L'adaptation thématique d'un modèle de langue fait-elle apparaître des mots thématiques ?

Gwénolé Lecorvé, Guillaume Gravier, Pascale Sébillot

IRISA, 263 av. Gén Leclerc Campus universitaire de Beaulieu, 35042 RENNES, France
{gwenole.lecorve,guillaume.gravier,pascale.sebillot}@irisa.fr

ABSTRACT

Whereas topic-based adaptation of language models (LM) claims to increase the accuracy of topic-specific words within automatic speech recognition, this paper investigates why this wish is not always verified. After outlining the mechanisms of LM adaptation and automatic speech recognition, diagnosing elements are proposed along with solutions. In addition to a better accuracy on topic-specific words, results show better graph error rates and word error rates on a set of spoken documents with various topics.
Keywords: language models, automatic speech recognition, topic-based adaptation

1. Introduction

L'adaptation thématique des modèles de langue (ML) vise à pallier le manque d'adéquation des ML généralistes appris une fois pour toute sur des textes aux sujets variés face à un document contenant de la parole thématiquement spécifique. Dans le cadre largement répandu des ML statistiques, basés sur des n-grammes, de multiples méthodes ont déjà été étudiées pour cette tâche [1]. Leur idée principale est d'acquérir automatiquement à partir d'un corpus des informations sur un thème considéré avant de les utiliser pour ré-estimer les probabilités d'un ML généraliste. Ceci permet que plus d'importance soit accordée aux n-grammes contenant des mots thématiques, mots porteurs d'une notion du thème en question.

Dans la littérature, les informations sur un thème sont souvent modélisées soit par des probabilités n-grammes estimées sur le corpus thématique [2], soit par une distribution unigramme obtenue par une technique d'analyse sémantique latente [3]. Par ailleurs, beaucoup de travaux récents recourrent à des techniques basées sur le minimum d'information discriminante (MDI) pour l'étape de ré-estimation du ML généraliste. C'est notamment le cas dans nos travaux [5]. Toutefois, alors que la technique d'adaptation en elle-même est donc largement étudiée, la majorité des travaux comparent des taux d'erreur sans s'interroger sur l'impact précis des ML adaptés dans le complexe processus de reconnaissance automatique de parole (RAP). Il n'est pourtant pas rare d'observer dans des transcriptions l'absence de mots thématiques prononcés dans le document sonore, alors que ceux-ci sont pourtant dans le vocabulaire du système et que leurs probabilités ont été adaptées dans le ML. Laissant donc de côté le problème des mots hors vocabulaire, cet article cherche à diagnostiquer pourquoi l'effet de l'adaptation thématique peut ainsi ne pas se faire ressentir et propose des éléments de solutions portant sur le système de RAP et sur l'adaptation du ML afin de réduire le nombre d'erreurs faites sur des mots thématiques.

Après avoir rappelé le fonctionnement de nos méthode d'adaptation et système de RAP, la section 2 présente un diagnostic général du manque d'effet de l'adaptation thématique, qui met au jour des réglages inadaptés du système de RAP et une adaptation parfois trop faible du ML. Les sections 3 et 4 présentent alors des solutions respectives à ces deux problèmes. Enfin, dans la section 5, nous présentons une manière de concilier représentation thématique et qualité de la modélisation du langage.

2. Vue d'ensemble

L'adaptation thématique dans un processus de RAP vise à ré-estimer le ML et à relancer un nouveau décodage sur la base de ce nouveau ML adapté. Dans cette section, nous présentons un aperçu du fonctionnement de ces deux étapes avant de donner quelques éléments de diagnostic quant à leurs insuffisances pour l'adaptation thématique.

