Un cadre prédictif pour l'évaluation des modèles de langage

P. Alain, O. Boëffard

IRISA / Université de Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
{pierre.alain,olivier.boeffard}@irisa.fr
http ://www.irisa.fr/cordial/

ABSTRACT

Perplexity is a widely used criterion in order to compare language models without any task assumptions. However, the main drawback is that perplexity supposes probability distributions and hence cannot compare heterogeneous models. As an evaluation framework, we propose in this article to abandon perplexity and to extend the Shannon's entropy idea which is based on model prediction performance using rank based statistics. Our methodology is able to predict joint word sequences being independent of the task or model assumptions. Experiments are carried out on the English language with different kind of language models. We show that long-term prediction language models are not more effective than the standard n-gram models.

1. INTRODUCTION

De nombreux systèmes de traitement de l'information s'appuient sur des modèles de langage dès lors que les données de traitement correspondent à une séquence de mots dans une langue naturelle. Il peut s'agir de thèmes aussi variés que celui du traitement de la parole, de la traduction automatique, de la recherche d'information ou encore d'interfaces en langue naturelle.

Un modèle de langage décrit, avec des hypothèses plus ou moins restrictives, l'apparition des mots dans une phrase. Méthodologiquement on peut distinguer deux approches pour évaluer la performance d'un modèle de langage : le modèle peut être évalué en dehors de tout contexte applicatif, ou être incorporé à une tâche, par exemple un système de reconnaissance de la parole, et évalué par la performance de la tâche. Le cadre d'évaluation le plus couramment admis est celui de l'entropie croisée, calculée sur un ensemble de test. Pour deux modèles différents, le meilleur est alors celui de plus faible entropie.

Le critère de perplexité souffre cependant de quelques inconvénients, notamment celui d'être fondé sur une hypothèse probabiliste, voir section 2. C'est pour cette raison que de nombreux protocoles expérimentaux privilégient une évaluation qui dépend de la tâche. Cette situation expérimentale est à notre avis aussi critiquable que d'opter pour le critère d'entropie. Il est en effet difficile de discerner dans une performance globale quelle est la part du modèle de langage de celle du reste du système. Il est de plus très difficile de transposer les résultats acquis entre différents domaines. Pour pallier ce défaut, nous proposons de revenir sur un protocole d'évaluation des modèles de langage indépendant de la tâche et d'abandonner le critère d'entropie croisée. Nous privilégions une évaluation fondée sur une statistique des rangs de prédiction. S'intéresser aux rangs de prédiction présente l'avantage de ne pas imposer d'hypothèses précises sur les modèles évalués (notamment d'être un modèle probabiliste).

Ce cadre prédictif n'est pas nouveau en soi, Shannon s'est en effet intéressé au début des années 1950 [7, 8] à une estimation de l'entropie des séquences de lettres pour l'anglais par une approche prédictive. Plus récemment Cover [3] propose un mécanisme de prédiction qui s'appuie sur une fonction de pari. Bimbot [1] a transposé les expériences de Cover à la prédiction de mots dans une phrase. Shannon, Cover et Bimbot ne se sont cependant intéressés qu'à la prédiction d'une seule lettre ou d'un seul mot.

Notre objectif consiste à étendre le cadre de Shannon à une prédiction conjointe de plusieurs mots consécutifs de manière à couvrir tous les type de modèles de langage, notamment ceux qui prédisent plusieurs mots à la fois comme les multigrammes [4, 6, 5]. La suite de cet article est organisée de la manière suivante. La section 2 présente un résumé des travaux de Shannon et de Cover sur l'estimation de l'entropie. Les expériences sont décrites dans la section 4, et les résultats expérimentaux sont présentés section 5.

2. TRAVAUX DE RÉFÉRENCE

Le jeu de Shannon [8] consiste à faire prédire par une personne une lettre à partir de la seule connaissance des $n-1$ lettres précédentes. Cette expérience a pour but d'estimer l'entropie de l'anglais caractérisée par une série de valeurs F_1, F_2, ..., F_N. Chaque F_i représente une approximation de l'entropie quand une personne soumise à l'expérience connaît les $i-1$ premières lettres et doit prédire la suivante. La suite F_i converge vers l'entropie de l'anglais, notée H.

