Décodage guidé par un modèle cache sémantique

Benjamin Lecouteux, Pascal Nocera, Georges Linarès

Décodage guidé par un modèle cache sémantique

Benjamin Lecouteux, Pascal Nocera, Georges Linarès

LIA-CERI, université d'Avignon (France)

ABSTRACT

This paper proposes an improved semantic based cache model: our method boils down to using the first pass of the automatic speech recognition (ASR) system, associated to confidence scores and semantic fields, for driving the second pass. We use a Driven Decoding Algorithm (DDA), which allows us to combine ASR systems, by guiding the search algorithm of a primary system with an auxiliary system. We propose a strategy that uses DDA to drive a semantic cache, according to the confidence measures. The method works like an unsupervised language model adaptation. Results show, on 8 hours, that semantic-DDA yields significant improvements to the baseline: we obtain a 4% word error rate relative improvement without acoustic adaptation, and 1.9% after adaptation.

Keywords: Speech recognition, driven decoding, Latent Semantic Analysis, cache model

1. Introduction

Bien que les modèles n-gramme aient démontré leur efficacité dans le cadre des systèmes de reconnaissance automatique de la parole (SRAP), ils sont limités dans les modélisations à long terme ou sémantiques. Quelques travaux ont adressé ces aspects, principalement en modifiant les probabilités n-gramme en fonction de dépendances distantes ou de thèmes spécifiques :

- Les modèles caches introduits par [1], augmentent la probabilité des mots apparus récemment. L'hypothèse initiale étant que si un mot particulier est utilisé, ce dernier a de fortes chances de réapparaître.
- Les modèles triggers, présentés par [2, 3], où le problème de dépendance sur le long terme est résolu en interpolant des n-grammes avec des paires de mots "déclencheurs" (*triggers*) sélectionnés en fonction de leur information mutuelle. Cette approche est une généralisation des modèles cache.
- Les mélanges de modèles à base de thèmes: [4, 5, 6] proposent des techniques pour mixer des modèles en fonction de thèmes spécifiques. Les données d'apprentissage sont partitionnées en ensembles de thèmes qui sont utilisés pour adapter le modèle. [7] a introduit une méthode pour construire des modèles de langage en exploitant à la fois des contraintes locales et globales basées sur une analyse sémantique LSA (Latent Semantic Analysis). Cette approche propose d'estimer deux modèles, l'un à base de n-grammes, l'autre basé sur LSA. Ces modèles sont combinés, afin d'introduire l'information sémantique au sein du modèle n-gramme.
- La combinaison entre des mélanges de modèles et un modèle cache proposée par [8], combine ces deux approches pour capturer des dépendances éloignées au sein du langage.

Les modèles cache manquent de robustesse car les mots mis en avant dépendent de l'hypothèse courante : une erreur peut facilement se propager. De plus, ces modèles se basent uniquement sur le passé (l'historique de l'hypothèse courante). Le mélange de modèles à base de thèmes estime généralement des poids sur une première transcription ou sur l'hypothèse courante, sans prise en compte des erreurs potentielles. De plus, le principal problème de ces modèles est la sélection/détection des thèmes destinés à l'apprentissage.

Dans les travaux cités, des informations telles que les mesures de confiance produites par le SRAP ne sont pas utilisées. Elles sont généralement exploitées pour l'adaptation non supervisée des modèles acoustiques et rarement pour l'adaptation du modèle de langage : dans [9], les scores de confiance associés aux mots sont utilisés directement dans le graphe d'exploration, améliorant ainsi le décodage.

Notre objectif est d'exploiter toute l'information issue d'une première passe au cours de la seconde, afin de d'obtenir un modèle de langage adapté automatiquement. Nous proposons d'appliquer un modèle cache durant le processus de décodage, qui exploite les informations sémantiques et les mesures de confiance issues de la passe précédente.

Cet article présente une méthode intégrée permettant de diriger un SRAP avec son hypothèse précédente, associée à des mesures de confiance ainsi qu'à un modèle cache sémantique. Notre stratégie se focalise uniquement sur les mots mal reconnus, afin de réduire le bruit introduit par le modèle cache. Récemment, dans [10], nous avons proposé un algorithme qui permet d'introduire la sortie d'un SRAP auxiliaire au sein de l'algorithme de recherche. Nous présentons une extension de cet algorithme dédiée à l'adaptation non supervisée d'un SRAP. La première section présente l'ensemble du système. La seconde présente le protocole expérimental sur 8 heures extraites de la campagne ESTER [11]. La dernière section présente les expériences liées à notre décodage sémantiquement guidé. Finalement, nous concluons et suggérons quelques améliorations.

