CHAPITRE VIII
Post-traitement syntaxique

1 Introduction

Lorsque l'on s'adresse à un correspondant distant par l'intermédiaire d'un téléphone, on est parfois amené à lui épeler un nom de personne ou de ville. Ainsi le correspondant a voulu récupérer des erreurs d'incompréhension par une tâche d'épellation. Dans le même contexte il est intéressant de récupérer les erreurs commises par un système de reconnaissance en utilisant l'épellation comme moyen de correction. Ainsi, si un usager des chemins de fer demande un billet pour "PARIS" et que le système reconnaît "BARIS", il peut lui demander dans un premier temps de confirmer son choix et dans un deuxième temps, si sa requête n'a pas abouti d'épeler sa ville. Cependant cette tâche n'est pas simple puisqu'elle suppose la connaissance préalable de toutes les villes possibles. Intégrer toutes cette connaissance dans un système de reconnaissance n'est parfois pas évident à faire, toutefois des solutions existent à différents niveaux.

Il faut noter dès le départ que l'on dispose d'une information utile qui est la liste des noms de villes possibles. Cette liste constitue un dictionnaire. Si la taille de ce dictionnaire n'est pas très élevée, on peut l'intégrer dans la phase de décodage. N'empêche que pour des dictionnaires moyens (5000 villes), l'utilisation d'un réseau compilé est exclue. Toutefois le développement du réseau pendant la phase de décodage, contrôlé par la grammaire et une heuristique sur le score de vraisemblance du chemin partiel exploré, permet encore de traiter de tels dictionnaires [Dupont, 93].

Une deuxième approche consiste à rechercher la ville la plus probable dans le dictionnaire connaissant la séquence de lettres reconnues. Cette recherche implique l'utilisation d'un modèle

de Markov discret [Jouvet, 93-a]. Elle peut être appliquée soit à la meilleure solution ou bien aux N meilleures solutions [Jouvet, 93-b] avec N de l'ordre de 4 ou 8.

Une troisième approche consiste à calculer les N meilleures solutions avec N assez grand (de l'ordre de 100). La première solution, parmi les N développées, qui appartient au dictionnaire fournit la réponse du système.

Dans ce chapitre nous allons nous intéresser plus particulièrement à cette troisième approche.

2 Description du post-traitement

Le système complet, comprenant la recherche des N meilleures solutions et le post-traitement, est représenté sur la figure VIII.1. Pour une ville épelée, on va rechercher les N meilleures solutions. Chaque solution est comparée avec la liste de villes contenue dans le dictionnaire. Si la ville correspondant à la solution développée se trouve dans le dictionnaire, on parle alors de solution syntaxiquement correcte. Cependant si cette solution ne correspond pas à la ville prononcée, on parle alors de solution syntaxiquement correcte et incorrectement reconnue (solution erronée), c'est une erreur de reconnaissance. Dans le cas où elle correspond à la ville prononcée, cette solution est déclarée syntaxiquement correcte et correctement reconnue (une bonne solution), c'est une réponse correcte. Une autre possibilité est qu'une solution développée n'appartienne pas au dictionnaire. Cette solution est donc syntaxiquement incorrecte. Dans le cas où on développe N solutions et où il n'existe aucune solution syntaxiquement correcte dans cette liste, la ville prononcée est rejetée (solution rejetée). Cette procédure de post-traitement syntaxique s'arrête dès qu'on trouve une solution syntaxiquement correcte dans la liste proposée ou bien dès qu'on atteint le nombre maximal de solutions.

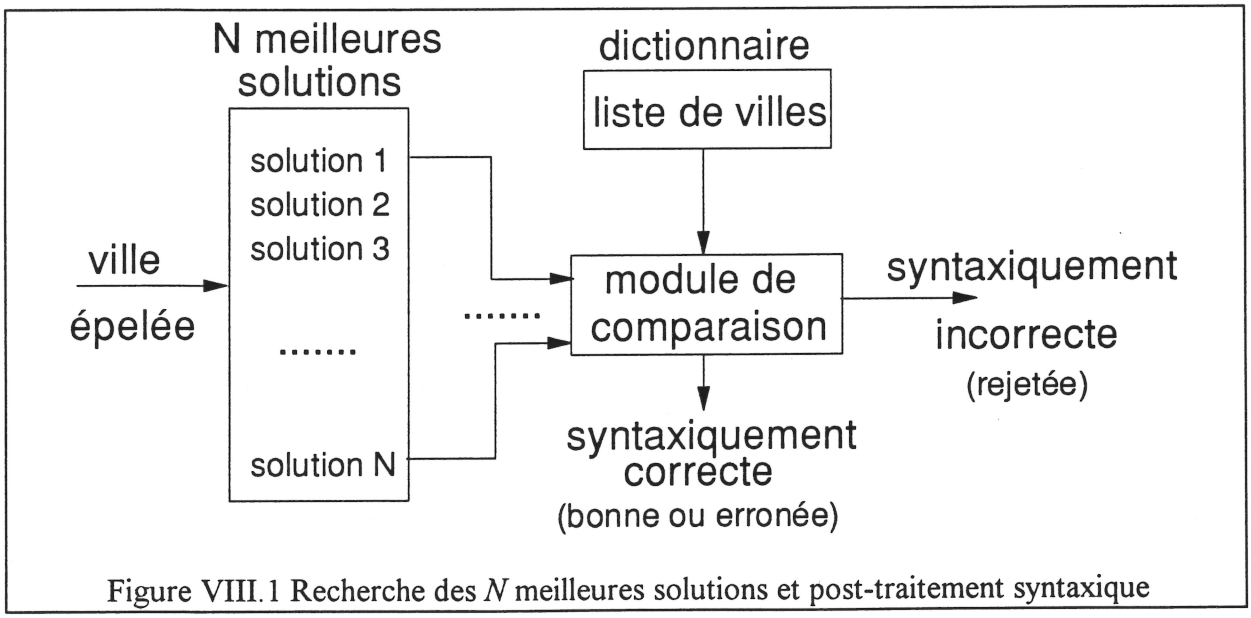

Figure VIII.1 Recherche des N meilleures solutions et post-traitement syntaxique

La procédure de calcul du taux d'erreur de substitution (syntaxiquement correcte et incorrectement reconnue) et du taux de rejet est décrite en annexe 4.

