Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques[1]

Philippe Bessières[1] —Adeline Nazarenko[2] — Claire Nédellec[3]

[1] Mathématique, Informatique et Génome (MIG) INRA,
Route de St-Cyr
78026 Versailles cedex
philb@biotec.jouy.inra.fr

[2]Laboratoire d'Informatique de Paris-Nord,
UPRESA 7030 CNRS
Université Paris Nord
93430 Villetaneuse
nazarenko@lipn.univ-paris13.fr

[3]Equipe Inférence et Apprentissage
LRI UMR 8623 CNRS
Université Paris-Sud,
91405 Orsay cedex
cn@lri.fr

1. Introduction

L'extraction d'information (EI) consiste à remplir automatiquement des formulaires ou une banque de données à partir de textes écrits en langue naturelle. Elle s'oppose classiquement à la recherche documentaire (ou recherche d'information RI) qui vise à retrouver dans une base de documents un ensemble de documents pertinents au regard d'une question. L'extraction met en œuvre une analyse du texte pour interpréter et construire une représentation formelle qui permettra d'apporter automatiquement des réponses précises à l'utilisateur. Il ne s'agit donc pas simplement de sélectionner un fragment brut du texte, mais de mettre des éléments en relation pour restituer une information complète et structurée. Sauf dans les cas très simples, c'est une tâche difficile qui requiert une part de compréhension et nécessite des connaissances, des ressources lexicales, sémantiques et conceptuelles adaptées aux documents et au domaine à traiter. L'acquisition automatique de ces ressources à partir de corpus de documents écrits en langage naturel constitue aujourd'hui un important défi pour l'automatisation de l'extraction d'Information.

Nous défendons l'idée que ces ressources peuvent être apprises automatiquement, et plus précisément par des méthodes d'apprentissage relationnel coopératif [Scott & Matwin, 99], [Cardie *et al.*, 2000]. De nombreuses tâches d'apprentissage dans ces domaines nécessitent en effet que des données et connaissances structurées puissent être représentées, que des connaissances préalables puissent être exploitées, quand elles sont disponibles, et que l'utilisateur du système puisse contrôler interactivement le processus d'apprentissage et apporter de nouvelles connaissances durant son exécution. Cette approche répond à une demande croissante des utilisateurs potentiels des méthodes de RI et de EI, en particulier des biologistes. Elles constituent en effet un moyen d'automatiser la recherche et l'extraction d'information dans les masses considérables de documents scientifiques dans lesquels sont représentées des informations utiles, mais difficiles d'accès [Craven & Kumlien, 99]. Grâce au développement d'Internet, ces documents sont en effet mis en commun dans de vastes bases de données spécialisées telles que MedLine, interrogeables via le web par des requêtes à base de mots-clef.

La section suivante de cet article situe le contexte de notre travail. Nous présentons ensuite les grandes lignes de notre approche (section 3) avant de montrer comment les ressources lexicales et sémantiques nécessaires à l'extraction peuvent être apprises (section 4). La dernière section présente les perspectives de ce travail.

2. Contexte

2.1 Identifier les interactions géniques, un problème d'extraction d'information en biologie

En biologie, nous avons choisi de nous attaquer à un problème particulier de la génomique fonctionnelle, celui de la modélisation des interactions géniques à partir de textes, problème décrit aussi dans [Blaschke *et al.*, 1999], [Pillet, 2000], et [Thomas *et al.*, 2000] et [Poibeau, 2001]. C'est un problème à la fois réaliste et suffisamment complexe pour permettre d'évaluer l'intérêt et la faisabilité de l'application de l'apprentissage à l'extraction d'information en biologie.

[1] Article accepté pour la conférence CIDE'2001 (Conférence Internationale sur le Document Electronique), Toulouse, octobre 2001.

La modélisation des interactions géniques présente un intérêt scientifique considérable pour les biologistes car elle constitue une étape fondamentale dans la compréhension du fonctionnement cellulaire. Aujourd'hui, la majeure partie de la connaissance biologique sur les interactions n'est pas décrite dans des banques de données mais uniquement sous la forme d'articles scientifiques. L'exploitation de ces articles est donc un enjeu central dans la construction des modèles d'interaction entre gènes. Les projets de génomique ont en effet généré de nouvelles approches expérimentales telles que les puces à ADN, à l'échelle globale de l'organisme étudié, et aujourd'hui, une équipe de recherche est capable de produire très vite des dizaines de milliers de mesures. Ce contexte très nouveau pour les biologistes impose un recours à l'extraction automatique de connaissances textuelles : pour être capable d'interpréter et de donner un sens à ces données élémentaires du laboratoire, il faut les relier à la littérature scientifique.

2.2 Une extraction d'information guidée par les connaissances

Dans ce cadre, si la recherche documentaire à l'aide de mots-clefs offre des performances intéressantes en termes de rapidité de traitement, ses résultats ne sont pas directement exploitables et nécessitent un travail d'analyse considérable des documents sélectionnés pour extraire l'information pertinente. L'homogénéité du domaine considéré permet de mettre en œuvre des techniques de recherche d'information plus sophistiquées que pour dans le cadre de recherches généralistes sur le web et sa spécialisation rend le besoin d'information précise plus pressant.

Les approches d'extraction automatiques appliquées jusque-là sont basées essentiellement sur des comptages statistiques de co-occurrences de mots-clefs [Stapley & Benoit, 2000], [Pillet, 2000] ou sur des règles ou automates d'extraction définis manuellement, à base de verbes significatifs et de noms de gènes [Blaschke *et al.*, 1999], [Thomas *et al.*, 2000], [Poibeau, 2001]. Les résultats obtenus présentent, soit une précision très faible, soit une couverture limitée. L'extraction automatique de connaissances pertinentes dans les documents sélectionnés nécessite donc la mise en œuvre de méthodes d'extraction d'information plus complexes qui s'appuient sur des ressources spécifiques au domaine étudié, de types lexical, syntaxique et sémantique comme des ontologies. Ces ressources spécialisées sont généralement difficiles et longues à acquérir manuellement [Riloff, 93], [Soderland, 99], [Nédellec, 2000]. L'aspect novateur de notre approche réside dans la définition et dans l'implémentation de techniques informatiques originales permettant leur acquisition automatique ou semi-automatique à partir de corpus textuels.

