Conférence TALN 2000, Lausanne, 16-18 octobre 2000

Vers un apprentissage en TALN dépendant du type de Texte

Gabriel ILLOUZ

Limsi -CNRS
BP133
91403 Orsay Cedex
France
gabrieli@limsi.fr

Résumé

Dans cet article, nous présentons la problématique de l'hétérogénéité des données textuelles et la possibilité d'utiliser cette dernière pour améliorer les traitements automatiques du langage naturel. Cette hypothèse a été abordée dans (Biber, 1993) et a donné lieu à une première vérification empirique dans (Sekine, 1998).

Cette vérification a pour limite de ne s'adapter qu'à des textes dont le type est explicitement marqué. Dans le cadre de textes tout venant, nous proposons une méthode pour induire des types de textes, apprendre des traitements spécifiques à ces types puis, de façon itérative, en améliorer les performances.

mot clés : annotation morpho-syntaxique, type de texte, linguistique de corpus, apprentissage, classification.

Introduction

En Traitement Automatique du Langage Naturel (TALN), les tâches à réaliser suivent un développement de plus en plus normé. On doit tout d'abord les **définir** (pour illustration, nous prendrons comme exemple récurrent l'assignation de parties du discours [1], puis **produire un étalon** de référence (Multitag, Brown Corpus,...), enfin **mettre au point des traitements** automatiques selon différentes technologies (Modèles de Markov Cachés, à base de règles,...) et enfin **les évaluer** (Adda *et al.*, 1999). Cette approche se retrouve dans différents domaines du TALN : désambiguïsation sémantique (Kilgariff, 1998), annotation d'actes de dialogue (Samuel *et al.*, 1999), reconnaissance d'entités nommées (Shumeet *et al.*, 1999), compréhension de messages (Hirshman, 1998), voire des tâches complexes comme l'extraction d'information en contexte multilingue (Harman, 1998).

La nécessité d'avoir des traitements adaptés au corpus à traiter a été soulignée dans (Péry-Woodley, 1995) et (Habert *et al.*, 2000). Or, l'évolution actuelle du TALN permet d'évaluer

[1] Son but est d'attribuer une partie du discours à chaque mot, selon le contexte (par exemple, annoter soit pronom soit déterminant le mot *le)*.

l'effet de l'hétérogénéité au sein des données textuelles et ce que cette dernière implique pour les traitements de corpus. En effet, les traitements produits par cette évolution relèvent souvent de l'apprentissage automatique. On peut alors tenter de les adapter à un *type de textes* : un sous-ensemble moins hétérogène que l'ensemble complet.

Dans cet article, les travaux abordant l'adaptation aux types de textes seront présentés. Des méthodes seront alors proposées pour obtenir des TALN adaptatifs aux types de textes de manière automatique ; elles seront alors mises en œuvre sur le BROWN CORPUS avec un assignateur de parties du discours stochastique : le treeTagger. Enfin, nous conclurons sur les limites observées lors des expériences ainsi que sur les perspectives des méthodes proposées.

1. Survol de l'existant

Par la suite, nous distinguerons *genre de textes* lorsque la classification des textes est *a priori* (donnée), de *type de textes* lorsqu'il est induit des données.

Dans (Slocum, 1986), sur un corpus en allemand, l'auteur montre que les jeux de règles pour l'analyse syntaxique devraient être différents selon le sous-langage utilisé. Ce corpus était composé de deux manuels écrits par des ingénieurs et de deux brochures écrites par des commerciaux. Il propose aussi un moyen de caractériser le type "manuel" (impératif, acronymes, suppression de déterminants) du type "brochure" (utilisation de pronoms, phrases longues, syntaxe plus riche). Cette approche semble particulièrement intéressante car elle permet de définir la notion de traitements spécialisés.

