Reconnaître les fragments de phrases pertinents pour l'extraction d'information dans les textes de génomique, un problème de classification

Claire Nédellec[*], Mohamed Ould Abdel Vetah[*+], Philippe Beissières[#], Christine Brun[°], Bernard Jacq[°]

* LRI UMR 8623 CNRS, Université Paris-Sud, cn/ould@lri.fr
+ ValiGen SA, Paris-La-Défense
# Mathématique, Informatique et Génome(MIG) INRA, philb@inra.fr
° LGPD-IBDM, Marseille, brun/jacq@lgpd.univ-mrs.fr

Résumé

Dans de nombreux domaines tels que celui de la génomique fonctionnelle, l'extraction d'information à partir de textes nécessite la mise en œuvre d'une analyse syntaxico-sémantique. Cette analyse est relativement coûteuse en temps et l'information recherchée dispersée. Une sélection rapide de fragments de phrases pertinents et basée sur des indices superficiels, préalablement à cette analyse permet de rendre l'extraction textuelle plus efficace. Nous présentons dans cet article les résultats obtenus en appliquant les méthodes de classification basées sur un Bayésien Naïf (BN), IVI et C4.5 au problème de sélection de fragments pertinents dans le domaine de la génomique fonctionnelle. Nous avons montré que IVI et BN avec élagage des attributs permettent d'obtenir les meilleurs résultats sur les corpus étudiés par rapport à l'application des mêmes méthodes sans présélection des attributs et comparativement à C4.5 après sélection des attributs.

Mots-clés : Classification, extraction d'information dans les textes, filtrage, génomique fonctionnelle.

1 INTRODUCTION

L'augmentation de la quantité d'information disponible sous forme de documents électroniques écrits en langue naturelle rend pressant le besoin de processus intelligents pour traiter de tels textes. Les méthodes de compréhension superficielle de textes comme l'Extraction d'Information (EI) apparaissent comme particulièrement attrayantes et utiles. L'EI a été définie restrictivement par le programme DARPA MUC (Message Understanding Conference, 92-98), comme la tâche consistant à extraire des informations spécifiques et bien définies à partir de textes écrits en langue naturelle dans des domaines restreints, avec l'objectif spécifique de remplir automatiquement des formulaires prédéfinis ou des bases de

données. Dans de nombreux domaines, les systèmes d'EI doivent reposer sur des méthodes d'analyse profondes qui sont locales aux fragments pertinents des textes et qui combinent l'analyse sémantico-conceptuelle des systèmes de compréhension automatique de textes avec l'extraction d'information par reconnaissance d'expressions régulières (Nédellec & Nazarenko, 2001). Dans une première étape, les fragments de texte sont extraits sur la base d'indices de surface. Dans une deuxième étape, une représentation conceptuelle du contenu sémantique des fragments de texte est 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 (Poibeau & Nazarenko, 99). L'interprétation résultante est finalement soumise, dans une troisième étape, 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 la forme appropriée, classiquement en remplissant un formulaire. Ces trois étapes diffèrent à la fois par les connaissances auxquelles elles font appel et par la complexité des méthodes mises en œuvre. La deuxième étape d'analyse syntaxico-sémantique est la plus coûteuse en temps et en moyens. Ce coût rend nécessaire la première étape de sélection préalable des fragments pertinents de manière à réduire l'analyse au strict nécessaire en la concentrant sur les fragments potentiellement porteurs d'information. Cette sélection est d'autant plus cruciale que les informations recherchées sont noyées dans une masse d'informations non pertinentes au regard de la tâche d'extraction. Ce problème de la dispersion de l'information a été noté par des recherches précédentes en EI, par exemple dans (Riloff, 93) et (Soderland, 99). La conséquence en est que la première étape de sélection des fragments pertinents doit être rapide, au prix éventuellement d'une perte de précision. Elle est donc basée sur des indices de surface du texte à traiter.

L'apprentissage pour la sélection de fragments pertinents a reçu relativement peu d'attention en EI, comparativement à l'apprentissage pour l'identification d'entités nommées (Bikel *et al*., 97) et à l'apprentissage de règles d'extraction (Riloff, 93) et (Soderland, 99). Ce manque d'intérêt s'explique par la nature des textes généralement traités par les recherches en EI, qui sont ceux proposés par les tâches des compétitions MUC. L'information recherchée est relativement dense dans des textes courts, rendant moins nécessaire un filtrage préalable. Le type d'information à extraire, tel que par exemple, repérer un nom de société, ne nécessite souvent qu'une analyse de surface dont le faible coût rend l'extraction applicable sans présélection. Ce n'est pas le cas dans de nombreuses tâches telles que celle que nous décrivons ici, en génomique fonctionnelle.

Le problème à résoudre du point de vue de l'apprentissage consiste donc à concevoir une méthode capable de repérer rapidement les zones de texte pertinentes dans des masses de documents avec une bonne couverture mais une précision moins exigeante. Ce problème peut être vu comme un problème de classification de zones de texte en deux classes, pertinente et non pertinente. Les exemples d'apprentissage représentent des fragments (des phrases dans cette application) et les attributs des exemples sont les mots significatifs des phrases, lemmatisés (sous forme canonique). La grande dispersion des exemples dans l'espace des attributs est une conséquence majeure de cette représentation puisque chaque exemple est décrit par peu d'attributs. Nous avons comparé expérimentalement une méthode de classification appelée IVI, pour Indice de Vraisemblance d'Interaction

et proposée par (Pillet, 2000) dans le cadre de la génomique, une méthode "Bayésien Naïf" (notée BN) (Mitchell, 97) et une méthode à base d'arbre de décision, C4.5, (Quinlan, 92). Notre étude, décrite au paragraphe 2, a été menée sur trois corpus de texte concernant des organimes biologiques différents, dans le même domaine de la génomique fonctionnelle. En plus des résultats de ces méthodes, nous avons étudié les effets de la sélection d'attribut comme étape de pré-apprentissage. L'objectif de cette étude est d'identifier la meilleure des méthodes de classification pour filtrer les phrases en génomique fonctionnelle et pour caractériser les corpus par rapport aux méthodes. Cet article rapporte les résultats des comparaisons de ces méthodes de classification. Les détails des méthodes de classification et des conditions d'évaluation sont décrits au paragraphe 3. Le paragraphe 4 rapporte et commente les résultats expérimentaux. Le paragraphe 5 débat des perspectives de ce travail.