2.1. Ré-estimation du modèle de langue

Notre technique d'adaptation thématique est fondée sur la méthode MDI. L'originalité et l'intérêt de cette méthode est de définir le ML adapté que nous cherchons à calculer comme la solution d'un système de contraintes dont l'entropie relative avec un ML de départ est minimale. Notre approche vise à n'augmenter que les probabilités des mots thématiques. Aussi, le système de contraintes utilisé ne porte que sur les probabilités des n-grammes se terminant par des mots appartenant à la terminologie du thème considéré, appelés *termes* par la suite. Pour un n-gramme hw, il en découle l'écriture suivante des probabilités conditionnelles P_A d'un ML adapté :

$$P_A(w|h) = \frac{P_B(w|h) \times \alpha(w)}{\sum_{\hat{w} \in V} P_B(\hat{w}|h) \times \alpha(\hat{w})} \quad (1)$$

$$\text{avec } \alpha(w) = \begin{cases} \frac{P_a(w)}{P_B(w)} & \text{si } w \text{ est un terme,} \\ 1 & \text{sinon,} \end{cases} \quad (2)$$

où V est le vocabulaire du système, P_B est la distribution du ML à adapter et P_a est une distribution estimée sur un petit corpus thématique.

En pratique, pour alléger le calcul du coefficient de normalisation dans (1), la masse de probabilité des évènements observés $\mathcal{E}(h)$ pour chaque historique h est contrainte à être conservée durant l'adaptation. Il en découle une nouvelle expression de $P_A(w|h)$:

$$P_A(w|h) = \frac{P_B(w|h) \times \alpha(hw)}{Z(h)} \quad (3)$$

$$\text{avec } Z(h) = \frac{\sum_{h\hat{w}\in\mathcal{E}(h)} P_B(\hat{w}|h) \times \alpha(h\hat{w})}{\sum_{h\hat{w}\in\mathcal{E}(h)} P_B(\hat{w}|h)}. \quad (4)$$

Outre la qualité variable de nos terminologies, due à un apprentissage automatique, notre méthode peut souffrir du fait que la présence d'un terme t dans la terminologie du thème n'implique pas que ce mot ait une probabilité $P_a(t)$ élevée dans notre corpus thématique. Ce phénomène conduit alors à une adaptation trop faible de certains n-grammes.

2.2. Système de RAP

Notre système de RAP est composé de plusieurs passes dont nous présentons seulement les principales du point de vue de l'utilisation du ML. Dans un premier temps, partant de paramètres acoustiques extraits d'un signal contenant de la parole, des graphes de mots sont générés par un algorithme de recherche en faisceau en utilisant des modèles acoustiques triphones et un ML 3-gramme sur un vocabulaire de 65 000 mots. Les graphes ainsi obtenus représentent, sous une forme plus ou moins compacte, l'intégralité des hypothèses de transcription que l'on peut s'attendre à obtenir en fin de décodage. Dans un second temps, ces graphes de mots sont réévalués en utilisant des modèles acoustiques triphones plus fins et un ML 4-gramme. Sur les nouveaux graphes ainsi obtenus, plus petits, un algorithme de recherche du meilleur chemin (Viterbi dans notre système) puis un décodage par consensus sont appliqués pour finalement obtenir une transcription du signal de départ.

Comme la potentialité d'apparition d'un mot dans le résultat final d'un décodage est avant toute chose conditionnée par sa présence dans les graphes de mots, l'étape de création des premiers graphes est particulièrement importante. Dans notre système, celle-ci est implémentée selon le principe de programmation dynamique détaillé dans [7] : chaque graphe de mots est construit au fur et à mesure que les hypothèses de phrases sont chronologiquement explorées. Pour que cette exploration s'effectue en un temps raisonnable, les hypothèses partielles les moins prometteuses sont itérativement écartées grâce à une stratégie d'élagage qui tient en trois points. Tout d'abord, à chaque trame t du signal décodé, seules sont conservées les hypothèses dont le score est suffisamment proche du score de la meilleure hypothèse :

