Abstract

Parallel text is a fundamental requirement for the development of corpus based or data driven machine translation systems and other Natural Language Processing applications. The unavailability of this valuable linguistic resource has greatly hampered the development of NLP applications in English and Igala language. This study is aimed at creating English – Igala parallel text. The result of the study in addition to providing linguistic resource will enhance language learning. Various algorithms for automatic construction of parallel text such as STRAND, PTMiner, PTI, WPDE, BITS were studied to determine their appropriateness in creating English – Igala parallel text. Wikipedia and the Bible which are excellent sources of parallel or comparable corpora were also gleaned. Existing algorithms and other sources of Parallel text were found to be unsuitable for the construction of English – Igala parallel text due to the unavailability of contents rendered in Igala language on the web and in electronic form. A combination of manual and machine assisted translation was used to generate the parallel text. English – Igala parallel corpora comprising of 50,000 aligned sentences was obtained.
References

Linguistics, 29(3)


Index Terms

Computer Science Information Sciences

Keywords

Parallel text, Natural Language Processing, Machine Translation, comparable corpora, corpus based or data driven machine translation systems, linguistic resource