

**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, 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 extended to model single-word quality phrases.