Vers une adaptation thématique non supervisée de modèles de langage : utilisation d'Internet comme un corpus ouvert

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

Since language models (LM) of automatic speech recognition systems are usually trained on multi-topic corpora, topic adaptation has been shown to be an effective way to improve the recognition accuracy, especially for broadcast news. This paper presents a new complete and unsupervised technique using information retrieval methods and based on the use of the Internet to retrieve thematically coherent corpora from which adapted LMs are trained. Experimental results demonstrate the validity of the proposed adaptation method with significant perplexity and word error rate reductions, and also show that topic adaptation should be included early in the recognition process.
Keywords: automatic speech recognition, language model adaptation, Web-based corpora

1. Introduction

Nombre de systèmes de reconnaissance automatique de la parole (RAP) se basent sur un modèle de langage (ML) à séquences de mots (n-grammes) appris sur une vaste collection de textes aux sujets variés. Ces ML synthétisent ainsi une bonne connaissance générale de la langue et permettent déjà d'obtenir des performances globalement satisfaisantes. Cependant, ils restent peu adaptés pour transcrire des documents très marqués thématiquement. En effet, il semble assez évident que des documents sonores abordant des sujets différents, par exemple le commerce de la drogue en Colombie et un concours de beauté en France, ne doivent pas être traités de la même manière. Ces sujets ont un vocabulaire et un emploi de ce vocabulaire qui leur sont propres. Pour résoudre ce problème, l'optique visée par plusieurs travaux, dont le nôtre, consiste justement à modifier le vocabulaire du système de RAP ou à adapter son ML en fonction de chaque sujet successivement abordé dans le document à transcrire. Dans cet article, nous nous intéressons uniquement au problème de l'adaptation du ML et laissons de côté la difficile question du vocabulaire.

Dans la majeure partie des travaux de ce domaine particulier, l'adaptation thématique d'un ML général consiste à mélanger celui-ci avec un ML spécialisé pour un thème précis. Pour cela, deux approches existent : l'une, supervisée, vise à choisir un ou plusieurs modèles parmi un ensemble de ML spécialisés *a priori* [6], alors que l'autre, non supervisée, cherche à apprendre un ML spécialisé à partir d'un corpus thématique construit dynamiquement [4]. Nous nous situons, pour notre part, dans la seconde vague.

Plus particulièrement, notre méthode d'adaptation vise à construire des corpora thématiques en collectant des textes à partir d'Internet. Outre la bonne modélisation de l'oral qu'offre cette ressource linguistique [8], la nature ouverte d'Internet présente l'intérêt majeur de lever toute restriction sur le nombre et la nature des sujets pour lesquels une adaptation peut être effectuée, ce qui n'est pas le cas lorsque l'on utilise des collections statiques de textes, comme dans [1, 4]. En pratique, notre travail s'intéresse à l'adaptation d'un ML pour des segments de transcription thématiquement cohérents. Ces segments peuvent provenir soit d'un long flux multimédia segmenté thématiquement, comme dans [1], soit de documents plus courts traitant d'un unique sujet, par exemple des *podcasts*.

Parmi les travaux proches des nôtres, [5] obtient des gains intéressants sur le taux de mots erronés (WER[1]) en collectant des pages Web sur la base de critères propres à la théorie de l'information et à la recherche d'information (RI). Cependant, pour un thème donné, la technique d'adaptation retenue se base sur un petit corpus spécialisé *a priori* pour ce même thème. [7] propose quant à lui une approche non supervisée utilisant Internet tant pour adapter un ML que pour enrichir le vocabulaire d'un système de RAP, ce qui interdit toute conclusion concernant l'intérêt des données récoltées pour la seule adaptation du ML. Plus généralement, les travaux connexes s'intéressant à Internet utilisent souvent plusieurs milliers de pages Web par adaptation et présentent des expériences sur un faible nombre de thèmes ou de segments de transcription[2]. Notre technique d'adaptation produit de bons résultats alors qu'elle ne manipule que quelques centaines de pages Web. De plus, elle est testée sur un grand nombre de segments de transcription dont les thèmes sont variés.

Dans cet article, la section 2 détaille les étapes-clés de notre technique d'adaptation, puis la section 3 présente nos résultats expérimentaux.

2. Méthode d'adaptation

Notre méthode d'adaptation thématique est présentée en figure 1. Étant donné un segment de transcription

[1] *Word Error Rate.*

[2] Dans [5], seul le thème des soins médicaux est traité et, dans [7], les expériences sont effectuées sur seulement 5 segments de transcription.

