MODÈLES PROBABILISTES ET ÉTIQUETAGE AUTOMATIQUE

Bernard MERIALDO*

Résumé - Abstract

L'étude et le traitement du Langage Naturel s'intéressent de plus en plus aux approches prenant en compte de grands volumes de textes d'apprentissage, pour y collecter des statistiques qui permettent de construire des modèles, par exemple basés sur des considérations probabilistes. Dans cette présentation, nous décrivons les fondements de l'approche probabiliste du traitement du langage naturel, en discutant ses avantages et inconvénients. Puis nous nous intéressons au problème de l'étiquetage grammatical automatique, pour lequel nous détaillons l'utilisation d'un modèle appelé triclasse. Nous terminons en commentant plusieurs expériences d'étiquetage.

The usage of large corpora is getting an increasing interest for various applications of Natural Language Processing. Statistics collected of large amounts of text can be used to derive complex models. In this paper, we describe the basics of the probabilistic approach to NLP, with a discussion of its advantages and drawbacks. Then we present its applications to the particular problem of text tagging, using a so-called "triclass" model. Finally, we present some tagging experiments.

Mots clefs - Keywords

modèle de langage probabiliste, étiquetage grammatical, corpus de texte.

probabilistic language model, text tagging, text corpus.

---

* Institut EURECOM, BP 193, 06904 Sophia-Antipolis,
email: merialdo@eurecom.fr

1. L'APPROCHE PROBABILISTE

L'utilisation de méthodes probabilistes pour le traitement de la langue naturelle a été proposée très tôt. Toutefois, les performances des ordinateurs n'ont permis une utilisation vraiment efficace de ces méthodes que depuis les années 80. En particulier, les méthodes probabilistes se sont révélées particulièrement efficaces pour résoudre les problèmes apparaissant aux différents niveaux de la reconnaissance de parole (aussi bien au niveau phonétique qu'aux niveaux syntaxique et sémantique) (Bahl et al., 1983, Derouault et Merialdo, 1986). Ce succès a entraîné le développement de ces méthodes vers de nombreuses autres applications (Charniak, 1994), telles l'étiquetage grammatical, le rattachement prépositionnel (Debili, 1994), la classification de mots (Brown et al., 1992), l'accentuation automatique (El-Beze et al., 1994), l'alignement de corpus bilingues, pour aller jusqu'à des applications aussi complexes que la traduction automatique (Brown et al., 1990).

La méthodologie pour mettre en oeuvre des méthodes probabilistes est simple:

1. on identifie un problème à résoudre, par exemple pour l'étiquetage grammatical il s'agira de choisir la bonne classe d'un mot dans un contexte donné parmi les classes possibles.
2. on modélise le problème en faisant apparaître les probabilités de certains événements, par exemple la probabilité qu'un mot appartienne à une classe donnée dans un certain contexte. En général les probabilités complexes seront calculées par composition de probabilités élémentaires, qui constituent les paramètres du modèle. La formule de calcul est obtenue en faisant des hypothèses simplificatrices: par exemple on suppose que l'apparition d'un mot ne dépend que des deux mots précédents. Le bon choix de ces hypothèses est un facteur primordial pour la qualité du modèle obtenu, et requiert une très bonne intuition du phénomène réel..
3. à partir de données d'apprentissage, corpus de textes éventuellement "décorés" par des informations lexicales ou syntaxiques, on construit des estimations des valeurs des probabilités élémentaires précédemment définies. Plusieurs techniques permettent de réaliser cet apprentissage, depuis la simple collecte de fréquences relatives, jusqu'au processus de réestimation des modèles cachés* (connu sous le nom d'algorithme de Baum-Welch) (Baum, 1972, Rabiner and Juang, 1986).
4. le modèle une fois appris, on peut utiliser les probabilités pour traiter de nouvelles données, et donc réaliser la fonction de traitement souhaitée.

Il est bien connu qu'une grande partie des difficultés rencontrées dans le

---

* dans un modèle caché, l'état du modèle n'est pas déterminé de façon unique par l'observation. Par exemple, si l'état dépend de la catégorie grammaticale du mot, l'observation d'un mot ne permet pas toujours de savoir l'état exact du modèle.

traitement de la langue naturelle concernent la résolution des ambiguïtés. Le grand intérêt des méthodes probabilistes est d'en fournir une solution simple et immédiate. Chaque solution potentielle se voit alors associée une probabilité que l'on peut interpréter comme une fréquence d'apparition de cette solution. Résoudre l'ambiguïté revient alors simplement à choisir l'hypothèse de plus forte probabilité (ce qui revient dans l'interprétation probabiliste à suivre le principe raisonnable de vouloir minimiser le risque d'erreur).

Un autre intérêt des méthodes probabilistes tient à leur capacité d'apprentissage. Les méthodes automatiques d'apprentissage permettent de construire des modèles complexes (comportant de très nombreux paramètres), chose qu'il est difficile d'envisager de faire manuellement. La qualité des modèles est souvent liée à la quantité de données utilisées dans l'apprentissage. Nous sommes heureusement à une époque où le développement de l'informatique et des télécommunications font que les données linguistiques deviennent disponibles sous forme électronique en quantité de plus en plus importante. Cette disponibilité de vastes corpus de données, ainsi que l'amélioration constante des performances des ordinateurs en terme de stockage et de calcul entraîne une amélioration quasi-automatique des modèles probabilistes, soit parce qu'on peut mieux les apprendre avec plus de données, soit parce qu'on peut construire des modèles plus complexes (utilisant plus de paramètres). Le progrès technologique devient donc un moteur de l'amélioration de la qualité des modèles et de la performance des applications qui les utilisent.

