Vers une nouvelle forme de recherche en biologie

Oriane Matte-Tailliez

oriane.matte@lri.fr

La mise à disposition en ligne quasi complète des articles scientifiques en biologie s'est largement développée grâce aux efforts considérables de la base Medline du NCBI. Aujourd'hui, tout biologiste accède donc à une littérature scientifique de taille considérable, et ne peut plus faire face à une gestion "manuelle". Pourtant cette abondante littérature est un outil de travail formidable, permettant au biologiste, au bio-informaticien de faire le point sur son domaine, et de proposer de nouvelles hypothèses de travail. Il faut donc développer des approches automatiques pour gérer ces masses de publications scientifiques.

Notre travail a pour but d'extraire de l'information à partir des textes de manière semi-automatique. Pour cela, nous construisons une chaîne logicielle qui permet à l'expert d'un domaine de spécialité comme la biologie moléculaire de pouvoir lui-même récupérer les textes natifs et les traiter pour en dégager de l'information. Notre approche est un juste mélange entre deux ingrédients : une grande convivialité des interfaces utilisées et des algorithmes inductifs. Nous avons développé un type d'induction nouveau, l'induction extensionnelle, qui permet une part d'automatisation avec intervention de l'expert. Les principes de cette forme d'induction que nous essayons de mettre en oeuvre à toutes les étapes de traitement sont les suivants : a) la création d'un modèle en extension (description complète de tous les objets à partir d'un noyau d'objets), b) l'utilisation d'une mesure spéciale qui permet de rejeter des propositions fausses, c) l'utilisation de plusieurs mesures d'entropie qui tiennent compte de la représentation de la connaissance de l'expert, d) un apprentissage itératif (avec interaction homme/machine). Pour aboutir à l'extraction d'information, plusieurs étapes de traitement sont nécessaires. Après recueil, les textes sont fondus en un seul, c'est le corpus de base. Ce corpus est ensuite normalisé, c'est à dire que le vocabulaire est standardisé, différents types de fautes sont repérés et corrigés, le corpus est découpé en phrases nettement séparées. Ensuite, tous les mots sont étiquetés grammaticalement, les termes pertinents pour le domaine sont repérés, les traces de concepts sont détectés puis classifiées. Et enfin, des patrons d'extraction (ce sont des "automates finis déterministes") sont construits afin d'extraire un type précis d'information. Cette dernière étape serait impossible sans les traitements préalables. Par notre méthodologie, un concept est représenté par des milliers d'instances dans les textes. Si ce n'était pas le cas, des milliers de patrons seraient nécessaires pour une information, ce qui est impossible à réaliser. Le problème est rendu plus difficile par le fait que ces étapes ne sont pas indépendantes et il est nécessaire de bien comprendre les actions et les rétro-actions des unes sur les autres. L'expert intervient à chaque étape, mais sa charge de travail est allégée par la semi-automatisation. Dans ces conditions, de l'information pertinente peut être extraite de ces grandes bases de données textuelles.