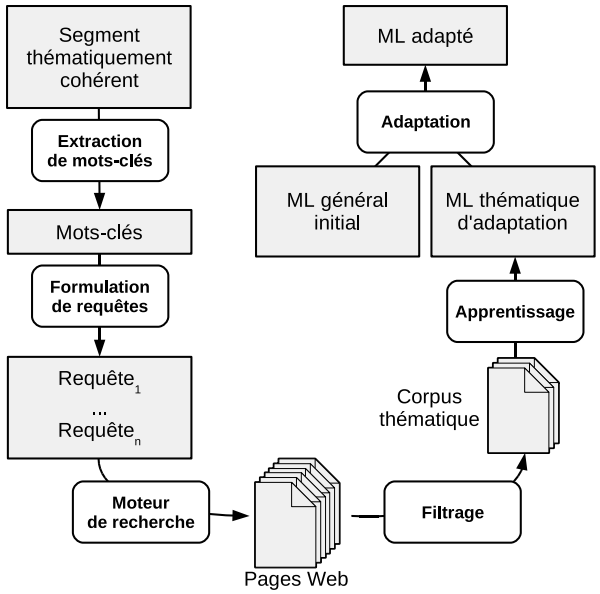

Fig. 1: Vue d'ensemble du processus d'adaptation.

thématiquement homogène obtenu à partir du ML général, des mots-clés sont extraits afin de caractériser le sujet abordé. Ces termes sont alors utilisés pour formuler et soumettre des requêtes à un moteur de recherche (*Yahoo !*). Les pages Web retournées sont ensuite parcourues de manière à construire un corpus d'adaptation à partir duquel un ML thématiquement spécialisé est appris. Celui-ci est alors combiné avec le ML général, le ML adapté ainsi obtenu étant utilisé pour produire une nouvelle transcription du segment.

De nombreuses questions émergent de cette chaîne. Comment extraire des mots-clés qui caractérisent suffisamment tous les aspects d'un segment mais qui, au sein d'une requête, restent suffisamment généraux pour aboutir à un nombre conséquent de pages Web ? Comment gérer le cas d'éventuelles erreurs de transcription ? Comment assembler les différents mots-clés pour former une ou plusieurs requêtes ? Il faut également s'interroger sur la manière de sélectionner les pages pertinentes parmi toutes celles récupérées pour obtenir un corpus d'adaptation suffisamment grand et thématiquement homogène. Enfin, la combinaison du ML spécialisé avec le ML général est également sujette à réflexion. Toutefois, cet article étant axé sur la faisabilité de l'approche proposée, nous avons choisi d'utiliser une technique d'interpolation linéaire, méthode simple même si loin d'être optimale. Dans la suite de cette section, nous présentons nos réponses à chacune des autres questions.

2.1. Extraction de mots-clés

Pour un segment de transcription donné, nous avons choisi de baser l'extraction de mots-clés sur un critère largement utilisé en RI : le *tf*idf*. Ce dernier consiste à attribuer un score à chaque mot d'un texte en fonction de sa fréquence dans le texte et par rapport à sa fréquence dans les documents d'un corpus de référence $\mathcal{C}$[3]. Les mots ayant les plus hauts scores sont alors considérés comme les plus discriminants. En pratique, nous normalisons additionnellement ces scores de manière à manipuler des valeurs entre 0 et 1. De plus, nous ne mesurons pas les scores *tf*idf* directement sur les mots du texte mais sur leurs lemmes, *i.e.*, leur forme canonique obtenue en ramenant, par exemple, les noms féminins pluriels à leur forme masculin singulier et les verbes conjugués à leur forme infinitive. Les mots de même lemme sont regroupés au sein d'une même classe ℓ pour laquelle un score $S(\ell)$ est calculé. Toutefois, étant donné que nous souhaitons former des requêtes à partir de mots, chaque classe est représentée par son mot le plus fréquent.

Enfin, les scores $S(\ell)$ sont modifiés afin de prendre en compte les spécificités d'un document transcrit. Premièrement, certains noms propres d'un segment n'apparaissent pas dans le corpus de référence $\mathcal{C}$ et sont ainsi saillants au regard du critère *tf*idf*. Or, inclure ces termes dans des requêtes a tendance à aboutir à des corpora trop spécifiques et trop petits pour être utilisables. Nous appliquons donc une pénalité à chaque nom propre en multipliant son score par un coefficient inférieur à 1, empiriquement fixé à $0,75$. Notons que cette pénalité ne doit pas être trop forte car certains noms propres peuvent aussi être très représentatifs pour des domaines donnés. Dans nos expériences, les noms propres sont détectés grâce à un dictionnaire des noms communs et un étiquetage morpho-syntaxique qui fournit la catégorie grammaticale de chaque mot en fonction de son contexte d'utilisation. Deuxièmement, la présence éventuelle d'erreurs de transcription peut biaiser le calcul du score de certaines classes de mots et impliquer la sélection de documents non pertinents dans le corpus d'adaptation. Contrairement à [7] où aucune stratégie n'est proposée pour pallier ce problème, nous modifions les scores *tf*idf* initiaux en fonction des mesures de confiance que le système de RAP fournit pour chaque terme transcrit. Le nouveau score d'un lemme ℓ est défini par

