Combinaison de l'approche inductive (progressive) et linguistique pour l'étiquetage morphosyntaxique des corpus de spécialité

Ahmed Amrani*,**, Yves Kodratoff**
*ESIEA Recherche, Pôle ECD, 11 rue Baudin, 74200 Ivry sur Seine, France
amrani@esiea.fr
**LRI, UMR CNRS 8623, Bât. 490, Université de Paris-Sud 11, 91405 Orsay, France
yk@lri.fr

Résumé. Les étiqueteurs morphosyntaxiques sont de plus en plus performants et cependant, un véritable problème apparaît lorsque nous voulons étiqueter des corpus de spécialité pour lesquels nous n'avons pas de corpus annotés. La correction des ambiguïtés difficiles est une étape importante pour obtenir un corpus de spécialité parfaitement étiqueté. Pour corriger ces ambiguïtés et diminuer le nombre de fautes, nous utilisons une approche itérative appelée *Induction Progressive*. Cette approche est une combinaison d'apprentissage automatique, de règles rédigées par l'expert et de corrections manuelles qui se combinent itérativement afin d'obtenir une amélioration de l'étiquetage tout en restreignant les actions de l'expert à la résolution de problèmes de plus en plus délicats. L'approche proposée nous a permis d'obtenir un corpus de biologie moléculaire « correctement » étiqueté. En utilisant ce corpus, nous avons effectué une étude comparative de quatre étiqueteurs supervisés.

1 Introduction

Dans le cadre d'un processus complet de fouille de textes (Kodratoff et al., 2003, Amrani et al., 2004a), nous nous sommes intéressés à l'étiquetage morphosyntaxique des corpus de spécialité. L'étiquetage morphosyntaxique consiste à affecter à chaque mot dans la phrase son étiquette morphosyntaxique, en prenant en considération le contexte et la morphologie de ce mot. L'étiquette morphosyntaxique est composée de la catégorie syntaxique du mot (nom commun, nom propre, adjectif, etc.) et souvent comporte des informations morphologiques (genre, nombre, personne, etc.). Les outils informatiques nécessaires à l'opération d'étiquetage sont appelés « étiqueteurs ».

Un problème se pose lorsque les étiquettes des mots sont ambiguës. Par exemple, le mot *functions* peut être un nom au pluriel ('*biological functions are…*') ou bien un verbe au singulier ('*this gene functions as…*'). Le problème à résoudre est celui de trouver l'étiquette correcte selon le contexte. La correction de ces ambiguïtés est une étape importante pour obtenir un corpus de spécialité « parfaitement » étiqueté. Pour lever ces ambiguïtés et donc diminuer le nombre de fautes d'étiquetage, nous proposons une approche interactive et itérative appelée *Induction Progressive*. Cette approche est une combinaison d'apprentissage automatique, de règles rédigées par l'expert et de corrections manuelles. L'induction pro-

gressive nous a permis d'obtenir un corpus de biologie moléculaire « correctement » étiqueté. Nous avons alors utilisé le corpus obtenu pour entraîner quatre étiqueteurs morphosyntaxiques supervisés, puis nous avons effectué une étude comparative.

2 Étiquetage morphosyntaxique

2.1 Les approches d'étiquetage morphosyntaxique

Il y a deux approches principales pour l'étiquetage morphosyntaxique : l'approche inductive et l'approche linguistique.

L'approche inductive nécessite la disponibilité d'un grand corpus annoté. L'annotation de corpus est tout apport d'information aux textes bruts. L'information requise ici est l'étiquette morphosyntaxique correcte de chaque mot. Parmi les étiqueteurs inductifs, nous pouvons citer : l'étiqueteur à base de transformation (Brill, 1995), l'étiqueteur à base de « Séparateurs à vaste marge (SVM)» (Giménez et Màrquez 2003) et les étiqueteurs probabilistes (Ratnaparkhi, 1996, Toutanova et al., 2003). Des étiqueteurs plus élaborés ont été développés comme les étiqueteurs basés sur la combinaison de plusieurs étiqueteurs individuels, permettant ainsi de pallier les déficiences de chacun des systèmes pris séparément (Brill et Wu 1998, Halteren et al., 2001). Les résultats publiés de ses étiqueteurs appliqués au corpus classique 'WSJ' sont de l'ordre de 96-97%.

L'approche linguistique, quant à elle, consiste à coder manuellement les connaissances linguistiques sous forme de règles. Les règles acquises sont ensuite utilisées pour l'étiquetage de nouveaux textes. L'un des travaux les plus important de cette approche est le développement d'une grammaire de contraintes (Karlsson et al., 1995) et son application à l'étiquetage morphosyntaxique (Voutilainen, 95). Cet étiqueteur peut être considéré comme le meilleur étiqueteur existant. En effet, il atteint une précision supérieure à 99% d'étiquettes correctes.