Enfin, on reproche souvent aux méthodes probabilistes de ne comporter qu'une vision mathématique des phénomènes étudiés, et de réduire la résolution d'un problème à la programmation des algorithmes d'apprentissage. Cela n'est que partiellement vrai car, à côté de ce travail de programmation, il faut noter l'importance du travail d'analyse lors de la définition du modèle à utiliser pour résoudre un problème. Lors de cette analyse sont faites des hypothèses simplificatrices dont la pertinence influe directement sur la qualité du modèle. Des hypothèses grossières ne permettront pas au modèle de distinguer la structure fine de certaines situations, des hypothèses inadaptées conduiront à des paramètres qui ne permettent pas de distinguer certains cas d'autres. La conception du modèle est donc une activité extrêmement importante, et, si les modèles les plus utilisés actuellement sont encore parfois rudimentaires dans leur conception (comme de considérer les seuls mots voisins comme caractéristiques d'un contexte), on admet généralement qu'une meilleure conception, avec en particulier l'introduction de considérations linguistiques fortes dans le choix des hypothèses, pourra seule permettre un progrès significatif de la qualité des modèles par rapport aux modèles actuels.

La mise en oeuvre des méthodes probabilistes se heurte à certaines difficultés. Nous en citons trois qui nous paraissent les plus importantes.

- La première est que la complexité de la programmation limite parfois la complexité des modèles que l'on peut raisonnablement employer. Par exemple, si l'on veut faire des statistiques sur des suites de 4 mots, il n'est pas possible d'utiliser un tableau à quatre dimensions, pour des raisons de taille de mémoire. Il faut alors construire des structures de données plus économiques, qui profitent du fait qu'il existe (en pourcentage) peu de quadruplets possibles, mais qui sont plus délicates à programmer. Cet argument diminue toutefois au fur et à mesure que les performances des ordinateurs augmentent, et qu'il devient plus aisé de disposer de grandes tailles de mémoire, d'espace disque etc...

- La seconde est que l'estimation des probabilités d'un modèle demande des données qui n'existent pas toujours. Par exemple, pour un modèle servant à l'étiquetage automatique, il est pratique de disposer de textes annotés où la classe d'un mot dans son contexte a été déterminée, afin de collecter les statistiques conduisant à l'estimation du modèle. Malheureusement, la constitution de grands volumes de tels textes est une opération nécessitant une intervention humaine (pour une qualité optimale), ce qui contraint le volume de ce type de production. Devant l'importance pratique de tels corpus, un certain nombre d'initiatives (Black et al., 1993, Marcus et al., 1993) ont été prises pour produire l'effort nécessaire à la mise en place de tels types de données, soit sous forme de textes étiquetés où chaque mot se voit associer la bonne classe, soit sous forme de textes contenant l'analyse grammaticale de chaque phrase. Même avec de telles bases, il n'est pas toujours possible de disposer des données correspondant au modèle que l'on souhaiterait expérimenter. Par exemple, lors de travaux sur la modélisation des rattachements prépositionnels (Debili, 1994), nous aurions pu utiliser avec profit des groupes nominaux analysés qui auraient permis de simples comptages; ne disposant pas de ces données, nous eûmes recours à d'autres techniques probabilistes pour estimer notre modèle. En effet, lorsqu'il n'est pas possible de disposer de données où les paramètres sont directement observables, les modèles probabilistes fournissent toutefois une alternative possible par l'utilisation de l'algorithme de Baum-Welch qui permet d'estimer la valeur de paramètres cachés. Cet algorithme reste toutefois délicat à comprendre et implémenter, et les utilisations en restent peu fréquentes.

- La troisième est d'ordre technique, liée à la qualité de l'estimation à partir d'un certain ensemble de données. Il arrive très souvent que les modèles que l'on utilise possèdent un grand nombre de paramètres (si l'on considère des suites de trois catégories grammaticales parmi un jeu de 100 catégories, il y a un million de triplets envisageables, dont certainement plusieurs milliers peuvent se réaliser dans des textes). Dans ces conditions, même si le corpus d'apprentissage est important, contenant des millions de mots, l'estimation des faibles probabilités est très imparfaite, c'est-à-dire que les valeurs relevées sur l'ensemble d'apprentissage ne correspondent qu'imparfaitement aux valeurs idéales. Si une probabilité est de la forme N/M, et qu'on observe que M vaut 1, que peut-on en conclure? Si N vaut 1,

est-ce pour autant que l'événement est sûr? et si N vaut 0, est-il vraiment impossible? Après tout, si l'on utilisait juste quelques données supplémentaires, on observerait peut-être des situations qui modifieraient la valeur de N de façon significative. Ce problème d'estimation des faibles probabilités, ou d'estimation en présence de données insuffisantes, est une cause majeure de certaines difficultés d'utilisation des modèles probabilistes.