$$\sigma(\ell) = \left[\alpha + (1-\alpha)\, c_\ell\right] S(\ell) \quad \text{avec} \quad c_\ell = \frac{1}{|\ell|} \sum_{w \in \ell} c_w \quad (1)$$

où c_w est la mesure de confiance moyenne de toutes les occurrences d'un mot w et $|\ell|$ est le nombre de mots dans la classe ℓ. Le paramètre α, fixé empiriquement à $0,25$, limite l'impact des mesures de confiance car celles-ci ne sont pas toujours fiables [3].

2.2. Formulation de requêtes

Sur la base de l'équation 1, nous obtenons une liste triée de mots-clés à partir desquels nous souhaitons construire une ou plusieurs requêtes. Comme mentionné dans [5], des requêtes comportant trop de mots-clés ne retournent pas assez de résultats pour construire un corpus d'adaptation. Nous considérons donc uniquement les 5 premiers mots-clés extraits et construisons des requêtes simples, formées d'un sous-ensemble de 2 ou 3 mots-clés parmi ces 5 mots-clés conservés. Par exemple, une première requête est constituée des deux meilleurs mots-clés alors qu'une autre est constituée des premier et troisième mots-clés. L'étude de l'utilisation successive de 1, 3, 5 et 15 requêtes simples a montré que la perplexité du ML adapté diminue d'autant plus que le nombre de re-

[3] 800 000 articles du journal Le Monde, 1987–2003.

quêtes est grand : une amélioration relative de 30 % de la diminution de la perplexité est obtenue avec 15 requêtes. Ce constat peut s'expliquer par le fait que cette stratégie maximise les chances d'avoir au moins une requête pertinente même si certains mots-clés sont non pertinents. Toutefois, cette analyse pourrait être affinée par des expériences complémentaires basées sur l'emploi de plus de 15 requêtes ainsi que sur l'optimisation des requêtes entre elles, tâche d'autant plus difficile que le fonctionnement précis des moteurs de recherche sur Internet reste pour le moins opaque.

2.3. Sélection des pages Web

L'utilisation de plusieurs requêtes simples se traduit par un grand nombre de résultats, fréquemment plusieurs millions, parmi lesquels seule une fraction est pertinente. Aussi est-il nécessaire d'adopter une stratégie de filtrage vis-à-vis du nombre et de la qualité des pages à conserver. Dans des tests préliminaires, nous avons mesuré les variations de la perplexité et du WER en fonction du nombre de pages Web conservées. Il apparaît qu'il faut entre 50 et 400 pages pour obtenir un bon corpus d'adaptation. Dans nos expériences, 200 pages sont conservées, ce qui correspond en moyenne à 240 000 mots[4], chiffres très inférieurs à ceux présentés dans d'autres travaux [5, 7]. Ces pages sont alors filtrées en fonction de leur similarité thématique avec le segment t pour lequel l'adaptation est en cours. En considérant p, le contenu d'une page, et t comme des vecteurs de scores, cette similarité $\mathrm{sim}(t,p)$ est calculée par une distance *cosinus* :

$$\mathrm{sim}(t,p) = \frac{\sum_{\ell \in t \cap p} \sigma_t(\ell) \times S_p(\ell)}{\sqrt{\sum_{\ell \in t} \sigma_t(\ell)^2 \times \sum_{\ell \in p} S_p(\ell)^2}} \qquad (2)$$

où $\sigma_t(\ell)$ et $S_p(\ell)$ sont les scores de la classe ℓ respectivement dans t et p. Les pages dont la similarité est inférieure à un seuil donné sont écartées du corpus d'adaptation. Les résultats de nos tests préliminaires sur le rapport entre le seuil de similarité et les gains en perplexité et WER mettent en évidence l'importance de ne pas choisir un seuil trop élevé car cela aboutit à des corpora d'adaptation trop petits pour estimer des probabilités n-grammes de manière fiable. Le réglage selectionné pour nos expériences finales est $0,08$.