Afin de bénéficier des avantages des deux approches, plusieurs chercheurs ont combiné les étiqueteurs inductifs et les règles linguistiques (Tapanainen et Voutilainen 1994, Samuelson et Voutilainen 1997). Il existe d'autres systèmes d'étiquetage qui utilisent de petits corpus annotés pour accélérer l'annotation d'un corpus plus grand ; ces systèmes combinent l'utilisation d'un étiqueteur appris sur un petit corpus et l'intervention d'un humain via une interface interactive. Nous pouvons citer par exemple, les systèmes ANNOTATE (Plaehn et al., 2000) et KCAT (Won-Ho et al., 2000) qui sont basés sur des étiqueteurs statistiques.

2.2 Problématique de l'étiquetage morphosyntaxique

Quelque soit le système sur lequel ils sont basés, les étiqueteurs actuels atteignent des performances très satisfaisantes mais il est difficile de dépasser la précision de 96-97%. Plusieurs chercheurs justifient cette difficulté par les incohérences dans le corpus d'apprentissage (Ratnaparkhi, 1996, Toutanova et al., 2003). Les corpus sont annotés manuellement, ils peuvent donc contenir des erreurs. Par conséquent, l'amélioration de la qualité des corpus et la correction des erreurs ont une importance capitale.

De plus, les bons résultats des étiqueteurs supervisés s'expliquent par le fait que les travaux en question se situent dans le domaine de l'apprentissage supervisé où le corpus de test est de nature similaire au corpus d'apprentissage. Un véritable problème apparaît lorsque nous voulons traiter des corpus de spécialité pour lesquels nous n'avons pas de corpus anno-

tés. L'acquisition d'un tel corpus est coûteuse et elle constitue le goulet d'étranglement pour construire un étiqueteur pour une nouvelle application ou un nouveau domaine.

La plupart des étiqueteurs utilisent des informations de nature essentiellement locale (une séquence de deux ou trois mots consécutifs). Par conséquent, ces étiqueteurs butent sur les ambiguïtés qui demandent la prise en considération d'un contexte large. Par exemple : l'ambiguïté relatif/conjonction pour *que*. Bien que l'approche linguistique engendre des modèles de très bonne qualité et traite efficacement les ambiguïtés difficiles, elle est coûteuse et laborieuse. Par exemple, le développement de l'étiqueteur ENGCG (Voutilainen, 95) a nécessité plusieurs années. Cependant, comme pour les étiqueteurs supervisés, les performances des étiqueteurs linguistiques se détériorent lorsqu'ils sont appliqués à de nouveaux corpus.

3 Induction progressive

La correction des ambiguïtés difficiles est une étape importante pour obtenir un corpus de spécialité parfaitement étiqueté. Pour corriger ces ambiguïtés et diminuer le nombre de fautes d'étiquetage, nous utilisons une approche itérative appelée *Induction Progressive*. L'induction progressive est une combinaison d'apprentissage automatique, de règles rédigées par l'expert et de corrections manuelles qui se combinent itérativement afin d'obtenir une amélioration de l'étiquetage tout en restreignant les actions de l'expert à la résolution de problèmes de plus en plus délicats. Le principe de l'induction progressive est le suivant : en utilisant le langage CorTag (détaillé dans la section suivante), l'expert écrit une règle (ou plusieurs règles) pour corriger une ambiguïté spécifique. Les règles de l'expert sont ensuite appliquées au corpus $Corp_0$ et engendrent un corpus $CorpRegExp_0$. Un algorithme d'apprentissage de règles est ensuite utilisé pour apprendre la modification engendrée par les règles de l'expert. Les règles apprises sont aussi appliquées au corpus $Corp_0$ et engendrent un corpus $CorpRegInd_0$. En utilisant une nouvelle version du logiciel interactif ETIQ (Amrani et al., 2004b, Amrani et al., 2005a), les différences entre les deux corpus ($CorpRegExp_0$ et $CorpRegInd_0$) sont alors visualisées pour faciliter leur analyse. Les points de désaccords sont souvent des cas particulièrement difficiles à étiqueter. Leur visualisation permet de :

- détecter les erreurs produites par les règles de l'expert.
- mettre à jour les règles de l'expert si elles se trompent. Pour ce faire, les règles induites ayant trouvé l'étiquette correcte sont une indication précieuse.
- confirmer ou corriger les étiquettes obtenues par les règles de l'expert, ainsi nous obtenons une base d'étiquettes sûres. Cette base servira pour améliorer progressivement la qualité des règles induites.

3.1 Le langage d'étiquetage : CorTag

Le langage CorTag (Kodratoff et Azé, non publié) traite principalement les problèmes d'étiquetage relationnel. Il offre la possibilité à l'utilisateur d'exprimer ses règles contextuelles sous forme de conditions qui doivent être vérifiées et d'actions associées. Les règles peuvent admettre des exceptions. La forme générale d'une règle est la suivante : si *conditions* alors *actions* sauf *exceptions*.

