

Research on Lao word segmentation algorithm and development of Lao-Vietnamese new word lookup support

Nguyen Minh Tam

Abstract: This study presents a novel approach to Lao word segmentation and the development of a Lao-Vietnamese dictionary lookup system for new words. Lao presents unique challenges for segmentation due to its lack of explicit word boundaries, continuous script, and complex syllable structures. We propose a hybrid algorithm combining a rule-based method utilizing syllable-level heuristics with a statistical model trained on a manually annotated Lao corpus. The algorithm effectively identifies word boundaries by analyzing character sequences and syllable patterns. Furthermore, we constructed a bilingual dictionary support system that automatically extracts and indexes new or unknown Lao words from input text, providing their Vietnamese equivalents. The system leverages the segmentation output to identify candidate terms and matches them against a base lexicon, with unknown terms flagged for user contribution or lookup. Experimental results demonstrate a segmentation accuracy of over 92% on a test set, and the lookup system successfully identifies and translates over 85% of new compound words. This work provides a foundational tool for Lao language processing and facilitates cross-lingual information access.

Keywords: Lao word segmentation, Lao-Vietnamese dictionary, lookup tool, natural language processing