Extraction d'information appliquée au domaine biomédical : apprentissage et traitement automatique de la langue

Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Lagarrigue, Adeline Nazarenko, Alain Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, et al.

Extraction d'Information appliquée au domaine biomédical

Apprentissage et traitement automatique de la langue

Erick Alphonse, Sophie Aubin*, Philippe Bessières**, Gilles Bisson****, Thierry Hamon*, Sandrine Lagarrigue***, Adeline Nazarenko*, Alain-Pierre Manine**, Claire Nédellec**, Mohamed Ould Abdel Vetah**, Thierry Poibeau* et Davy Weissenbacher***

**Laboratoire d'Informatique de Paris-Nord CNRS UMR 7030Av. J.B. Clément 93430 F-Villetaneuse*
{prénom.nom}@lipn.univ-paris13.fr

***Laboratoire Mathématique, Informatique et Génome (MIG), INRA, Domaine de Vilvert, 78352 F-Jouy-en-Josas*
{prénom.nom}@jouy.inra.fr

****Laboratoire de Génétique Animale, INRA-ENSARRoute de Saint Brieuc, 35042 Rennes Cedex*
lagarrig@roazhon.inra.fr

*****Laboratoire Leibniz – UMR CNRS 5522 46 Avenue Félix Viallet - 38031 F-Grenoble Cedex*
Gilles.Bisson@imag.fr

RÉSUMÉ. *Cet article présente une vue d'ensemble du projet Caderige. Ce projet pluridisciplinaire fait intervenir des équipes de compétences complémentaires (biologie, apprentissage artificiel, traitement automatique des langues naturelles) afin de développer des outils automatiques d'analyse destinés à extraire des informations structurées de bases de données bibliographiques spécialisées en génomique (en particulier Medline). Cet article donne un aperçu de notre approche et la compare à l'état de l'art..*

ABSTRACT. *The instructions put together below fall into three categories. The publisher would be grateful to authors for respecting these indications. The length of this summary may attain a dozen lines. It is to be written in size 9 italic Times. An abstract in French will be joined.*

MOTS-CLÉS : *fouille de texte, extraction d'information, génomique.*

KEYWORDS: *text mining, information extraction, genomics.*

Actes de CIFT 2004, pages 7 à 20

1. Introduction

Les résultats de recherches biologiques ou biomédicales sont consignés au sein de vastes bases de données bibliographiques spécialisées (e.g. Flybase, spécialisée sur Drosophilia Melanogaster) ou généralistes telle que . Medline. Il s'agit de sources d'information cruciales pour les biologistes. Cependant, il n'existe pas d'outils automatiques permettant de les explorer et d'en extraire des informations pertinentes. Alors que la reconnaissance d'entités nommées a acquis un certain succès dans le domaine biomédical, l'Extraction d'Information (IE) reste un défi.

Le projet Caderige a pour but de concevoir et d'intégrer des outils de Traitement Automatique des Langues Naturelles (TALN) et d'Apprentissage Artificiel (AA) afin d'explorer et d'analyser des bases de données biologiques, et d'en extraire des informations ciblées. Nous promouvons une approche à base de corpus, centrée sur une analyse préliminaire du texte et sur sa normalisation afin de nous affranchir, autant que possible, des variations linguistiques. Les conférences MUC (1995) ont montré une meilleure efficacité des tâches d'extraction appliquées au texte normalisé : les patrons d'extraction sont plus faciles à réaliser ou à apprendre, plus génériques et plus aisément maintenables.

De plus, comme cela sera exposé par la suite, il est également possible d'acquérir automatiquement une partie des connaissances nécessaires à la normalisation du texte, ceci à partir de corpus à l'aide de méthodes d'apprentissage.

Cet article présente une vue d'ensemble du projet Caderige et de ses résultats. Dans un premier temps, nous introduirons notre approche puis, nous présenterons les techniques de filtrage de phrases et de normalisation employées dans Caderige : résolution des synonymes pour les entités nommées, analyse syntaxique et apprentissage d'ontologies. Pour finir, nous montrerons comment des patrons d'extraction peuvent être appris à partir de textes normalisés et annotés.

2. Description de l'approche

Nous présentons dans cette partie les principes centraux de notre approche.

