Conférence TALN 2000, Lausanne, 16-18 octobre 2000

Enrichissement automatique de lexique de noms propres à partir de corpus

F. Béchet *, A. Nasr **, F. Genet*

* LIA - Université d'Avignon - BP1228 - Avignon Cedex 9
* LIM - Université Aix-Marseille 2 - 163, avenue de Luminy - 13288 Marseille Cedex 9

Résumé

Cet article présente une méthode d'étiquetage sémantique de noms propres fondé sur la technique des arbres de décision. Ces derniers permettent de modéliser les éléments saillants dans les contextes d'occurrence de noms propres d'une classe donnée. Les arbres de décision sont construits automatiquement sur un corpus d'apprentissage étiqueté, ils sont ensuite utilisés pour étiqueter des noms propres apparaissant dans un corpus de test. Les résultats de l'étiquetage du corpus de test est utilisé pour enrichir un lexique de noms propres. Ce dernier peut être utilisé à son tour pour réestimer les paramètres d'un étiqueteur stochastique. Nous nous intéressons en particulier au cas où le corpus de test a été glané sur le Web.

1. Introduction

Cette étude se situe dans le cadre de l'enrichissement automatique de lexiques de noms propres. L'enrichissement est effectué à l'aide d'un arbre de décision qui permet d'attribuer une étiquette sémantique à un nom propre, en fonction de son contexte d'apparition. Les lexiques produits par une telle méthode contiennent, pour chaque entrée e, une série de couples <C,Nb>où Nb représente le nombre fois où e s'est vue attribuer l'étiquette C. De tels lexiques peuvent être utilisés pour estimer les paramètres du modèle d'un étiqueteur statistique de type N-Classe.

Nous nous intéressons ici aux noms propres dans un contexte journalistique bien que la méthode puisse être étendue à n'importe quelle catégorie de mots.

Le choix de se focaliser sur les noms propres se justifie à deux niveaux : quantitatif et qualitatif.

– D'un point de vue quantitatif, même s'ils ne représentent qu'une faible part des occurrences des mots de notre corpus (3,67%), les noms propres sont par contre majoritaires dans les formes inconnues de notre dictionnaire de référence. Une étude réalisée sur un corpus de textes issus du journal *Le Monde Diplomatique* (Béchet & Yvon, 2000) a montré que 72% des formes inconnues d'un dictionnaire de 265K mots sont potentiellement des noms propres. De plus, la même étude montre que 30% des phrases contiennent au moins l'un de ces mots inconnus. Ainsi, accroître automatiquement des lexiques de noms

propres permet d'augmenter significativement la couverture des systèmes utilisant de tels lexiques.
- D'un point de vue qualitatif, les noms propres sont des unités particulièrement intéressantes dans des applications telles que la recherche documentaire, l'alignement de textes multilingues et la compréhension automatique de textes popularisée par les campagnes d'évaluation MUC. Dans ce cadre, la tâche d'extraction d'entités nommées existe maintenant aussi bien dans les campagnes MUC que dans les campagnes DARPA d'évaluation des systèmes de transcription de données sonores (Broadcast News). Cette tâche a bien évidemment besoin de vastes dictionnaires de noms propres étiquetés.

La technique présentée ici se décompose en deux étapes : tout d'abord un corpus d'apprentissage étiqueté est utilisé pour construire un arbre de décision d'un type particulier, appelé *Arbre de classification sémantique*, décrit en 2. Ces arbres permettent de classer les noms propres, en fonction de leur contexte d'occurrence, en 5 catégories : prénom (PRENOM), nom de famille (FAMILLE), ville (VILLE), pays (PAYS) et organisation (ORG). Le processus de construction des arbres est décrit en 3.

L'arbre de décision est utilisé pour étiqueter des noms propres dans un corpus de test lorsque ces derniers apparaissent dans des contextes discriminants. Ainsi, l'arbre peut être vu comme un filtre permettant de n'étiqueter un nom propre que lorsqu'il apparaît dans un contexte discriminant. Enfin, les étiquettes attribuées aux noms propres peuvent être utilisées pour mettre à jour un lexique. Ces deux modes d'utilisation de l'arbre de décision sont décrits en 4. L'influence sur la qualité de l'étiquetage de nouveaux exemples glanés sur le Web est évaluée en 5.

2. Les arbres de classification sémantiques

Les arbres de classification sémantiques (ACS) ont été introduits par (Kuhn & de Mori, 1996) comme un moyen général de *classifier, à partir d'un corpus de chaînes étiquetées, des chaînes non encore vues*. Nous nous en servirons comme un moyen d'attribuer à un nom propre une étiquette sémantique en fonction de groupes nominaux (GN) dans lesquels il apparaît.