2 LE DOMAINE D'APPLICATION : LA GENOMIQUE FONCTIONNELLE

2.1 Problème du point de vue génomique

Le problème applicatif traité ici est celui de la modélisation des interactions géniques à partir de textes, dans le domaine de la génomique fonctionnelle. Ce problème a été décrit dans (Blaschke et al., 1999), (Pillet, 2000), et (Thomas et al., 2000). 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 les méthodes seront transposables et exploitables dans l'ensemble du contexte de l'extraction de connaissance à partir de documentation scientifique et technique.

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. En effet, après la prédiction de l'ensemble des gènes contenus dans le génome d'une espèce donnée à partir de l'analyse de sa séquence, la compréhension des interactions que ces gènes peuvent établir entre eux est essentielle. Bien que les interactions entre gènes aient été étudiées depuis des décennies dans certains cas, la majeure partie de la connaissance biologique sur les interactions n'est pas décrite aujourd'hui 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, à l'échelle globale de l'organisme étudié, comme les puces à ADN, et aujourd'hui, une équipe de recherche équipée 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, afin de relier ces données nouvelles issues du laboratoire à la littérature scientifique, pour les interpréter et leur donner un sens.

Dans le cadre de notre étude, la spécification des besoins biologiques et la validation des outils et des résultats d'apprentissage ont été réalisées principalement en collaboration avec le laboratoire MIG 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, 98). 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. Les nombreuses données déjà connues et disponibles sur la transcription chez cette bactérie modèle vont permettre dans un premier temps de juger de la pertinence de l'extraction de l'information, mais à terme, ce sont l'identification de l'ensemble des interactions et des régulations moléculaires qui sont visées. 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 drosophile d'une part (en collaboration avec le LGPD-IBDM). et la souris et l'homme, (avec la société ValiGen et le LGPD-IBDM) d'autre part. Deux bases bibliographiques, MedLine, la principale base du domaine biologique (elle rassemble plus de 16 millions de résumés d'articles biologiques scientifiques et médicaux), et FlyBase une base spécialisée dans les gènes de drosophile ont été exploitées. Leur exploration repose sur des requêtes via Internet, qui portent classiquement 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 résumés d'articles effectivement pertinents, de l'ordre de quelques centaines. Par exemple, la requête "*Bacillus subtilis transcription*" en sélectionne quelque 2209. La Figure 1 présente un extrait d'un tel résumé comportant une information d'interaction entre une protéine et un gène. Il reste à sélectionner les zones pertinentes pour cette interaction, en gras dans l'exemple, puis à en extraire les connaissances utiles relatives aux interactions entre gènes, et à les enregistrer de manière structurée, de telle sorte qu'un biologiste puisse obtenir des réponses à une requête spécifique, (voir l'exemple de formulaire pour la phrase, Figure 2.).

```
UI  - 99175219 [...]
AB  - [...] Most cot genes, and the gerE gene, are transcribed by
sigmaK RNA polymerase. Previously, it was shown that the
GerE protein inhibits transcription in vitro of the
sigK gene encoding sigmaK. Here, we show that GerE binds
near the sigK transcriptional start site, to act as a repressor.
[...]
```

FIG. 1 – *Exemple d'extrait de résumé de MedLine concernant Bacillus subtilis.*

Interaction **Type :** négatif
Agent : GerE protein
Cible : **Expression** **Source :** sigK gene
Produit : sigmaK protein

FIG. 2 – *Exemple de formulaire d'interaction génique, rempli.*

Ce domaine est bien représentatif du champ de notre étude sur le filtrage de fragments pertinents pour l'extraction d'information : l'information est locale, principalement localisée dans des phrases isolées ou des portions de phrase. Elle

est très dispersée dans les documents. Par exemple, dans les 2209 résumés de *Bacillus subtilis*, mentionnés ci-dessus, seules de l'ordre de 3 % des phrases contiennent de l'information sur les interactions géniques. Or, l'extraction d'information doit reposer sur une analyse profonde. Les techniques basées sur des indices superficiels – techniques issues de l'extraction d'information telles que des transducteurs définis manuellement et basés sur des verbes significatifs et des noms de gènes (Blaschke et al., 1999), (Thomas et al., 2000), (Poibeau, 2001), (Ono *et al*., 2001) – ou des mesures statistiques de cooccurrences de mots-clefs (Stapley & Benoit, 2000), (Pillet, 2000) – techniques issues de la recherche documentaire et de l'analyse des données – n'obtiennent en effet que des résultats limités dans le cadre de la recherche d'interactions. Prenons un exemple illustratif de certains des problèmes rencontrés.

"**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 encodes sigma K. ".

