Amélioration des modèles de repli par des sacs de mots et des n-grammes à variables

Raphaël Rubino, Benjamin Lecouteux, Georges Linares

Amélioration des modèles de repli par des sacs de mots et des n-grammes à variables *

Raphaël Rubino[1], Benjamin Lecouteux[2], Georges Linarès[1]
[1]Laboratoire Informatique d'Avignon, [2]Laboratoire Informatique de Grenoble, équipe GETALP
raphael.rubino@univ-avignon.fr, benjamin.lecouteux@imag.fr,
georges.linares@univ-avignon.fr

RÉSUMÉ

Les modèles classiques de n-grammes manquent de robustesse sur évènements non observés. La littérature suggère des méthodes de lissage, la plus utilisée d'entre elles étant le Kneyser-Ney modifié. Nous proposons d'améliorer ce modèle en réordonnant les possibilités de replis par rapport à l'information mutuelle portée par les mots ; ainsi que par l'utilisation de n-grammes à variables. Nos résultats montrent un gain significatif par rapport un modèle Kneyser-Ney modifié : 0.6% de gain absolu sans adaptation des modèles acoustiques et 0.4% après adaptation.

ABSTRACT

Improving back-off models with bag of words and hollow-grams

Classical n-grams models lack robustness on unseen events. The literature suggests several smoothing methods : empirically, the most effective of these is the modified Kneser-Ney approach. We propose to improve this back-off model : our method boils down to back-off value reordering, according to the mutual information of the words, and to a new hollow-gram model. Results show that our back-off model yields significant improvements to the baseline, based on the modified Kneser-Ney back-off. We obtain a 0.6% absolute word error rate improvement without acoustic adaptation, and 0.4% after adaptation.

MOTS-CLÉS : modèles de langage, modèles de replis.

KEYWORDS: language model, low-order interpolation, back-off.

1 Introduction

Les modèles de langage (ML) à base de n-grammes ont démontré leur efficacité dans les systèmes de reconnaissance de la parole (SRAP), mais ils restent corrélés au taux d'erreur mot (TEM) dans le cas des évènements non observés (Berdy *et al.*, 1997). Les modèles de lissage des n-grammes furent introduits pour aborder ce problème. Ainsi sont apparus les modèles de repli redistribuant une masse de la probabilité des évènements vus pour les évènements peu ou non observés. Beaucoup de modèles ont été proposés : le lissage additif (Lidstone, 1920), le repli sur les n-grammes d'ordre inférieur (Katz, 1987) ou encore l'interpolation avec les n-grammes d'ordres inférieurs (Kneser et Ney, 1995)...

Dans (Chen et Goodman, 1999), une étude sur les méthodes de lissage est proposée. Il en ressort que le Kneyser-Ney (KN) modifié est la plus performante. Mais un problème demeure : tous les

*. Ce travail est financé par l'Agence Nationale pour la Recherche, projet ASH (ANR-09BLAN-0161-02)

modèles de repli surestiment certains évènements. (Rosenfeld et Huang, 1992) propose donc de les recalculer par rapport à certaines paires mots déclencheuses présentes dans le ML.

Dans (Schwenk et Gauvain, 2002), les modèles de repli sont calculés dans un espace continu, permettant un meilleur lissage. Cependant, les calculs et les optimisations nécessaires sont assez couteux.

D'autres approches se basent sur la similarité des mots. Elle est utilisée par (Rosenfeld, 1996) comme contrainte dans le cadre de l'apprentissage d'un modèle à maximum d'entropie, réduisant au final le TEM et la perplexité sur le corpus de test. Cependant ce modèle semble plus adapté pour capturer des dépendances linguistiques distantes que courtes.

Dans (Brown *et al.*, 1990), une mesure de similarité inter-mots est utilisée afin des les regrouper en classes. Une autre approche est le repli calculé via des classes hiérarchique de n-grammes (Zitouni, 2007). Ceci permet d'avoir de meilleures estimations pour les évènements non observés en se basant sur des informations syntaxiques ou sémantiques. Mais ces méthodes nécessitent des connaissances *a priori* sur les mots.