Chaque nœud de l'arbre est associé à une expression régulière[1] construite sur un alphabet composé d'items lexicaux, d'étiquettes de parties de discours, des deux symboles $<$ et $>$ matérialisant le début et la fin d'un GN et du symbole $+$[2]. Une expression régulière est vue comme une question, celle de l'appartenance d'un GN au langage reconnu par l'expression régulière. De plus, chaque feuille de l'arbre est associée à une distribution de probabilités sur le jeu des 5 étiquettes sémantiques. L'étiquette de probabilité maximale dans cette distribution est appelée le *candidat* de la feuille.

Un parcours de l'arbre depuis la racine jusqu'à une feuille s'effectue en répondant aux différentes questions se trouvant le long du parcours. Le choix d'une question dépend de la réponse à la question précédente. Lorsqu'à l'issue d'un parcours un GN comportant un nom propre inconnu atteint une feuille, la distribution de cette dernière permet d'estimer la probabilité que le nom propre appartienne à chacune des classes du jeu d'étiquettes.

Une partie d'un ACS a été représenté dans la figure 1. Chaque nœud est étiqueté par l'expression régulière à laquelle il correspond. Considérons le GN *le président du groupe Lafarge*

1. Il s'agit en fait d'un sous-ensembles des expressions régulières.

2. La signification du symbole $+$ est un peu différente de celle qu'il possède dans le formalisme des expressions régulières, il correspond ici à une séquence quelconque composée d'un mot au moins.

contenant un nom propre inconnu, ici *Lafarge*, auquel on désire attribuer une étiquette sémantique. La question correspondant à la racine porte sur la présence du nom *président* suivi et précédé d'un mot au moins dans le GN. La réponse étant positive, la question suivante est celle correspondant à l'expression régulière < +président+groupe+> qui donne lieu à une réponse positive, menant à une feuille donnant une estimation de l'appartenance de du nom *Lafarge* aux différentes classes sémantiques dans un tel contexte.

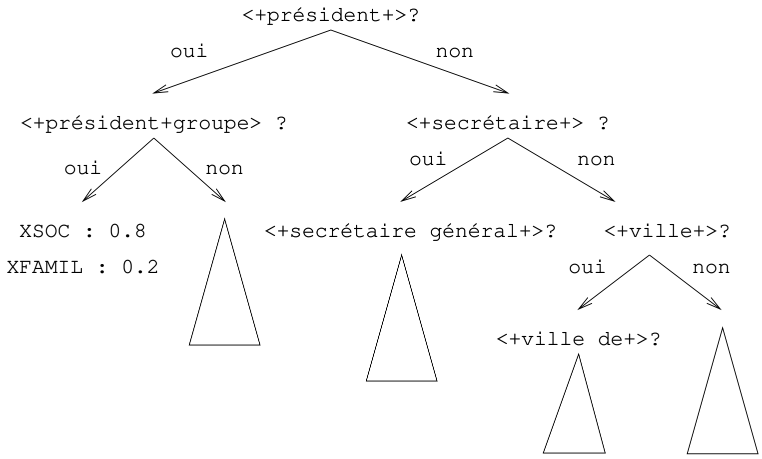


FIG. 1 – *Un arbre de classification sémantique*

3. Construction de l'arbre de décision

La construction de l'arbre de décision nécessite un corpus d'exemples, un ensemble de questions, un critère de division d'un nœud et une condition d'arrêt. Ces différents éléments sont décrits ci-dessous.

3.1. Le corpus d'exemples

Le corpus d'exemples est constitué de GN comprenant chacun un nom propre appartenant à une classe sémantique connue. Les GN sont reconnus grâce à un analyseur syntaxique de surface puis sont étiquetés par la classe du nom propre qu'ils contiennent. Les limites du GN définissent la taille du contexte duquel seront sélectionnés les éléments les plus pertinents permettant de caractériser les noms propres d'une catégorie donnée. Le GN comme unité de contexte pertinent est un compromis entre d'une part une *fenêtre* de taille arbitraire autour d'un mot, s'étendant souvent au mot ou aux deux mots entourant le nom propre et d'autre part la phrase complète, qui introduit en général trop de bruit dans le processus d'apprentissage. Afin de limiter les GN incorrects dus à de mauvais rattachements prépositionnels lors de l'analyse syntaxique, la couverture de la grammaire des GN à été limitée aux GN comportant au plus un rattachement prépositionnel. Les GN du corpus d'exemples ont une longueur comprise entre 1 et 12 mots.