Dans (DeRose, 1988), des performance s différentes sont obtenues pour un seul étiqueteur selon les genres de textes présents dans le Brown Corpus. Ce résultat montre la sensibilité d'un traitement au type de textes. À partir des résultats de la campagne d'évaluation GRACE (Adda *et al.*, 1997), il a été montré que des systèmes d'annotation morpho-syntaxiques pouvaient être complémentaires, c'est-à-dire qu'un texte peut entraîner des variations de performances (par rapport à sa performance moyenne) positives pour un, négatives pour un autre.

Dans (Biber, 1988), l'auteur propose une méthode pour retrouver le genre de textes à l'aide de statistique multi-dimensionnelle, mais ne dit rien quant à leur corrélation avec des traitements automatiques. Dans (Biber, 1993), il montre que le genre du texte peut avoir un effet sur les résultats d'un étiqueteur stochastique. Pour ce faire, il prend deux genres présents dans le LOB corpus (texte de fiction et d'exposition). Il montre les différences de probabilités pour les séquences de deux étiquettes (par exemple ADJECTIF suivi de NOM) dans ces deux sous-ensembles.

Suivant ce paradigme, dans (Sekine, 1998), l'auteur entraîne un analyseur syntaxique pour chaque genre de textes présent dans le Brown Corpus. De là, il l'évalue sur un ensemble de tests, et conclut que les performances sont toujours meilleures si la classe sur laquelle l'analyseur a été entraîné est la même que celle sur laquelle il est testé.

Ces travaux confirment l'hypothèse suivante : il est possible de s'adapter aux types de données traitées pour améliorer les performances. Néanmoins, aucune méthode n'est donnée lorsque l'on cherche à s'adapter à des textes non marqués en genre. Le cas est très fréquent lorsque l'on veut enrichir d'importants volumes de données textuelles, en recherche d'information par exemple.

Les méthodes et expériences qui suivent ont donc pour but de construire de façon automa-

tique et optimale des traitements spécialisés.

2. Présentation des méthodes Adaptatives

Dans cette partie, trois méthodes sont envisagées. La première est celle utilisée dans Sekine reposant sur un corpus où le genre de chaque texte est donné. La seconde est une méthode induisant des types de textes à partir des données. La troisième est une méthode qui, partant d'une partition, cherche à améliorer itérativement celle-ci.

2.1. Partitionnement existant

Le modèle de TALN Adaptatif (TALNA) sous-jacent au travail de Sekine est présenté en figure 1. Ce modèle se décompose en deux phases : une d'apprentissage et une de test ou d'utilisation. Durant la phase d'apprentissage, la classification partitionne le corpus en sous-corpus et des traitements spécialisés sont créés par entraînement sur chaque partition. Durant la phase de test (d'utilisation), pour un nouveau texte, on utilise le traitement spécialisé correspondant ,dans la classification trouvée, à la classe à laquelle ce texte appartient.

