Acquisition de la théorie ontologique d'un système d'extraction d'information

Alain-Pierre Manine*

*LIPN, Université Paris 13/CNRS UMR7030
99 ave. Jean-Baptiste Clément
F93430 Villetaneuse
manine@lipn.univ-paris13.fr

Résumé. La conception de systèmes d'Extraction d'Information (EI) destinés à extraire les réseaux d'interactions géniques décrits dans la littérature scientifique est un enjeu important. De tels systèmes nécessitent des représentations sophistiquées, s'appuyant sur des ontologies, afin de définir différentes relations biologiques, ainsi que les dépendances récursives qu'elles présentent entre elles. Cependant, l'acquisition de ces dépendances n'est pas possible avec les techniques d'apprentissage automatique actuellement employées en EI, car ces dernières ne gèrent pas la récursivité. Afin de palier ces limitations, nous présentons une application à l'EI de la Programmation Logique Inductive, en mode multi-predicats. Nos expérimentations, effectuées sur un corpus bactérien, conduisent à un rappel global de 67.7% pour une précision de 75.5%.

1 Introduction

La modélisation des interactions géniques présente un considérable intérêt scientifique pour les biologistes ; pourtant, la majeure partie de la connaissance la concernant n'est pas présente dans des banques de données génomiques, mais dans la littérature scientifique. De fait, de nombreux travaux (ex. Craven et Kumlien (1999); Krallinger et al. (2007)) ont été entrepris afin de concevoir des systèmes d'Extraction d'Information (EI) visant à extraire un réseau d'interactions géniques à partir de la bibliographie. Dans la plupart de ces systèmes, des patrons d'extraction permettent l'extraction d'une *unique* relation d'interaction binaire (ex. Saric et al. (2005)). De tels modèles ne rendent pas compte de la complexité des données biologiques, telles que les voies métaboliques. En effet, l'EI nécessite des représentations complexes, fondées sur des ontologies, et impliquant de multiples relations interdépendantes (Berardi et Malerba (2006)), éventuellement récursives.

Afin de modéliser ce type de connaissances, Manine et al. (2008) ont récemment introduit une architecture dans laquelle l'EI est considérée comme une tâche de *population d'ontologie*[1]. Dans ce contexte, la théorie logique de l'ontologie subsume les patrons d'extraction, et le problème de l'apprentissage de patrons devient alors une tâche d'*apprentissage d'ontologie*[2].

[1] Ontology Population
[2] Ontology Learning

Ce cadre soulève le problème de l'acquisition de la théorie logique de l'ontologie. En effet, les méthodes d'apprentissage habituellement mises en jeu en IE ne sont pas adaptées à un contexte ontologique : l'emploi de classifieurs binaires rend impossible la prise en compte de multiples relations ; quant à l'apprentissage multi-classes, plus rarement mis en oeuvre (Craven et Kumlien (1999) ou Rosario et Hearst (2004)), il fait l'hypothèse de l'indépendance des prédicats cibles, et ne peut donc pas gérer la récursivité.

Afin de palier ces limitations, nous présentons dans cet article une application à l'EI d'un cadre riche — quoique relativement peu exploré — de l'apprentissage relationnel : l'apprentissage *multi-prédicats* (Malerba (2003)). Ce paradigme offre la possibilité d'apprendre des théories récursives, telle que la théorie logique d'une ontologie. Cette approche autorise une richesse inductive allant au-delà de celle offerte par les systèmes d'EI existants, en produisant des règles qui combinent les niveaux sémantiques et syntaxiques, et ce d'une façon potentiellement récursive.

Cet article est organisé comme suit. Nous introduisons brièvement la plateforme d'EI que nous avons utilisée dans le paragraphe 2, suivie de notre système d'apprentissage d'ontologie en 3. Nos résultats d'apprentissage sont exposés et commentés dans le paragraphe 4. Enfin, dans le paragraphe 5, nous discutons notre approche et esquissons quelques perspectives.

