Construction automatique du vocabulaire d'un système de transcription

Alexandre Allauzen, Jean-Luc Gauvain*

Groupe Traitement du Langage Parlé (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex,
{allauzen,gauvain}@limsi.fr

ABSTRACT

This article investigates the problem of automaticaly building the vocabulary of a broadcast news transcription system. A specific evaluation framework is presented, and the original manual method is described. An algorithm is introduced which minimizes the Kullback-Leibler divergence between the target distribution and an interpolated distribution estimated on training corpora. Experiments show a relative reduction of out-of-vocabulary rate of 17% on the development set and 6% on the test set compared to the reference vocabulary.

1. INTRODUCTION

Le vocabulaire est la liste close des N mots pouvant être reconnus par un système de reconnaissance automatique de la parole (RAP). Dans les systèmes de transcription d'émissions télé-radiodiffusées, sa taille N est souvent fixée à 65 535 mots lorsque la langue le permet (comme c'est le cas pour l'anglais ou pour le français). Le choix du vocabulaire a un fort impact sur les performances du système de transcription automatique puisque tout mot absent du vocabulaire ne peut être reconnu, et que sa présence entraîne des erreurs dans son voisinage.

La plupart des méthodes de construction du vocabulaire sont manuelles et consistent à sélectionner les mots en appliquant des seuils sur leurs fréquences observées sur des corpus d'entraînement. Des considérations plus ou moins quantifiables interviennent dans le choix des seuils, telles que la taille ou la nature des corpus, l'époque qu'ils couvrent et la corrélation de contenu supposée avec la tâche [6, 2]. La procédure est alors déterminée de manière *ad hoc*, en fonction des ressources disponibles et de la tâche. Tous ces aspects entravent donc la reproductibilité de la démarche. Et s'il existe des méthodes d'adaptation du vocabulaire [3], l'étape première de construction du vocabulaire reste manuelle et laborieuse.

Dans cet article, il est proposé de formaliser plus avant la construction du vocabulaire afin de l'automatiser. Pour cela, un cadre d'évaluation est d'abord introduit. Il définit les objectifs d'un point de vue théorique, et décrit les corpus disponibles ainsi que les critères d'évaluation. Ensuite, la procédure manuelle ayant servit à la construction du vocabulaire de référence est présentée puis évaluée. Une typologie des mots hors vocabulaire (MHV) est alors effectuée. Enfin un algorithme de construction automatique est proposé puis expérimenté. Cet algorithme interpole linéairement les modèles unigrammes estimés sur des corpus d'entraînement, de manière à minimiser la divergence de Kullback-Leibler entre la distribution estimée sur le corpus de développement et la distribution interpolée.

*Ces travaux ont été effectués lors de ma thèse, en collaboration avec la Direction de la Recherche et de l'Expérimentation de l'Institut National de l'Audiovisuel (http://www.ina.fr)

2. CADRE D'ÉVALUATION

Le vocabulaire est construit de manière à obtenir la meilleure couverture lexicale sur la tâche qui est la transcription automatique de documents télé-radiodiffusés. Cependant, le contenu lexical de la tâche n'est pas accessible directement, et un corpus de développement Y est constitué à partir de transcriptions de journaux télévisés afin de le modéliser.

2.1. Cadre théorique

Un vocabulaire V est couramment évalué par la mesure duale de la couverture lexicale : le taux de MHV mesuré sur un texte ou un corpus donné. Appliqué au corpus de développement Y, il s'écrit :

$$T_{MHV}(V,Y) = \frac{\text{nombre d'occurrences dans } Y \notin V}{\text{nombre d'occurrences total dans } Y}$$

Le corpus de développement est représenté par sa distribution unigramme sur les mots p_Y. L'objectif est alors de choisir le vocabulaire V, parmi l'ensemble des vocabulaires possibles, qui minimise l'espérance du taux de MHV connaissant la distribution p_Y :

$$\begin{aligned} V &= \text{argmin}_v \, E[T_{MHV}(v,Y)|p_Y] \\ &= \text{argmin}_v \sum_w \delta_v(w) p_Y(w). \end{aligned}$$

Pour estimer l'espérance, la fonction δ_v est définie. Elle indique si un mot w est hors vocabulaire : $\delta_v(w) = 1$ si $w \notin v$, et $\delta_v(w) = 0$ sinon.