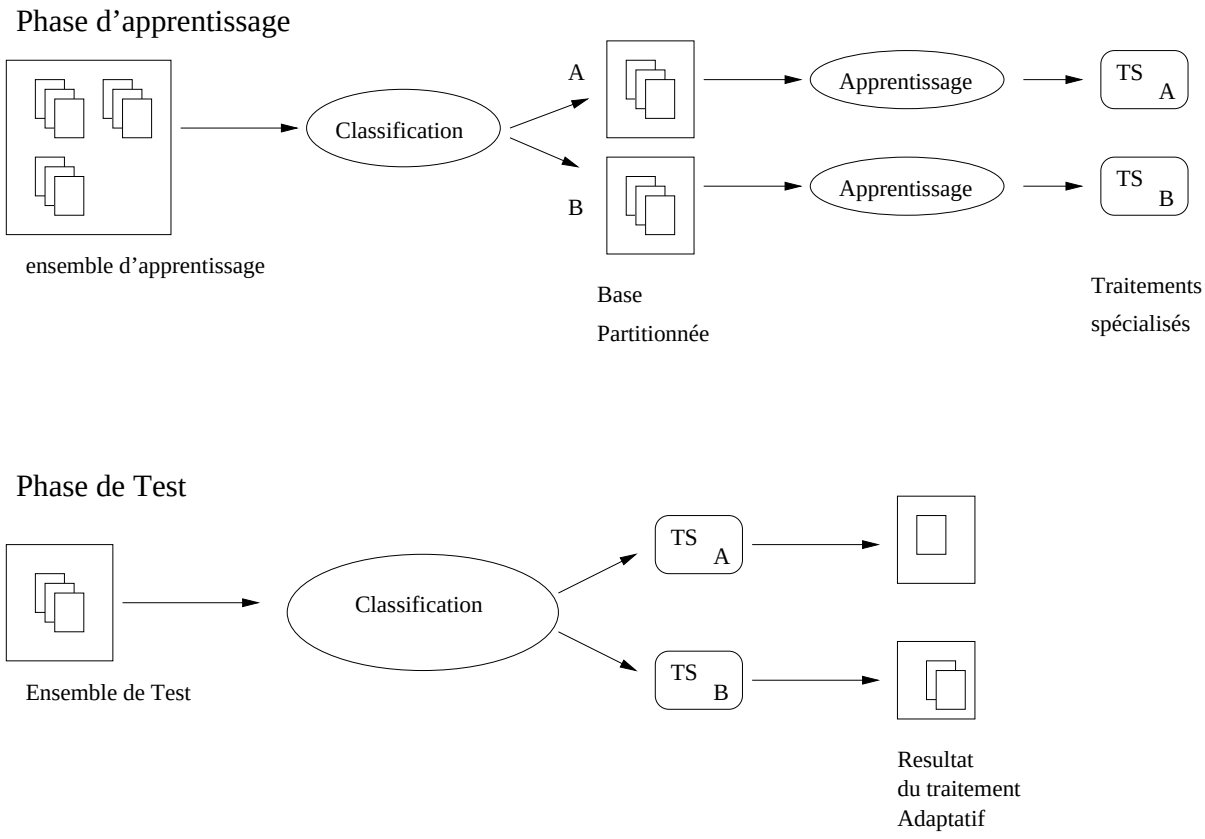


FIG. 1 – modèle de TALN Adaptatif

2.2. Partitionnement induit

Dans l'objectif de traiter du texte tout-venant, la méthode utilisée est la même si ce n'est que la classification est induite au lieu d'être *a priori*. Pour créer des *types de textes*, on a besoin d'un classifieur, et d'un ensemble de traits pouvant être extrait de chaque texte.

Le choix des traits est un enjeu important dans le cadre des données textuelles et nombres de choix sont possibles, partant de traits surfaciques (comme les caractères) jusqu'à des traits linguistiques (comme par exemple la fréquence des verbes transitifs sans objet). Le coût de l'extraction doit aussi être prise en compte, comme cela est souligné dans (Karlgren & Cutting, 1994). Ce choix doit aussi prendre en compte la possibilité de corrélation entre le jeu de traits et le traitement qu'on a pour objectif d'améliorer.

2.3. Partitionnement itératif

Pour ajuster la classification induite à la tâche demandée, nous proposons une méthode itérative pour trouver la partition optimale. Durant la phase d'entraînement, chaque texte est évalué par l'ensemble des traitements spécialisés. Si le meilleur traitement n'est pas celui qui correspond au type existant pour le texte, ce dernier est reclassé. L'algorithme est donné ci-dessous :

Algorithme de recherche de la partition optimale partant d'une partition donnée

Soit $P = \{P_i\}$ une partition de la base de textes $T = \{T_j\}$

On cherche une partition P'

$E = \{E_k\}$: l'ensemble des traitements possibles

une fonction d'apprentissage $app\ :\ P \rightarrow E$

pour chaque $P_i \in\ P$

$E_i \leftarrow app(P_i)$

pour chaque $T_i\ \in\ T$

si $\arg\max_k eval[E_k(T_i)]$ **alors** $P'_k \ni T_i$

On recommence, jusqu'à ce que P et P' soient identiques.