Les conditions, actions et exceptions s'expriment généralement sous la forme de triplet : (*Pos*, *Mot*, *Etiquette*) où *Pos* est la position relative du mot dans la phrase, *Mot* est le mot situé à la position indiquée et *Etiquette* est l'étiquette morphosyntaxique du mot. Les posi-

tions s'expriment relativement à un élément central que nous nommons le pivot. Le pivot est l'élément autour duquel la règle va s'articuler et il s'agit très souvent du mot qui doit être ré-étiqueté. Le pivot doit obligatoirement être présent dans la partie *conditions* de la règle et s'exprime de la manière suivante : (0, *Mot*, *Etiquette*). Les informations *Mot* et *Etiquette* peuvent ne pas être toutes les deux renseignées.

Par exemple, la règle « ***si*** *(-1,,RB) (0,,NN) (+1,,JJ)* ***alors*** *(-1,,JJ)* ***sauf*** *(-2,,JJ)* » exprime le fait que si l'élément central est étiqueté comme un nom *(0,,NN)* et qu'il est précédé d'un adverbe *(-1,,RB)* et suivi d'un adjectif *(+1,,JJ)* alors l'adverbe est ré-étiqueté en adjectif sauf s'il est lui-même précédé d'un adjectif.

Le langage offre la possibilité d'écrire des règles complexes en permettant à l'expert de manipuler des éléments dont la position peut être inconnue lors de l'écriture de la règle mais qui seront instanciés lors de l'application de celle-ci. Par exemple, le langage permet d'écrire la règle relationnelle suivante : Si le premier verbe avant le mot *that* appartient à une liste donnée alors le mot *that* est étiqueté IN (conjonction de subordination) sauf s'il y a un non commun (singulier ou pluriel) entre le mot *that* et ce verbe.

Le langage dispose aussi d'une bibliothèque de fonctions intégrées qui permettent à l'expert d'exprimer des contraintes sur les mots, les étiquettes, les positions et la phrase.

3.2 Induction Progressive

L'étape 0 consiste à obtenir un corpus ($Corp_0$) de spécialité étiqueté par un étiqueteur morphosyntaxique généraliste. Pour ce faire, nous utilisons l'étiqueteur de Brill (Brill, 1995). Puis, avec l'aide de ETIQ (Amrani et al., 2004b, Amrani et al., 2005b), l'expert adapte les règles morphologiques et contextuelles au domaine étudié. Nous avons constaté que les erreurs dues aux mots inconnus de la spécialité sont facilement et efficacement résolues par les règles morphologiques d'ETIQ. Cependant, les confusions contextuelles qui nécessitent des règles relationnelles sont difficilement résolues. Pour les résoudre, nous utilisons l'induction progressive.

L'expert identifie alors des erreurs qui lui paraissent importantes, et rédige des règles de correction pour chaque ambiguïté. Ces règles peuvent être rédigées au sein de ETIQ, ou bien, s'il s'agit de règles fortement contextuelles, en utilisant un langage de programmation dédié à la rédaction de ces règles (le langage CorTag). Chaque règle s'applique à un contexte précis et sert à corriger une erreur spécifique. L'expert produit ainsi un certain nombre de règles, qu'il applique au corpus $Corp_0$ et il obtient ainsi le corpus $CorpRegExp_0$. Un problème quasi insurmontable se présente lorsque l'expert travaille sur de gros corpus : le nombre d'application des règles peut atteindre plusieurs milliers, et l'expert ne peut alors vérifier la validité de ses règles que sur un sous-corpus du corpus initial.

Une fois que l'expert estime qu'il a franchi une étape et qu'il pense avoir résolu un problème à peu près correctement, il fait alors appel à un algorithme d'apprentissage automatique de règles. Cet algorithme sert à apprendre les modifications produites par les règles de l'expert pour résoudre une erreur spécifique. Bien entendu, l'apprentissage correspondant peut se faire sans problème sur le corpus complet. Le logiciel ETIQ permet d'apprendre ces modifications en comparant deux versions du même corpus :

- Le corpus $Corp_0$ avant l'application des règles de l'expert.
- Le corpus $CorpRegExp_0$ après l'application des règles de l'expert.

Une base d'apprentissage est engendrée à partir de ces deux corpus. La base contient les exemples modifiés par les règles de l'expert (exemples positifs) et les exemples non modifiés

(exemples négatifs). Les règles induites sont aussi appliquées au corpus $Corp_0$ et engendrent le corpus $CorpRegInd_0$.

Enfin, les corpus $CorpRegInd_0$ et $CorpRegExp_0$ sont présentés à l'expert lorsque les corrections apportées par les règles de l'expert diffèrent de celles apportées par les règles induites (Voir les exemples du tableau 1). L'expert examine ces cas. Un point important est de remarquer que l'expert doit soit confirmer 'son' étiquetage (Voir tableau 1, les exemples 1, 2, 3 et 4), soit constater que les règles induites ont 'raison' (Voir tableau 1, les exemples 5, 6, 7 et 8). Lorsque le cas se produit assez fréquemment (c'est-à-dire que le processus inductif est de bonne qualité), ceci produit une sorte de compétition entre lui-même et l'induction, si bien que son attention reste éveillée sans problème.

