Des textes aux associations entre les concepts qu'ils contiennent

Yves Kodratoff*, Jérôme Azé*,
Mathieu Roche*, Oriane Matte-Tailliez* **

* CNRS, LRI ** CNRS, IGM
Université Paris Sud, 91405 Orsay Cedex
{yk,aze,roche,oriane}@lri.fr

Résumé. Nous présentons dans cet article, une chaîne originale d'outils allant de l'acquisition du corpus à l'extraction d'information. Ces outils permettent de faciliter le travail de l'expert en automatisant une partie des traitements. Nous étudions l'automatisation d'une étape clef préalable à la construction d'une ontologie terminologique, à savoir l'acquisition des termes pertinents qui constitueront les noeuds de l'ontologie. Nous avons obtenu la terminologie complète de quatre corpus différents par la langue et par la taille. La validation de ces terminologies par des experts montre que notre méthode fournit un très grand nombre de termes de qualité satisfaisante. Des classes de concepts ont été construites avec ces termes de façon semi-automatique. Celles-ci nous permettent de représenter chaque corpus sous une forme plus compacte, à partir desquelles un processus d'extraction de règles d'association peut être appliqué. Nous avons validé les règles d'association obtenues en comparant nos résultats avec ceux d'une amélioration récente de l'Intensité d'Implication sur trois corpus. Deux de ces corpus sont issus de données réelles et un expert du domaine a discuté l'intérêt des règles obtenues avec les deux mesures.

1 Introduction

Nous présentons une approche originale permettant d'extraire des connaissances à partir de corpus spécialisés. La description des quatre corpus étudiés est abordée dans la section 2.1.

Pour traiter la diversité des formes linguistiques, par exemple le problème de la polysémie, nous avons choisi d'effectuer le travail de reconnaissance d'occurences de concepts au sein des textes. La présence d'un concept est reconnue par la présence d'une forme syntaxique particulière, ou bien d'un terme particulier (Kodratoff 2001; Fontaine et Kodratoff 2003).

Ainsi, dans la section 2.4, nous expliquerons notre méthodologie d'extraction de la terminologie du domaine. Les relations syntaxiques qui sont également considérées comme des instances des concepts sont traitées dans (Fontaine et Kodratoff 2003).

L'utilisation des ontologies construites dans la première phase du processus de fouille de textes permet une représentation condensée et simplifiée des corpus que nous étudions. L'utilisation de telles ontologies permet de représenter le corpus selon une matrice numérique *texte* $\times$ *concept*. La seconde phase de notre étude consistera à extraire des informations à partir d'une telle matrice. Nous nous intéresserons ici à l'extraction des règles d'association.

Nous présenterons dans la partie 4 un algorithme d'extraction de règles d'association fondé sur l'utilisation d'une mesure de qualité permettant d'extraire les règles les moins-contredites dans les corpus étudiés.

2 Méthodologie

2.1 Description des corpus

L'utilisation de quatre corpus de taille, de langue et de technicité différentes permet de vérifier la généricité de notre méthodologie. Le corpus de biologie moléculaire (9424 Ko) a été obtenu par une requête au *NIH* sur Medline (PubMed) avec les mots-clés *DNA-binding, proteins, yeast* obtenant de ce fait un corpus de 6119 résumés d'articles scientifiques en anglais. Il illustre le problème du traitement d'un grand nombre de textes écrits dans une langue étrangère fortement technique. Le second corpus, représenté par les introductions d'articles traitant de la fouille de données se compose de 100 textes (369 Ko). Le corpus en ressources humaines d'une taille de 3784 Ko (Compagnie PerformanSe), a été rédigé par un psychologue qui sert d'expert pour ce corpus. Le corpus de Curriculum Vitæ, d'une taille de 2470 Ko, contient 1144 CVs (Groupe VediorBis). Ces textes sont écrits dans un mode semi-télégraphique avec beaucoup de fautes d'orthographe.

2.2 Nettoyage du corpus

Chaque corpus exige un type particulier de nettoyage. Un grand nombre de règles sont appliquées, leur nombre et leur variété dépendent de chaque corpus. Par exemple, sur le corpus de biologie moléculaire, nous avons appliqué deux grands types de règles. Nous avons remplacé par "N-term" toutes les occurrences de *amino-terminal*, *amino-termini*, *N-terminal*, *N-termini*, *NH2-terminal* et *NH2-termini*. Ce type d'opération, consistant à uniformiser le vocabulaire employé, est effectué par environ 100 groupes de règles. Le deuxième type de traitement, représentant 1932 règles, consiste à remplacer les alias de gènes par leur nom générique.