2. LES MODÈLES DE MARKOV

Les modèles de Markov constituent un des types de modèles probabilistes les plus utilisés, en raison de leur simplicité et de leur efficacité. Au travers d'un exemple simple, nous allons montrer les principes de base de leur fonctionnement. Considérons un exemple météorologique, dans lequel on observe le temps qu'il fait chaque jour, selon les trois possibilités:

beau, couvert et pluie.

.

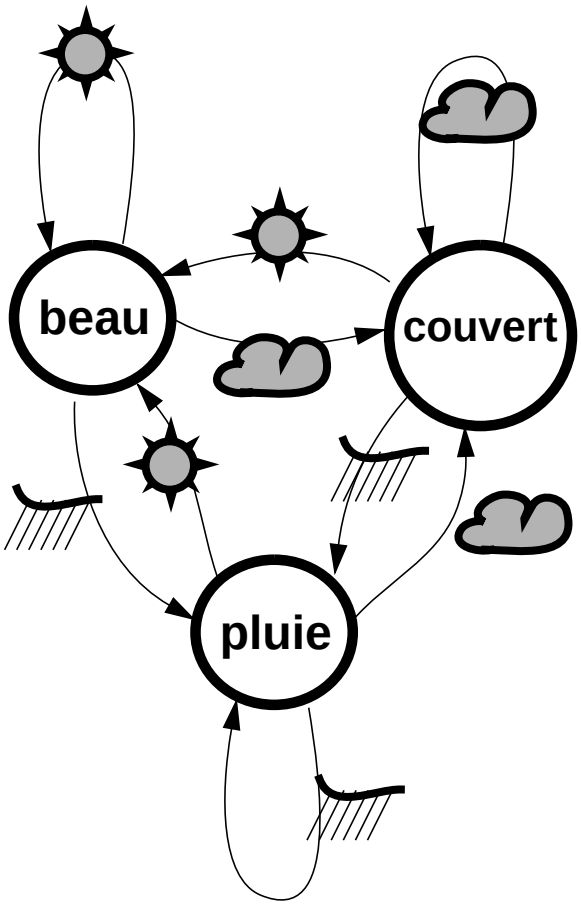

Figure 1: un modèle de Markov de l'évolution possible du temps

On peut construire un modèle de Markov à trois états, chaque état correspondant à une situation donnée, où une nouvelle observation provoque une transition qui fait passer dans l'état correspondant. Chaque transition est affectée d'une probabilité. Une suite de transitions correspond à l'observation d'une suite donnée de symboles.

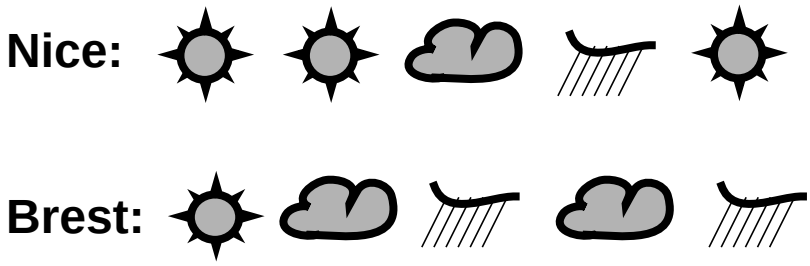

Figure 2: des exemples d'observation en des lieux connus

La phase d'apprentissage consiste à estimer les valeurs des probabilités des différentes transitions. Pour cela, on observe l'évolution de la météo dans un lieu donné. Cela permet de suivre l'état du modèle de Markov, et de déterminer les transitions utilisées. On peut donc calculer la fréquence relative de chaque transition, ce qui fournit une estimation de la probabilité de cette transition. Bien sûr, selon que l'observation est faite à Nice ou à Brest, les probabilités de transitions seront très différentes, même si la structure (états et transitions possibles) du modèle reste la même. Dans le modèle de la Figure 1, en supposant que l'on part de l'état initial beau, l'observation faite à Nice correspond au chemin:

(beau initial) : beau - beau - couvert - pluie - beau

et celle faite à Brest au chemin:

(beau initial) : beau - couvert - pluie - couvert - pluie

A partir de l'état beau, on aura donc observé, à Nice 2 fois du beau temps et une fois un temps couvert, et à Brest une fois du beau temps et une fois un temps couvert. Les estimations des probabilités de transition seront donc:

à Nice, p(beau/beau) = 2/3, p(couvert/beau) = 1/3...

à Brest, p(beau/beau) = 1/2, p(couvert/beau) = 1/2...

Dans la phase de reconnaissance, on peut à partir d'une observation donnée (suite de symboles beau, pluie et couvert), se demander en quel lieu elle a été observée. Il suffit pour cela de calculer la probabilité que cette observation ait été produite par chacun des modèles de Nice et de Brest. Cette probabilité se calcule comme le produit des transitions empruntées dans le modèle pour produire l'observation. En comparant les probabilités fournies par chacun des deux modèles, on peut décider quel est le lieu d'observation le plus probable.

Figure 3: un exemple d'observation en un lieu inconnu

Par exemple, la suite observée dans la Figure 3 a une probabilité d'être émise (à partir de l'état initial beau) qui se calcule comme:

p(beau/beau) x p(couvert/beau) x p(pluie/couvert),

ce qui, en fonction de l'endroit, donne:

à Nice, 2/3 x 1/3 x 1/1, soit 2/9

à Brest, 1/2 x 1/2 x 1/1, soit 1/4

Il est donc plus probable que cette observation ait été faite à Brest plutôt qu'à Nice. Si un choix doit être fait, le modèle probabiliste indique Brest comme meilleure réponse.