Pour un nouveau texte, il suffit de chercher le traitement qui lui correspond et de l'appliquer. L'algorithme correspondant est donné ci-dessous.

Algorithme de recherche pour un nouveau texte du traitement spécialisé

Soit $T_a$le nouveau texte à traiter

Soit E le traitement spécialisé recherché

$P = \{P_i\}$ une partition de la base de textes $T = \{T_i\}$

une fonction $nvoisins\ :\ T \rightarrow P$, qui associe à un nouveau texte une le nom d'un sous ensemble P_i de la partition P, reposant sur une mesure de distance $d\ :\ \Re^m \times \Re^m \rightarrow \Re$

si $P'_k \equiv nvoisins(T_a)$ **alors** $E \leftarrow E_k$

Les propriétés à vérifier lors de cette phase sont :

(1) la performance sur une partition est meilleure avec le traitement correspondant qu'avec un autre.

(2) la performance globale de cette méthode est meilleure que la performance du traitement entraîné sur l'ensemble des textes. Nous nommerons ce dernier *traitement généraliste*.

3. Expériences et résultats

3.1. Données utilisées

Dans cet article, le BROWN Corpus annoté par le PENN TREE BANK PROJECT (Marcus Mitchell, 1993) est utilisé avec sa classification entre textes à visée informative *(Informative prose)* et textes de fiction *(Imaginative prose)* comme point de repère. Il est composé de 500 échantillons dont chacun comprend environ 2000 mots. Dans le tableau 1, la taille de chaque sous-ensemble est donnée ainsi que la séparation entre Apprentissage et Test. Les tailles sont données en nombre de mots.

	Apprentissage	Test
Informative Prose	767 260	97 237
Imaginative Prose	272 570	33 736
Total	1 039 830	130 973

TAB. 1 – Taille des sous-ensembles en nombre de mots.

3.2. Classifications existantes

Pour donner un point de comparaison, l'expérience de (Sekine, 1998) est ici reproduite, en utilisant le TreeTagger (Schmid, 1995), un étiqueteur stochastique fondé sur les arbres de décision. Celui-ci est entraîné sur chaque sous-ensemble pour produire deux traitements spécialisés. Leurs résultats sont comparés à ceux obtenus par le traitement généraliste. Ils sont donnés sous forme de pourcentages d'étiquettes correctes ainsi qu'en pourcentages de phrases où toutes les étiquettes sont correctes dans le tableau 2. On remarque que les propriétés (1) et (2) sont vérifiées, c'est-à-dire que les traitements spécialisés ont des performances comparables ou meilleures que celles du traitement généraliste entraîné sur l'ensemble (donc sur plus de données).

Les spécificités de la tâche d'étiquetage morpho-syntaxique (faibles différences des résultats et le niveau des performances) rendent difficile l'utilisation des tests statistiques classiques permettant de vérifier si les différences de résultats sont significatives. D'autres approches sont à l'étude.

Apprentissage sur :	Info.	Imag.	Total
Test sur Inf.	94.04	93.16	94.04
Test sur Ima.	93.83	94.21	94.16

(a)

Info.	Imag.	Total
34.38	29.45	33.90
48.49	50.27	49.36

(b)

TAB. 2 – Précisions en pourcentages (a) d'étiquettes correctes (b) de phrases correctes

3.3. Classifications induites

Une autre approche possible est de considérer qu'une classification induite à base d'un ensemble de traits extraits des textes peut permettre d'obtenir le même type de partitionnement.

3.3.1. Mise en œuvre

Avant d'utiliser des traits plus évolués et de mettre au point les outils correspondants, l'induction sera effectuée sur le jeu de caractères. Cette information peut sembler très pauvre, elle fournit néanmoins des indices sur la forme du texte. En effet, la fréquence du point est corrélée à la longueur des phrases (ou plus précisément à une combinaison de la longueur des phrases, de la présence d'acronymes, et de nombres). L'espace est un indicateur de la longueur des mots, la proportion de lettres capitales, le nombre de noms propres. À titre d'exemple, les échantillons textuels sont représentés en figure 2 dans l'espace à deux dimensions formé par les fréquences des caractères espace et point.