Le domaine de la génomique fonctionnelle permet de mettre en œuvre cette approche. Les articles sont écrits dans une langue de *spécialité*, correcte d'un point de vue grammatical et dont le vocabulaire est relativement limité. Les informations recherchées sont locales (exprimées sur quelques lignes au plus) et les critères de pertinence des biologistes sont précis. Comme l'a mis en effet en évidence Harris [Harris *et al.* 89] en immunologie, la variabilité des sous-langages utilisés dans les domaines de recherche spécifiques est limitée à la fois du point de vue du vocabulaire, de la polysémie, des formes syntaxiques et du nombre de concepts représentés. Dans ces conditions, il est réaliste de vouloir acquérir automatiquement les ressources lexicales, sémantiques et conceptuelles nécessaires à une analyse profonde, ceci à partir des régularités observées dans un corpus [Staab *et al.*, 2000]. Ensuite, grâce à ces ressources, par l'intermédiaire de transformations, on peut ramener toute phrase d'un corpus de spécialité à une forme canonique qui représente son contenu informationnel en termes d'opérateurs (ou prédicats) et d'arguments.

La génomique fonctionnelle constitue un domaine réel d'application qui est loin d'être un problème jouet. À partir de là de très nombreuses applications biologiques sont ouvertes sous la condition que la tâche d'extraction soit clairement spécifiée et que des exemples des informations recherchées puissent être identifiés sans ambiguïté par les biologistes. L'identification claire du besoin permettra également de valider précisément les résultats obtenus. L'existence de nombreux domaines scientifiques et techniques présentant de forts points communs d'un point de vue documentaire avec celui de la génomique fonctionnelle permettra d'adapter hors de ce champ, les méthodes développées. C'est typiquement le cas d'un domaine connexe comme la protéomique, mais plus généralement nos méthodes seront exploitables dans l'ensemble du contexte de l'extraction de connaissance à partir de documentation scientifique et technique.

2.3 Une approche multidisciplinaire

Dans le cadre proposé, une approche multidisciplinaire est nécessaire qui fait collaborer étroitement des biologistes avec des informaticiens respectivement spécialistes du traitement automatique de la langue, de l'extraction d'information et de l'apprentissage automatique. Le travail décrit ici s'inscrit dans le cadre du projet bio-informatique *Caderige*. La spécification des besoins biologiques et la validation des outils et des

résultats d'apprentissage s'effectuent principalement en collaboration avec le laboratoire MIG[2] de l'INRA par l'étude de notices de MedLine portant sur la transcription des gènes chez la bactérie modèle *Bacillus subtilis* [Haldenwang, 95], [Wosten *et al.*, 98]. Les nombreuses données déjà connues et disponibles sur la transcription des gènes chez cette bactérie modèle vont permettre dans un premier temps de juger de la pertinence de l'extraction de l'information, mais il s'agira de s'intéresser rapidement à l'ensemble des interactions et des régulations moléculaires. À ce stade de notre recherche, nous nous assurons de la généralité de l'approche en la confrontant à la modélisation des interactions géniques chez d'autres espèces : la souris et l'homme, (avec la société Valigen) d'une part, et la drosophile d'autre part (en collaboration avec le LGPD, Université de Marseille).

3. Notre approche

3.1 Le problème à résoudre

L'exploration des bases documentaires telles que MedLine repose sur des requêtes portant sur un ensemble de termes connectés à l'aide d'opérateurs logiques. Ce type d'accès permet au biologiste de ramener un sur-ensemble des documents effectivement pertinents, de l'ordre de quelques centaines ou milliers. Par exemple, la requête "*Bacillus subtilis transcription*" sélectionne quelque 2200 résumés. Il reste ensuite à en extraire toutes les connaissances utiles relatives aux interactions entre les gènes et à les enregistrer de manière structurée de telle sorte qu'un biologiste puisse obtenir des réponses à une requête spécifique sous une forme logique ou libre, par exemple, "`Quel facteur sigma contrôle l'expression du gène dacB ?`"[3].
Aujourd'hui l'extraction s'effectue soit à la main, soit automatiquement sur la base d'indices superficiels (techniques issues de l'extraction d'information) ou de l'indexation par mots-clefs (techniques issues de la recherche documentaire et de l'analyse des données). Ces techniques ne peuvent obtenir que des résultats limités dans le cadre de la recherche d'interactions pour de multiples raisons dont voici quelques exemples :

- Les noms des gènes ne sont pas toujours repérables sans ambiguïté. Ainsi, chez certains organismes tels *Drosophila*, des noms de gènes et protéines ont des homonymes dans la langue courante, comme *Red* ou *Giant*. D'autres ne sont pas répertoriés dans une nomenclature stable ou possèdent des variantes non inventoriées (*Sigma E*, *Sig(E)*, *Sigma-E*, etc.). Ce problème est étudié depuis peu dans des travaux comme [Fukuda, 98], [Proux *et al.* 98], [Collier *et al.*, 2000] et [Humphreys *et al.*, 2000]. Une fois ce problème de terminologie résolu, restent les problèmes liés à des questions de compréhension.
- Des noms de gènes, de protéines ou d'organismes peuvent être cités dans un article sans en constituer le sujet principal, soit par exemple parce qu'ils apparaissent dans une expérimentation ou parce qu'ils correspondent à des modèles de régulation classiques repris comme référence ou simple illustration (par exemple, "l'opéron lactose" ou "tryptophane"). Le simple repérage de ces noms ne saurait donc suffire comme indice de la description d'une interaction.
- De manière plus fondamentale, l'identification des relations de causalité nécessite une analyse plus approfondie de fragments de texte. Ainsi, bien que le sens de la phrase 1, "`la` **`protéine A`** `agit sur l'expression de la` **`protéine B`** `qui inhibe le` **`gène C`**" soit éloigné de celui de la phrase 2, "`Les` **`protéines A`** `et` **`B`** `contrôlent l'expression du` **`gène C`** `dans deux contextes différents,[..]`", une analyse superficielle ne les distingue pas. Dans une méthode à base de mots-clefs, la limitation à moins de 10, du nombre de mots qui apparaissent entre l'agent, la protéine A, et sa cible potentielle, le gène C, permet d'éviter d'identifier à tort l'interaction *protéine A - gène C* dans la phrase 1, mais cette limitation empêche du même coup de repérer l'interaction entre `GerE` et `sigK` qui sont distants de 28 mots dans la phrase "**`GerE`** `stimulates cotD transcription and y cotA transcription in vitro by sigma K RNA polymerase, as expected from in vivo studies, and, unexpectedly, profoundly` **`inhibits`** `in vitro transcription of the gene (`**`sigK`**`) that encode sigma K.`".

3.2 Extraire à partir de textes analysés et normalisés

La méthode que nous proposons consiste à focaliser l'analyse sur les fragments pertinents des textes et combine la démarche des systèmes d'extraction et l'analyse sémantico-conceptuelle des systèmes de compréhension automatique de textes [Poibeau & Nazarenko, 99].

