Proposition d'une nouvelle méthodologie pour la sélection automatique du vocabulaire d'un système de reconnaissance automatique de la parole

Brigitte Bigi

CLIPS/IMAG Lab. CNRS,
BP 53, 38041 Grenoble cedex 9, France
Tél. : ++33 (0)4 76 51 45 26 - Fax : ++33 (0)4 76 63 55 52
Mél : {brigitte.bigi}@imag.fr

ABSTRACT

The vocabulary of an Automatic Speech Recognition (ASR) system is a significant factor in determining its performance. The goal of vocabulary selection is to construct a vocabulary with exactly those words that are the most likely to appear in the test data. This paper proposes a new measure to evaluate the quality of a vocabulary regarding a domain-specific ASR application. This Q_α-measure is based on the trade off between the target lexical coverage and vocabulary size. Experiments were carried out on French Broadcast News Transcriptions using the Q_α-measure compared to the state-of-the-art method. Results of these two methods favor systematically the proposed methodology.

INTRODUCTION

Dans le cadre d'une amélioration des performances des systèmes de Reconnaissance Automatique de la Parole (RAP), nos travaux visent à développer une méthodologie pour sélectionner efficacement le vocabulaire du système. La sélection du vocabulaire est un processus appliqué à un ou plusieurs corpus de différentes sources, origines ou périodes, dont le résultat est la définition d'une liste de mots. Le but est de déterminer la combinaison optimale des vocabulaires produits par les différents corpus.

La couverture lexicale indique le taux des mots du vocabulaire présents dans le corpus de test. Le problème de la construction d'un vocabulaire consiste à obtenir la meilleure couverture lexicale, sur un corpus de développement. Les méthodes manuelles de construction du vocabulaire consistent à sélectionner les K mots les plus fréquents observés sur des corpus d'apprentissage. Les valeurs de K sont définies manuellement en fonction de la taille, la période ou la nature des corpus disponibles.

Une méthode automatique qui permet d'ordonner les mots à partir de plusieurs corpus a été proposée dans [4] puis dans [1]. Le principe consiste en une interpolation linéaire des modèles unigrammes estimés sur des corpus d'apprentissage, de manière à minimiser la perplexité du modèle interpolé sur le corpus de développement. Le vocabulaire choisi contient les K mots les plus probables, K étant fixé manuellement. La méthode présentée dans cet article repose sur la proposition d'une nouvelle mesure de qualité qui intègre la couverture lexicale et une notion de pertinence relative à la taille du vocabulaire. Cette mesure s'intègre dans une démarche de recherche du meilleur rapport entre le taux de mots hors-vocabulaire et la taille du vocabulaire.

En section 1, nous montrons une étude qui met en avant la problématique relative à la sélection du vocabulaire et argumentons sur la nécessité de développer une méthode automatique de sélection. La méthodologie proposée est décrite en section 2. Elle repose sur la proposition d'une nouvelle mesure de qualité d'un vocabulaire. Le principe général de cette approche consiste à utiliser les corpus disponibles pour créer un ensemble de vocabulaires possibles et d'utiliser la mesure Q_α pour les comparer. La section 3 expérimente la méthodologie de sélection sur les données du projet ESTER dont le corpus se compose du journal "Le Monde", de transcriptions d'émissions radiophoniques, auxquelles nous avons ajouté des pages téléchargées sur des sites de radios et journaux sur l'Internet. Dans la dernière section, les résultats obtenus sont confrontés à la méthode automatique couramment utilisée dans le domaine [4, 1]. Le gain obtenu par la méthode proposée concerne la réduction systématique du taux de mots hors-vocabulaire sur les corpus de développement et de test.

1. PROBLÉMATIQUE

1.1. État de l'art

Le problème de la sélection du vocabulaire est notamment abordé dans [3]. L'auteur montre que la taille du vocabulaire d'un système de RAP a deux effets. D'une part, le taux de mots hors vocabulaire (MHV) est réduit, aidant à faire moins d'erreurs de substitutions dues aux mots inconnus. D'un autre côté, les entrées lexicales ajoutées augmentent la confusion acoustique sur les mots, induisant de nouvelles erreurs de reconnaissance. Il conclut que la taille optimale du vocabulaire dépend de la tâche pour laquelle le système est dédié et du système lui-même. Cependant, il ne propose pas de solution pour obtenir ce vocabulaire optimal.