2.1. *Organisation du projet*

Le projet Caderige (2000 - 2003) est un projet multidisciplinaire concernant la fouille automatique de texte biomédical. Il est principalement destiné à un usage exploratoire. Les principaux partenaires sont des équipes biologiques (INRA), informatiques (LIPN, INRA, et Leibniz-IMAG), et de traitement des langues naturelles (LIPN). Le LRI et l'INRIA ont également été impliqués.

2.2. *Motivation du projet*

Les biologistes exploitent les bases de données bibliographiques en soumettant des requêtes constituées de mots-clefs retournant un ensemble de références. Il leur est également possible d'utiliser les références présentes dans les bases de données génomiques sous la forme d'hyperliens vers les bases de données bibliographiques. Afin d'extraire les connaissances recherchées, il leur faut identifier les résumés ou les paragraphes pertinents de ces références. Une telle démarche manuelle est répétitive et coûteuse en temps. Du fait de la taille de la bibliographie, les données pertinentes sont très éparpillées. De plus, les bases de données sont en constante mise à jour. Ainsi, la requête « Bacillus subtilis and transcription » soumise en 2002 à la base de données Medline retournait 2209 résumés ; à présent, cette requête en retourne 2693. Cet exemple a été choisi car Bacillus subtilis est une bactérie modèle et la transcription un phénomène fondamental en génomique fonctionnelle qui implique de nombreuses interactions géniques. L'extraction d'interactions géniques pose un problème d'EI attractif et très étudié.

GerE stimulates **cotD** transcription and inhibits **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.**

Figure 1: **Phrase décrivant une interaction génique**

Une fois les résumés pertinents sélectionnés, les formulaires décrivant les interactions doivent être remplis manuellement car il n'existe à l'heure actuelle aucun outil d'EI opérationnel dans le domaine de la génomique.

Interaction	**Type:** positive
	Agent: *GerE*
	Cible: transcription of the gene **sigK**

Figure 2: **Patron décrivant une interaction génique.**

L'adaptation à la génomique (et, plus généralement, à la biologie) des méthodes d'EI employées pour MUC n'est pas une tâche triviale : les systèmes d'EI nécessitent des méthodes d'analyse complexes pour localiser les fragments pertinents. Comme le montrent les figures 1 et 2, extraire *GerE* en tant qu'agent de l'inhibition de la transcription du gène *sigK* nécessite au minimum l'analyse des dépendances syntaxiques et du mécanisme de coordination. De surcroît, pour la plupart des tâches d'EI en génomique (fonction, localisation, homologie), ces méthodes doivent être combinées à l'analyse sémantique et conceptuelle du texte issue du domaine de la compréhension de texte.

2.3 Principes généraux

Notre approche, comme celle du projet Genia (Collier et al., 1999) s'appuie sur une analyse linguistique profonde comme préalable à l'EI. Le type d'information recherchée est similaire, et concerne principalement les interactions géniques et protéiques comme la plupart des activités de recherche dans ce domaine. Le corpus de Genia (Ohtae et al., 2001) n'est pas spécialisé sur une espèce donnée alors que le nôtre concerne Bacillus subtilis.

Ces deux projets développent des outils d'annotation et des DTD (Document Type Definition) relativement compatibles. Leur but conjoint est de construire un corpus d'apprentissage et d'y appliquer diverses techniques de TALN et de AA afin d'acquérir des patrons efficaces d'extraction d'événements. Le choix des méthodes de AA et de TALN diffèrent entre les deux projets, mais leur objectif reste le même : normaliser le texte pour le structurer sous forme prédicative pour un meilleur apprentissage des patrons d'extraction. Genia applique une analyse de type HPSG en combinant plusieurs analyseurs, alors que l'analyse syntaxique de Caderige est basée sur une spécialisation de Link Parser (Sleator et Temperley, 1993, voir paragraphe 4) pour le domaine de la génomique.

Dans les deux sections suivantes, nous détaillerons nos méthodes de filtrage de texte et de normalisation. Le filtrage a pour but de supprimer les parties non pertinentes du corpus tandis que la normalisation bâtit une représentation abstraite du texte pertinent. La section 4 est consacrée à l'acquisition de patrons d'extraction à partir du texte filtré et normalisé.