---

[2] Nous profitons ainsi de l'implication de l'unité MIG dans les programmes internationaux de génomique fonctionnelle de la bactérie et de son rôle clé dans la structuration et l'organisation du partage de ces connaissances par le biais de la base de données génomique MICADO déjà existante.

[3] Les exemples extraits de notices réelles sont en anglais, les autres sont en français.

Dans un premier temps, les fragments de texte sont extraits sur la base d'indices de surface. Quand la complexité de l'expression la rend nécessaire, une *représentation conceptuelle du contenu sémantique* des fragments de texte est ensuite construite par une série d'opérations d'interprétation qui s'appuient sur des lexiques syntaxico-sémantiques selon une approche classique dans les travaux de compréhension. La représentation conceptuelle obtenue est confrontée au modèle biologique et réinterprétée, complétée, voire corrigée à la lumière de celui-ci. L'interprétation résultante est soumise à l'application de règles d'extraction de manière à identifier les éléments de l'interaction et à les stocker dans la base de données sous une forme appropriée. Prenons un exemple. De la phrase "`The expression of` *`ac`* `stimulates the expression of` *`sc.`*", nous voulons extraire les faits que nous notons

```
gene(ac), gene_expression(ac,protein1?), protein(protein1?), gene(sc),
positive_interaction(protein1?, gene_expression(sc,protein2?)), protein(protein2?)
```

et qui signifient que, ac est un gène, qu'il exprime une protéine inconnue dans la phrase, notée `protein1?`, qui agit positivement sur l'expression, par le gène sc, d'une protéine non mentionnée dans la phrase et notée `protein2?`. Le modèle biologique, ici très simple, permet de compléter les faits une fois qu'ils sont extraits de la phrase, par les références aux deux protéines objets des expressions de ac et de sc.

Dans une première phase d'interprétation, l'étiquetage grammatical et l'analyse syntaxique identifient les rôles des différents groupes de mots les uns par rapport aux autres (en italique, figure 1[4]). `NprépN` indique ici une relation entre un nom et son complément.

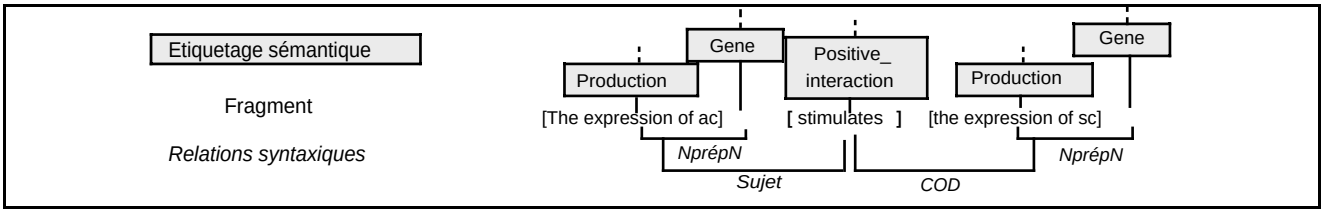

Figure 1. Analyse syntaxique et étiquetage sémantique.

Des *hiérarchies conceptuelles* permettent ensuite un étiquetage sémantique des termes tels que les gènes et protéines et les types d'interaction (encadrés, figure 1). Elles sont formées de classes sémantiques de termes, qui représentent les concepts, ordonnés par une relation de généralité. Des *cadres de sous-catégorisation* limitent les ambiguïtés possibles de l'étiquetage sémantique. Ils représentent pour chaque nom ou verbe, (ou prédicat), l'ensemble de ses relations de dépendances syntaxiques avec les concepts des hiérarchies, par exemple, **`stimulate`**⇒`sujet`⇒`<Protein>` **`stimulate`**⇒`COD`⇒`<expression>`.

Ensuite, des expressions régulières, par exemple des automates, extraient de la phrase ainsi analysée tous les faits recherchés. La figure 2 présente, à titre d'exemple, deux automates qui s'enchaînent pour reconnaître une protéine en jeu dans l'interaction. Ces automates s'appuient sur des caractéristiques discriminantes du texte analysé comme par exemple des catégories grammaticales (Préposition : `Prép`), des liens syntaxiques (complément du nom : `NprépN`) et des étiquettes sémantiques (classe sémantique `Gene`).

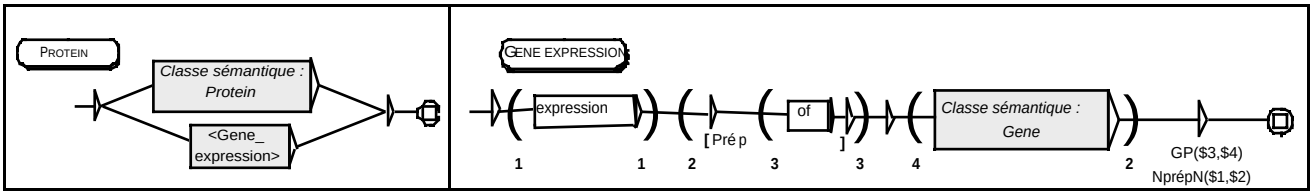

Figure 2. Automates d'extraction : `Protein` et `Gene expression`.

La figure 3 présente l'automate qui reconnaît, au sein d'une proposition, une interaction positive et identifie les éléments en jeu dans cette interaction, c'est-à-dire les valeurs des champs à remplir : les deux protéines ou gènes et l'expression de l'interaction.

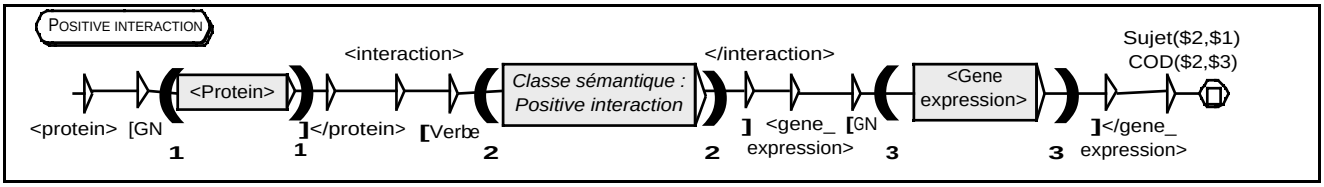

Figure 3. Automate d'extraction : `Positive interaction`.

[4] Pour simplifier, nous n'avons pas fait apparaître les catégories grammaticales dans la figure 1.

