IDIAP RESEARCH REPORT

IMPACT DU DEGRé DE SUPERVISION SUR L'ADAPTATION à UN DOMAINE D'UN MODèLE DE LANGAGE à PARTIR DU WEB

Gwénolé Lecorvé John Dines Thomas Hain
Petr Motlicek

Idiap-RR-23-2012

JULY 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Impact du degré de supervision sur l'adaptation à un domaine d'un modèle de langage à partir du Web

Gwénolé Lecorvé[1] John Dines[1,2] Thomas Hain[3] Petr Motlicek[1]
(1) Idiap Research Institute, Martigny, Suisse (2) Koemei, Martigny, Suisse
(3) University of Sheffield, Sheffield, Royaume-Uni
glecorve@idiap.ch, dines@idiap.ch, t.hain@dcs.shef.ac.uk, motlicek@idiap.ch

RÉSUMÉ

L'adaptation à un domaine d'un modèle de langage consiste à réestimer ses probabilités afin de mieux modéliser les spécificités linguistiques d'un thème considéré. Pour ce faire, une approche désormais classique est de récupérer des pages Web propres au domaine à partir d'un échantillon textuel représentatif de ce même domaine, texte appelé noyau. Cet article présente une étude originale sur l'importance qu'a le choix du noyau sur le processus d'adaptation et sur les performances des modèles de langage adaptés en reconnaissance automatique de la parole. Le but de cette étude est d'analyser les différences entre une adaptation supervisée, au sein de laquelle le noyau est généré manuellement, et une adaptation non supervisée, où le noyau est une transcription automatique. Nos expériences, menées sur un cas d'application réel, montrent que les différences varient selon les scénarios d'adaptation et que l'approche non supervisée est globalement convaincante, notamment au regard de son faible coût.

ABSTRACT

Impact of the level of supervision on Web-based language model domain adaptation

Domain adaptation of a language model aims at re-estimating word sequence probabilities in order to better match the peculiarities of a given broad topic of interest. To achieve this task, a common strategy consists in retrieving adaptation texts from the Internet based on a given domain-representative seed text. In this paper, we study the influence of the choice of this seed text on the adaptation process and on the performances of adapted language models in automatic speech recognition. More precisely, the goal of this original study is to analyze the differences between supervised adaptation, in which the seed text is manually generated, and unsupervised adaptation, where the seed text is an automatic transcript. Experiments carried out on videos from a real-world use case mainly show that differences vary according to adaptation scenarios and that the unsupervised approach is globally convincing, especially according to its low cost.

MOTS-CLÉS : Modèle de langage, adaptation à un domaine, supervision, données du Web.

KEYWORDS: Language model, domain adaptation, supervision, Web data.

1 Introduction

Le modèle de langage (ML) *n*-gramme de la plupart des systèmes de reconnaissance automatique de la parole (RAP) est habituellement appris sur une vaste collection de textes de domaines variés. Par conséquent, ce ML généraliste n'est plus optimal dès lors qu'il s'agit de transcrire des documents oraux traitant d'un domaine précis. Pour résoudre ce problème, l'adaptation à un domaine d'un ML cherche à réestimer les probabilités *n*-grammes du ML généraliste de manière à prendre en compte les spécificités linguistiques du domaine considéré, le but final étant d'améliorer les taux de reconnaissance du système de RAP.

Une approche d'adaptation désormais commune consiste à utiliser le Web comme un corpus ouvert afin de récupérer des données propres au domaine et d'extraire des statistiques pour la réestimation des probabilités n-grammes (Zhu et Rosenfeld, 2001; Wan et Hain, 2006; Bulyko *et al.*, 2007; Lecorvé *et al.*, 2008). Ce processus fondé sur le Web se scinde principalement en trois étapes : tout d'abord, des requêtes sont extraites à partir d'un texte supposé représentatif du domaine considéré – nous parlerons de *texte noyau* ou plus simplement de *noyau* ; ensuite, des pages Web sont récupérées en soumettant les requêtes à un moteur de recherche sur Internet ; et finalement, un ML adapté est construit grâce aux données d'adaptation récupérées.