3. Filtrage du texte

La Recherche d'Information (RI) ainsi que le filtrage de texte sont deux conditions préalables à l'EI. Les méthodes d'EI (incluant la normalisation et l'apprentissage) ne peuvent être appliquées à des corpus de taille trop importante ou trop bruités (elles ne sont pas assez robustes et sont trop coûteuses en temps de calcul). La RI s'effectue via l'interface de Medline grâce à des requêtes (sous forme de mots-clefs) renvoyant un ensemble de documents appropriés. Le filtrage de texte réduit alors la variabilité des données textuelles selon les hypothèses suivantes :

l'information recherchée est locale aux phrases;

les phrases pertinentes contiennent au moins deux noms de gènes.

Ces hypothèses peuvent conduire à éliminer certaines interactions géniques. Cependant, nous faisons l'hypothèse que la redondance de l'information est telle qu'au moins une instance de chaque interaction sera conservée.

Les documents sont automatiquement segmentés en phrases et les phrases contenant au moins deux noms de gènes sont sélectionnées.

Afin d'identifier les phrases pertinentes au sein de cet ensemble, des méthodes classiques de AA ont été appliquées à un corpus dédié à *Bacillus subtilis*. Chaque phrase de ce corpus a été préalablement annotée comme pertinente ou non pertinente par un expert biologiste. Parmi les SVMs, Naive Bayes (NB), réseaux de neurones, arbres de décision, (Marcotte *et al*., 2001; Nedellec *et al*., 2001), (Nedellec *et al.,* 2001) ont démontré qu'une méthode simple à base de Naive Bayes couplée à des algorithmes de sélection d'attributs affichait de bonnes performances (environ 75% de précision et de rappel). De plus, nos expérimentations préliminaires ont montré que les changements de représentation basés sur des outils linguistiques (lemmatisation, terminologie et entités nommées) n'apportent aucune amélioration significative.

Les phrases filtrées à cette étape sont utilisées comme entrées par les tâches suivantes (normalisation et Extraction d'Information).

4. Normalisation

Cette partie propose une brève présentation de trois tâches de normalisation des textes : normalisation des entités nommées, mise en évidence des relations entre éléments du texte grâce à une analyse syntaxique et à un étiquetage sémantique. Une telle normalisation facilite l'acquisition et l'apprentissage de règles d'extraction en offrant une représentation plus abstraite des phrases.

4.1 Normalisation des noms d'entités

Les textes de biologie sont un terrain intéressant pour la reconnaissance des Entités Nommées (EN) et ont déjà suscité de nombreux travaux autour de la détection des noms de gènes (Proux et al., 1998), (Fukuda et al., 1998). Caderige n'a pas pour objectif la création d'un nouvel extracteur d'EN, mais plutôt l'étude d'un problème particulier qui est la gestion des synonymes d'EN. En plus des classiques variations typographiques et autres abréviations, une même entité biologique peut être désignée par plusieurs noms. La synonymie des noms de gènes est un problème reconnu. En effet, un nom temporaire est souvent donné à un gène lors de sa découverte. Plus tard, il est renommé d'après les différentes informations recueillies à son sujet. Par exemple, SYGP-ORF50 est le nom temporaire attribué au gène PMD1 de la levure dans le cadre d'un projet de séquençage. Nous avons pu montrer qu'en plus des informations disponibles dans les bases de données de génomique (GenBank, SwissProt, etc.), il est possible d'acquérir des relations de synonymie dans les textes avec une bonne précision. A partir d'amorces de synonymie telles que « also called » ou « formerly », nous pouvons extraire des fragments du type : gene amorce gene.

Toutefois, ces amorces sont elles-mêmes sujettes à variation et les arguments de la relation de synonymie doivent être précisément identifiés au préalable. Des patrons ont été définis et entraînés sur un échantillon représentatif extrait de Medline puis testés sur un nouveau corpus de 106 phrases présentant l'amorce « formerly ». La précision obtenue est de 97,5% et le rappel de 75%. Nous avons privilégié la précision puisque l'information ainsi obtenue doit être valide pour les étapes suivantes de la chaîne de traitement.