Une contrainte importante pour un vocabulaire est sa taille N. Si le vocabulaire ne contient qu'un seul mot, ce mot est choisi tel que $E[T_{MHV}(v,Y)|p_Y]$ soit minimum. Il s'agit donc du mot le plus probable selon p_Y. Le second mot sera le deuxième mot le plus probable. Ainsi, par récurrence, le vocabulaire de N mots qui minimise l'espérance du taux de MHV connaissant p_Y, est celui qui contient les N mots les plus probables selon p_Y.

Cependant il n'existe généralement pas assez de transcriptions pour constituer un corpus de développement permettant la construction d'un vocabulaire de $N = 65\,535$ mots. Dans notre cas, le corpus de développement permet au

Source	*Période*	*Total*
Transcriptions	1994-1999	1,6
Service de presse	1997,1998,2000	72,7
Agence de presse	1994-1996	66,8
Le Monde	1987-1998	257,4
Le Monde Diplo.	1990-1996	6,7

TAB. 1: Description des Données d'entraînement : chaque source est caractérisée par la période couverte, et le nombre total de mots (en million de mots).

mieux la sélection de 7 247 mots (voir paragraphe 2.2). Mais même si cela est possible, le danger est de construire un vocabulaire spécifique au corpus de développement qui est une représentation lacunaire de la tâche. Il est donc préférable que le corpus de développement serve uniquement pour calculer le taux de MHV, et que des corpus d'entraînement soient utilisés pour la sélection des mots.

2.2. Description des corpus

Le corpus de développement choisi contient 17 026 occurrences pour 7 247 formes et contient des transcriptions manuelles du journal télévisé *6 minutes* de M6 datant de l'année 2000. Ce corpus fut constitué lors de la construction du vocabulaire de référence (voir paragraphe 3) avec les transcriptions qui étaient alors les plus récentes. Afin d'étudier l'impact des différences d'époques, un corpus de test est choisi, qui est au moins de 2 ans plus récent que les textes d'entraînement ou de développement. Il est constitué à partir de transcriptions manuelles de journaux télévisés du mois de janvier 2002 et contient 77 541 occurrences.

Les textes d'entraînement disponibles se répartissent selon 5 sources. La première contient les transcriptions manuelles de journaux radio-télévisés provenant de 6 diffuseurs. La deuxième source nommée "Service de presse" contient des transcriptions d'extraits d'émission. Les trois dernières sources proviennent de la presse écrite : dépêches d'agences de presse, articles du journal *Le Monde* et des articles du journal *Le Monde Diplomatique*. La répartition de ces textes est donnée dans le tableau 1 en spécifiant pour chaque source la période couverte, et le nombre total de mots. Les transcriptions représentent une faible quantité de mots, justifiant ainsi l'utilisation d'autres sources palliatives. Tous les textes utilisés ont subit une étape de normalisation décrite dans [1].

2.3. Critères d'évaluation

Le premier critère d'évaluation du vocabulaire est le taux de MHV sur le corpus de développement. C'est ce critère qui a été manuellement optimiser lors de la construction du vocabulaire de référence. Ce critère caractérise l'aptitude de la méthode à générer un vocabulaire "pour le corpus de développement" sous certaines contraintes détaillées à la section 2.1. Afin d'étudier la pérennité des vocabulaires construits, le second critère choisi est le taux de MHV mesuré sur le corpus de test qui date de janvier 2002.

2.4. Le vocabulaire exhaustif

Si la taille du vocabulaire n'est pas limitée, on peut envisager de sélectionner toutes les formes apparaissant dans tous les corpus d'entraînement. Dans notre cas, ce vocabulaire exhaustif contient 765 442 formes distinctes. La couverture lexicale obtenue avec ce vocabulaire, que ce soit sur le corpus de développement ou sur le corpus de test, est bien entendu la meilleure possible pour l'ensemble des corpus d'entraînement donné. Ainsi, la borne inférieure du taux de MHV est de 0,25% sur le corpus de développement et de 0,35% sur le corpus de test. Avec ce vocabulaire, les MHV sont principalement des entités nommées qui apparaissent très ponctuellement dans l'actualité.

3. LE VOCABULAIRE DE RÉFÉRENCE

Le vocabulaire de référence a été construit en 2000, selon la méthode décrite dans les articles [2, 1], et à partir des textes décrits au paragraphe 2.2.

