Mise à jour automatique du modèle de langage d'un système de transcription

Alexandre Allauzen[1,2], *Jean-Luc Gauvain*[1]

[1] Groupe Traitement du Langage Parlé (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France

[2] Institut National de l'Audiovisuel (http://www.ina.fr/)
4 Avenue de l'Europe, 94366 Bry-sur-Marne cedex, France

{allauzen,gauvain}@limsi.fr

RÉSUMÉ

This paper investigates the problem of automatic adaptation of the vocabulary and the language models (LM) of a broadcast news speech transcription system. We propose to make use of written Internet news sources which are available on a daily basis to model the thematic changes typical of the news domain. For each news source a specific normalization is needed. The lexicon is updated daily and an up-to-date LM is estimated using only recent data. Adaptation is performed by interpolating the up-to-date LM with a standard (and fixed) LM. Each day the data collected from one of the sites is reserved as an evaluation corpus. Experiments carried during the month of January 2002 show a relative reduction in out-of-vocabulary rate of 32% and a 13% reduction in perplexity compared to using a fixed language model.

1. INTRODUCTION

Les progrès récents en matière de reconnaissance automatique de la parole permettent d'ores et déjà de transcrire des émissions radio ou télévisées avec une qualité suffisante pour indexer ces documents automatiquement. L'état de l'art pour la transcription automatique de journaux télédiffusés correspond à un taux d'erreur sur les mots de l'ordre de 25% (pour le français), soit un mot sur quatre incorrectement transcrit. Les erreurs sont dues d'une part à l'imperfection des modèles utilisés par le système de transcription, et d'autre part à la grande variabilité rencontrée dans les documents audio. Étant donnés les coûts liés à la préparation des corpus d'apprentissage, le lexique et les modèles de langages (ML) utilisés par un système de transcription sont généralement développés à partir d'un corpus figé qui n'est pas contemporain du document à transcrire (un décalage d'un an est commun). Or les caractéristiques linguistiques des documents à traiter dépendent fortement de l'actualité. On observe en particulier une augmentation du taux de mots hors vocabulaire (MHV) lorsque que le document à traiter et le corpus d'entraînement sont plus éloignés dans le temps. En moyenne chaque mot hors vocabulaire est la cause de 1.5 à 2 erreurs.

Une des solutions consiste à réactualiser le corpus d'apprentissage et à réestimer le vocabulaire et le modèle de langage. Cette démarche est longue et nécessite une intervention manuelle importante. Une autre solution, développée dans cet article, est d'utiliser des méthodes automatiques d'adaptation [2], [4]. Pour modéliser le contenu lexical et linguistique de l'actualité, nous proposons d'utiliser des textes diffusés sur Internet qui sont contemporains des documents à transcrire. Il existe en effet de nombreux sites Internet d'actualités qui diffusent des articles et des dépêches écrites. Notre procédure d'adaptation est évaluée en mesurant la couverture lexicale et la perplexité du modèle de langage sur des textes de développement provenant également d'Internet. Un lexique et un ML adaptés ont été construits et testés pour chaque jour du mois de janvier 2002.

Dans la suite de cette article, nous décrivons tout d'abord la construction des corpus d'adaptation et de développement. Puis, après une description succinte du système de transcription automatique, nous décrivons et discutons les principaux résultats expérimentaux.

2. CORPUS

De nombreux sites Internet permettent de suivre l'actualité quasiment en temps réel. Certains contiennent une forme réduite de quotidiens nationaux, et d'autres diffusent des dépêches d'agences de presse. À partir d'un ensemble de sites judicieusement sélectionnés, on peut donc collecter quotidiennement des textes reflettant le contenu linguistique des actualités radio ou télévisées.

2.1. Textes du Web

Les pages HTML collectées chaque jour sur les sites choisis sont "nettoyées" automatiquement de manière spécifique à chaque source. L'objectif est de les convertir dans un format unique et d'en extraire les parties utiles : la date, les titres et le texte des articles. Le codage des caractères est uniformisé et le formatage HTML inutile est supprimé.