Le texte noyau est un point-clé du processus car celui-ci doit permettre d'extraire des informations permettant de caractériser le domaine considéré. Dans la littérature, deux approches sont proposées : l'une supervisée où le domaine est connu *a priori* et le noyau est fait de textes collectés manuellement, typiquement des transcriptions manuelles (Sethy *et al.*, 2005; Wan et Hain, 2006) ; l'autre non supervisée où le noyau est construit automatiquement à partir de documents oraux à transcrire, généralement des transcriptions automatiques (Suzuki *et al.*, 2006; Lecorvé *et al.*, 2008). Alors que l'approche supervisée semble intuitivement la plus performante car celle-ci est exempte d'erreurs de transcription, peu de travaux ont toutefois cherché à déterminer clairement l'impact du niveau de supervision sur les performances des ML adaptés. Seul (Tür et Stolcke, 2007) semble s'y être intéressé de près. Cependant, la méthode étudiée dans ce dernier travail ne s'appuie pas sur Internet. Ainsi, notre article vise à comparer l'emploi de différents degrés de supervision sur une même technique d'adaptation fondée sur le Web. Nous cherchons à comprendre quels gains peuvent être attendus en RAP pour des scénarios d'adaptation donnés et, particulièrement, quels impacts peut avoir la présence d'erreurs de transcription dans le noyau.

Cet article s'organise comme suit : la section 2 présente notre technique d'adaptation d'un ML, la section 3 décrit notre cadre expérimental et introduit différents scénarios d'adaptation, puis la section 4 étudie l'impact de ces scénarios sur différents aspects de la technique d'adaptation.

2 Technique d'adaptation du modèle de langage

Notre technique d'adaptation d'un ML tient en trois temps. Étant donné un texte noyau représentatif d'un domaine visé, des requêtes sont tout d'abord extraites. Puis, en soumettant ces requêtes à un moteur de recherche en ligne, des pages Web sont récupérées et un corpus d'adaptation est construit. Finalement, un ML adapté est appris en ajoutant ces données d'adaptation à l'ensemble des autres données textuelles ayant initialement servi à apprendre le ML généraliste. Le nouveau ML est alors censé conduire à des transcriptions automatiques meilleures que celles que fournirait le ML généraliste pour des documents oraux traitant du domaine en question. Cette section décrit notre stratégie d'extraction de requêtes avant d'expliquer comment les pages Web sont récupérées et comment le ML adapté est estimé en pratique.

2.1 Extraction des requêtes à partir du texte noyau

Le principe de notre méthode d'extraction de requêtes, telle que présentée dans (Wan et Hain, 2006), est d'analyser quels sont les n-grammes les plus mal modélisés par le ML généraliste d'après le texte noyau et d'utiliser ces n-grammes comme des requêtes. Concrètement, étant donné le texte noyau T, tous les trigrammes de T qui n'ont pas été observés lors de l'apprentissage du ML généraliste sont considérés comme des requêtes potentielles, c.-à-d. tous les trigrammes de T dont la probabilité se calcule par le mécanisme de *back-off*. Comme ces n-grammes peuvent

être nombreux selon la taille du noyau T, ce qui conduirait à une trop longue récupération des pages Web, et comme beaucoup d'entre eux sont simplement des séquences de mots sans importance pour le domaine, ces trigrammes sélectionnés sont filtrés en supprimant tous ceux qui contiennent au moins un mot vide. La liste des mots vides est faite d'environ 600 mots-outils anglais [1]. Dans nos expériences, cette stratégie conduit à l'extraction de quelques centaines de requêtes.

