Parallel Corpus Creation

Pavan Mandava, Isaac Riley, Jasvinder Singh

January 26, 2020

Overview

1. Motivation

2. Project overview
    - Collect
    - Prepare
    - Access

3. Summary

What is a parallel corpus?

A corpus containing

Uses of parallel corpora

- machine translation
- translation analysis & studies
- comparative linguistics
- digital humanities
- technology-assisted language learning

Examples of parallel corpora

- Rosetta Stone
- OPUS (collection - incudes OpenSubtitles, TED Talks, etc.)
- EuroParl Corpus
- EUR-Lex Corpus
- Tatoeba

Our contribution

There is a need for

Project overview

Collect

Prepare

¡ difficulties with data collection; reasons for doing it by hand so far ¿

Bitext alignment in general

Alignment algorithm

Example

Data format and query types

Example