Dans cet exemple simple, l'observation permet de savoir directement quelles transitions sont empruntées, et donc de calculer directement les fréquences relatives. Un des grands intérêts des modèles de Markov est l'existence d'un algorithme d'apprentissage pour les modèles cachés, c'est-à-dire où les transitions ne sont pas directement observables. Supposons que l'observation consiste dans le relevé de température, valeur qui est corrélée avec l'ensoleillement, sans pour autant en donner directement la valeur. En précisant quelques contraintes pour rompre la symétrie du problème, (par exemple si la température dépasse 20 degrés, il fait beau, si elle est en-dessous de 5, le temps est couvert), l'algorithme de Baum-Welch permet de calculer des valeurs des probabilités de transition qui maximisent un principe de vraisemblance. Il devient donc possible d'estimer les valeurs des probabilités de transition, même lorsqu'on ne sait pas exactement par quels états et transitions on est passé pour produire une observation donnée. C'est ce principe qui peut être utilisé, par exemple, pour estimer les probabilités d'apparition de suites de classes grammaticales à partir de textes non étiquetés.

3. L'ÉTIQUETAGE GRAMMATICAL

L'étiquetage grammatical, c'est-à-dire le choix pour chaque mot d'un texte de la catégorie grammaticale correcte dans le contexte où il apparaît, est un problème important qui a fait l'objet de nombreuses études. Deux types d'approches ont été proposées:

- utilisation de règles contextuelles, avec les travaux de (Klein et Simmons, 1963), (Brodda, 1982), (Paulussen et Martin, 1992), (Brill, 1992), (Brill, 1994a),
- utilisation de modèles probabilistes, comme dans (Bahl et Mercer, 1976),

(Beale, 1988), (Church, 1989), (Cutting et al., 1992), (Garside et Leech, 1985), (Kupiec, 1992), (Leech et al., 1983), (Merialdo, 1994), (Schmid, 1994b), (Schutze and Singer, 1994).

Mentionnons également le fait que des réseaux de neurones ont été utilisés dans (Benello et al., 1989), (Nakamura et Shikano, 1989), (Schmid, 1994a).

Dans la plupart des cas, seul le contexte proche est utilisé (mots avoisinants, à une distance de un, deux, voire exceptionnellement plus).

Ces différentes approches ont généralement obtenu de bons résultats, sans que l'on puisse dire si l'une est fondamentalement meilleure que l'autre. Les chiffres avancés sont souvent supérieurs à 95% de réussite (mots dont la classe a été correctement choisie). Rappelons toutefois que les évaluations numériques ne peuvent être vraiment comparées que si elles portent sur des conditions d'expérimentation similaires. En effet, le choix du texte utilisé, le jeu de catégories grammaticales possibles, la qualité du dictionnaire, le volume de données d'apprentissage disponible, sont autant de facteurs importants dans l'évolution du taux de réussite. Notons la pratique occasionnelle d'évaluer les performances sur un texte faisant partie des données d'apprentissage, ce qui donne un bonus certain au système par rapport à une utilisation réelle sur des données nouvelles. On retiendra qu'il ne faut pas s'attacher à la valeur absolue des taux de réussite publiés, mais plutôt aux variations de ce taux lorsqu'on expérimente plusieurs méthodes dans un environnement de données constant (Chanod et Tapanainen, 1995b).

Un système d'étiquetage est composé en général de trois éléments:

1. un dictionnaire, qui permet de connaître pour chaque mot la liste des catégories grammaticales possibles dans le jeu de catégories considéré,
2. un modèle contextuel, soit sous forme de règles, soit sous forme de probabilités, permettant de prendre en compte le contexte pour l'étiquetage,
3. un algorithme qui utilise ce modèle pour réaliser l'étiquetage.

Le dictionnaire est bien sûr un élément capital pour une bonne performance du système, que ce soit en termes de couverture ou de précision. De même, le jeu de catégories grammaticales utilisé a une forte influence sur la qualité de l'étiquetage (Chanod et Tapanainen, 1995a). Il est clair qu'un système réduit de catégories conduira souvent à de meilleurs taux de réussite qu'un système plus détaillé (mais ce n'est pas garanti et il est certainement facile de construire des contre-exemples). Il est clair également que certains traits sont plus difficiles à identifier (si on souhaitait par exemple une catégorie "substantif représentant un objet animé", ou si l'on voulait distinguer l'indicatif et le subjonctif en français) et peuvent rendre un jeu de catégories plus difficile à désambiguiser que d'autres.

4. L'ÉTIQUETAGE PROBABILISTE

L'étiquetage d'un texte consiste à déterminer la classe grammaticale de chaque mot du texte dans le contexte où il est utilisé. Pour une phrase M, on veut donc construire une suite de classes C:

$M = m_1\ m_2\ ...\ m_n$

$C = c_1\ c_2\ ...\ c_n$

Le principe de l'étiquetage probabiliste est de définir un modèle probabiliste permettant de calculer la probabilité d'une suite de mots-classe: p(M,C). Lorsqu'on veut étiqueter un nouveau texte M', on utilise ce modèle pour trouver la suite de classes C' qui maximise la probabilité p(M',C').