Les méthodes d'EI basés sur des mots clefs, des noms de gène (en gras dans l'exemple) et des verbes d'interaction (encadrés dans l'exemple), ne sont pas capables d'identifier l'interaction inhibitrice entre GerE et sigK gene transcription (28 mots plus loin) ou, si elle l'identifie, identifie aussi par erreur l'interaction entre cotD et sigK et entre cotA et sigK 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 lexicaux, syntaxiques et sémantiques[1]. Les caractéristiques de cette application en font donc un exemple tout à fait approprié à notre étude des méthodes d'apprentissage pour la sélection de fragments de texte pertinents.

2.2 Corpus textuels et exemples d'apprentissage

Nous avons cherché à évaluer la robustesse des méthodes de classification destinées à sélectionner les fragments de texte pertinents en fonction des styles rédactionnels et en fonction des espèces et donc des modèles d'interaction géniques. Les méthodes de classification ont été appliquées, évaluées et comparées à l'aide de trois ensembles d'apprentissage. Les résumés dont ils sont issus concernent des espèces différentes. Le premier ensemble noté *Dro* concerne une mouche, *Drosophila melanogaster*[2], le second noté *Bs*, concerne une bactérie, *Bacillus subtilis*[3] et le dernier noté *HM*[4], la souris et l'homme. Ils sont issus de bases documentaires différentes, avec des traditions de rédactions différentes : l'ensemble Dro provient de la base de donnée spécialisée sur *Drosophila*, FlyBase, dont les résumés très concis contiennent 2 à 3 phrases. Les deux autres proviennent de MedLine, la base de données bibliographique généraliste dont les résumés plus

[1] Thème du projet *Caderige* dont cette étude est un sous-ensemble (Nédellec & Nazarenko, 2001).

[2] La base Dro a été fournie telle que par B. Jacq et V. Pillet du LGPD-IBDM (Pillet, 2000).

[3] Cet ensemble a été construit en étroite collaboration avec P. Beissières (MIG, INRA) dans le cadre du projet Caderige).

[4] It a été fourni par le LGPD-IBDM et la société ValiGen.

longs contiennent une dizaine de phrases dans une forme syntaxique plus complexe. Les résumés dont sont issus les exemples d'apprentissage ont été sélectionnés par les requêtes "Bacillus subtilis transcription" pour Bs et et "Telomere", "Apoptose", "DNA replication", "DNA repair","cell cycle control", "two-hybrid" et "interaction" pour HM. Dans le cas de Dro, tous les résumés de la base référençant des gènes ont été pris en compte.

Les exemples d'apprentissage de Bs et HM ont ensuite été sélectionnés dans les résumés. Nous avons fait pour tous les cas, la même hypothèse de localité de l'information au niveau de la phrase suivant en cela (Soderland, 99) et (Pillet, 2000). Pour Bs et HM, nous avons supposé que les phrases potentiellement pertinentes contenaient au moins deux noms de gènes ou de protéines. Pour Dro, les phrases potentiellement pertinentes contiennent exactement deux noms de gènes ou de protéines. Cette nuance ne doit pas avoir de conséquence sur la classification, mais sur l'extraction, uniquement. L'identification des noms de gènes pour la présélection des phrases de HM a été faite manuellement par les biologistes du LGPD-IBDM. La sélection pour Bs a été faite automatiquement dans le cadre de notre étude à l'aide d'une liste de noms de gènes et de protéines de *Bacillus subtilis* et de leurs variantes lexicales fournies par MIG et complétée manuellement. Le problème de la reconnaissance des noms de gènes dans les textes est un problème qui a été étudié depuis peu comme préalable à tout traitement textuel automatique, en raison du manque de standardisation dans la nomenclature et de stabilité dans les notations (Fukuda *et al*., 98), (Proux *et al*., 98), (Collier *et al*., 2000) et (Humphreys *et al.*, 2000).

Les exemples d'apprentissage ont été décrits à partir des phrases des trois ensembles au moyen des mêmes attributs, à savoir, leurs mots significatifs et lemmatisés à l'aide du "shallow parser" de Xerox. Les mots "vides" tels que les déterminants ont été supprimés comme non significatifs à l'aide de la liste de 620 mots fournie par Patrice Bonhomme (LORIA) et révisée par nos soins en fonction de la tâche. Par exemple, le mot *act* qui appartenait à cette liste a été conservé comme potentiellement discriminant. Les attributs décrivant les exemples sont booléens dans le cas de C4.5 (présence - absence du mot), et ils représentent le nombre d'occurrence dans les phrases, pour les autres méthodes. Les exemples ont ensuite été classés par les biologistes en deux catégories, positives et négatives, comme décrivant ou non, *au moins une interaction génique*. La Figure 3 présente un exemple de phrase et d'exemple d'apprentissage du corpus Bs. Le tableau de la Figure 4 résume les caractéristiques des ensembles d'apprentissage.

Phrase : In addition, 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 encodes sigma K.
Ex : addition stimulate transcription inhibit transcription vitro RNA polymerase expected vivo study unexpectedly profoundly inhibit vitro transcription gene encode
Classe : Positif

FIG. 3 – *Exemples d'apprentissage du corpus Bs.*

Comme illustré par la Figure 5 pour le corpus Dro, les données des trois ensembles sont très dispersées dans l'espace des descripteurs, phénomène bien connu dans ce type d'application textuelle (Yang & Pedersen, 97).

	Dro	Bs	HM
Base documentaire	FlyBase	MedLine	
Nombre de références bibliographiques	> 100 000	Env. 16 Millions	
Nombre moyen de phrases par résumé	2, 3	une dizaine	
Nombre de résumés sélectionnés	530	2209	105
Nombre de phrases dans les résumés	5 244	Env. 20 000	962
Nombre d'attributs	1701	2340	1789
Nb total d'ex. (phr. avec 2 noms de gènes)	1197	932	407
Nombre d'exemples positifs	653	470	240
Nombre d'exemples négatifs	544	462	167

FIG. 4 – *Caractéristiques des ensembles d'apprentissage.*

La capacité des méthodes à traiter la dispersion est donc un critère important pour choisir la méthode de classification.