Cette méthode d'extraction se justifie sur un plan théorique. En effet, celle-ci garantit que, une fois que les statistiques des pages Web récupérées auront été intégrées dans le ML adapté, la probabilité conditionnelle de chaque trigramme-requête (w_1, w_2, w_3) sera supérieure pour le ML adapté que pour le ML généraliste. Mathématiquement, cela s'exprime ainsi :

$$P_A(w_3|w_1, w_2) > P_G(w_3|w_1, w_2) \qquad \forall (w_1, w_2, w_3) \in Q \ , \tag{1}$$

où Q est l'ensemble des requêtes alors que P_A et P_G désignent respectivement les distributions de probabilités du ML adapté recherché et du ML généraliste. Puisque les requêtes sont toutes extraites du noyau T, il en découle que la vraisemblance du noyau est plus grande pour le modèle adapté que pour le modèle généraliste :

$$P_A(T) > P_G(T) \ . \tag{2}$$

L'utilisation du ML adapté sur la base de ces requêtes doit donc profiter au système de RAP pour le domaine considéré, sous l'hypothèse que T est suffisamment caractéristique de ce domaine.

2.2 Récupération des pages Web et apprentissage du modèle adapté

Pour récupérer des données d'adaptation propres au domaine, les requêtes sont soumises à un moteur de recherche sur Internet (en l'occurrence, Bing) et les liens retournés sont téléchargés selon un algorithme en tourniquet, c.-à-d. que les i-èmes résultats de chaque requête sont téléchargés avant de télécharger les $(i+1)$-èmes résultats... Cette stratégie a l'avantage de donner une importance égale à chaque requête, ce qui semble une pratique raisonnable en l'absence d'information *a priori* sur le domaine et sur la pertinence des requêtes. Les pages Web sont nettoyées, normalisées puis rassemblées au sein d'un corpus d'adaptation [2]. Ce processus s'arrête dès qu'un certain nombre de mots est atteint. Dans nos expériences, ce seuil est arbitrairement fixé à 5 millions de mots, ce qui requiert de télécharger entre 20 et 40 pages par requête.

Le corpus d'adaptation est ensuite ajouté à l'ensemble des corpora initialement utilisés pour apprendre le ML généraliste. Un ML est appris indépendamment pour chacune de ces sources, puis ces ML individuels sont interpolés linéairement de manière à ce que leur combinaison minimise la perplexité sur le texte noyau. Notons que le vocabulaire de chacun de ces ML reste le même que celui d'origine car nous nous concentrons dans cet article sur la seule adaptation du ML. Pour finir, le ML résultant de l'interpolation est élagué afin d'obtenir un modèle de taille comparable à celle du ML généraliste.

3 Cadre expérimental et scénarios d'adaptation

Avant de présenter l'impact sur ce processus d'adaptation du choix supervisé ou non supervisé du noyau, cette section présente notre cadre expérimental, c.-à-d. le système de RAP et les données

1. En français, ce filtrage devrait sans doute être assoupli car l'articulation des mots est différente. Cependant, cette dépendance à la langue ne peut être tenu pour de la supervision car le filtrage reste indépendant du domaine traité.
2. Le processus de nettoyage des pages Web étant abouti, le corpus d'adaptation n'est que peu bruité.

utilisées, puis introduit les scénarios d'adaptation étudiés.

3.1 Cadre expérimental

Le système de RAP utilisé est un système multi-passes pour l'anglais, largement décrit dans (Hain *et al.*, 2012). Principalement, il s'appuie sur un vocabulaire de 50 000 mots et un ML quadrigramme interpolé à partir de ML appris indépendamment sur de nombreux corpora formant un total d'environ un milliard de mots.

Le domaine considéré est représenté par 57 vidéos provenant de la chaîne YouTube d'un établissement d'enseignement supérieur spécialisé. Alors que les thèmes abordés sont homogènes car centrés sur le contenu des cours, ces vidéos sont de types variés (cours magistraux, autopromotion, conférences, interviews...), elles ont été enregistrées dans des conditions acoustiques différentes et certains intervenants ne sont pas d'origine anglophone. Les transcriptions manuelles des vidéos forment un total de 40 000 mots. Les vidéos sont séparées en deux ensembles : un ensemble de développement de 29 vidéos à partir duquel des informations peuvent être extraites pour l'adaptation et un ensemble d'évaluation de 28 vidéos uniquement dédié aux tests. Les références respectives de ces ensembles sont de longueurs équivalentes, soit environ 20 000 mots chacune.