La construction du corpus d'exemples s'effectue en quatre étapes : Le corpus d'apprentissage est étiqueté grâce à un étiqueteur stochastique (Spriet & El-bèze, 1995) dont le jeu d'étiquettes comporte des étiquettes morpho-syntaxique (PREP, ADJ...) ainsi que des étiquettes sémantiques (PAYS, VILLES...). Le corpus étiqueté est ensuite analysé à l'aide d'un analyseur à états finis afin de détecter les GN. Les GN comportant un nom propre (étiqueté comme tel lors de la première étape) sont alors stockés dans le corpus d'exemple. Chaque GN est associé à l'étiquette du nom propre qu'il contient. Finalement, seuls les GN dont le nombre d'occurrences est supérieur à un seuil sont conservés.

Dans la phrase *le président de SONY a déclaré dans une interview . . .* , le GN *le président de SONY* est détecté. Il sera stocké dans le corpus d'exemple, si son nombre d'occurrences est supérieur au seuil, sous la forme :
`(le, DET) (président, N) (de, PREPDE) (X, NP) = ORG`

A l'issue de ce processus, un ensemble de GN pour chaque classe de noms propres a été constitué. Ce corpus d'exemples constitue le corpus d'apprentissage sur lequel l'arbre de décision sera construit.

3.2. L'ensemble de questions

L'aspect original des ACS est la façon dont les expressions régulières associées aux nœuds de l'arbre sont générées. Lors du processus de construction de l'arbre, chaque nœud est associé à une partie du corpus d'exemples ainsi qu'à une expression régulière appelée la *structure connue* (SC). Au début de la construction de l'arbre, la racine est associée à l'intégralité du corpus d'exemples ainsi qu'à la SC $< + >$, qui reconnaît l'ensemble des GN possibles. Les différentes expressions régulières pouvant être associées à un nœud sont construites à partir de la SC du nœud et de l'ensemble E constitué par le vocabulaire et les étiquettes morpho-syntaxiques. La construction consiste à remplacer successivement dans la SC, chaque symbole $+$ par les quatre expressions suivantes i, $+\ i$,$i\ +$, $+\ i+$ dans lesquelles i décrit l'ensemble E. Si un nœud est associé, par exemple à la SC <+président+>, lorsque i vaut *de*, les huit expressions régulières suivantes sont générées :

$< de\ president+ >$	$< +de\ president+ >$	$< de\ +president+ >$	$< +de + president+ >$
$< +president\ de >$	$< +president + de >$	$< +president\ de+ >$	$< +president + de+ >$

Chaque expression régulière vue comme une question sépare la partie du corpus d'exemples associé à un nœud en deux : les GN qui appartiennent au langage généré par l'expression régulière et ceux qui n'appartiennent pas.

3.3. Choix de la question la plus discriminante

Parmi les différentes questions générées à l'issue du processus décrit ci-dessus, une question est choisie en fonction du critère d'impureté de Gini (Breiman *et al.*, 1984), qui est une mesure de l'homogénéité d'un ensemble. La question choisie est celle qui provoque la baisse maximale d'*impureté* entre un nœud et ses deux descendants directs. Etant donné un nœud N et ses deux descendants directs, appelés N_{oui} et N_{non}, la baisse d'impureté ΔI est définie de la façon suivante[3] :

$$\Delta I = I(N) - \frac{|N_{oui}|}{|N|} I(N_{oui}) - \frac{|N_{non}|}{|N|} I(N_{non})$$

l'impureté d'un nœud N est calculée selon la formule suivante :

$$I(N) = \sum_{j \neq k} p(j|N) p(k|N)$$

3. $|N|$ dénote le nombre d'exemples associés au nœud N.

où j et k décrivent l'ensemble des étiquettes. La probabilité qu'un GN étiqueté j appartienne au nœud N, $p(j|N)$, est estimée par la fréquence relative des GN étiquetés j dans le nœud N. Lorsque tous les éléments du nœud possèdent la même étiquette, l'impureté est minimale, elle vaut zéro. La question maximisant ΔI sera associée au nœud N. L'expression régulière lui correspondant constituera la SC des descendants directs de N.

3.4. Condition d'arrêt

Lorsqu'un nœud de l'arbre ne contient plus qu'un GN ou que son impureté est inférieure à un seuil, il n'est plus scindé en deux.