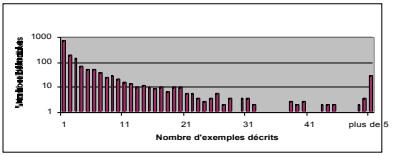

FIG. 5 – *Distribution des attributs de Dro, en fonction des nombres d'exemple.*

3 LES METHODES DE CLASSIFICATION

3.1 Description des méthodes

La méthode de classification *IVI* proposée par (Pillet, 2000) et précédemment appliquée au corpus Dro, repose une mesure de poids des attributs calculée par la formule Figure 5a et sur une mesure de poids des exemples, calculée par la formule Figure 5b. La classe de l'exemple est déterminée en fonction d'un seuil fixé à 0 expérimentalement et tel que les phrases dont les poids sont supérieurs au seuil sont classées positives et inversement.

$$(a)\ Poids(Att_i) = \frac{Nbocc(Att_i, ExPos) - Nbocc(Att_i, ExNég)}{Nbocc(Att_i, Ex)} \qquad (b)\ IVI(Ex) = \sum_{i=1}^{Nbatt(Ex)} Poids(Att_i)$$

FIG. 5 – *Mesures des poids des attributs et des exemples par IVI.*

La méthode Bayésien Naïf (BN) telle que définie par (Mitchell, 97) apparaît adaptée à notre problème en raison de sa capacité à traiter la dispersion des données. L'hypothèse selon laquelle les attributs devraient être indépendants n'est évidemment pas vérifiée ici, mais d'autres auteurs avant nous avaient montré d'étonnantes bonnes performances de BN en dépit de cette contrainte

(Domingos & Pazzani, 96). BN consiste à estimer, pour chaque attribut, la probabilité de décrire un exemple positif et la probabilité de décrire un exemple négatif, ceci en fonction du nombre d'occurrences des attributs observé dans l'ensemble d'apprentissage. La probabilité pour un exemple d'appartenir à une classe est estimée par le produit des probabilités pour cette classe, des attributs qui décrivent l'exemple (Figure 7b). L'exemple est affecté à la classe pour laquelle la probabilité est la plus élevée, Figure 7c. L'estimateur utilisé ici est basé sur la loi de Laplace (Figure 7a) qui donne de meilleur résultat que l'estimateur brut dans le cas qui est le nôtre, (Nédellec & Ould Abdel Vetah, 2001), à cause de sa capacité de lissage dans le cas de données dispersées (à estimer la probabilité d'un mot qui n'apparaît pas dans une phrase comme non nulle).

(a) $$\Pr(Att_j|Cl_i) = \frac{NbOcc(Att_j, Ex \quad Cl_i)}{\sum_{k=1}^{NbClasses} NbOcc(Att_j, Ex \quad Cl_k) + NbClasses}$$

(b) $$\Pr(Ex|Cl_i) = \prod_{j=1}^{NbAtt(Ex)} \Pr(Att_j|Cl_i)$$ (c) $$Classe(Ex) = Max_{i\in[1..NbClasse]} \Pr(Ex|Cl_i)$$

FIG. 7 – *Estimation des probabilités par BN pour les attributs et pour les exemples.*

La troisième méthode mise en œuvre est C4.5 avec C4.5Rules. L'intérêt de C4.5 ici par rapport à BN est que l'analyse des arbres de décision ou des règles permet d'observer l'effet de conjonction d'attributs sur la classification et d'espérer en tirer des conclusions sur les mots reliés par des dépendances syntaxiques et qui seraient potentiellement porteurs d'information pour l'extraction d'information ultérieure. Par contre, la faible densité des exemples est un handicap potentiellement élevé pour C4.5.

La sélection d'attribut apparaît comme un bon moyen de réduire le nombre d'attributs en conservant les attributs les plus pertinents pour améliorer la classification (Yang & Pedersen, 97), mais aussi pour sélectionner le corpus approprié pour d'autres tâches d'apprentissage tel que l'apprentissage de classes sémantiques (§ 5). Ce but a motivé le choix d'une méthode de filtrage des attributs par opposition à une méthode à base de "wrapper". La mesure de la pertinence des attributs est basée sur la formule de la Figure 8. Elle mesure la capacité de l'attribut à caractériser une classe. Les attributs sont ordonnés par rapport à cette mesure et les meilleurs d'entre eux sont sélectionnés pour décrire les ensembles d'apprentissage pour la classification.

$$\text{Précision}(Att) = \sum_{i=1}^{NbClasses} Max\{\Pr(Att, Cl_i), 1 - \Pr(Att, Cl_i)\}$$

FIG. 8 – *Mesure de la pertinence des attributs à des fins de sélection d'attributs.*

3.2 Mesures d'évaluation

Les méthodes ont été évaluées et comparées en fonction des critères classiques de mesure des taux de rappel, ou couverture, des taux de précision, ou correction, et de la F-mesure, obtenus sur les différents jeux de données (Figure 9), ceci globalement pour les deux classes et pour la classe des Positifs.

$$\text{Rappel}(\text{Classe}_i) = \frac{|\text{Ex} \quad \text{Classe}_i \text{ et affecté à Classe}_i|}{|\text{Ex} \quad \text{Classe}_i|}$$

$$\text{Précision}(\text{Classe}_i) = \frac{|\text{Ex} \quad \text{Classe}_i \text{ et affecté à Classe}_i|}{|\text{Ex affecté à Classe}_i|}$$

$$F = \frac{(\beta^2 + 1) * \text{Précision} * \text{Rappel}}{(\beta^2 * \text{Précision}) + \text{Rappel}}$$

FIG. 9 – *Mesures de rappel, de précision et F-mesure pour la classification.*

Ce sont en effet les exemples classés comme Positif qui feront ensuite l'objet d'extraction d'information et pour lesquels le rappel devrait donc être le plus élevé possible, même au prix d'une perte de précision. En conséquence, le facteur β qui permet de pondérer le rappel et la précision dans la F-mesure a été fixé à 1.65 de manière à favoriser légèrement le rappel par rapport à la précision.