3.1. Construction du vocabulaire de référence

Construire le vocabulaire se résume à sélectionner les N formes lexicales les plus fréquentes provenant de plusieurs corpus d'apprentissage afin de maximiser la couverture lexicale sur un corpus de développement. Comme le montre le tableau 1, il est préférable de ne pas regrouper toutes les données dans un seul corpus. Sinon, étant donné les différences de taille, les transcriptions manuelles peuvent être noyées dans les millions de mots provenant des autres corpus.

Pour chaque corpus, un seuil sur la fréquence des mots élimine les mots les moins fréquents. Une liste de mots par corpus est alors obtenue, et le vocabulaire final résulte de la fusion de ces listes. Le vocabulaire est déterminé par le choix de ces seuils qui sont le résultat d'un compromis entre trois critères : le nombre de mots N que doit contenir le vocabulaire final, la couverture lexicale sur le corpus de développement et le filtrage de bruit éventuel. Puisqu'un nombre important de sources est disponible, il est impératif d'effectuer des regroupements, car il est difficile de construire manuellement un vocabulaire (choix des seuils) avec un nombre important de sources. Il a été choisi de regrouper les textes selon leur nature en ne tenant pas compte de leur date. Trois corpus ont été ainsi constitués : le premier regroupe tous les textes de la presse écrite (*Le Monde*, *Le Monde diplomatique*, Agence de presse), le deuxième les textes de services de presse et le troisième les transcriptions fines de journaux radio-télévisés.

Le vocabulaire de référence ainsi construit contient 65 333 formes. Son taux de MHV sur le corpus de développement est de 0,91%. Sur le corpus de test, les performances sont un peu dégradées avec un taux de MHV de 1,13%. En deux ans, la population des MHV a changé quasiment du tout au tout, et seules 2 formes hors vocabulaire sont communes aux deux corpus "tombola" et "coupe-feu". Les MHV sont en grande majorité des mots qui apparaissent une seule fois.

3.2. Typologie des mots hors vocabulaire

Les corpus de développement et de test ont été étiquetés en catégories morpho-syntaxiques, à l'aide du tagger de Brill[1] [4] adapté au français par l'INALF[2]. Le jeu d'étiquettes a été réduit après l'opération de marquage, en

[1] Disponible à http ://www.cs.jhu.edu/brill

[2] Disponible à http ://jupiter.inalf.cnrs.fr/WinBrill

supprimant l'information de nombre. La répartition des MHV par catégorie morpho-syntaxique est donnée dans le tableau 2. Plus de la moitié des MHV sont des noms propres et presque 15% sont des noms communs. Les MHV proviennent tous de catégories lexicales ouvertes, et les catégories closes (déterminants, préposition, ...) n'interviennent pas. Peu de différences sont observées dans la répartition des MHV entre le corpus de développement et le corpus de test, hormis un augmentation significative de la part des noms propres et des verbes conjugués.

	% des MHV	
Catégorie	Dev	Test
Nom Propre	50,3	55,4
Nom commun	14,2	14,3
Verbe conjugué	8,4	13,1
Adjectif	8,4	4,1
Verbe infinitif	5,2	3,1
Participe passé adjectif	4,5	4,3
Participe passé (être)	3,9	2,2
Participe passé (avoir)	1,9	1,4

TAB. 2: Répartition des catégories morpho-syntaxiques les plus fréquentes des MHV du vocabulaire de référence observés sur le corpus de développement et de test.

4. CONSTRUCTION AUTOMATIQUE DU VOCABULAIRE

L'objectif est de construire un vocabulaire de N mots à partir de K corpus d'entraînement notés $(X_1, X_2, \ldots, X_K)$, qui optimise la couverture lexicale sur le corpus de développement Y. Selon le cadre théorique présenté au paragraphe 2.1, la combinaison des corpus d'entraînement doit permettre d'approcher la distribution observée sur le corpus de développement. L'algorithme de construction automatique du vocabulaire nécessite donc de définir au préalable un opérateur de mélange, et de choisir une fonction coût.

4.1. Algorithme de construction automatique du vocabulaire

Le mélange des distributions d'entraînement est réalisé par interpolation linéaire. L'objectif est alors de calculer le jeu de coefficients d'interpolation

$$\Lambda = (\lambda_1, \lambda_2, \ldots \lambda_K), \text{avec } \sum_{i=1}^{K} \lambda_i = 1,$$

qui minimisent une fonction de coût entre la distribution estimée sur le corpus de développement p_Y et la distribution interpolée $\sum_{i=1}^{K} \lambda_i \, p_{X_i}$.

