Adaptation de modèles de langage à l'utilisateur et au registre de langage : expérimentations dans le domaine de l'aide au handicap

Tonio Wandmacher, Jean-Yves Antoine

Université François-Rabelais de Tours – Laboratoire d'Informatique
{tonio.wandmacher ; jean-yves.antoine}@univ-tours.fr

Résumé

Les modèles markoviens de langage sont très dépendants des données d'entraînement sur lesquels ils sont appris. Cette dépendance, qui rend difficile l'interprétation des performances, a surtout un fort impact sur l'adaptation à chaque utilisateur de ces modèles. Cette question a déjà été largement étudiée par le passé. En nous appuyant sur un domaine d'application spécifique (prédiction de texte pour l'aide à la communication pour personnes handicapées), nous voudrions l'étendre à la problématique de l'influence du registre de langage. En considérant des corpus relevant de cinq genres différents, nous avons étudié la réduction de cette influence par trois modèles adaptatifs différents : (a) un modèle cache classique favorisant les *n* derniers mots rencontrés, (b) l'intégration au modèle d'un dictionnaire dynamique de l'utilisateur et enfin (c) un modèle de langage interpolé combinant un modèle général et un modèle utilisateur mis à jour dynamiquement au fil des saisies. Cette évaluation porte un système de prédiction de texte basé sur un modèle trigramme.

Mots-clés : modèles de langage, adaptation, utilisateur, thème, aide au handicap.

Abstract

Statistical language models (LM) are highly dependent on their training resources. This makes it not only difficult to interpret evaluation results, it also has a strong impact on users of a LM-based application. This question has already been studied by others. Focussing on a specific domain (text prediction in a communication aid for handicapped persons) we want to extend it to the influence of the language register. Considering corpora from five different registers, we discuss three methods to adapt a language model to its actual language resource hereby reducing the effect of training dependency: (a) a simple cache model augmenting the probability of the *n* last inserted words, (b) a dynamic user dictionary, keeping every unseen word and (c) an interpolated LM combining a base model with a currently updated user model. Our evaluation is based on the results obtained from a text prediction system working on a trigram LM.

Keywords: language model, dynamic adaptation, user, theme, AAC systems.

1. Introduction

Les modèles markoviens de langage (N-grammaires) sont étroitement liés aux données sur lesquelles a été réalisé leur apprentissage. Ce qui signifie qu'ils ont des capacités de généralisation très réduites en dehors du style de langage sur lesquels ils ont été entraînés. Leurs performances dépendent donc d'une manière aiguë de la similarité du corpus d'apprentissage avec la tâche considérée. Prenons l'exemple de la reconnaissance de parole. Les évaluations autour de la tâche *broadcast news* ont montré qu'un corpus journalistique comme celui du journal *Le Monde* est adapté à la transcription automatique des journaux télévisés. Au contraire, un modèle appris sur ce corpus sera totalement inopérant pour la reconnaissance en dialogue oral spontané.

Une réponse simple à ce problème serait de construire des ressources spécifiques pour chaque nouvelle situation. La constitution de corpus d'apprentissage a cependant un coût (pour l'écrit

mais plus encore pour l'oral) qui obère une telle solution. C'est pourquoi il est d'usage d'interpoler un modèle de langage général avec un second appris sur un corpus plus petit spécifique à la tâche (Woodland *et al.,* 1998). Pour certaines applications, il peut même être envisagé de faire évoluer le modèle résultant en cours d'utilisation : le modèle est donc construit par adaptation dynamique au cours du temps.

Bien que sous-optimale, cette approche donne des résultats satisfaisants et elle est largement utilisée dans les systèmes actuels. Dans cet article, nous nous interrogeons sur la pertinence de ces techniques d'adaptation dans des situations limites où le langage à modéliser varie fortement suivant le contexte d'utilisation. C'est le cas de l'aide à la communication pour personnes handicapées, où chaque association patient / registre de communication constitue un problème spécifique. En partant de ce cadre applicatif particulier, nous nous intéresserons donc à des questions aussi générales que l'influence du registre de langage (Biber 1988, 1993) sur les modèles, ce qui pose la question de leur adaptation sur des données limitées.

