Enrichissement dynamique du vocabulaire à partir du Web

*Stanislas Oger, Georges Linarès, Frédéric Béchet, Pascal Nocéra**

LIA - Université d'Avignon, BP1228 84911 Avignon Cedex 09 - France
{*stanislas.oger,georges.linares,frederic.bechet,pascal.nocera*}*@univ-avignon.fr*

ABSTRACT

Most of the Web-based methods for lexicon augmenting consist in capturing global semantic features of the targeted domain in order to collect relevant documents from the Web. We suggest that the local context of the out-of-vocabulary words contains relevant information on the OOV words. With this information, we propose to use the Web to build locally-augmented lexicons which are used in a final local decoding pass. We first demonstrate the relevance of the Web for the OOV word retrieval. Then, different methods are proposed to retrieve the hypothesis words. Finally we present the integration of new words in the transcription process based on part-of-speech models. This technique allows to recover 7.6% of the significant OOV words and the accuracy of the system is slightly improved.

Keywords: Lexical modeling, Speech recognition, Information retrieval, Natural languages

1. Introduction

Malgré la grande quantité de données utilisées pour l'entraînement des modèles de langage, le problème des mots hors-vocabulaires (HV) est toujours difficile à traiter en reconnaissant automatique de la parole (RAP) continue grand vocabulaire.

L'augmentation de la taille des lexiques pour améliorer la couverture lexicale tend à augmenter considérablement les ressources nécessaires aux systèmes de RAP. De plus, le monde réel est une source inépuisable de nouveaux mots qui ne peuvent pas être tous répertoriés dans un lexique fermé.

De récentes recherches se sont intéressées à l'utilisation de sources textuelles variées lors de l'adaptation lexicale. Certains de ces travaux traitent de la recherche *a posteriori* de nouveaux mots dans de grandes bases de données [9], mais ce type d'approches statiques ne convient pas pour traiter des documents récents, où les sujets abordés et les entités nommées inconnus sont fréquents.

Cependant, le Web est une source de données linguistiques quasi-infinie et constamment mise à jour. Nous émettons donc l'hypothèse que toutes les séquences de mots possibles se trouvent sur le Web. Cette idée a déjà été explorée dans le domaine de la RAP grand vocabulaire. Les méthodes proposées consistent généralement à collecter une grande quantité de documents relativement proches du contexte sémantique et linguistique du document audio [5][1][8][3]. Malheureusement, les données issues du Web sont mal structurées et les techniques utilisant de grand corpus bien ciblés sont plus performantes que les modèles de langages basés sur le Web [6].

Du point de vue de l'enrichissement lexical, la recherche de mots HV rencontre deux principales difficultés : comment les mots manquants peuvent être automatiquement retrouvés sur le Web et comment ces mots peuvent être intégrés dans le processus de RAP. Habituellement, pour résoudre ce problème, des paramètres de la sémantique globale du document sont capturés puis utilisés pour collecter des documents pertinents. Les documents ainsi extraits sont utilisés pour la modélisation du langage.

Dans cet article, nous supposons que le contexte local contient des informations caractéristiques du mot HV et que ces informations peuvent être utilisées pour retrouver le mot manquant dans la collection de documents présents sur le Web. Nous proposons ici plusieurs méthodes pour collecter les mots HV manquants en se basant sur l'hypothèse précédente.

La section suivante présente le cadre expérimental dans lequel nous avons travaillé. Ensuite, nous évaluons l'hypothèse selon laquelle le Web est pertinente pour la recherche des mots HV et nous proposons des méthodes de recherche basées sur les patrons de mots. Ces méthodes sont ensuite évaluées sur la référence et sur les sorties du moteur de RAP. La troisième partie décrit le problème de l'intégration des lexiques localement augmentés dans le processus de RAP. Pour réaliser cette intégration, nous proposons une méthode basée sur un modèle de langage gérant les catégories morphosyntaxiques des mots. Finalement, nous concluons sur l'intérêt des méthodes locales basées sur le Web pour la correction *a posteriori* des transcriptions automatiques.

2. Contexte expérimental

Notre approche générale consiste à corriger *a posteriori* la transcription automatique réalisée par un système de RAP. Nous utilisons ici le système *broadcast news* du LIA, SPEERAL [7]. Ce système utilise un décodeur utilisant l'algorithme A* sur des modèles de

*Ces recherches sont financées en partie par l'Agence Nationale de la Recherche, projet SIGMUND ANR-05-RIAM-0903.

