Induction de règles de correction pour l'étiquetage morphosyntaxique de la littérature de biologie en utilisant l'apprentissage actif

Ahmed Amrani (1), Yves Kodratoff (2) et Oriane Matte-Tailliez (2)

(1) ESIEA Recherche, 9 rue Vésale, 75005 Paris, France.
amrani@esiea.fr
(2) LRI, UMR CNRS 8623, Bât. 490, Université Paris 11, 91405 Orsay.
{yk, oriane}@lri

Mots-clés : Etiquetage morphosyntaxique, Apprentissage de règles, Apprentissage actif, fouille de textes.

Keywords: Part-of-speech tagging, rule learning, active learning, text-mining.

Résumé Dans le contexte de l'étiquetage morphosyntaxique des corpus de spécialité, nous proposons une approche inductive pour réduire les erreurs les plus difficiles et qui persistent après étiquetage par le système de Brill. Nous avons appliqué notre système sur deux types de confusions. La première confusion concerne un mot qui peut avoir les étiquettes 'verbe au participe passé', 'verbe au passé' ou 'adjectif'. La deuxième confusion se produit entre un nom commun au pluriel et un verbe au présent, à la 3ème personne du singulier. A l'aide d'interface conviviale, l'expert corrige l'étiquette du mot ambigu. A partir des exemples annotés, nous induisons des règles de correction. Afin de réduire le coût d'annotation, nous avons utilisé l'apprentissage actif. La validation expérimentale a montré une amélioration de la précision de l'étiquetage. De plus, à partir de l'annotation du tiers du nombre d'exemples, le niveau de précision réalisé est équivalent à celui obtenu en annotant tous les exemples.

Abstract In the context of Part-of-Speech (PoS)-tagging of specialized corpora, we proposed an approach focusing on the most 'important' PoS-tags because mistaking them can lead to a total misunderstanding of the text. After tagging a biological corpus by Brill's tagger, we noted persistent errors that are very hard to deal with. As an application, we studied two cases of different nature: first, confusion between past participle, adjective and preterit; second, confusion between plural nouns and verbs, 3rd person singular present. With a friendly user interface, the expert corrected the examples. Then, from these well-annotated examples, we induced rules. In order to reduce the cost of annotation, we used active learning. The experimental validation showed improvement in tagging precision and that on the basis of the annotation of one third of the examples we obtain a level of precision equivalent to the one reached by annotating all the examples.

1 Introduction

L'étiquetage morphosyntaxique est une étape importante pour la tâche d'extraction d'informations à partir de textes bruts et spécialisés. Cette étape consiste à associer à chaque mot son étiquette grammaticale en fonction de sa morphologie et de son contexte. Les étiqueteurs morphosyntaxiques actuels atteignent des performances très satisfaisantes en précision (plus de 95%) (Paroubek, Rajman, 2000). Ces bons résultats 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 même nature que le corpus d'apprentissage. Un pré-requis pour la construction d'un étiqueteur est la disponibilité d'un corpus annoté de taille importante. L'acquisition d'un tel corpus est coûteuse. D'autre part, les systèmes d'étiquetage ont tous des difficultés sur les cas difficiles. Les décisions sont le plus souvent fondées sur l'examen des contextes locaux (tels que les trigrammes de mots), qui résolvent mal les cas qui demanderaient une analyse plus globale et approfondie (Valli, Veronis, 1999).

Il existe deux approches principales pour l'apprentissage de règles : la création de règles à partir d'arbres de décision (Quinlan, 1993) et la technique d'apprentissage directe de règles comme dans l'algorithme RIPPER (Cohen, 1995). L'algorithme d'apprentissage de règles propositionnelles, PART (Frank, Witten, 1998), combine les deux approches précédentes. Chaque règle induite par PART a la forme d'une conjonction de conditions : Si T_1 et T_2 et ... T_n alors la classe est C_x. (Si T_1 et T_2 et ... T_n) est appelé le corps de la règle et (Cx) est la classe cible à apprendre. Chaque condition T_i teste une valeur particulière d'un attribut. La condition a la forme suivante : $A_i = v$, où A_i est un attribut symbolique et v est une valeur possible de A_i.

L'apprentissage actif est une technique qui permet de réduire le nombre d'exemples à annoter. Cette technique consiste à sélectionner les exemples les plus instructifs, pour lesquels le modèle courant est le plus incertain. L'apprentissage actif est de plus en plus utilisé dans des applications de traitement du langage naturel telles que l'étiquetage morphosyntaxique (Engelson, Dagan, 1999), le parsage stochastique (Tang et al, 2002) et la reconnaissance d'entités nommées (Shen et al, 2004).