Dans un premier temps, nous présenterons la problématique de la prédiction de mots pour l'aide à la communication. Ensuite, nous détaillerons plusieurs expérimentations qui cherchent à montrer l'influence du registre de langage sur les comportement des modèles markoviens (N-grams). Nous comparerons alors des techniques d'adaptation bien connues telles que l'interpolation EM de modèles ou encore l'utilisation d'un modèle cache. En guise de perspectives, nous réfléchirons à l'utilisation de l'analyse sémantique latente (ASL) pour une meilleure adaptation de la prédiction dans le cas de l'aide à la communication.

2. Aide à la communication pour personnes handicapées

Les communicateurs, ou systèmes d'aide de communication assistée (AAC pour *Alternative and Augmentative Communication* en anglais) ont pour objectif de restaurer les capacités de communication de personnes souffrant d'un handicap moteur sévère s'accompagnant d'une perte de l'usage de la parole (Infirmités Motrices Cérébrales, Scléroses Latérales Amyotrophiques, *locked-in syndrom,* etc.). Quelle que soit la maladie considérée, la communication est privée de son support oral habituel et les capacités de contrôle physique de l'environnement par la personne handicapée sont très limitées.

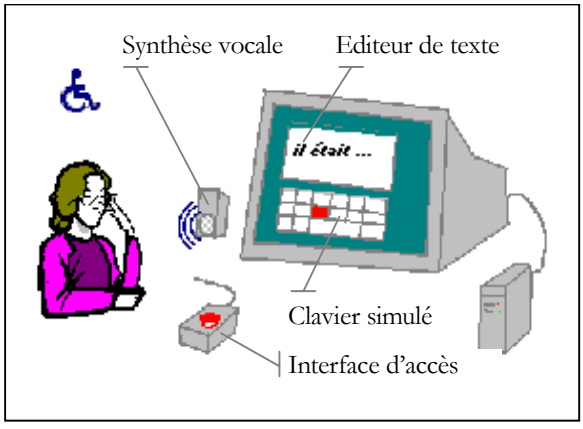

Figure 1. Système de communication assisté

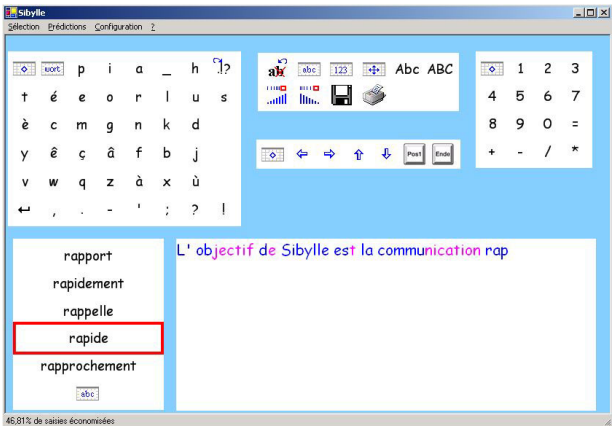

Figure 2. Interface du système SIBYLLE

Le principe des systèmes d'aide à la communication repose sur l'écriture de phrases à l'aide d'un tableau de symboles (mots, lettres voire icônes pour les patients aphasiques). Le message est construit en sélectionnant successivement sur le tableau les différents symboles qui le

composent (figures 1 et 2). L'intervention de la personne handicapée se limite à la désignation des symboles à l'aide d'un clavier virtuel simulé sur un écran d'ordinateur.

D'une manière générale, un système AAC se compose de quatre composants principaux. Tout d'abord, un dispositif physique joue le rôle de périphérique d'entrée de l'ordinateur. Cette interface matérielle dépend du geste libre laissé par le handicap. Il peut s'agir d'un joystick, d'une commande oculaire, d'une commande par souffle, d'un simple bouton poussoir, etc. Une caractéristique importante est le degré de liberté qu'elle permet pour manipuler l'ordinateur. Le plus souvent, le patient n'a plus que la possibilité de réaliser l'équivalent d'un simple clic (commande de l'environnement de type « tout ou rien »). C'est donc par clics successifs que la personne handicapée va sélectionner les éléments de son message sur un clavier simulé présenté à l'écran. Par exemple, dans le cas d'un clavier à défilement linéaire, un curseur se déplace sur le clavier virtuel, la personne n'ayant plus qu'à cliquer lorsque le curseur arrive sur la lettre recherchée. Enfin, les deux derniers composants du système sont un éditeur de texte (pour rédiger des courriels ou documents écrits) et une synthèse de parole pour la vocalisation de messages en cas de communication orale.