Par ailleurs, les travaux exposés ne prennent pas en considération les erreurs intrasèques d'un SRAP et ne dépendent pas du contexte. Par exemple avec un repli KN, les 3-grammes non vus se calculent ainsi : "explosion de pneu" $= \alpha(explosionde)p(pneu)$ et "explosion de peu" $= \alpha(explosionde)p(peu)$. Ces deux 3-grammes partagent la même valeur de repli $\alpha()$ dans des contextes différents. Pour être plus fin, il est nécessaire d'ordonner ces replis en fonction de leur historique court (3-gramme).

(Bilmes et Kirchhoff, 2003) introduisent les modèles de langage factorisés. Ils proposent un modèle de repli généralisé qui peut dépendre de chaque évènement observé et de ses attributs. Ce modèle permet de mettre en compétition plusieur replis basés sur des connaissances linguistiques, des critères statistiques... Nos travaux s'inspirent de cette approche dans le sens où plusieurs replis peuvent se concurrencer.

Cet article présente un modèle de repli simple basé soit sur les cooccurrences de mots soit sur des modèles n-grammes à variables. De plus, notre méthode s'applique facilement aux traditionnels modèles de langage n-grammes. La première section présente notre approche. La seconde présente le cadre expérimental. Quant aux deux dernières sections, elles présentent les expériences menées et leurs résultats. Finalement nous concluons et présentons quelques perspectives.

2 Approches proposées

(Berdy *et al.*, 1997) montre la corrélation entre le comportement des modèles de repli et le TEM. Le Tableau 1 contient le TEM en fonction des situations de repli rencontrées sur notre corpus d'apprentissage. La couverture du ML n'est pas l'unique problème, mais des expériences avec un ML biaisé par les données de test améliore grandement les résultats de la reconnaissance (**?**).

Actuellement, dans la majorité des SRAP état de l'art, le modèle de repli utilisé est basé sur le KN modifié. Cependant, ce dernier ne prend pas en compte les éléments du contexte court (comme vu dans l'exemple de l'introduction). Nous proposons une mesure alternative permettant de déterminer la vraisemblance d'apparition de combinaisons de mots.

Nous proposons ensuite de réévaluer les valeurs de repli par rapport à la possibilité d'apparition

des évènements : notre méthode est alors utilisée pour vérifier l'existence d'une séquence de mots. Dans le cas négatif, la valeur de repli est légèrement dégradée.

La valeur de repli est alors réestimée selon l'équation suivante :

$\tilde{\alpha}(w_{i-n},..,w_i) = \alpha(w_{i-n},..,w_{i-1})^{1-\beta} p_\phi(w_{i-n},w_i)^\beta$
Où $\tilde{\alpha}$ est la valeur réestimée du repli, α est la valeur initiale du repli, $p_\phi(w_i,w_{i-n})$ est la fonction de lissage basée sur la co-occurrence (section 4) ou les *n-grammes à variables* (section 3). $p_\phi(w_{i-n},w_i) = \delta$ si (w_i,w_{i-n}) la co-occurrence n'est pas observée. β est un facteur d'échelle calculé empiriquement et δ est une pénalité également empirique basée sur la possibilité binaire d'existence du n-gramme courant. β et α sont estimés via un simplex sur le corpus de développement. Nous introduisons le paradigme du n-gramme à variables dans la prochaine section.

Cas de repli	meilleure hypothèse
le 3-gramme existe	17.7%
repli sur le bigramme	28.5%
repli sur l'unigramme	50.0%

Émission	TEM	TEP	TCM
Inter (4h)	33,1	75,0	69,7
Tvme (1h)	31,3	67,6	71,2
Rfi (1h)	18,7	66,6	84,0

TABLE 1 – TEM en fonction du niveau de repli ainsi que les TEM, Taux d'Erreur Phrase (TEP) et Taux Mots Corrects (TMC) sur la référence.