Soit h_{N-1} une suite de $N-1$ lettres consécutives, j une lettre quelconque qui suit la séquence h_{N-1}, $P(h_{N-1}, j)$ la probabilité du N-gramme $[h_{N-1}, j]$, et $P(j|h_{N-1})$ la probabilité d'apparition de la lettre j après le bloc h_{N-1}, est :

$$\begin{aligned} F_N &= -\sum_{N,j} P(h_{N-1}, j) \times \log_2(P(j|h_{N-1})) \\ H &= \lim_{N\to\infty} F_N \end{aligned}$$

Avec H l'entropie de la séquence. En réitérant ce mécanisme de prédiction un grand nombre de fois, Shannon estime une distribution des rangs de prédiction (le rang R

signifie que la personne a dû prédire $R-1$ lettre erronées avant d'obtenir la bonne prédiction).

Cover [3] propose une approche différente pour l'estimation de l'entropie. Il définit une fonction de pari $b(.|.)$: $\mathcal{S} \times \mathcal{S}^* \rightarrow \mathbb{R}$, avec les contraintes : $b(.|.) \geq 0$ et $\sum_j b(j|i_1, i_2, \cdots i_{N-1}) = 1$, $\mathcal{S}$ est l'ensemble des symboles (l'espace des lettres augmenté du caractère $\lrcorner$) et $\mathcal{S}^*$ l'ensemble des suites finies formées à partir de ces symboles. L'utilisation d'une fonction de pari est intéressante, car elle permet de ne pas faire d'hypothèses sur la nature du modèle (notamment comme c'est le cas pour Shannon, de faire l'hypothèse d'un modèle probabiliste).

Nous venons d'évoquer deux cadres d'évaluation qui permettent d'encadrer, selon Shannon, ou de définir en limite, selon Cover, l'entropie du langage écrit. Comme déjà proposés dans Bimbot [1], nous appliquons les principes de Shannon et Cover à la prédiction des mots d'une langue naturelle. Cependant, l'approche proposée par Cover et Bimbot présente le défaut de nécessiter un modèle probabiliste afin de permettre le calcul de $P(W_1, \cdots, W_N)$.

Nous cherchons un cadre d'évaluation qui permette la comparaison de modèles de langage sans faire d'hypothèses sur la nature de ces modèles. Nous souhaitons de plus que la performance du modèle ne puisse évoluer dans le même sens que l'augmentation du nombre de paramètres (critère de parcimonie). Ce qui n'est clairement pas le cas de la perplexité.

3. UN CADRE PRÉDICTIF

On considère une séquence de mots $(W_1\ W_2 \cdots W_n)$, et on s'intéresse dans cet article à la prédiction de plusieurs mots consécutifs. On définit une fenêtre de prédiction de longueur l qui correspond à la prédiction des mots $W_i\ W_{i+1} \cdots W_{i+l-1}$, en ayant connaissance d'un historique partiel issu de l'historique complet $W_1\ W_2 \cdots W_{i-1}$. On s'intéresse alors au rang de prédiction d'une séquence de test par le modèle soumis à évaluation. Plus le rang moyen de prédiction est faible plus le modèle est considéré comme performant. Pour calculer un rang sur une fenêtre de longueur l, le problème est malheureusement de nature combinatoire.

Nous proposons une solution exacte qui s'appuie sur une représentation par un graphe multivalué et la mise en oeuvre d'un algorithme admissible de calcul des R meilleurs chemins si R est le rang de prédiction d'une séquence. La détermination du rang de prédiction d'une séquence de mots de test correspond à la recherche de tous les meilleurs chemins de prédiction du modèle de langage jusqu'au chemin qui correspond à l'émission des mots de test.

La propriété de multivaluation du graphe entraîne un *non déterminisme*. Il existe en effet plusieurs arcs qui permettent de prédire le mot W_i. Une même séquence de test peut donc être prédite selon plusieurs chemins. Nous n'avons pas la place dans cet article de présenter les détails de l'algorithme A^* que nous proposons, notamment une heuristique d'élagage admissible qui permet de contraindre l'exécution de l'algorithme dans un temps raisonnable. Nous proposons ici la mise en œuvre de ce cadre d'évaluation sur différents modèles de langage.