2.3 Étiquetage des mots

L'étiqueteur de Brill (Brill 1994) attribue automatiquement une étiquette grammaticale aux mots du corpus. Seul le corpus de biologie moléculaire pose vraiment problème car 70% des mots n'ont pas été reconnus en utilisant le lexique standard de l'étiqueteur de Brill. Nous avons donc développé GenoBrill, une version de Brill adaptée à la biologie moléculaire et à la génomique. Cette application consiste à enrichir le lexique de base en utilisant des règles lexicales et contextuelles établies par un expert du domaine.

2.4 Acquisition des termes

En nous appuyant sur le corpus étiqueté, nous pouvons extraire les termes (nom-nom, adjectif-nom, etc.) les plus pertinents pour le domaine à l'aide d'une mesure décrite ci-après. La mesure de pertinence se fonde sur une mesure d'association favorisant l'association des mots qui sont aussi rarement associés que possible à tous les autres mots. Par exemple, dans nos corpus, examinons les termes *single-strand-DNA* ou *data-mining*. Les mots *ADN* et *data* sont liés à beaucoup d'autres mots, et ceci diminue la pertinence des termes, tandis que *single-strand* est pratiquement toujours associé à *DNA*, *mining* toujours précédé de *data*, et ceci augmente la pertinence des termes. Nous détaillerons le choix de la mesure dans la section 2.4.2.

2.4.1 Mesure d'évaluation pour l'extraction

Pour évaluer la qualité des termes extraits automatiquement, nous utiliserons la mesure de *Précision* qui est la seule mesure dont nous puissions disposer en apprentissage non supervisé et son complément la courbe d'élévation. La précision représente la proportion de termes corrects parmis les termes qui sont extraits automatiquement. Cette évaluation de la pertinence des termes doit s'effectuer par un expert du domaine.

La définition de la précision est indiquée ci-dessous, où $\mathcal{L}$ est l'ensemble des termes appartenant à la classification conceptuelle. Cette liste $\mathcal{L}$ a été constituée par l'expert du domaine.

$$Précision = \frac{\text{nombre de termes extraits présents dans } \mathcal{L}}{\text{nombre de termes extraits}}$$

Les courbes d'élévation consistent à donner la variation de la précision en fonction du nombre de termes extraits par le système (voir Figure 1).

2.4.2 Mesure utilisée pour l'extraction

Il existe un certain nombre de mesures dans la littérature et nous en avons examiné plusieurs (Church et Hanks 1990; Dunning 1993; Daille et al. 1998). Notre observation principale est que la courbe d'élévation obtenue en employant le rapport de vraisemblance ("Loglike") (Dunning 1993) donne les meilleurs résultats sur l'ensemble de nos corpus, c'est pourquoi nous l'avons retenue.

La Figure 1 montre la courbe d'élévation avec deux mesures traditionnelles pour l'extraction de la terminologie du domaine, l'information mutuelle (Church et Hanks 1990) et le rapport de vraisemblance (Dunning 1993). Les expérimentations pésentées ici ont été effectuées sur le corpus des Ressources Humaines avec la relation nom-adjectif ayant un nombre d'occurences supérieur à trois.

Au rapport de vraisemblance, nous pouvons ajouter différents paramètres, expliqués dans (Roche 2003), afin d'améliorer la précision obtenue. Notre algorithme d'extraction de la terminologie s'effectue en différentes itérations. A chaque itération, les termes binaires extraits (nom-nom, adjectif-nom, etc.) sont introduits dans le corpus avec un trait d'union. Lors des itérations suivantes, ces termes sont donc reconnus comme des mots à part entière et permettent alors la formation d'autres termes. Un des paramètres essentiels que nous avons ajouté dans notre approche consiste à privilégier les termes