2. Approche intégrée : Le décodage guidé dédié à un cache sémantique

Le décodage guidé (Driven Decoding Algorithm, DDA) initial consiste à intégrer la sortie d'un

système auxiliaire dans l'algorithme de décodage d'un système primaire. Nous proposons de modifier l'algorithme pour obtenir un modèle cache amélioré. Les prochaines sous-sections présentent les différentes parties du système.

2.1. L'algorithme A* du système Speeral

Le SRAP du LIA est utilisé comme système primaire. Il est basé sur un algorithme de recherche A* opérant sur un treillis de phonèmes. Le processus de décodage repose sur une fonction d'estimation $F(h_n)$ qui évalue la probabilité de l'hypothèse h_n passant par le noeud n:

$$F(h_n) = g(h_n) + p(h_n), \quad (1)$$

Où $g(h_n)$ est la probabilité de l'hypothèse partielle au noeud n, qui résulte de l'exploration partielle du graphe. $p(h_n)$ est une sonde qui estime la probabilité restante entre le noeud n et le noeud final. Afin d'intégrer l'information issue du système auxiliaire, la partie linguistique de g dans (1) est modifiée en fonction de l'hypothèse auxiliaire, comme décrit ci-après.

2.2. Le décodage guidé

Le SRAP Speeral génère des hypothèses au fur et à mesure de l'exploration du treillis de phonèmes. La meilleure hypothèse à un temps t est étendue en fonction de la probabilité de l'hypothèse courante et du résultat de la sonde. Afin de combiner l'information issue de la transcription auxiliaire H_{aux} avec le processus de recherche, un point de synchronisation doit être trouvé pour chaque mot que le système évalue. Ces points sont trouvés en alignant dynamiquement la transcription fournie avec l'hypothèse courante; cette tâche est effectuée en minimisant la distance d'édition entre les deux hypothèses. Ce processus permet d'identifier dans la transcription auxiliaire H_{aux}, la meilleure sous-séquence qui correspond à l'hypothèse courante h_{cur}. Cette sous-séquence h_{aux} est utilisée pour une ré-estimation du score linguistique en fonction des probabilités *a posteriori* $\phi(w_i)$:

$$L(w_i|w_{i-2}, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})^{1-\beta}.\phi(w_i)^{\beta}$$
$$\beta = 0 \text{ si } w_i \text{ n'est pas trouvé dans } H_{aux} \quad (2)$$

Où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ est la probabilité initiale du trigramme, β est un facteur d'échelle estimé empiriquement et $\phi(w_i)$ est le score de confiance du mot w_i.

2.3. Les mesures de confiance

Les mesures de confiance sont estimées en deux étapes. La première extrait des paramètres de bas niveau relatifs à l'acoustique et à la topologie du graphe, ainsi que des paramètres plus haut niveau liés à la linguistique. A partir de ces paramètres, un classifieur basé sur l'algorithme boosting assigne à chaque mot une probabilité d'être correct, comme détaillé dans [12]. Chaque mot de l'hypothèse est au final représenté par un vecteur de 23 paramètres, qui se regroupent en 3 classes :

- Les **paramètres acoustiques** tels que la log-vraisemblance acoustique du mot et la log-vraisemblance moyenne par trame.
- Les **paramètres linguistiques** sont basés sur les probabilités estimées par le modèle de langage 3-gramme utilisé dans le SRAP. Nous utilisons la probabilité 3-gramme, la perplexité du mot dans une fenêtre définie et la probabilité unigramme. Nous ajoutons un index représentant le repli actuel du mot au niveau du modèle de langage.
- **les paramètres liés au graphe** se basent sur l'analyse du mot dans le réseau de confusion. Nous utilisons le nombre de chemins alternatifs ainsi que la probabilité *a posteriori*. Nous incluons également des valeurs relatives à la distribution des probabilités *a posteriori* dans le réseau de confusion.

Le classifieur a été entraîné sur un corpus annoté en mots décodés correctement ou non. Le taux d'erreur confiance (CER) est de 19.5% sur le corpus de développement et 18.6% sur le corpus de test pour un seuil de 0.5. L'entropie normalisée croisée est quant à elle de 0.373 sur le corpus de développement et 0.282 sur le corpus de test. Un seuil de 0.85 a été choisi pour maximiser la confiance de décision dans le module sémantique : 55% des mots sont sélectionnés comme corrects avec seulement 2.7% d'erreurs.

