Applying the methodology WaCky in constructing Portuguese corpus

Rodrigo Boos - Federal University of Rio Grande do Sul

Motivation

-Scarce resources on corpora available in Portuguese;

-Useful for

- corpus analysis
- dictionary and thesaurus construction
- identifying Multiword Expressions (MWEs). Ex.: construção de programas, engenharia de software

Related Work

-Web as a Corpus article;
-Boilerplate removal;

Construction

1) List of content words;
2) Get links from web searches on these words;
3) Craw links;
4) Remove non content elements.

List of Content Words

-List of words from Linguateca;
-Remove high and low frequency words;
-Remove stopwords;
-Generate random pairs:
    -assiste toneladas;
    -psicologia criciúma;
    -panos assuma;

Get links and crawling

-Using web search API, get 10 seed URL's for each pair;
-Crawl with seed URL's -> removal

Removal

-First, check size of file (5kb to 200kb);
-Boilerplate:
- Boilerpipe;
- Stopwords density;

-Duplicates:
- N-grams comparison

What we have done

-2Gb corpus;
-52 million words;
-875000 types;

Ceten Folha:
-24 million words;
-343000 types.

Next Steps

Process Corpus

- PALAVRAS parser
- Format output as dependency parsing table

Evaluate quality in application

- MWE identification

Doubts

References

M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3): 209-226.

Christian Kohlschütter, Peter Fankhauser and Wolfgang Nejdl,Boilerplate Detection using Shallow Text Features,WSDM 2010 -- The Third ACM International Conference on Web Search and Data Mining New York City, NY USA.

Pomikalek, Phd (2011). Removing Boilerplate and DuplicateContent from Web Corpora. Ph.D. Thesis. Masaryk University: Czech Republic.