BN et IVI ont été évalués par "leave-one-out" pour chaque jeu de donnée de manière à obtenir une évaluation proche des conditions réelles. C4.5 et C4.5Rules ont été évalués sur des ensembles de test formant 10 % des ensembles initiaux, les 90 % restant formant l'ensemble d'apprentissage. Les résultats présentés ici sont obtenus par la moyenne des évaluations pour 10 tirages indépendants. Ce type d'évaluation pour C4.5 a dû être choisi pour des raisons de performances. Nous avons cependant évalué BN sur la même base et les résultats obtenus se sont avérés comparables. Les erreurs sont calculées en utilisant les formules classiques de calcul d'erreur (Mitchell 1997).

4 ÉVALUATION

4.1 Comparaison des méthodes IVI, C4.5 et BN

La première expérience a consisté à comparer C4.5, C4.5Rules et BN sur les trois corpus, du point de vue de la seule classe Positif et pour les 2 classes (Figure 10).

Corpus	Dro				Bs				HM			
Méthode	C4.5	C4.5 R	BN	IVI	C4.5	C4.5 R	BN	IVI	C4.5	C4.5 R	BN	IVI
Rappel Pos	**88,9** ±2.4	86,8 ±2.6	75,3 ±2.9	69,1 ±3.5	63,9 ±4.3	71,4 ±4.1	**85,7** ±3.2	82,6 ±3.4	88,3 ±4.1	84,5 ±4.1	**97,1** ±2.1	90 ±3.8
Précision Pos	68,1 ±3.6	70,5 ±3.5	82 ±3.2	**83,1** ±2.8	63,4 ±4.3	62,8 ±4.4	66,6 ±4.3	**67,4** ±4.2	63,7 ±6.1	64,2 ±6.1	68,5 ±5.9	**70,3** ±5.8
Rap.-préc. tte classe	72 ±2.5	73.6 ±2.5	**77,5** ±2.4	75,4 ±2.4	62,4 ±3.1	62,9 ±3.1	**71,1** ±2.9	71 ±2.9	63,7 ±4.1	63,4 ±4.7	**72** ±4.4	71,5 ±4.4

FIG. 10 – *Comparaison de C4.5, C4.5Rules, BN et IVI sur les trois corpus*

Pour les trois corpus, les résultats de BN et IVI sont meilleurs que ceux de C4.5 et C4.5Rules. Ce succès peut s'expliquer par la difficulté pour C4.5 de traiter des ensembles d'exemples peu denses et hétérogènes. La précision globale est de 5 à 8 % meilleure et celle pour la classe Positif est de 4 à 12 % meilleure. Les résultats sont plus contrastés pour le rappel des positifs, où le taux de C4.5 est meilleur que la famille BN-IVI pour Dro (13 %), mais moins bon pour Bs et

HM (-12 à -13 %). L'origine du corpus Dro peut expliquer ce comportement. Il provient de la base bibliographique FlyBase où les phrases sont plus courtes et le vocabulaire moins riche que celui de MedLine, d'où sont issus HM et Bs, car les résumés de FlyBase sont réécrits par trois ou quatre annotateurs seulement. Ainsi les phrases de Dro sont décrites proportionnellement par moins d'attributs que les autres corpus. Ceci peut expliquer la surgénéralisation faite par C4.5 sur le corpus, illustrée par un taux de rappel élevé et une mauvaise précision. L'analyse des résultats de BN et IVI montre que BN a des performances légèrement meilleures à un niveau global que IVI. Cependant, leur comportement sur la classe Positif est très différent : BN a un rappel plus élevé que celui de IVI (3 à 7%) tandis que la précision de IVI est meilleure que celle de BN (1 à 2%) mais la différence n'est pas significative. Les bons résultats obtenus sur HM sont expliqués par la façon dont a été construit ce corpus. La sélection des phrases dans les résumés a été faite manuellement par les biologistes parmi un grand nombre de phrases candidates (Figure 4). Ce choix peut expliquer l'homogénéité de ce corpus comparé à Bs où cette sélection a été faite automatiquement. Un meilleur rappel étant préférable pour notre application, la conclusion de ces expériences serait de préconiser BN pour les données issues de MedLine (Bs et HM). Pour FlyBase (Dro), ce choix va dépendre de la capacité du module d'EI à traiter les phrases qui ont été filtrées avec une précision moyenne. C4.5 pourra être choisie pour son meilleur rappel tandis que BN pourra être choisie pour son meilleur compromis rappel-précision.

4.2 Élagage des attributs

Comme décrit au paragraphe 3, les attributs de chaque ensemble d'apprentissage ont été ordonnés en fonction de leur pertinence. Par exemple, les meilleurs attributs de Dro sont, `downstream`, `interact`, `modulate`, `autoregulate`, et `eliminate`. Nous avons étudié l'influence de l'élagage des attributs sur les deux méthodes, C4.5Rules et BN, en faisant varier le nombre d'attributs sélectionnés de 100 en 100, de 100 au nombre maximum d'attributs.