Le problème majeur des systèmes de communication assistée est la lenteur de la composition des messages. La tâche de saisie est généralement longue (1 à 5 mots par minute en moyenne) et extrêmement fatigante pour les patients. Pour accélérer la saisie, deux approches complémentaires sont envisageables. La première vise à optimiser la sélection sur le clavier simulé. La seconde consiste à minimiser le nombre de saisies en tentant de prédire les mots qui peuvent survenir à la suite de ceux déjà saisis. Plusieurs méthodes peuvent être utilisées pour réaliser cette prédiction, parmi lesquelles l'utilisation de modèles de langage markoviens qui peuvent fournir une liste d'hypothèse lexicales en fonction des (N-1) derniers mots saisis. D'autres modèles stochastiques plus complexes comme le modèle structurel (Schadle *et al.* 2004) par exemple, peuvent être élaborés, mais nous nous limiterons ici aux N-grammaires qui donnent des résultats déjà satisfaisants.

Après chaque saisie, une liste de prédictions lexicales est présentée à l'écran (voir figure 2). Si l'utilisateur retient une de ces propositions, le texte est automatiquement complété, ce qui évite la saisie des dernières lettres du mot. L'efficacité des systèmes de prédiction est évaluée par le taux d'économie de saisies, suivant une métrique appelée *ksr* (*keystroke saving rate*) :

$$ksr = (1 - k_p/k_a) \cdot 100 \quad (1)$$

avec k_p et k_a le nombre d'appuis sur le dispositif d'entrée respectivement avec et sans prédiction.

3. AAC et adaptation des modèles de langage

Dans le cas d'un apprentissage sur le corpus du *Monde*, puis d'un test sur une autre partie de ce corpus, notre système de prédiction *SIBYLLE* est capable d'atteindre des taux d'économie variant entre 50 % et 55 % (Schadle *et al.* 2004). Lorsqu'on teste ce système sur des productions de personnes handicapées en gardant le même corpus d'apprentissage, ce taux peut baisser extrêmement fortement, surtout dans le cas de patients qui sont en outre agrammatiques.

La question de l'adaptation des modèles de langage se pose donc avec acuité dans le cas de la prédiction pour les systèmes AAC. Ce problème est en outre renforcé par le fort particularisme de chaque situation d'utilisation. Étant donné que les patients répondent à des tableaux cliniques très variés, mais également qu'ils/elles vont utiliser le système pour des

usages variés relevant de registres différents (communication écrite ou orale, rédaction de documents écrits voire d'œuvres littéraires), nous devons faire face à des demandes d'adaptation multi-factorielles. Si d'autres travaux ont déjà montré l'intérêt des modèles adaptatifs en AAC (Trost *et al.,* 2005), cette multiplicité d'influences se traduit par un éparpillement des données nécessaires à l'adaptation : pour chaque contexte d'adaptation, les données d'apprentissage dynamique seront très limitées. C'est pour quantifier ce phénomène que nous avons conduit les expérimentations présentées dans cet article. Nous limiterons ici notre étude à l'influence du registre de langage (Biber, 1988, 1993), problématique qui n'a jusqu'ici été que peu abordée dans la littérature sur l'adaptation des modèles de langage.

4. Méthode d'évaluation

4.1. Le modèle de langage et le corpus d'apprentissage

Bien que *SIBYLLE* comporte un modèle stochastique plus élaboré se basant sur une analyse syntaxique superficielle (Schadle *et al.,* 2004), nous avons choisi d'utiliser un modèle trigramme afin d'étendre la portée de cette étude (plus grand généricité en termes d'application).

Le modèle trigramme a été entraîné sur un corpus journalistique (*Le Monde* de l'année 1999) comprenant 5,58 millions de mots ; notre vocabulaire comprend 141.022 mots. Pour éviter le problème d'événements non-observés (*zero frequency problem*) en cours de prédiction, notre prédicteur utilise un lissage absolu (*absolute discounting method*). Celui-ci retire une partie fixe de chaque événement observé ; la masse globale retirée est cependant variable et dépend de la proportion entre le nombre d'événements observés et le nombre de contextes différents.