2 Plateforme d'Extraction d'Information

La plate-forme d'EI que nous avons utilisée est décrite dans Manine et al. (2008). Le processus d'extraction est vu comme une tâche de population d'ontologie. Il consiste à extraire de nouvelles instances de l'ontologie à partir du texte grâce à des modules de traitement automatique de la langue naturelle (TALN). Afin de combler le fossé entre la langue naturelle et les structures ontologiques, une *couche lexicale* est introduite, qui permet la normalisation des sorties des modules de TALN dans le langage de l'ontologie (ex. en définissant une relation syntaxique entre deux instances de l'ontologie). La figure 1 illustre une ontologie simplifiée de la transcription chez les bactéries. L'ontologie traduit, par exemple, que la "transcription" d'un gène (*"et"*) à partir d'un promoteur (*"t_from"*) peut survenir suite à l'action d'une protéine (*"t_by"*). Les lignes en pointillés illustrent la définition déclarative de la couche lexicale.

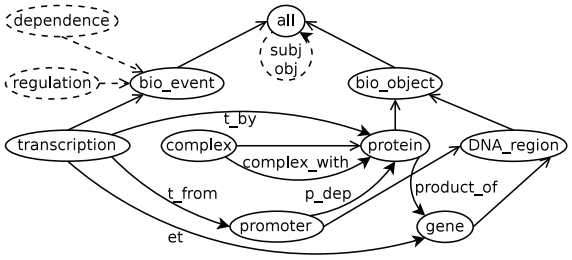

FIG. 1 – *Exemple d'une ontologie (lignes pleines) et de sa couche lexicale (pointillés). Les étiquettes des relations "is-a" sont omises.*

Une sortie du module de population d'ontologie est présentée figure 2 (traits fins). La représentation résultante n'est que faiblement normalisée, et, contrairement à d'autres travaux en EI, ce sont les règles d'inférence de l'ontologie qui, en subsumant les patrons d'extraction, permettent de dériver de nouvelles instances et d'aboutir à une représentation plus normalisée (figure 2, traits gras).

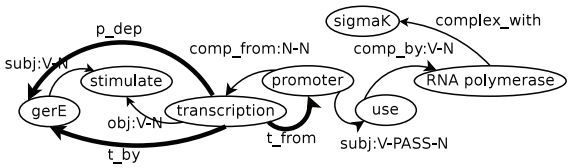

FIG. 2 – *Sortie du module de population d'ontologie (traits fins), et exemples de relations inférées via la théorie logique de l'ontologie (traits gras).*

Dans l'exemple, les relations "p_dep" (promoter_dependance) et "t_by" (transcription_by) (figure 2, traits gras), ont été instanciées en utilisant les règles suivantes :

$p\_dep(B, A) \leftarrow t\_by(C, A), t\_from(C, B), protein(A), promoter(B), transcr(C).$

$t\_by(B, A) \leftarrow subj(A, C), obj(B, C), transcr(B), protein(A), regulation(C).$

La première règle stipule que : "si une protéine A est responsable d'un événement de transcription C à partir d'un promoteur B, alors B est dépendant de ("peut se lier à") la protéine A". Ici, *p_dep(promoter, GerE)* est vrai car *t_by(transcription, GerE)* et *t_from(transcription, promoter)* sont tous deux vrais.

À noter que la deuxième règle exploite la couche lexicale (relations "subj" et "obj") afin d'effectuer un raisonnement à un niveau syntaxico-sémantique.

3 Apprentissage d'Ontologie

Pour acquérir automatiquement la théorie logique de l'ontologie, il est nécessaire de l'apprendre à partir d'un corpus du domaine. Contrairement à Manine et al. (2008), ici, notre système est capable d'apprendre une théorie récursive. L'apprentissage se déroule dans le langage de l'ontologie, et un expert du domaine fournit les exemples d'apprentissage sous forme d'instances de l'ontologie.

Pour apprendre à partir d'un tel langage relationnel, nous avons utilisé le système ATRE Malerba (2003), qui permet l'apprentissage de théories logiques *récursives*. Dans ce système, les exemples sont représentés sous forme de clauses liées dont la tête comporte une conjonction de littéraux, appelés *objets* (nous reportons le lecteur à Malerba (2003) pour une description détaillée d'ATRE). Ici, à chaque phrase correspond un objet, et les exemples négatifs sont générés via l'hypothèse du monde clos.

Un exemple d'objet figure ci-dessous[3] :

[3] Certains exemples négatifs ont été omis

$$t\_by(id2, id1), p\_dep(id4, id1), t\_from(id2, id4),$$
$$\neg t\_by(id1, id2), \neg t\_by(id1, id3), [\ldots] \leftarrow$$
$$subj(id1, id3), obj(id2, id3), comp\_from(id4, id2), transcription(id2),$$
$$protein(id1), regulation(id3), promoter(id4).,$$

À noter que l'ensemble de la connaissance ontologique, comme la relation de généralisation entre les concepts, est fournie à l'algorithme de PLI comme connaissance du domaine.