L'impact de l'adaptation au domaine est principalement analysé en comparant les perplexités du ML généraliste avec celles des ML adaptés à partir de différents textes noyaux, tant sur l'ensemble de développement que sur celui d'évaluation. Pour les configurations les plus intéressantes, nous rapportons aussi des taux d'erreurs sur les mots (WER).

3.2 Scénarios d'adaptation

Le but de cet article est d'étudier l'importance qu'a le choix du noyau sur l'efficacité de l'adaptation du ML. Cette adaptation peut principalement s'inscrire au sein de deux scénarios. Soit l'adaptation vise à fournir une nouvelle transcription de documents ayant déjà été transcrits une première fois sur la base du ML généraliste – nous parlerons d'*auto-adaptation*. Soit l'adaptation est dédiée à l'usage à long terme du ML adapté pour transcrire de futures vidéos traitant d'un même domaine – nous parlerons d'*adaptation à long terme*. Sur la base de nos deux ensembles de vidéos traitant d'économie, nous définissons trois principales valeurs possibles que le texte noyau peut prendre afin de mettre en œuvre ces scénarios. Ces valeurs, listées ci-dessous, s'échelonnent du cas le plus supervisé à celui complètement non supervisé.

1. Le noyau est la *référence de l'ensemble de développement*. Il s'agit du cas le plus supervisé et, du fait de la génération de transcriptions manuelles, également du plus coûteux à mettre en place, en temps comme en argent.
2. Le noyau est constitué d'un *ensemble de pages Web aspirées sur le site de l'établissement produisant les vidéos* à partir d'un point d'entrée fourni manuellement. Ces pages représentent un total de 400 000 mots. Ce cas est moins supervisé car le contenu de ces pages Web ne coïncident pas complètement avec celui des vidéos, surtout au niveau du style.
3. Le noyau est la *transcription automatique de l'ensemble de développement*. Il s'agit du cas non supervisé. Mis à part le temps nécessaire à la génération des transcriptions automatiques, cette solution s'avère la moins coûteuse. Elle reste néanmoins peu fiable car le WER est de 29,6 %. Dans les tableaux de résultats, ce cas est désigné par « RAP ».

La prochaine section présente comment la méthode d'adaptation du ML se comporte pour chacun de ces trois cas au sein des deux étapes de la méthode qui font intervenir le texte noyau.

4 Expériences et résultats

Le texte noyau joue un rôle durant deux étapes du processus d'adaptation à un domaine : il est utilisé pour extraire des requêtes propres au domaine et il sert à déterminer l'importance des données d'adaptation au moment de combiner ces dernières avec celles utilisées initialement pour estimer le ML généraliste. Dans cette section, nous étudions ainsi tout d'abord l'impact du noyau sur l'extraction des requêtes avant d'analyser son rôle lors de l'interpolation linéaire.

4.1 Impact du noyau sur l'extraction de requêtes

L'extraction de requêtes est la première étape du processus d'adaptation. Aussi, la qualité du texte noyau est probablement cruciale. Pour tester cette hypothèse, la table 1 compare les perplexités obtenues en utilisant soit le ML généraliste soit les ML adaptés à partir de différents noyaux. Les résultats sur les ensembles de développement et d'évaluation illustrent respectivement les scénarios d'auto-adaptation et d'adaptation à long terme. Pour chaque ML, l'interpolation linéaire est effectuée en minimisant la perplexité de la référence afin de mettre en avant les meilleures perplexités possibles pour chaque noyau. Nous constatons que, sur l'ensemble de développement, les meilleures améliorations sont de loin obtenues quand le noyau est la référence. Ce résultat est logique car cette configuration (en italique) est un cas artificiel où le noyau et le texte à prédire par le ML adapté sont le même. Les résultats moindres de la référence sur l'ensemble d'évaluation sont donc logiques. Ensuite, les perplexités obtenues par les pages Web collectées manuellement sont les meilleures sur l'ensemble d'évaluation, probablement car ce noyau est plus grand que la référence tout en étant fiable, il permet donc une bonne caractérisation du domaine. De manière plus surprenante, l'emploi des transcriptions automatiques conduit à des améliorations proches de celles des cas supervisés.