Dans [4], trois méthodes sont proposées pour obtenir une liste de mots triés par priorité décroissante, à partir de plusieurs corpus. La meilleure des trois méthodes repose sur la maximisation de la vraissemblance de l'estimation des comptes, à partir de m corpus. Le but est de trouver les λ_m coefficients de l'interpolation linéaire des m modèles unigrammes. De la même manière, [1] propose de calculer un jeu de coefficients d'interpolation des unigrammes des différents corpus. Ils vérifient l'hypothèse 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. Dans cette méthode, il reste à l'utilisateur à choisir la valeur de K, taille du vocabulaire final.

1.2. Description des corpus

Les données sur lesquelles nous travaillons dans cet article proviennent de la campagne ESTER qui vise à l'évaluation des performances des systèmes de transcription d'émissions radiophoniques. La campagne est organisée dans le cadre du projet EVALDA sous l'égide scientifique de l'Association Francophone de la Communication Parlée avec le concours du Centre d'Expertise Parisien de la Délégation Générale de l'Armement et de ELDA (Evaluations and Language resources Distribution Agency).

TAB. 1: Description des corpus d'apprentissage

Source	Période	$\|W\|$	$\|V\|$	MHV
Le Monde	1987-2003	413M	855K	0,21
Transcriptions	1998-2000,2003	1M	34K	2,26
Web	2003-2004	2,5M	52K	3,16

Deux corpus écrits ont été utilisés pour la campagne ESTER Phase 2. Le premier est un corpus audio manuellement transcrit. Ces transcriptions proviennent principalement de France-Inter, France-Info, Radio France International et Radio Télévision Marocaine. Le second est un corpus de textes du journal "Le Monde". Le corpus des transcriptions a été divisé en 3 parties pour l'apprentissage, le développement et les tests. Le corpus de développement concerne l'année 2003 ; il contient 97K occurrences pour un vocabulaire de 9800 mots. Le corpus de test concerne les mois d'octobre et décembre 2004 ; il contient 119K occurrences pour un vocabulaire de 11317 mots. Enfin, nous avons ajouté un corpus web qui provient de l'aspiration quotidienne de données ciblées du web durant la période de juin 2003 jusqu'à avril 2004 (de 20 à 200 pages web par jour, de radios et journaux). Le tableau 1 indique la période concernée par chacun des corpus d'apprentissage, le nombre total de mots $|W|$, la taille du vocabulaire complet $|V|$, et le pourcentage de mots hors-vocabulaire sur le corpus de développement.

1.3. Couverture lexicale : discussion

Sélectionner un vocabulaire est une tâche difficile dans la mesure où elle présuppose des besoins de l'application. Dans la plupart des cas, le vocabulaire "complet" est impossible à atteindre, ou alors, il implique que celuici soit d'une très grande taille, bien supérieure à la limite imposée par le système auquel il est dédié. Notons $V = \{w_1, w_2, \cdots, w_K\}$ l'ensemble des K mots d'un vocabulaire V ; notons également $c(w, C)$ le nombre d'occurrences du mot w dans le corpus C. La couverture lexicale L d'un vocabulaire V sur un corpus C s'écrit :

$$L(V, C) = \frac{\sum_{w \in V} c(w, C)}{\sum_{w \in C} c(w, C)}$$

Plus le vocabulaire est grand, plus la couverture lexicale est élevée. On utilise de façon identique le pourcentage de MHV, calculé tel que : $MHV = 100 - (L(V, C) \times 100)$.

La figure 1 indique le taux de MHV du corpus du journal "Le Monde", divisé par années, sur un corpus de transcriptions de 2003. Elle montre la pertinence d'avoir des données récentes.

La figure 2 indique le taux de MHV du corpus du journal "Le Monde", divisé par années, sur un corpus de transcriptions de 2003, en regroupant les années. Elle montre le rôle négligeable de la quantité de données : une seule année suffit pour obtenir une bonne couverture lexicale, et l'ajout de données plus anciennes n'apporte rien, voire dégrade la qualité du vocabulaire.

La figure 3 indique la couverture lexicale des corpus de transcription, "Le Monde" et du Web. Elle met en exergue l'importance d'avoir des données de la même origine que celles de l'application visée.

