**Automated Phrase Mining from Massive Text Corpora**

**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 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 extend to model single-word quality phrases.

**Index Terms**—Automatic Phrase Mining, Phrase Mining, Distant Training, Part-of-Speech tag, Multiple Languages

**Existing system:**

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 SegPhrase outperforms many other approaches but still needs domain experts to first carefully select hundreds of varying quality phrases from millions of candidates, and then annotate them with binary labels. 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 “termhood”.

**Disadvantages:**

* 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.
* SegPhrase, additional manual labeling effort needed.

**Proposed system:**

A novel automated phrase mining framework, Auto Phrase has two new techniques.1. Robust Positive-Only Distant Training,2. POS-Guided Phrasal Segmentation. AutoPhrase can support any language as long as a general knowledge base in that language is available, the POS-guided phrasal segmentation can be applied in many scenarios. It is worth mentioning that for domain-specific knowledge bases (e.g., MeSH terms in the biomedical domain) and trained POS taggers, the same paradigm applies. In this study, without loss of generality, we only assume the availability of a general knowledge base together with a pre-trained POS tagger.

**Advantages:**

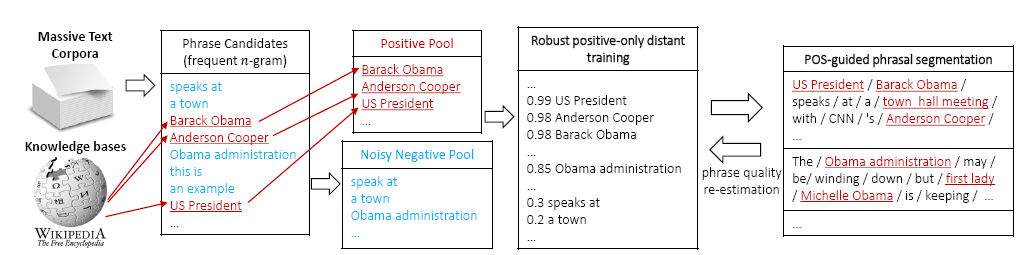
Our main contributions are highlighted as follows:

1. Automated phrase mining

2. A robust positive-only distant training method for phrase quality estimation to minimize the human effort.

3Develop a novel phrasal segmentation model to leverage POS tags to achieve further improvement, when a POS tagger is available.

**Architecture:**



**Modules:**

**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, to get rid of additional manual labeling 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.

**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.

**Label Pools:**

Public knowledge bases 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 synonyms5 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 ), 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**.

**Software Requirements:**

Operating System : Windows 7

User Interface : HTML, CSS

Client-side Scripting : JavaScript

Programming Language : Java

Web Applications : JDBC, Servlets, JSP

IDE/Workbench : My Eclipse 8.6

Database : Oracle 10g xe

Server Deployment : Tomcat 7.0

**Hardware Requirements (Minimum):**

Processor : Intel core i3 or above

Hard Disk : 40GB or more

RAM : 2GB or more

**CONCLUSIONS**

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 effectively, with minimal human effort.