Évaluation des Modèles de Langage *n*-gramme et *n*/*m*-multigramme

P. Alain, O. Boëffard

Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)
groupe Cordial

09 juin 2005

Introduction

Présentation des modèles

Méthodologie expérimentale

Résultats expérimentaux

Travaux complémentaires

Conclusion

Introduction

- Méthodologie de construction :
  - déterministe
  - probabiliste
- Modèles de langage (sous forme d'arbre) :
  - n-gramme à horizon fixe
  - n-gramme à horizon variable
  - n/m-multigramme
- Méthodologie d'évaluation : perplexité
  - Comparaison à nombre de paramètre, et taux hors vocabulaire constant.
  - Complexité spatiale.
  - Complexité temporelle.

Présentation des modèles

- Séquence de mots $W = w_1, w_2, \cdots, w_N$.
- Estimation de la perplexité

$$\begin{aligned} PP &= 2^{H^*} \quad \text{avec} \\ H^* &= -\frac{1}{N} \log_2 (P(W)) \end{aligned}$$

- Probabilité conjointe $P(W)$ peut se développer en :

$$P(W) = p(w_1) \times \prod_{i=2}^{N} p(w_i | w_1, \cdots w_{i-1})$$

Modèle n-gramme à horizon fixe

Apprentissage :

- Tête à 1 mot.
- Historique à $(n-1)$ mots.

Décodage :

- Si n-gramme absent dans l'apprentissage, réduit au $(n-1)$ puis multiplié par le coefficient de backoff.
- On prend la probabilité correspondant au chemin le plus long dans l'arbre du modèle.
- On stocke les $p(w_i|w_{i-(j)}, \cdots w_{i-1}), 1 \leq j \leq n-1$ dans le modèle.

$$p(w_i|w_1, \cdots w_{i-1}) \quad = \quad p(w_i|w_{i-(n-1)}, \cdots w_{i-1})$$

Modèle *n*-gramme à horizon variable

Apprentissage :

- Tête à 1 mot.
- *Historique de 1 à* $(n-1)$ *mots.*

Décodage :

- Si 2-gramme absent dans l'apprentissage, la tête est considérée comme non prédictible.
- On recherche la probabilité maximum.
- Par rapport aux *n*-grammes à horizon fixe, la complexité temporelle est plus forte due à la variation de l'horizon.
- La complexité spatiale est la même, on ne garde pas plus de probabilités.

$$p(w_i|w_1,\cdots w_{i-1}) \quad = \quad \max_{1\leq j\leq n-1} p(w_i|w_{i-j},\cdots w_{i-1})$$

Modèle n/m-multigramme

```
 w_1    w_2       w_3   w_4
-----------------------------
 w_1 , w_2   |   w_3
       w_2   |   w_3
 w_1 , w_2   |   w_3, w_4
[w_1 , w_2]  | [w_3, w_4]
```

Apprentissage :

- *Tête à 1 à m mots.*
- *Historique de 1 à* $m \times (n-1)$ *mots.*

Modèle n/m-multigramme

Décodage :

- Si 2-gramme absent dans l'apprentissage, la tête est considérée comme non prédictible.
- On recherche la probabilité maximum.
- Complexité temporelle due à l'horizon (beaucoup plus long horizon par rapport au n-gramme à horizon variable), et sur la tête.
- Complexité spatiale importante due au nombre de fragments de mots.
- Recherche du meilleur chemin dans la phrase selon l'algorithme de Dijkstra (garantie de la découpe optimale pour le calcul de la perplexité).

$$p(w_i|w_1, \cdots w_{i-1}) = \max_{1 \leq j \leq m \times (n-1)} \max_{0 \leq k \leq m-1} p([w_i, \cdots w_{i+k}]|w_{i-j}, \cdots w_{i-1})$$

Méthodologie expérimentale

```
w_1   w_2   w_3   w_4
--------------------
w_1 , w_2 | w_3 , w_4
w_1 , w_2 | w_3
      w_2 , w_3 | w_4
      w_2 | w_3 , w_4
```

- Contrainte sur la longueur des modèles.
- Modifications de la chaîne HTK.
- Corpus utilisé.
- Réduction du nombre de paramètre des modèles.