	mot -3	**mot -2**	**mot -1**	**mot 0**	**mot 1**	**mot 2**	**mot 3**
1	*that*	*rad59*	*Delta*	***exhibits***	*synergistic*	*effects*	*with*
	IN	FRM	NNP	**NNS**	JJ	NNS	IN
	IN	FRM	NNP	**VBZ**	JJ	NNS	IN
2	*,*	*pathway*	*that*	***functions***	*to*	*bypass*	*the*
	,	NN	IN	**NNS**	TO	VB	DT
	,	NN	WDT	**VBZ**	TO	VB	DT
3	*suggesting*	*that*	*TANK2*	***kills***	*cells*	*by*	*necrosis*
	VBG	IN	FRM	**NNS**	NNS	IN	NN
	VBG	IN	FRM	**VBZ**	NNS	IN	NN
4	*transcription*	*activator*	*that*	***modulates***	*,*	*with*	*Aft1p*
	NN	NN	DT	**NNS**	,	IN	FRM
	NN	NN	WDT	**VBZ**	,	IN	FRM
5	*hypoplasia*	*congenila*	*and*	***blocks***	*steroid*	*biosynthesis*	*by*
	NN	NN	CC	**VBZ**	JJ	NN	IN
	NN	NN	CC	**NNS**	JJ	NN	IN
6	*the*	*CFTR*	*gene*	***augments***	*intestinal*	*expression*	*in_vivo*
	DT	NNP	NN	**VBZ**	JJ	NN	NN
	DT	NNP	NN	**NNS**	JJ	NN	NN
7	*gln3*	*gat1*	*mutant*	***displays***	*a*	*pronounced*	*sensitivity*
	FRM	FRM	JJ	**VBZ**	DT	VBN	NN
	FRM	FRM	JJ	**NNS**	DT	JJ	NN
8	*contains*	*Six*	*membrane*	***spans***	*,*	*two*	*of*
	VBZ	CD	NN	**NNS**	,	CD	IN
	VBZ	CD	NN	**VBZ**	,	CD	IN

TAB. 1 – *Une liste d'exemples relatifs à la confusion VBZ-NNS. Dans ces exemples, le mot central (mot 0) est étiqueté différemment par les règles induites et les règles de l'expert (les étiquettes attribuées par les règles de l'expert sont soulignées). Nous constatons que pour certains exemples, les règles de l'expert ont 'raison' (les quatre premiers exemples) et que pour d'autres, les règles induites ont 'raison' (les quatre derniers exemples). Par exemple, l'expert constate que ses règles ont incorrectement étiqueté le mot 'displays' comme (NNS) (le mot central du 7ème exemple du tableau). En utilisant notre logiciel ETIQ, l'expert peut corriger sa faute en cliquant sur un menu.*

Exemple. Dans la phrase "…that rad59 Delta **exhibits** synergistic effects...", le mot *exhibits* est étiqueté incorrectement comme NNS par les règles induites et il est étiqueté correctement comme VBZ par les règles de l'expert (le 1er exemple du tableau 1). Cependant, dans la phrase "…gene responsible for adrenal hypoplasia congenita and **blocks** steroid biosynthesis

by…", le mot *blocks* est étiqueté incorrectement comme NNS par les règles de l'expert et étiqueté correctement comme VBZ par les règles induites (le 5ème exemple du tableau 1).

Ainsi nous disposons de trois versions successives du même corpus : $CorpRegExp_0$, $CorpRegInd_0$ et $CorpSûr_0$.

- $CorpRegExp_0$ est le corpus de départ sur lequel nous appliquons les règles de l'expert.
- $CorpRegInd_0$ est le corpus de départ sur lequel nous appliquerons les règles induites.
- $CorpSûr_0$ est le corpus dans lequel nous gardons les étiquettes corrigées manuellement (si l'induction 'gagne' comme pour les exemples 5, 6, 7 et 8 du tableau 1) ou confirmées par l'expert (si c'est lui qui a 'gagné' comme pour les exemples 1, 2, 3 et 4 du tableau 1) durant le processus.

Encore une fois, si le corpus est suffisamment petit, l'expert peut corriger toutes les erreurs et il n'est pas vraiment nécessaire d'itérer ce processus. Comme nous partons du principe que le corpus est volumineux, l'expert ne peut pas examiner les milliers de cas où on peut voir une différence entre ses règles et les règles induites. Par contre, il peut noter certains des cas où il a 'perdu' par rapport à l'induction et afficher dans ETIQ la règle induite qui a 'gagné' sur lui (voir exemple ci-dessous). Il analyse ces règles et reçoit ainsi une indication sur la façon d'améliorer ses propres règles pour ne plus faire les erreurs qu'il a corrigées à la main. Il applique alors ces nouvelles règles à $CorpSûr_0$ qui devient le corpus de départ de l'itération suivante, pour engendrer $CorpRegExp_1$ en lui appliquant les nouvelles règles déduites au cours de l'itération « 1 ».