$$Q_h(s,t) > \delta_{AC} \times \max_{(h',s')} \{Q_{h'}(s',t)\} \quad (5)$$

où s est un état de la copie d'arbre lexical de l'historique h. La constante δ_{AC} est appelé *seuil acoustique*. Ensuite, le même type de seuillage est appliqué pour chaque hypothèse de fin de mot. Seules seront explorées les nouvelles copies d'arbre des historiques dont le score est suffisamment proche du score $Q_{LM}(t)$ de la meilleure hypothèse de fin de mot à la trame t :

$$Q_{h'}(s_0,t) > \delta_{LM} \times Q_{LM}(t) \quad (6)$$

avec

$$Q_{h'}(s_0,t) \simeq \max_{h,s_w} \left\{Q_h(s_w,t) \times P(w|h)^\lambda \times I^{-1}\right\}, \quad (7)$$

où w est un mot supposé se terminer à un état s_w, h' est le nouvel historique dont l'état initial est s_0, $P(w|h)$ est la probabilité du ML, λ est le poids du ML et I est la pénalité d'insertion d'un mot. Le facteur δ_{LM} est appelé *seuil linguistique*. Enfin, parmi ces hypothèses de fin de mot ayant survécu, seules sont conservées les M hypothèses ayant le meilleur score. En pratique, les constantes λ, I, δ_{AC}, δ_{LM} et M sont des valeurs empiriques généralement obtenues par la recherche d'un compromis optimal entre différents critères comme le taux oracle des graphes générés, la durée de cette génération et la taille des graphes.

2.3. Premiers éléments de diagnostic

Dans cet article, nos expériences sont basées sur 91 documents sonores thématiquement homogènes issus de 3h d'émissions d'actualités du corpus ESTER 1 [4]. Ces segments, provenant de 3 radios différentes et datés de la même période, sont variés en terme de thème (guerre en Irak, politique nationale et internationale, sports...) et de longueur (de 30 à 2 000 mots). Pour chaque segment, les ML généralistes de notre système sont adaptés selon la méthode décrite en 2.1 puis un décodage basé sur ces nouveaux ML adaptés est exécuté. Nous présentons un bref bilan de ce décodage quant à son impact sur les mots thématiques.

L'intérêt d'utiliser un ML adapté durant un décodage est de favoriser en sortie l'émergence d'hypothèses comportant des séquences probables dans le thème considéré, séquences jusqu'alors sous-estimées par le ML généraliste. Si cet effet est bien observé lorsque les mots thématiques à corriger sont déjà dans les graphes de mots générés en première passe avec le ML généraliste, il apparaît au contraire que l'impact de l'adaptation thématique est souvent insuffisant pour insérer dans les graphes de mots des hypothèses comportant de nouveaux mots thématiques. Nous identifions principalement deux raisons à cela. Tout d'abord, nous avons observé que les termes d'un thème font particulièrement les frais de l'élagage de l'espace de recherche utilisé lors de la génération des graphes de mots en première passe – ceci en dépit de l'importance linguistique accrue que leur apporte un ML adapté. Nous expliquons ce phénomène par la tendance des termes à être des mots rares, des mots techniques, des entités nommées..., caractéristique qui accentue les risques pour ces mots d'être mal prononcés ou mal phonétisés. Ensuite, il semblerait que, malgré l'adaptation du ML, certains n-grammes se terminant par un terme aient toujours des probabilités trop faibles pour que les hypothèses de phrase qui les contiennent survivent au seuillage linguistique. La suite de cet article revient sur ces deux problèmes et présente des solutions envisageables.

3. Favoriser le modèle de langue

Malgré l'importance plus grande donnée aux termes dans les ML adaptés, le calcul des scores Q_h accorde toujours la même importance au ML, conduisant les hypothèses de phrases contenant ces termes à être élaguées. À ce phénomène peuvent s'ajouter

Tab. 1: GER et WER pour les réglages d'origine et pour les nouveaux réglages, avec le ML généraliste (ML_B) ou le ML adapté (ML_A).