Markov à états cachés pour la modélisation linguistique. Les modèles de langage sont des tri-grammes classiques estimés sur environ 200M mots issus du journal Français Le Monde et du corpus d'environ 1M mots de *broadcast news* fourni lors de la campagne ESTER [4]. Le lexique utilisé comporte les 65000 mots les plus fréquents de ces corpus.

Les expériences sont basées sur l'environnement et les outils de la campagne d'évaluation ESTER. Le moteur de recherche Google est utilisé pour accéder aux données Web [1]. Tous les résultats ont été obtenus sur environ 6 heures de *broadcast news* en français issues du corpus de test de la campagne ESTER 2005. Le taux d'erreur mot (WER) de ce corpus avec le système décrit précédemment est de 25.5% après la première passe (sans adaptation acoustique en aveugle). Il y a 645 mots HV sur les 62024 mots de ces 6 heures de parole, le taux de mots HV est d'environ 1.03%. 73% des mots HV sont des entités nommées (EN), 24% sont des termes techniques spécifiques au domaine, et les 3% restants sont des formes verbales peu fréquentes ou des mots mal orthographiés dans la référence. Il est important de noter que les EN et les termes techniques sont importants pour la compréhension du document et représentent 97% des mots HV.

3. Apprentissage des nouveaux mots

Nous vérifierons tout d'abord l'hypothèse selon laquelle le Web peut être considéré comme une source infinie de mots et qu'il est pertinente de l'utiliser pour la tâche de recherche des mots HV. Nous étudierons ensuite l'impact des erreurs de transcription automatique sur les méthodes de recherche de mots HV.

3.1. Le web, une source infinie de mots

Dans le but de retrouver les mots HV en utilisant le contexte local et en émettant l'hypothèse que le Web contient toutes les séquences de mots, nous le considérerons comme un modèle n-gramme infini qui contient tous les n-grammes possibles, y compris ceux qui contiennent le mot recherché, w_t.

Afin d'évaluer cette hypothèse, nous mesurons ici le taux de présence sur le Web des n-grammes qui contiennent le mot HV recherché, w_t. Le contexte utilisé est celui de la référence afin de s'affranchir des erreur de RAP. Des requêtes du type $w_{t-n-1}...w_{t-1}w_t$, où $w_{t-n-1}...w_{t-1}$ est l'historique de w_t, sont soumises à Google afin de vérifier que le n-gramme existe sur le web. Les mots composés HV sont considérés comme un seul mot. Les résultats présentés dans le tableau 1 montrent que la majeure partie des 2-grammes contenant w_t se trouvent sur le Web et le rappel décroît quand n augmente. Ces résultats indiquent que le Web a un potentiel intéressant dans la tâche de récupération des mots HV, mais à condition de formuler les requêtes de manière pertinente.

3.2. Collecte des lexiques augmentés

Nous proposons ici plusieurs méthodes pour la collecte des mots HV et évaluons leurs performances en utilisant le contexte de référence et celui issu d'une première passe de RAP. Les segments contenant des mots HV sont étiquetés manuellement afin de simuler le processus de détection.

n-gramme	1	2	3	4	5
Rappel	100.0	88.2	50.5	27.3	16.1

Tab. 1: Rappel (en pourcentages) de la récupération des n-grammes contenant le mot HV sur le Web, avec Google, en fonction de la taille n.

Stratégie à base de N-grammes Cette première technique consiste à construire des requêtes en prenant le n-gramme contenant w_t et de remplacer ce mot par un joker. Dans le cas du contexte de transcription automatique, c'est le mot qui a été substitué au mot HV qui est remplacé par un joker. Cette stratégie est évaluée sur le contexte de référence et de transcription automatique afin de mesurer l'impact des erreurs de RAP sur les performances de la méthode. Les résultats sont présentés dans le tableau 2 pour le rappel et 3 pour la taille moyenne des listes de mots candidats.

Naturellement, le rappel est généralement meilleur avec le contexte de référence qu'avec le contexte de RAP. Par ailleurs, on remarque que plus les requêtes sont petites, moins elles sont discriminantes. Quand leur taille augmente, la précision augmente rapidement, mais le rappel chute, ce qui les rend inexploitables.