Les trois types de connaissances qui interviennent, *cadres de sous-catégorisation*, *hiérarchies conceptuelles* et *automates d'extraction* sont spécifiques au domaine, coûteuses à acquérir et peu réutilisables [Califf & Mooney, 98], [Freitag, 98], [Nédellec, 2000], d'où l'intérêt de les apprendre à partir de corpus. Comme ces connaissances mettent en jeu des relations, il est intéressant d'utiliser des méthodes d'apprentissage relationnelles dans ce cadre.
Dans la suite, nous allons simplement montrer comment les résultats actuels de nos recherches permettent de traiter l'exemple ci-dessus. Nous appellerons ontologie, l'ensemble, cadres de sous-catégorisation et hiérarchies conceptuelles pour un domaine. Des schémas prédicatifs permettront si nécessaire une extraction et une interprétation conceptuelle d'expressions plus complexes. Nous reviendrons sur ce point au paragraphe 4 qui présente plus largement les perspectives de ces travaux.

4 L'apprentissage des ressources

L'objectif de l'étude en cours est la validation de l'approche par l'extraction dans les textes, d'interactions géniques pour *Bacillus subtilis* sur la base d'un ensemble connu de 353 interactions entre gènes et protéines, facteurs sigma, identifiées par l'équipe de Philippe Bessières dans une étude précédente. Le cœur du problème est l'apprentissage d'une ontologie sur les interactions géniques puis d'automates d'extraction de ces interactions à l'aide de cette ontologie.
Des expérimentations antérieures dans le domaine des articles de journaux sur des attentats terroristes (tâche MUC-4) [Faure & Poibeau, 2000] ont déjà montré qu'une ontologie apprise par le système Asium permet de réduire notablement, de l'ordre de 30 %, le temps de mise au point manuelle des automates d'extraction de connaissance, temps d'acquisition coopératif avec Asium inclus, [Faure & Poibeau, 2000].

4.1 Apprendre des ontologies

Nous avions montré dans [Nédellec & Faure, 99] que des ontologies sont apprenables à partir de corpus analysés syntaxiquement, par une méthode de classification telle que celle qui est définie dans Asium. Il s'agit d'acquérir des classes sémantiques à partir de l'observation de régularités syntaxiques dans des corpus [Grishman & Sterling, 94], [Dagan *et al.*, 96] [McMahon & Smith, 96]. Pour évaluer l'applicabilité d'Asium à notre problème biologique, nous avons constitué un corpus initial d'apprentissage d'environ 2200 résumés de MedLine traitant de *Bacillus subtilis* et de transcription.
La première étape a consisté à isoler à l'aide de l'atelier Mo'K [Bisson et Nédellec, 2001], un sous-ensemble de ce corpus pertinent pour l'apprentissage. Les relations syntaxiques les plus porteuses d'information ont été sélectionnées : il s'agit essentiellement de la relation Nom - Complément du nom. Les premières expériences d'apprentissage d'une ontologie avec Asium en fonction de ces contraintes syntaxiques se sont révélées très prometteuses. Les classes sémantiques obtenues sont de bonne qualité, la précision, c'est-à-dire le taux de termes bien classés, est de l'ordre de 95 %, aux erreurs d'analyse syntaxique près. Les erreurs d'analyse syntaxique constituent en moyenne de l'ordre de 30 % des données d'entrée, mais n'apparaissent qu'à 15 % environ dans les meilleures classes conservées dans l'ontologie. L'essentiel du travail des biologistes a consisté à valider et affiner l'ontologie en découpant les classes obtenues en concepts, ceci à l'aide d'Asium. Par exemple, la classe reliée à `[transcription from]` a été partagée en cinq concepts, `region`, `gene`, `phage`, `promoter` et `reporter`.

4.2 Constituer un corpus pour apprendre des classes sémantiques sur les interactions géniques

Les concepts obtenus dans cette première expérience ne sont pas nécessairement les classes les plus pertinentes pour exprimer des interactions géniques. Ils peuvent aussi se rattacher, par exemple, aux conditions d'expérimentation et à l'environnement, auxquels les références sont nombreuses dans le corpus.
Dans le but d'acquérir des ontologies spécialisées sur le sujet de l'interaction, nous avons sélectionné le vocabulaire utilisé dans les contextes relatant des interactions à l'aide de méthodes de classification. Les méthodes utilisées sont de type arbres de décision (C4.5), *bayésien naïf* et IVI [Pillet, 2000]. Elles ont été appliquées à trois corpus de FlyBase et MedLine relatifs à la Drosophile, à Bacillus subtilis et à l'homme et la souris [Nédellec *et al.* 2001]. Les phrases de ce corpus avaient été préalablement classées par les biologistes comme décrivant des interactions ou non. À titre d'exemple, les cinq noms les plus discriminants obtenus pour la drosophile sont `downstream`, `interact`, `modulate`, `autoregulate`, et `eliminate`. La méthode semble fiable : sur une tâche de classification utilisant les probabilités calculées par la méthode bayésien naïf, nous obtenons des taux de précision et de rappel élevés, de l'ordre de 85 à 95 % (évalués par "leave-one-out").
Une fois analysé et *élagué* à l'aide de ce vocabulaire spécifique, le corpus d'apprentissage pour Asium sera normalisé à l'aide des termes identifiés au moyen du système *Acabit* [Daille, 96]. L'ontologie qui pourra être

apprise à partir de ce corpus devrait alors être focalisée sur les seuls termes pertinents pour décrire les interactions.

4.3 Apprendre des automates d'extraction

A partir du texte analysé syntaxiquement et étiqueté sémantiquement à l'aide de l'ontologie, il s'agit d'apprendre des règles d'extraction (dans notre cas des automates). Les approches récentes en la matière convergent vers l'utilisation de méthodes d'apprentissage descendantes relationnelles de type *FOIL* [Quinlan, 90], appliquées à des exemples de phrases *annotés*, [Califf & Mooney, 98], [Freitag, 98], [Soderland, 99], c'est-à-dire dont les valeurs des champs à remplir sont marquées, par exemple, les protéines, gènes et natures des interactions. Comme il n'est pas envisageable d'annoter l'ensemble du corpus pour *Bacillus subtilis* et que ce serait inutile, puisque la grande majorité n'est pas pertinente pour la tâche qui nous intéresse, nous avons automatiquement sélectionné 1300 phrases qui contenaient au moins deux noms de gènes ou de protéines et leurs variantes et synonymes, comme potentiellement porteuses d'information[5].
Nous déterminons actuellement les consignes de choix et d'annotation des exemples d'apprentissage en collaboration avec les biologistes. Ces consignes sont déterminantes car elles ont un impact considérable sur l'efficacité des méthodes qui seront ensuite mises en œuvre. Les consignes sont fonction de la puissance de l'analyse syntaxique, du type de méthode d'apprentissage et de la complexité des modèles biologiques utilisés dans l'interprétation automatique. Par exemple, dans le fragment, "`[..] but suggested that SpoIIID represses sigma K-directed transcription of genes encoding spore coat proteins.`", faut-il que les biologistes annotent l'interaction ternaire complexe entre les protéines `SpoIIID` et `Sigma K` et `genes encoding spore coat proteins` ? Cela dépend évidemment de la sophistication des méthodes qui vont être mises en œuvre.
Inversement, afin de choisir les représentations et les méthodes les plus appropriées nous travaillons à déterminer la nature de la redondance de l'information recherchée et les formes qu'elle prend. Ainsi, si toutes les interactions entre deux gènes donnés A et B sont exprimées au moins une fois dans des formes simples, telles que `A expression represses B expression`, il est inutile de mettre en œuvre des mécanismes sophistiqués, comme par exemple de résolution d'anaphores, pour traiter les formes plus complexes[6].
Une fois l'ensemble d'apprentissage constitué, annoté, analysé et interprété à l'aide de l'ontologie, des méthodes d'apprentissage relationnel de règles d'extraction pourront être mises en œuvre.