4. MODÈLES DE LANGAGE

Cette section présente les 3 types de modèles de langage utilisés au cours de nos expériences. Tous les modèles de langage soumis à expérience utilisent des distributions de probabilité, mais ne sont pas tous probabilistes. Nous avons volontairement repris la notation d'une fonction de pari $b(.|.)$ introduite par Cover. Cette fonction est quant à elle probabilisée *a posteriori*.

4.1. Le modèle n-gramme fixe

Le premier modèle correspond à un n-gramme classique (modèle probabiliste). Ce modèle utilise un historique de $n-1$ mots et prédit un seul mot. En notant $h(W_i)$ l'historique connu, et W_i le mot à prédire, on définit la fonction de pari ainsi :

$$\begin{aligned} b(W_i|h(W_i)) &= P(W_i|h(W_i)) \\ &= P(W_i|W_{i-n+1}, \cdots, W_{i-1}) \end{aligned}$$

Lorsque l'historique de longueur $n-1$ n'est pas observé à l'apprentissage, on peut calculer une valeur de probabilité à partir d'un coefficient de *back off* et d'un historique réduit. La fonction bow définit le poids de back off, et permet de conserver une valeur de probabilité dans la distribution d'origine ($1 \leq k \leq n$) :

$$\begin{aligned} &P(W_i|W_{i-n+k}, \cdots, W_{i-1}) = \\ &\quad \text{bow}_{W_{i-n+k}, \cdots, W_{i-1}} \times P(W_i|W_{i-n+k+1}, \cdots, W_{i-1}) \end{aligned}$$

4.2. Le modèle n-gramme variable

Le second modèle utilise des probabilités de co-occurrences de mots pour effectuer ses prédictions. Ce modèle n'est pas probabiliste car pour un historique donné, les valeurs de la fonction de pari ne somment pas à 1. La prédiction d'un seul mot est calculée à partir d'un historique de $n-1$ mots.

La modèle de prédiction utilise la fonction $\max$ pour déterminer le score d'un mot en considérant tous ses historiques possibles, $h_k(.)$, avec k une longueur variable d'historique

$$b(W_i|h(W_i)) = \frac{\max_k P(W_i|h_k(W_i))}{\sum_{v_j} \max_k P(W_i = v_j|h_k(W_i))}$$

L'unigramme n'entre pas dans le calcul. Les distributions de probabilité ne sont plus comparables car pondérées par la taille du vocabulaire (valeur absolue) et non la fréquence d'apparition de l'historique (valeur relative).

4.3. Le modèle n/m-multigramme

Le dernier modèle s'appuie sur des distributions de probabilité, et n'est pas non plus probabiliste. Comme le n-gramme variable, il manipule des distributions de probabilité issues de distributions différentes.

L'unité de prédiction n'est plus un mot, mais un fragment, c'est-à-dire un groupe de mots conjoints de longueur 1 à m. L'historique qui sert à la prédiction comporte $n-1$ fragments, ou encore $m \times (n-1)$ mots connus. La prédiction du modèle porte sur un seul fragment mais correspond à un prédiction de 1 à m mots consécutifs. On peut montrer que, sans être probabiliste, ce modèle *multigramme*

est plus général que le modèle définit par [4, 5] (c'est-à-dire, tous les chemins vus par le modèle proposé par ces derniers seront vus par ce modèle n/m-multigramme). Il a le même comportement que le n-gramme variable mais sur un ensemble de fragments et non plus de mots. Avec F_i définit par $[W_i, W_{i+1}, \cdots, W_{i+m}]$, ou $[W_i, W_{i+1}, \cdots, W_{i+m-1}]$, ..., ou $[W_i]$, on obtient

$$b(F_i|h(W_i)) \quad = \quad \frac{\max_k P(F_i|h_k(W_i))}{\sum_{f_j} \max_k P(F_i = f_j|h_k(W_i))}$$

4.4. Protocole expérimental

Les expériences ont été réalisées à partir des textes de l'année 1987 du *Wall Street Journal*. Le corpus d'apprentissage comporte 500 455 phrases, avec un vocabulaire de 94 198 mots et 11 260 184 occurrences. Le vocabulaire retenu est composé des 30 000 mots les plus fréquents. Pour ces 30 000 mots, 11 118 995 occurrences sont observées. Le corpus de test est composé de 50 037 phrases, différentes du corpus d'apprentissage, comportant 38 564 mots différents et 1 127 831 occurrences.

