

Plateforme d'annotation terminologique de corpus des résumés d'articles de biologie

Thierry Hamon*

LIPN – UMR CNRS 7030, Université Paris-Nord, Avenue J.B. Clément, 93430 Villetaneuse
thierry.hamon@lipn.univ-paris13.fr, <http://www-lipn.univ-paris13.fr/~ht>

Résumé

Nous décrivons la plateforme que nous avons développée pour annoter des termes dans des corpus d'articles scientifiques du domaine de la génomique. L'architecture de cette plateforme se veut modulaire afin d'exploiter et d'intégrer rapidement des outils de TAL et de terminologie mais aussi des ressources et des corpus

Mots clef : Terminologie, annotation de termes, corpus

1 Introduction

Nous présentons ici l'architecture de la plateforme d'annotation terminologique de corpus développée dans le cadre du projet ExtraPloDocs¹. Ce projet vise notamment à sélectionner des phrases décrivant des interactions entre gènes et protéines à l'aide de techniques d'apprentissage et de connaissances du domaine. Notre objectif est d'identifier les termes du domaine dans les corpus de travail et de les annoter dans la perspective de l'apprentissage de règles de sélection de fragments pertinents.

Les ressources terminologiques étant largement disponibles en génomique (le MeSH² [3] ou Gene Ontology³ [2]), nous avons développé deux axes pour annoter terminologiquement les corpus : (1) projection en corpus de termes issus de ces ressources, (2) l'extraction automatique de candidats termes et leur filtrage. L'architecture de la plateforme que nous présentons reflète ces choix.

2 Architecture de la plateforme

Nous avons conçu la plateforme pour répondre à l'objectif initial du projet : l'annotation terminologique de corpus en vue de la sélection de phrases mentionnant des interactions entre gènes et protéines. Cependant, nous avons également voulu tenir compte d'extensions futures comme la construction de classes sémantiques pour l'élaboration d'ontologie à l'aide d'ASIUM [5]. Ainsi, l'architecture de la plateforme reflète essentiellement nos besoins actuels. Mais, dans la perspective d'intégration de nouveaux traitements, nous l'avons conçue aussi modulaire que possible. Ce choix doit nous permettre d'intégrer rapidement et à moindre coût

une nouvelle ressource terminologique, un nouvel outil ou un nouveau corpus. Les nombreux résultats que peut produire la plateforme nous a également incité à conserver une trace des traitements effectués sur les données.

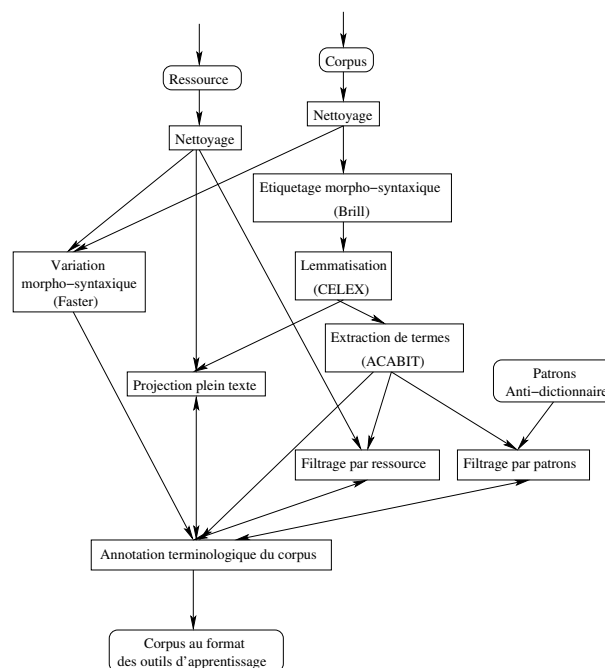


FIG. 1 – Architecture de la plateforme.

La figure 1 présente l'architecture et les modules actuellement implantés. Les flèches indiquent le flux des données et les dépendances entre les modules. Chaque module réalise un type de traitement. Compte tenu de nos besoins actuels, les modules dont les tâches sont similaires ne sont pas utilisables de manière complémentaire, mais uniquement interchangeables (extraction automatique de termes à la place d'une projection plein texte).