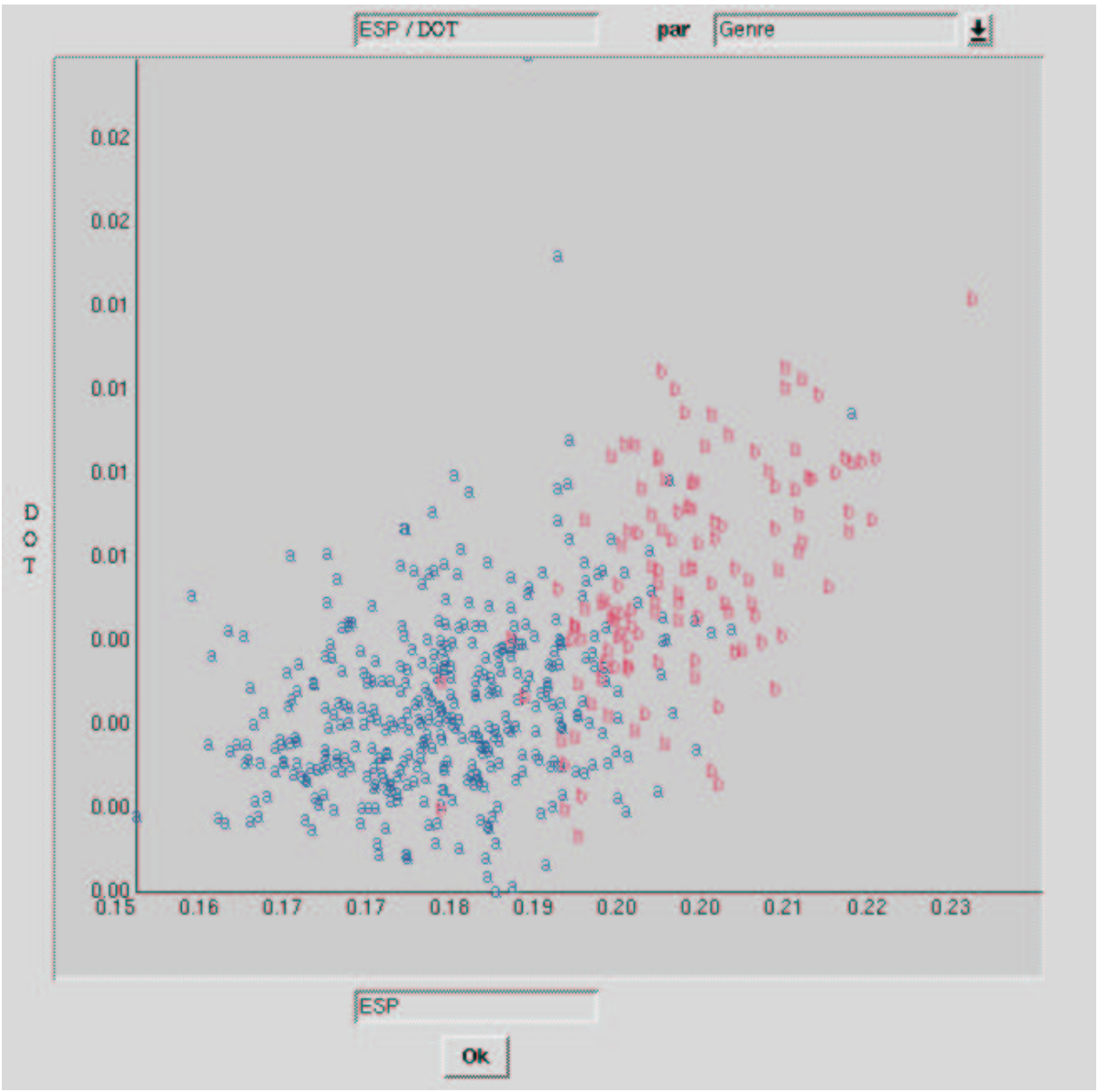


FIG. 2 – Espace et Point pour les échantillons Informative(a) et Imaginative(b)