3 Modèle de repli basé sur des n-grammes à variable

Dans le cas d'un modèle 3-gramme, la probabilité $p(w_3|w_1,w_2)$ d'une séquence non observée (w_1,w_2,w_3) est calculée avec la valeur de repli de (w_1,w_2) et la probabilité conditionnelle de $p(w_3|w_2)$. Cette méthode fait l'hypothèse d'indépendance de la co-occurrence de w_3 et w_1. Nous proposons un modèle n-gramme à variable $p_\phi(w_3|w_1)$ permettant d'introduire une hypothèse de dépendance à court terme sur des séquences non observées. Dans le cadre d'un modèle 3-gramme, la méthode consiste à estimer une modèle 2-gramme basé sur la paire de mots débutant et finissant tous les 3-grammes. Ce modèle peut être assimilé à une expression régulière : $(w_1,*,w_3)$ où $*$ est un mot non observé. Dans le cas d'un évènement non observé, le modèle se replie sur un 2-gramme utilisant la valeur de repli : $p(w_3|w_1,w_2) = \alpha(w_1,w_2)p(w_3|w_2)$. L'équation de repli devient :

$$\tilde{p}(w_3|w_1,w_2) = \alpha(w_1,w_2)^{1-\beta} p_\phi(w_3|w_1)^\beta p(w_3|w_2) \quad (1)$$

Où $\tilde{p}(w_3|w_1,w_2)$ est la probabilité recalculée du 3-gramme, $\alpha(w_1,w_2)$ est la valeur de repli initiale de la paire de mots (w_1,w_2), $p_\phi(w_3|w_1)$ la probabilité du n-gramme à variable, β est un facteur d'échelle. Cependant, dans notre modèle le repli dépend uniquement d'évènements observés. Nous généralisons ce modèle afin d'améliorer le repli pour les évènements non observés.

4 Modèle de repli utilisant la co-occurrence des mots

L'idée sous jacente est de combiner l'information mutuelle entre les mots avec les valeurs de repli classiques. Beaucoup de travaux ont proposé des mesures pour évaluer la co-occurrence des mots (Dagan *et al.*, 1999). Parmis ces travaux : l'information mutuelle, les probabilités conditionnelles ou des mesures statistiques telles que Chi^2. Dans cet article nous avons décidé d'utiliser un rapport de vraisemblance introduit par (Dunning, 1993). Cette mesure s'est avérée, dans la

pratique, très appropriée aux données incomplètes afin d'y trouver des paires de mots qui sont corrélés. La log-vraisemblance entre deux mots w_a et w_b est décrite dans l'équation 2.

$$\psi(w_a;w_b) = \sum_{ij\in\{1;2\}} log\frac{k_{ij}N}{C_iR_j} = k_{11}log\frac{k_{11}N}{C_1R_1} + k_{12}log\frac{k_{12}N}{C_1R_2} + k_{21}log\frac{k_{21}N}{C_2R_1} + k_{22}log\frac{k_{22}N}{C_2R_2} \quad (2)$$

Où $C_1 = k_{11} + k_{12}$, $C_2 = k_{21} + k_{22}$ et $R_1 = k_{11} + k_{21}$, $R_2 = k_{12} + k_{22}$
$N = k_{11} + k_{12} + k_{21} + k_{22}$ et k_{11} = le compte des co-occurrences des mots w_a et w_b
k_{12} = le compte du mot $w_a - k_{11}$ et k_{21} = le compte du mot $w_b - k_{11}$
k_{22} = le comptes de tous les mots dans le corpus $- k_{12} - k_{21} + k_{11}$