Dans cet article, nous proposons une méthodologie basée sur l'apprentissage de règles de correction. Ces règles sont employées pour résoudre les erreurs d'étiquetage qui persistent après l'application de l'étiqueteur de Brill (Brill, 1994) et d'ETIQ (Amrani et al, 2004).

2 Méthodologie d'Etiquetage morphosyntaxique

L'approche proposée consiste à adapter un étiqueteur induit à partir d'un corpus généraliste à un corpus de spécialité. Notre système est basé sur l'étiqueteur de Brill (Brill, 1994). Cet étiqueteur utilise un apprentissage supervisé à base de transformations pour engendrer deux listes ordonnées de règles : règles lexicales et règles contextuelles. ETIQ[1](Amrani et al, 2004), l'étiqueteur que nous avons conçu, permet à l'expert de détecter les erreurs de l'étiqueteur de Brill, produites sur les corpus de spécialité. A l'aide d'ETIQ, l'expert visualise le résultat de l'étiquetage de Brill; il peut faire des requêtes lexicales ou contextuelles pour visualiser des

[1] http://ww.lri.fr/ia/genomics/. Téléchargement d'une version de démonstration du logiciel ETIQ.

groupes de mots (et leurs étiquettes) ayant des caractéristiques morphologiques ou contextuelles similaires. En fonction des erreurs détectées, l'expert insère des règles lexicales et contextuelles pour les corriger.

Après l'application de l'étiqueteur de Brill et d'ETIQ (Amrani et al, 2004; Amrani et al, 2005), nous avons remarqué, à l'aide du logiciel ETIQ, que certaines confusions spécifiques et difficiles à résoudre persistent. Voici les confusions les plus sérieuses : (1) JJ (adjectif) et NN (nom commun, singulier) pour quelques mots très fréquents comme *complex*. (2) VBN (verbe participe passé), JJ et VBD (verbe au passé) comme *transformed*. (3) VBZ (verbe au présent, troisième personne du singulier) et NNS (nom commun, pluriel) comme *functions* et *contacts*.

L'expert annote les exemples correspondant aux confusions identifiées. Il corrige ou il confirme l'étiquette du mot cible de chaque exemple. Afin de réduire le nombre d'exemples à annoter, nous utilisons l'apprentissage actif. Pour étudier l'impact de la représentation des exemples sur la performance, nous avons fait varier la taille des contextes aussi bien que les attributs utilisés pour représenter les exemples. Ces exemples servent à apprendre automatiquement des règles qui corrigent l'étiquette du mot en fonction de son contexte. Ces règles sont appliquées à la suite des règles contextuelles existantes.

2.1 Apprentissage actif

Nous calculons une mesure de distance entre chaque couple d'exemples. Puis, un ensemble initial d'exemples est sélectionné puis annoté. A partir de cet ensemble, nous apprenons un modèle. A chaque itération, un nouvel ensemble d'exemples pertinents est sélectionné puis annoté. La stratégie de sélection est basée sur la confiance et la diversité. Chaque étape est détaillée dans les sections suivantes.

2.1.1 Mesure de distance entre deux exemples

Chaque exemple est représenté comme suit: le mot cible est pris dans une fenêtre de *n* mots de chaque côté. Chaque mot est représenté par un ensemble d'attributs correspondant à son étiquette morphosyntaxique et à ses caractéristiques morphologiques. Soit l'exemple *x* représenté comme suit, où ($m = 2n+1$) est le nombre d'attributs et $V_{x,y}$ est la valeur de l'attribut qui est à la position *y* de l'exemple *x*.

$$Exemple\ x: [V_{x,-n}\ V_{x,-(n-1)}\ \ldots\ V_{x,0}\ \ldots\ V_{x,(n-1)}\ V_{x,n}]$$

La mesure globale de distance entre deux exemples A et B ($G\_dist(ex_A,ex_B)$) est basée sur les distances (($L\_dist(V_{A,k},V_{B,k})$)) entre les valeurs de chaque attribut. Pour chaque attribut (k), nous comparons ses valeurs ($V_{A,k}$ and $V_{B,k}$) dans les exemples: si les valeurs sont égales alors la distance est de 0; si les valeurs sont différentes alors la distance est de 1.

$$si\ (V_{A,k} = V_{B,k})\ alors\ \ L\_dist(V_{A,k},V_{B,k}) = 0,\ \ si\ (V_{A,k} = V_{B,k})\ alors\ \ L\_dist(V_{A,k},V_{B,k}) = 1$$

La mesure globale de distance (G_dist) entre deux exemples A (ex_A) et B (ex_B) est calculée comme suit, où W_k sont les poids donnés aux attributs de sorte que les attributs des mots les plus près du mot central soient les plus importants dans la mesure:

$$G\_dist(ex_A, ex_B) = \frac{\sum_{k=-n}^{n} W_k * L\_dist(V_{A,k}, V_{B,k})}{\sum_{k=-n}^{n} W_k}.$$