La normalisation des textes est une étape majeure dans la préparation d'un corpus d'entraînement [1]. L'une des opération de normalisation consiste à segmenter le texte en mots en effectuant un compromis entre la couverture lexicale et la capacité discriminante du ML. Contrairement à ce qui a été fait dans [3], nous ne pouvons plus conformer les textes à un vocabulaire défini, car un de nos objectifs est l'enrichissement du vocabulaire. Nous devons utiliser une normalisation autorisant une forte discrimination afin de repérer les mots nouveaux tout en restant robuste aux bruits des sources de textes. La première étape de la normalisation comprend le traitement des unités fréquentes, des séparateurs ambigus (" ' " et " - ") hors mots composés, et la réécriture des abréviations usuelles. La deuxième étape s'appuie sur le vocabulaire le plus récent pour déterminer la casse de la première lettre de chaque phrase et pour traiter les mots composés. Les nombres sont ensuite convertis en mots, et seules les phrases de plus de huit mots sont conservées afin de filtrer les énumérations

Source	*Moyenne*	*Min*	*Max*
D	5.6k	0.2k	10.9k
$A^{(1)}$	37.2k	11.7k	68.9k
$A^{(2)}$	1.2k	0k	4.6k
$A^{(3)}$	51.3k	1.3k	95.8k
$A^{(4)}$	34.9k	0k	59.1k
D_{14}	78.0k	67.6k	88.6k
A_{28}	3.5M	3.1M	3.9M

TAB. 1 – *Moyenne, minimum et maximum du nombre de mots par jour et par source sur les mois de décembre 2001 et janvier 2002.*

et les phrases structurelles.

2.2. Constitution des corpus

Cinq sites ont été sélectionnés pour leurs contenu et leur facilité de téléchargement. Le tableau 1 résume dans sa partie supérieure, le comportement de chaque source par la valeur moyenne, minimale et maximale du nombre de mots collectés par jour sur l'ensemble des mois de décembre 2001 et janvier 2002. Force est de constater que la quantité de données que l'on peut attendre de chaque site est très variable (par exemple certaines sources peuvent ne diffuser aucun document nouveau les dimanches ou jours fériés). Quatre sites ($A^{(1)}$ à $A^{(4)}$) constituent le corpus d'adaptation. Le cinquième site (D) délivre en un seul fichier, un résumé de l'actualité avec une moyenne de 6 000 mots par jour. Il consitue le corpus de développement et d'évaluation. Par la suite, nous notons $S(j)$ l'ensemble des textes du jour j venant du site S.

Pour constituer des ensembles de textes de taille suffisante et homogène, il est nécessaire de considérer plusieurs jours consécutifs. Pour chaque jour j du mois de janvier 2002, nous utilisons un corpus d'adaptation $A_k(j)$ contenant les textes des jours j-k à k :
$A_k(j) = \bigcup[A(j-k)\cdots A(j)]$ avec $A(j) = \bigcup_i A^{(i)}(j)$.
De même, le corpus de développement (et de test) pour le jour j comprend les textes $D(j)$ des jours j-k à j : $D_k(j)$.
Pour estimer un modèle trigramme, le corpus d'entraînement doit être de taille suffisante. Nous avons donc choisi k pour que le corpus d'adaptation comprenne au moins 3 millions de mots. Pour le corpus de test, une image précise de l'actualité est recherchée, mais nous désirons au moins 60k mots pour que nos mesures ne soient pas trop bruitées. En prenant en compte ces contraintes, nous avons fixé k à 28 pour les données d'adaptation et à 14 pour les données de test. La partie inférieure du tableau 1 montre qu'on obtient en moyenne 78 k mots pour $D_{14}(j)$ et 3.5 M mots pour $A_{28}(j)$ avec une variation acceptable.

3. SYSTÈME DE RÉFÉRENCE

Le système de transcription automatique de documents audiovisuel du LIMSI comporte deux traitements : le processus de segmentation et le système de reconnaissance [6]. Le processus de segmentation [5] permet de diviser le flux audio continu en segments acoustiques homogènes étiquetés en genre, en bande passante et en locuteurs.