En utilisant cet ensemble de traits et un classifieur inductif décrit dans (Jardino & Beaujard, 1997), on répartit 500 échantillons dans deux classes, C1 et C2. Dans le tableau 3, les résultats de cette classification sont présentés en croisé avec la classification existante pour permettre

la comparaison. On retrouve une classification proche de la classification existante, seuls 54 échantillons sont classés différemment, si l'on suppose que C1 correspond à Imaginative et C2 à Informative. Sur la figure précédente, ces 54 textes se trouvent précisément à la frontière entre les deux classes.

	Informative Prose	Imaginative Prose
C1	53	125
C2	321	1

TAB. 3 – Nombre d'échantillons selon les classifications

3.3.2. Résultats

La méthode présentée pour la classification existante est utilisée sur cette partition. Nous avons gardé la même répartition entre classes d'apprentissage et de test. Les tailles pour les sous-ensembles sont différentes. La taille de chaque sous-ensemble est donnée en tableau 4.

	Apprentissage	Test
C1	385 822	43 095
C2	654 008	87 886
Total	1 039 830	130 973

TAB. 4 – Taille des deux classes induites en nombre de mots

Deux traitements spécialisés, reposant sur le TreeTagger, sont entraînés sur l'ensemble correspondant. Leurs résultats sont donnés dans le tableau 5.

Apprentissage sur :	C1	C2	Total
Test sur C1	94.05	93.83	94.19
Test sur C2	93.47	93.98	94.03

(a)

C1	C2	Total
47.99	47.15	48.10
29.39	32.75	32.80

(b)

TAB. 5 – Précision en pourcentages (a) d'étiquettes correctes (b) de phrase correctes

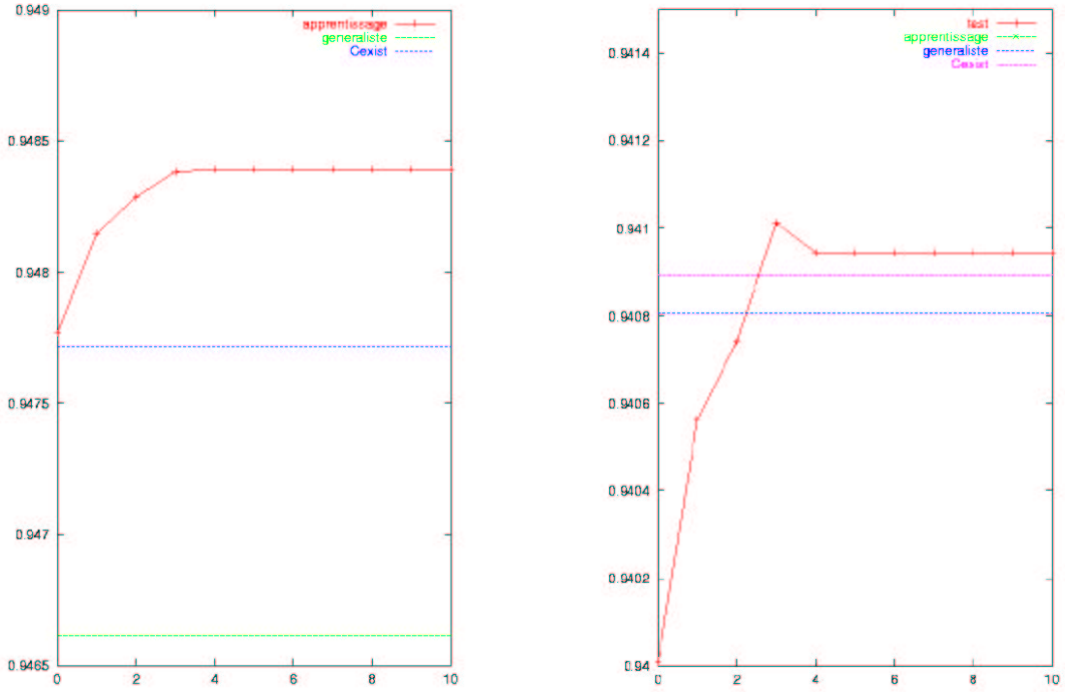
On constate que les résultats sont moins bons que précédemment. Les deux traitements spécialisés ne réussissent pas à faire aussi bien que le généraliste. Néanmoins, on retrouve la propriété (1) : les traitements sont meilleurs lorsque classes de test et classes d'apprentissage sont les mêmes. De cette expérience, on peut conclure qu'il est nécessaire de mettre au point le jeu de traits utilisé pour obtenir une corrélation forte entre performance et classification.