5. Perspectives

Au-delà de l'apprentissage d'ontologies et de règles d'extraction, beaucoup reste à faire pour concevoir des méthodes d'extraction d'information avec la précision et la couverture souhaitée. Nous avons illustré à travers des exemples, l'intérêt de l'analyse syntaxique couplée avec l'étiquetage sémantique pour atteindre un niveau d'interprétation des textes suffisant pour extraire de l'information utile. Nous avons vu que les connaissances nécessaires pour interpréter (cadres de sous-catégorisation et hiérarchies conceptuelles) et pour extraire (règles ou automates d'extraction) sont apprenables semi-automatiquement. Dans certains cas pourtant, une interprétation plus *conceptuelle* est nécessaire car le niveau d'interprétation syntaxico-sémantique peut se révéler insuffisant pour apprendre des règles capables d'extraire des informations avec le degré de précision souhaitable.

5.1 L'extraction à un niveau d'interprétation plus conceptuel

L'interprétation à un niveau plus conceptuel permet tout d'abord de réduire considérablement le nombre de règles d'extraction et donc de faciliter la maintenance de la base de connaissance et sa compréhension par le biologiste. Elle peut surtout se révéler indispensable pour l'apprentissage de règles d'extraction. Dans le cas où les interprétations au niveau syntaxico-sémantique des exemples d'apprentissage qui représentent un même type d'interaction, sont très variées et mélangées à d'autres formes au sein des phrases, la méthode d'apprentissage de règles d'extraction peut ne pas découvrir suffisamment de régularités dans les exemples pour être capable d'apprendre des connaissances fiables. Ce problème est aggravé dans le cas où les règles d'extraction portent sur des relations, ici les interactions. Ce problème est connu sous le nom de problème d'apprentissage "multi-slot", [Soderland, 99], [Ciravegna, 2000].
Dans les cas où il n'est pas possible d'étendre la base d'exemples d'apprentissage, la seule solution efficace, selon nous, consisterait à réduire le nombre de formes syntaxico-sémantiques en les abstrayant à un niveau

[5] Cette sélection est facilitée par le fait que les noms de gènes et de protéines de *Bacillus subtilis* suivent des règles de construction globalement respectées.

[6] Notre point de vue serait différent s'il s'agissait de constituer une base bibliographique, car dans ce cas, il faudrait retrouver *toutes* les mentions des interactions, une ne suffirait pas.

conceptuel [Harris *et al.*, 89], de la même façon que, précédemment, nous avions réduit le nombre de formes et augmenté le nombre de régularités dans la base d'apprentissage en étiquetant les termes par des concepts.
Considérons l'exemple de la figure 4 qui présente un ensemble de 30 paraphrases[7]. L'extraction d'information devrait aboutir aux même faits (figure 5) à partir de chacune de ces paraphrases. Si on considère également les formes nominales comme, "`The` **`stimulation`** `of X by Y has been shown [..]`", le nombre de combinaisons augmente encore et il faut également compter avec des phénomènes de type métonymique qui reposent sur le modèle biologique (les protéines sont par exemple exprimées par les gènes en jeu). C'est donc un ensemble important de formulations qui peuvent aboutir à la même représentation au niveau conceptuel (figure 6)[8].

The expression of spoIIID spoIIID expression The spoIIID gene product The production of SpoIIID SpoIID SpoIIID production	stimulates	the expression of sigma K. sigma K expression. the sigma K gene product the production of Sigma K. Sigma K production.

Figure 4. Exemples de phrases représentant la même information.

```
gene(spoIIID).  protein(SpoIIID). gene_expression(spoIIID,SpoIIID). gene(sigma_K).
protein(Sigma K). positive_interaction(SpoIIID, gene_expression(sigma_K,Sigma K)).
```

Figure 5. Faits extraits.

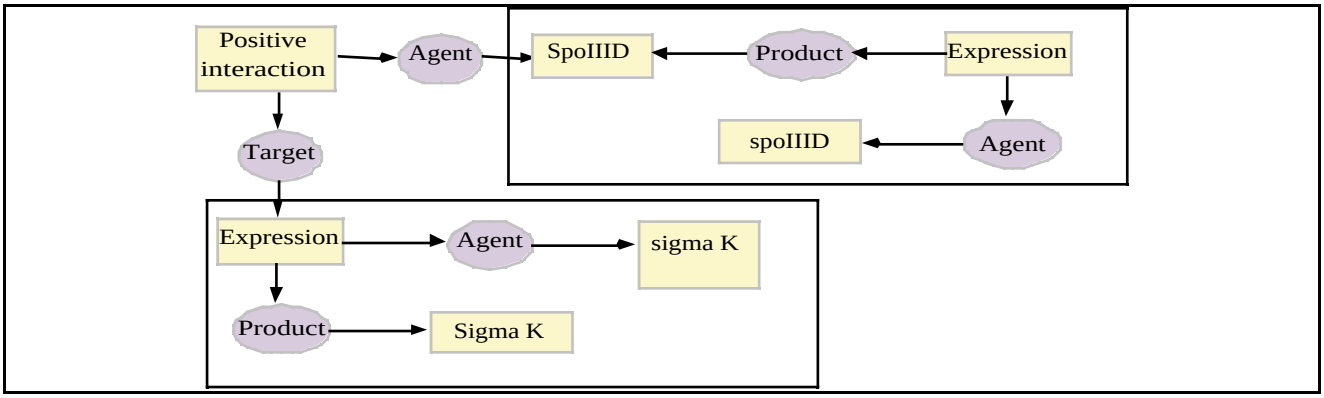

Figure 6. Interprétation conceptuelle.

