**Introduction**

Wikipedia has over 260 versions in different languages, but the disparity in article numbers and content quality is remarkable. Human authoring takes time, while machine-translated texts are hardly readable. This calls for an automatic approach that efficiently generates articles for weaker Wikipedia versions.

In this paper, we propose an unsupervised approach to automatically synthesize Wikipedia articles in multiple languages. Taking an existing high-quality version of any entry as content guideline, we extract keywords from it and use the translated keywords to query the monolingual web of the target language. Candidate excerpts or sentences are selected based on an iterative ranking function and eventually synthesized into a complete article that resembles the reference version closely.

The algorithm does not assume article domain or language pair. Evaluation results support the claim of domain- and language-independence. Both a web interface and a desktop application have been released for public use.

**Methodology**

Suppose English is the reference version and Chinese is the target version. The structured English article is first extracted from the Wikipedia page. A separate synthesis task is performed on each section and all synthesized sections are eventually combined in the original order to form the Chinese article.

**Synthesis Workflow**

- Extract keywords by tf-idf and TextRank scores, also extract named entities, time indicators and terms with hyperlinks.
- Use Chinese translation of keywords to search online Chinese texts. Section keywords alternatively form query pairs with the subject term.
- Generate excerpt objects with both Chinese and English versions.
- Filter retrieved Chinese excerpts based on cosine similarity and Jaccard index computed against the source.
- Split remaining excerpts into segments as synthesis units by the TextTiling algorithm.
- Synthesize the current section using candidate segments by an iterative segment selection algorithm.

**Evaluation**

Human evaluation on 20 entries across 5 domains showed that the synthesized article outperforms machine translated texts and in many cases more comprehensive than existing human-authored versions.

**Applications**

The system is implemented with Python. The web interface is written in PHP and the desktop GUI in wxPython.

**Conclusion**

For Wikipedia entries that have not yet been created, our system efficiently collects relevant information and fills the void. For those in 'sub' status, the synthesized articles can act as either replacements or supplements. For other relatively well-written ones, our system can provide content prototypes for missing sections and missing topics, facilitating later human editing.