Nous avons mené une seconde série d'expériences afin de déterminer d'où viennent les différences observées entre la référence et sa transcription automatique. Les résultats de ces expériences sont présentés par les trois dernières lignes de la table 1. Trois nouveaux ensembles de requêtes ont été dérivés des cas précédents. À partir des requêtes issues de la transcription automatique, un premier ensemble regroupe les requêtes sans aucune erreur de transcription (RAP sans erreur) alors qu'un second contient les autres requêtes où au moins une erreur est présente (RAP avec erreur(s)). Le dernier ensemble liste ce qu'aurait dû être ces requêtes erronées si le système de RAP ne faisait pas d'erreur (Référence des erreurs). Sur l'ensemble de développement, il apparait que les forts gains amenés par la référence étaient dus aux requêtes que le système a du mal à transcrire. Ceci se révèle logique car il s'agit aussi implicitement des trigrammes les plus mal modélisés par le ML généraliste. Sur l'ensemble d'évaluation, ce constat n'est plus vrai. Seule une différence entre les requêtes avec ou sans erreurs de transcription persiste. On remarque toutefois que les requêtes erronées conduisent malgré tout à des diminutions significatives de la perplexité. Après analyse, ce résultat surprenant s'explique, d'une part, par le fait que les moteurs de recherche actuels transforment automatiquement certaines requêtes peu vraisemblables vers d'autres plus communes [3] et, d'autre part, par l'absence de résultats lorsque les requêtes sont vraiment dénuées de sens, aucune page Web ne venant alors biaiser le corpus d'adaptation.

Pour résumer, nous pouvons conclure que, pour l'adaptation à long terme, il est préférable de s'appuyer sur des pages Web pour extraire des requêtes de manière supervisée. Cette solution est d'autant plus valable qu'elle est moins coûteuse que le recours à des transcriptions manuelles. Dans un contexte non supervisé, l'emploi de transcriptions automatiques est peu pénalisant

3. Ces transformations sont facilitées par le fait que certaines erreurs de transcription n'altèrent pas la racine des mots de la référence et correspondent donc malgré tout à des mots caractéristiques du domaine.

Extraction de requêtes	Interpolation linéaire	Développement	Évaluation
ML généraliste		165	170
Référence	Référence	*119 (-27,9 %)*	139 (-18,2 %)
Pages Web	Référence	129 (-21,8 %)	137 (-19,4 %)
RAP	Référence	133 (-19,4 %)	143 (-15,9 %)
RAP sans erreur	Référence	134 (-18,8 %)	143 (-15,9 %)
RAP avec erreur(s)	Référence	142 (-13,9 %)	150 (-11,8 %)
Référence des erreurs	Référence	120 (-27,3 %)	140 (-17,6 %)

TABLE 1 – Perplexités obtenues sur les ensembles de développement et de test avant et après adaptation à partir de différents textes noyaux pour l'extraction de requêtes. Pour chaque configuration, le corpus d'adaptation récupéré sur le Web contient 5 millions de mots. Entre parenthèses, les variations relatives par rapport à l'utilisation du ML généraliste.

car l'impact des erreurs de transcription se limite à brider l'information disponible sans biaiser l'adaptation.

4.2 Impact du noyau sur l'interpolation linéaire

Le second aspect concerné par le choix du texte noyau est l'estimation des poids de l'interpolation linéaire finale. La table 2 présente l'impact des différentes possibilités en terme de perplexité. Pour commencer, les lignes (a) montrent que l'adaptation n'a presque aucun effet lorsque l'interpolation est guidée par le texte hors domaine ayant servi à construire le ML généraliste. À l'inverse, les résultats (b) montrent que, sans récupérer de données d'adaptation sur Internet, la simple réinterpolation des corpora généralistes sur la base d'un texte propre au domaine conduit à de légères améliorations. Parmi ces résultats, l'utilisation des pages Web semblent néanmoins moins judicieuse.