4.2.1 Effet de l'élagage des attributs sur BN

Pour les trois corpus, la croissance du rappel et la décroissance de la précision sont notables (ce qui était attendu). La F-mesure croît sur le premier quart, se stabilise plus ou moins sur un plateau sur la moitié, éventuellement en montant légèrement (ce qui est dû à la prédominance du rappel sur la précision dans notre choix des paramètres de la F-mesure), puis redescend pour le dernier quart ou cinquième, après un petit pic dans le cas de Dro et Bs. Les points idéaux du point de vue du compromis rappel-précision mesuré par la mesure F sont donc à la fin du plateau autour de 3/4 à 4/5 du nombre total d'attributs. Pour Dro, ce point est autour de 1400. Dans le cas de Bs, il se situe aux alentours de 1800 pour les positifs, 1900 pour l'ensemble. On note que le rappel pour les positifs est plus élevé de 10 à 15 % que le rappel général, et c'est le contraire pour la précision, plus élevée pour les négatifs, ce qui est souhaitable (Figure 14) Pour HM, ce phénomène est encore plus accentué, le rappel pour les positifs est très élevé, presque 100 %, et plus élevé de 20 % que le rappel global (Figure 13). Comparé aux autres, le plateau est plus horizontal entre 400 et 1900, après une très faible

remontée entre 400 et 800 et ne présente pas de pic avant la décroissance. Cette fois donc, le rappel et la précision étant stables entre 800 et 1400, tous les points se valent. Cela pourrait être expliqué par l'homogénéité du corpus HM.

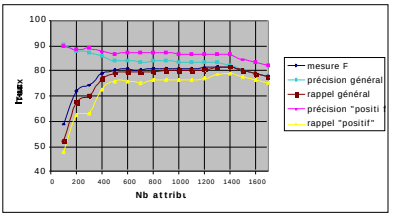

FIG. 11 – *Effet de l'élagage des attributs sur BN pour le corpus Dro*

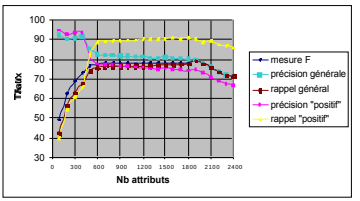

FIG. 12 – *Effet de l'élagage des attributs sur BN pour le corpus Bs*

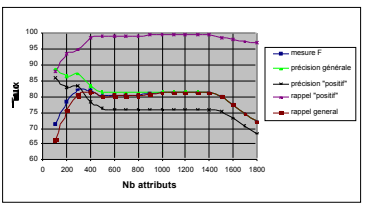

FIG. 13 – *Effet de l'élagage des attributs sur BN pour le corpus Bs*

La Figure 14 présente les résultats obtenus avec BN pour les meilleurs élagages des attributs d'une part, et pour tous les attributs. Ils sont clairement améliorés par l'élagage des attributs. Le gain de l'élagage est considérable pour HM, 10 % environ, moindre pour Bs, 6 à 7 %, et pour Dro de 4-5 %.

Corpus	Dro		Bs		HM	
Nb attributs	Tous 1701	1400	Tous 2340	1800	Tous 1789	900-1300
Rappel Pos	75,3 ± 2.9	**79**±3.1	85,7±3.2	**90,8**±2.6	97,1±2.1	**99,6**±0.8
Précision Pos	82 ±3.2	**86,4**±2.6	66,6±4.3	**74,1**±4.00	68,5±5.9	**76,1**±5.4
Rappel-précision Toute classe	77,5 ±2.4	Rappel : **81,8** ±2.2 Précision **82,1** ± 2.2	71,1±2.9	Rappel **77,5**±2.7 Précision **79,9**±2.6	72±4.4	Rappel **81,1** ±3.8 Précision **81,3** ±3.8

FIG. 14 – *Comparaison entre tous les attributs et le meilleur élagage pour BN.*

4.2. Effet de l'élagage des attributs sur C4.5 et IVI

Le même type d'expérience a été réalisé avec C4.5. Les résultats de ces expérimentations sont présentés dans la Figure 15. Les conclusions sont les mêmes que pour BN: l'élagage des attributs améliore les résultats de la classification pour les trois corpus. Ce gain est considérable pour Bs et HM (9%), mais moindre pour Dro (1.6%) pour les mêmes raisons que celles que nous avions mises en évidence précédemment. Pour la classe des positifs, nous observons le même phénomène qu'avec BN, la sélection des attributs améliore considérablement les taux de rappel (5 à 13 %) mais donne des taux de précision comparables ou moins bons (-2 %).

Corpus	Dro		Bs		HM	
Nb attributs	Tous 1701	1400	Tous 2340	1600	Tous 1789	1300
Rappel Pos	**86,8** ±2.6	84,5 ±2.8	**71,4** ±4.1	70,1 ±4.2	84,5 ±4.6	**84,6** ±4.6
Précision Pos	70,5 ±3.5	**75**±3.33	62,8 ±4.4	**71,4** ±4.13	64,2 ±6.1	**78,8** ±4.6
Rappel-préc. tte classe	73.7 ±2.5	**75,3** ±2.4	62,9± 3.1	**71,1**±3	63,4 ±4.7	**74,9** ±5.2

FIG. 15 – *Comparaison entre tous les attributs et le meilleur élagage pour C4.5Rules.*

Corpus	Dro		Bs		HM	
Nb attributs	Ts 1701	1300	Ts 2340	1900	Ts 1789	1400
Rappel Pos	69 ±3.5	**77,9** ±3.2	82,6 ±3.42	**91,5**±2.5	90 ±3.8	**98,3** ±1.6
Précision Pos	83,6 ±2.9	**88,4** ±2.5	67,4 ±4.23	**78,3**±3.7	70,3 ±5.8	**83,4** ±4.7
Rappel-préc. Tte classe	75,4 ±2.4	Rappel **81,9**±2.2 Préc. **84,1** ±2.1	71 ±2.91	Rappel **82,8**±2.4 Préc. **83,2**±2.4	71,5 ±4.4	Rappel **87,5**±1.6 Préc. **87,5**±4.7