Méthodologie expérimentale : Longueur des modèles

Pour limiter la complexité spatiale, et comparer des modèles de taille équivalente : limitation à une somme maximum la longueur tête+historique pour le modèle de multigramme. Ici on compare un 3-gramme avec un 2/2-multigramme somme à 3.

$$
\begin{aligned}
p([w_3\,w_4]|[w_1\,w_2]) &= \frac{c(w_1, w_2, w_3, w_4)}{c(w_1, w_2)} \\
p(w_4|w_1, w_2, w_3) &= \frac{c(w_1, w_2, w_3, w_4)}{c(w_1, w_2, w_3)} \\
p([w_3\,w_4]|[w_2]) &= \frac{c(w_2, w_3, w_4)}{c(w_2)} \\
p(w_3|w_1, w_2) \times p(w_4|w_2, w_3) &= \frac{c(w_1, w_2, w_3) \times c(w_2, w_3, w_4)}{c(w_1, w_2) \times c(w_2, w_3)}
\end{aligned}
$$

Méthodologie expérimentale : HTK

Utilisation de la chaîne htk :

- Extraction des n-uplets et accumulation des statistiques.
- Lissage des statistiques (Katz).
- Construction du modèle de langage (arbre).
- Décodage de la perplexité pour les n-gramme fixe (et gestion des mots hors vocabulaires).

Modification apportées :

- Lors du test, itérations supplémentaires pour les n-gramme à horizon variable.
- Lors de l'apprentissage, itérations pour les multigrammes (prendre en compte les probabilités des multigrammes).
- Lors du test, itérations supplémentaires pour l'algorithme de Dijkstra.

Méthodologie expérimentale : Corpus

- Corpus le monde 1997 : articles parus dans le journal "Le Monde" pendant l'année 1997.
- Corpus normalisé sur la casse et la ponctuation.
- Environ 1 million de phrases : 70% pour l'apprentissage et 30% pour le test.

<s> LES TéMOIGNAGES MANQUENT QUI PERMETTRAIENT D éTAYER L HYPOTHèSE </s>

<s> MANIOC éPINARDS ET LéGUMES DIVERS POUSSENT EN VILLE HORS DES PARCELLES ET DES JARDINS </s>

Méthodologie expérimentale : Contrôle du nombre de paramètres

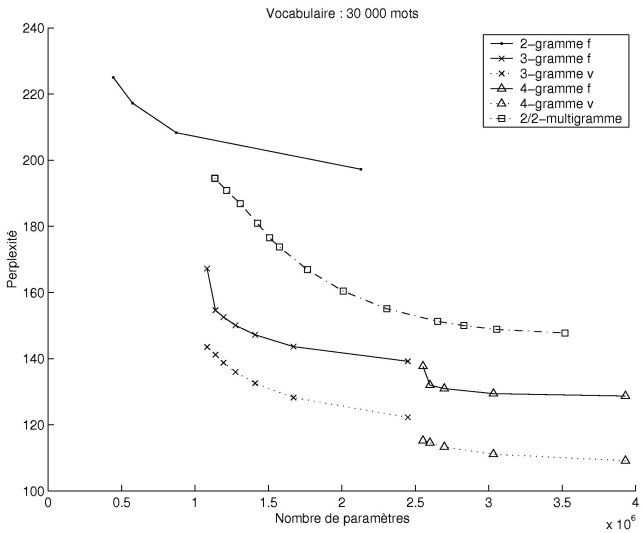
- Même nombre de paramètres pour les modèles *n*-gramme fixe et variable ; utilisation différente des paramètres (gestion du phénomène de backoff).
- Nécessité d'un nombre de paramètres constant pour la comparaison des modèles.
- Baisse du nombre de paramètres pour les multigrammes vers celui des deux autres modèles.
- Pour faire baisser celui des multigrammes :
  - moins de fragments reconnus,
  - seuils d'apparition plus élevés (cutoff).

