Swath Manual

Swath (**S**mart **W**ord **A**nalysis for **TH**ai) is a word segmentation program for Thai. It provides three different algorithms; Longest Matching, Maximal Matching and Bigram Model which are described in [1, 2]. Swath can support utf8 and accept four types of input file which are text format, LaTex, RTF and HTML.

Author: Paisarn Charoenpornsawat
http://www.links.nectec.or.th/~yai

USAGE

swath [mule|-v] [-b "wordseperator"] [-d wordsegdatadir] [-f html|rtf|lat
ex|lambda|winlatext|maclatex] [-m long|max|bi|bip] [-l] [-help] < inputfile > outputfile

Option mule : for mule
Option -v : verbose mode
Option -b : user define a word seperator
Option -d : set a new data path (containing *.tri files)
Option -f : specify a format of an input file
html : html file
rtf : rtf file
latex : laTex file
lambda : An input and output are same as latex but only
word break strings are ^^^^^^^^200c
winlatex : laTex file shaping on Windows
maclatex : laTex file shaping on Macintosh
Option -m : choose an algorithm of word segmentation
long : longest matching algorithm
max : maximal matching algorithm
bi : bigram algortihm without part-of-speech tag
bip : bigram algortihm with part-of-speech tag (described in [3])
Option -l : line processing(effect only in a bigram algo.)
Option -help: Help

Example

To display help.
swath –help
To input file an output file (inputfile.txt and outputfile are an input file and output file, respectively)
swath < inputfile.txt > outputfile.txt
To use bigram alogirthm
swath –m bi < inputfile.txt > outputfile.txt
To use bigram alogirthm and also output Part-of-Speech Tags.
swath –m bip < inputfile.txt > outputfile.txt

Reference

[1] Paisarn Charoenpornsawat. 1999. Feature-based Thai Word Segmentation. Master's Thesis. Computer Engineering. Chulalongkorn University, Bangkok, Thailand. (in Thai)
[2] Surapant Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijsirikul, 1997. Feature-based Thai Word Segmentation. In Proceedings of the Natural Language Processing Pacific Rim Symposium 1997(NLPRS'97), Phuket, Thailand.
[3] Virach Sornlertlamvanich, Thatsanee Charoenporn and Hitoshi Isahara. *ORCHID: Thai Part-Of-Speech Tagged Corpus.* Technical Report Orchid TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand, pp. 5-19, Dec 1997.