2.4. Le module d'analyse sémantique

L'analyse sémantique latente (LSA) [7] est une technique permettant d'associer des mots qui sont corrélés sémantiquement à travers plusieurs documents. L'hypothèse formulée est que les mots co-occurrents dans un même document sont sémantiquement corrélés.

Dans notre système, une séquence de mot sémantiquement pertinente peut être considérée comme incohérente par le modèle de langage du SRAP en raison de la limite du modèle de langage n-gramme. Pour cette raison, nous ajoutons un estimateur de consistance sémantique qui permet de valider ou rejeter certaines hypothèses.

Dans nos expériences, le module LSA a été entraîné sur les données d'apprentissage du modèle de langage. Pour une meilleure couverture, le corpus a été lemmatisé et le vocabulaire réduit au lexique lemmatisé (environ 33K mots). De plus, une stop-liste a été appliquée pour filtrer les mots non porteurs de sens.

Lorsqu'un mot est présenté au module, ce dernier retourne les 100 meilleurs mots associés avec leurs scores de confiance LSA.

2.5. Sélection de mots pour LSA

Les mots sont sélectionnés en fonction de leur score de confiance, afin de ne pas introduire de bruit dans le module sémantique. Le seuil a été fixé à 0.85 où 55% du corpus est sélectionné tandis que le taux d'erreur de sélection est de 2.7%. Pour chaque mot sélectionné, 100 mots sont extraits avec le module LSA, générant au final un groupe de mots destiné à notre modèle cache.

2.6. Décodage guidé avec LSA

L'utilisation du décodage guidé est indispensable dans ce contexte, car un simple model cache LSA introduirait trop de bruit. Ainsi, le système est dirigé par ses hypothèses précédentes pour limiter les déviations des mots corrects. Le *trigger* LSA est appliqué uniquement sur les mots à faible confiance (< 0.5) : les mots corrects sont ainsi préservés. De plus, les mots associés à de faibles mesures de confiance sont sous-évalués, permettant au SRAP d'explorer des chemins alternatifs.

Le système final tel que détaillé dans la figure 1 fonctionne comme un modèle cache amélioré. Le DDA-LSA devient :

$$\phi(w_i) \geq 0.5 : \begin{cases} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\beta}.\phi(w_i)^{\beta} \\ \beta = 0 \text{ si } \phi(w_i) \text{ est trouvé dans } H_{aux} \end{cases}$$

$$\phi(cw_i) < 0.5 : \begin{cases} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\alpha}.\theta(w_i)^{\alpha} \\ \alpha = 0 \text{ si } \theta(w_i) \text{ non trouvé} \end{cases} \quad (3)$$

Où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ est la probabilité initiale du trigramme, β et α sont des facteurs d'échelle calculés empiriquement, $\phi(w_i)$ est le score de confiance du mot w_i, cw_i est le mot aligné après l'historique (w_{i-2}, w_{i-1}) dans la transcription auxiliaire et $\theta(w_i)$ est le score LSA de w_i.

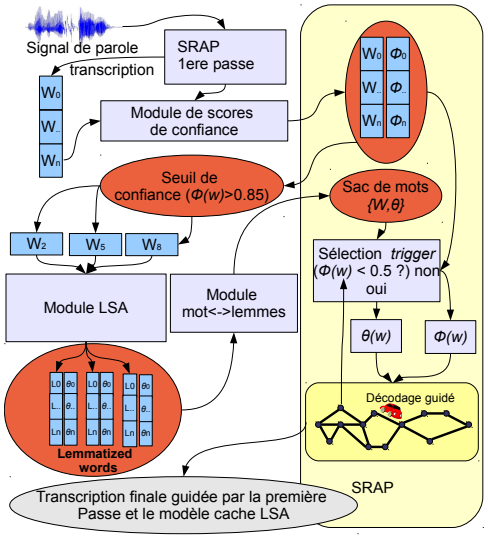

Figure 1: Principe du décodage guidé par LSA

2.7. *Le système de transcription du LIA*

Le système du LIA, Speeral, est un décodeur asynchrone opérant sur un treillis de phonèmes ; les modèles acoustiques utilisent des HMM et sont contextuels à base de tri-phones. Le modèle de langage est 3-gramme, estimé sur environ 200 millions de mots issus du journal *Le Monde* ainsi que du corpus ESTER (environ 1 million de mots). Le lexique est composé de 67000 mots. Dans ces expériences, une seule passe est effectuée en 3 fois le temps réel (3xRT).

