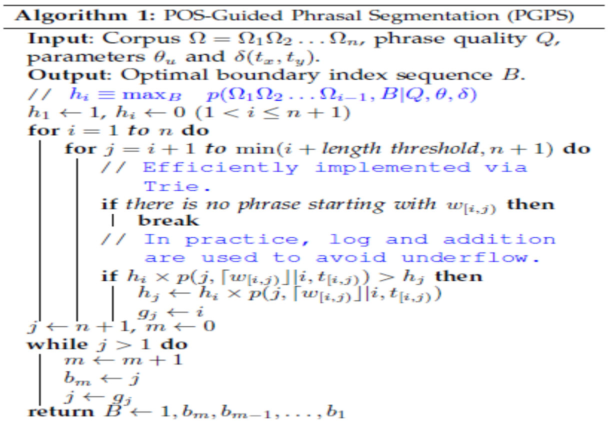
- ✓ **Concordance**: The collocation of tokens in quality phrases occurs with significantly higher probability than expected due to chance.
- ✓ **Informativeness**: A phrase is informative if it is indicative of a specific topic or concept.
- ✓ **Completeness**: Long frequent phrases and their subsequences within those phrases may both satisfy the 3 criteria above. A phrase is deemed complete when it can be interpreted as a complete semantic unit in some given document context. Note that a phrase and a subphrase contained within it, may both be deemed complete, depending on the context in which they appear. For example, "relational database system", "relational database" and "database system" can all be complete in certain context.
- ✓ **Popularity**: Quality phrases should occur with sufficient frequency in the given document collection.

Algorithm:

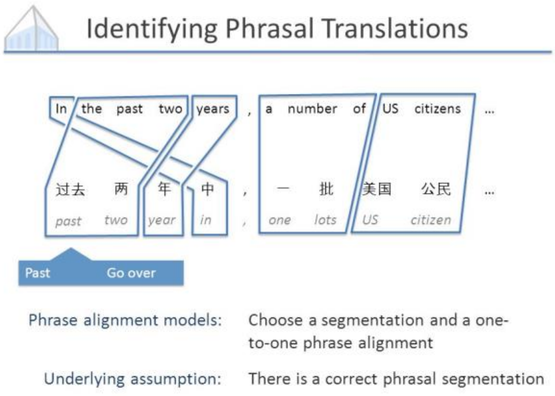
- ✓ **POS-Guided Phrasal Segmentation:**
- ✓ **Text indexing algorithms:**

POS-Guided Phrasal Segmentation:

There is a trade-off between the accuracy 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 well, 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.



```
Algorithm 1: POS-Guided Phrasal Segmentation (PGPS)
Input: Corpus $\Omega = \Omega_1\Omega_2 \ldots \Omega_n$, phrase quality $Q$,
parameters $\theta_u$ and $\delta(t_x, t_y)$.
Output: Optimal boundary index sequence $B$.
// $h_i \equiv \max_B \; p(\Omega_1\Omega_2 \ldots \Omega_{i-1}, B|Q, \theta, \delta)$
$h_1 \leftarrow 1$, $h_i \leftarrow 0$ $(1 < i \leq n+1)$
for $i = 1$ to $n$ do
    for $j = i + 1$ to $\min(i + \textit{length threshold}, n + 1)$ do
        // Efficiently implemented via
           Trie.
        if there is no phrase starting with $w_{[i,j)}$ then
            break
        // In practice, log and addition
           are used to avoid underflow.
        if $h_i \times p(j, \lceil w_{[i,j)} \rfloor | i, t_{[i,j)}) > h_j$ then
            $h_j \leftarrow h_i \times p(j, \lceil w_{[i,j)} \rfloor | i, t_{[i,j)})$
            $g_j \leftarrow i$
$j \leftarrow n + 1$, $m \leftarrow 0$
while $j > 1$ do
    $m \leftarrow m + 1$
    $b_m \leftarrow j$
    $j \leftarrow g_j$
return $B \leftarrow 1, b_m, b_{m-1}, \ldots, b_1$
```