On peut calculer cette probabilité de gauche à droite, en faisant le produit de la probabilité du premier mot-classe, puis du second connaissant le premier, etc...:

$p(M,C) = p(m_1,c_1)\ x\ p(m_2,c_1/m_1,c_1)\ x...\ p(m_n,c_n/m_1,c_1,m_2,c_2,...\ m_{n-1},c_{n-1}).$

Définir le modèle revient donc à définir les probabilités élémentaires:

$p(m_i,c_i/m_1,c_1,m_2,c_2,...\ m_{i-1},c_{i-1})$

Un des modèles les plus utilisés est le modèle triclasse, obtenu en faisant les hypothèses suivantes:

1. la probabilité d'un mot-classe ne dépend que des deux mots précédents:

   $p(m_i,c_i/m_1,c_1,m_2,c_2,...\ m_{i-1},c_{i-1}) = p(m_i,c_i/m_{i-2},c_{i-2},m_{i-1},c_{i-1})$

2. la probabilité d'apparition du mot (dans un contexte) ne dépend que de sa classe:

   $p(m_i,c_i/m_{i-2},c_{i-2},m_{i-1},c_{i-1}) = p(m_i/c_i)\ .\ p(c_i/m_{i-2},c_{i-2},m_{i-1},c_{i-1})$

3. la probabilité d'apparition de la classe ne dépend que des classes précédentes:

   $p(c_i/m_{i-2},c_{i-2},m_{i-1},c_{i-1}) = p(c_i/c_{i-2},c_{i-1})$

Les probabilités élémentaires sont alors:

- p(m/c), dont le nombre est le nombre de mots-classes possibles dans le dictionnaire,
- $p(c_3/c_1,c_2)$, dont le nombre est au maximum le cube du nombre de classes.

Si l'on dispose de textes d'apprentissage étiquetés, la façon la plus simple de construire ce modèle est de calculer les fréquences relatives

$f(m/c) = N(m,c) / N(c)$

$f(c_3/c_1,c_2) = N(c_1,c_2,c_3) / N(c_1,c_2)$

où N(m,c) est le nombre de fois où le mot m est rencontré avec la classe c dans le corpus d'apprentissage, $N(c_1,c_2)$ est le nombre d'occurrences de la

succession de classes c_1, c_2, $N(c_1,c_2,c_3)$ le nombre d'occurrences du triplet de classes.

Les fréquences relatives d'ordre 2 donnent une approximation incomplète de la probabilité, car si une suite de triclasses n'a pas été rencontrée pendant l'apprentissage, sa probabilité sera nulle. En pratique, il est donc avantageux de considérer que la probabilité est une combinaison linéaire des fréquences relatives de plusieurs ordres (on appelle ce processus un 'lissage' des probabilités):

$$p(c_3/c_1,c_2) = l_1 \cdot f(c_1) + l_2 \cdot f(c_3/c_2) + l_3 \cdot f(c_3/c_1,c_2)$$

Les coefficients de la combinaison linéaire l_1, l_2 et l_3 peuvent être déterminés en utilisant l'algorithme de Baum-Welch, comme expliqué ci-dessous. Dans ce cas particulier, l'implémentation de l'algorithme est simplifiée:

1. *on calcule les fréquences relatives sur la quasi-totalité du corpus d'apprentissage, en laissant de côté ("held-out") une petite partie des données pour l'optimisation des coefficients. Sur cette partie, on relève alors les comptes $N'(c_1,c_2,c_3)$,*
2. *on donne des valeurs initiales non nulles aux coefficients l_1, l_2, l_3 (en général, 1/3),*
3. *on commence une phrase d'itération en créant trois compteurs k_1, k_2, k_3 que l'on initialise à 0,*
4. *pour chaque triplet c_1,c_2,c_3, on incrémente les compteurs k_1, k_2, k_3 des valeurs respectives:*

- $N'(c_1,c_2,c_3) \cdot l_1 \cdot f(c_1) / (l_1 \cdot f(c_1) + l_2 \cdot f(c_3/c_2) + l_3 \cdot f(c_3/c_1,c_2))$,
- $N'(c_1,c_2,c_3) \cdot l_2 \cdot f(c_3/c_2) / (l_1 \cdot f(c_1) + l_2 \cdot f(c_3/c_2) + l_3 \cdot f(c_3/c_1,c_2))$,
- $N'(c_1,c_2,c_3) \cdot l_3 \cdot f(c_3/c_1,c_2) / (l_1 \cdot f(c_1) + l_2 \cdot f(c_3/c_2) + l_3 \cdot f(c_3/c_1,c_2))$.

5. *on normalise les valeurs de k_1, k_2, k_3 (pour que leur somme fasse 1), et cela fournit les nouvelles estimations des valeurs des coefficients l_1, l_2, l_3,*
6. *on itère cette phase d'itération autant de fois que nécessaire (par exemple en arrêtant lorsque les valeurs des coefficients subissent des modifications inférieures à un seuil).*

Le lissage des fréquences relatives permet d'éviter que trop de probabilités soient nulles (donc considérées comme impossibles) simplement parce que le cas correspondant n'a pas été observé dans le texte d'apprentissage.