	Réglages d'origine		**Réglages modifiés**	
	ML_B	ML_A	ML_B	ML_A
GER	8.9	8.6 (-0.3)	8.5	**8.1 (-0.4)**
WER	21.8	21.0 (-0.8)	21.0	**20.5 (-0.5)**

Tab. 2: Exemple d'alignement d'un groupe de souffle de référence (Réf) pour les réglages d'origine et modifiés, avec un ML généraliste (ML_B) ou adapté (ML_A).

Réf : `cas probable de la maladie`
Réglages d'origine
ML_B : `cas probable de la MÊLÉES`
ML_A : `cas probable de la MALAISIE`
Réglages modifiés
ML_B : `cas probable de la MÊLÉES`
ML_A : `cas probable de la maladie`

des conditions acoustiques difficiles, par exemple une mauvaise phonétisation, une prononciation erronée ou encore un locuteur parlant avec un accent régional ou étranger. Pour pallier ce problème, nous proposons de rendre le seuillage acoustique plus tolérant et de donner plus d'importance au ML dans le calcul des scores Q_h. En pratique, ceci consiste à diminuer δ_{AC} et à augmenter λ. Par ailleurs, pour contrebalancer l'augmentation du nombre d'hypothèses actives engendrée par la diminution de δ_{AC}, le nombre d'hypothèses de fin de mot pour chaque trame M est abaissé de manière à ce que le temps global de calcul soit conservé par rapport aux réglages d'origine.

Le tableau 1 présente les taux d'erreur en mots sur les graphes en première passe (GER) et sur les transcriptions finales (WER) obtenues, avec ou sans adaptation thématique, pour les réglages d'origine et nos réglages modifiés. Tout d'abord, il ressort clairement que la nouvelle configuration produit des taux d'erreur nettement meilleurs. Si les gains sur le GER s'expliquent par une augmentation constatée du nombre d'hypothèses de phrase dans les graphes générés en première passe, les résultats en WER montrent que nos anciens réglages n'étaient pas optimaux. Ensuite, il apparaît que les gains initiaux impliqués par l'adaptation thématique (colonne 2) se cumulent en partie mais pas entièrement avec ceux obtenus par nos nouveaux réglages sans adaptation (colonne 3). Ceci s'explique par le fait que les nouveaux réglages permettent notamment de corriger des erreurs sur des termes. Toutefois, les gains reportés pour l'utilisation conjointe des nouveaux réglages et de l'adaptation thématique (colonne 4) sont le signe d'une certaine complémentarité entre ces deux mécanismes. Le tableau 2 illustre ce propos en présentant des alignements d'un groupe de souffle tiré d'un document parlant de la pneumonie atypique où le locuteur prononce « *malédie* » au lieu de « *maladie* ». Alors que « *maladie* » n'apparaît dans aucun des trois premiers cas, soit à cause du problème acoustique, soit à cause d'une probabilité linguistique trop faible, la combinaison de nos nouveaux réglages et de l'adaptation thématique permet d'obtenir la bonne sortie. En contrepartie, l'augmentation de λ semble provoquer la disparition de nombreux mots courts (prépositions, articles...). De plus, certains termes n'apparaissent toujours pas dans les graphes de mots. Nous pensons que ceci est dû à la probabilité linguistique trop faible que peut leur attribuer notre technique d'adaptation.

Tab. 3: WER pour différents ML utilisés lors de la création des graphes, puis avec un ML adapté avec P_a pour le reste du décodage. Entre parenthèses, le type de poids utilisé lors de l'adaptation du premier ML.

	Réglages d'origine	**Réglages modifiés**
ML_B	21.8	20.9
ML_A ($P_a(t)$)	21.0 (-0.8)	**20.5 (-0.4)**
ML_A ($K = 10^{-8}$)	21.5 (-0.3)	20.8 (-0.1)
ML_A ($K = 10^{-5}$)	21.6 (-0.2)	21.0 (+0.1)

Tab. 4: Exemple d'alignement des graphes d'un groupe de souffle de référence (Réf) pour différents ML utilisés lors de la création des graphes.