2.1.2 Stratégie de sélection des exemples

Tout d'abord, nous sélectionnons un échantillon initial représentatif de tous les exemples. Pour ce faire, nous utilisons l'algorithme des k-moyennes (Jain et al, 1999; Tang et al., 2002). Cet algorithme est basé sur la mesure de distance définie précédemment. Nous obtenons un ensemble composé de *nbc* groupes. Chaque groupe contient des exemples similaires. L'échantillon initial est constitué à partir d'une sélection aléatoire d'un pourcentage α d'exemples de chaque groupe. Cet échantillon nous sert à apprendre un modèle initial. Ensuite, les autres exemples sont sélectionnés de manière itérative. A chaque itération, nous utilisons deux critères pour la sélection : la confiance et la diversité.

L'utilisation du critère de la confiance consiste à choisir les exemples pour lesquels le modèle courant n'est pas satisfaisant. L'incertitude du modèle au sujet d'un exemple peut être due au fait que les exemples semblables sont sous-représentés dans l'ensemble d'apprentissage, ou bien que les exemples semblables sont intrinsèquement complexes. Nous tirons profit de la disponibilité de la confiance en classification du modèle courant. L'algorithme d'apprentissage de règles (par exemple PART (Frank, Witten, 1998)) assigne un degré de confiance à chaque règle induite. Pour chaque exemple non annoté, nous affectons le degré de confiance de la règle de laquelle il vérifie les conditions.

Le but du critère de la diversité (Shen et al., 2004) est de maximiser l'utilité inductive d'un ensemble d'exemples. Nous préférons les ensembles d'exemples hétérogènes. En choisissant un nouvel exemple non annoté, nous le comparons avec tous les exemples précédemment choisis dans l'ensemble courant. Si la similitude entre eux est au dessus d'un seuil β, l'exemple n'est pas ajouté dans l'ensemble. De cette façon, nous évitons de choisir les exemples trop semblables (valeur de similitude ≥ β) dans un ensemble.

La stratégie globale de sélection des exemples est décrite comme suit : les exemples non-annotés sont ordonnés selon la confiance. A chaque itération, nous choisissons un ensemble de *nb* exemples de la manière suivante: D'abord, nous sélectionnons un exemple candidat ($Exemple_i$) avec une valeur de confiance minimale. Ensuite, nous évaluons le critère de diversité et nous ajoutons l'exemple candidat $Exemple_i$ à l'ensemble si seulement $Exemple_i$ est assez différent de n'importe quel exemple précédemment inséré dans l'ensemble. Le seuil β est fixé à une valeur comprise entre la valeur maximale de similitude et la moyenne des similitudes par paires dans l'ensemble des exemples non annotés.

3 Validation expérimentale

Pour les expérimentations, nous avons utilisé un corpus de 600 résumés d'articles MEDLINE (Amrani et al, 2004) de biologie moléculaire. Ce corpus a été étiqueté par l'étiqueteur de Brill, puis par ETIQ. A partir de ce corpus, nous avons présenté à l'annotateur 4133 exemples où le

mot cible est étiqueté VBN et 3298 exemples où le mot cible est étiqueté NNS. Le nombre total d'exemples NNS était de 7708 dont 4410 sont des mots non-ambigus. L'annotateur a classé les mots cibles en VBN, JJ ou VBD pour le premier jeu d'exemples et NNS ou VBZ pour le deuxième jeu. Pour améliorer la précision, nous avons représenté les exemples comme suit : pour le cas des VBN, le mot cible est pris dans une fenêtre de 10 mots (5 mots à gauche et 5 mots à droite) et chaque mot du contexte est représenté par : son étiquette morphosyntaxique, le groupe auquel appartient son étiquette (verbal, nominal ou autre) et le mot est un verbe auxiliaire ou non. Pour le cas des NNS, le mot cible est pris dans une fenêtre de 6 mots : 3 mots à droite et 3 mots à gauche. En plus des attributs utilisés pour représenter les exemples des VBN, nous avons utilisé les suffixes et les préfixes les plus fréquents des mots. A partir de ces exemples, nous avons induit des règles avec les algorithmes PART (pour les VBN) et RIPPER (pour les NNS). Nous avons calculé les précisions de l'étiqueteur de Brill, d'ETIQ et d'ETIQ enrichi par les règles induites (voir Figure 1). La précision des règles induites a été calculée par la méthode «validation croisée 10 fois».

Confusion / %de précision	Brill	Brill+ ETIQ	Brill+ ETIQ+Règles induites
VBN→VBN-VBD-JJ (PART)	54	76	94
NNS→NNS-VBZ (RIPPER)	92	96	97,5

Figure 1 : Précisions obtenues sur deux jeux d'exemples de confusions d'étiquettes.