Pour conclure, On peut noter que bien que les mots HV cherchés se trouvent sur le Web, l'utilisation de n-grammes strictes ne semble pas être assez discriminant pour les retrouver. Dans la section suivante, nous proposons de relâcher les contraintes syntaxiques grâce à l'utilisation de patrons au lieu des n-grammes pour construire les requêtes.

Stratégie à base de patrons Cette méthode consiste à construire les requêtes en extrayant des patrons du contexte. Les patrons sont réalisés en remplaçant les mots-outils du contexte par des jokers qui se substitueront à un groupe de un à cinq mots lors de la recherche. Par exemple, pour la phrase *"Cette région a été touchée par le tsunami du 26 décembre"* avec le mot *"tsunami"* HV, le patron correspondant est *"région * touchée (*) 26 décembre"*. Dans les documents résultats, les mots qui auront été substitués au joker spécial *"(*)"* constitueront les mots HV candidats.

Les résultats présentés dans le tableau 2 et 3 montrent que le rappel est généralement meilleur qu'avec la méthode basée sur les n-grammes et ce quelque soit le contexte utilisé. De plus, la taille moyenne des listes de candidats augmente légèrement. Ces résultats laissent supposer que le fait de relâcher les contraintes sur les mots outils de la langage permet de récupérer des variantes du segment de phrase, ce qui introduit du bruit dans les listes de candidats mais permet d'accroître le rappel.

De plus, nous pouvons remarquer que pour les patrons de taille 2, le rappel décroît moins qu'avec la précédente méthode lorsque le contexte de transcription

[1] http ://www.google.fr

automatique est utilisé, ce qui indique une meilleure robustesse dans cette configuration.

Stratégie à base de sémantique locale Afin de réduire l'impact des erreurs de reconnaissance automatique, nous proposons de ne conserver que les mots porteurs de sens et de retirer toute contrainte de séquentialité lors de la construction des requêtes. Les mots présents dans une courte fenêtre temporelle autour du mot HV sont extraits et les n mots les moins fréquents dans la langue, donc les plus discriminants, sont utilisés pour construire les requêtes sans contrainte. Les listes de mots candidats sont construits en prenant l'ensemble des mots des documents les mieux classés par le moteur de recherche.

Nous pouvons voir dans le tableau 2 et 3 que le rappel augmente fortement avec le nombre de mots-clés, qui varie de 2 à 5. Avec la meilleure configuration, le rappel est plus de deux fois supérieur à celui de la meilleure configuration de la précédente méthode, mais que la précision décroît beaucoup. Ce dernier point est un inconvénient majeur pour l'intégration des candidats dans le processus de décodage. Considérant la complémentarité de la précédente approche avec celle-ci, nous avons essayé de combiner ces deux méthodes. Cette approche est présentée dans la section suivante.

	n-grammes		Patrons		Sémantique	
n	REF	RAP	REF	RAP	REF	RAP
2	14.0	4.7	20.0	7.3	32.6	18.5
3	18.1	5.1	20.3	5.0	39.7	27.8
4	16.4	2.3	17.5	2.0	45.9	35.2
5	13.8	1.9	12.3	1.2	50.2	40.9

Tab. 2: Rappel (en pourcentage) de la collecte de mots HV sur les 100 premiers documents retournés par Google, en fonction de la taille des requêtes n, avec le contexte de référence et de RAP.

	n-grammes		Patrons		Sémantique	
n	REF	RAP	REF	RAP	REF	RAP
2	145	322	411	475	16.0k	13.7k
3	49	207	139	166	19.0k	38.1k
4	13	34	34	21	37.9k	42.6k
5	4	9	15	8	44.9k	45.0k

Tab. 3: Taille moyenne des listes collectées dans les 100 premiers documents retournés par Google, en fonction de la taille des requêtes n, avec le contexte de référence et de RAP.

