Rapport de stage du 12/04/2010 au 09/07/2010

Apprentissage automatique d'un étiqueteur du français par analyse discriminante

Lieu : Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Bâtiment IIIA - 6, Rue Léonard de Vinci
B.P. 6759 - 45067 ORLEANS Cedex 2

Yoann DUPONT
Licence 3 informatique

Maître de stage : Isabelle TELLIER
Fonction : professeur des universités

Remerciements

Je remercie tout d'abord M. Jérôme DURAND-LOSE, Directeur du LIFO, de m'avoir permis d'effectuer mon stage au sein de son laboratoire. Je remercie également toute l'équipe « contraintes et apprentissage » de m'avoir accueilli et de m'avoir accordé leur confiance.

Plus particulièrement, je remercie Mesdames Isabelle TELLIER et Sylvie BILLOT, Messieurs Denys DUCHIER, Jean-Philippe PROST et Samer TAALAB pour leur disponibilité et leurs explications.

Je tiens à remercier également Messieurs Thomas LAVERGNE et François YVON de m'avoir accordé l'utilisation de leur apprentisseur Wapiti en alternative à CRF++.

Je remercie toutes les personnes que je n'ai pas citées mais qui ne sont pas oubliées pour autant.

Sommaire

Remerciements..... 2

Sommaire..... 3

Lexique..... 4

Introduction..... 5

Contexte du stage..... 6

Lexique

Analyse discriminante / Conditional Random Field (CRF) : technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis d'un ensemble d'observations à partir d'une série de variables. (source : Wikipedia)

Maximum d'entropie : *besoin d'aide*

Gold standard / étalon d'or : fichier de référence supposément parfaitement annoté, il est employé dans l'apprentissage automatique pour que le programme apprenne à classer les éléments et/ou à évaluer la qualité de la sortie.

Apprentissage automatique (supervisé) : si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement. (source : Wikipedia)

Introduction

Ce stage, d'une durée de trois mois, a consisté à entraîner un apprentisseur afin d'obtenir un étiqueteur de la langue française. Ce genre de programme a pour but de reconnaître la fonction de chaque mot dans une phrase.

Ce rapport présente le travail que j'ai effectué pendant trois mois au LIFO. L'objectif principal de ce projet était d'obtenir le meilleur étiqueteur possible. Le second objectif a été d'évaluer la "puissance" du modèle des CRF par rapport au maximum d'entropie. Je me suis familiarisé avec le langage "python" et j'ai exploité des données réelles à traiter afin de les rendre utilisables.

Après une rapide présentation du contexte de mon stage, j'évoque en détail le corpus et le système d'apprentissage, puis j'explique les différentes étapes à réaliser et comment en améliorer le résultat.

Contexte du stage

Présentation du LIFO, cf site. Donner l'activité principale, les différentes équipes

Mon travail se composait de deux parties principales : la récupération des données d'un gold standard dans le but de les mettre dans un format utilisable par l'apprentisseur (cf annexe 1), puis entraîner ces exemples en utilisant des programmes recourant aux CRF. Je devais trouver les informations utiles à extraire des mots et les utiliser au mieux. J'ai ensuite eu à utiliser une ressource extérieure (le dictionnaire lefff) pour évaluer son impact sur l'apprentissage.

Conclusion

Ce stage s'est avéré intéressant en premier lieu par le fait d'apprendre un langage de programmation extérieur à ma formation, et en second lieu qu'il soit lié à ma langue natale. J'en ai d'autant plus apprécié ce stage. Il m'a aussi donné un premier regard sur l'intelligence artificielle en me servant de programmes utilisant l'apprentissage automatique.

Annexe

I.

En_tout_cas/ADV est/V
-il/CLS plus/ADV
nuancé/ADJ ./PONCT

En ADV
tout _ADV
cas _ADV
est V
-il CLS
plus ADV
nuancé ADJ
. PONCT