Les tâches d'apprentissage et d'extraction seraient considérablement simplifiées si les fragments de texte étaient représentés au bon niveau d'abstraction. Pour ce faire, il faut disposer d'une connaissance supplémentaire, les *schémas prédicatifs*, ou *schémas actanciels* ("case frames"), [Fillmore, 68]. Ils permettent d'apparier, sous une forme canonique conceptuelle, des formes de surface très différentes au niveau textuel. Ces schémas sont parfois utilisés en extraction d'information, [Gomez, 98], [Sasaki, 2000], mais aucun moyen ne permet à ce jour de les apprendre sans annotation coûteuse [Thomson, 95][9].

5.2 L'utilisation de schémas prédicatifs

La figure 7 présente deux exemples de schémas prédicatifs, `Positive interaction` et `Gene expression`.

[7] Sachant que selon le modèle biologique, le gène `spoIIID` exprime `SpoIIID` et que le gène `sigma K` exprime `Sigma K.`

[8] Notons que dans cet exemple très simple, l'interprétation au niveau conceptuel est très proche des faits à extraire, ce qui n'est pas le cas en général, d'où l'intérêt de l'apprentissage et de l'utilisation de règles d'extraction.

[9] Au-delà de l'extraction d'information, leurs applications sont pourtant très nombreuses et touchent, entre autres, à la recherche documentaire, aux systèmes dits *question-réponse* et à la traduction automatique.

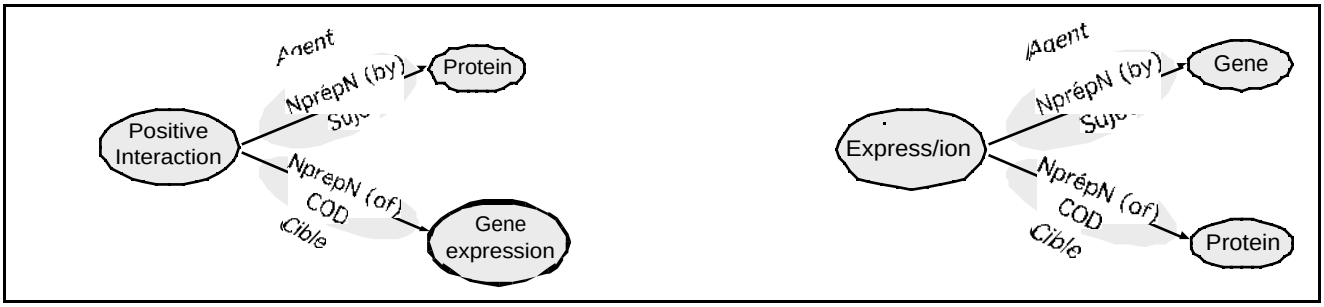

Figure 7. Schémas prédicatifs de `Positive interaction` et de `Gene expression`.

Ils lient des *concepts prédicatifs*, `Positive interaction` et `Express/ion` à leurs arguments conceptuels `Protein` et `Gene`, par des relations sémantiques, `Agent` et `Cible`. Tels que nous les définissons, les schémas prédicatifs précisent également quelles sont les formes qui leur correspondent à un niveau syntaxico-sémantique. À chaque concept prédicatif, comme `Positive Interaction`, correspondent deux ensembles de prédicats nominaux et verbaux, comme (`stimulate`, `interact`) et (`stimulation`, `interaction`). Pour chaque prédicat nominal ou verbal, le schéma prédicatif décrit quelles sont les dépendances syntaxiques correspondant aux relations sémantiques. Par exemple, la relation sémantique `Agent` entre `Positive Interaction` et `Protein` correspond à la dépendance syntaxique `Sujet` entre les formes verbales du prédicat, et l'argument `Protein`. Ces schémas prédicatifs utilisés avec l'ontologie qui définit les classes de protéines, gènes et types d'interaction, permettraient donc de construire des interprétations conceptuelles telles que celle de la figure 6.

5.3 L'apprentissage de schémas prédicatifs

Une fois acquis les cadres de sous-catégorisation verbaux et nominaux du domaine, ainsi que les hiérarchies conceptuelles, l'apprentissage de schémas prédicatifs consistera à classer les schémas de sous-catégorisation en fonction de leurs similarités, chaque classe représentant un schéma prédicatif. Les similarités qui doivent être prises en compte sont,

- les similarités sémantiques entre arguments, mesurées grâce aux hiérarchies conceptuelles, par exemple entre `protein1` et `protein2` (figure 8),
- les variantes dérivationnelles associant des formes verbales et leurs nominalisations, comme `repress` et `repression` (figure 8),
- les transformations syntaxiques possibles, par exemple, la dépendance `sujet` peut être tranformée en complément en `by` (figure 8),
- les similarités sémantiques entre prédicats, mesurées à l'aide des hiérarchies conceptuelles, par exemple, entre `inhibit` et `repress` (figure 9).

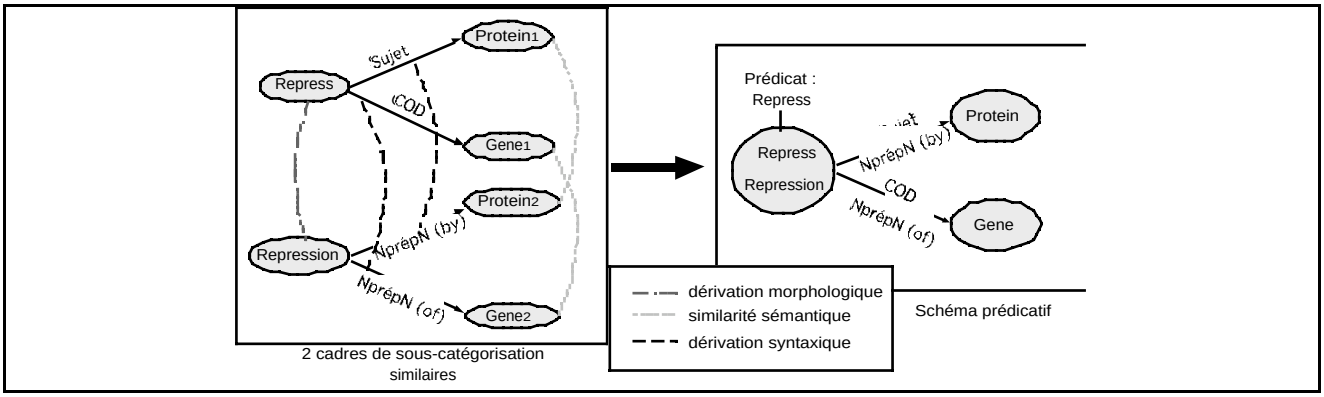

Figure 8. Apprentissage de schémas prédicatifs sur la base de similarités sémantiques entre arguments et de dérivations morphologiques et syntaxiques.