L'approche décrite ici se veut très modulaire car de tels patrons (gene amorce gene, où amorce est un marqueur linguistique ou une ponctuation) peuvent être instanciés de nombreuses manières. Un score est calculé pour chaque instanciation lors de la phase d'apprentissage sur un large corpus représentatif. L'utilisation d'un corpus étiqueté de taille restreinte et d'un large corpus non étiqueté implique la mise en œuvre de méthodes d'apprentissage semi-supervisé. Cette question est encore actuellement en cours d'étude au LIPN.

4.2 Analyse syntaxique

L'extraction d'informations structurées nécessite une analyse précise des dépendances syntaxiques existantes entre les différentes entités du domaine. A la différence de (Akane et al., 2001), nous avons choisi de procéder à une analyse syntaxique partielle : seules certaines relations calculées à partir de phrases pertinentes pré-sélectionnées nous intéressent. Les raisons pour lesquelles nous avons fait ce choix sont, d'une part, que les informations pertinentes apparaissent généralement au sein de structures syntaxiques identifiables et, d'autre part, que nous acquérons les ontologies à partir de relations syntaxiques particulières (Faure et Nedellec, 2000 ; Bisson et al., 2000).

Nous avons procédé à une première série de tests sur des analyseurs syntaxiques partiels. Il en est ressorti que les stratégies basées sur les constituants ne permettent pas toujours l'extraction de relations pertinentes entre groupes. Les analyseurs basés sur des grammaires de dépendances sont plus adaptés à notre tâche, puisqu'ils calculent les liens entre les têtes des différents groupes syntaxiques. De plus, et comme le décrit (Schneider, 1998), ces derniers sont généralement moins contraints au niveau de l'ordre des mots, ce qui nous semble être un avantage lorsqu'on travaille sur des textes de spécialité.

Deux analyseurs en dépendance ont été plus longuement étudiés (Aubin 2003) : un outil commercial fonctionnant à la fois sur les constituants et les dépendances (nous l'appellerons ACD) et un analyseur en dépendances : Link Parser.

Concernant l'évaluation des analyseurs, (Prasad et Sarkar, 2000) recommandent de procéder aux tests sur un corpus existant et sur un ensemble de phrases créées artificiellement. L'idée est de connaître le comportement de l'analyseur à la fois sur des données réelles (du domaine) et sur des constructions syntaxiques particulières.

•		Link Parser				ACD			
•Rel	nbRel	relOK	R.	RelTot	P.	RelOK	R	RelTot	P.
Sujet	18	13	0.72	19	0.68	14	0.78	20	0.65
Objet	18	16	0.89	17	0.94	9	0.5	13	0.69
Prep	48	25	0.52	55	0.45	20	0.42	49	0.41
V-GP1	14	13	0.93	15	0.87	9	0.64	23	0.39
O-GP	16	7	0.43	12	0.58	12	0.75	28	0.43
NofN	16	13	0.81	15	0.87	14	0.87	26	0.54
VtoV	10	9	0.9	9	1	7	0.7	7	1
VcooV	10	8	0.8	9	0.89	6	0.6	6	1
NcooN	10	8	0.7	10	0.8	4	0.4	6	0.67
nV-Adj	10	8	0.8	9	0.89	0	0	0	1
PaSim	18	17	0.94	18	0.94	17	0.94	22	0.77
PaRel	12	11	0.92	11	1	8	0.67	11	0.73

Table 1: **Résultats des évaluations par relations syntaxiques**

Relations : Sujet = sujet-verbe, Objet = verbe-objet, Prep = groupe prépositionnel, V-GP = verbe-groupe prep., O-GP = Objet- groupe prep., NofN = Nom of nom, VtoV = Verbe to Verbe, VcooV = Verbe coord. Verbe, NcooN = Nom coord. Nom, nV-Adj = not + Verbe ou Adjectif, PaSim = passif, PaRel = prop. relative avec passif

Nous avons opté pour le compromis en constituant un jeu de test formé de phrases extraites du corpus Medline que nous avons organisées selon leur spécificité syntaxique.

