Web as a Corpus: creation and annotation of very large corpora

André Trevisol Trindade

What's a WaC

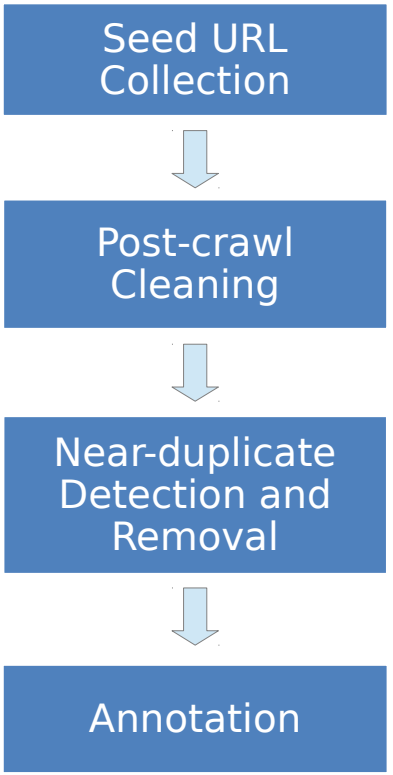
- Short for "Web as a Corpus"
- Apporach for using the web as basis for building very large corpora.
- These initiatives enables the creation of corpora for languages that have scarce freely distributed resources.

Uses for brWaC

Large corpora are necessary for

- Spelling Correction
- Case Frame Acquisition
- Information Retrieval
- Machine Translation
- Distributional Thesaurus Construction
- Subcategorization Frame Acquisition
- Identification of Multiword Expressions

Construction



As seeds to create the URL collection were used medium frequency words.

Tokenized, lemmatised, part-of-speech information.

Future Works

Future Works

- Format annotation in CoNLL model.
- Distribute the corpus in a easily accessible way.