Identifying Phrasal Translations

Phrase alignment models: Choose a segmentation and a one-to-one phrase alignment

Underlying assumption: There is a correct phrasal segmentation

✓ Text indexing algorithms:

Text indexing algorithms typically filter out stop words and restrict candidate terms to noun phrases. With predefined part-of-speech (POS) rules, one can identify noun phrases as term candidates in POS-tagged documents. Supervised noun phrase chunking techniques exploit such tagged documents to automatically learn rules for identifying noun phrase boundaries. Other methods may utilize more sophisticated NLP technologies such as dependency parsing to further enhance the precision. With candidate terms collected, the next step is to leverage certain statistical measures derived from the corpus to estimate phrase quality.

CONCLUSION

In this paper, we present an automated phrase mining framework with two novel techniques: the robust positiveonly distant training and the POS-guided phrasal segmentation incorporating part-of-speech (POS) tags, for the development of an automated phrase mining framework AutoPhrase. Our extensive experiments show that AutoPhrase is domain-independent, outperforms other phrase mining methods, and supports multiple languages (e.g., English, Spanish, and Chinese) effectively, with minimal human effort. Besides, the inclusion of quality single-word phrases (e.g., dUIUCc and dUSAc) which leads to about 10% to 30% increased recall and the exploration of better indexing strategies and more thorough parallelization, which leads to about 8 to 11 times running time speedup and about 80% to 86% memory usage saving over SegPhrase. Interested readers may try our released code at GitHub.

Future work:

For future work, it is interesting to

(1) Refine quality phrases to entity mentions,

(2) Apply AutoPhrase 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. Our extensive experiments show that AutoPhrase is domain-independent, outperforms other phrase mining methods, and supports multiple languages (e.g., English, Spanish, and Chinese) effectively, with minimal human effort. Besides, the inclusion of quality single-word phrases (e.g., dUIUCc and dUSAc) which leads to about 10% to 30% increased recall and the exploration of better indexing strategies and more thorough parallelization, which leads to about 8 to 11 times running time speedup and about 80% to 86% memory usage saving over SegPhrase. Interested readers may try our released code at GitHub.

REFERENCES

[1] K. Ahmad, L. Gillam, L. Tostevin, etal. University of surreyparticipation in trec8: Weirdness indexing for logical documentextrapolation and retrieval (wilder). In TREC, pages 1–8, 1999.

[2] A. Allahverdyan and A. Galstyan. Comparative analysis of viterbitraining and maximum likelihood estimation for hmms. In NIPS,pages 1674–1682, 2011.

[3] T. Baldwin and S. N. Kim. Multiword expressions. Handbook ofNatural Language Processing, second edition. Morgan and Claypool,2010.

[4] S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, and G. Weikum.Interesting-phrase mining for ad-hoc text analytics. Proc. VLDBEndow., 3(1-2):1348–1357, Sept. 2010.

[5] L. Breiman. Randomizing outputs to increase prediction accuracy.Machine Learning, 40(3):229–242, 2000.

[6] K.-h. Chen and H.-H. Chen. Extracting noun phrases from largescaletexts: A hybrid approach and its automatic evaluation.In Proceedings of the 32Nd Annual Meeting on Association forComputational Linguistics, ACL '94, pages 234–241, Stroudsburg,PA, USA, 1994. Association for Computational Linguistics.

[7] M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, and J. Han.Automatic construction and ranking of topical keyphrases oncollections of short documents. In SDM, 2014.

[8] M.-C. De Marneffe, B. MacCartney, C. D. Manning, etal.Generatingtyped dependency parses from phrase structure parses. InProceedings of LREC, volume 6, pages 449–454, 2006.

[9] P. Deane. A nonparametric method for extraction of candidatephrasal terms. In Proceedings of the 43rd Annual Meeting onAssociation for Computational Linguistics, ACL '05, pages 605–613, Stroudsburg, PA, USA, 2005. Association for ComputationalLinguistics.

[10] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalabletopical phrase mining from text corpora. Proc. VLDB Endow.,8(3):305–316, Nov. 2014.