Réf : `accès à la frontière du libéria`
Réglages modifiés
ML_B : `À SERT la ANTIENNE du DÉLIRE`
ML_A (P_a) : `MERCI à la frontière NOUVELLE`
ML_A (10^{-8}) : `MERCI à la frontière LIBÉRIENNE`
ML_A (10^{-5}) : `MERCI à la frontière libéria`

4. Surpondérer les termes

Comme évoqué en 2.1, la probabilité $P_a(t)$ d'un terme t peut conduire à une adaptation trop faible et inhiber l'apparition d'hypothèses de phrases contenant ce terme dans les graphes de mots. Pour pallier ce problème, nous proposons alors d'utiliser des poids K arbitrairement élevés et identiques pour tous les termes. Le facteur de mise à l'échelle $\alpha(w)$ d'un n-gramme hw dans (2) se réécrit alors :

$$\alpha(w) = \begin{cases} \frac{K}{P_B(w)} & \text{si } w \text{ est un terme,} \\ 1 & \text{sinon.} \end{cases} \quad (8)$$

Les résultats sur le GER obtenus pour différentes valeurs de K nous montrent, d'une part, que des poids trop élevés dégradent vite les performances et qu'un poids $K \approx 10^{-8}$ semble optimal quel que soit le réglage. En poursuivant le reste du décodage avec un ML adapté avec P_a, nous obtenons les taux d'erreur de la table 3. Ces résultats sont reportés pour deux valeurs de K : la valeur optimale en terme de taux oracle (10^{-8}) et une valeur beaucoup plus grande (10^{-5}). Dans l'ensemble, il ressort que l'utilisation de poids fixes n'apporte rien voire dégrade les taux d'erreur obtenus avec un ML généraliste. Cependant, une analyse qualitative des résultats montre qu'une adaptation à poids fixes et forts produit bien, lors de la création des graphes, une apparition de termes qui n'apparaissaient jusqu'alors pas, même en utilisant une adaptation classique avec P_a (*cf.* exemple de la table 4).

Nous expliquons ce comportement par deux raisons. D'une part, nous pensons que certaines hypothèses introduites lors de la création des graphes ne survivent pas aux différents réglages d'élagage des étapes postérieures. Il faudrait alors modifier ces réglages, notam-

Tab. 5: GER mesurés après fusion des graphes de mots issus d'un premier décodage avec le ML généraliste et d'un second décodage avec différents ML.

1er décodage ► / **▼ 2nd décodage**		Réglages d'origine + ML_B	Réglages modifiés + ML_B
Réglages d'origine	+ ML_B	8.9	7.8 (-1.1)
	+ ML_A	8.5 (-0.4)	7.7 (-1.2)
Réglages modifiés	+ ML_B	7.8 (-1.1)	8.5 (-0.4)
	+ ML_A	**7.6 (-1.3)**	8.1 (-0.8)

ment encore une fois au niveau acoustique. D'autre part, la normalisation utilisée en pratique dans MDI (formule 4) ne permet pas une modification globale de la distribution d'un ML mais aboutit seulement à des modifications locales de celle-ci, historique par historique. Lorsque les poids considérés sont importants, ceci tend à rendre moins probables, en moyenne, les historiques adaptés que ceux qui ne le sont pas.

5. Fusionner les graphes de mots

Comme le montrent les expériences et observations précédentes, il est difficile de concilier au niveau d'une adaptation de ML les informations que l'on possède sur un thème avec celles plus généralistes d'un ML initial. Il semble alors intéressant de s'orienter vers une intégration *a posteriori* de celles-ci, notamment via la fusion de graphes de mots [6]. Pour cela, nous définissons la fusion de deux graphes de mots comme le graphe déterminisé représentant l'union de l'ensemble des phrases codées par chaque graphe. Nous appliquons cette méthode de fusion entre les graphes générés avec le ML généraliste et ceux obtenus grâce à un ML adapté.