Afin de compter les co-occurrences de deux mots, nous utilisons une fenêtre glissante de taille s. Dans nos expériences, sa taille est de 5 mots. L'ordre des mots n'est pas pris en compte dans la fenêtre : elle peut être vue comme un sac de mots. Par ailleurs, nous pondérons le compte des co-occurrences de w_a et w_b par le nombre de mots qui les sépare dans la fenêtre : $\tilde{k}_{11} = \frac{1}{d(w_a;w_b)}k_{11}$. Le rapport de vraisemblance est calculé avec tout le lexique sur le corpus complet. Les valeurs sont normalisées afin d'en extraire des probabilités. Le modèle obtenu $p_\psi()$ est alors utilisé pour pondérer les valeurs de repli comme dans l'équation 1. Contrairement au n-gramme à variable, cette approche peut être appliquée à tous les ordres de repli et dans le cas d'un repli complet, la probabilité du 3-gramme devient :

$$\tilde{p}(w_3|w_1,w_2) = \alpha(w_1,w_2)^{1-\beta}p_\psi(w_1,w_3)^\beta\alpha(w_2)^{1-\beta}p_\psi(w_2,w_3)^\beta p(w_3) \quad (3)$$

Où $\alpha(w_1,w_2)$ est la probabilité de repli initiale des mots, β est un facteur d'échelle et $p_\psi()$ est la fonction de lissage basée sur la co-occurrence des mots, obtenue par l'équation 2.

5 Combinaison des deux approches

Le modèle de n-gramme à variable est incapable de fonctionner sur des valeurs non observées. Dans ces cas là nous proposons de nous replier sur le modèle de co-occurrences. Le modèle qui en découle est alors défini comme suit :

$$\tilde{\alpha}(w_1,w_2) = \begin{cases} \alpha(w_1,w_2)^{1-\beta}p_\phi(w_3|w_1)^\beta \text{ si le n-gramme à variable existe} \\ \alpha(w_1,w_2)^{1-\beta}p_\psi(w_1,w_3)^\beta \text{ sinon} \end{cases} \quad (4)$$

6 Modèle compact

Le modèle de co-occurrence initial nécessite beaucoup de mémoire ($\frac{|V|^2}{2}$ où V est la taille du vocabulaire). Nous proposons un modèle basé sur la possibilité binaire de la co-occurrence de mots. Si w_a et w_b ont été observés dans une fenêtre, le repli initial est utilisé tel quel sinon une pénalité est appliquée.

7 Cadre experimental

7.1 Le système de transcription du LIA

Les expériences présentées utilisent le SRAP du Laboratoire Informatique d'Avignon (LIA) entraîné sur les données de la campagne d'évaluation ESTER (Galliano *et al.*, 2005). Ce système repose sur

Émission	TEM	TEP	TMC
Inter (4h)	32,8 (- 0,3)	74,5 (- 0,5)	70,5 (+ 0,8)
Tvme (1h)	31,3 (+ 0,0)	67,0 (- 0,3)	71,7 (+ 0,5)
Rfi (1h)	18,5 (- 0,2)	65,7 (- 0,9)	84,4 (+ 0,4)
Moyenne	- 0,2	- 0,5	+ 0,7

Émission	TEM	TEP	TMC
Inter (4h)	32,7 (- 0,4)	74,5 (- 0,5)	70,5 (+ 0,8)
Tvme (1h)	31,0 (- 0,3)	66,8 (- 0,8)	71,8 (+ 0,6)
Rfi (1h)	18,4 (- 0,3)	65,6 (- 1,0)	84,5 (+ 0,5)
Moyenne	- 0,4	- 0,6	+ 0,7

TABLE 2 – TEM, TEP et TMC pour les replis sur n-gramme à variable et à base de co-occurrences.

Speeral (Nocera *et al.*, 2004), un décodeur A^* opérant sur un treillis de phonèmes. Les modèles acoustiques utilisent des MMC et sont contextuels à base de tri-phones. Le ML est 3-grammes et entraîné sur environ 1,3G mots du journal *Le Monde*, les corpus issus d'ESTER et de Gigaword. Le lexique contient 86K mots. Dans ces expériences, la première passe est effectuée en trois fois le temps réel (3xRT), suivie d'une seconde (5xRT) après adaptation acoustique par maximum de vraisemblance par régression linéaire (MLLR).

