Wikicorpus: A Trilingual Corpus

- Trilingual corpus of Catalan, English, and Spanish
- Contains over 750 million words
- Lemma and POS tagged; Sense annotated
- Licensed under the GNU Free Documentation License
- Available for download (both raw and tagged version)
- Possibly the largest automatically sense-tagged corpus
- Produced by Universitat Politècnica De Catalunya. Talp Research Center

Creation

- Used a large portion of a 2006 dump of Wikipedia

 - About 10-15% not used in the English and Spanish corpus
  - Filtering: Not all articles were suitable (exclusion of articles without a category)
  - Parser errors

- Catalan articles not affected by parser errors as they were usually shorter

-Encyclopedic text

Annotation

- Lemma and POS tags were done using FreeLing, an open source library

- Sense annotated using Word Sense Disambiguation (WSD) algorithm, UKB

- However, the current corpus has no alignment between the three languages, so more work could be done there

Usage

- Using Wikicorpus and NLTK to build a Spanish POS tagger
  - Tom De Smedt (Computational Linguistics Research Group, University of Antwerp

- Two byproducts of compiling the corpus as well
  - An open source Java-based parser for Wikipedia
  - Integration of the WSD Algorithm into FreeLing

"The Wikipedia corpora allow us to test different methods for the automatic acquisition of semantic knowledge (a) at a large scale, (b) on a multilingual dimension."

References

Reese, S., Boleda, G., Cuadros, M., Padró, L., & Rigau, German. (2010). *Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus*. Paper presented at the 7th Language Resources and Evaluation Conference, LaValleta, Malta. (http://www.cs.upc.edu/~nlp/papers/reese10.pdf)

De Smedt, T. (n.d.). *Using Wikicorpus & NLTK to build a Spanish part-of-speech tagger*. Retrieved from http://www.clips.ua.ac.be/pages/using-wikicorpus-nltk-to-build-a-spanish-part-of-speech-tagger

ISLRN Metadata

Title	Wikicorpus
Full Title	Wikicorpus
Resource Type	Text (Encyclopedic)
Source/URL	http://www.cs.upc.edu/~nlp/wikicorpus/
Format/MIME Type	Text/xml
Size/Duration	More than 750 million words
Access Medium	Free download online

Description The Wikicorpus is a trilingual corpus (Catalan, Spanish, English) that contains large portions of the Wikipedia (based on a 2006 dump) and has been automatically enriched with linguistic information. In its present version, it contains over 750 million words.

The corpora have been annotated with lemma and part of speech information using the open source library FreeLing. Also, they have been sense annotated with the state of the art Word Sense Disambiguation algorithm UKB..

Version	1.0
Media Type	Text
Language	Spanish, Catalan, and English
Resource Creator	Universitat Politècnica De Catalunya. Talp Research Center
Distributor	Universitat Politècnica De Catalunya. Talp Research Center
Rights Holder	Universitat Politècnica De Catalunya. Talp Research Center