3.4. Classifications induites itérativement

De l'expérience précédente, on note l'importance du choix des traits. Nous utilisons alors le modèle proposé en section 2.4 pour augmenter la corrélation entre performance et classification, en utilisant une méthode des *plus proches voisins*[2] avec des poids pondérés par l'inverse des distances comme proposé dans (Mitchell, 1997), et une mesure de distance euclidienne.

[2]On prend les n plus proches voisins dans les espace à k dimension, et on déclare le point examiné du type majoritaire, on peut donc donner des poids représentant l'importance de chaque point. Dans cette expérience n vaut 7.

Les résultats sont donnés en figure 3, où est représentée pour chaque itération, l'évaluation sur le corpus d'apprentissage et sur celui de test des performances globales du traitement généraliste, du traitement reposant sur un partitionnement existant, du partitionnement itératif. L'étape 0 du traitement itératif correspond à la partition induite. L'apprentissage s'améliore jusqu'à ce qu'il y ait convergence, ce qui est l'objectif de l'algorithme présenté. Le point le plus important est que l'on a pas obtenu de sur-apprentissage fort, c'est-à-dire quand le traitement est tellement spécialisé qu'il améliore les performances d'apprentissage en diminuant celles du test. De plus, on note sur le test que cette méthode permet de dépasser les performances obtenues par le partitionnement existant.



(a) (b)

FIG. 3 – Résultats pour la phase d'apprentissage (a) et de test (b)

4. Conclusion et Perspectives

Dans cet article, nous avons vu la possibilité de partitionner les données pour améliorer les résultats d'étiquetage en Parties du Discours (première expérience). Nous avons tenté d'obtenir des résultats similaires avec le jeu de caractères en utilisant une méthode induite directe. L'utilisation de celle-ci ne s'est pas révélée suffisante. Nous pouvons faire l'hypothèse que l'un des facteurs sous-tendant ces résultats est le choix du jeu de traits utilisé. Nous nous proposons d'expérimenter d'autres jeux de traits a priori plus corrélés avec les performances d'étiquetage comme par exemple : les proportions de parties du discours, ou les séquences de n parties du discours. Toutefois, la question de l'extraction de ces traits doit être traitée avec le plus grand soin. Utiliser les parties du discours marquées dans le corpus de référence revient à utiliser une information nécessairement inconnue lors de l'apprentissage. Si un étiqueteur sert à les produire, pourquoi celui-ci ne serait-il pas sensible au type de textes ? Une méthode envisageable est d'utiliser uniquement les mots du lexique non-ambigus en parties du discours.

Une méthode tentant de découvrir la partition optimale a donné des résultats plus prometteurs. Nous devons tout d'abord souligner que, par rapport à la taille des données employées, les résultats présentés se situe dans l'état de l'art. De plus, contrairement à ce qui est prévu par Schmid (Schmid, 1994) : *plus la taille du corpus d'entraînement est grande, meilleurs sont les résultats*, nous arrivons à augmenter les performances en diminuant la taille d'entraînement. Cela fournit une piste pour améliorer encore les performances. Mais cela nécessite de travailler sur un corpus dans lequel le genre de textes est varié, et où chaque genre est suffisamment représenté, comme par exemple le BNC (British National Corpus).