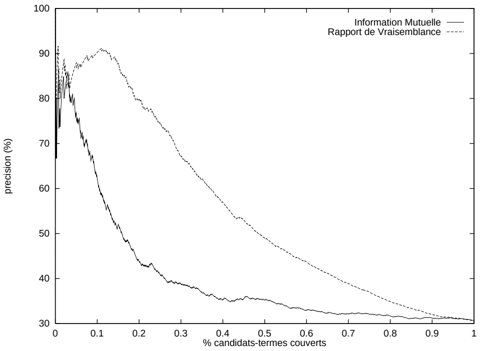

FIG. 1 – *Courbe d'élévation avec la relation nom-adjectif du corpus des Ressources Humaines. Nous avons sélectionné ici les termes qui apparaissent plus de 3 fois.*

	Corpora			
	Bio. Mol.	F. de D.	Res. Hum.	CV
Nb total de termes extraits (précision)	**9014 (88.4%)**	**372 (80.0%)**	**1844 (97.3%)**	**489 (91.1%)**
Nb de termes de longueur ≥ 4 (précision)	1921 (94.7%)	13 (76.9%)	160 (71.2%)	47 *TNE*
Nb de termes de longueur = 3 (précision)	3912 (86.2%)	86 (86.0%)	656 (99.5%)	174 (82.2%)
Nb de termes de longueur = 2 (précision)	3181 (88.9%)	366 (78.7%)	1028 (99.9%)	363 (95.3%)

TAB. 1 – *Nombre de termes et précision estimée pour chaque corpus. TNE: termes non expertisés*

formés avec les mots inclus dans les termes des itérations précédentes. Un tel paramètre permet de privilégier plus spécifiquement le vocabulaire du domaine.

3 Analyse de la pertinence des termes

Les résultats sont analysés selon deux axes complémentaires. Une analyse statistique montre les résultats en précision pour chacun des quatre corpus étudiés (section 3.1). En l'absence de terminologies standards, la validation sémantique des termes est impossible pour trois de nos corpus. Par contre, pour les termes obtenus à partir du corpus de biologie (Matte-Tailliez et al. 2002), nous avons pu comparer nos résultats à des bases de terminologie telles que Gene Ontology (GO) (section 3.2).

3.1 Analyse statistique sur les quatre corpus

Le Tableau 1 montre que la précision obtenue sur les quatre corpus est globalement très satisfaisante (selon les corpus de 80% à 97.3%).
Nous avons également effectué une expérimentation consistant à évaluer la précision

selon la longueur des termes. Ceci montre que, selon les corpus, les résultats de la précision sont sensiblement différents. Les termes les plus pertinents concernant le domaine de la Biologie Moléculaire semblent être les termes les plus longs.

3.2 Analyse sémantique par comparaison à Gene Ontology

Gene Ontology (GO, `http://www.genontology.org`) est l'ontologie terminologique de référence dans le domaine de la Génomique. GO propose une aide précieuse pour l'annotation des génomes (Gene Ontology Consortium 2001). Les trois ontologies de GO contiennent 13000 termes. La pertinence des termes y est régulièrement vérifiée, par rapport à l'évolution en Génomique, ainsi l'ontologie est soigneusement construite et vérifiée par des experts. Afin de valider les termes que nous avons extraits, nous les avons comparés à ceux de GO. Les grandes variations que nous avons observées nous ont conduit à consacrer un effort significatif pour comprendre ces différences. GO est visiblement construit par des experts. GO pourrait néanmoins servir de point de départ à l'extraction automatique des connaissances. Une recherche automatique de tous les termes biologiques pertinents à l'annotation génomique serait une amélioration intéressante. Afin d'effectuer une comparaison à grande échelle entre notre terminologie et GO, nous avons dû transformer tous ses nœuds en termes, selon une syntaxe normalisée. Puisqu'ils contiennent divers symboles comme “(” , “<”, “\” et mots d'anglais comme “and”, “associated to”, etc. Il est très difficile de les éliminer systématiquement (et correctement!) d'une liste de termes (Collier et al 2001). De cette façon, nous construisons une liste de termes que nous appelons “GO-nostermes” qui ne contient pas certains termes, présents en fait dans GO, mais présents de façon compréhensible à un expert humain seulement. Par exemple, GO: 0007001 indexe *chromosome organization and biogenesis (sensu Eukaria)* qui désigne deux termes, au moins pour le spécialiste humain, *chromosome-organization-sensu-Eukaria* et *biogenesis-sensu-Eukaria*. Notez que la connaissance du domaine est nécessaire pour disposer les mots ainsi autour d'un “et”. Un autre problème vient des synonymes courants dans les textes biologiques, ou même de simples variantes syntaxiques, telles que notre terme : *initiation-of-transcription* correspondant trivialement à G0:0006352,: *transcription-initiation*. Notez cependant que, puisque nous n'avons pas trouvé d'occurrences significatives du terme *transcription-initiation* dans les textes, les auteurs s'expriment de fait en utilisant notre terme.