A l'issue de ce processus, un arbre de décision a été créé. Chaque nœud est associé à une expression régulière et chaque feuille comporte un certain nombre de GN étiquetés. Ces GN vont permettre de calculer une distribution de probabilité sur l'ensemble des étiquettes sémantiques des noms propres. Si, par exemple, une feuille comporte 100 exemples, dont 90 sont étiquetés VILLE et 10 sont étiquetés ORG, la classe VILLE se verra attribuer la probabilité $0{,}9$ et la classe ORG, la probabilité $0{,}1$. Plus la distribution sera uniforme, moins l'expression régulière associée à la feuille sera représentative d'une classe sémantique.

L'allure de la distribution d'une feuille permet de classer ces dernières selon leur capacité à modéliser un contexte correspondant à une classe sémantique particulière. Les feuilles possédant cette propriété sont appelées des feuilles *discriminantes*. Une feuille est considérée discriminante lorsque la probabilité de son candidat dépasse un certain seuil. La probabilité du candidat est appelée la *discriminance* de la feuille et le seuil, *seuil de discriminance minimum*. Lorsque le seuil de discriminance minimum est fixé à zéro, toutes les feuilles sont considérées discriminantes.

4. Utilisation de l'arbre de décision

A l'issue de la construction d'un ACS, chaque nœud de ce dernier est associé à une question. De plus, chaque feuille possède une distribution de probabilités. Cet arbre va permettre d'étiqueter des noms propres apparaissant dans des GN. Le principe consiste à présenter à la racine de l'arbre un GN contenant un nom propre inconnu et de parcourir l'arbre en fonction des réponses aux questions associées aux nœuds, jusqu'à atteindre une feuille. Il est alors possible d'étiqueter le nom propre par le candidat de la feuille si toutefois la discriminance de celle-ci est supérieure au seuil de discriminance minimum. Il s'agit là de la façon la plus naturelle d'utiliser un ACS pour effectuer de l'étiquetage sémantique.

Il est aussi possible d'utiliser l'étiquette attribuée par l'ACS à un nom propre du corpus de test pour mettre à jour l'entrée lexicale correspondant à ce dernier, toujours si la discriminance de la feuille dépasse un certain seuil. Le lexique ainsi mis à jour est utilisé pour estimer les paramètres d'un étiqueteur stochastique. Deux séries d'expériences explorant ces deux modes d'utilisation des ACS sont décrites en 4.1 et en 4.2.

Le corpus d'apprentissage utilisé est constitué de textes issus du journal *Le Monde* entre les années 1987 et 1991, ce qui constitue un ensemble d'environ 98M de mots. De ce denier sont extraits $107K$ GN différents sur lesquels l'arbre de décision est construit. L'arbre comprend $10{,}5K$ nœuds internes et $10{,}5K$ feuilles. Chaque feuille correspond à une expression régulière composée de graphies, de catégories syntaxiques et de symboles $+$. A titre d'exemples, voici quelques unes de ces expressions régulières :

```
+ président +  + administration  de  DET  XXXX* +  => XSOC   =1.00
+ président  PREP  gouvernement  de  DET  XXXX* +  => XPAY   =1.00
+ président  PREP  directoire  de  DET  XXXX* +    => XSOC   =1.00
+ le +  + président  PREP +  + de  DET  XXXX* +    => XSOC   =1.00
```

Ces expressions régulières correspondent à des feuilles de discriminance maximale : la probabilité de leur candidat est en effet égale à 1.