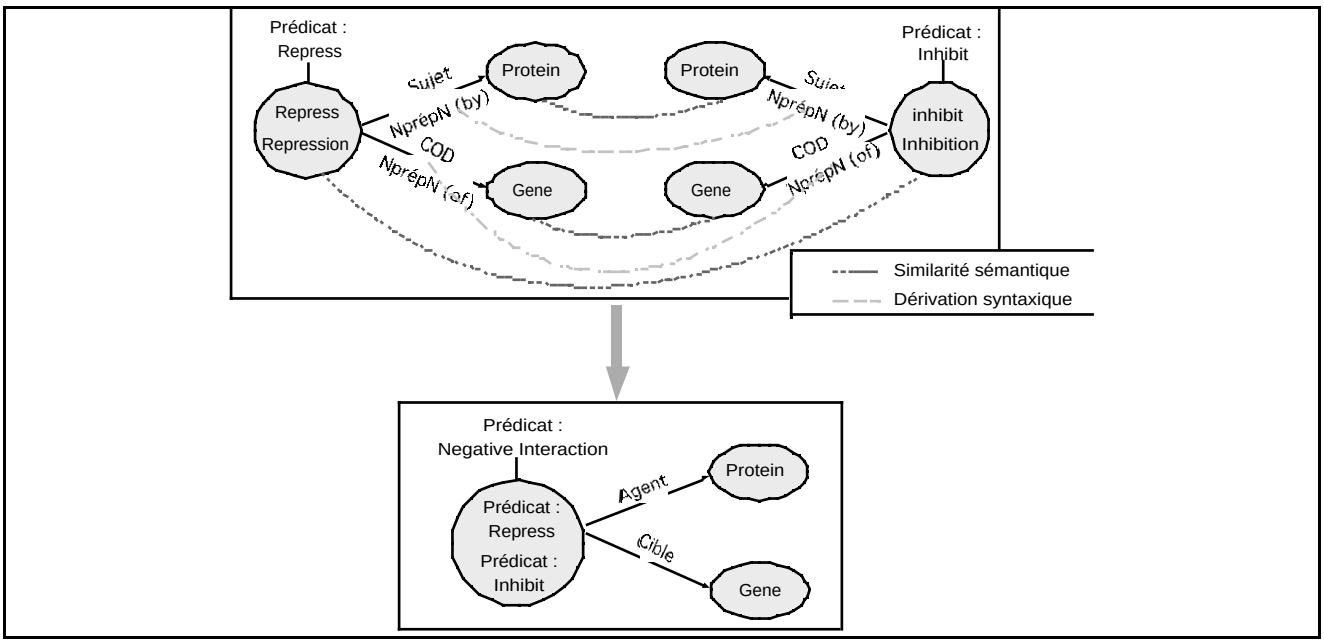

Figure 9. Apprentissage de schémas prédicatifs sur la base de similarités sémantiques entre prédicats et de dérivations syntaxiques.

Reste ensuite éventuellement à identifier les relations sémantiques, par exemple, `Agent` et `Cible` (figure 9), ce qui peut nécessiter une intervention manuelle pour lever les ambiguïtés résiduelles. Cet apprentissage nécessite d'une part, la conception de nouvelles méthodes de classification conceptuelle relationnelle qui pourront s'inspirer des rares exemples de méthodes existants [Bisson, 92], et d'autre part, l'étude des dérivations et transformations morphologiques et syntaxiques spécifiques au domaine. Nous envisageons l'automatisation de cette étude grâce au système FASTR. Il permet le repérage systématique de variantes nomino-verbales telles qu'ici `inhibition-inhibit`, `expression-express`, etc. [Jacquemin & Tzoukermann, 97].

L'apprentissage automatique d'ontologies associé à la relative spécificité et homogénéité des corpus sur lesquels nous travaillons permettent d'envisager d'apprendre ces schémas prédicatifs semi-automatiquement par la méthode proposée.

6. Conclusion

L'approche adoptée jusqu'à présent en extraction d'information repose sur une analyse locale de surface basée sur des mots déclencheurs connus et des patrons lexico-syntaxiques encodés dans des règles ou des automates. Dans des domaines scientifiques et techniques, tels que la génomique, la nature de la tâche d'extraction change en raison des informations recherchées et du type des documents qui rendent nécessaire l'utilisation de méthodes d'analyse plus profondes. Celles-ci font appel à des ressources lexicales, sémantiques et conceptuelles spécialisées. Nous avons montré que les méthodes d'apprentissage relationnel coopératif possédaient d'excellentes propriétés pour l'apprentissage de telles ressources à partir de corpus de textes, là où des approches manuelles ou statistiques se révèlent impraticables ou limitées. Nous avons également montré qu'elles étaient appropriées pour apprendre les règles d'extraction elles-mêmes. Les résultats préliminaires déjà obtenus, en collaboration avec des biologistes de MIG, sont très prometteurs en ce sens qu'ils illustrent la faisabilité de l'approche tout en produisant des résultats d'ores et déjà utilisables pour l'extraction d'information sur les interactions géniques.

Remerciements

Ce travail est financé partiellement par le Ministère de l'Economie, des Finances et de l'Industrie à travers le contrat RNRT Astuxe, et par le CNRS, l'INRA, l'INRIA et l'INSERM à travers le projet bioinformatique Caderige. Les auteurs remercient Mohammed Ould Abdel Vetah et Bernard Weiss pour leur participation aux expérimentations décrites dans cet article.

Références

Bisson G., "Learning in FOL with a similarity measure". In Proccedings of the *11th National Conference on Artificial Intelligence (AAAI),* San Jose, CA., p. 82-87, AAAI Press, 1992.

Bisson G. et Nédellec C., "Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier Mo'K" in Actes des *Journées Francophones d'Extraction et de Gestion des Connaissances (EGC'2001)*, Briand H. (Ed.), Hermès (Pub.), Nantes, Janvier 2001.

Blaschke C., Andrade M. A., Ouzounis C. and Valencia A., "Automatic Extraction of biological information from scientific text: protein-protein interactions", in Proceedings of *International Symposium on Molecular Biology, (ISMB'99)*, 1999.

Califf M. E. and Mooney R. J., "Relational Learning of Pattern-Match Rules for Information Extraction." In Proceedings of *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, Stanford, CA, Mars, 1998.

Cardie C., Daelemans W., Nédellec C. and Tjong Kim Sang E., Proceedings of the *Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, Omni Press (Pub.), Lisbonne, Septembre 2000.

Ciravegna F., "Learning to Tag for Information Extraction from Text". In Proceedings of the *ECAI-2000 Workshop on Machine Learning for Information Extraction*, F. Ciravegna *et al.* (Eds.), Berlin, August 2000.