Notre liste ”GO-nostermes” contient 13641 termes. Il est néanmoins évident que notre procédé introduit une certaine erreur puisque certains termes qui seraient reconnus par un expert dans GO n'ont pas été engendrés par notre méthode automatique. Quand nous comparons nos résultats et la terminologie de GO, nous devons donc introduire une nouvelle source d'erreur. Après analyse manuelle, nous estimons ce taux d'erreur autour 4%. Nous pouvons donc dire que nous combinons deux sources d'erreur différentes : d'une part en produisant des termes non significatifs à partir des textes, d'autre part en ne produisant pas certains des termes de GO. Cette erreur globale est de l'ordre de 15%. Puisque le nombre total de termes que nous produisons et qui ne sont pas dans GO-nostermes est de 8553, ceci signifie que 1283 d'entre eux peuvent être des erreurs, c'est-à-dire que nous produisons au moins 7270 termes valides qui ne sont pas dans GO.

chaperone (GO:0003754)	Hsp90-chaperone		
DNA-damage	double-strand-break	**DNA-double-strand -break-processing (GO:0000729)**	
meiosis (GO:0007126)	meiotic-prophase	**meiotic-prophase-I (GO:0007128)**	
		meiotic-prophase-II (GO:0007136)	
RNA-splicing (GO:0008380)	***	alternative-splicing	
zinc-finger-protein	zinc-finger-domain	LIM-homeodoamin	
		zinc-finger-motif	C4-zinc-finger-motif
DNA-binding (GO:0003677)	DNA-binding -protein-domain	DNA-binding -protein-motif	basic-helix-loop-helix -leucine-zipper-motif

TAB. 2 – *Une sélection de nos quelques termes montrant comment ils peuvent s'insérer dans GO. Les termes sont ordonnés par généralité décroissante de la droite vers la gauche. Les termes appartenant aussi à GO sont en caractères gras, ceux qui ne le sont pas sont en lettres normales. Nous ajoutons un *** quand nous pensons que des termes intermédiaires seraient nécessaires dans une ontologie.*

Parmi les 1428 termes validés par un expert et non présents dans GO-nostermes, nous montrons ici un petit ensemble de termes illustrant comment ceux-ci pourraient s'insérer dans GO (voir Tableau 2). Quelques centaines d'autres exemples de termes, que nous avons extraits, sont disponibles sur le site `http://www.lri.fr/ia/Genomics/`.

Nos termes pourraient être facilement incorporés dans GO. Par exemple *Hsp90-chaperone* est une instance évidente de *chaperone*, GO:0003754. Quelques-autres sont très loin de tout concept de GO, et il serait intéressant de créer des liens entre ces termes et ceux de GO. Par exemple, *basic-helix-loop-helix-leucine-zipper-motif* est un descendant de *DNA-binding-protein-motif*, lui-même un descendant de *DNA-binding-protein-domain*, lui-même un descendant *DNA-binding* , GO:0003677. Le Tableau 2 montre que nos termes contiennent les intermédiaires nécessaires. Quelques-uns de nos termes pourraient être insérés entre deux concepts de GO. Par exemple, *meiotic-prophase* peut être placé entre *meiosis* (GO:0007126) et *meiotic-prophase-I* (GO:0007128)). Finalement, quelques termes de GO, comme *DNA-double-strand-break-processing* (GO:0000729) pourraient recevoir plusieurs parents, comme nos termes *double-strand-break et DNA-damage.*

Un effet de la façon dont GO complète son ontologie apporte une validation supplémentaire à notre approche. Les termes requêtés non trouvés sont étudiés par les créateurs de GO, et ceux considérés comme intéressants sont ajoutés à l'ontologie. Nous avons remarqué que les termes que nous avons requêtés, absents de GO en octobre 2002 y sont maintenant insérés. Nous ne prétendons pas que notre requête est la raison pour laquelle ils ont été inclus, mais cela montre bien que les termes que nous avons découverts dans les textes sont d'intérêt pour la communauté de Génomique.

