Abstract

Objective

Parallel corpus is the key resource for English Punjabi machine translation. At wide level there is no availability of English-Punjabi corpora. There is a primary requirement of parallel corpus for the training of statistical machine translation.

Methods/Analysis

In this paper, authors focus on building English-Punjabi corpus at large scale. It posed difficulties and the intensive labor to develop the corpus. We are intricate on the collection as well as the flow of work for the construction of parallel corpus. Now after getting the raw text, we need to refine the corpus in such a way that every source language sentence should have corresponding target language sentence.
Findings

The paper attempts to explore existing tools as well as building new tools. One of the goals is alignment of bilingual corpus. The alignment algorithms are used to tune the sentences. The accuracy depends on the type of corpus.

Novelty/Improvement

A cautious endeavor has been made to capture different types of texts.

References

16. www.pseb.ac.in
17. www.christos-c.com/bible
18. www.tdil.mit.gov.in

Index Terms

Computer Science
Information Systems

Keywords

Bilingual corpora, Machine-translation, English, Punjabi, NLP.