Le système de reconnaissance de la parole utilise des modèles de Markov cachés pour la modélisation acoustique et des modèles n-grammes pour la modélisation du langage. Le décodage est effectué en trois passes : génération d'une hypothèse initiale, puis génération d'un graphe de mots, et génération de l'hypothèse finale. L'hypothèse initiale est utilisée pour adapter les modèles acoustiques pour l'étape suivante. Un ML trigramme est utilisé pour les deux premières passes. Pour l'hypothèse finale le calcul est effectué grâce à un ML 4-gramme et à des modèles acoustiques adaptés sur les hypothèses de l'étape 2.

Les modèles de langages de référence ont été obtenus par interpolation de modèles n-grammes avec *back-off* estimés sur différents type de données : 332 M mots provenant des journaux *Le Monde* et *Le Monde Diplomatique*, 63 M de mots de l'*Agence France Presse*, 22 M d'agence de service de presse et 0.75 M de mots correspondant aux transcriptions manuelles des données d'entraînement acoustique. Les coefficients d'interpolation ont été estimés avec l'algorithme EM en minimisant la perplexité sur un corpus de développement. Le ML 4-gramme contient 15 M de bigrammes, 15 M de trigrammes et 13 M de 4-grammes.

Le lexique contient 65533 mots et a une couverture lexicale de 98.8% sur un corpus de développement. Chaque entrée est décrite à l'aide d'une séquence d'éléments choisit parmi 33 phonèmes, plus 3 modèles pour les hésitations, les respirations et les silences. Les prononciations ont été obtenues à partir de règles graphèmes-phonèmes, et ont été vérifiées manuellement.

4. EXPÉRIMENTATION

Après un étude des performances du système de référence, nous proposons une méthode d'adaptation du vocabulaire et du modèle de langage. Le vocabulaire est évalué en mesurant le pourcentage de MHV dans $D_{14}(j)$. Une typologie des MHV est faite pour quantifier les améliorations de la couverture lexicale des entités nommées (les les noms propres, les acronymes et les sigles). Le modèle de langage est évalué en mesurant sa perplexité sur $D_{14}(j)$.

4.1. Performances du système de référence

Le pourcentage de MHV pour le vocabulaire référence sur $D_{14}(j)$ varie suivant les jours entre 2.3% et 1.9% avec une valeur moyenne de 2.15% (voir figure 1). Il est de 1.2% sur son corpus de développement (cf. section 3). Considérons l'ensemble du corpus d'évaluation pour le mois de janvier que nous notons $D_{tot} = \bigcup_{j=1}^{31} D(j)$ et sur lequel nous avons effectué manuellement la typologie suivante. Dans ce corpus, 34% des MHV n'apparaissent qu'une fois, 45% sont des entités nommées, 20% sont dus à des erreurs typographiques, 20% sont des noms communs (en grande majorité des formes fléchies de mots du vocabulaire) et 10% sont imputables à la normalisation des premières lettres (principalement des majuscules emphatiques). La normalisation des majuscules emphatiques n'a pas été jugée nécessaire car elle implique une intervention manuelle importante : le fait de savoir si une majuscule en début de mot relève d'une emphase ou traduit un nom propre nécessite une analyse syntaxique, voire sémantique [1]. La perplexité du modèle de référence sur $D_{14}(j)$ varie assez peu avec j (voir figure 1). Elle évolue de 120 à 127 et est de 122.5 si l'on considère D_{tot}. Ces résultats montrent une dégradation des performances par rapport à celles observées lors du développement du lexique et des ML. Ces dégradations sont dues en grande partie au fait

que le corpus d'entraînement n'est pas contemporain des données traitées.

4.2. Adaptation du vocabulaire

Nous cherchons à modéliser le contenu lexical de l'actualité du jour à partir des textes récents en adaptant le lexique référence. La sélection des mots pour un système de reconnaissance de la parole se fait généralement en prenant les N formes les plus fréquentes d'un corpus d'apprentissage. Ici nous disposons d'un lexique initial que nous souhaitons enrichir. Pour chaque jour, nous identifions les candidats à l'entrée dans le lexique en analysant la fréquence des mots hors vocabulaire sur le corpus d'adaptation.