4 Extraction de connaissances

Nous présentons dans cette section le type de connaissances que nous recherchons. Puis, les sections suivantes présentent les algorithmes utilisés pour extraire les connaissances ainsi que les mesures de qualité étudiées.

4.1 Connaissances recherchées

Nous avons choisi d'extraire des connaissances du type : règles d'association entre concepts présents dans les textes. La reconnaissance de la présence d'un concept dans un texte est en soi un processus complexe, utilisant Rowan, un logiciel en cours de création dans notre équipe et que nous ne pouvons décrire ici faute de place.

Chaque règle $concept_1 \rightarrow concept_2$ est accompagnée de son support et de sa confiance (Agrawal et al. 1993). Notre approche est basée sur la recherche des règles les moins-contredites par les données, c.-à-d. que implicitement, et dans notre cas, la règle $concept_1 \rightarrow concept_2$ signifie aussi que nous ne rencontrons presque jamais $concept_1 \rightarrow \neg concept_2$.

La classification conceptuelle composée de termes et de relations syntaxiques définie par l'expert nous permet de réécrire le corpus sous la forme d'une matrice $\mathcal{M}$ contenant les fréquences d'apparition de chaque concept dans les textes du corpus.

Avant d'appliquer les algorithmes d'extraction de règles d'association, la matrice $\mathcal{M}$ est discrétisée en collaboration avec un expert du domaine. Cette discrétisation a pour but d'introduire pour chaque concept un ensemble de modalités représentant l'absence d'un concept, la faible présence ou la forte présence de chaque concept discrétisé. Le nombre de modalités et leur signification sont contrôlés par l'expert.

A partir de la nouvelle matrice obtenue par discrétisation de $\mathcal{M}$, nous pouvons appliquer un algorithme d'extraction de règles d'association que nous détaillons dans la section suivante.

4.2 Extraction des règles d'association

De nombreux algorithmes peuvent être utilisés pour extraire des règles d'association à partir d'une matrice booléenne.

L'extraction des règles d'association se déroule en deux étapes : extraction des motifs fréquents puis utilisation d'une ou plusieurs mesures de qualité pour obtenir les règles d'association à partir des motifs fréquents. Un motif fréquent correspond à un ensemble d'attributs prenant la valeur vrai sur un ensemble de transactions de la base de données. La fréquence de ces motifs est appelée le support et l'utilisation de contrainte minimale sur le support permet d'élaguer l'espace de recherche des motifs fréquents. En effet, la propriété d'anti-monotonie du support[1] permet de parcourir efficacement l'espace des motifs (Agrawal 1993).

L'inconvénient majeur de cette approche repose précisément sur le besoin de définir un support minimum. Lorsque nous recherchons des connaissances dans des données

1. $\mathcal{S}$ est anti-monotone si ($si\ \mathcal{S}(a,b) < T\ alors\ \forall c, \mathcal{S}(a,b,c) < T$).

(issues de textes ou non), nous disposons rarement de suffisamment de connaissances pour pouvoir fixer un seuil minimal en dessous duquel nous sommes sûrs de ne pas trouver de connaissances intéressantes.

Si nous levons la contrainte liée à l'utilisation du support, nous ne pouvons plus utiliser les algorithmes classiques d'extraction de règles d'association.

Nous proposons donc d'utiliser directement des mesures de qualité pour extraire les règles. Notre objectif étant lié à l'extraction des règles les moins contredites par les données, nous avons choisi d'étudier le comportement de deux mesures de qualité prenant explicitement en considération le nombre de contre-exemples dans le calcul de la mesure. Ces deux mesures sont la moindre-contradiction (Azé 2003) et l'intensité d'implication normalisée (Lerman et Azé 2003).

Dans le cadre de l'étude comparative réalisée entre ces deux mesures dans la suite de l'article, nous nous sommes focalisés sur l'extraction de règles du type $A \rightarrow B$, telles que A et B soient réduits à un seul attribut.