Stratégie combinant N-grammes et sémantique Comme nous l'avons vu dans la section 3.2, seulement 14% des 2-grammes contenant le mot recherché peuvent être retrouvés avec le contexte de référence tandis que environ 88% sont présents sur le Web. On peut supposer que la plupart des 2-grammes contenant le mot recherché sont dans les documents retournés par le moteur de recherche, mais qu'ils sont au delà du 100ème document, qui est la limite que avons fixée. Nous supposons qu'ajouter aux requêtes n-grammes précédemment décrites des mots porteurs de sens, extraits du contexte proche, aiderait le moteur de recherche à mieux ordonner les documents résultats et ainsi ferait remonter les documents contenant le mot HV recherché dans les 100 premiers résultats. Nous appelleront ces mots des mots-pilotes.

Les mots-pilotes sont extraits d'une fenêtre de taille fixe autour du mot HV. Les mots ayant la plus petite fréquence dans la langue sont sélectionnés comme mots-pilotes. Ces mots sont ajoutés aux requêtes sans aucune contrainte syntaxique. Les listes d'hypothèses sont construites comme pour la stratégie basée sur les n-grammes. Les résultats avec le contexte de référence sont présentés dans le tableau 4.

Les résultats dépassent significativement ceux obtenus avec la même méthode sans les mots-pilotes (tableaux 2 et 3). Un bon rappel peut être obtenu avec une augmentation minime du lexique. Par exemple, avec la configuration 2/2 nous obtenons environ 26% de rappel pour une augmentation du lexique de moins de 1000 mots. Comme toujours, l'utilisation du contexte de transcription automatique dégrade le rappel. Cette méthode reste cependant la meilleure en terme de compromis entre rappel et augmentation du lexique.

	Stratégie n-gram pilotés			
	REF		RAP	
n/m	Rappel	listes	Rappel	listes
2/1	24.0	268	8.7	292
2/2	26.1	789	8.1	306
2/3	27.0	1.3k	6.5	295
3/1	19.1	16	4.0	87
3/2	15.0	15	3.9	79
3/3	13.3	19	3.1	98

Tab. 4: Rappel (en pourcentage) et taille moyenne des listes d'hypothèses pour les 100 premiers documents retournés par Google, en utilisant la stratégie combinant N-grammes et sémantique avec n la taille du n-gramme et m le nombre de mots-pilotes sur le contexte de référence et de RAP.

4. Décodage avec les lexiques augmentés

Nous évaluons ici les performances des méthodes proposées pour corriger les sorties du système de RAP. Pour chaque segment contenant un mot HV, une seconde passe de décodage est réalisée avec le lexique adapté au segment.

4.1. Le modèle du mot inconnu

Nous avons tout d'abord cherché un moyen d'incorporer les listes d'hypothèses sans modifier le modèle de langage. Les mots candidats sont intégrés dans le lexique de décodage comme variantes de prononciation du mot inconnu. La phonétisation des nouveaux mots est réalisée automatiquement par LIA_PHON [2]. Les mots introduits auront donc tous la même probabilité, celle du mot inconnu. Les résultats sur le corpus de test montrent que 5% du nombre total de mots HV sont correctement décodés avec le lexique augmenté (appelé rappel dans le tableau 5). 8.7% du nombre total de mots HV étaient dans les listes de mots hypothèse (voir le tableau 4). La faible précision indique que beaucoup de mauvais mots sont in-

	Rappel	Précision	WER
baseline	0.0	0.0	24.5
mot inconnu	5.0	22.0	24.6
morpho.	6.1	55.1	24.3

Tab. 5: précision, rappel et WER (en pourcentage) du décodage avec la méthode du mot inconnu et des classes morphosyntaxiques pour l'augmentation des lexiques.

troduits mais cela ne perturbe pas significativement les performances. Cependant, 71.9% des mots correctement introduits sont des EN, 25.0% sont des termes techniques et les 3.1% restant sont des formes verbales peu fréquentes ou des noms communs. L'augmentation du WER est compensée par le nombre de mots porteurs de sens introduits, ce qui améliore la compréhension du document.

De plus, ces résultats montrent que quand le mot HV est dans le lexique augmenté, la probabilité que le décodage le fasse ressortir est d'environ 57%. Ces résultats sont résumés dans le tableau 5.

La mauvaise précision de cette méthode suggère une modélisation plus précise du langage.