Si le corpus de développement est représentatif de la tâche, sa distribution est celle attendue dans les documents qui vont être traités. Les coefficients d'interpolation peuvent être calculés afin de minimiser le coût d'émettre comme hypothèse la distribution interpolée, sachant que la distribution réelle est p_Y. En théorie de l'information, la divergence de Kullback-Leibler quantifie le coût de cette hypothèse. Les coefficients d'interpolation sont déterminés de la manière suivante :

$$\Lambda^* = \underset{\Lambda}{\operatorname{argmin}} \sum_{w \in \mathcal{V}} p_Y(w) \log \left(\frac{p_Y(w)}{\sum_{i=1}^{K} \lambda_i \, p_{X_i}} \right)$$

Puisque les modèles utilisés sont des modèles unigrammes obtenus avec l'estimateur du maximum de vraissemblance, cette approche est équivalente à celle couramment rencontrée pour calculer les coefficients d'interpolation qui minimise la perplexité via l'algorithme E.M [5]. La construction automatique du vocabulaire s'effectue de la manière suivante :

1. Estimation à partir de chaque corpus d'entraînement X_i d'une distribution unigramme p_{X_i} pour $i = 1$ à K.
2. Application de l'algorithme E.M pour calculer les coefficients d'interpolation afin de minimiser la perplexité du corpus de développement.
3. Construction du vocabulaire à partir de la distribution interpolée, en sélectionnant les $N = 65\,333$[3] mots les plus fréquents

Dans chaque distribution, les mots du corpus de développement qui ne figurent pas dans le vocabulaire définit par X_i sont considérés comme des MHV et sont projetés sur la forme unique <UNK>. Sa probabilité est fixée à 10^{-99} de manière à pénaliser la présence de MHV. Les mots du corpus de développement qui n'appartiennent à aucun des corpus d'entraînement ne sont pas pris en compte. En ce qui concerne l'algorithme E.M, la condition d'arrêt a été fixée en terme de précision sur les coefficients d'interpolation. Cette précision dépend du nombre de corpus d'entraînement utilisés.

5. EXPÉRIMENTATIONS

L'algorithme décrit précédemment permet de manipuler autant de corpus que nécessaire. À partir des corpus décrits au paragraphe 2.2, plusieurs segmentations sont envisagées, et chacune est évaluée sur le corpus de développement (Dev 2000) et sur le corpus de test (Test 2002).

5.1. Segmentation des données d'entraînement

Les quatre segmentations les plus intéressantes qui ont été évaluées sont :
– la segmentation *originale* est celle utilisée pour la construction du corpus de référence,
– la segmentation par *année et source* sépare les données selon leur source et l'année,
– la segmentation *hybride* sépare les données par diffuseurs pour les transcriptions, et par année pour les autres sources,
– la segmentation par *année et type de source* sépare les textes selon le type de sources (presse écrite, service presse, et transcription) et par année.

5.2. Résultats

Le tableau 3 rassemble l'ensemble des résultats y compris ceux obtenus avec le vocabulaire de référence et avec le vocabulaire exhaustif. La première expérimentation vise à valider l'algorithme décrit à la section précédente avec les mêmes conditions que celles de la construction du vocabulaire de référence et avec la même segmentation des données d'entraînement. La méthode de construction automatique améliore nettement le vocabulaire à corpus identiques, que ce soit sur le corpus de développement avec

[3] Cette taille correspond à celle du vocabulaire de référence.

	Dev 2000	Test 2002
Vocabulaire	%MHV	%MHV
Standard	0,91	1,13
Exhaustif (765 442 mots)	0,25	0,35
Original	0,79	1,06
Année et source	0,79	1,05
Hybride	0,78	1,05
Année et type de source	0,76	1,05

TAB. 3: Évaluation des segmentations des corpus, en terme de taux de MHV mesurés sur le corpus de développement "Dev 2000" et sur le corpus de test "Test 2002"

un gain relatif de 12%, ou sur le corpus de test dans une moindre mesure, avec un gain relatif de 6%.

La segmentation par année est la segmentation la plus fine : pour chaque année de chaque source un modèle est construit, obtenant ainsi 30 modèles. Sur le corpus de développement, les taux de MHV observés avec ce vocabulaire est équivalent à celui observé avec le vocabulaire "original". L'examen des coefficients d'interpolation permet de quantifier la participation de chacun des corpus au vocabulaire et met en exergue l'impact de la date des corpus. Ainsi le seul corpus contemporain du corpus de développement (*Service de presse* de l'année 2000) est prépondérant puisque son coefficient d'interpolation est de 0,79, alors que le second corpus est affecté d'un coefficient 0,05.