Nous avons appliqué la stratégie de l'apprentissage actif aux exemples correspondent à l'ambiguïté VBN-VBD-JJ. Parmi les 4133 exemples disponibles, nous avons pris 3100 exemples pour l'apprentissage actif, et 1033 exemples pour le test. Le modèle initial a été construit à partir de 423 exemples. A chaque itération, nous avons sélectionné 100 exemples. L'expérience a été répétée 5 fois. La courbe (figure 2) représente les valeurs moyennes obtenues. La précision obtenue avec tous les exemples (3100) est de 93,5.

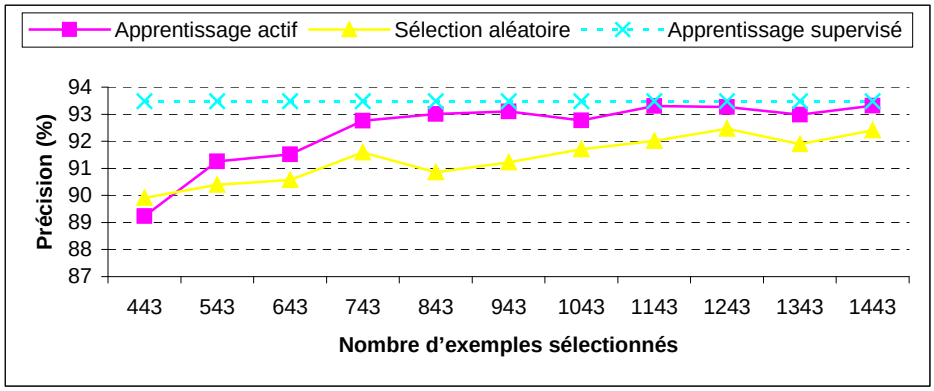

Figure 2 : Apprentissage actif *versus* sélection aléatoire.

4 Conclusions et perspectives

Dans le cadre d'une méthodologie globale pour l'étiquetage morphosyntaxique des corpus de spécialité, nous avons complété notre approche pour traiter efficacement les problèmes

d'étiquetage pointus. Après la détection des contextes ambigus et particuliers, les mots cibles sont annotés (exemples). A partir de ces exemples, nous avons induit des règles de correction. Nous avons obtenu une nette amélioration de la précision d'étiquetage. Pour réduire le nombre d'exemples à annoter, nous avons utilisé l'apprentissage actif avec une stratégie de sélection basée sur la confiance et la diversité. En annotant seulement un tiers des exemples, nous obtenons des performances équivalentes à celles obtenues en annotant tous les exemples. Nous étendrons cette approche à d'autres classes d'ambiguïtés. Nous envisageons également de considérer d'autres méthodes d'apprentissage, par exemple : la Programmation Logique Inductive. La combinaison optimale des règles obtenues par différents algorithmes pourrait améliorer les performances. Le critère de diversité, utilisé pour l'apprentissage actif, peut être amélioré en utilisant une valeur de similitudes (β) optimale.

Références

AMRANI, A., AZE, J., KODRATOFF, Y. (2005) ETIQ: Logiciel d'aide à l'étiquetage morpho-syntaxique de textes de spécialité. *Dans la revue RNTI, numéro spécial EGC'2005*.

AMRANI, A., KODRATOFF, Y., MATTE-TAILLIEZ, O. (2004) A Semi-automatic System for Tagging Specialized Corpora, *PAKDD 2004*, Sydney, LNAI, Vol. 3056, pp 670-681.

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

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

ENGELSON, S.A., DAGAN, I. (1999) Committee-Based Sample Selection for Probabilistic Classifiers. *Journal of Artifical Intel-ligence Research*.

FRANK, E., WITTEN, I.H. (1998) Generating Accurate Rule Sets Without Global Optimization, Shavlik, J. Eds., *Proceedings of the 15th ICML*, Madison, Wisconsin, pp 144-151.

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323.

PAROUBEK, P., RAJMAN, M. (2000) Chapitre 5: Etiquetage morpho-syntaxique, Ingénierie des Langues, sous la direction de Jean-Marie Pierrel, Collection "*Information Commande Communication*", aux Editions Hermes Science, 2000 pp 131-148.

QUINLAN, J.R (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann San Mateo.

SHEN, D., ZHANG, J., SU, J., ZHOU, G., TAN, C-L. (2004) Multi-Criteria-based Active Learning for Named Entity Recognition. *Proceedings of ACL* 2004.

TANG, M., LUO, X., ROUKOS, S., 2002. Active Learning for Statistical Natural Language Parsing. *In Proceedings of the ACL* 2002.

VALLI, A., & VERONIS, J. (1999). Etiquetage grammatical de corpus oraux: problèmes et perspectives. *Revue Française de Linguistique Appliquée*, IV(2), 113-133.