Le test de prédiction est réalisé sur 12 000 fenêtres de 2 mots. Les fenêtres sont sélectionnées par un décalage de 7 mots à partir du premier mot du fichier. 7 est le nombre maximum de mots incluant une longueur maximum d'historique. Les expériences proposées mettent en concurrence un modèle 3-gramme (fixe ou variable) et un $2/2$-multigramme (il s'agit du n/m-multigramme le moins complexe). Ce $2/2$-multigramme peut, dans certains cas, développer, un paramètre de type 4-gramme, cependant l'historique *réel* connu par le modèle reste de même taille que celui connu par le modèle 3-gramme.

Les modèles 3-gramme à horizon fixe et variable sont générés à partir du corpus d'apprentissage et sur le vocabulaire sélectionné. Ces modèles 3-gramme contiennent 1 646 823 paramètres (hors coefficients de back off). Les modèles $2/2$-multigramme sont construits à partir du même vocabulaire de 30 000 mots auquel on a adjoint un ensemble de fragments de 2 mots. La variation du nombre de fragments implique un nombre plus ou moins important de paramètres. L'ensemble des fragments varie de 125 000 (modèle de langage à 2 662 174 paramètres) à la totalité des fragments existant dans le corpus d'apprentissage et construits sur le vocabulaire (ce qui fait 1 815 271 fragments et 5 069 602 paramètres pour le modèle de langage). Nous sommes dans un cadre expérimental plus général que les protocoles expérimentaux proposés par [4, 5] car nous incorporons dans le modèle tous les fragments possibles. Lors de l'apprentissage, un cutoff à 1 [2] est appliqué pour ne garder comme paramètres que les apparitions conjointes présentes au moins 2 fois dans le corpus d'apprentissage.

Toutes les expériences font intervenir le calcul de la valeur moyenne du rang de prédiction. Ces valeurs moyennes sont présentées avec des intervalles de confiance à 95% établis par ré-échantillonnage à partir des distributions empiriques (méthode de monte-carlo, bootstrap).

5. RÉSULTATS ET DISCUSSION

Le rang maximal autorisé dans l'algorithme A^* est fixé à 7 000 000 ; au delà de ce rang, la fenêtre est déclarée non prédite par le modèle de langage. Ce rang maximal est évidement à déterminer en fonction de la longueur des fenêtres de prédiction. Nous avons pu observer sur des expériences préliminaire que le rang maximal atteint est de 659 473 pour le modèle de 3-gramme fixe, 659 370 pour le modèle 3-gramme variable, et 1 814 435 pour le modèle multigramme qui tient compte du maximum de fragments (le rang maximal n'a donc jamais été atteint).

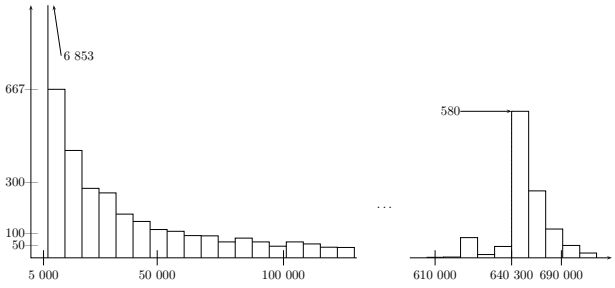

FIG. 1: Un exemple de distribution de rangs. Il s'agit de la distribution obtenue lors du test de prédiction pour le modèle $2/2$-multigramme, avec 500 000 fragments autorisés dans le vocabulaire. L'histogramme est représenté pour 100 intervalles.