2.8. *Le corpus ESTER*

Les expériences ont été effectuées sur les données issues de la campagne d'évaluation ESTER [11] : des émissions radio Marocaines et Françaises, des interviews, des locuteurs non natifs et des transcriptions à la volée. Les résultats sont issus de 8 heures extraites du test d'ESTER (sur 6 radios différentes). Les données d'apprentissage sont composées de 80 heures annotées manuellement, soit environ 1M mots ainsi que de 200M de mots issus du journal *Le Monde*.

3. Expériences

3.1. *Le SRAP guidé par sa 1ère passe*

Nous avons testé une seconde passe guidée par la première, associée à ses mesures de confiance. Le résultat n'a pas été surprenant : le SRAP converge sur la première passe et aucun changement significatif n'est observé. Ceci est du au fait qu'aucun chemin alternatif n'est proposé à l'algorithme d'exploration.

3.2. *Modèle cache sans contrainte*

Ces expériences testent l'utilisation d'un cache LSA sans contrainte. Pour chaque segment de parole, un ensemble de mots est associé via le module LSA. Nous avons implémenté un simple modèle cache : durant le processus d'exploration, chaque mot de l'hypothèse est recherché dans cet ensemble. Si le mot est trouvé, sa probabilité est augmentée en fonction de son score LSA. Les résultats sont peu intéressants, car une dégradation est observée, malgré l'information apportée par le LSA.

Cette dégradation est due à la quantité de mots dont la probabilité est modifiée : le bruit introduit impacte le résultat final. Cet aspect met en évidence la difficulté liée au réglage d'un modèle cache.

3.3. *Décodage guidé avec un cache LSA*

Ces expériences sont réalisées avec un système 3xRT, sans adaptation acoustique. Le SRAP utilise sa meilleure hypothèse précédente associée à ses scores de confiance. Le cache sémantique n'est appliqué qu'aux mots de faible confiance. Nous espérons guider le système avec les bons mots tout en réévaluant les moins bons avec le module LSA. Les résultats sont présentés dans le tableau 1. Le taux d'erreur mots (TEM) est réduit de 4% relatifs par rapport à la première passe : le cache LSA permet donc de se focaliser uniquement sur les erreurs. Cette stratégie réduit l'introduction de bruit.

Heure	P1 3xRT	P2 3xRT	P2-LSA DDA 3xRT
Classique 1h	21.4%	20.8%	20.9%
Culture 1h	34.0%	31.9%	33.3%
INTER 1h	22.7%	22.0%	21.6%
INFO 2h	25.8%	24.6%	25.0%
RFI 1h	28.6%	26.0%	27.1%
RTM 2h	35.4%	32.3%	33.6%

Table 1: DDA-LSA sans adaptation acoustique : *baseline* en première passe (P1 3xRT), *baseline* en seconde passe avec adaptation acoustique (P2 3xRT) et seconde passe avec DDA-LSA sans adaptation acoustique (P2-LSA DDA 3xRT)

3.4. *DDA-LSA avec adaptation mllr 3RT*

Ces expériences combinent l'adaptation acoustique par maximum de vraisemblance par régression linéaire (Maximum Likelihood Linear Regression, MLLR) pour la seconde passe (3xRT) avec le décodage guidé sémantique, afin de tester la complémentarité du DDA-LSA avec l'adaptation acoustique. Les résultats présentés dans le tableau 2 montrent un TEM réduit de 1.9% relatifs. Ceci montre la complémentarité avec le processus d'adaptation acoustique. Nous observons un meilleur gain sur la plus mauvaise heure (4.6% relatifs).

Heure	P2 3RT	P2-LSA DDA 3RT
Classique 1h	20.8 %	20,5 %
Culture 1h	31.9 %	31.8 %
INTER 1h	22.0 %	21.6 %
INFO 2h	24.6 %	24.5 %
RFI 1h	26.0 %	25.5 %
RTM 2h	32.3 %	30.8 %

Table 2: *baseline* en seconde passe 3xRT (P2 3xRT), décodage guidé par la sémantique avec adaptation acoustique en 3xRT (P2-LSA DDA 3xRT)

3.5. DDA-LSA avec adaptation mllr 10RT

Ces dernières expériences testent l'approche avec une exploration maximale du graphe de recherche au cours de la seconde passe en 10xRT. Les résultats sont présentés dans le tableau 3 : le TEM est réduit de seulement 1.1% relatifs ; la complémentarité entre l'adaptation acoustique et l'adaptation linguistique devient faible. Cependant, la plus mauvaise heure (RTM) présente une amélioration relative de 3.8%. Plus globalement, notre stratégie est intéressante dans le contexte d'un système 3xRT où les résultats convergent vers ceux d'un système 10xRT.