Résultats expérimentaux

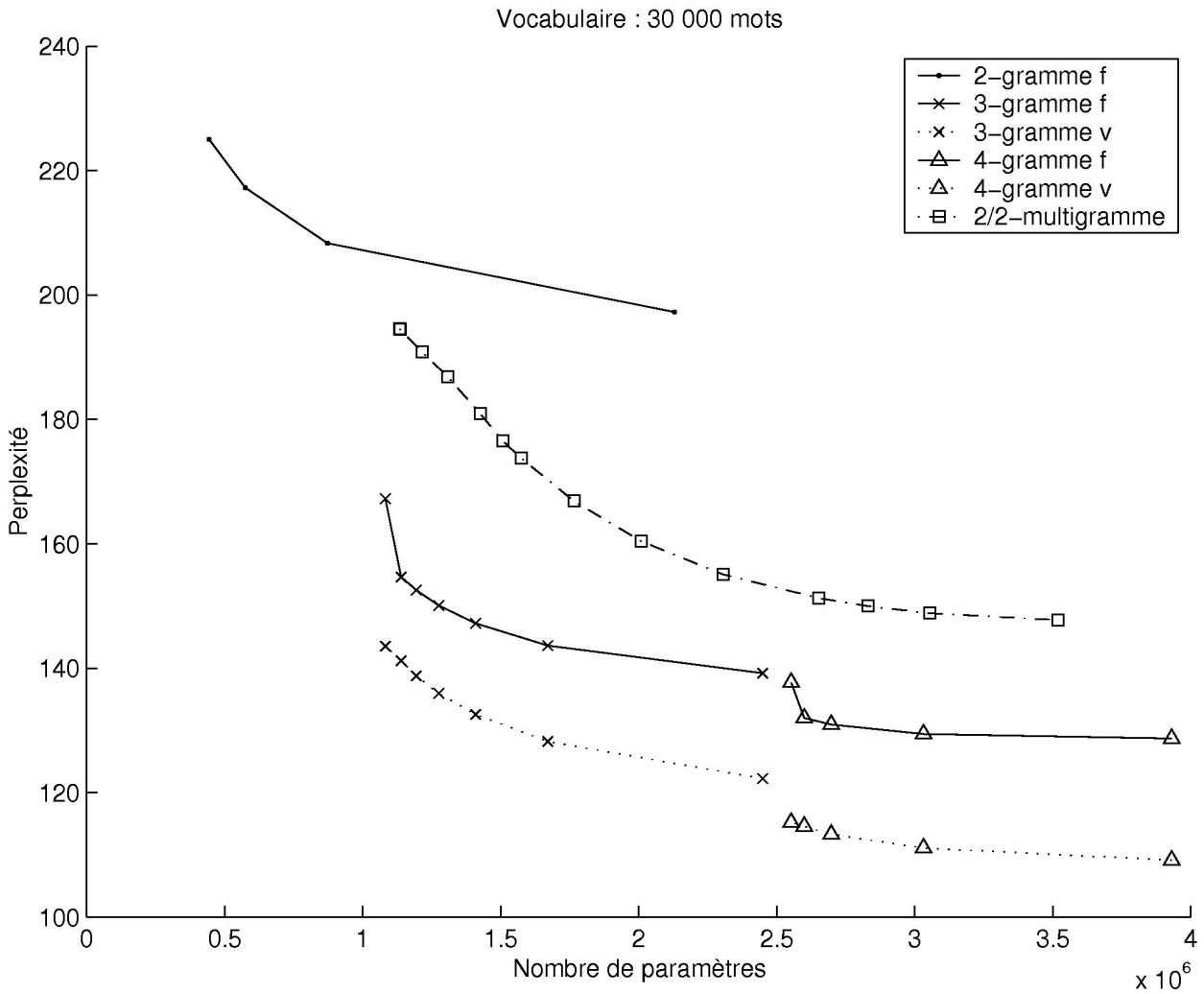
- Vocabulaire de 30 000 mots.
- Vocabulaires de 3 000 mots et 60 000 mots.
- Méthodes de réduction du nombre de paramètres.

Résultats expérimentaux : vocabulaire de 30 000 mots

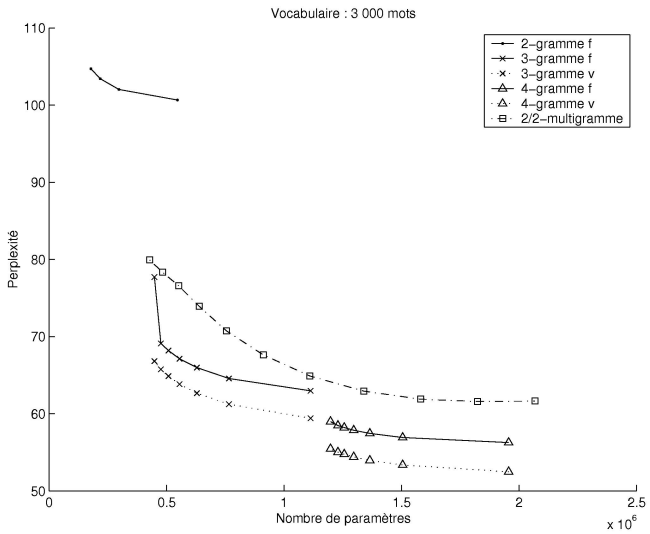
- Vocabulaire de base de 30 000 mots.
- Baisse du nombre de fragments construits sur le vocabulaire reconnus par le modèle de langage pour diminuer le nombre de paramètres.