5.1. Analyse des distributions de rang

L'algorithme A^* utilisé permet de déterminer un rang pour une fenêtre de 2 mots. Il est donc nécessaire d'appliquer ce calcul sur un grand nombre de fenêtres de manière à construire une distribution stable des rangs de prédiction. La figure 1 est un exemple de distribution obtenue lors du test. Toutes les distributions sont de même nature. Des changements apparaissent sur le rang maximum atteint ou sur la valeur moyenne des distributions. On comparera les distributions sur ces critères. On peut noter que les distributions possèdent deux modes. Le premier suit une loi puissance, ce mode a déjà été observé par Shannon dans ses expériences sur les lettres. On note également la présence d'un second mode, ce mode correspond aux fenêtres prédites avec difficulté. Par exemple dans les phrases :

...RETAILERS BUSINESS HAS SOFTENED UP | SAID STACY | RUCHLAMER...

...115 6 MILLION NET LOSS AND | 183 4 | MILLION LOSS...

La fenêtre de prédiction correspond aux mots entre 'l'. Ce type de fenêtre possède un coût très important (c'est-à-dire une probabilité très faible pour nos modèles probabilistes). Avec un tel coût, la fenêtre réelle est prédite dans un lot de fenêtres *improbables*. Sur ce mode, la prédiction se fait le plus souvent sans tenir compte de l'historique. Dans ce cas de figure, le modèle de langage classe les mots candidats selon leurs fréquences d'apparition dans la langue. Dans cette situation, il est normal de trouver un rang très mauvais.

5.2. Comparaison des modèles

Pour comparer les moyennes associées à leurs intervalles de confiance, on peut soit estimer la moyenne sur l'ensemble de la distribution, soit limiter le calcul de la moyenne au premier mode. Si on considère l'ensemble de la distribution, on peut se rendre compte que les modèles de langage qui manipulent un grand nombre de paramètres sont défavorisés. En effet, lors de la prédiction d'une fenêtre difficilement prédictible, avoir un grand nombre de

Modèle	moyenne [ci $\pm 5\%$]
3-gramme fixe	**9.18** [8.82 – 9.52]
3-gramme variable	**10.08** [9.73 – 10.42]
$2/2$-multigramme 1.8M	**20.65** [19.69 – 21.53]
$2/2$-multigramme 1M	**13.69** [13.17 – 14.21]
$2/2$-multigramme 500k	**10.21** [9.82 – 10.56]
$2/2$-multigramme 250k	**9.84** [9.47 – 10.19]
$2/2$-multigramme 125k	**9.67** [9.30 – 10.03]

TAB. 1: Rangs moyens et intervalles de confiance à 95%, (en 10^4), en intégrant les deux modes de la distribution expérimentale des rangs.

Modèle	moyenne [ci $\pm 5\%$]	% masse
3-gramme fixe	**1.65** [1.58 – 1.72]	84.6%
3-gramme variable	**1.79** [1.71 – 1.86]	83.1%
$2/2$-multigramme 1.8M	**1.71** [1.64 – 1.78]	83.2%
$2/2$-multigramme 1M	**1.71** [1.64 – 1.78]	83.3%
$2/2$-multigramme 500k	**1.70** [1.63 – 1.77]	83.3%
$2/2$-multigramme 250k	**1.72** [1.65 – 1.79]	83.3%
$2/2$-multigramme 125k	**1.95** [1.88 – 2.03]	84.6%

TAB. 2: Rangs moyens, intervalles de confiance à 95% pour le 1^{er} mode seulement (en 10^4) et pourcentage de fenêtres avec un rang inférieur à 200 000.

paramètres signifie avoir un grand nombre de candidats pas forcément très bons. Sur le plan de l'interprétation des rangs, on peut se demander si un modèle de langage qui prédit une fenêtre après 600 000 propositions est vraiment plus performant qu'un modèles qui prédit cette même fenêtre après 1 800 000 autres propositions. Nous pensons qu'il y a un seuil à partir duquel la prédiction est mauvaise, et qu'un régime normal est décrit par le premier mode.