4.1. Etiquetage avec l'ACS

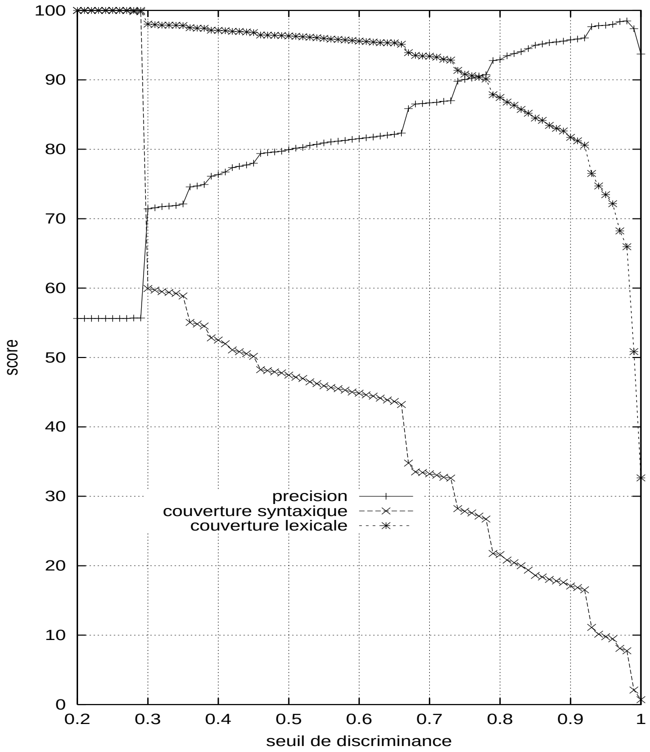
Afin d'évaluer les capacités de l'ACS à correctement étiqueter des noms propres, deux corpus de test, C_0 et C_1, composés de GN contenant des noms propres ont été constitués. Les expériences ont consisté à attribuer au noms propres contenus dans les GN composant ces corpus une étiquette en effectuant un parcours de l'ACS, depuis la racine, jusqu'à atteindre une feuille. Si la discriminance de cette dernière est supérieure au seuil de discriminance minimum, le nom propre est étiqueté par le candidat de la feuille. Sinon il n'est pas étiqueté. Les deux corpus de test sont décrits ci-dessous :

- C_0 est un corpus de 1,2M de GN extraits d'articles du journal *Le Monde* des années 91-92. Il est constitué de l'ensemble des GN possédant un nom propre connu de notre dictionnaire référence de 265K mots. Cette évaluation massive s'effectue donc sur des données très proches de celles utilisées pour la construction de l'arbre (sans chevauchement toutefois avec le corpus d'apprentissage).
- C_1 est un corpus de 695 GN contenant 282 nom propres différents. Il s'agit des noms propres, qui apparaissent au plus 4 fois dans le corpus du *Monde* 91-92. Ce corpus nous permet d'évaluer les performances de notre méthode sur des noms propres ayant une fréquence d'occurrence très faible. Nous vérifions ainsi la capacité de *généralisation* de l'ACS.

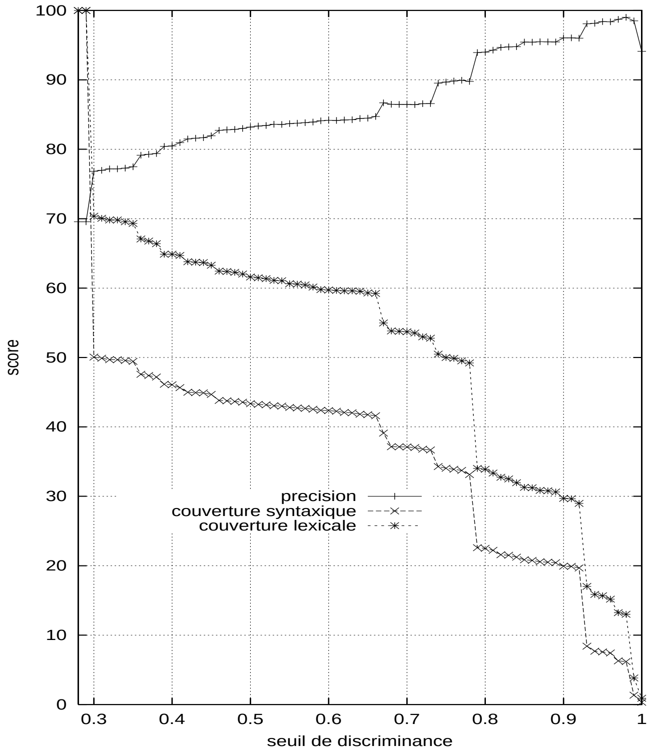
Les résultats de ces expériences sont présentés dans le tableau 1 qui indique pour différentes valeurs du seuil de discriminance minimum les scores de précision et de couverture. La précision est le rapport du nombre de noms propres correctement étiquetés par l'ACS sur le nombre de noms propres étiquetés. La couverture syntaxique est la proportion des occurrences de GN du corpus de test qui sont considérées comme discriminantes par l'ACS (qui sont reconnus par les expressions régulières de feuilles discriminantes). La couverture lexicale indique la proportion de noms propres ayant apparu au moins une fois dans un contexte discriminant.

Les résultats obtenus montrent qu'une bonne précision de l'étiquetage ne peut être obtenue qu'en imposant un seuil de discriminance élevé, correspondant à une couverture syntaxique faible.