FIG. 16 – *Comparaison entre tous les attributs et le meilleur élagage, pour IVI*

Des expériences similaires ont été réalisées avec IVI (Figure 16). L'amélioration obtenue par l'élagage est plus importante pour IVI que pour les

deux autres méthodes. Le gain obtenu est approximativement de +6% pour Dro, +10% pour Bs et +16% pour HM. Cette amélioration est due à une augmentation du rappel de la classe Positif (+8-9%) comme dans les autres méthodes, mais surtout à une augmentation de la précision de cette classe (5 à 13%).

4.2.3 Conclusion sur l'effet de l'élagage des attributs sur la classification

La comparaison entre les meilleurs résultats de classification de C4.5, IVI et de BN en utilisant l'élagage des attributs, montre que les performances globales de IVI sont meilleures (voir l'annexe pour une tentative d'interprétation). Cependant, pour le rappel de la classe Positif, BN donne des résultats sensiblement meilleurs que IVI (1 à 2%) alors que les taux de précision d'IVI sont meilleurs que ceux de BN (2 à 7 %). Si le choix est de favoriser un meilleur taux de rappel pour la classe Positif, mieux vaut appliquer BN avec la sélection d'attributs, ceci pour tous les corpus exceptés pour ceux qui comme Dro sont plus denses et plus homogènes et pour lesquels C4.5 sans sélection d'attribut est meilleur. Si le meilleur compromis rappel-précision est recherché, IVI avec élagage des attributs est préférable.

5 PERSPECTIVES

Cette étude a porté sur la classification des exemples, des phrases, représentés par les mots lemmatisés et significatifs. Les méthodes étudiées permettent d'obtenir des taux de rappel et de précision de l'ordre de 80 % pour les deux classes et de bien meilleurs taux de rappel pour les positifs avec sélection des attributs. D'autres critères de sélection des attributs devront être testés, tels que le gain d'information et la mesure d'information mutuelle. Nous envisageons maintenant d'appliquer une approche à base de "wrapper" pour la sélection des attributs (John & Kohavi, 97), de telle sorte que les algorithmes de classification seraient appliqués et évalués successivement sur des sous-ensembles d'attributs de manière à identifier le meilleur sous-ensemble (Langley & Sage, 94).

Pour obtenir de meilleurs résultats, des mesures plus globales de gain d'information devraient être envisagées, afin de prendre en compte les dépendances entre les mots formant des expressions langagières significatives. Par exemple, l'étude terminologique en cours au LIPN devrait permettre de réduire le nombre d'attributs tout en prenant en compte cette dépendance. La réduction du nombre d'attributs par le remplacement des mots par leurs concepts ou par les classes sémantiques auxquels ils appartiennent est également une approche prometteuse. L'apprentissage de telles classes avec les systèmes Asium (Nédellec & Faure, 1999) et Mo'K (Bisson & Nédellec, 2001) à partir des corpus de biologie est en cours. Symétriquement, une classification conceptuelle des exemples de manière à regrouper préalablement à la classification les exemples décrits par les mêmes attributs devrait permettre de diminuer l'hétérogénéité des données et d'apprendre des classifieurs adaptés à l'expression d'interactions géniques de différentes natures. En effet, sous le vocable générique "interactions génétiques" sont regroupées des interactions physiques (entre protéines, entre protéines et gènes, entre protéines et

ARN) et des interactions génétiques. Dans le même ordre d'idée, une approche à base de "boosting" focaliserait successivement l'apprentissage de classifieurs sur les exemples mal classés.

Du point de vue de la tâche d'extraction, l'hypothèse simplificatrice mais raisonnable selon laquelle seules les phrases comportant au moins deux noms de gènes ou de protéines décrivent des interactions devrait être levée. Les attributs les plus pertinents selon la mesure proposée devraient permettre d'identifier d'autres phrases potentiellement pertinentes, et ce, de façon automatique. Enfin, l'apprentissage de règles d'extraction d'information passe par l'apprentissage de classes sémantiques. L'identification des attributs, des mots, les plus pertinents pour repérer des interactions permettra de sélectionner les expressions contenant ces mots dans le corpus d'apprentissage des classes sémantiques et de focaliser l'apprentissage sur les concepts potentiellement plus utiles à l'expression de règles d'extraction d'interactions géniques.

Remerciements

Ce travail est financé partiellement par le Ministère de l'Économie, 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.* Le travail du LPPD-IBDM est en partie financé par la société ValiGen.

RÉFÉRENCES

BIKEL D. M., MILLER S., SCHWARTZ R. AND WEISCHEDEL R. (1997). Nymble: a High-performance Learning Name-finder. *Conference on Applied Natural Language Processing*.

BISSON G. ET NEDELLEC C. (2001). Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier Mo'K. In H. Brian Eds. Actes des *Journées Francophones d'Extraction et de Gestion des Connaissances (EGC'2001)*, Hermès (Pub.) Nantes.

BLASCHKE C., ANDRADE M. A., OUZOUNIS C. AND VALENCIA A. (2001). Automatic Extraction of biological information from scientific text: protein-protein interactions. In Proceedings of *International Symposium on Molecular Biology, (ISMB'99)*.

COLLIER N, NOBATA C. AND TSUJII (2000). 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.

CRAVEN M. AND KUMLIEN J.(1999). 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)*.