4.2. Les corpus de test

Nous avons effectué notre évaluation sur 5 corpus de test issus de registres de langage assez distincts pour montrer l'influence des corpus (et de leur genre) d'une manière étendue (tableau 1).

Nous avons calculé la *ksr* à l'aide d'un module de simulation de la manière suivante : le module charge le corpus de test et l'insère lettre par lettre dans le système. Si le prédicteur affiche le bon mot dans la liste de mots prédits, le simulateur ajoute le mot (ou la partie restante) au texte. Ce simulateur représente ainsi l'utilisateur idéal choisissant toujours (et dès sa première apparition) le bon mot de la liste. En comptant le nombre de caractères ajoutés, la *ksr* peut être calculée. L'étape de sélection de mot coûte un caractère de plus.

Corpus	Description	Nombre de mots
Le Monde	Extrait du journal *le Monde* (année 1999) différente du corpus d'apprentissage.	20.009
Scientifique	Article scientifique (non-publié) références bibliographiques comprises	8.766
Littéraire	Premier chapitre de *Germinal* d'Émile Zola (1885).	20.928
Parole	Transcription de dialogues oraux spontanés dans le domaine du renseignement touristique (corpus *OTG* (Antoine *et al.* 2002)). Énoncés prononcés par l'hôtesse d'accueil dans 315 dialogues.	15.435
Courriel	Ensemble de courriels personnels des auteurs ; en-têtes, réponses attachées et hyperliens filtrés.	8.874

Tableau 1. Corpus de test utilisés

Pour tous les tests, nous avons fixé la taille de la liste de prédiction à 5 éléments, taille de fenêtre fréquemment utilisée dans le domaine d'aide à la saisie (Trost *et al.,* 2005).

4.3. Résultats – Prédicteur de base

La figure 3 montre les résultats en termes de *ksr* pour les 5 corpus de test considérés, ainsi que les dégradations de performances entre la situation de contrôle (test et apprentissage dans le même registre) et celles correspondant à des tests sur d'autres registres

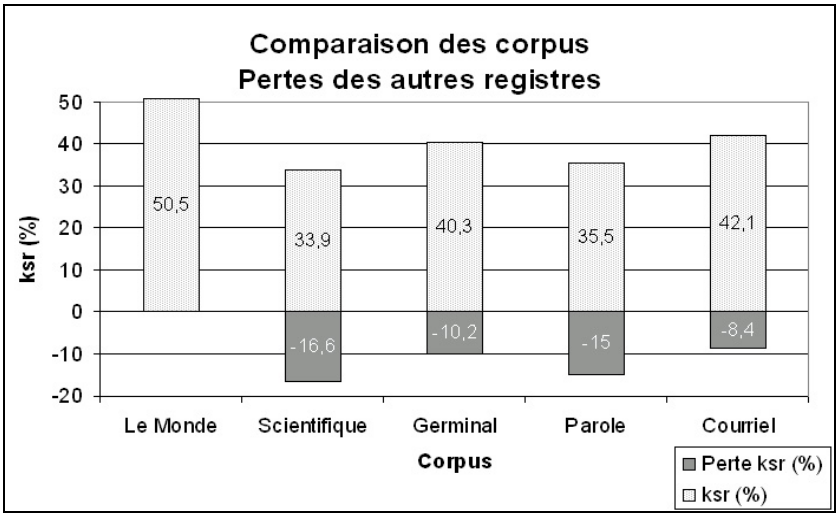

Figure 3. Résultats (ksr) pour les 5 corpus de test et les pertes des corpus issus d'autres registres par rapport au corpus du Monde

La *ksr* de plus que 50 % observée dans la situation de contrôle était attendue. Schadle *et al.* (2004), ainsi que Trost *et al.* (2005) obtiennent des taux similaires pour un modèle trigramme. On observe au contraire une perte très sensible de performances pour les autres corpus (8,4 % - 16,6 %), ce qui témoigne d'une influence importante du corpus d'apprentissage et surtout du registre correspondant Dans les expérimentations qui suivent, plusieurs modèles adaptatifs ont été testés pour réduire cette influence sur la prédiction.