Il en résulte que l'ACS ne peut être utilisé pour étiqueter directement les noms propres d'un corpus avec une bonne précision, car dans la plupart des cas, les noms propres n'apparaissent pas dans des contextes suffisamment discriminants. Cette faible proportion de contextes discriminants s'explique par le fait que d'une part certains contextes sont réellement ambigus (par exemple *le président de X*) et que d'autre part, le choix que nous avons fait de limiter le contexte d'un nom propre au GN dans lequel il apparaît ne permet pas de prendre en compte des collocations entre éléments n'appartenant pas au même GN. Il serait intéressant pour cela de recourir à



Corpus C_0



Corpus C_1

TAB. 1 – *Résultat de l'étiquetage sur les corpus C_0 et C_1*

une analyse syntaxique plus étendue ou plus fine, à l'image de (Collins & Singer, 1999), ce qui aurait par contre pour effet d'accroître les erreurs d'analyse.

D'autre part, pour des niveaux élevés de discriminance, la précision de l'étiquetage est très bonne, pour un seuil de discriminance de $0{,}9$, la précision atteint 95%. De plus, à ce niveau de discriminance minimale, le rappel lexical reste élevé, plus de 80%. C'est la raison pour laquelle l'ACS va être utilisé non pour étiqueter directement un corpus mais d'une façon *indirecte*, pour mettre à jour un lexique, comme le décrit la section 4.2.

Il est intéressant de constater que les performances de l'étiquetage sont similaires sur les deux corpus C_0 et C_1 (à l'exception de la courbe de couverture lexicale qui est proche de la couverture syntaxique pour C_1, du fait du faible nombre d'occurrences des noms propre dans ce corpus). Cela permet de vérifier la capacité de l'arbre à traiter correctement des noms propres peu fréquents. Les expressions régulières construites lors de l'apprentissage semblent bien caractériser une *classe* de noms propres, et non pas uniquement les contextes des noms propres les plus fréquents.

4.2. Réestimation des paramètres d'un étiqueteur stochastique

Les expériences décrites ci-dessus ont montré qu'un bon niveau de précision de l'étiquetage par un ACS ne pouvait être atteint qu'au prix d'un faible niveau de couverture, ce qui ne permettait pas d'envisager l'utilisation d'un ACS comme méthode d'étiquetage d'un corpus. Il est néanmoins possible de recourir un ACS pour étiqueter d'une façon indirecte un corpus. L'idée consiste à utiliser l'ACS pour attribuer des étiquettes à un nom propre donné, lorsque ce dernier apparaît dans un contexte discriminant, comme dans les expériences précédentes. Par contre, cet étiquetage est utilisé pour mettre à jour un lexique qui lui même sera utilisé par un étiqueteur stochastique du type n-gram. L'utilisation est dite indirecte dans la mesure où l'étiquetage du corpus n'est pas effectuée par l'ACS, mais par un étiqueteur stochastique dont une partie des paramètres a été, elle, estimée par un ACS.

Les expériences effectuées ont porté sur le corpus de test C_1. Dans un premier temps, les noms propres sont étiquetés en fonction de leur contexte d'occurrence, comme dans les expériences de la section précédente, à la différence que, lorsqu'à l'issue de la traversée de l'arbre un GN contenant le nom propre m aboutit dans une feuille discriminante F, la distribution de probabilité de cette dernière est utilisée pour mettre à jour l'entrée lexicale de m. Si F attribue à l'étiquette FAMILLE la probabilité $0{,}6$ et à l'étiquette ORG la probabilité $0{,}4$, par exemple, le couple $<$FAMILLE,$n>$ de l'entrée lexicale m est incrémentée de 6 ($<$FAMILLE,$n+6>$) et le couple $<$ORG,$n>$ de 4. Le lexique ainsi mis à jour est alors utilisé pour réestimer les probabilités conditionnelles $P(m|c)$ dans le modèle de l'étiqueteur stochastique (voir (Charniak *et al.*, 1993) pour plus de détails). Le nouveau modèle est alors utilisé pour étiqueter le corpus C_1. Le résultat de cet étiquetage est appelé T_{ACS}, c'est sur ce dernier que le taux d'étiquetage correct est calculé.

Afin de disposer d'un point de comparaison, nous avons eu recours à une technique classique de prise en compte des mots inconnus dans un étiqueteur stochastique, décrite dans (Weischedel *et al.*, 1993). Cette technique consiste à considérer que les mots inconnus ont la même probabilité d'appartenir à n'importe quelle classe. Dans notre cas, cette hypothèse consiste à considérer que les 282 noms propres différents de C_1 ont la même probabilité d'appartenir aux 5 différentes classes sémantiques. Le résultat de l'étiquetage de C_1 selon cette technique constitue T_{base}.