Deux autres variantes de segmentation ont été expérimentées. La première consiste à regrouper les transcriptions non plus par année, mais par diffuseurs. Cette configuration est notée "hybride" dans le tableau 3 et améliore sensiblement le taux de MHV sur le corpus de développement. La deuxième segmentation référencée "Année et type de source" permet d'obtenir le meilleur taux de MHV. Elle regroupe tous les textes de la presse écrite par année, les extraits d'émissions par année également et les transcriptions par diffuseur comme précédemment.

5.3. Équivalence entre perplexité et couverture lexicale

Au paragraphe 4.1, l'hypothèse implicite est faite que minimiser la perplexité du corpus de développement calculée avec un modèle interpolé, implique la minimisation du taux de MHV du vocabulaire construit à partir de ce modèle. Afin de vérifier cette hypothèse, la figure 1 trace le nombre de MHV obtenu sur le corpus de développement en fonction des 200 premières itérations de l'algorithme E.M. La fonction est globalement monotone décroissante comme cela était supposé. Quelques irrégularités sont observées, mais leurs amplitudes sont de 1 mot à une exception près (de 2 mots). Elles sont par ailleurs excessivement rares (8 fois sur les 1 520 itérations). Ces irrégularités peuvent donc être imputées à la sélection des mots du vocabulaire par seuillage sur le rang lexical. Au delà de l'itération 200, la fonction est effectivement monotone décroissante.

CONCLUSION

Nous avons exploré dans cet article la problématique de la construction du vocabulaire d'un système de reconnaissance automatique de la parole. Un cadre d'évaluation

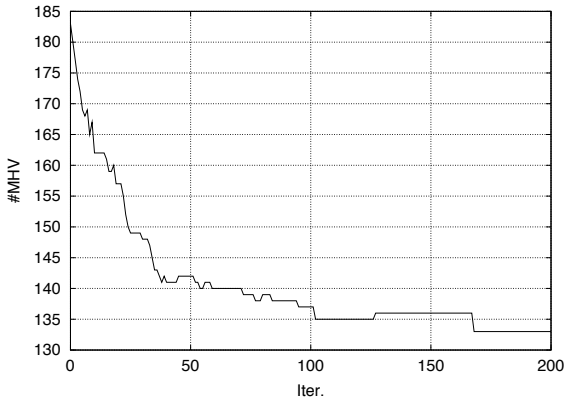

FIG. 1: Nombre de MHV en fonction des itérations de l'algorithme E.M.pour la segmentation *Année et type de source* (24 modèles sont interpolés)

spécifique a été proposé permettant de définir les attentes théoriques et pratiques d'un tel objectif. Le processus manuel de construction du vocabulaire a été décomposé afin de mettre en place un algorithme de construction automatique du vocabulaire. Cet algorithme consiste en l'interpolation d'un ensemble de modèles unigrammes estimés sur les données d'entraînement, les coefficients d'interpolation sont calculés grâce à l'algorithme E.M. afin de minimiser la perplexité du corpus de développement. Différentes segmentations des données d'entraînement ont été expérimentées, et des gains relatifs de 17% sur le corpus de développement et 6% sur le corpus de test sont obtenus. Cet algorithme repose sur l'hypothèse que minimiser la perplexité du corpus implique une amélioration de la couverture lexicale. Cette hypothèse est vérifiée expérimentalement, validant ainsi cette approche. De plus, les résultats obtenus avec le vocabulaire exhaustif montre que l'utilisation des ressources peut être encore améliorée.

RÉFÉRENCES

[1] G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel. Text normalization and speech recognition in French. In *Proc. Eurospeech*, volume 5, pages 2711–2714, Rhodes, September 1997.

[2] M. Adda-Decker, G. Adda, J.L Gauvain, and L. Lamel. Large vocabulary speech recognition in french. In *Proc. ICASSP*, Phoenix, March 1999.

[3] A. Allauzen and J.L Gauvain. Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés. *Traitement Automatique des langues*, 44(1) : 11–31, 2003.

[4] E. Brill. Some advances in rule based part-of-speech tagging. In AAAI, editor, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722–727, Seattle, WA, 1994.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1) :1–38, november 1977.

[6] R. Rosenfeld. Optimizing lexical and n-gram coverage via judicious use of linguistic data. In *Proceedings of Eurospeech 95*, pages 1763–1766, 1995.