Automatic Corpus-based Acquisition of Binary Terms

Last Updated jeudi, 23 octobre 2014

ACABIT is under Licence GPL.
ACABIT is a terminology extraction program which takes as input a linguistic annotated corpus and proposes as output a list of multi-word term (MWT) candidates ranked from the most representative of the corpus to the least using loglike score. For each MWT candidate, a XML structure is provided which gathers all the base structures and the variations encountered.

ACABIT uses the following programs :

-
Perl

-
For French :

-
Brill's POS tagger for French ATILF

-
French lemmatizater FLEMM (WARNING : the output data of FLEMM has been modified. You need to use FLEMM-v2.0 (1999))

-
XML format. This perl script could be used brill2xml.pl

-
For English :

-
Brill's POS BRILL

-
Lemmatiser : lexical database CELEX

Loading

-
Encoding UTF-8 - Perl version >= 5.8 - Acabit_UTF8 - FR_V4.5

-
Encoding UTF-8 - Perl version >= 5.8 - English ACABIT, version v4.3 tgz

-
Japanese ACABIT by Koichi Takeuchi, University of Okayama, Japan JACABIT

-
ACABIT for Malagazy, please contact me

Old versions

- Encoding UTF-8 - Perl version >= 5.8 - French ACABIT, version v4.3 tar.gz

- Encoding ISO-8889-1 - French ACABIT, version tar.gz

-
Encoding ISO-8889-1 - English ACABIT, version tgz

To understand ACABIT, please read some of my publications, for example :
[Daille, B. 2003b]. B. DAILLE, "Conceptual structuring through term variations". In F. Bond, A. Korhonen, D. MacCarthy and A. Villacicencio (eds.), Proceedings ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 9-16, 2003. Version PDF.