4 Résultats

L'apprentissage de la théorie de l'ontologie a été validé en réutilisant le corpus employé par Manine et al. (2008). Les relations conceptuelles utilisées sont les suivantes : i (relation d'interaction générique), promoter dependence (p_dep), promoter of (p_of), bind to (b_to), site of (s_of), regulon member (rm), regulon dependence (r_dep), transcription from (t_from), transcription by (t_by), event target (et). Leur sémantique est illustrée par le tableau 1 qui fournit, pour chaque relation, une expression dans laquelle cette dernière est nécessaire pour normaliser le texte.

Name	Example
p_dep	*sigmaA* recognizes **promoter elements**
p_of	the *araE* **promoter**
b_to	**GerE** binds near the sigK *transcriptional start site*
s_of	*-35 sequence* of the **promoter**
rm	*yvyD* is a member of sigmaB **regulon**
r_dep	*sigmaB* **regulon**
t_from	**transcription** from the Spo0A-dependent *promoter*
t_by	**transcription** by final *sigma(A)-RNA polymerase*
et	**expression** of *yvyD*
i	**KinC** was responsible for Spo0A~P *production*

TAB. 1 – *Liste des relations définies dans l'ontologie, et exemples d'expression (les sous-termes de la relation sont représentés en italique et en gras).*

Nous avons évalué le rappel et la précision du processus d'EI à l'aide d'une 10-validation croisée. Les résultats sont présentés dans le tableau 2. Nos performances sont satisfaisantes, et confirment la pertinence de notre approche. Les relations les plus spécifiques (et, t_from, r_dep), qui présentent une faible variabilité lexicale, atteignent des scores élevés ; au contraire, les relations plus génériques, telle que *i*, font montre d'une plus grande variabilité, et sont plus difficiles à apprendre. Les scores faibles de *rm* sont probablement dûs à une mauvaise distribution de cette relation parmi les "objets" d'ATRE.

Dans la suite de cette section, nous allons présenter les avantages de l'apprentissage multi-prédicats au travers quelques exemples de règles apprises par notre système.

Avant tout, certaines règles permettent de raisonner à un niveau purement sémantique :

$$i(X2, X1) \leftarrow t\_by(X2, X3), et(X3, X1). \quad (1)$$

Ainsi, (1) exprime que si X1 est transcrit par X2, alors X1 et X2 interagissent (ex. "gspA" et "sigma B" dans "transcription of gspA is sigma B dependent").

Relation	Recall (%)	Prec. (%)	Number
i	50.2	70.6	225
rm	33.3	41.7	15
r_dep	100.0	100.0	12
b_to	69.6	75.3	79
p_dep	69.8	71.2	53
s_of	61.2	61.2	67
p_of	69.8	55.6	43
et	95.7	96.9	164
t_from	73.3	84.6	15
t_by	52.6	62.5	38
Global	67.7	75.5	711

TAB. 2 – *Résultats pour l'apprentissage multi-prédicats. La dernière colonne présente le nombre d'exemples.*

D'autre part, le multi-prédicats est particulièrement bien adapté aux structures hiérarchiques des ontologies. Ainsi, la règle (2) encode une relation *is-a* entre p_of et s_of, tandis que la règle (3) permet la spécialisation d'une relation s_of en une relation p_of, si X2 est un promoteur et X1, un gène. À noter que les règles (2) et (3) constituent une théorie récursive.

$$s\_of(X2, X1) \leftarrow p\_of(X2, X1). \quad (2)$$

$$p\_of(X2, X1) \leftarrow s\_of(X2, X1), gene\_entity(X1), promoter(X2). \quad (3)$$

Les règles précédentes s'appuient sur d'autres règles (ex. règle (4)), dont certains littéraux sont définis dans la couche lexicale (attributs syntaxico-sémantiques). À défaut d'apprentissage multi-prédicats, seul ce type de règles peut être appris par Manine et al. (2008).

$$i(X2, X1) \leftarrow subj\_v\_n(X3, X1), obj\_v\_n(X3, X2), term(X3, require). \quad (4)$$

Notre approche possède la capacité de combiner niveaux syntaxiques et sémantiques pour inférer de nouvelles relations. Ainsi, la règle récursive (5) combine l'usage de relations sémantiques (b_to, s_of) et syntaxiques (comp_n_n) afin de déduire que X2 se lie à X1.