Afin de filtrer les éventuelles erreurs provenant des sources ou de notre normalisation, un seuil est appliqué au nombre d'occurrence de chaque forme. Que ce soit pour les journaux radio et télédiffusés ou pour la presse papier, le contenu de l'actualité est la somme de thèmes d'actualité dont les durées de vie peuvent être considérées comme aléatoires. Pour un jour j, les journaux traitent de thèmes ponctuels qui n'apparaissent que ce jour, mais aussi de thèmes récurrents présents dans l'actualité depuis plusieurs jours. En prenant $A_1(j)$ comme corpus d'adaptation (c'est-à-dire uniquement les textes du jour j), la quantité de données est très réduite mais les thèmes ponctuels sont bien représentés. Par contre si l'on prend $A_{28}(j)$ (les textes sur les quatre semaines précédentes), les thèmes récurrents masquent les thèmes ponctuels. Nous notons $E_n(A_k(j))$, l'ensemble des mots apparaissant plus de n fois dans $A_k(j)$ et ne faisant pas partie du vocabulaire de référence. La valeur de n est choisie de manière à obtenir suffisamment de mots nouveaux tout en rejetant les erreurs typographiques. Pour réaliser ces compromis, l'ensemble des mots candidats à l'entrée dans le vocabulaire pour le jour j ($E_{in}(j)$) est obtenu ainsi:

$$E_{in}(j) = E_5(A_{28}(j)) \cup E_1(A_1(j))$$

c'est-à-dire en faisant l'union des MHV apparaissant au moins deux fois dans $A_1(j)$ et des MHV apparaissant plus de 5 fois dans $A_{28}(j)$. Le nombre de mots dans $E_{in}(j)$ est noté $N_{in}(j)$.

Comme nous désirons garder la taille du vocabulaire constante (65k mots), nous devons pour chaque candidat en l'entrée choisir un candidat en sortie. Pour cela le vocabulaire de référence est trié dans l'ordre des probabilités unigrammes, et les $N_{in}(j)$ mots les moins probables qui n'apparaissent pas dans $A_{28}(j)$ sont remplacés par les candidats à l'entrée. Nous imposons également que les mots retirés ne fasse pas partie des 30 000 les plus fréquents du vocabulaire de référence. On obtient ainsi pour chaque jour j un vocabulaire de 65 333 mots : $V(j)$.

La figure 1 permet de visualiser l'évolution du pourcentage de MHV dans $D_{14}(j)$ obtenu pour le vocabulaire de référence et le vocabulaire adapté $V(j)$. L'adaptation du vocabulaire permet un gain absolu qui varie entre 0.47% et 1% soit une diminution relative entre 23.5% et 44% avec une moyenne de 32%. Le taux de MHV pour le vocabulaire adapté est très proche du taux de 1.2% qui avait été obtenu par le vocabulaire de référence sur son corpus de développement. La part de mots retirés du vocabulaire et apparaissant dans $D_{14}(j)$ ne dépasse pas 0.02%. L'algorithme d'actualisation du vocabulaire permet l'ajout de 1775 mots par jour en moyenne. Cette quantité peut varier

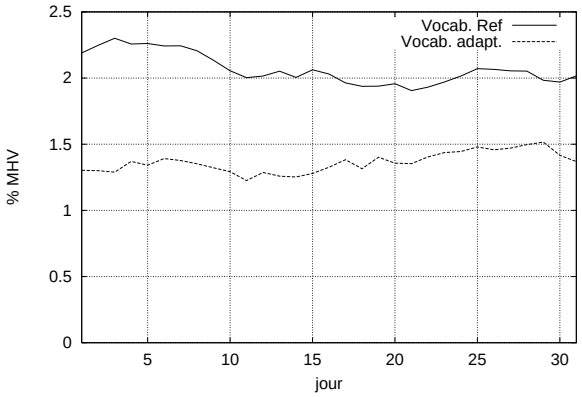

FIG. 1 – *% de MHV pour chaque jour j sur $D_{14}(j)$ avec le vocabulaire de référence puis le vocabulaire adapté.*