La procédure d'étiquetage peut s'effectuer de plusieurs manières. La façon optimale consiste à énumérer pour une phrase donnée toutes les suites de classes possibles (du moins celles autorisées par le dictionnaire), à évaluer leur probabilité, et à garder la meilleure. Cette procédure a l'inconvénient de dépendre de la longueur de la phrase, le nombre de possibilités augmentant

en général de manière exponentielle en fonction de cette longueur. Une implémentation efficace demande alors d'éliminer des hypothèses partielles dont la probabilité est faible.

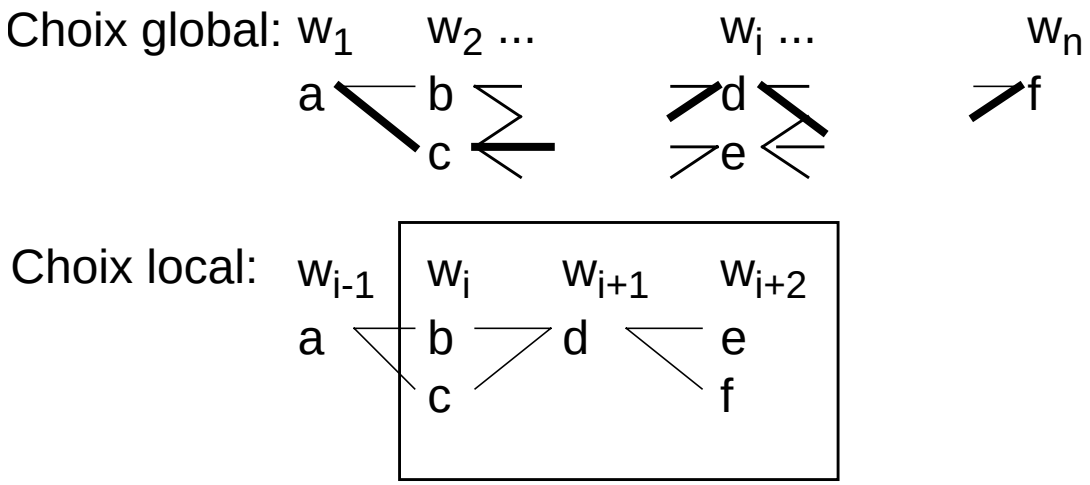

Figure 4: Choix d'une classe, local ou global

Nous avons également proposé (Merialdo, 1994) et utilisé un algorithme qui choisit la classe du premier mot en examinant toutes les possibilités des N mots suivants, la fixe, puis choisit la classe du second mot en examinant les N mots suivants, la fixe, etc... Avec cet algorithme, l'étiquetage peut se faire sans attendre la fin de la phrase, et le volume de calcul ne dépend plus de la longueur de la phrase. Son implémentation est plus simple que la forme optimale, mais utilise plus de calculs. Nous avons trouvé expérimentalement qu'avec N=3, c'est-à-dire en ne regardant que les 3 mots suivants, le résultat de l'étiquetage était identique à celui d'un étiquetage optimal.

Les résultats de l'étiquetage par des modèles probabilistes peuvent être considérés comme "très bons" (tout au moins par leurs auteurs), puisque les taux affichés dépassent souvent 95% de mots correctement étiquetés (Merialdo, 1994). Toutefois, comme nous l'avons déjà mentionné, ces chiffres ne peuvent généralement pas être comparés d'un système à l'autre, car les conditions d'expérience (jeu de classes, quantité et qualité des textes d'apprentissage, dictionnaire...) sont trop différentes. Parmi les quelques expériences comparatives, nous avons montré pour l'anglais (Merialdo, 1994) que les modèles construits à partir de textes étiquetés à la main sont bien meilleurs que ceux construits en utilisant l'algorithme de Baum-Welch, même avec de grands volumes de texte. Dans (Chanod et Tapanainen, 1995b), une comparaison entre l'approche probabiliste et l'approche par règles semble donner un avantage à la seconde, tout au moins sous la contrainte d'un temps de mise au point limité. Toutefois, le degré de satisfaction de tels taux dépend en grande partie de l'utilisation que l'on fera du texte ainsi étiqueté. Lorsqu'il

s’agit de construire des modèles de langage pour la reconnaissance de parole ou la traduction, il semble que les erreurs résiduelles ne constituent pas une source importante de problèmes. Si par contre, l’étiquetage n’est qu’une première phase qui est suivie par une analyse plus approfondie, une erreur d’étiquetage peut entraîner une erreur quasi-systématique de l’analyse. Dans ce cas, un taux d’étiquetage correct de 95% sur les mots peut s’interpréter comme fournissant pratiquement une erreur par phrase (si les phrases ont une longueur moyenne de 20 mots), et donc très problématique pour certaines applications.

5. PROBLÈMES OUVERTS

Nous terminons cet exposé en mentionnant quelques points donnant lieu à des difficultés particulières pour l’étiquetage automatique, et dont la mise en oeuvre est encore un sujet de réflexion.

Quelque soit la qualité et l’étendue d’un dictionnaire, on ne peut espérer qu’il ait une couverture complète des futurs textes à étiqueter. Un système d’étiquetage doit donc intégrer le problème des mots inconnus rencontrés lors de l’utilisation. Plusieurs solutions sont alors possibles. On peut avoir créé une classe grammaticale “mot inconnu” pour contenir tous ces mots, et, en agissant de façon adéquate pendant l’apprentissage, avoir collecté des statistiques sur le comportement de cette classe. Si le dictionnaire est suffisamment complet, les mots nouveaux seront souvent des noms propres, le comportement de la classe “mot inconnu” sera donc assez stable. Une autre solution est de supposer que les mots inconnus ont certaines classes possibles, et de laisser les contraintes de contexte décider laquelle est la meilleure. Finalement, on peut également rajouter des considérations morphologiques pour déterminer une liste de classes possibles pour un mot nouveau rencontré.