5 Validation des règles d'association obtenues

Nous avons validé les règles obtenues sur trois corpus. L'un d'eux, "mushrooms" est un corpus académique qui présente l'avantage d'avoir été déjà très étudié, mais le désavantage évident de ne pas être ancré dans la vie réelle et donc de ne pas attirer l'intérêt des experts. Il produit de nombreuses règles, et permet de comparer le comportement de notre algorithme de génération de règles, la moindre contradiction, avec un algorithme bien connu, celui de l'intensité d'implication (Gras 1979) qui permet de détecter des règles présentant étonnamment peu de contre exemples par rapport à la valeur attendue sous l'hypothèse d'indépendance des attributs présents dans la règle considérée. Cet algorithme vient de connaître de récentes améliorations (Lerman et Azé 2003), et nous les avons utilisées. Ces améliorations permettent de prendre en considération le contexte dans lequel les règles sont placées. Cette nouvelle mesure, appelée intensité d'implication normalisée, est moins sensible à la taille du corpus dont sont issues les règles d'association.

La moindre contradiction (mc) prend en considération l'ensemble des règles étudiées et permet de détecter les règles les plus surprenantes parmi l'ensemble des règles étudiées (c.-à-d. celles présentant une valeur de mc supérieure à la somme de la valeur moyenne observée et de l'écart-type de la mc).

Nous avons choisi de nous comparer à l'intensité d'implication parce qu'elle a un comportement assez semblable à la moindre contradiction : toutes deux permettent de détecter des règles sans limitation de support, et avec comme seule limitation le fait que la confiance des règles doit dépasser 50%. Ces deux méthodes sont donc conçues pour éviter le piège des règles triviales presque toujours vérifiées sur la majorité du corpus. De plus, dans la comparaison que nous avons effectuée, nous nous sommes limités aux règles contenant une seule prémisse et une seule conclusion : les deux méthodes ont des problèmes de temps calcul quand on augmente trop le nombre de prémisses et de conclusions possibles. Enfin, pour faciliter la comparaison de ces approches, nous appelons "règle trouvée par une méthode" une règle dont la valeur (soit en moindre contradiction, soit en intensité d'implication) est strictement supérieure à la valeur moyenne + l'écart type des règles trouvées par cette méthode, une présentation classique en Analyse de Données. Sur la base "mushrooms" et avec cette convention de présentation, la moindre contradiction trouve 224 règles et l'Intensité d'Implication Normalisée 363. Parmi celles-ci, 108 règles sont communes aux deux mesures, ce qui confirme bien notre hypothèse qu'elles se comportent de façon comparable.

Parmi ces règles communes, le cas général est celui d'une règle avec $support \in [0,2..0,5]$ et $confiance \in [0,7..1]$. Notez que $0,2$ de support signifie que tout de même plus de 1600 exemples sur les 8124 de la base vérifient cette règle.

Quelques règles ont un support "minuscule" de $0,024$ ce qui correspond à 192 instances dans la base. Elles ont alors comme propriété de n'être que très peu contredites. La plus contredite d'entre elles est telle que $A \rightarrow B$ est soutenue par 256 exemples et contredite par 36, ceux-ci soutenant donc $A \rightarrow \neg B$. De telles règles avec un si petit support sont peut-être bien dues au bruit comme la plus contredite d'entre elles, mais il faut avouer que les règles affirmées par 192 cas et contredites par aucun méritent au moins un examen sérieux par un expert, dans le cas d'une base ancrée. Nous ne pouvons donc que nous féliciter d'avoir été capables de détecter de telles règles représentant de possibles pépites de connaissance. Inversement, nous ne trouvons qu'une seule règle un peu triviale, avec un support de $0,97$. Cette règle triviale présente une intensité d'implication de $0,758$, ce qui en fait la règle acceptée avec la plus faible des valeurs. Ceci confirme bien que l'intensité d'implication tend à éliminer ce type de règles. Quant aux règles qui ne sont détectées que par l'une des deux méthodes, elles présentent une tendance générale à avoir un support très faible dans le cas de l'intensité d'implication. Dans ce cas, sur les 255 règles concernées, 69 ont un support inférieur à $0,1$; les autres règles ont un support compris entre $0,1$ et $0,31$. Inversement, les règles détectées par la moindre contradiction ont tendance à avoir un support un peu plus élevé. Sur les 116 règles concernées, 32 ont un support inférieur à $0,1$ (c'est-à-dire une proportion comparable à l'intensité d'implication), mais 71 ont un support supérieur à $0,4$.