5. Modèles adaptatifs

5.1. Modèle cache

5.1.1. Fonctionnement

Le modèle cache repose sur l'idée suivant laquelle les mots récemment apparus dans un contexte ont une probabilité d'occurrence plus élevée (Kuhn et De Mori, 1990 ; Rosenfeld, 1996). Ainsi, on garde les n (env. 50 – 200) derniers mots qui ont été insérés et on augmente leur probabilité d'un facteur p constant. Clarkson et Robinson (1997) utilisent un facteur décroissant, mais les avantages par rapport au facteur constant sont minimes.

Dans notre approche, nous avons adopté un modèle de cache assez simple : la taille maximale du cache était de 100, avec un $p = 0{,}0005$. Ainsi, la masse de probabilité retirée pour le cache est de 0,05 au maximum. Les mots outils (*stopwords*) n'étaient pas considérés pour le cache.

5.1.2. Évaluation

Corpus :	*Le Monde*	*Scientifique*	*Littéraire*	*Parole*	*Courriel*
ksr absolu	51,07	35,11	40,76	39,0	42,98
Av. baseline	+0,55	+1,14	+0,47	+3,50	+0,87

Tableau 2. Résultats (ksr) basés sur le modèle trigramme + cache pour les 5 corpus de test et leurs avantages par rapport au niveau de base (trigramme)

Les résultats du modèle trigramme + cache (tableau 2) montrent une amélioration constante mais limitée des résultats sur les corpus écrits. À l'opposé, le test sur le dialogue oral présente une progression bien plus importante (+3,5 %). Ces résultats peuvent être attribués à une plus grande contextualité du dialogue oral qu'à l'écrit. Plus précisément, cet effet est renforcé par le caractère finalisé du dialogue : comme il s'agit de dialogues entre une hôtesse et des clients, certaines séquences prototypiques (« *Bonjour madame* », « *Je vous en prie* » …) se répètent fréquemment. La prédiction de ces mots est ainsi facilitée par le modèle cache.

5.2. Dictionnaire utilisateur adaptatif

5.2.1. Fonctionnement

Comme toutes les applications TALN, les systèmes de prédiction sont confrontés au problème des mots inconnus (MHV : mots hors vocabulaire). Comme un mot hors vocabulaire n'est pas prédictible, la *ksr* pour ce mot reste désespérément à 0 %. Le taux de MHV dépend de la similarité entre le registre du corpus d'apprentissage et celui de l'utilisation, mais il reste toujours important, même en considérant un corpus du même registre (4,83 % dans le cas du corpus *Le Monde*).

Pour réduire ce taux de mots non-prédictibles, nous avons conçu un dictionnaire d'utilisateur (DU) adaptatif, qui intègre immédiatement tout mot inconnu avec une fréquence initiale de 1. Au cours de l'utilisation, les fréquences des MHV sont adaptées. À la fin d'une session, ces mots ainsi que leurs fréquences sont enregistrés dans un fichier séparé.

5.2.2. Évaluation

Comme le montrent les résultats du tableau 3, l'utilisation d'un dictionnaire utilisateur permet une réduction significative de la proportion de mots hors vocabulaire.

Corpus :	*Le Monde*	*Scientifique*	*Littéraire*	*Parole*	*Courriel*
% MHV sans DU	4,83	16,33	4,44	2,18	8,83
% MHV avec DU	1,84	9,34	2,93	0,96	5,19

Tableau 3. Taux de mots hors vocabulaire (MHV) sans/avec un dictionnaire d'utilisateur pour les 5 corpus considérés

Sans surprise, cette réduction a un effet bénéfique sur la *ksr* comme le montre le tableau 4 ci-dessous. On peut ainsi atteindre des gains de 2,45 %. On notera cependant qu'il n'y a pas de correspondance absolue entre réduction du taux de MHV et augmentation de la *ksr*. Ainsi, la nette réduction du nombre de MHV dans le cas du corpus courriel se traduit par une progression assez limitée de la prédiction : Une partie importante des mots ainsi intégrés au DU sont des hapax legomena, cependant les mots inconnus ne sont prédictibles par le DU qu'à partir de leur deuxième occurrence.