3 Résultats

Les résultats présentés dans ce paragraphe sont obtenus en développant les 100 meilleures solutions pour chaque ville épelée sur le corpus de test. On dispose pour les tests effectués de plusieurs dictionnaires contenant chacun un nombre fini de villes. Chaque dictionnaire comporte au moins 120 villes épelées et est complété d'une manière aléatoire par d'autres villes. Les résultats sont présentés pour des dictionnaires de taille 120, 500, 1000, 5000, 10000 et 30000 villes. Une ville se trouvant dans le dictionnaire des 500 villes peut ne pas se trouver dans le dictionnaire des 1000 villes (à cause du tirage aléatoire). Par contre 120 villes identiques se trouvent dans les 6 dictionnaires utilisés.

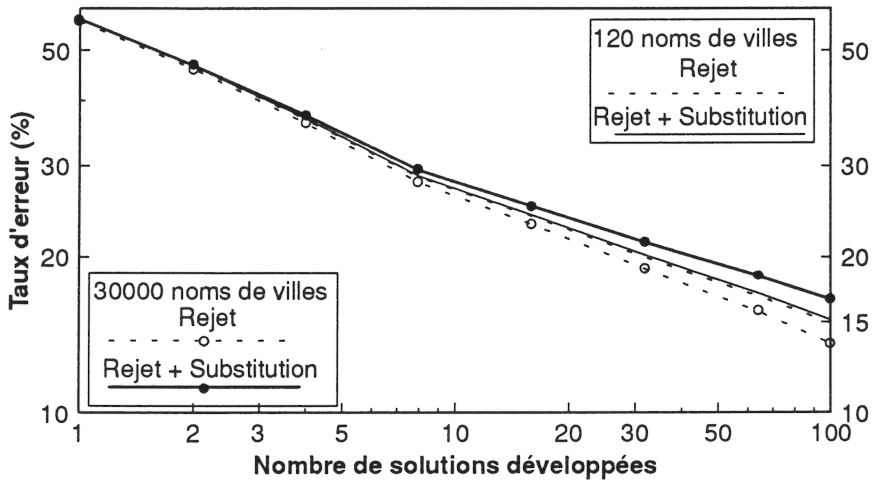

Figure VIII.2 Taux de rejet et de substitution sur les villes épelées

La figure VIII.2 présente les courbes de taux d'erreur de rejet d'une part et de rejet plus substitution d'autre part pour le dictionnaire des 120 villes et le dictionnaire des 30000 villes. On remarque qu'en augmentant le nombre de villes dans le dictionnaire, on commet plus d'erreurs. Ce qui est normal puisque certaines solutions syntaxiquement correctes vont se confondre avec d'autres. Un exemple relevé concerne les villes "C.A.N.N.E.S" et "V.A.N.N.E.S" qui ne diffèrent que par une seule lettre.

Le taux de rejet avoisine les 15% dans le cas d'un dictionnaire de 120 villes et 13.6% dans le cas d'un dictionnaire de 30000 villes, et cela pour 100 solutions développées. Le taux de substitution quant à lui est de 0.2% pour un dictionnaire de 120 villes et de 2.9% pour un dictionnaire de 30000 villes. Il est à noter que ce taux de substitution est relativement bas.

Dans ce cas, on peut envisager de réaliser des applications impliquant l'épellation des villes s'il est possible de renvoyer la communication vers un opérateur en cas de rejet.

Le tableau suivant résume les différents taux d'erreur obtenus en développant 100 solutions et en utilisant 6 dictionnaires de tailles différentes.

Taille du dictionnaire	Taux (%) épellation villes N=100		
	Taux de rejet	Taux de substitution	Taux d'erreur global
120	14.9%	0.2%	15.1%
500	14.9%	0.2%	15.1%
1000	14.9%	0.3%	15.2%
5000	14.4%	0.8%	15.2%
10000	14.3%	1.5%	15.8%
30000	13.6%	2.9%	16.5%

Tableau VIII.1 Performances du post-traitement syntaxique sur les villes épelées pour N=100

Sans traitement spécifique, les résultats obtenus sur la reconnaissance des villes épelées ont montré que le taux d'erreur sur le corpus de test avoisine les 60 % (voir chapitre VI). Avec un post-traitement syntaxique et pour un nombre de solutions développées égal à 100 les résultats obtenus montrent que lorsque la solution est syntaxiquement correcte elle est incorrectement reconnue dans 3% des cas uniquement, si la taille du dictionnaire est de 30000 noms de villes. Cependant le taux d'erreur global (solution rejetée + solution incorrectement reconnue) reste très élevé autour de 16.5%. Pour réduire ce taux d'erreur, une solution [Jouvet, 93-b] consiste à rechercher l'épellation la plus probable en fonction des lettres reconnues (HMM discret).