Le modèle triclasse considère une phrase comme une suite de mots. Les mots sont en général construits à partir de la phrase par une étape de segmentation déterministe, se basant sur des considérations typographiques et éventuellement sur des dictionnaires d’entrées lexicales complexes. Cela ne permet pas de prendre en compte les ambiguités sur les entrées lexicales composées, entre le choix d’une entrée composée ou d’une suite de mots simples, comme pour “bien que”. En théorie, le modèle triclasse pourrait être utilisé à partir d’une segmentation en mots ambigue, ce qui fournirait à la fois une segmentation de la phrase en mots et un étiquetage de chaque mot. Toutefois, nous ne connaissons pas de système où ce niveau ait été mis en oeuvre.

Un problème symétrique concerne les contractions telles que “aux = à + les”. Elles sont parfois traitées en considérant que “aux” possède sa classe grammaticale propre, parfois en modifiant le texte de départ pour faire apparaître le couple “à + les”. Dans ce dernier cas, il faut également prévoir d’interdire l’interprétation de “les” comme pronom dans cette configuration.

Le dernier point concerne l'utilisation de plusieurs jeux de classes grammaticales différents. Chaque application développée utilise en général son propre jeu, qui dépend des contraintes spécifiques de l'application, et de l'approche de modélisation choisie par son concepteur. La question se pose alors de savoir comment, si on a défini son propre jeu de classe, pouvoir profiter de données linguistiques construites selon un autre jeu de classes, que ce soient des dictionnaires ou bien des textes étiquetés manuellement. Dans ces conditions, il est possible de modéliser la correspondance entre plusieurs jeux de classes (de façon déterministe ou probabiliste), et donc d'utiliser au mieux des données d'apprentissage construites avec d'autres jeux. Cette fonctionnalité importante est toutefois rarement présente dans les systèmes d'étiquetage.

CONCLUSION

Les modèles probabilistes sont devenus un des outils usuels du traitement du Langage Naturel. Les probabilités s'introduisent de façon naturelle dans de nombreuses applications, l'étiquetage grammatical, l'accentuation automatique, et même la traduction automatique. L'utilisation des probabilités donne un moyen simple de résolution des ambiguïtés, ce qui constitue une part importante des difficultés de ce domaine. Leur efficacité provient essentiellement de leur capacité d'apprentissage à partir de grands volumes de textes, qui autorise la construction de modèles complexes. La poursuite de l'évolution technologique, avec l'augmentation de la puissance des ordinateurs et de leurs capacités de stockage, ainsi que la plus grande disponibilité de données linguistiques sous forme informatique, en particulier de données vérifiées manuellement, vont encore favoriser l'utilisation de ces méthodes. Il reste toutefois qu'une performance optimale ne pourra vraisemblablement être obtenue que lorsque ces modèles probabilistes sauront intégrer des connaissances linguistiques suffisamment fines, ce qui constitue encore un sujet de recherche à peine effleuré.

RÉFÉRENCES

BAHL, L. R., JELINEK, F. et MERCER, R. L. (1983): “A maximum likelihood approach to continuous speech recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5: pp. 179–190.

BAHL, L. R. et MERCER, R. L. (1976): “Part of speech assignment by a statistical decision algorithm”, In *IEEE International Symposium on Information Theory*, pp. 88–89, Ronneby (Sweden).

BAUM, L. E. (1972): “An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process”, *Inequalities*, 3: pp. 1–8.

BEALE, A. D. (1985): “A probabilistic approach to grammatical analysis of written english by computer”, In *Proc. 2nd conference of the European Chapter of the ACL*, Geneva, Switzerland.

BEALE, A. D. (1988): “Lexicon and grammar in probabilistic tagging of written english”, In *Proc. 26th Annual Meeting of the ACL*, pp. 211–216, Buffalo, NY.

BENELLO, J., MACKIE, A. W. et ANDERSON, J. A. (1989): “Syntactic category disambiguation with neural networks”, *Computer Speech and Language*, (3): pp. 203–217.

BLACK, E., GARSIDE, R. et LEECH, G. (1993): *Statistically driven computer grammars of English: the IBM-Lancaster approach*, Rodopi.

BRILL, E. (1992): “A simple rule-based part of speech tagger”, In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*, Trento, Italy.

BRILL, E. (1994): “Some advances in transformation-based part of speech tagging”, In *Proceedings of the AAAI*.

BRILL, E., MAGERMAN, D., MARCUS, M. et SANTORINI (1990): “Deducing linguistic structure from the statistics of large corpora”, *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 275–282.

BRODDA, B. (1982): “Problems with tagging and a solution”, *Nordic Journal of Linguistics*, pp. 93–116.

BROWN, P. F., COCKE, J., PIETRA, S. A. D., PIETRA, V. J. D., JELINEK, F., LAFFERTY, J. D., MERCER, R. L. et ROOSSIN, P. S. (1990): “A statistical approach to machine translation”, *Computational Linguistics*, 16(2): pp. 79–85.