Corpus :	*Le Monde*	*Scientifique*	*Littéraire*	*Parole*	*Courriel*
ksr absolu	51,31	36,42	41,26	35,65	42,46
Av. baseline	+0,79	+2,45	+0,97	+0,15	+0,35

Tableau 4. Résultats (ksr) basés sur le modèle trigramme + DU pour les 5 corpus de test et leurs avantages par rapport au niveau de base (trigramme)

Enfin, on observe que les meilleurs cas d'améliorations par genre (*scientifique*, *littéraire*, *journal*) sont assez complémentaires de ceux observées avec le modèle cache (*parole*, *courriel*, *scientifique*). Ces deux approches peuvent être utilement envisagées de concert.

5.3. Le modèle interpolé

5.3.1. Fonctionnement

Dans ce modèle, deux prédicteurs sont utilisés en même temps : un prédicteur de base, appris sur le corpus *Le Monde* et un prédicteur utilisateur (PU) qui apprend un trigramme spécifique en cours de saisie. Ce dernier est donc en charge de l'adaptation aux productions de l'utilisateur, voire du registre de langue utilisé. En gardant tout mot inconnu et en ne considérant que le texte inséré au préalable, ce modèle remplit les fonctions de dictionnaire d'utilisateur et, dans une moindre mesure, de cache. Les deux modèles sont interpolés de la manière suivante :

$$P'(m_i \mid c_j) = \lambda_1 \cdot P_{Pbase}(m_i \mid c_j) + \lambda_2 \cdot P_{PU}(m_i \mid c_j) \quad (2)$$

où c_j est le contexte actuel (= $m_{i-1}\ m_{i-2}$) et λ_1, λ_2 sont les facteurs de pondération ($\lambda_1 + \lambda_2 = 1$). Ceux-ci sont déterminés par l'algorithme EM qui estime à partir des contributions passées (en termes de probabilité) les contributions futures des deux modèles (Jelinek et Mercer, 1980).

5.3.2. Évaluation

Le tableau 5 donne la *ksr* du modèle interpolé sur les 5 corpus en précisant à chaque fois les améliorations de performances observées par rapport au modèle trigramme.

Au total, nous observons une forte amélioration (6,6 % - 14,6 %) apportée par ce modèle dans tous les registres considérés. On relève en particulier une progression assez étonnante (+ 11 %) pour le corpus contrôle (*Le Monde*), de même qu'on observe une amélioration maximale pour le corpus de dialogue oral finalisé (+14,6 %). Comme le corpus du *Monde* est assez similaire au corpus d'apprentissage, nous nous interrogeons encore sur l'amélioration observée.

Corpus :	*Le Monde*	*Scientifique*	*Littéraire*	*Parole*	*Courriel*
ksr absolu	61,58	43,09	46,89	52,14	51,62
Av. baseline	+11,06	+9,12	+6,60	+14,64	+9,51

Tableau 5. Résultats (ksr) basés sur le modèle interpolé pour les 5 corpus de test et leurs avantages par rapport au niveau de base (trigramme)

Au contraire, la progression pour la parole s'explique aisément : différence marquée de genre entre la parole spontanée et l'écrit préparé du journal *Le Monde*, mais également degrés de finalisation très différents (vocabulaire beaucoup moins étendu dans le cas du corpus oral). Le modèle utilisateur peut ainsi apprendre très vite sur ce type de registre.

Une autre particularité intéressante est le comportement des facteurs de pondération. Dès que le PU a appris sur environ 1000 mots, les facteurs ne changent quasiment plus, restant de 0,44 à 0,49 pour le prédicteur de base, et de 0,56 à 0,51 pour le PU : l'adaptation des modèles est très rapide, ce qui est très intéressant dans le cadre de l'aide au handicap : cela montre qu'il est possible de réaliser une adaptation à la fois au registre de communication et à l'utilisateur.

6. Conclusion

Ces expérimentations montrent combien un modèle de langage est dépendant du registre du corpus d'apprentissage. Nos avons observé une baisse très significative des performances lorsque les registres des corpus d'apprentissage et de test ne correspondaient pas. Cette perte confirme ici encore la nécessité des méthodes adaptatives.

Nous avons testé trois approches pour l'adaptation d'un modèle de langage à son utilisateur et au registre de communication : modèle cache, dictionnaire utilisateur adaptatif et interpolation d'un modèle général avec un modèle spécifique qui apprend exclusivement sur le texte saisi par l'utilisateur. Les bénéfices dus au cache et au DU sont constant sur tous les corpus de test, mais restent limités (de + 0,5 à +3,5 %). Par contre, le modèle interpolé autorise une amélioration des performances bien plus intéressante (entre +6,6 % et 14,6 %, cf. figure 4). L'adaptation du modèle interpolé est par ailleurs très rapide (1000 mots le plus souvent).