Un ensemble de relations syntaxiques a donc été sélectionné puis évalué à la main. Le tableau 1 présente les résultats pour les principales relations, indiquant à chaque fois le rappel et la précision (rappel : # de relations pertinentes trouvées / # de relations à trouver; précision : # de relations pertinentes trouvées / # total de relations trouvées par l'analyseur).

Link Parser obtient généralement de meilleurs résultats que ACD. Nous expliquons ceci en partie par le fait que le corpus Medline contient beaucoup de phrases très longues (27 mots en moyenne) souvent constituées de plusieurs propositions. Alors que Link Parser parvient à repérer les têtes de groupes syntaxiques et à les mettre en relation, ACD est mis en échec par la complexité des phrases.

Nous avons donc opté pour Link Parser qui présente également l'avantage d'avoir une grammaire et un lexique modifiables (voir la section suivante). Link Parser est actuellement utilisé à l'INRA pour analyser les textes afin d'apprendre des ontologies par analyse distributionnelle (Harris 1951, Faure et Nédellec, 1999) et d'apprendre des patrons d'extraction.

4.3 Adaptation d'un analyseur syntaxique général au domaine biologique

L'évaluation des analyseurs, a permis d'identifier les modifications qui devaient concerner -l'analyseur et celles qui devaient être appliquées au texte lui-même afin d'améliorer les performances. L'analyse syntaxique échoue en effet sur des structures qui peuvent être repérées a priori par des pré-traitements efficaces du texte adaptés au domaine, par exemple, la suppression des références bibliographiques.

Link Parser a été modifié notamment par l'ajout de dictionnaires et de nouvelles règles afin de permettre l'analyse de structures spécifiques au domaine biologique. Ainsi, le nom de la bactérie Bacillus Subtilis est formé à partir d'éléments lexicaux hérités du latin qui doivent être introduits dans le dictionnaire.

Les règles syntaxiques doivent par ailleurs être affaiblies afin d'accepter des structures parfois écrites dans un anglais approximatif (Medline est alimenté par des auteurs du monde entier). Une des erreurs les plus fréquentes est l'absence de déterminant devant des noms qui en requièrent pourtant un.

En ce qui concerne le corpus, une étape de normalisation a été ajoutée afin de permettre la reconnaissance des entités nommées, la segmentation en phrases, la lemmatisation et l'analyse de la terminologie. Cette dernière tâche a pour conséquence de supprimer du texte un grand nombre d'ambiguïtés, ce qui améliore les performances de l'analyseur tant en qualité qu'en durée d'exécution. Un analyseur de termes est actuellement en cours d'élaboration au LIPN ; celui-ci est en partie fondé sur des ressources existantes comme Gene Ontology (ce point est détaillé dans Hamon et Aubin, 2004).

4.5 Etiquetage sémantique

Le logiciel Asium est utilisé afin d'acquérir de manière semi-automatique des catégories sémantiques grâce à une analyse distributionnelle de corpus.

Ces catégories contribuent à la normalisation du texte sur deux points : elles aident à la désambiguïsation de structures syntaxiques ayant reçu plusieurs analyses possibles tels que les attachements verbaux/adjectivaux et elles permettent le typage des entités pertinentes pour la tâche d'extraction. Asium est fondé sur une méthode de classification hiérarchique ascendante qui construit une hiérarchie de classes à partir des dépendances syntaxiques analysées dans un corpus d'apprentissage. Une validation manuelle est nécessaire afin de distinguer les différents sens de mots partageant les mêmes structures syntaxiques.

5. Apprentissage des patrons d'extraction

L'apprentissage permet d'envisager une acquisition semi-automatique de patrons d'extraction. Cet apprentissage nécessite un corpus d'entraînement à partir duquel des régularités pertinentes et discriminantes sont identifiées. Cette tâche est elle-même basée sur deux processus fondamentaux : une normalisation des textes décrite au paragraphe 4 (cette normalisation est dépendante du domaine mais pas de la tâche visée) et l'annotation de données pertinentes par rapport à la tâche

5.1 Procédure d'annotation

Le langage d'annotation retenu dans le cadre de Caderige est XML. Une DTD (Définition de Type de Documents) spécifique a été définie par un groupe d'experts incluant des biologistes. Cette DTD convient aussi bien aux organismes procaryote qu'eucaryotes ; elle comporte 50 étiquettes pouvant compter jusqu'à 8 attributs. Une telle précision est nécessaire à l'apprentissage et à la tâche d'extraction visée. Concrètement, l'annotation permet de mettre en évidence des séquences textuelles décrivant :

· les agents (A): entités activant ou contrôlant l'interaction

· les cibles (T): entités produites ou contrôlées

· les interactions (I): type de contrôle en jeu dans le processus

· la confiance (C): niveau de confiance accordé à l'interaction décrite.