$$b\_to(X2, X1) \leftarrow b\_to(X2, X3), s\_of(X3, X4), comp\_n\_n\_of(X4, X1). \quad (5)$$

D'autre part, raisonner selon de multiples niveaux d'abstraction autorise la factorisation d'un ensemble de variations lexicales au sein d'une unique étiquette sémantique. Il en résulte de la part de l'algorithme d'apprentissage des théories plus compactes. Cela est illustré par la règle (6). Cette dernière permet de capturer, d'une part, des expressions telles que "the cwlB operon is transcribed by E sigma D" ou, d'autre part, d'autres formules telles que "transcription of cotD by sigmaK RNA polymerase". En effet, les deux expressions "transcription of A" et "A is transcribed" sont factorisées par les règles (7) et (8). L'apprentissage multi-classes aurait nécessité deux règles en lieu et place de la seule règle (6).

$$i(X2, X1) \leftarrow comp\_n\_n\_by(X3, X2), et(X3, X1). \quad (6)$$

$$et(X2, X1) \leftarrow comp\_n\_n\_of(X2, X1), event(X2). \quad (7)$$

$$et(X2, X1) \leftarrow subj\_v\_pass\_n(X2, X1), transcription(X2). \quad (8)$$

5 Conclusion et perspectives

Les applications d'EI émergentes, telles que celles liées au domaine biomédical, nécessitent des représentations complexes du texte, fondées sur des ontologies, et autorisant la dé-

finition de plusieurs relations, ainsi que des dépendances (éventuellement récursives) qu'elles présentent entre elles. Dans cet article, nous avons exploité ATRE, un algorithme de PLI en mode multi-prédicats permettant l'apprentissage de théories récursives, afin d'apprendre la théorie logique d'une ontologie. Cela nous a conduit à la conception d'une plateforme d'EI dans laquelle la théorie logique de l'ontologie subsume les patrons d'extraction. Notre système est ainsi capable de combiner récursivement les niveaux syntaxiques et sémantiques, ce qui lui confère des capacités d'inférence allant au-delà des systèmes d'EI existants.
Dans le futur, nous prévoyons de comparer plusieurs représentations du texte en élaborant différentes couches lexicales. Nous projetons également d'étudier la gestion du bruit d'ATRE, le bruit étant d'une grande importance dans le contexte du TALN.

L'auteur remercie Erick Alphonse et Philippe Bessières pour leur contribution à ce travail.

Références

M. Berardi, and D. Malerba. Learning recursive patterns for biomedical information extraction. In S. Muggleton, R. P. Otero, and A. Tamaddoni-Nezhad, editors, *ILP*, volume 4455 of *Lecture Notes in Computer Science*, pages 79–93. Springer, 2006.

M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proc. 7th Intl. Conf. Intelligent Systems for Molecular Biology*, pages 77–86. AAAI Press, 1999.

Krallinger, M., F. Leitner, et A. Valencia (2007). Assessment of the second BioCreAtIvE PPI task : Automatic extraction of protein-protein interactions. In *Proceedings of the Second BioCreAtIvE Challenge Evaluation Workshop*, pp. 41–54.

D. Malerba. Learning recursive theories in the normal ILP setting. *Fundamenta Informaticae*, 57(1) :39–77, 2003.

A.-P. Manine, E. Alphonse, and P. Bessières. Information extraction as an ontology population task and its application to genic interactions. In *20th IEEE Intl. Conf. Tools with Artificial Intelligence, ICTAI 2008.*, 2008.

B. Rosario and M. A. Hearst. Classifying semantic relations in bioscience texts. In *ACL'04 : Proc. 42nd Ann. Meet. Association for Computational Linguistics*, page 430, 2004. Association for Computational Linguistics.

J. Saric, L. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Large-scale extraction of protein/gene relations for model organisms. In *1st Intl. Symp. Semantic Mining in Biomedicine*, 2005.

Summary

Numerous works aim at designing Information Extraction (IE) systems able to extract genic interaction networks from text. IE systems need sophisticated representations, encoded with ontologies, allowing the definition of multiple relations, recursively dependent. ML techniques usually involved in IE are unfitted, as they do not handle recursion. In this paper, we use ILP in a multi-predicate setting to overcome these limitations. We experimented our OL framework on a bacteria corpus, in which we reach a global recall of 67.7% and a precision of 75.5%.