Les lignes (c) correspondent aux configurations où le même noyau est utilisé pour l'extraction des requêtes et pour l'interpolation linéaire, comme cela serait vraisemblablement le cas dans une vraie application. On remarque que, pour les transcriptions automatiques, les écarts précédemment observés avec la référence se cumulent. Au contraire, l'emploi des pages Web collectées manuellement conduit à des résultats anormalement moins bons. Une analyse plus poussée nous montre que ce phénomène s'explique par le fait que certaines pages Web automatiquement récupérées sont les mêmes ou sont très proches de celles ayant servi à extraire les requêtes. Il en découle un poids très élevé associé au ML appris sur les données d'adaptation et, comme ces dernières ne représentent que 5 millions de mots, une perplexité moindre du ML interpolé. Bien que nous n'ayons conduit aucune expérience pour résoudre ce problème, il est probable que l'exclusion de ces pages gênantes du corpus d'adaptation aboutirait à un ML adapté de meilleure qualité. Enfin, la ligne (d) montre quels seraient les résultats si l'on était capable de supprimer les portions mal transcrites dans la transcription automatique, pour l'extraction de requêtes et pour l'interpolation linéaire [4]. Cette suppression ne conduit qu'à une très faible amélioration par rapport aux résultats obtenus *via* la transcription automatique complète.

Les WER obtenus par les configurations (c) et (d) sont présentés par la table 3. Des diminutions significatives sont observées par rapport au ML généraliste et les tendances sont les mêmes que celles notées pour la perplexité : l'emploi des pages Web collectées manuellement conduit aux

4. Pour l'interpolation linéaire, les erreurs de transcription ont été remplacées par des mots hors vocabulaire.

	Extraction de requêtes	Interpolation linéaire	Développement	Évaluation
	ML généraliste		165	170
(a)	Référence	Texte hors domaine	159 (-3,6 %)	168 (-1,2 %)
	Pages Web	Texte hors domaine	164 (-0,6 %)	169 (-0,6 %)
	RAP	Texte hors domaine	163 (-1,2 %)	169 (-0,6 %)
(b)	Aucune donnée	Référence	154 (-6,7 %)	159 (-6,5 %)
	Aucune donnée	Pages Web	158 (-4,2 %)	164 (-3,5 %)
	Aucune donnée	RAP	155 (-6,1 %)	161 (-5,3 %)
(c)	Référence		119 (-27,9 %)	139 (-18,2 %)
	pages Web		141 (-14,5 %)	151 (-11,2 %)
	RAP		136 (-17,6 %)	145 (-14,7 %)
(d)	RAP sans erreur		135 (-18,2 %)	143 (-15,9 %)

TABLE 2 – Perplexités obtenues avant et après adaptation à partir de différents noyaux pour l'estimation des poids de l'interpolation linéaire. Entre parenthèses, les variations relatives.

Extraction de requêtes	Interpolation linéaire	Développement	Évaluation
ML généraliste		29,6 %	25,8 %
Référence		26,8 % (-2,8)	24,1 % (-1,7)
Pages Web		27,7 % (-1,9)	24,9 % (-0,9)
RAP		27,3 % (-2,3)	24,6 % (-1,2)
RAP sans erreur		27,5 % (-2,1)	24,4 % (-1,4)

TABLE 3 – WER obtenus avec ou sans adaptation. Entre parenthèses, les variations absolues.

améliorations les plus faibles et la référence aux plus élevées. Les différences d'amélioration de la perplexité entre les ensembles de développement et d'évaluation sont néanmoins amplifiées, probablement en raison d'un WER de départ plus bas sur l'ensemble d'évaluation. Dans le détail, les écarts en terme de WER sont relativement faibles entre l'emploi de la référence et celui de la transcription automatique, y compris dans le cas d'une auto-adaptation (ensemble de développement). En outre, la suppression des passages mal transcrits ne produit toujours pas d'écart significatif par rapport à l'utilisation de la transcription complète. Nous pouvons en conclure que les erreurs de transcription n'empêchent pas la réussite de l'adaptation. Leur impact essentiel est de supprimer des informations qui permettraient de mieux caractériser le domaine.