2.4. Mélange des modèles de langage

L'interpolation linéaire du ML général avec un modèle d'adaptation est régie par un facteur qui permet d'attacher plus ou moins d'importance au ML général. La figure 2 présente la variation de la diminution de la perplexité en fonction du coefficient d'interpolation pour deux valeurs du seuil de similarité sur un petit ensemble de segments de transcription. Les résultats obtenus avec un coefficient d'interpolation constant quel que soit le segment de transcription (lignes continues) montrent clairement que le choix de ce facteur est crucial pour le bon déroulement de l'adaptation thématique. Une variation de $0,1$ de ce paramètre peut se traduire par une variation relative

[4]Nombre de mots après nettoyage des pages Web, *i.e.*, élimination des balises HTML et suppression automatique des zones correspondant à des publicités, mentions légales, menus...

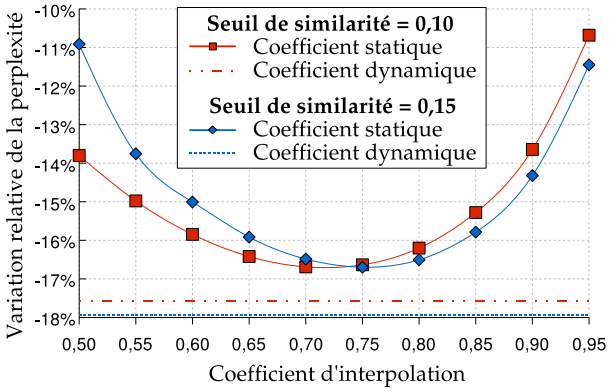

Fig. 2: Impact du coefficient d'interpolation sur la perplexité selon qu'il soit constant pour tout segment (lignes continues) ou fixé de manière optimale pour chaque segment.

de 20 % de la diminution de la perplexité. Cependant, la valeur optimale de ce coefficient dépend largement de la qualité du corpus d'adaptation et devrait être choisie séparément pour chaque segment. Dans une telle hypothèse, un gain relatif de 18 % de la diminution de la perplexité pourrait encore être obtenu. Toutefois, dans nos expériences, un coefficient d'interpolation fixé à 0,8 est utilisé quel que soit le segment.

3. Expériences et résultats

Nos expériences sont menées sur 172 segments issus de 6h d'émissions d'actualités du corpus ESTER [2]. 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). Cette collection est répartie entre un ensemble de développement (91 segments) et un ensemble de test (81 segments). Pour chaque segment, un ML adapté est appris selon le processus décrit en section 2.

Notre système de RAP, brièvement décrit dans [3], est un système multi-passes basé sur un ML 4-grammes général couvrant un vocabulaire de 64 000 mots. Deux passes nous intéressent principalement. Une première décode le signal à partir du ML général et d'un modèle acoustique général. Des graphes de mots comportant beaucoup d'hypothèses sont alors générés. Une seconde passe vise à adapter le modèle acoustique à chaque locuteur et à élaguer les graphes. Il en découle des graphes de mots beaucoup plus petits.

Pour obtenir de nouvelles transcriptions tirant parti de l'adaptation thématique, nous réévaluons, dans un premier cas, les graphes de mots issus de la première passe et, dans un second cas, ceux issus de la seconde passe en remplaçant le ML initial par celui adapté. Les WER mesurés dans ces 2 expériences sont comparés à ceux obtenus avant adaptation (tableau 1). Les résultats présentés montrent globalement une diminution du WER. Toutefois, les gains observés sur le corpus de test sont moins bons que ceux obtenus sur le corpus de développement. Cette différence est d'autant plus surprenante que les données de ces ensembles sont très proches tant sur le plan des thèmes que de la

Tab. 1: WER obtenus sans adaptation thématique et avec adaptation après les 1^{re} et 2^{nde} passes.

	Dev	Test
Sans adaptation	22,40	21,66
Adaptation après la 1^{re} passe	21,94 (-0,46)	21,49 (-0,17)
Adaptation après la 2^{nde} passe	22,02 (-0,38)	21,65 (-0,01)

Tab. 2: LER obtenus sans adaptation thématique et avec adaptation après les 1^{re} et 2^{nde} passes.