Ceci confirme donc le rôle joué par chacune de ces deux mesures. L'intensité d'implication va détecter des règles à très faible support, mais peut accepter des règles assez contredites (nous en verrons un exemple sur les données ancrées). La moindre contradiction aura tendance à accepter des règles à plus fort support et à éliminer celles à très faible support mais, bien entendu, sélectionne des règles très peu contredites dans les données.

De cette expérimentation sur des données non ancrées, il ressort que ces deux mesures sont capables de détecter des pépites de connaissance de nature légèrement différentes. Un premier passage devrait se concentrer sur les règles considérées comme les meilleures par les deux méthodes, mais on peut envisager de les utiliser en séquence pour ne pas risquer d'oublier une pépite de connaissance.

Les deux bases de données ancrées que nous avons utilisées sont issues de textes. La chaîne de traitement décrite dans cet article est appliquée aux textes, et des associations entre concepts présents dans les textes sont détectées. Un des corpus est constitué de 100 introductions à des articles en anglais et traitant de fouille de données. Ce corpus a été constitué par l'un des auteurs et peut être consulté sur demande. Les concepts sont du type “nature des entrées”, “algorithme de l'auteur”, etc. L'autre est constitué de 378 commentaires en français de tests psychologiques en ressources humaines correspondant aux 378 types d'individus prévus par les tests. Ce corpus est propriété de la société PerformanSe. Les concepts sont du type “activité dans l'entreprise”, “relations dans l'entreprise”, etc.

5.1 Résultats relatifs au corpus de PerformanSe.

La moindre contradiction trouve 25 règles et l'Intensité d'Implication Normalisée 38. Parmi celles-ci, 22 règles sont communes aux deux mesures. Il se trouve qu'aucune de ces règles n'a un support très petit, et donc elles représentent bien les *a priori* utilisés dans les tests. Par exemple, quand le concept de stress est fortement évoqué, alors le concept d'environnement l'est aussi, ce qui est normal dans la mesure où le stress s'exerce par l'intermédiaire de l'environnement. Les trois règles trouvées par la moindre contradiction seule sont en fait à la limite des valeurs détectées par l'intensité d'implication et donc ne sont pas essentiellement uniques à la moindre contradiction. Elles expriment, d'une part, une forme d'équivalence entre le relationnel et l'environnement, ce qui n'est guère étonnant, et d'autre part le fait que le concept d'implication dans l'entreprise implique celui de relationnel. Ceci est au contraire une surprise et exprime peut être un *a priori* discutable des tests. Les règles détectées par l'intensité d'implication seule présentent toutes la propriété d'être presque autant confirmées qu'infirmées, ce qui les rend peu fiables.

5.2 Résultats relatifs au corpus d'introductions en anglais.

La moindre contradiction trouve une seule règle, en commun avec l'Intensité d'Implication Normalisée qui en trouve six. La règle en commun affirme que lorsque l'auteur décrit des méthodes connues, alors il parle aussi de la nature des sorties de son système, et ce pour environ 25% des articles concernés. Cette règle est assez surprenante et mérite un examen plus approfondi. Les cinq règles trouvées par l'intensité d'implication seule sont encore, presque autant confirmées qu'infirmées. On ne peut pas en tirer la conclusion que ceci est une propriété de l'intensité d'implication, mais plutôt un caractère propre à nos corpus. En particulier, nos résultats sur “mushrooms” ne présentent pas du tout cette propriété.

6 Conclusions et perspectives

Le processus de fouille de textes que nous avons présenté comporte plusieurs étapes issues du domaine du traitement du langage naturel et de l'analyse de données. A l'issue du processus de fouille de textes, nous obtenons, d'une part la terminologie précise d'un domaine spécialisé, ainsi qu'un ensemble de connaissances spécifiques au domaine étudié.

Dans notre approche, la présence de l'expert, à tous les niveaux, permet d'améliorer la qualité des outils utilisés pour (récursivement) optimiser le temps de l'expert. L'intervention la plus importante de l'expert se situe au niveau de la construction des concepts (Fontaine et Kodratoff 2003). Afin de minimiser son intervention, nous avons mis en place des méthodes automatiques permettant d'extraire les termes les plus pertinents pour les domaines étudiés. Les différentes validations réalisées sur des corpus de nature très différentes montrent que les termes extraits de manière automatique sont utiles et pertinents pour les experts.

