Utilisation des CRF pour construire un segmenteur

Apprentissage

L'entrainement utilisé est celui donné en exemple dans pour l'API sanford nlp dans le ficher trainsegmenter

Utilisation

Les commandes suivantes font appel à l'API de Stanfrd pour les CRF décrite dans "A Conditional Random Field Word Segmenter." Fourth SIGHAN Workshop on Chinese Language Processing. 2005

Il est possible de teser le segmenteur avec deux modèles différents pour l'apprentissage: Chinese Penn Treebank standard (ctb) et le Peking University standard (pku).

Il est possible d'ajouter l'option -kBest <int> pour préciser les k meilleurs segmentations.

commande pour utiliser le segmenteur basé sur l'apprentissage avec le ctb:

java -mx2g -cp seg.jar edu.stanford.nlp.ie.crf.CRFClassifier -sighanCorporaDict data -testFile test1.utf8 -inputEncoding UTF-8 -sighanPostProcessing true -keepAllWhitespaces true -loadClassifier data/ctb.gz -serDictionary data/dict-chris6.ser.gz

commande pour pour utiliser le segmenteur basé sur l'apprentissage avec le pku:

java -mx2g -cp seg.jar edu.stanford.nlp.ie.crf.CRFClassifier -sighanCorporaDict data -testFile test1.utf8 -inputEncoding UTF-8 -sighanPostProcessing true -keepAllWhitespaces true -loadClassifier data/ctb.gz -serDictionary data/dict-chris6.ser.gz