DOMINGOS P. AND PAZZANI M. (1996). Beyond independence: conditions for the optimality of the simple Bayesian classifier. In proceedings of *ICML'96*, Saitta L. (ed.). p. 105-112.

FAURE D. AND NEDELLEC C. (1999). 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. et Studer R. (Ed.), Springer Verlag, Karlsruhe, Allemagne.

FUKUDA K., TSUNODA T., TAMURA A. AND TAKAGI T. (1998). Toward Information Extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on biocomputing (PSB'1998)*.

HALDENWANG W. G. (1995). The sigma factors of Bacillus subtilis. *Microbiol. Rev*. vol 59, p. 1-30.

HUMPHREYS K., DEMETRIOU G, AND GAIZAUSKAS R. (2000). 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.

JOHN G. AND KOHAVI R. (1997). Wrappers for feature subset selection. In *Artificial Intelligence Journal*.

LANGLEY P. AND SAGE S. (1994). Induction of selective Bayesian classifiers. In proceedings of the *10th conference on UAI*,. Lopez de Mantaras R. (Ed.), p. 399-406, Morgan Kaufmann.

MITCHELL, T. M. (1997). *Machine Learning*, Mac Graw Hill, 1997.

MUC. (1192-1998). Proceedings of the *Message Understanding Conference* (MUC-4-7), Morgan Kaufman, San Mateo, USA.

NÉDELLEC C. ET NAZARENKO A. (2001). Application de l'apprentissage à la recherche et à l'extraction d'information - Un exemple, le projet *Caderige* : identification d'interactions géniques. In Actes de la *Journée thématique Exploration de données issues d'Internet*, Bennani Y., et al. (Eds).

NÉDELLEC C. (2000). 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.

NEDELLEC C. ET OULD ABDEL VETAH M. (2001). Modélisation des interactions géniques à partir de textes. *Journée Post-Génomique de la Doua (JPGD)*, Lyon.

ONO T., HISHIGAKI H., TANIGAMI A., AND TAKAGI T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. In Bioinformatics, vol 17 no 2 2001, pp. 155-161.

PILLET V. (2000). *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.

POIBEAU T. (2001). Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états. *Conférence Terminologie et Intelligence Artificielle (TIA'2001)*.

POIBEAU T. ET NAZARENKO A. (1999). L'extraction d'information, une nouvelle conception de la compréhension de texte ?. *Traitement Automatique des Langues*, vol 40 n°2, p. 87-115.

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

QUINLAN J. R. (1992). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

RILOFF E. (1993). 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.

SODERLAND S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. In *Machine Learning Journal*, vol 34.

STAPLEY B. J. AND BENOIT G. (2000). 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. (2000). Automatic Extraction of Protein Interactions from Scientific Abstracts. In Proceedings of the *Pacific Symposium on biocomputing (PSB'2000)*, vol.5, p. 502-513, Honolulu.

WOSTEN M. M. (1998). Eubacterial sigma-factors. FEMS *Microbiol. Rev*. vol 3, 127-50.

YANG Y.AND PEDERSEN J. (1997). a comparative study on feature selection in text categorization. In *International Conference on ML*.

Annexe

Les expériences décrites ici ont montré que IVI et BN ont un comportement similaire avant élagage des attributs. En revanche, après élagage, les résultats d'IVI sont meilleurs. Afin d'expliquer ces résultats, commençons par réécrire les formules de BN et IVI. Soit une phrase S = (w_1, w_2, w_3, .., w_n) qu'on se propose de classer soit dans la classe positive (notée ici Y) ou dans la classe négative (notée ici N). BN classera cette phrase dans la classe Y si P(S|Y)>P(S|N). En réécrivant cette condition, on obtient :

$$\sum_{i=1}^{n} \log(1+Occ(w_i,Y)) - \log(1+Occ(w_i,N)) > \log\frac{P(N)}{P(Y)} \quad (1)$$

IVI classera cette phrase dans la classe Y si :

$$\sum_{i=1}^{n} \frac{Occ(w_i,Y) - Occ(w_i,N)}{Occ(w_i)} > 0 \quad (2)$$

$Occ(w_i,Y)$: le nombre d'occurrences de w_i dans la classe Y
$Occ(w_i,N)$: le nombre d'occurrences de w_i dans la classe N
$Occ(w_i)$: le nombre total d'occurrences de w_i

Ces deux formules sont relativement similaires, à part la division par un coefficient de normalisation dans IVI et l'utilisation du log et d'un seuil dans BN.

$$\text{seuil} = \log\frac{P(N)}{P(Y)}$$

L'usage du seuil dans BN est probablement à l'origine du rappel élevé pour la classe positive: comme le nombre d'exemples positifs est légèrement supérieur au nombre d'exemples négatifs pour tous les corpus (Figure 4), ce seuil est négatif. De ce fait, BN a tendance à privilégier la classe la plus probable, ce qui n'est pas étonnant. IVI, au contraire utilise un seuil fixe et ne tient donc pas compte de l'ensemble d'apprentissage pour fixer son seuil.

En revanche, après élagage, les résultats d'IVI sont meilleurs. Ceci est probablement dû à la division par le coefficient de normalisation. Ces attributs ont en général *$Occ(w_i)$* qui est petit. Ainsi, en divisant par ce facteur, on privilégie parmi les attributs discriminants ceux qui sont *les plus* discriminants. Il est à noter que toutes ces remarques ont été confirmées par nos expérimentations ultérieures. Nous avons modifié *l'équation 1* relative à BN en la divisant par le facteur de normalisation de IVI. Les résultats obtenus sont à la fois meilleurs que ceux de IVI et de BN.