Les résultats de ces expériences sont reportés dans le tableau 2 qui indique le taux d'étiquetage correct des 695 occurrences de noms propres de C_1 pour plusieurs valeurs du seuil de discriminance minimal S_d. La seconde ligne indique le gain de performance par rapport à T_{base}.

S_d	0.0	0.2	0.4	0.6	0.8	1
T_{ACS}	71.3	71.3	73.0	72.31	70.8	67.5
$\%gain$	5.94	5.94	8.47	7.44	5.2	0.3

TAB. 2 – *Taux d'étiquetage correct en fonction du seuil de discriminance*

La baisse de performance se produisant pour des valeurs de S_d supérieure à $0{,}5$ s'explique par le fait que pour ces valeurs de S_d peu des 282 noms propres différents ont été vus dans des contextes discriminants, leur entrée lexicale n'a, par conséquent, pas été mise à jour. Par contre les expériences ont montré que plus la valeur de S_d est élevée, meilleures sont les performances rapportées aux seuls noms propres dont les entrées lexicales ont été mises à jour. Pour une valeur de S_d de $0{,}8$, moins de 30% des noms propres sont apparus dans des contextes discriminants mais le taux d'étiquetage correct sur ces derniers dépasse 95%. Il est par conséquent intéressant d'augmenter le nombre de contextes dans lequel apparaît un nom propre afin d'augmenter la probabilité que ce dernier apparaîsse dans des contextes discriminants. Cette voie sera explorée dans la partie 5.

Il est important de remarquer que le processus de mise à jour du lexique présenté ci-dessus peut favoriser certaines catégories. En effet, si certaines catégories apparaissent plus souvent que d'autre comme candidat des feuilles de l'ACS, il leur correspondra plus de contextes discriminants et par conséquent la probabilité qu'un tel contexte apparaisse dans le corpus de test augmentera. Les performance d'étiquetage sur ces catégories s'en ressentira probablement. Nous n'avons pas à l'heure actuelle étudié plus précisément ce phénomène.

5. Acquisition de nouveaux exemples sur le Web

Il a été montré dans le paragraphe précédent que la présence d'un mot inconnu dans au moins un contexte discriminant permet d'augmenter très fortement la qualité de son étiquetage (95% d'étiquetage correct pour un seuil à $0{,}8$). Lorsqu'un nom propre est peu représenté dans le corpus de test, il est tentant d'acquérir plus d'occurrences de ce dernier en glanant sur le Web de nouveaux exemples. Ce processus a l'avantage de pouvoir être totalement automatisé et de n'engendrer aucun coût supplémentaire pour l'acquisition de corpus. Néanmoins, étant donné la masse et l'hétérogénéité des données disponibles sur le Web, un filtrage s'avère indispensable pour éliminer les données non pertinentes (liste, textes en langues étrangères, tableaux, etc.). Cette acquisition et ce filtrage se font de la façon suivante :

Tout d'abord, pour chaque nom propre inconnu à traiter, une requête est effectuée à l'aide d'un moteur de recherche en spécifiant la langue désirée et un mot-clef : le nom propre. Chaque page correspondant aux résultats de la requête est alors téléchargée puis nettoyée. Ce nettoyage consiste à éliminer des données rapatriées les parties non textuelles. Le texte obtenu est traité par une chaîne de nettoyage de corpus qui comprend un module de traitement des accents, un segmenteur en mots, un segmenteur en phrase et enfin un étiqueteur statistique utilisant un jeu de 105 classes morpho-syntaxiques. Les noms propres hors vocabulaire se verront attribuer l'étiquette INC. Les GN du texte étiqueté sont détectés grâce à l'analyseur syntaxique de surface en utilisant la même grammaire que celle employée en 3, dans laquelle les 5 catégories correspondant aux noms propres ont été remplacées par l'étiquette INC. Enfin, les GN obtenus sont analysés par l'arbre de décision et seuls ceux tombant dans des feuilles ayant un seuil de discriminance supérieur à un seuil fixé sont conservés.

Cette procédure a été mise en œuvre sur les 282 noms propres du corpus C_1. Les requêtes effectuées nous ont permis de rapatrier, en moyenne, un corpus de 3Mo de texte HTML pour chaque nom propre. A l'issue du processus de nettoyage, seulement 110Ko de texte ont été conservé pour chaque mot. Les 5000 GN extraits de ces corpus ont été ajoutés aux 695 GN du corpus C_1 pour constituer le corpus C_2. Il faut noter que C_2 ne contient que 247 des 282 entrées soit $87{,}6\%$ des formes. Enfin, le nombre moyen d'occurrences de chaque forme dans le corpus obtenu sur le Web est de $14{,}5$.