4.2. Le modèle des classes grammaticales

De récents travaux se sont intéressés à l'utilisation de classes morphosyntaxique pour capturer le profile linguistique des mots peu fréquents [1]. Nous proposons d'estimer la probabilité du mot candidat à partir de la probabilité de sa classe morphosyntaxique. Cette méthode se base sur l'approximation n-gramme suivante :

$$P(w_u|w_i, ..., w_{i-n}) \approx \alpha * P(POS_w|w_{i-1}, ..., w_{i-n})$$

où w_u est le mot ajouté au lexique et POS_w la classe morphosyntaxique de w_u. α est le facteur d'échelle qui réduit la probabilité de la classe morphosyntaxique. La classe de w_u est déterminée par LIA_TAGG, un étiqueteur morphosyntaxique utilisant des HMM[2]. La probabilité n-gramme $P(POS_w|w_{i-1}, ..., w_{i-n})$ peut être estimée directement sur le corpus d'entraînement. Le facteur d'échelle α est estimé de manière expérimentale en essayant différentes valeurs et en les testant sur le corpus de développement.

Les résultats sont présentés dans le tableau 5, avec ceux de la précédente méthode. Nous observons que la méthode basée sur les classes morphosyntaxiques apporte un gain en terme de WER, de rappel et de précision de décodage. Le WER décroît de 0.2% absolu et 92.3% des mots HV retrouvés sont des EN. De plus, 7.7% du nombre total d'EN manquantes sont correctement retrouvées avec une bonne précision.

5. Conclusion et perspectives

Nous avons proposé une méthode pour augmenter la couverture lexicale en utilisant des lexiques localement augmentés. Cette méthode se base sur une stratégie de décodage en deux passes où la première sert à construire des requêtes Google. Des lexiques augmentés sont construits avec les documents retournés. Nous avons présenté plusieurs méthodes de formulation des requêtes. Nos résultats valident l'hypothèse initiale selon laquelle le contexte local renferme des informations sur les mots manquants. Les meilleures performances sont obtenues en combinant des patrons de mots et des mots-pilotes qui capturent la sémantique locale. Finalement, nous avons proposé d'intégrer les lexiques augmentés à l'aide d'un modèle de langage contenant des probabilités de classes morphosyntaxiques, ce qui constitue la seconde passe du processus de décodage. Ceci permet une réduction absolue de 0.2% du WER par rapport à la première passe. De plus, la plupart des mots HV retrouvés sont des EN et sont donc indispensables pour la compréhension du document. De manière générale, notre approche permet de retrouver 7.7% des mots HV importants pour le document tout en augmentant très légèrement les performances du système.

[2]http ://lia.univ-avignon.fr

Références

[1] A. Allauzen and J. Gauvain. Open Vocabulary ASR for Audiovisual Document Indexation. In *Proceedings of the ICASSP*, volume 1, pages 1013–1016, 2005.

[2] F. Béchet. LIA_PHON : Un système complet de phonétisation de textes. In *Proceedings of Traitement Automatique des Langues*, volume 42, pages 47–67, 2001.

[3] N. Bertoldi and M. Federico. Lexicon adaptation for broadcast news transcription. In *Proceedings of ISCA ITRW workshop on AMSR*, pages 187–190, 2001.

[4] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri. The ESTER evaluation campaign of rich transcription of french broadcast news. In *Proceedings of Language Resources and Evaluation Conference*, 2004.

[5] Y. Kajiura, M. Suzuki, A. Ito, and S. Makino. Generating search query in unsupervised language model adaptaion using www. *The Journal of the Acoustical Society of America*, 120(5) :3043–3044, 2006.

[6] M. Lapata and F. Keller. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.*, 2(1) :1–30, 2005.

[7] G. Linares, P. Nocera, D. Massonie, and D. Matrouf. The LIA Speech Recognition System : From 10xRT to 1xRT. *LECTURE NOTES IN COMPUTER SCIENCE*, 4629 :302, 2007.

[8] G.A. Monroe, J.C. French, and A.L. Powell. Obtaining language models of web collections using query-based sampling techniques. In *Proceedings of the 35th Annual Hawaii International Conference on*, pages 1241–1247, 2002.

[9] K. Ohtsuki, N. Hiroshima, M. Oku, and A. Imamura. Unsupervised vocabulary expansion for automatic transcription of broadcast news. In *Proceedings of the ICASSP*, pages 1021–1024, 2005.