Découverte automatique de règles pour l'annotation de corpus

Sujet de stage master recherche informatique - 2007/08

Encadrants : T. Charnois, B. Crémilleux
Lieu : GREYC - CNRS - UMR 6072

Contexte

L'objectif général de ce projet est de rendre complémentaire les techniques du traitement automatique des langues (TAL) et de fouille de textes pour améliorer la recherche ou l'extraction d'information, et la découverte de connaissances.

Au niveau du TAL, les méthodes d'extraction d'information visent à identifier un type d'information ciblé : par exemple en biologie médicale de nombreux travaux ont pour objet de repérer dans des corpus des noms de gènes, de protéines, de fonctions biologiques, etc. Les méthodes utilisées sont essentiellement basées sur des règles élaborées manuellement, et traitent de corpus spécifiques. Appliquer les méthodes d'EI à des textes variés et de grande taille suppose d'adapter les règles ; l'acquisition automatique des règles par des méthodes d'apprentissage automatique et de fouille de textes est une voie prometteuse pour prendre en compte la diversité des corpus et leur évolution. C'est l'un des objectifs du stage. Au niveau de la fouille de textes, il s'agit de tester sur des applications réelles la faisabilité de méthodes usuelles et, à moyen terme, de concevoir et développer des méthodes de fouilles adaptées aux données séquentielles et intégrant les spécificités du texte.

Ce travail s'appuie sur le projet "Annodis" (Annotation discursive) soutenue par l'ANR et qui rassemble trois laboratoires.

Ce stage peut être poursuivi en thèse.

Cadre de travail

Le but de ce stage est de concevoir et réaliser un ensemble de traitements pour découvrir automatiquement des règles annotant des corpus. Pour cela, le stagiaire s'appuiera sur la démarche présentée ci-dessous (les outils de fouille de textes seront aussi à sa disposition).

Un travail préliminaire d'EI a été réalisée au sein de l'équipe DoDoLa du GREYC et a été appliqué sur des "summaries" (C1), ce sont des synthèses de résumés en biologie médicale. Ces textes étant très spécialisés, ils présentent des régularités linguistes intéressantes qui ont permis de concevoir des règles d'extraction simples sous la forme de "patrons syntaxiques". Ces règles indiquent par exemple comment repérer un nom de protéine : ainsi ce type d'entité peut être précédé d'un verbe comme 'product', 'generate' et d'un détermi-

nant, et suivi par une ponctuation ou un pronom relatif. En repérant ce type de contexte, on va, par hypothèse identifier un nom de gène.

Ces premiers résultat serviront de "boostrap" pour la poursuite du travail qui est suggéré comme suit :

- l'utilisation de ces règles sur le corpus C1 permet de marquer les textes de ce corpus (c'est une annotation XML qui fournit un typage des entités identifiées). On obtient donc immédiatement un ensemble de noms de gènes, de protéines, etc. qui va constituer un dictionnaire d'entités biologiques.
- puis, l'utilisation de ce dictionnaire sur un corpus plus complexe (C2) permet de repérer des entités biologiques. Le corpus "complexe" est donc marqué sans qu'il soit nécessaire de développer ou d'appliquer des règles spécifiques à ce type de corpus. Évidemment, à cette étape, on ne repère que ce qui est déjà connu (c'est-à-dire contenu dans le dictionnaire). L'intérêt est d'aller plus loin en utilisant le corpus C2 ainsi marqué pour générer *automatiquement* des règles d'annotation.
- pour découvrir automatiquement de telles règles :
  - on extrait un "contexte" pour chaque nom d'entité (par exemple, les mots précédents et les mots suivants). On obtient ainsi une base d'exemples où chaque exemple est un contexte.
  - puis, découverte de règles à partir de cette base d'exemples :
    - découverte d'épisodes (i.e., séquences de mots) : sélection et généralisation pour leur utilisation permettant d'annoter un nouveau corpus,
    - intégration de contraintes du TAL (e.g., présence d'un verbe, nombre de mots) dans le processus de fouille.