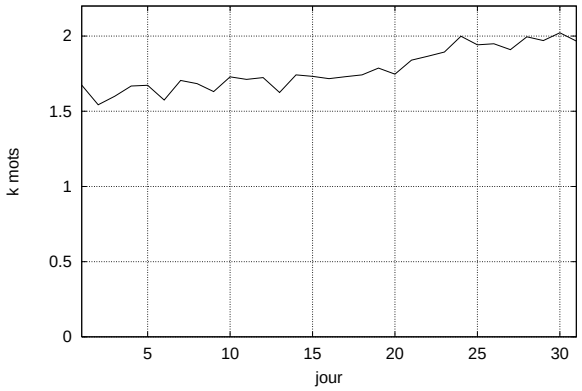

FIG. 2 – *Nombre de mots nouveaux pour chaque jour*

de 1500 à 2000 (voir figure 2), et elle n'est pas corrélée à la couverture lexicale de l'un ou de l'autre des vocabulaires. Sur l'ensemble des mots ajoutés au mois de janvier, on compte 5418 mots distincts. Seuls 323 mots (6%) sont présents sur l'ensemble du mois et 3134 mots n'apparaissent qu'une fois.

Si nous reprenons la typologie des MHV donnée en section 4.1, les entités nommées ne représentent plus que 35% des MHV, alors que la part de noms communs est en augmentation à 32%. Les erreurs typographiques ont été relativement peu incorporées dans le vocabulaire car elles représentent désormais 30% des MHV. La proportion de MHV due à la normalisation des premières lettres en majuscules est réduite à 3%. Majoritairement, ce sont les entités nommées et les mots avec majuscule emphatique qui ont été incorporés au vocabulaire.

4.3. Adaptation des ML

Le modèle de langage de référence (M_{ref}) et son lexique ont été construits à partir de textes datant des années 1987 à 1999. Il y a donc au minimum 2 ans d'écart avec l'actualité du mois de janvier 2002. Étant donné la taille du corpus d'apprentissage et le temps nécessaire à la construction d'un tel modèle, nous préférons ne pas réestimer les probabilités de M_{ref}. Pour incorporer à ce modèle les informations linguistiques présentes dans le corpus $A_{28}(j)$, nous construisons pour chaque jour un modèle $M_a(j)$ que l'on peut qualifier de modèle roulant [2]. Ce modèle utilise

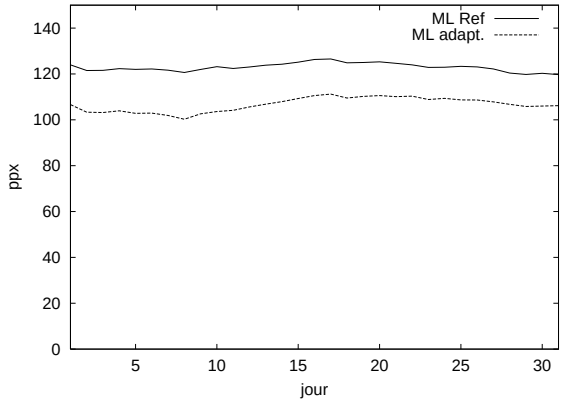

FIG. 3 – *Perplexité (ppx) calculée pour chaque jour j sur $D_{14}(j)$ avec le ML référence et le ML adapté du jour.*

le vocabulaire adapté $V(j)$. La perplexité de $M_a(j)$ varie entre 315 et 365 sur les données $D_{14}(j)$ avec une moyenne de 344. Ces perplexités sont très élevées mais ne laisse en rien présager le potentiel d'adaptation de ces modèles [3]. L'adaptation se fait par l'interpolation linéaire entre M_{ref} et le modèle du jour : Pour chaque mot w_i et chaque historique $h_k = w_{i-1}w_{i-2}$ (pour un modèle trigramme) nous avons :