Dans le cadre de recherche en typologie de textes, et en se situant du point de vue d'une tâche de TAL, la méthode proposée fournit un moyen objectif de comparer les jeux de traits employés pour la classification. En effet, ici le maximum de précision possible pour l'ensemble de test n'est pas atteint, celui-ci est facilement calculable en utilisant la méthode du jeu d'apprentissage sur le jeu de test : c'est à dire en évaluant pour chaque texte les performances de chaque traitement et en gardant la meilleure.

Références

ADDA G., LECOMTE J., MARIANI J., PAROUBEK P. & RAJMAN M. (1997). Les procédures de mesures automatique de l'action grace pour l'évaluation des assignateurs de parties de discours pour le fran cais. In *JST-FRANCIL*, Avignon.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M. & LECOMTE J. (1999). Métrique et premiers résultats de l'évaluation grace des éttiquetteurs morpho-sytaxiques pour le fran cais. In *TALN*, Cargèse.

BIBER D. (1988). *Variation accross speech and writing*. Cambridge : Cambridge University Press.

BIBER D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 243–258.

DEROSE S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, **14**(1), 31–39.

HABERT B., ILLOUZ G., LAFON P., FLEURY S., FOLCH H., HEIDEN S. & PRÉVOST S. (2000). Profilage de textes : cadre de travail et expérience. In *JADT2000(5th International Conference on the Statistical Analysis of Textual Data)*, Lausanne, Suisse.

HARMAN D. (1998). The text retrieval conference (trecs) and the cross-language track. In *LREC*, Grenade, Espagne.

HIRSHMAN L. (1998). Language understanding evaluations : Lessons learned from muc and atis. In *LREC*, Grenade, Espagne.

JARDINO M. & BEAUJARD C. (1997). Rôle du contexte dans les modèles de langage 'n-classes' application et évaluation sur mask et railtel. In *Actes des 1 ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'Aupelf-Uref*.

KARLGREN J. & CUTTING D. (1994). Recognizing text genres with simple metrics using discriminant analysis. p. 1071–1075, Kyoto, Japon : COLING.

KILGARIFF A. (1998). Senseval : An excercise in evaluating word sense disambiguation programs. In *LREC*, Grenade, Espagne.

MARCUS MITCHELL, BEATRICE SANTORINI M. A. M. (1993). Building a large annoted corpus of english : the penn treebank. *Computational Linguistics*, **19**(1), 313–330.

MITCHELL T. M. (1997). *Machine Learning*. McGraw-Hill.

PÉRY-WOODLEY M.-P. (1995). Quel corpus pour quels traitements automatiques ? *t.a.l. (traitement automatique des langues)*, **36**(1-2), 213–232.

SAMUEL K., CARBERRY S. & VIJAY-SHANKER K. (1999). Automatically selecting useful phrases for dialogue act tagging. In N. CERCONE, Ed., *PACLING*, Waterloo, Ontario, Canada.

SCHMID H. (1994). Probalistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New methods in Language Processing*, Manchester, UK.

SCHMID H. (1995). Improvements in part of speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*.

SEKINE S. (1998). The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, p. 96–102, Washington.

SHUMEET B., MITTAL V. O. & SUKTHANKAR R. (1999). Applying machine learning for high performance named-entity extraction. In N. CERCONE, Ed., *PACLING99*, Waterloo, Ontario, Canada.

SLOCUM J. (1986). How one might automatically identify and adapt to a sublanguage. In R. GRISHMAN & R. KITTREDGE, Eds., *Analyzing Language in Restricted Domains*, chapter 11, p. 195–210. Hillsdale, NJ : Lawrence Erlbaum Ass.