	Dev	Test
Sans adaptation	20,11	19.61
Adaptation après la 1^{re} passe	19,26 (-0,85)	19,11 (-0,50)
Adaptation après la 2^{nde} passe	19,19 (-0,92)	20,00 (+0,39)

période. Toujours est-il que ces gains sont cohérents avec les baisses de perplexité de 16,7 %, sur le corpus de développement, et de 14,5 %, sur le corpus de test, mesurées sur les ML adaptés par rapport au ML général. Une lecture verticale des résultats montre que les WER des transcriptions obtenues à partir des graphes de mots comportant le plus d'hypothèses sont nettement meilleurs que ceux obtenus à partir des graphes de mots plus petits. Ceci souligne la nécessité d'intégrer l'adaptation thématique relativement tôt dans le processus de reconnaissance.

Une analyse plus appronfondie des résultats montre que notre méthode permet principalement une amélioration de la transcription des mots « thématiques », *i.e.*, les mots ayant trait au sujet d'un segment, au détriment des mots grammaticaux, comme les prépositions ou les auxiliaires. Par exemple, dans un reportage traitant des maladies infectieuses, le groupe de souffles « *de la gorge et des bronches* » est transcrit, avec le ML général, par « *l'accord de branche* » puis, avec le ML adapté, par « *la gorge les bronches* ». Dans cet exemple, l'hypothèse de départ est grammaticalement correcte mais thématiquement incohérente et inversement pour la transcription après adaptation thématique. Plus généralement, le tableau 2 présente les différents taux d'erreurs mesurés sur les mots pleins (LER[5]), *i.e.*, sur les transcriptions lemmatisées et débarrassées entre autres des auxiliaires, des verbes de modalité et des prépositions. Ces résultats corroborent l'exemple précedemment cité : les gains obtenus sur les mots pleins sont environ deux fois supérieurs à ceux mesurés sur les transcriptions brutes, mis à part pour les graphes élagués de l'ensemble de test. Dans ce dernier cas, l'augmentation franche du LER vient encore souligner la nécessité d'utiliser les graphes de mots de la première passe. Par ailleurs, pour atteindre le niveau d'amélioration du LER en terme de WER, ajouter une phase de traitement morpho-syntaxique en fin de processus de reconnaissance, comme présenté dans [3], semble être une solution intéressante.

4. Conclusion et perspectives

Dans cet article, nous avons présenté une méthode d'adaptation thématique d'un ML général pour des segments de transcriptions thématiquement cohérents. Cette technique présente un emploi original de techniques de RI pour manipuler de manière non supervisée le délicat concept de thème. Plus particulièrement, nous mettons en avant une méthode de collecte de textes à partir d'Internet qui se base sur la notion de similarité thématique. Nos résultats présentent des gains en perplexité et en WER malgré des corpora de taille moyenne et le recours à des interpolations linéaires. Ces résultats soulignent plus particulièrement l'intérêt d'intégrer l'adaptation thématique en amont du processus de reconnaissance et de coupler celle-ci avec un post-traitement visant à améliorer la grammaticalité des transcriptions.

Ces conclusions présagent la suite de nos travaux sur l'adaptation thématique d'un ML. D'une part, la technique d'interpolation linéaire, tributaire de son coefficient d'interpolation, devrait être remplacée par une méthode plus fiable : l'adaptation MDI, qui se base sur des rapports de probabilités unigrammes, semble être une bonne alternative. D'autre part, puisque l'emploi des ML adaptés doit se faire tôt lors de la reconnaissance, il serait bon d'amorcer l'adaptation thématique au niveau des graphes de mots et non plus au niveau d'une transcription issue d'une première exécution complète du processus de RAP.

[5] *Lemmas Error Rate.*

Références

[1] L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda. Unsupervised language model adaptation for broadcast news. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 220–223, 2003.

[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 *Proc. Interspeech*, pages 1149–1152, 2005.

[3] S. Huet, G. Gravier, and P. Sébillot. Morpho-syntactic processing of N-best lists for improved recognition and confidence measure computation. In *Proc. Interspeech*, pages 1741–1744, 2007.

[4] D. Klakow. Selecting articles from the language model training corpus. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 3, pages 1695–1698, 2000.

[5] A. Sethy, P. G. Georgiou, and S. Narayanan. Building topic specific language models from webdata using competitive models. In *Proc. Interspeech*, pages 1293–1296, 2005.

[6] K. Seymore and R. Rosenfeld. Using story topics for language model adaptation. In *Proc. Eurospeech*, pages 1987–1990, 1997.

[7] M. Suzuki, Y. Kajiura, A. Ito, and S. Makino. Unsupervised language model adaptation based on automatic text collection from WWW. In *Proc. Interspeech*, pages 2202–2205, 2006.

[8] D. Vaufreydaz. *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue.* PhD thesis, Université Joseph Fourier, Grenoble, 2002.