$$P(w_i|h_k) = \lambda P_a(w_i|h_k) + (1-\lambda)P_{ref}(w_i|h_k)$$

où $P_a(w_i|h_k)$ est la probabilité assignée par le modèle $M_a(j)$, et $P_{ref}(w_i|h_k)$ est la probabilité assignée par le modèle de référence. Le coefficient d'interpolation λ peut être calculé de manière à minimiser la perplexité sur un corpus de développement via l'algorithme EM. Des tests effectués sur le mois de décembre 2001, ont montré que λ varie de 0.26 à 0.33. Pour les expériences qui suivent, λ a donc été fixé à 0.3.

La figure 3 montre l'évolution de la perplexité calculée sur $D_{14}(j)$ avec le ML référence et le ML interpolé au cours du mois de janvier. L'adaptation du ML permet une réduction relative de la perplexité allant de 11% à 16% (elle varie de 100 à 111). Sur l'ensemble du corpus D_{tot}, la perplexité est égale à 107.8, ce qui implique un gain relatif de 13%. Nous avons noté que la perplexité du ML adapté est plus variable que celle du ML de référence avec une variation de 10%.

5. CONCLUSION

Nous avons présenté une méthode d'adaptation entièrement automatique permettant à un lexique et à un modèle de langage anciens d'être actualisés. Cette méthode utilise des textes d'actualité disponibles quotidiennement sur des sites Internet.

Pour chaque jour du mois de janvier 2002, nous avons constitué un corpus d'adaptation et un corpus de développement grâce à une méthode de normalisation développée spécifiquement pour les textes du Web. Une analyse des MHV obtenus avec les différents vocabulaires a permis de montrer que les mots de l'actualité changent suffisamment rapidement pour justifier une adaptation quotidienne du vocabulaire. Cette adaptation entraîne l'ajout d'une faible quantité de mots (environs 2000) et permet de réduire le taux de MHV de 32% en moyenne le rapprochant ainsi de la valeur obtenue avec le vocabulaire de référence lors de son développement. Les mots sortis du vocabulaire de référence par l'algorithme d'adaptation lexicale, apparaissent peu dans les corpus de développement (un peu moins de 0.2%). Cet algorithme améliore la couverture lexicale particulièrement pour les entités nommées et les mots avec majuscule emphatique. Après interpolation avec le modèle de référence, le modèle construit avec le nouveau vocabulaire permet de réduire de façon significative la perplexité (entre 11% et 16%).

Les résultats présentés dans cette article doivent encore être validés par des résultats effectifs de reconnaissance de la parole. Si le lien entre mots hors vocabulaire et reconnaissance est direct, une forte diminution de la perplexité n'implique pas forcément une réduction équivalente du taux d'erreur de reconnaissance. De plus nous pensons phonétiser automatiquement les mots nouveaux. La phonétisation automatique pose quelques problèmes avec les noms propres étrangers qui représentent une part importante des mots qui sont ajoutés au vocabulaire.

REMERCIEMENTS

Les auteurs souhaitent remercier Gilles Adda pour avoir construit le lexique et le modèle de langage de référence ainsi que pour ces précieux conseils.

RÉFÉRENCES

[1] G. Adda and M. Adda-Decker. Normalisation de textes en français : une étude quantitative pour la reconnaissance de la parole. In *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, pages 289–296, Avignon, France, April 1997.

[2] Fiscus Jonathan G. Auzanne Cedric, Garofolo John S. and Fisher William M. Automatic language model adaptation for spoken document retrieval. In *SDR 2000*. TREC 9, 2000.

[3] C. Barras, A. Allauzen, and L. Lamel J.L. Gauvain. Transcribing audio-video archives. In *Proc. ICASSP*, 2002. to appear.

[4] Marcello Federico and Nicola Bertoldi. Broadcast news adaptation using contemporary texts. In *Proc. Eurospeech*, pages 239–242, Aalborg, Denmark, Sep 2001.

[5] J.L. Gauvain, L. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. In *Proc. ICSLP*, 1998.

[6] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 2002. to appear.