Le résultat de l'annotation de « A low level of GerE activated transcription of CotD by GerE RNA polymerase in vitro ... » est donné ci-dessous. Les attributs associés à l'étiquette <GENIC-INTERACTION> reflètent le fait qu'il s'agit d'une activation transcriptionnelle qui a ici un statut avéré. Les autres étiquettes (<IF>, <AF1>, …) concernent l'agent (AF1 and AF2), la cible (TF1) et l'interaction (IF).

```
<GENIC-INTERACTION
    id="1"
    type="transcriptional"
    assertion="exist"
    regulation="activate"
    uncertainty="certain"
    self-contained="yes"
    text-clarity="good">
  <IF>A<I> low level </I>of</IF>
  <AF1><A1
        type=protein
       role=modulate
       direct=yes> GerE
  </A1></AF1>,
  <IF><I>activated</I> transcription
     of</IF>
  <TF1><T1 type=protein> CotD </T1>
     </TF1> by
```

```
      <AF2><A2
            type=protein
                role=required>
        GerE RNA polymerase
    </A2></AF2>,
    <CF>but<C>in vitro</C></CF>
</GENIC-INTERACTION>
```

5.2 L'éditeur d'annotation

Les annotations ne peuvent pas être directement réalisées par les biologistes dans un format textuel. Caderige offre un cadre d'annotation avec un éditeur XML général, doté d'une interface graphique pour créer, vérifier et modifier les documents annotés. Sur la base d'une feuille de style XML, l'éditeur affiche le texte avec des attributs graphiques, permet l'ajout de balises sans contraintes fortes sur l'ordre d'insertion, et effectue automatiquement un certain nombre de vérifications d'intégrité.

L'interface est composée de quatre parties principales (voir Figure 3). La zone éditable d'annotation (text), la liste des balises disponibles XML (tags), la zone pour éditer les attributs des balises, attributes, et le code XML engendré (XML code). Dans la zone text, la phrase présentée plus haut est affichée de la façon suivante :

A low level of GerE activated transcription of CotD by GerE RNA polymerase but in vitro

Actuellement; l'éditeur est utilisé par plusieurs partenaires du projet Caderige, ainsi que par SIB (Swiss Institute of BioInformatics), dans le cadre du projet européen BioMint, qui utilise une autre DTD. Plusieurs corpus portant sur plusieurs espèces ont déjà été annotés avec cet outil, principalement par les biologistes de l'INRA.