Le tableau 5 présente les résultats GER obtenus par cette méthode de fusion pour différents couples de réglages. Il apparaît que les gains les plus importants sont ceux où sont utilisés deux réglages différents de l'algorithme de création des graphes (cellules grisées). Ces gains dépassent nettement les gains sans fusion de la table 1. L'effet de la fusion est tel que celui de l'adaptation thématique semble quasi gommé. Nos résultats sur le WER montrent cependant que ces différences sont lissées à la sortie du système. Seuls ressortent des écarts notables pour chaque couple de réglages entre l'utilisation d'un ML généraliste et d'un ML adapté.

Ces résultats quelque peu décevants nous apparaissent cependant comme logiques. En effet, notre méthode de fusion n'aboutit qu'à considérer l'union des hypothèses des graphes fusionnés et n'introduit donc aucune nouvelle hypothèse. Ainsi, quel que soit le ML utilisé, celui-ci privilégiera quasi toujours les mêmes hypothèses qu'il privilégiait déjà sans fusion. Il serait intéressant d'étudier l'impact de stratégies de fusion plus élaborées permettant d'introduire de nouvelles hypothèses, par exemple une stratégie basée sur la combinaison de réseaux de confusion.

6. Conclusions

Dans cet article, nous avons cherché à diagnostiquer l'impuissance à transcrire des mots thématiques dont souffre parfois l'intégration d'un ML adapté thématique dans le processus de RAP. Pour cela, deux mécanismes ont été proposés au niveau de la génération des graphes de mots en première passe : une prise en compte moins forte de l'acoustique au profit des scores du ML et une adaptation plus marquée du ML utilisé en première passe. De plus, pour concilier qualité de représentation thématique et modélisation du language, nous avons proposé une technique de fusion de graphes. Outre les gains GER et WER qui peuvent être relevés, ces mécanismes tendent à améliorer la transcription des mots thématiques au détriment de portions de texte plus généralistes.

Ces résultats ouvrent probablement la voie à de meilleurs résultats dans des tâches où les mots discriminants ont une importance particulière comme, par exemple, l'indexation. Au-delà de cela, il est intéressant de réfléchir plus en avant au principe d'adaptation thématique. En effet, ce dernier est trop souvent centré sur la tâche de ré-estimation du ML dont le résultat serait censé se suffire à lui-même. À notre sens, ce postulat est illusoire. Nous pensons que différentes solutions peuvent être envisagées. Par exemple, il serait bon de systématiser les mécanismes de fusion *a posteriori* ou de systèmes de RAP collaborant en parallèle. De même, il serait intéressant de réfléchir à une modification des algorithmes du processus de RAP pour pouvoir intégrer directement des modèles indépendants de différentes sources d'information. Enfin, bien que la question des mots hors vocabulaire n'ait pas été traitée dans cet article, celle-ci joue un rôle dans l'impact d'une adaptation thématique car la rareté et la spécificité générale des mots thématiques implique souvent leur absence dans le vocabulaire du système. L'adaptation de ce dernier est donc un enjeu majeur pour l'adaptation thématique d'un système.

Références

[1] J. R. Bellegarda. Statistical language model adaptation : review and perspectives. *Speech Communications*, 2004.

[2] M. Federico. Efficient language model adaptation through MDI estimation. In *Proc. Eurospeech*, 1999.

[3] M. Federico. Language model adaptation through topic decomposition and MDI estimation. In *Proc. ICASSP*, 2002.

[4] S. Galliano, É. 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 *Proc. Eurospeech*, pages 1149–1152, 2005.

[5] G. Lecorvé, G. Gravier, and P. Sébillot. Constraint selection for topic-based MDI adaptation of language models. In *Proc. Interspeech*, pages 368–371, 2009.

[6] X. Li, R. Singh, and R. M. Stern. Combining search spaces of heterogeneous recognizers for improved speech recognition. In *Proc. ICSLP*, pages 405–408, 2002.

[7] H. Ney and S. Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, pages 64–83, 1999.