7.2 Le corpus ESTER-2

Le corpus ESTER-2 est composé d'émissions de radio francophones issues du groupe *Radio-France*. Le corpus nécessaire à l'entraînement et à l'optimisation du SRAP provient de la campagne d'évaluation ESTER-2 : 100 heures annotées manuellement. Notre approche est évaluée sur 6 heures de parole extraites des données de test ESTER-2. Le TEM moyen est de 32,7% lors de la première passe, puis de 27,3% après adaptation par MLLR.

Le Tableau 1 contient les résultats du système de référence selon trois critères d'évaluation : le taux d'erreur mot (TEM), le taux d'erreur phrase (TEP) et le taux de mots corrects (TMC). Ces trois mesures, permettent d'observer le comportement du SRAP. Ces résultats sont donc présentés séparément, tout comme les émissions de radio. Ces dernières permettent de tester le système sur différents types de données, faisant ainsi varier le nombre de locuteurs, leur niveau de spontanéité, etc.

8 Expériences

Pour toutes les expériences, nous effectuons les mesures présentées dans la section 7.2. En intégrant notre méthode de repli au ML, la dynamique de ce dernier est modifiée. Nous observons notamment une forte augmentation du nombre d'insertions. Les résultats en terme de TMC nous permettent alors de constater le nombre de mots corrects et l'impact réel de notre approche.

8.1 Repli basé sur des n-grammes à variable

Cette approche se base sur des 3-grammes comportant en leur centre un élément inconnu, noté $*$, et se présentant sous la forme $(w_1, *, w_3)$. L'ordre des mots composant ce motif est pris en compte, permettant d'utiliser cette approche comme une expression régulière pour capturer des occurrences dans le corpus d'entraînement. Les résultats des expériences utilisant le modèle n-gramme a variable sont présentés dans le Tableau 2 à gauche.

Ces résultats montrent une légère amélioration du TEM et du TMC en comparaison avec un repli classique. Le gain global sur le TMC est deux fois plus important que celui obtenu sur TEM. Cela indique que le modèle de repli utilisé corrige des mots et en introduit des nouveaux. Ces expériences montrent qu'une ré-estimation des valeurs de repli permet d'améliorer le comportement du ML. En pondérant ces valeurs il est alors possible de désambiguïser certaines hypothèses

Émission	WER	SER	CWR
Inter (4h)	32,7 (- 0,3)	74,8 (- 0,2)	70,3 (+ 0,6)
Tvme (1h)	31,1 (- 0,2)	67,5 (- 0,1)	71,6 (+ 0,3)
Rfi (1h)	18,6 (- 0,1)	65,7 (- 0,9)	84,2 (+ 0,2)
Moyenne	- 0,25	- 0,3	+ 0,5

Émission	WER	SER	CWR
Inter (4h)	32,4 (- 0,7)	74,6 (- 0,4)	70,8 (+ 1,1)
Tvme (1h)	30,9 (- 0,4)	67,2 (- 0,4)	72,0 (+ 0,8)
RFI (1h)	18,4 (- 0,3)	65,6 (- 1,0)	84,5 (+ 0,5)
Moyenne	- 0,6	- 0,5	+ 0,9

TABLE 3 – TEM, TEP et TMC avec le modèle compact (gauche) et le modèle combiné (droite).

produites par le SRAP. Cependant, le modèle présenté est limité par sa topologie, car seules les valeurs de repli des bigrammes sont repondérées. Nous proposons donc, dans la prochaine section, d'exploiter la co-occurrence des mots.