BROWN, P. F., PIETRA, V. J. D., DESOUZA, P. V., LAI, J. C. et MERCER, R. L. (1992): “Class based n-gram models of natural language”, *Computational Linguistics*, 18: pp. 467–479.

CHANOD, J.-P. et TAPANAINEN, P. (1995a): “Creating a tagset, lexicon and guesser for a french tagger”, In *Proceedings of EACL SIGDAT workshop on From Texts To Tags: Issues In Multilingual Language Analysis*.

CHANOD, J.-P. et TAPANAINEN, P. (1995b): “Tagging french – comparing a statistical and a constraint-based method”, In *Proceedings of EACL-95*.

CHARNIAK (1993): *Statistical Language Learning.*, MIT Press.

CHURCH, K. W. (1989): “A stochastic parts program noun phrase parser for unrestricted text”, In *Proc. ICASSP*, pp. 695–698, Glasgow (Scotland).

CODOGNO, M., FISSORE, L., MARTELLI, A., PIRANI, G. et VOLPI, G. (1987): “A hierarchical, mutual information based probabilistic language model”, In *Experimental evaluation of Italian language models for large-dictionary speech recognition*, pp. 159–162, Edinburgh (Scotland).

CUTTING, D., KUPIEC, J., PEDERSEN, J. et SIBUN, P. (1992): “A practical part-of-speech tagger”, In *Proceedings of the Third Conference on Applied Natural Language Processing*.

DE MARCKEN, C. G. (1990): “Parsing the LOB corpus”, In *Annual Meeting of the ACL*, pp. 243–251, Pittsburg, Pa.

DEBILI, F. (1994): “Rattachement prepositionnel dans les groupes nominaux”, *Seminaire ATALA*.

DEROSE, S. (1988): “Grammatical category disambiguation by statistical optimization”, *Computational Linguistics*, 14(1).

DEROUAULT, A.-M. et MERIALDO, B. (1986): “Natural language modeling for phoneme-to-text transcription”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6): pp. 742–749.

EL-BEZE, M., MERIALDO, B., ROZERON, B. et DEROUAULT, A.-M. (1994): “Accentuation automatique de textes par des methodes probabilistes”, *Techniques et Sciences Informatiques*, 13(6): pp. 797–815.

ELWORTHY, D. (1994): “Does Baum-Welch re-estimation help taggers?”, In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing (13–15 October 1994, Stuttgart)*: Association for Computational Linguistics.

GARSIDE, R. et LEECH, F. (1985): “A probabilistic parser”, In *Proc. 2nd conference of the European Chapter of the ACL*, Geneva, Switzerland.

JELINEK, F. (1985): “Markov source modeling in text generation”, In Skwirzinski, J. K., editor, *The Impact of Processing Techniques on Communications*: Nijhoff, Dordrecht, Nijhoff, Dordrecht.

KEMPE, A. (1994): “Probabilistic tagging with feature structures”, In *COLING-94*.

KLEIN, S. et SIMMONS, R. F. (1963): “A grammatical approach to grammatical coding of english words”, *JACM*, 10: pp. 334–347.

KUPIEC, J. (1992): “Robust part-of-speech tagging using a Hidden Markov Model”, *Computer Speech and Language*, 6: pp. 225–242.

LEECH, G., GARSIDE, R., et ATWELL, E. (1983): “The automatic grammatical tagging of the LOB corpus”, *Newsletter of the International Computer*

Archive of Modern English, 7: pp. 13–33.

MARCUS, M. P., SANTORINI, B. et MARCINKIEWICZ, M. A. (1993): “Building a large annotated corpus of English: The Penn Treebank”, *Computational Linguistics*, 19(2): pp. 313.

MARSHALL, I. (1983): “Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus”, *Computers and the Humanities*, pp. 139–150.

MERIALDO, B. (1994): “Tagging english text with a probabilistic model”, *Computational Linguistics*, 20(2).

NAKAMURA, M. et SHIKANO, K. (1989): “A study of english word category prediction based on neural networks”, In *Proc. ICASSP*, pp. 731–734, Glasgow (Scotland).

PAULUSSEN, H. et MARTIN, W. (1992): “Dilemma-2: a lemmatizer-tagger for medical abstracts”, In *Proceedings of the 3rd conference on applied language processing*, pp. 141–146, Trento, Italy.

RABINER, L. R. et JUANG, B. H. (1986): “An introduction to Hidden Markov Models”, *IEEE ASSP Magazine*, 3(1): pp. 4–16.

RAMSHAW, L. A. et MARCUS, M. P. (1994): “Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging”, In *Proceedings of the ACL Balancing Act workshop*.

SCHMID, H. (1994a): “Part-of-speech tagging with neural networks”, In *COLING*, pp. 172–176.

SCHMID, H. (1994b): “Probabilistic part-of-speech tagging using decision trees”, In *International Conference on New Methods in Language Processing*, Manchester, UK.

SCHUTZE, H. et SINGER, Y., (1994) “Part-of-speech tagging using a variable context Markov model”, In Mozer, M., Smolensky, P., Touretzky, D., Elman, J., et Weigend, A., editors, *Proceedings of the 1993 Connectionist Models Summer School*, Hillsdale, NJ.