Adaptation d'un modèle CRF à un corpus différent pour étiquetage morpho-syntaxique et l'extraction d'information

Directeur de thèse : Isabelle TELLIER, Marco DINARELLI
Doctorante : Tian TIAN

13 octobre 2014

Plan

- Cadre de travail
- Analyse de tâche
- Méthodes possibles
- Les CRF
- Références

- Thèse en CIFRE
- Synthésio : e-réputation
- Tâches :
  - Tagger morpho-syntaxique automatique
  - Extraction d'entités nommées
- Spécificités :
  - Multi-sources : journal, blog, forum, tweet
  - Multi-langues : englais, français, espagnol, chinois, etc

Analyse de tâche

- Tagger morpho-syntaxique
  - Partie d'une chaîne d'analyse syntaxique automatique
  - Construire un étiqueteur morpho-syntaxique
  - Associer automatiquement une étiquette à chaque mot
  - Difficultés
    - tokenisation : notion de mot
      exemples : pomme de terre, 'est-il possible', aujourd'hui, etc
    - désambiguisation : séquences d'étiquettes possibles
      exemples : voile (V,N), souris (N,V)
- Extraction d'entités nommées
  - Types d'entités nommées : noms de personnes, lieux, organisations, fonctions, entités cibles
  - Désambiguisation : J'adore (Dior)

Méthodes possibles

- méthodes par règles
  - ensemble de règles composées par les experts linguistiques
  - nécessite des différentes règles pour d'autres langues
- méthodes d'apprentissage automatique
  - supersivé VS non-supervisé
  - nécessite des exemples de traitement (corpus annoté)
  - relativement facile à adapter à une autre langue

Les CRF

- méthode d'apprentissage automatique supervisé (statistique)
- meilleur méthode séquencielle (précision/rappel et temps d'apprentissage)
- phase d'apprentissage
  - Entrée : exemples annotés et patron
  - Sortie : modèle (les fonctions caractéristiques et leurs poids)
- phase de test
  - Entrée : textes à tagger
  - Sortie : textes taggés

Référence


Anne Abeillé, Lionetl Clément, and François Toussenel.
Treebanks : Building and Using Parsed Corpora, chapter Building a Treebank for French, pages 165–188.
Abeillé ed. edition, 2003.


Yoann Dupont Iris Eshkol, Isabelle Tellier and Ilaine Wang.
Peut-on bien chunker avec de mauvaises étiquettes pos ?
TALN, 2014.


A. Cornuéjols & L. Miclet.
Apprentissage artificiel, 2ème édition.
Eyrolles, 2010.


Y. Tsuruoka, J. Tsujii, and S. Ananiadou.
Fast full parsing by linear-chain conditional random fields.
In *Proceedings of EACL 2009*, pages 790–798, 2009.