5 Conclusion

Dans cet article, nous avons mené une étude originale sur l'impact du niveau de supervision lors d'une adaptation à un domaine d'un ML. Concrètement, plusieurs scénarios ont été testés sur notre méthode d'adaptation fondée sur le Web afin de mettre en évidence l'influence qu'a le choix du texte noyau utilisé pour extraire des requêtes et pour prendre en compte les probabilités *n*-grammes issues des données d'adaptation. Il apparait logiquement que l'emploi de transcriptions manuelles produit le ML adapté ayant les meilleures performances, en perplexité comme en matière de taux d'erreurs sur les mots. Cependant, d'autres conclusions intéressantes ont émergé. Tout d'abord, les erreurs de transcriptions n'affectent pas beaucoup notre méthode d'adaptation, que ce soit pour l'extraction de requêtes ou pour l'interpolation linéaire. Au lieu

de faire échouer l'adaptation, ces erreurs semblent simplement brider son degré de réussite en limitant l'information disponible. Ce résultat est d'autant plus intéressant que l'estimation de mesures de confiance fiables et le repérage automatique de zones mal transcrites dans une transcription automatique sont des tâches difficiles. À l'inverse, l'étude sur l'estimation des poids de l'interpolation linéaire a montré que certains effets de bord pouvaient dégrader significativement les performances de l'adaptation. L'utilisation de pages Web collectées manuellement a en effet conduit à une adaptation trop forte du ML.

D'autres aspects ayant trait à la supervision du processus mériteraient d'être étudiés. Par exemple, il serait bon de savoir quel impact a le WER des transcriptions automatiques fournies initialement par le ML généraliste ou encore quel serait le comportement de la méthode si la taille du texte noyau était limitée. Par ailleurs, bien que nous ayons volontairement laissé de côté le problème de l'adaptation du vocabulaire, un travail complémentaire pourrait consister à analyser la propension des différents textes noyaux à conduire à des corpora d'adaptation permettant l'ajout de mots hors vocabulaire au système de RAP.

Remerciements

Ce travail est financé par le projet n° CTI 12189.2 PFES-ES de la Commission pour la Technologie et l'Innovation (CTI, Suisse) et a été effectué en collaboration avec Koemei.

Références

Bulyko, I., Ostendorf, M., Siu, M., Ng, T., Stolcke, A. et Çetin, O. (2007). Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing*, 5(1):1–25.

Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., van Leeuwen, D., Lincoln, M. et Wan, V. (2012). Transcribing meetings with the AMIDA systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:486–498.

Lecorvé, G., Gravier, G. et Sébillot, P. (2008). An unsupervised Web-based topic language model adaptation method. *In Proceedings of ICASSP*, pages 5081–5084.

Sethy, A., Georgiou, P. G. et Narayanan, S. (2005). Building topic specific language models from Webdata using competitive models. *In Proceedings of Eurospeech*, pages 1293–1296.

Suzuki, M., Kajiura, Y., Ito, A. et Makino, S. (2006). Unsupervised language model adaptation based on automatic text collection from WWW. *In Proceedings of Interspeech*, pages 2202–2205.

Tür, G. et Stolcke, A. (2007). Unsupervised language model adaptation for meeting recognition. *In Proceedings of ICASSP*, pages 173–176.

Wan, V. et Hain, T. (2006). Strategies for language model Web-data collection. *In Proceedings of ICASSP*, volume 1, pages 1520–6149.

Zhu, X. et Rosenfeld, R. (2001). Improving trigram language modeling with the World Wide Web. *In Proceedings of ICASSP*, volume 1, pages 533–536.