PARALLEL ANNOTATED SYNTHETIC CORPORA (PASC)

注釈付き対訳合成コーパス

Jack Halpern (jack@cjki.org)
The CJK Dictionary Institute, Inc.

Summary

The Parallel Annotated Synthetic Corpora (PASC) project, led by Jack Halpern and developed by The CJK Dictionary Institute (CJKI), focuses on creating comprehensive synthetic corpora for various applications in natural language processing (NLP) and machine translation (MT).

Artificial corpora, including augmented and synthetic corpora, are gaining attention in the machine translation community. Synthetic data represents observations mimicking real-world situations, used to train machine learning models when actual data is scarce or expensive to obtain. Synthetic corpora, known as SC or artificial corpora, show promise in improving translation quality in NMT systems.

The PASC project aims to create Parallel Annotated Synthetic Corpora using supervised generation techniques. Unlike augmented corpora, which expand existing data, PASC constructs synthetic corpora from scratch using predefined sentence templates, ensuring adherence to linguistic rules. This meticulous approach yields precise translations, accurate alignment, grammatical annotation, phonetic transcription, and more.

PASC stands out due to features like full alignment, translation accuracy, accurate transcription (especially for non-Latin scripts), multilingual formats, full annotation, and consistent formatting. The project spans domains like personal names, place names, points of interest, and technical terms.