Heure	P2 10xRT	P2-LSA DDA 3RT	P2-LSA DDA 10RT
Classique 1h	20.2 %	20,5 %	20.0 %
Culture 1h	31.7 %	31.8 %	31.5 %
INTER 1h	21.6 %	21.6 %	21.6 %
INFO 2h	24.0 %	24.5 %	23.9 %
RFI 1h	25.4 %	25.5 %	25.3 %
RTM 2h	31.7 %	30.8 %	30.5 %

Table 3: *baselines* issues de la seconde passe après adaptation des modèles acoustiques en 10xRT (P2 10xRT), décodage guidé par la sémantique avec adaptation acoustique en 3xRT (P2-LSA DDA 3xRT) et décodage guidé par la sémantique avec adaptation acoustique en 10xRT (P2-LSA DDA 10xRT)

Le DDA-LSA améliore significativement le système 3xRT et plus particulièrement l'heure RTM. Avec le système 10xRT l'amélioration est faible, excepté sur l'heure RTM. Contrairement aux autres heures, RTM est une radio Marocaine, tandis que les données d'apprentissage du modèle de langage sont dérivées d'un journal Français. Il en résulte une couverture moins bonne pour RTM. Ceci explique les gains plus significatifs sur RTM et montre la contribution du modèle sémantique.

4. Conclusion

Nous avons proposé un décodage guidé par de l'information sémantique extraite de la première passe d'un SRAP. Notre stratégie s'est concentrée sur un modèle cache sémantique, qui s'applique en fonction du score de confiance de chaque mot.

Les expériences montrent que cet algorithme améliore le système initial et complète la phase d'adaptation acoustique. L'enrichissement de la première passe avec les scores de confiance est nécessaire pour orienter correctement l'algorithme de recherche, tandis que l'information sémantique permet de sélectionner des chemins alternatifs corrects quand les scores de confiance sont bas : cette stratégie s'assimile à une adaptation non-supervisée du modèle de langage. Un gain de 4% relatifs de TEM est obtenu sur la première passe sans adaptation acoustique, 1.9% sur le système 3xRT après adaptation acoustique et 1.1% sur le système 10xRT. La stratégie est plus intéressante dans le cadre d'un système 3xRT où le TEM est proche du système 10xRT, tout en réduisant les coûts de calcul. Malgré tout, les meilleurs résultats sont obtenus sur les heures les plus éloignées des données d'apprentissage.

Actuellement notre méthode est limitée à la granularité des segments. Nous souhaitons l'étendre à des ensembles de segments (discussion entre plusieurs locuteurs etc.). Nous envisageons également d'intégrer des données sémantiques externes telles que des résumés de documents audio ou des méta données (titres, locuteur, etc.).

Bibliographie

[1] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[2] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, vol. 10, no. 3, pp. 187–228, 1996.

[3] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modeling," in *Computer Speech and Language, vol. 8, pp. 1-38,*, 1994.

[4] Yoshihiko Gotoh and Steve Renals, "Topic-based mixture language modelling," *Natural Language Engineering*, vol. 5, pp. 355–375, 1999.

[5] N. Singh-Miller and C. Collins, "Trigger-based language modeling using a loss-sensitive perceptron algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, 15–20 April 2007, vol. 4, pp. 25–28.

[6] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here ?," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, Aug. 2000.

[7] J.R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[8] P.R. Clarkson and A.J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, 1997, vol. 2, pp. 799–802 vol.2.

[9] P. Fetter, F. Dandurand, and P. Regel-Brietzmann, "Word graph rescoring using confidence measures," in *Fourth International Conference on Spoken Language ICSLP*, 1996, vol. 1, pp. 10–13 vol.1.

[10] B. Lecouteux, G. Linares, Y. Esteve, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008*, 2008, pp. 1549–1552.

[11] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase 2 evaluation campaign for the rich transcription of french broadcast news," in *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[12] P. Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," in *Interspeech, Aalborg, Denmark*, 2001, pp. 2109–2112.