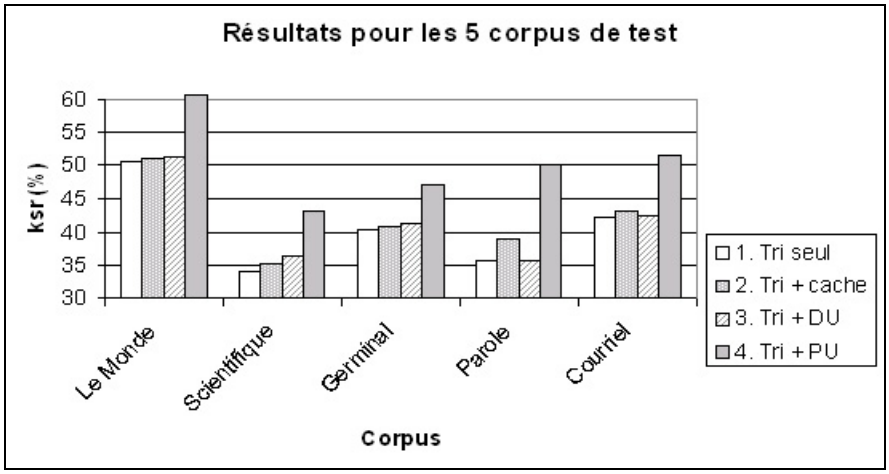

Figure 4. Aperçu des résultats pour chaque modèle testé

Ces résultats montrent que l'adaptation des modèles de langage est envisageable sur des situations (couple utilisateur + registre de langage) où les données d'adaptations sont relativement limitées (cas de l'aide au handicap).

7. Perspectives – Adaptation thématique

Une méthode ultérieure pour adapter dynamiquement le modèle au contexte immédiat est l'adaptation thématique. Il est évident que la probabilité d'occurrence de mots de contenu dépend fortement du thème actuel (Lesher *et al.,* 2002). Par exemple, un mot a priori assez rare (donc improbable) comme '*contrepoint*' aura une probabilité d'occurrence très élevée dans un contexte de musique baroque.

Pour l'identification du thème du contexte plusieurs approches ont été proposées, dont celui de Bigi *et al.* (2001). Une autre approche, le modèle *trigger* (Rosenfeld, 1996 ; Matiasek et Baroni, 2003) utilise des collocations pour s'adapter au mode de l'utilisation actuel. Dans ce

modèle un mot déclencheur augmente (dès qu'il est utilisé) la probabilité d'autres mots associés. Cependant, les gains apportés par ces modèles ne sont pas très importants.

Nous nous proposons d'étudier un nouveau modèle d'adaptation thématique, basé sur l'analyse sémantique latente (ASL) (Deerwester *et al.,* 1990). Avec ce modèle, tout mot m_i du vocabulaire est représenté par un vecteur de haute dimension, la distance entre les vecteurs correspond (selon la théorie) à la similarité sémantique des mots $\vec{m}_i$. Un contexte peut être représenté par la somme des vecteurs des mots qu'il contient et a ainsi la même dimensionnalité que tout vecteur mot (cf. Landauer *et al.,* 1997). On peut calculer le vecteur du contexte actuel c (= m_1, ..., m_n) de la manière suivante :

$$\vec{c} = \sum_{i=1}^{n} \vec{m}_i \tag{3}$$

Comme le vecteur $\vec{c}$ reflète la sémantique de la section déjà insérée, on peut supposer que les vecteurs des termes proches sont sémantiquement liés à ce contexte. Ce vecteur est comparable à tout mot du vocabulaire par une des mesures de similarité (produit scalaire, distance euclidienne, cosinus etc.). La formule suivante montre comment cette distance peut être incorporée dans le modèle de langage (Coccaro et Jurafsky, 1998):

$$P'(m_i) = \lambda_1 \cdot P(m_i) + \lambda_2 \cdot \frac{\cos(\vec{c}, \vec{m}_i) - \cos_{\min}}{\sum_{j=1}^{N} \cos(\vec{c}, \vec{M}_j) - \cos_{\min}} \tag{4}$$

où $P(m_i)$ est la probabilité initiale d'un mot m_i (issue du modèle de base) ; λ_1, λ_2 sont des facteurs de pondération ($\lambda_1+\lambda_2 = 1$), le dénominateur (M_j étant n'importe quel mot du vocabulaire) et $\cos_{\min}$ normalisent la valeur du cosinus afin d'obtenir un facteur probabiliste.