Exemple. Nous avons expliqué dans le tableau 1 que l'expert peut modifier 'à la main' ses erreurs. Il peut aussi apprendre de nouvelles règles qui lui sont suggérées par le programme d'induction de règles. Illustrons ce procédé sur le 7ème exemple du tableau 1, relatif à la confusion NNS-VBZ. En examinant $Corp_0$ (le corpus étiqueté par un étiqueteur généraliste), l'expert avait écrit, en utilisant le langage CorTag (voir Section 3.1), quelques règles pour améliorer l'étiquetage. En appliquant ces règles au corpus $Corp_0$, nous obtenons le corpus $CorpRegExp_0$. En utilisant ETIQ, nous collectons les différences relatives à la confusion NNS-VBZ entre $Corp_0$ et $CorpRegExp_0$. Puis, en nous basant sur ces différences, nous apprenons des règles en utilisant l'algorithme d'apprentissage de règles RIPPER (Cohen 1995). Pour cette confusion, l'algorithme a engendré 50 règles. L'attention de l'expert est attirée par son erreur sur l'exemple 7. Il peut alors cliquer sur le mot incorrectement étiqueté (*displays*), et faire ainsi apparaître la règle de RIPPER qui a trouvé, elle, la bonne étiquette. Dans ce cas, la règle qui s'affiche est : « SI le mot courant est étiqueté VBZ ou NNS ET SI il est suivi par un déterminant (DT) ALORS attribuer l'étiquette VBZ au mot courant » (c'est-à-dire qu'un verbe est souvent suivi d'un déterminant). Sur le corpus de départ $Corp_0$ (c'est-à-dire avant l'application des règles induites), cette règle s'applique correctement 630 fois et engendre 20 erreurs. Après avoir affiché les 630 phrases où un VBZ est suivi d'un DT, nous avons constaté que ces 630 étiquettes sont correctes. Cette règle doit donc contenir une certaine vérité linguistique. Nous avons ensuite affiché les 20 phrases où un mot étiqueté NNS est suivi par un déterminant (DT). Nous avons observé quatre cas :

- cas 1 : Le déterminant (DT) est *both* ou *each*. Par conséquent, la règle était un peu trop générale et devrait tenir compte de ces deux exceptions (*both* et *each*).
- cas 2 : Une virgule devrait apparaître après le NNS mais elle a été omise.
- cas 3 : L'étiquette déterminant (DT) est attribuée au 'A' (comme dans : *Proteins A and B*) et le 'A' est incorrectement étiqueté comme DT.

- cas 4 : Le mot d'intérêt a été incorrectement étiqueté comme NNS, ce type d'erreurs sera corrigé par la règle courante.

Par cet exemple, nous avons illustre le fait que l'ensemble des règles induites permet souvent d'optimiser les règles l'expert et de détecter des exceptions.

Il est bien entendu théoriquement possible que l'expert se trompe complètement et que l'itération « i+1 » contienne plus d'erreurs que l'itération « i », et ce serait un cas d'échec de notre méthode qui se traduirait d'ailleurs par un découragement et un abandon de l'expert. En fait, mis face à ses erreurs, l'expert a plutôt tendance à les corriger et à obtenir des règles plus efficaces. Le danger serait alors que le système induise des règles faisant les mêmes erreurs qu'à l'itération précédente, et que l'expert revoie ces mêmes erreurs à chaque itération. Nous n'avons jamais constaté ce comportement de la part de notre système inductif qui apprend des règles très différentes quand les ensembles d'apprentissage sont différents. On notera au passage l'importance des corrections manuelles de l'expert, même si elles sont relativement peu nombreuses, afin de partir d'étiquetage vraiment différent.

En conséquence, et en pratique, l'expert constate à chaque itération que le nombre de fois où il est mis en défaut par le système inductif diminue, au point qu'après quelques itérations il peut examiner toutes les différences et, éventuellement corriger toutes ses erreurs manuellement, sans recourir à des règles. Le corpus final, $CorpSûr_p$ est alors parfaitement étiqueté du point de vue de l'erreur d'étiquetage considérée au départ. Il est évidemment nécessaire de répéter le processus entier pour chacune des confusions que l'on désire corriger.

3.2.1 Les types de confusions traités par l'induction progressive

Les fautes les plus importantes sont celles qui faussent la compréhension de la phrase en détruisant la structure syntaxique. Les confusions les plus importantes que nous ayons résolu avec l'induction progressive sont les suivantes :

- Les erreurs d'étiquetage du mot « that » qui peut être étiqueté comme un déterminant « DT », une conjonction de subordination « IN », un pronom relatif « WDT », un pronom « PRP » (e.g. *I do not like **that***.) ou un prédéterminant « PDT ».
- Les confusions entre les noms et les verbes, notamment la confusion entre un nom commun au pluriel « NNS » et un verbe au présent, à la troisième personne du singulier « VBZ ». Par exemple : le mot « functions ».
- La confusion qui concerne un mot qui se termine par « ed » qui peut être un verbe au participe passé « VBN », un verbe au passé « VBD» et adjectif « JJ ».
- La confusion qui concerne un mot se terminant par « ing » qui peut être principalement un nom au singulier « NN », un verbe au gérondif « VBG » ou un adjectif « JJ ».