La table 1 montre le rang moyen obtenu sur chacune des distributions de rang en intégrant les deux modes. On remarque que les modèles qui utilisent un grand nombre de paramètres obtiennent de mauvais scores. Cela est principalement dû au rang maximum atteint par ces modèles. En effet, le modèle $2/2$-multigramme à 1.8M fragments autorisés a un rang maximum de 1 814 435, le même modèle avec 1 000 000 fragments a un rang maximum de 999 015. Les trois autres modèles multigramme ont un rang maximum proche de 670 000. L'ordre de grandeur du rang maximum pour les deux modèles 3-gramme (fixe et variable) est de 660 000. On peut enfin remarquer que le modèle 3-gramme variable, qui avait obtenu le meilleur score par une mesure de perplexité, ne se distingue plus des modèles $2/2$-multigramme qui intègre jusqu'à 500 000 fragments. Le modèle 3-gramme fixe obtient le meilleur rang moyen de prédiction.

Le tableau 2 présente les valeurs des rangs moyens ainsi que leurs intervalles de confiance pour les différents modèles en considérant uniquement le premier mode des distributions de rang. On n'intègre dans le calcul des moyennes que les rangs inférieurs à 200 000. La dernière colonne du tableau présente la masse de ce premier mode par rapport à l'ensemble de la distribution. Ainsi, les modèles $2/2$-multigramme avec plus d'un million de fragments ne sont plus pénalises par les mauvais rangs de fenêtres de test difficilement prédictibles. On constate que le modèle 3-gramme variable reste toujours moins bon prédicteur, tout comme le modèle $2/2$-multigramme avec 125 000 fragments. Plus les modèles s'approchent d'un 2-gramme (un 2-gramme est un $2/2$-multigramme sans fragments autre que les mots du vocabulaire), moins le pouvoir de prediction est bon. Tous les autres modèles donnent des résultats statistiquement non significatifs ; on ne peut donc les différencier. Il faut tout de même garder présent à l'esprit que leurs nombres de paramètres fluctuent de manière importante.

6. CONCLUSION

Dans cet article, nous avons présenté une extension originale du jeu de Shannon pour l'évaluation des modèles de langage appliquée à la prédiction conjointe de plusieurs mots. Ce type d'évaluation prédictive permet une comparaison fiable de modèles de langage en dehors de toute tâche applicative. Nous avons mis en œuvre ce protocole avec différents modèles de langage : un modèle n-gramme standard, un modèle n-gramme variable et un modèle multigramme. Nous avons montré que le nombre de paramètres des modèles évalués n'a pas une influence importante sur la moyenne du premier mode des distributions de rangs. Pour conclure, la méthodologie d'évaluation proposée montre qu'il est possible de comparer avec raison le pouvoir de prédiction des modèles de langage sans nécessiter l'évaluation d'une tâche particulière. Nous avons proposé une méthodologie efficace de recherche des distributions de rangs, et cela sans faire d'hypothèses sur la nature des modèles soumis à l'évaluation.

RÉFÉRENCES

[1] F. Bimbot, M. El-Beze, S. Igounet, M. Jardino, K. Smaili, and I. Zitouni. An alternative scheme for perplexity estimation and its assessment for the evaluation of language models. *Computer Speech and Language*, 15(1) :1–13(13), 2001.

[2] S.F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4) :359–394, 1999.

[3] T. Cover and R.C. King. A convergent gambling estimate of the entropy of english. *IEEE Transactions on Information Theory*, 24(4) :413–421, 1978.

[4] S. Deligne and F. Bimbot. Language modeling by variable length sequences : theoretical formulation and evaluation of multigrams. In *IEEE International Conference on Acoustics and Speech Signal Processing*, pages 169–172, 1995.

[5] S. Deligne and Y. Sagisaka. Statistical language modeling with a class-based n-multigram model. *Computer Speech and Language*, 14 :261–279, 2000.

[6] P.A. Heeman and G. Damnati. Deriving phrase-based language models. *IEEE Workshop on Speech Recognition and Understanding*, pages 41–48, 1997.

[7] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 :379–423, 1948.

[8] C. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30 :50–64, 1951.