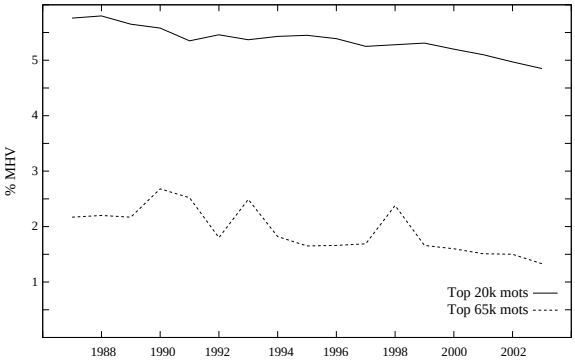

FIG. 1: MHV des vocabulaires du journal "Le Monde", selon l'année du corpus

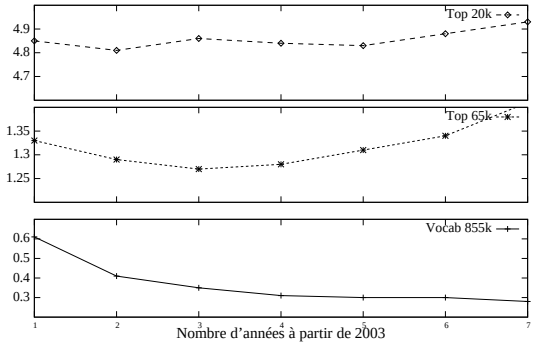

FIG. 2: MHV des vocabulaires du journal "Le Monde", en augmentant les données chaque année

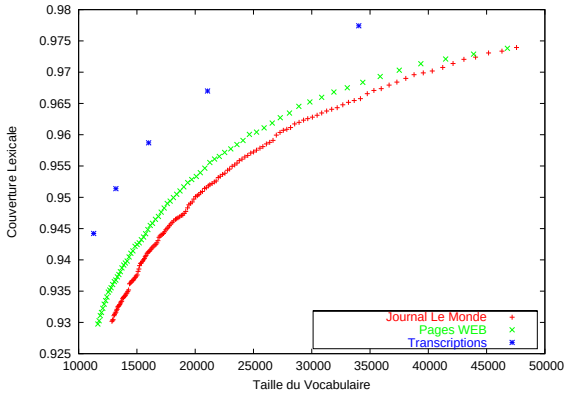

FIG. 3: Relation entre couverture lexicale, taille du vocabulaire et origine du corpus

2. MÉTHODE DE SÉLECTION AUTOMATIQUE

2.1. Production de vocabulaires candidats

La sélection d'un vocabulaire s'appuie sur l'utilisation de la fréquence ; les mots sont triés par ordre décroissant de leur nombre d'apparitions et le choix se porte sur les K mots les plus fréquents. Cette méthode, bien que très largement utilisée, pose un problème de coupe "brute" car des mots peuvent avoir la même fréquence autour de la valeur de K. C'est classiquement le rang des mots dans l'ordre alphabétique qui détermine les mots choisis. Nous proposons que les vocabulaires candidats se limitent à ceux qui ont des mots de fréquence supérieure à N. Par exemple, le vocabulaire du corpus des transcriptions contient 34037 mots dont 21059 ont plus d'une occurrence ; ainsi on choisira soit un vocabulaire de 34037 mots, soit de 21059 mots, mais on ne choisira pas une taille intermédiaire car les 12978 mots de différence ont la même fréquence (égale à 1). Pour un corpus donné, on dispose de plusieurs vocabulaires candidats, en faisant varier N.

On peut également segmenter les grands corpus selon des périodes de temps, ou de thèmes, selon l'application à laquelle le système est dédié. De nombreuses possibilités existent lorsque l'on dispose de corpus de tailles, origines et périodes différentes. Dans ce cas, il est souhaitable de segmenter au maximum les corpus.

Tous ces vocabulaires candidats, obtenus directement à partir des corpus, ou sous-corpus, peuvent être combinés (union ou intersection), afin d'obtenir de nouveaux vocabulaires candidats. Il reste alors à définir une mesure de qualité pour les comparer et décider du vocabulaire final de l'application.

2.2. Sélection par une mesure de qualité

La définition de la mesure proposée nécessite l'introduction des notations suivantes :
- M_c est la masse des mots communs, i.e. la somme des occurrences dans le corpus $\mathcal{C}$, de tous les mots présents à la fois dans le vocabulaire V et dans le corpus $\mathcal{C}$;
- M_f est la masse des mots exclus, i.e. la somme des occurrences dans le corpus $\mathcal{C}$, de tous les mots hors du vocabulaire V qui sont présents dans le corpus $\mathcal{C}$;
- M_i est la masse des mots ignorés, i.e. le nombre de mots présents dans le vocabulaire mais absents du corpus.