Il est important de noter que la qualité de chaque étape dépend de la qualité de l'étape précédente dans notre processus de fouille de textes. Ainsi, si après nettoyage, les textes contiennent toujours du bruit, l'étiqueteur grammatical utilisé n'associera pas les bonnes étiquettes aux mots bruités. Cependant, même en améliorant considérablement les pré-traitements des données textuelles (phases de nettoyage et d'étiquetage), il semble impossible de supprimer totalement le bruit dans les corpus. Les travaux que nous comptons mener sont donc orientés, non seulement sur les méthodes à mettre en place afin d'améliorer la qualité de nos pré-traitements, mais concernent également le renforcement des algorithmes d'extraction de règles d'association en présence de données bruitées.

Références

Agrawal R., Imielinski T. et Swami A.N. (1993), Mining Association Rules between Sets of Items in Large Databases. In Peter Buneman and Sushil Jajodia (eds.), Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., pp 207-216.

Azé J. (2003), Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances, Revue RIA-ECA numéro spécial EGC03, Volume 17, pp 171-182.

Brill E. (1994), Some Advances in Transformation-Based Part of Speech Tagging, AAAI, Vol. 1, pp 722-727.

Church K.W. et Hanks P. (1990), Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, Vol. 16, pp 22-29.

Collier N., Nobata C. et Tsujii J. (2001), Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain, Journal of Terminology, Vol. 7, pp 239-258.

Daille B., Gaussier E. et Langé J.M. (1998), An Evaluation of Statistical Scores for Word Association, J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications, pp 177-188.

Dunning T.E. (1993), Accurate Methods for the Statistics of Surprise and Coincidence In Computational Linguistics, Vol. 19, pp 61-74.

Fontaine L. et Kodratoff Y. (2003), Comparaison du rôle de la progression thématique et de la texture conceptuelle chez les scientifiques anglophones et francophones s'exprimant en Anglais, Journée de Rédactologie scientifique : L'écriture de la recherche, Nantes, *Publication ASP acceptée.* Une version en anglais est disponible à l'adresse http://www.lri.fr/~yk/.

The Gene Ontology Consortium (2001), Creating the Gene Ontology Resource: Design and Implementation, Genome Research, Vol 11, pp 1425-1433.

Gras R. (1979), Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Thèse, Université de Rennes 1.

Kodratoff Y. (2001), Comparing Machine Learning and Knowledge Discovery in DataBases: An Application to Knowledge Discovery in Texts, Machine Learning and Its Applications, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos (Eds.), Springer Verlag, LNAI 2049, pp 1-21.

Lerman I.C. et Azé J. (2003), Une mesure probabiliste contextuelle discriminante de qualité des règles d'association, Revue RIA-ECA numéro spécial EGC03, Vol 17, pp 247-262.

Matte-Taillez O., Roche M. et Kodratoff Y. (2002), A Precise Automatic Extraction of Terminology in Genomics, Research Report n°1344, LRI UMR CNRS 8623 - Université de Paris XI - Orsay.

Roche M. (2003), Extraction paramétrée de la terminologie du domaine. Revue RIA-ECA numéro spécial EGC03, Vol 17, pp 295-306.

Summary

Performing knowledge extraction from texts requests the completion of successive steps. The amount of time requested by an expert to structure knowledge causes a relative lack of such tools in the various specialized fields. Automation is therefore necessary, and this paper present some progress on the topic of automating one fundamental step in the process of ontology building, namely the gathering of significant terms ("terminology") that will constitute the nodes of the ontology. We obtained the complete terminology of four homogeneous sets of texts (corpus) different by the language and the size. The validation of these terminologies by experts showed that our method provides a very great number of terms of satisfactory quality. These terms made it possible to build classes of concepts in a semi-automatic way. Using this knowledge, we extract association rules specific to the fields. The rules thus obtained where validated on three corpora by comparing our results with the ones given by a new measure called "Normalized Implication Intensity." Two of these corpora were real-life, and a field expert discussed the interest of the rules generated by the two methods.