8.2 Modèle de repli basé sur la co-occurrence de mots

Dans ces expériences, les valeurs de repli sont combinées avec la probabilité de co-occurrence des mots. Afin de calculer les probabilités de co-occurrence, une matrice symétrique est construite (l'ordre des mots n'étant pas pris en compte) et la méthode présentée dans la section 4 est appliquée sur l'ensemble du corpus d'entraînement. Les valeurs de repli de tous les ordres sont concernées par la pondération. Les résultats obtenus sont présentés dans le Tableau 2 à droite.

Cette approche atteint des gains plus importants en terme de TEM et TEP, par rapport à l'approche précédente basée sur les n-grammes à variable. Ces résultats sont liés au fait que les valeurs de repli des unigrammes et des bigrammes sont réestimés. Il apparaît ici aussi, qu'un modèle simple permet d'améliorer les performances du SRAP en modifiant à peine le ML classique. Dans la prochaine section, nous combinons les deux approches présentées.

8.3 Combinaison des co-occurrences et des n-grammes à variable

Pour évaluer la complémentarité des approches présentées dans les sections 8.2 et 8.1, nous les combinons ainsi :
– si le motif $(w_1, *, w_3)$ existe, alors les n-grammes à variable sont utilisés,
– sinon, le modèle de co-occurrences est appliqué, comme présenté dans la section 4.
Les résultats de cette combinaison sont présentés dans le Tableau 3 à droite.

Ces résultats montrent la complémentarité des deux modèles, par les gains obtenus sur les 3 mesures. En combinant les deux approches, le modèle de repli capture plus d'informations sur le contexte des mots. Ces contextes sont de tailles variables et dépendent directement de la proximité des mots dans le corpus d'apprentissage. Cependant, le modèle baséesur les co-occurrences devient coûteux si tout le lexique est couvert. Nous proposons donc, un modèle compact basé sur la possibilité binaire de co-occurrences.

8.4 Un modèle de co-occurrences compact

Le modèle compact représente, la possibilité d'un repli selon la co-occurrence de deux mots. A partir de la matrice de co-occurrences initiale, toute valeur non nulle rend vraie la possibilité de co-occurrence. Nous reportons les résultats obtenus avec à le modèle compact dans le Tableau 3 à gauche.

Ces résultats montrent une dégradation des résultats par rapport aux approches présentées dans la section 8.3. Cependant, la taille du modèle représente la principale motivation de son

utilisation : un seul *bit* est utilisé par paire de mots pour représenter la possibilité d'un repli. Malgré tout, ce modèle permet l'amélioration d'un modèle de repli KN modifié.

9 Comportement du SRAP en seconde passe

Dans cette section, nous étudions le comportement de notre nouveau ML. Le Tableau 4 contient les taux d'insertions, substitutions et suppressions sur l'ensemble du corpus de test. Nous observons que les substitutions restent stables entre le modèle de référence et le modèle de combinaison proposé. Le taux d'insertions a fortement augmenté et le nombre de suppressions a chuté.

	insertions	suppressions	substitutions
référence	2,73	8,58	18,86
combinaison	3,14	7,57	18,98
gains	+ 15%	- 12%	+ 0.6%

TABLE 4 – Repli : comportement du SRAP en terme d'insertions, suppressions et substitutions.

D'une manière générale, ces résultats montrent que les gains selon le TMC sont deux fois supérieurs aux gains en TEM. Notre modèle insère donc une grande quantité de mots dans les transcriptions produites, ce qui pourrait être régulé par une stop-liste et une plus grande pénalité au niveau mot. Nous n'utilisons donc pas ce nouveau modèle de manière optimale.

Après avoir effectué l'adaptation acoustique par MLLR en utilisant la transcription issue de la première passe, nous proposons d'appliquer notre modèle de repli pendant la seconde passe. Les résultats sont présentés dans la Table 5 et indiquent des gains sur les trois ensembles de test utilisés. Cette expérience démontre l'impact de notre approche : une ré-estimation simple des valeurs de repli permet d'améliorer le modèle KN initial.