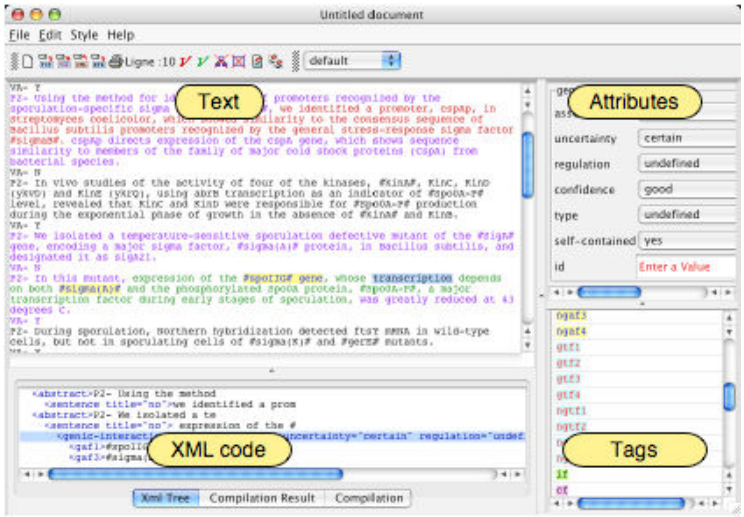

Figure 3**: L'éditeur d'annotation de Caderige**

5.3 Méthodes d'apprentissage de patrons d'extraction

La majorité des approches sont fondées sur des patrons d'extraction écrits manuellement, appliqués à des représentations superficielles des phrases (par exemple les expressions régulières de Ono et al., 2001) dont la bonne précision est au détriment du rappel. Dans Caderige, les méthodes d'analyse profonde augmentent la complexité de la représentation des phrases, et ainsi des patrons d'extractions. Les méthodes d'apprentissage apparaissent alors très intéressantes pour automatiser le processus d'acquisition des patrons (Freitag, 1998; Califf et al., 1998; Craven et al., 1999).

L'apprentissage de patrons est vu ici comme une tâche de classification où le concept à apprendre est une relation n-aire entre les différents arguments correspondant aux champs du formulaire à remplir. Par exemple; le formulaire présenté à la figure 2 peut être rempli en apprenant la relation ternaire interaction-génique(X,Y,Z), où X,Y et Z sont respectivement le type, l'agent et la cible de l'interaction. Un ensemble d'exemples et de contre-exemples sont construits à partir des phrases annotées et normalisées pour l'apprentissage. Dans Caderige, nous utilisons le système d'apprentissage relationnel Propal (Alphonse et al., 2000). L'intérêt des méthodes relationnelles est qu'elles permettent de représenter naturellement la structure relationnelle des dépendances syntaxiques des phrases normalisées, ainsi que les connaissances du domaine éventuelles telles que les relations sémantiques.

A titre d'exemple, les patrons d'extraction appris par Propal permettent à partir de la phrase suivante : "In this mutant, expression of the spoIIG gene, whose transcription depends on both sigA and the phosphorylated Spo0A protein, Spo0AP, a major transcription factor during early stages of sporulation, was greatly reduced at 43 degrees C.", d'extraire correctement les deux relations interaction-génique(positive, sigA, spoIIG) et interaction-génique(positive, Spo0AP, spoIIG). Comme expérimentation préliminaire, nous avons sélectionné un sous-ensemble de phrases similaires à celle-ci pour l'apprentissage. La performance des patrons évaluée par 10 validation-croisée est de 69±6.5% pour le rappel, et de 86±3.2% pour la précision. Ce résultat préliminaire est très encourageant, montrant que la normalisation permet d'obtenir une bonne représentation pour l'apprentissage de patrons d'extraction ayant à la fois un fort rappel et une forte précision.

6. Conclusion

Nous avons présenté dans cet article quelques résultats du projet Caderige. Les deux développements majeurs sont la définition d'un éditeur d'annotation pour des experts du domaine, et la mise au point de méthodes de TALN spécialisées pour le domaine de la biologie.

Les développements actuels se penchent sur l'utilisation de méthodes d'apprentissage dans le processus d'extraction à différents niveaux de l'architecture : une première utilisation est l'acquisition des ressources linguistiques spécialisées ; une deuxième est l'adaptation dynamique des modules déjà existants durant l'analyse, à partir des caractéristiques du texte traité.

7. Bibliographie

Composées en Times 9 romain, interligné 11 points, les références sont rassemblées en fin d'article par ordre alphabétique, espacées les unes des autres de 6 points. Leur référence est du type (Kolski, 1997) pour un auteur, (Kolski *et al.,* 1998) pour plusieurs auteurs. Elles sont justifiées avec un alinéa négatif de 5 mm (Format > Paragraphe > Retrait de 1re ligne Négatif de 0,5 cm).

– Pour les ouvrages : titre en italique, le reste en romain.

– Pour les revues et actes de conférences publiés : titre de la revue ou de la conférence en italique, le reste en romain.

– Pour les rapports internes et les thèses : texte tout en romain.

Voici, en guise d'exemple, quelques cas de figures parmi les plus courants :

Kolski C., *Interfaces homme-machine*, Paris, Editions Hermès, 1997.

E. Alphonse et C. Rouveirol (2000). Lazy propositionalisation for relational learning. In Horn W. (ed.). *14th European Conference on Artificial Intelligence (ECAI'2000)*, Berlin, Allemagne, pp. 256-260, IOS Press.