Après avoir réestimé les paramètres de notre étiqueteur stochastique par la méthode décrite dans le paragraphe 4.2 grâce au corpus C_2, nous avons étiqueté le corpus de phrases de C_1. Le résultat de cet étiquetage est appelé T_{web}. Une comparaison entre le taux d'étiquetage correct entre T_{base} et T_{web} est donnée dans le tableau 3.

S_d	0.0	0.2	0.4	0.6	0.8	1
T_{web}	72.6	72.6	73.5	73.7	73.9	70.9
%gain	7.8	7.8	9.2	9.5	9.8	5.3

TAB. 3 – *Taux d'étiquetage correct en utilisant le corpus du Web*

Même si ces résultats montrent une légère amélioration par rapport aux performances sur T_{ACS}, le gain reste marginal si l'on considère l'ensemble des 282 entrées de C_1. En ne considérant que les noms propres ayant été vu au moins une fois dans un contexte correspondant au seuil de discriminance choisi, l'apport des exemples du Web devient significatif : un taux d'étiquetage de 90% est obtenu sur 50% des entrées grâce à l'introduction des GN issus du Web alors qu'il n'était que de 40% en utilisant uniquement les GN de C_1.

Une étude manuelle des données récoltées nous donne quelques explications sur ces faibles performances : d'une part, même après la phase de nettoyage de corpus, les données récoltées restent très bruitées (mauvaise ponctuation, tags HTML incomplets, erreurs de segmentation en mots et en phrases, . . .) ; d'autre part, le style et le domaine sémantique des corpus rapatriés diffèrent très souvent de ceux présent dans les textes ayant servis à la construction de l'arbre (articles du journal *Le Monde*). En effet, les expressions régulières apprises sur du corpus journalistiques ne sont pas à même de traiter correctement des extraits de texte littéraire, du texte issus de forum de discussion, ou encore des pages personnelles écrites dans un style relâché.

6. Conclusion

Les résultats des expériences menées dans cette étude ont montré que l'acquisition de corpus sur le Web pouvait améliorer les performances d'un étiqueteur stochastique. Ils ont aussi montré la grande variété des corpus issus du Web et les difficultés découlant de cette variété pour la tâche d'enrichissement de lexiques de noms propres. Ces conclusions montrent la voie à des évolutions des techniques présentées dans cet article. Ces évolutions portent en particulier sur deux points :

D'une part la nécessité de mieux contrôler les capacités de généralisation des arbres de classification sémantiques. La voie que nous comptons explorer découle de l'observation que les éléments qu'intègrent les expressions régulières produites lors de la construction de l'arbre sont soit trop précis (des mots) soit trop vastes (des classes syntaxiques). Des travaux futurs porteront sur la possibilité d'intégrer dans les expressions régulières des classes de mots constituées dynamiquement lors de la construction des ACS.

D'autre part le traitement de données issues du Web nous amène à être confronté à des corpus appartenant à des domaines sémantiques très vastes. Les arbres appris sur des seules données journalistiques ne sont pas toujours pertinents pour traiter l'ensemble des corpus collectés. Une phase de filtrage spécifique à ce type de données semblent donc nécessaire afin de faire correspondre les données d'apprentissage et d'utilisation des arbres.

Références

BÉCHET F. & YVON F. (2000). Les noms propres en traitement automatique de la parole. *to appear in Traitement Automatique des Langues*.

BREIMAN L., FRIEDMAN J., OHLSEN R. & STONE C. (1984). *Classification and Regression Trees*. Wadsworth.

CHARNIAK E., HENDRICKSON C., JACOBSON N. & PERKOWITZ M. (1993). Equations for part-of-speech tagging. In *11th National Conference on Artificial Intelligence*, p. 784–789.

COLLINS M. & SINGER Y. (1999). Unsupervised models for named entity classification. In *Empirical Methods in NLP processing and Very Large Corpora - EMNLP-VLC'99*, University of Maryland.

KUHN R. & DE MORI R. (1996). The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(5), 449–460.

SPRIET T. & EL-BÈZE M. (1995). Etiquetage probabiliste et contraintes syntaxiques. In *TALN*.

WEISCHEDEL R., SCHWARTZ R., PALMUCCI J., METEER M. & RAMSHAW L. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, **19**(2), 359–382.