	TEM		TEP		TMC	
Émission	*référence*	*combinaison*	*référence*	*combinaison*	*référence*	*combinaison*
Inter (4h)	30,4	29,9	73,1	72,3	72,2	73,0
Tvme (1h)	25,3	24,8	62,8	62,5	77,8	78,6
Rfi (1h)	17,0	16,9	64,4	64,0	85,6	86,1
Moyenne	- 0,4		- 0,6		+ 0,7	

TABLE 5 – TEM, TEP et TMC après adaptation acoustique et combinaison des approches proposées.

10 Conclusion et perspectives

Dans cet article, nous introduisons un nouveau modèle de repli basé sur des n-grammes à variable et la co-occurrence des mots. Notre approche se combine facilement avec les modèles de langage classiques à base de n-grammes. Dans nos expériences, un modèle de repli KN modifié est interpolé avec nos deux approches.

Nous évaluons nos modèles sur 6h d'émissions radiophoniques Françaises. En ré-estimant uniquement les valeurs de repli, le nouveau modèle permet de gagner jusqu'à 0,6% de TEM absolus et 0,9% de TMC lors de la première passe de décodage. Après adaptation acoustique, les gains atteignent 0,4% de TEM absolu et 0,7% de TMC. Les meilleurs résultats sont obtenus en combinant les deux approches, co-occurrences et n-grammes à variable.

Nous avons aussi présenté une version compacte de notre modèle, permettant une faible consommation de mémoire : seul un *bit* représente la possibilité d'existence d'une paire de mots ou non.

Ces expériences préliminaires confirment que la prise en charge d'événements non vus n'est pas un problème résolu. Nous envisageons d'étendre notre modèle à des séquences de mots plus longues. Dans le cadre de notre modèle de co-occurrences, nous souhaitons étudier l'impact de différentes heuristiques introduisant des notions de distances entre les mots.

Références

BERDY, U., UHRIK, C. et WARD, W. (1997). Confidence metrics based on n-gram language model backoff behaviors. *In in Proc. EUROSPEECH*, pages 2771–2774.

BILMES, J. A. et KIRCHHOFF, K. (2003). Factored language models and generalized parallel backoff. *In in Proceedings of HLT/NACCL*, pages 4–6.

BROWN, P. F., PIETRA, V. J. D., DESOUZA, P. V., LAI, J. C. et MERCER, R. L. (1990). Class-based n-gram models of natural language. *Computational Linguistics*, 18:18–4.

CHEN, S. et GOODMAN, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer speech and language*, 13:359–394.

DAGAN, I., LEE, L. et PEREIRA, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *In Machine Learning*, pages 34–1.

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19:74.

GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K., BONASTRE, J.-F. et GRAVIER, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. *In Eurospeech*.

KATZ, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech, and Signal Processing*, 35:400–401.

KNESER, R. et NEY, H. (1995). Improved backing-off for m-gram language modeling. *In Proc. Int Acoustics, Speech, and Signal Processing ICASSP-95. Conf*, volume 1, pages 181–184.

LIDSTONE, G. (1920). Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *In Transactions of the Faculty of Actuaries, 8 :182-192.*

NOCERA, P., FREDOUILLE, C., LINARES, G., MATROUF, D., MEIGNIER, S., BONASTRE, J.-F., MASSONIÉ, D. et BÉCHET, F. (2004). The lia's french broadcast news transcription system. *In SWIM : Lectures by Masters in Speech Processing*.

ROSENFELD, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228.

ROSENFELD, R. et HUANG, X. (1992). Improvements in stochastic language modeling. *In HLT '91 : Proceedings of the workshop on Speech and Natural Language*, pages 107–111, Morristown, NJ, USA. Association for Computational Linguistics.

SCHWENK, H. et GAUVAIN, J.-L. (2002). Connectionist language modeling for large vocabulary continuous speech recognition. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '02)*, volume 1.

ZITOUNI, I. (2007). Backoff hierarchical class n-gram language models : effectiveness to model unseen events in speech recognition. *Computer Speech and Language*, 21:88–104.