Selon cette notation $M_c + M_f$ représente la somme de toutes les occurrences de tous les mots de $\mathcal{C}$. La couverture lexicale L d'un vocabulaire V sur un corpus $\mathcal{C}$ se réécrit comme suit :

$$L(V, \mathcal{C}) = \frac{M_c}{M_c + M_f}$$

Dans cet article, nous introduisons une "mesure de pertinence du vocabulaire" qui tient compte des mots du vocabulaire qui ne sont pas utiles. Cette pertinence, notée R pour *relevance* en anglais, s'évalue telle que :

$$\mathcal{R}(V, \mathcal{C}) = \frac{M_c}{M_c + M_i}$$

La mesure Q_α combine la couverture lexicale L et la pertinence R, suivant le même principe que la F_α-mesure, largement utilisée en recherche d'information, par exemple, pour combiner le rappel et la précision des systèmes. Q_α permet la sélection des V_α vocabulaires dont la taille augmente au fur et à mesure qu'α croît. Q_α se calcule comme suit :

$$Q_\alpha = \frac{e^{\frac{\alpha+1}{2}} \times \mathcal{R}(V, \mathcal{C}) \times L(V, \mathcal{C})}{\left(e^{\frac{\alpha+1}{2}} \times \mathcal{R}(V, \mathcal{C})\right) + L(V, \mathcal{C})}$$

3. EXPÉRIMENTATION

Cette section détaille l'utilisation de la méthodologie proposée, sur les données décrites en section 1.2.

3.1. Vocabulaires candidats

La première étape de la méthodologie consiste à définir un ensemble de vocabulaires candidats. Pour cette expérimentation, nous avons fait les choix suivants :
- "Le Monde", avec N variant de 50 à 1800 par pas de 5 (soit 351 vocabulaires possibles) ;
- Transcriptions, avec N variant de 0 à 8 par pas de 1 (soit 9 vocabulaires possibles) ;
- Web, avec N variant de 0 à 50 par pas de 1 (soit 51 vocabulaires possibles).

ainsi que toutes les unions de 2 ou 3 de ces vocabulaires.

3.2. Sélection par la mesure Q_α

La table 2 montre les vocabulaires choisis par la mesure Q_α parmi les milliers de vocabulaires candidats. Nous avons fait varier α de 1 à 9, ne reste à l'utilisateur qu'à choisir celui dont la taille est la plus adaptée à l'application visée (ou au système qui l'utilise). La première colonne indique la mesure utilisée, les colonnes suivantes indiquent le détail des vocabulaires utilisés pour créer le vocabulaire choisi, dont la taille est indiquée dans la dernière colonne.

3.3. Validation

La validation consiste à comparer les taux de MHV des vocabulaires obtenus par notre méthode automatique avec celle proposée dans [4, 1]. Dans [1], il est prouvé que la méthode manuelle donne de moins bons résultats que la méthode automatique. Nous comparerons donc nos résultats directement à celle-ci. Chacun des corpus $\mathcal{C}_i$ est utilisé pour apprendre une distribution de probabilités $P(w, \mathcal{C}_i)$. Le problème consiste à trouver les λ coefficients de l'interpolation linéaire entre ces unigrammes :

$$P(w, \mathcal{C}_1, \cdots, \mathcal{C}_n) = \sum_{i=1}^{n} \lambda_i P(w, \mathcal{C}_i)$$

où $\sum_{i=1}^{n} \lambda_i = 1$. Les meilleures valeurs de λ optimisent la perplexité estimée sur le corpus de développement des transcriptions. Les mots obtenant les K meilleures probabilités $P(w, \mathcal{C}_1, \cdots, \mathcal{C}_n)$ sont sélectionnés pour le vocabulaire final. Pour notre expérimentation, l'interpolation optimale sur le corpus de développement est la suivante :

$$\begin{aligned} P(w, ester) = \quad & 0,758 \times P(w, transcriptions) + \\ & 0,139 \times P(w, lemonde) + \\ & 0,103 \times P(w, web) \end{aligned}$$

Les résultats sont présentés dans la table 3, pour le corpus de développement, et dans la table 4 pour le corpus de test. Dans ces deux tables, il est intéressant de noter que pour les 9 vocabulaires proposés par la méthode Q_α, le