E. Agichtein et H. Yu (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, vol. 19 Suppl.1.

S. Aubin (2003). Évaluation comparative de deux analyseurs produisant des relations syntaxiques. *In workshop TALN and multilinguisme*. Batz-sur-Mer.

Y. Akane, Y. Tateisi, Y. Miyao et J. Tsujii. (2001). Event extraction from biomedical papers using a full parser. In *Proceedings of the sixth Pacific Symposium on Biocomputing (PSB 2001)*. Hawaii, U.S.A.. pp. 408-419.

G. Bisson, C. Nedellec, L. Cañamero 2000. Designing clustering methods for ontology building: The Mo'K workbench. In *proceedings of Ontology Learning workshop (ECAI 2000)*, Berlin, 22 août 2000.

M. E. Califf, 1998. Relational Learning Techniques for Natural Language Extraction. Ph.D. Dissertation, Computer Science Department, University of Texas, Austin, TX. AI Technical Report 98-276.

N. Collier, Hyun Seok Park, Norihiro Ogata, Yuka Tateisi, Chikashi Nobata, Takeshi Sekimizu, Hisao Imai et Jun'ichi Tsujii. (1999). The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the European Association for Computational Linguistics* (EACL 1999).

M. Craven et J. Kumlien, 1999. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. *Proc. ISMB* 1999: 77-86

D. Faure et C. Nedellec (1999). Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: the system ASIUM. In *Proc. EKAW'99*, pp. 329-334, Springer-Verlag.

D. Freitag, 1998, Multistrategy learning for information extraction. In *Proceedings of the Fifteenth International Conference on Machine Learning, 161-169*. Madison, WI: Morgan Kaufmann

K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi (1998). Toward information extraction : identifying protein names from biological papers. In *Proceedings of the Pacific Symposium of Biocomputing*, pp. 707-718.

T. Hamon et S. Aubin (2004). Evaluating terminological resource coverage for relevant sentence selection and semantic class building. In *BioNLP Coling 2004 Workshop*.

Z. Harris (1951). *Methods in Structural Linguistics*. Chicago. University of Chicago Press.

E.M. Marcotte, I. Xenarios I., et D. Eisenberg (2001). Mining litterature for protein-protein interactions. In *Bioinformatics*, vol. 17, n° 4, pp. 359-363.

MUC (1995). *Proceeding of the 6th Message understanding Conference*. Morgan Kaufmann. Palo Alto.

C. Nédellec, M. Ould Abdel Vetah et P. Bessières (2001). Sentence Filtering for Information Extraction in Genomics: A Classification Problem. In *Proceedings of the International Conference on Practical Knowledge Discovery in Databases (PKDD'2001)*, pp. 326–338. Springer Verlag, LNAI 2167, Freiburg.

T. Ohta, Yuka Tateisi, Jin-Dong Kim, Hideki Mima et Jun'ichi Tsujii. (2001). Ontology Based Corpus Annotation and Tools. In *Proceedings of the 12th Genome Informatics 2001*. pp. 469-470.

T. Ono, H. Hishigaki, A. Tanigami, T. Takagi (2001) Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*. 17(2): 155-161.

Park JC, Kim HS, Kim JJ (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Proceedings of PSB'2001*.

B. Prasad et A. Sarkar (2000) Comparing Test-suite based evaluation and Corpus-based evaluation of a wide-coverage grammar for English. In the *Proceedings of the workshop on 'Using Evaluation within Human Language Technology'*. LREC. Athens.

D. Proux, F. Rechenmann, L. Julliard, V. Pillet, B. Jacq (1998). Detecting gene symbols and names in biological texts : a first step toward pertinent information extraction. In Genome *Informatics*, vol. 9, pp. 72-80.

G. Schneider (1998). A Linguistic Comparison of Constituency, Dependency and Link Grammar. PhD thesis, Institut für Informatik der Universität Zürich, Switzerland.

D. Sleator et D. Temperley (1993). Parsing English with a Link Grammar. In *Third International Workshop on Parsing Technologies*. Tilburg. Netherlands.

A. Yakushiji, Y. Tateisi, Y. Miyao Y, J-I Tsujii, (2001). Extraction from biomedical papers using a full parser. In the *Proceedings of PSB'2001*.