L'analyse morpho-syntaxique d'un corpus est réalisée par l'étiqueteur de Brill [1] et la lemmatisation à l'aide de la base CELEX. Ces résultats sont ensuite fournis à l'extracteur de termes ACABIT [4], ou bien en entrée du module que nous avons développé pour réaliser la projection plein texte de termes issus de ressources terminologiques. Faster [8] nous permet d'identifier les variantes morpho-syntaxiques des

1. Projet RNTL 2002 avec le LIPN, l'unité MIG de l'INRA et la société Isoft.

2. Medical Subject Headings. <http://www.nlm.nih.gov/mesh>

3. <http://www.geneontology.org/>

termes issus de ressources terminologiques, ceux-ci étant proposés sous une forme normalisée généralement proche de la forme lemmatisée. Les candidats termes extraits automatiquement peuvent être filtrés dans deux modules distincts à l'aide de ressources terminologiques ou de patrons lexico-syntaxiques et d'un anti-dictionnaire. Les résultats de ces analyses sont ensuite exploités par un module réalisant l'annotation des termes dans le corpus.

Format des données Les informations permettant d'annoter le corpus courant sont stockées dans des formats propres à la plateforme. Les termes sont identifiés en corpus à l'aide de balises de type SGML. L'intégration de ressources ou le traitement d'un nouveau corpus ne nécessite ainsi que le développement d'un module de traduction dans le format interne approprié. Actuellement, nous avons intégré à la plateforme cinq ressources (le MeSH, Gene Ontology, une liste de mots clés utilisés pour indexer la base de séquençage SwissProt⁴, un glossaire de Biochimie et de Biologie Moléculaire⁵, ainsi qu'un glossaire de terminologie de la biologie moléculaire⁶). Nous travaillons sur trois corpus de formats assez proches. Un module spécifique produit le corpus annoté dans un format exploitable par les outils d'apprentissage.

Implémentation Nous avons choisi d'implémenter la plateforme à l'aide de l'utilitaire make. Les fonctionnalités de chaque module consiste en la résolution de cibles définies dans un fichier Makefile. Ainsi, les paramètres définissant le corpus annoté (choix de l'analyseur morpho-syntaxique, origine des termes, etc.) conduisent à un nom de fichier unique. Ce nom correspond à l'exécution d'un ensemble de modules précis et constitue l'historique d'une expérience.

Les modules correspondent à des scripts SHELL et Perl ou à des appels à des outils linguistiques et terminologiques disponibles (Faster, ACABIT, Brill). Ces modules peuvent faire appel à des données stockées dans des formats internes ou des formats d'entrée/sortie propres aux outils linguistiques et terminologiques.

3 Discussion

L'utilisation d'outils linguistiques et terminologiques relativement courants et exploitant un jeu d'étiquettes morpho-syntaxiques unique, nous a mis à l'abri d'un certain nombre de problèmes rencontrés dans d'autres travaux : format de sortie différents, lemmatisation variable d'un outil à un autre [6], découpage des termes différent suivant les outils [7].

Toutefois, lors du développement de la plateforme, nous avons dû faire face au problème de l'enchâssement et de l'in-

clusion de termes qu'ils soient proposés par ACABIT ou par une ressource terminologique. Cet écueil a eu une incidence importante sur le format permettant l'annotation des termes. Le balisage SGML classique n'était plus possible. Il nous a été nécessaire d'associer correctement les balises de début et de fin d'un terme à l'aide d'un identifiant spécifique. L'utilisation de la plateforme nous amène à reconsidérer le choix du balisage SGML qui apparaît coûteux en temps de calcul. Nous envisageons ainsi de redéfinir le format d'annotation. Celui-ci devra décrire les opérations à réaliser ou les informations à associer à chaque élément du corpus initial.

Références

- [1] Eric Brill. Automatic grammar induction and parsing free text : A transformation-based approach. In *Proceedings the 31th meeting of the ACL*, pages 259–265, Columbus, Ohio, June 1993.
- [2] The Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Res.*, 11(8):1425–1433, 2001.
- [3] Roger A. Côté. *Répertoire d'anatomopathologie de la SNOMED internationale*, v3.4. Université de Sherbrooke, Sherbrooke, Québec, 1996.
- [4] Béatrice Daille. Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *T.A.L.*, 36(1-2):101–118, 1995.
- [5] David Faure et Claire Nédellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *Proceedings of the Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications (LREC'98)*, pages 5–12, Granada, 1998.
- [6] Natalia Grabar et Blandine Jeannin. Contribution de différents outils à la construction d'une terminologie pour la recherche d'information. In *Actes de IC'2002*, Rouen, 2002.
- [7] Thierry Hamon. *Variation sémantique en corpus spécialisés : Acquisition de relations de synonymie à partir de ressources lexicales*. Thèse en informatique, Université Paris 13, 2000.
- [8] Christian Jacquemin. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, 1997.

4. <ftp://ftp.expasy.org/databases/swiss-prot/release/keywlist.txt>

5. http://www.portlandpress.com/pcs/books/prod_det.cfm?product=1855780887

6. <http://www.asheducationbook.org/cgi/content/full/2002/1/490>