TAB. 2: Description des vocabulaires proposées par la méthodologie

Mesure	Transcriptions		Le Monde		Web		Union
	N	$\|V\|$	N	$\|V\|$	N	$\|V\|$	$\|V\|$
Q_1	6	8927	1735	13416	49	4247	15055
Q_2	3	13196	1725	13478	49	4247	17200
Q_3	2	16006	1250	16668			20944
Q_4	1	21059	735	23300			28898
Q_5	1	21059	415	32662			36544
Q_6	1	21059	330	37217			40519
Q_7	0	34037	230	45483			54220
Q_8	0	34037	140	61210			67963
Q_9	0	34037	75	84140			89139

taux de MHV est inférieur à celui de la méthode utilisant l'interpolation linéaire. Même si le gain est parfois négligeable, il n'en est pas moins significatif car systématique (c'est-à-dire pour toutes les valeurs de α).

TAB. 3: Comparaison des taux de MHV (développement)

	% MHV	
$K = \|V\|$	Q_{alpha}	$P(w, ester)$
15055	4,07	4,04
17200	3,51	3,60
20944	2,89	2,95
28898	2,07	2,16
36544	1,59	1,90
40519	1,41	1,59
54220	1,06	1,13
67963	0,88	0,90
89139	0,68	0,71

TAB. 4: Comparaison des taux de MHV (test)

	% MHV	
$K = \|V\|$	Q_{alpha}	$P(w, ester)$
15055	4,82	4,98
17200	4,37	4,45
20944	3,72	3,75
28898	2,70	2,83
36544	2,13	2,51
40519	1,92	2,16
54220	1,47	1,45
67963	1,18	1,21
89139	0,92	1,00

Pour des raisons de clarté de la présentation de cette section, nous avons choisi de ne présenter qu'une seule expérience effectuée sur les trois corpus dont nous disposons. De nombreuses autres expériences ont été menées, notamment en divisant le corpus "Le Monde" par années, ou en regroupant les années les plus récentes. Dans tous les cas, les résultats sont en faveur de la méthode Q_α. A titre d'exemple, le vocabulaire issu de notre participation à la phase 2 du projet ESTER était composé de 21010 mots, créé comme suit :

$$(\text{"}LeMonde\text{"}, 2003, N = 30$$
$$\wedge\ \text{"}LeMonde\text{"}, 2001 - 2003, N = 241)$$
$$\vee\ Transcriptions, N = 2$$

Le taux de MHV sur le corpus de développement est de 2,84 %, contre 2,94 % pour la méthode à base d'interpolation linéaire. De même le taux de MHV sur le corpus de test est de 3,66 %, contre 3,74 % pour l'autre méthode.

TAB. 5: *Description du corpus anglais*

	$\|W\|$	$\|V\|$	MHV
Meetings - Test	16,5K	1943	-
Meetings - Train	1,0M	14669	3,2
Fisher - Train	5,3M	33110	3,3
Broadcast News - Train	131M	229535	2,2

Nous avons également conduit la même expérience sur des données en langue anglaise pour une application de transcriptions de réunions (table 5). Dans cette expérience, on observe un gain pour chacune des valeurs α, de 1 à 9. La mesure Q_6 propose un vocabulaire de 10124 mots avec 3,40% de MHV, tandis que l'interpolation linéaire obtient 3,50% de MHV pour un vocabulaire de même taille.

CONCLUSION

Cet article a proposé une méthode qui peut déterminer entièrement automatiquement le vocabulaire d'un système de RAP. Contrairement à la méthode à base d'interpolation linéaire, la méthode proposée ne nécessite pas de fixer *a priori* la taille du vocabulaire final. De plus, les résultats montrent une réduction systématique des taux de MHV des vocabulaires choisis par Q_α.

RÉFÉRENCES

[1] A. Allauzen and J-L. Gauvain. Construction automatique du vocabulaire d'un système de transcription. In *XXV Journees d'Etudes sur la Parole*, Fès (Maroc), 2004.

[2] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of Interspeech 2005*, 2005.

[3] R. Rosenfeld. Optimizing lexical and n-gram coverage via judicious use of linguistic data. In *Proceedings of Eurospeech 1995*, 1995.

[4] A. Venkataraman and W. Wang. Techniques for effective vocabulary selection. In *Proceedings of Eurospeech 2003*, 2003.