Ainsi, le processus doit être répété 3 ou 4 fois entièrement, ce qui n'est pas exagéré au vu des avantages qu'il y a à posséder un corpus spécialisé « correctement » étiqueté.

Nous avons appliqué notre approche à corpus de biologie moléculaire composé de 600 résumés d'intérêt parmi un corpus initial (de 6119 résumés) obtenu par requête sur *Medline* (http://www.ncbi.nlm.nih.gov) avec les mots-clés *DNA-binding, proteins, yeast*. Ce qui nous a permis d'obtenir un corpus « correctement » étiqueté.

De plus, à la fin du processus nous obtenons un ensemble de règles relationnelles intelligibles. Ces règles pourraient être utilisées pour résoudre les ambiguïtés qui ne sont pas bien traités par les étiqueteurs supervisés.

4 Comparaison des étiqueteurs supervisés

Dans la mesure où il existe déjà d'excellents étiqueteurs avec comme défaut, que nous l'avons déjà évoqué, de ne pas être capable de s'adapter à de nouveaux corpus. La tâche fondamentale d'un chercheur abordant un nouveau domaine consiste à créer un corpus correctement étiqueté. C'est cette tâche que nous venons d'illustrer. Maintenant, il est tout de même intéressant de comparer les performances des étiqueteurs existants. En utilisant le corpus « correctement » étiqueté obtenu, nous avons comparé les étiqueteurs suivants :

1- L'étiqueteur à base de transformation (Brill 1994).

2- L'étiqueteur à base de SVM (Giménez et Màrquez 2003). Pour entraîner cet étiqueteur nous avons utilisé la boite à outils *SVMTool* (Giménez et Màrquez 2004).

3- L'étiqueteur probabiliste de Stanford (Toutanova et al. 2003). Cet étiqueteur atteint une précision de 97,22% d'étiquettes correctes sur le corpus WSJ.

4- L'étiqueteur à base d'Entropie Maximale (Ratnaparkhi 1996).

Dans le tableau 2 nous présentons une comparaison entre les quatre étiqueteurs utilisés sur une partie du corpus du WSJ: (i) Les précisions publiées des étiqueteurs appliqués à un corpus test issu du corpus WSJ. (ii) Les précisions que nous avons calculées en appliquant ces étiqueteurs (appris sur le corpus WSJ) à notre corpus de biologie moléculaire.

Etiqueteurs	**SVM**	**Brill**	**Stanford**	**ME**
Précision (corpus : WSJ)	97.0	96.6	97.22	96.6
Précision (corpus : biologie)	88.97	89.47	85.73	82.81

TAB. 2 – *Les précisions relatives de quatre étiqueteurs entraînés sur le corpus WSJ.*

4.1 Les critères d'évaluation

Pour comparer les étiqueteurs nous avons utilisé les mesures de précisions et de rappel pour chaque étiquette. La précision est définie par la formule suivante :

Précision = nombre d'exemples positifs couvets / nombre d'exemples couvets.

Le rappel est défini par la formule suivante :

Rappel = nombre d'exemples positifs couverts / nombre d'exemples positifs.

Il est important de déterminer un compromis entre le rappel et la précision. Pour ce faire, nous utilisons la mesure du F_{score} (avec β=1) :

$$Fscore(\beta)=\frac{(\beta^2+1)\times \text{Pr}écision\times Rappel}{\beta^2\times \text{Pr}écision+Rappel}$$

4.2 Résultat de la comparaison

Pour comparer les quatre étiqueteurs, nous les avons entraînés sur le même sous-corpus. Le sous-corpus d'apprentissage a été constitué à partir de deux tiers de notre corpus (87965 mots) de biologie moléculaire. Pour l'évaluation, nous avons utilisé le tiers restant du corpus de biologie moléculaire. Le corpus entier comporte 131 346 mots. Le tableau 3 présente les F_{score} obtenus pour chaque étiquette par les différents étiqueteurs.

Etiq.	Nbr	F_{score} (Brill)	F_{score} (ME)	F_{score} (SVM)	F_{score} (Stan.)
WRB	97	99.48	98.96	99.48	99.48
WPP	9	100.00	100.00	100.00	100.00
WP	1	100.00	0.00	100.00	100.00
WDT	340	98.20	97.03	99.85	98.83
VBZ	1197	97.96	97.84	98.67	98.58
VBP	817	92.33	92.97	93.73	94.13
VBN	1193	90.16	91.73	92.75	92.66
VBG	540	91.87	88.59	91.31	91.76
VBD	595	86.21	90.71	89.64	92.35
VB	562	87.30	90.15	90.00	92.63
TO	760	99.80	97.09	100.00	99.67
RP	3	0.00	0.00	40.00	0.00
RBS	12	100.00	100.00	100.00	100.00
RBR	23	66.67	78.95	88.37	82.05
RB	1163	97.62	95.75	97.58	97.36
PRPP	154	99.67	99.67	99.67	99.67
PRP	383	99.87	99.21	99.87	99.87
POS	13	100.00	100.00	100.00	100.00
PDT	13	63.15	46.15	86.95	63.15
NNS	2257	98.25	97.40	98.69	98.56
NNPS	185	88.95	79.56	91.88	89.39
NNP	5841	97.63	96.32	97.83	98.46
NN	7215	97.27	95.89	97.85	97.93
MD	160	100.00	99.69	100.00	100.00
JJS	23	100.00	95.65	100.00	97.78
JJR	27	72.73	77.19	87.27	86.21
JJ	3900	94.78	93.31	95.48	95.32
IN	5584	99.71	98.80	99.86	99.80
EX	13	100.00	100.00	100.00	100.00
DT	3890	99.79	99.69	99.90	99.87
CD	393	89.91	92.41	86.89	97.16
CC	1399	100.00	99.82	100.00	100.00
: ou "	141	100.00	100.00	100.00	100.00
.	1790	100.00	99.86	100.00	100.00
,	1641	100.00	99.97	100.00	100.00
SYM	531	100.00	99.53	100.00	100.00
(	516	100.00	99.90	100.00	100.00
total	43381	97.41	96.66	97.87	98.06