Nous ne pouvons pas encore présenter des résultats pour cette approche, mais nous pensons qu'elle augmentera encore la *ksr*. En outre, comme la liste de prédiction devrait ainsi contenir des mots sémantiquement liés au contexte, nous espérons que l'utilisateur sera ainsi aidé dans le processus de recherche de mots.

8. Remerciements

Cette recherche a été partiellement supporté par le DAAD (service allemand des échanges universitaires).

Références

ANTOINE J.-Y., LETELLIER-ZARSHENAS S., NICOLAS P., SCHADLE I. (2002). « Corpus OTG et ÉCOLE_MASSY : vers la constitution d'un collection de corpus francophones de dialogue oral diffusés librement ». In *Actes de TALN 2002*. Nancy : 319-324.

BIBER D. (1988). *Variations across speech and writing*. Cambridge University Press, Cambridge.

BIBER D. (1993). « Using Register-Diversified Corpora for General Language Studies ». In *Computational linguistics* 19 (2) : 219-241.

BIGI B., BRUN A., HATON J., SMAILI K., ZITOUNI I. (2001). « Dynamic Topic Identification: Towards Combination of Methods ». In *Proceedings of the Recent Avances in NLP workshop* : 255-257.

CLARKSON P.R., ROBINSON A.J. (1997). « Language Model Adaptation using Mixtures and an Exponentially Decaying Cache ». In *Proceedings of IEEE International Conference on Speech and Signal Processing*. Munich.

COCCARO N., JURAFSKY D. (1998). « Towards better integration of semantic predictors in statistical language modelling ». In *Proceedings of ICSLP-98*. Sydney.

DEERWESTER S.C., DUMAIS S., LANDAUER T., FURNAS G., HARSHMAN R. (1990). « Indexing by Latent Semantic Analysis ». In *Journal of the American Society of Information Science* 41 (6) : 391-407.

JELINEK F., MERCER R. (1980). « Interpolated estimation of Markov source parameters from sparse data ». In *Pattern Recognition in Practice* : 381-397.

KUHN R., DE MORI R. (1990). « A Cache-Based Natural Language Model for Speech Reproduction ». In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (6) : 570-583.

LANDAUER T.K., LAHAM D., REHDER B., SCHREINER M.E. (1997). « How well can passage meaning be derived without using word order ? A comparison of LSA and humans ». In *Proceedings of the 19th annual meeting of the Cognitive Science Society.* Erlbaum Mawhwah, NJ : 412-417.

LESHER G.W., MOULTON B.J, HIGGINBOTHAM D.J., ALSOFROM B. (2002). « Limits of human word prediction performance ». In *Proceedings of the CSUN 2002*. California State University, Northridge.

MATIASEK H., BARONI M. (2003). « Exploiting long distance collocational relations in predictive typing ». In *Proceedings of the EACL-03 Workshop on Language Modeling for Text Entry Methods*. Budapest.

ROSENFELD R. (1996). « A maximum entropy approach to adaptive statistical language modelling ». In *Computer Speech and Langua*ge 10 (1) : 187-228.

SCHADLE I., ANTOINE J.-Y., LE PÉVÉDIC B., POIRIER F. (2004). « SibyMot : Modélisation stochastique du langage intégrant la notion de chunks ». In *Actes de TALN 2002*. Fès.

TROST H., MATIASEK J., BARONI M. (2005). « The Language Component of the FASTY Text Prediction System ». In *Applied Artificial Intelligence* 19 (8) : 743-781.

WOODLAND P.C., ODELL J.J., HAIN T., MOORE G.L., NIELSER T.R., TUERK A., WHITTAKER E.W.D. (1998). « Improvements in Accuracy and Speed in the HTK Broadcast News Transcription System ». In *Proceedings of the Eurospeech'98*. Budapest.