Collier N, Nobata C. and Tsujii, "Extracting the names of genes and gene products with a hidden Markov model. In Proceedings of the *18th International Conference on Computational Linguistics (COLING'2000)*, Saarbrück, Allemagne, Juillet-Août 2000.

Craven M. and Kumlien J., "Constructing Biological Knowledge Bases by Extracting Information from Text Sources.", In Proceedings of the *7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, 1999.

Dagan I., Lee L., and Pereira F., "Similarity-Based Methods For Word-Sense Disambiguation", in Proceedings of the *Annual Meeting of the Association for Computational Linguistics, ACL'96*, 1996.

Daille B., "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology ", in P.

Faure D. and Nédellec C.,"Knowledge Acquisition of Predicate-Argument Structures from technichal Texts using Machine Learning" in Proceedings of *Current Developments in Knowledge Acquisition: EKAW-99*, p. 329-334, Fensel D. and Studer R. (Ed.), Springer Verlag, Karlsruhe, Allemagne, Avril 1999.

Faure D. et Poibeau T., "Extraction d'information utilisant Intex et des connaissances sémantiques apprises par Asium, premières expérimentations " in actes du *12ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle (RFIA'2000)*, Schmitt F. et Bloch I. (Eds.), Paris, Février 2000.

Fillmore C., "The case for case." In E. Bach and R. T. Harms (Eds), *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, 1968.

Freitag D., "Toward General-Purpose Learning for Information Extraction.", in *Proceedings of the Seventeenth International Conference on Computational Linguistics (COLING-ACL-98)*, 1998.

Fukuda K., Tsunoda T., Tamura A. and Takagi T., "Toward Information Extraction: Identifying protein names from biological papers". In *Proceedings of the Pacific Symposium on biocomputing (PSB'1998)*, 1998.

Gomez F., "Linking WordNet Classes to Semantic Interpretation", in Proceedings of the *COLING-ACL Workshop on the usage of WordNet in NLP Systems*, 1998.

Grishman R. and Sterling J., "Generalizing Automatically Generated Selectional Patterns", in Proceedings of the *16th International Conference on Computational Linguistics (COLING'94)*, 1994.

Haldenwang W. G., « The sigma factors of Bacillus subtilis. » Microbiol. Rev. vol 59, 1-30, 1995.

Harris Z., Gottfried M., Ryckman T., Mattick Jr P., Daladier A., Harris T. N. and Harris S. *The Form of Information in Science, Analysis of Immunology Sublanguage*. Volume 104 of Boston Studies in the Philosophy of Science. Kluwer Academic Publisher, Boston, 1989.

Humphreys K., Demetriou G, and Gaizauskas R., "Two applications of information extraction to biological science article: enzyme interaction and protein structure. In Proceedings of the *Pacific Symposium on biocomputing (PSB'2000)*, vol.5, p. 502-513, Honolulu, 2000.Jacquemin C. and Tzoukermann E., "NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax." T.Strzalkowski (Ed.), Natural Language Information Retrieval. Klüwer, Boston, MA, 1997.

McMahon J. G. and Smith F. J., "Improving statistical language model performance with automaticlly generted word hierarchies". *Computationnal Linguistics*, 22(2), 217-247.

Nédellec C., "Knowledge Extraction from Text, a Machine Learning Approach". In Proceedings of the *Third International Conference on Human-System Learning, CAPS'3, Learning WWW*, Europia Production (Pub.), Paris, France, Décembre 2000.

Nédellec C., Ould Abdel Vetah M., Bessières P., Brun C. et Jacq B., "Filtrage de phrases pour l'extraction d'information en génomique, un problème de classification." In *Actes de la Conférence Francophone d'Apprentissage (CAP'2001)*, à paraître, 2001..

Pillet V., *Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information*, thèse de l'Université de droit, d'économie et des sciences d'Aix-Marseille, 2000.

Poibeau T., "Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états" In *Actes de la Conférence Française de Traitement Automatique de la Langue, (TALN'2001)*, à paraître, 2001.

Poibeau T. et Nazarenko A., "L'extraction d'information, une nouvelle conception de la compréhension de texte ?", Traitement Automatique des Langues, vol 40 n°2, p. 87-115, 1999.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V., Jacq, B., "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Exctraction". In *Genome Informatics 1998*, S. Miyano and T. Takagi, (Eds), Universal Academy Press, Inc, Tokyo, Japan, p. 72 - 80, 1998.

Quinlan J., "Learning logical definitions from relations.", In *Machine Learning Journal*, vol 5(3) p. 239-266, 1990.

Riloff E., "Automatically constructing a Dictionary for Information Extraction Tasks". In Proceedings of the *Eleventh National Conference on Artificial Intelligence (AAAI-93)*, p. 811-816, AAAI Press / The MIT Press, 1993.

Rindflesch T.C., Tanabe L., Weinstein J. N., and Hunter L., "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature." In Proceedings of the *Pacific Symposium on Biocomputing (PSB'2000)*, vol 5, p 514-525, 2000.

Sasaki Y. and Matsuo Y., "Learning Semantic-Level Information Extraction Rules by Type-Oriented ILP", in Proceedings of the *18th International Conference on Computational Linguistics, COLING-2000*, Kay M. (Ed.), Saarbrücken, 2000.

Scott S. and Matwin S., "Feature Engineering for Text Classification", in Proceedings of *ICML'99*, 1999.

Soderland S., "Learning Information Extraction Rules for Semi-Structured and Free Text" in *Machine Learning Journal*, vol 34, 1999.

Staab S., Mädche A., Nédellec C. and Wiemer-Hastings P., ECAI Workshop Notes of the *Ontology Learning*, workshop of the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Août 2000.

Stapley B. J. and Benoit G., "Bibliometrics: Information Retrieval and Visualization from co-occurrence of gene names in MedLine abstracts.". In Proceedings of the *Pacific Symposium on biocomputing (PSB'2000)*, 2000.Thomas, J., Milward, D., Ouzounis C., Pulman S. and Caroll M., "Automatic Extraction of Protein Interactions from Scientific Abstracts". In Proceedings of the *Pacific Symposium on biocomputing (PSB'2000)*, vol.5, p. 502-513, Honolulu, 2000.

Thompson C. A., "Acquisition of a Lexicon from Semantic Representations of Sentences", in Proceedings of the *33rd Annual Meeting of the Association of Computational Linguistics, (ACL'95)*, p. 335-337, Boston, M A, Juillet, 1995.

Wosten M. M., "Eubacterial sigma-factors." FEMS Microbiol. Rev. vol 3, 127-50, 1998.