TAB. 3 – *Comparaison des quatre étiqueteurs supervisés. Les étiquetages ont été appris et testés sur un corpus de biologie moléculaire. Le F_{score} a été calculé pour chaque étiquette.*

A partir de cette expérience, nous remarquons que les performances réalisées par les étiqueteurs entraînés sur un corpus de la même spécialité sont significativement meilleures que celles obtenues avec des étiqueteurs généralistes. De plus, nous pouvons constater qu'un petit corpus d'apprentissage de la même spécialité est plus bénéfique qu'un grand corpus généraliste de plus grande taille. En effet, la taille du corpus d'apprentissage de biologie moléculaire (87965 mots) est inférieure au dixième du corpus du WSJ (1 million de mots).

Comme pour le corpus du WSJ, nous constatons que pour le corpus de biologie moléculaire les deux meilleurs étiqueteurs sont l'étiqueteur de Stanford et l'étiqueteur à base de SVM, avec un léger avantage pour l'étiqueteur de Stanford.

Bien que les étiqueteurs que nous avons utilisés sont parmi les meilleurs étiqueteurs existants et leur performance sont globalement très satisfaisantes, nous constatons que les F_{score} de certaines étiquettes sont encore assez faibles. Par exemple : les meilleurs F_{score} obtenus pour les étiquettes VBN, VBG, VBD et VB sont respectivement de 92.75, 91.87, 92.35 et 92.63. En effet, même les étiqueteurs les plus performants sont incapables de résoudre ces ambiguïtés difficiles. Par contre, les règles relationnelles écrites par l'expert sont très efficaces pour corriger ces ambiguïtés.

Voici des exemples de règles relationnelles (CorTag) relatives à la correction des erreurs des VBD et des VBN:

- *si (?1,been,) (*1,,RB) (0,@BEENedVBP,) alors (0,,VBD)*

« si le mot *been* est suivi par une suite d'adverbes et cette dernière est suivie par un mot (position 0) appartenant au groupe *BEENedVBP* alors ce dernier mot est étiqueté VBD »

- si (?1,are,) (*1,,RB) (0,@AREedVBP,) alors (0,,VBN)

« s'il existe un 'are' suivi par un certain nombre (éventuellement 0 mots) d'adverbes (RB) et un mot appartenant à la classe AREedVBP suit les adverbes alors ce dernier mot est étiqueté comme VBN »

D'après ces exemples, nous constatons que la correction des ambiguïtés difficiles nécessite souvent des règles relationnelles. Les prémisses de ces règles sont fondées sur des contraintes pourtant sur les mots contextuels. Pour que ces règles soient efficaces, il est donc nécessaire que les mots du contexte soient bien étiquetés.

D'autre part, les précisions des étiqueteurs entraînés et évalués sur le corpus de biologie moléculaire sont meilleures que celles des étiqueteurs entraînés et évalués sur le corpus général du WSJ. Les étiqueteurs se comportent donc mieux lorsqu'il s'agit d'un corpus très spécifique. En effet, les corpus spécifiques sont plus homogènes et contiennent plus de régularités que les corpus généralistes (le WSJ corpus ou le *Brown* corpus).

Le tableau 4 présente les précisions obtenues par les différents étiqueteurs supervisés sur les corpus du WSJ et de biologie moléculaire :

Corpus	Brill	ME	SVM	Stanford
WSJ	96.60	96.60	97.00	97.22
Bio. Mol.	97.43	96.68	97.88	98.07

TAB. 4 – *Précisions obtenues par les différents étiqueteurs sur les corpus de biologie moléculaire et du WSJ.*

5 Conclusion

L'induction progressive est une combinaison de l'approche inductive et l'approche linguistique. L'approche inductive aide l'expert à détecter les erreurs engendrées par ces règles et lui propose de nouvelles règles pour optimiser et traiter les exceptions de ses règles. De plus, les corrections manuelles permettent d'améliorer progressivement la qualité des règles induites.

L'étude comparative des étiqueteurs a montré que les étiqueteurs supervisés sont performants lorsqu'ils sont appliqués à des corpus de spécialité. Cependant, les étiqueteurs inductifs ont tous des difficultés pour traiter les cas épineux. Ces cas sont efficacement traités par les règles relationnelles utilisées dans l'approche linguistique.

L'induction des règles effectuée par ETIQ se limite aux règles propositionnelles, l'extension de l'induction aux méthodes d'apprentissage de la programmation logique inductive permettra d'étendre la famille de règles pouvant être apprises. Nous envisageons aussi de développer des algorithmes pour détecter les incohérences dans le corpus.

Références

Amrani, A., J. Azé et Y. Kodratoff (2005a). ETIQ: Logiciel d'aide à l'étiquetage morpho-syntaxique de textes de spécialité. Dans la revue RNTI, numéro spécial EGC'2005, pages 673-678 (Session démonstration de logiciel).

Amrani, A., J. Azé, T. Heitz, Y. Kodratoff et M. Roche (2004a). From the texts to the concepts they contain: a chain of linguistic treatments. In proceedings of Text REtrieval Conference 2004 (TREC'04), Maryland USA, p 712-722.

Amrani, A., Y. Kodratoff et O. Matte-Tailliez (2004b). A Semi-automatic System for Tagging Specialized Corpora, H. DAI, R. SRIKANT, C. ZHANG Eds., Advances in Knowledge Discovery and Data Mining, PAKDD, May, Sydney, LNAI, Vol. 3056, pp 670-681.

Amrani, A., M. Roche, Y. Kodratoff et O. Matte-Tailliez (2005b). Inductive Improvement of Part-of-Speech Tagging and its Effect on a Terminology of Molecular Biology. Canadian Conference on AI 2005, Canada, LNAI, Vol. 3501, pp 366-376.

Brill, E (1995). Transformation-Based Error-Driven Learning and Naturel Langage Processing : A case Study in Part of Speech Tagging. Computational Linguistics. Decembre.

Brill, E. et J. Wu (1998). Classifier Combination for Improved Lexical Disambiguation. Proceedings of the Thirty-Sixth ACL and Seventeenth COLING.

Cohen, W. W. (1995), Fast Effective Rule Induction, Proceedings of the 12th International Conference on Machine Learning, 1995.

Giménez, J. et L. Marquez (2003). Fast and accurate part-of-speech tagging: The SVM approach revisited. RANLP- 2003. pages 153-163.

Halteren, V., J. Zavrel et W. Daelemans (2001). Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems. *Computational Linguistics*, 27(2):199-230. 2001.

Karlsson, F., A. Voutilainen, J. Heikkila et A. Anttila (1995) Constraint Grammar: A Language–Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin and New York, 1995.

Kodratoff, Y., J. Azé, M. Roche et O. Matte-Tailliez (2003). Des textes aux associations entre les concepts qu'ils contiennent. Dans les actes des XXXVièmes Journées de Statistique, Volume 2, p599-602 (résumé) et dans le numéro spécial "Entreposage et Fouille de Données" de la Revue RNTI. 1:171-182, 2003.

Plaehn, O. et T. Brants (2000). Annotate - An Efficient Interactive Annotation Tool. In Proceedings of the 6th Conference on ANLP, Seattle, 2000.

Ratnaparkhi, A (1996). A maximum entropy model for part-of-speech tagging. In EMNLP 1, pages 133–142.

Samuelsson, C. et A. Voutilainen (1997). "Comparing a Linguistic and a Stochastic Tagger". In Procs. Joint 35th Annual Meeting of the ACL and 8th Conf. of the European Chapter of the Association for Computational Linguistics, pp. 246-253.

Tapanainen, P. et A. Voutilainen (1994). Tagging accurately - Don't guess if you know. In Proceedings of the 4th ACL Conference on Applied Natural Language Processing, pp 47-52, 1994. Stuttgart.

Toutanova, K., D. Klein, C. Manning et Y. Singer. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proc. of HLT-NAACL 2003 pages 252-259.

Voutilainen, A (1995). A syntax-based part-of-speech analyser. Proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin, 1995.

Won-Ho, R., L. Heui-Seok, K. Jin-Dong et R. Hae-Chang (2000). KCAT : A Korean Corpus Annotating Tool Minimizing Human Intervention, Proc. of the 19th COLING 2000.

Summary

The part-of-Speech taggers are more and more powerful. Their accuracy, however, drops down dramatically when applied in a domain the genre of which differs from the one they have been trained upon. The correction of difficult part-of-speech ambiguities is a significant stage to obtain a 'perfectly' tagged specialized corpus. To correct these ambiguities and to decrease the number of tagging mistakes, we use an approach we call: *Progressive Induction*. The process of correction is iterative. This approach is a combination of machine learning, of rules written by the expert, and of manual corrections that are iteratively combined in order to obtain an improvement of the tagging while restricting the actions of the expert to the resolution of increasingly delicate problems. The proposed approach enabled us to obtain a "correctly" tagged molecular biology corpus. By using this corpus, we carried out a comparative study of four supervised tagger.