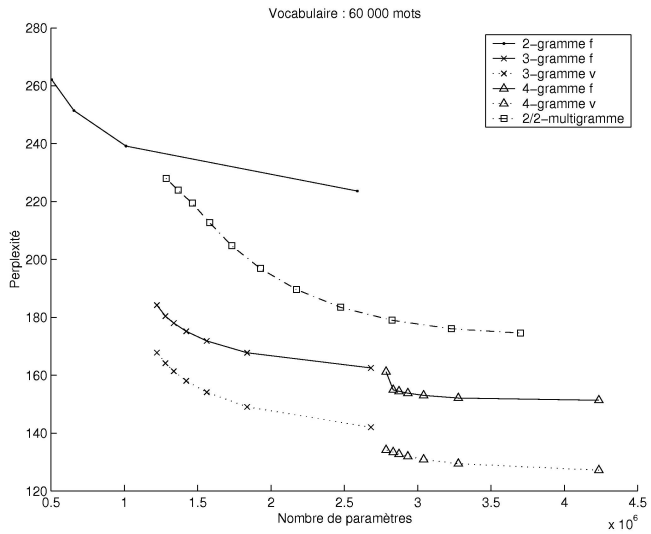


Résultats expérimentaux : vocabulaire de 30 000 mots

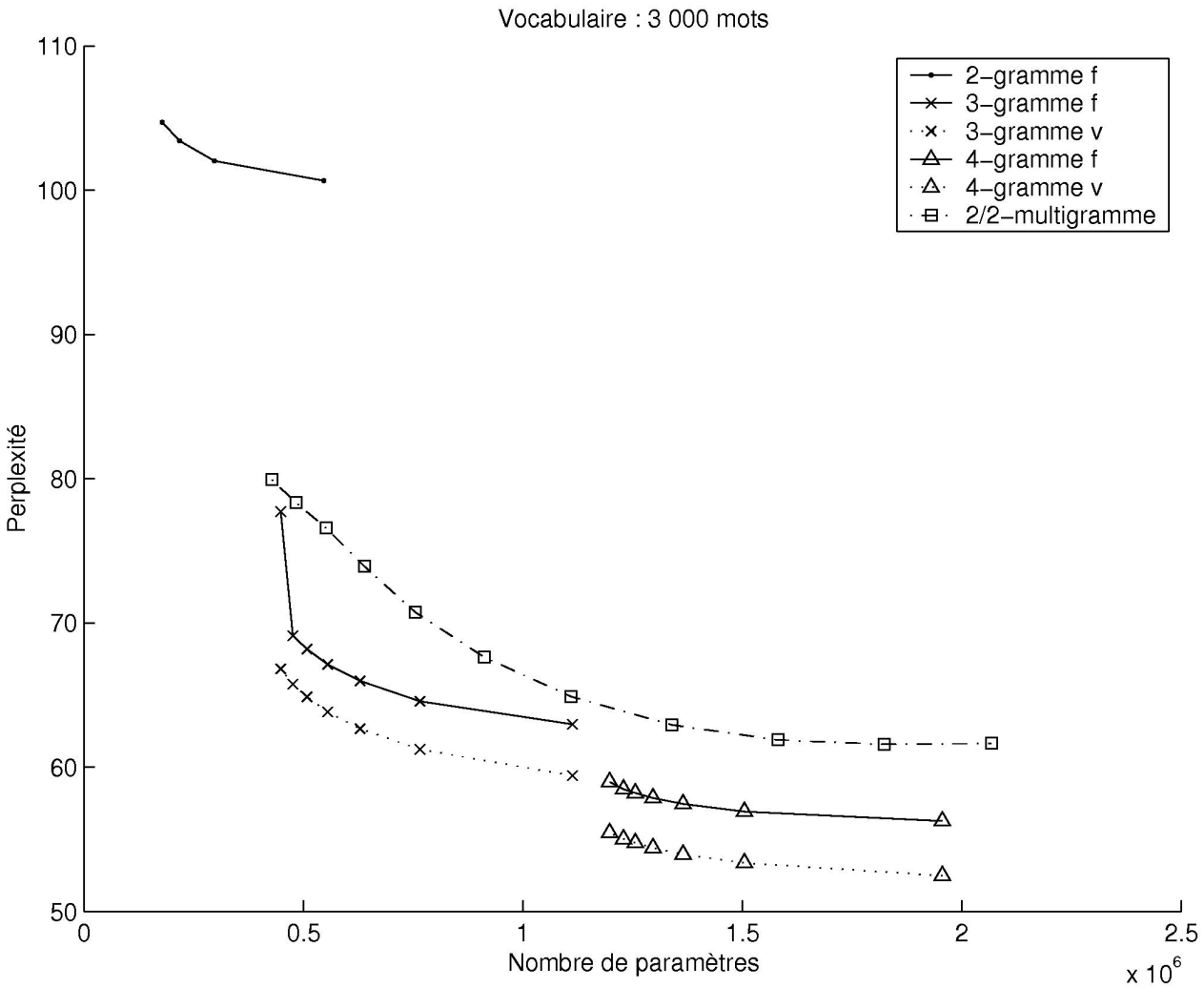


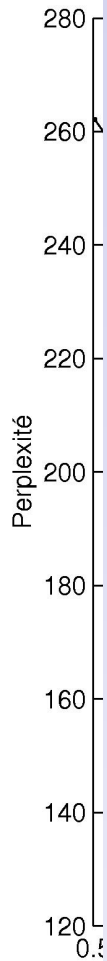
Résultats expérimentaux : vocabulaire de 3 000 et 60 000 mots



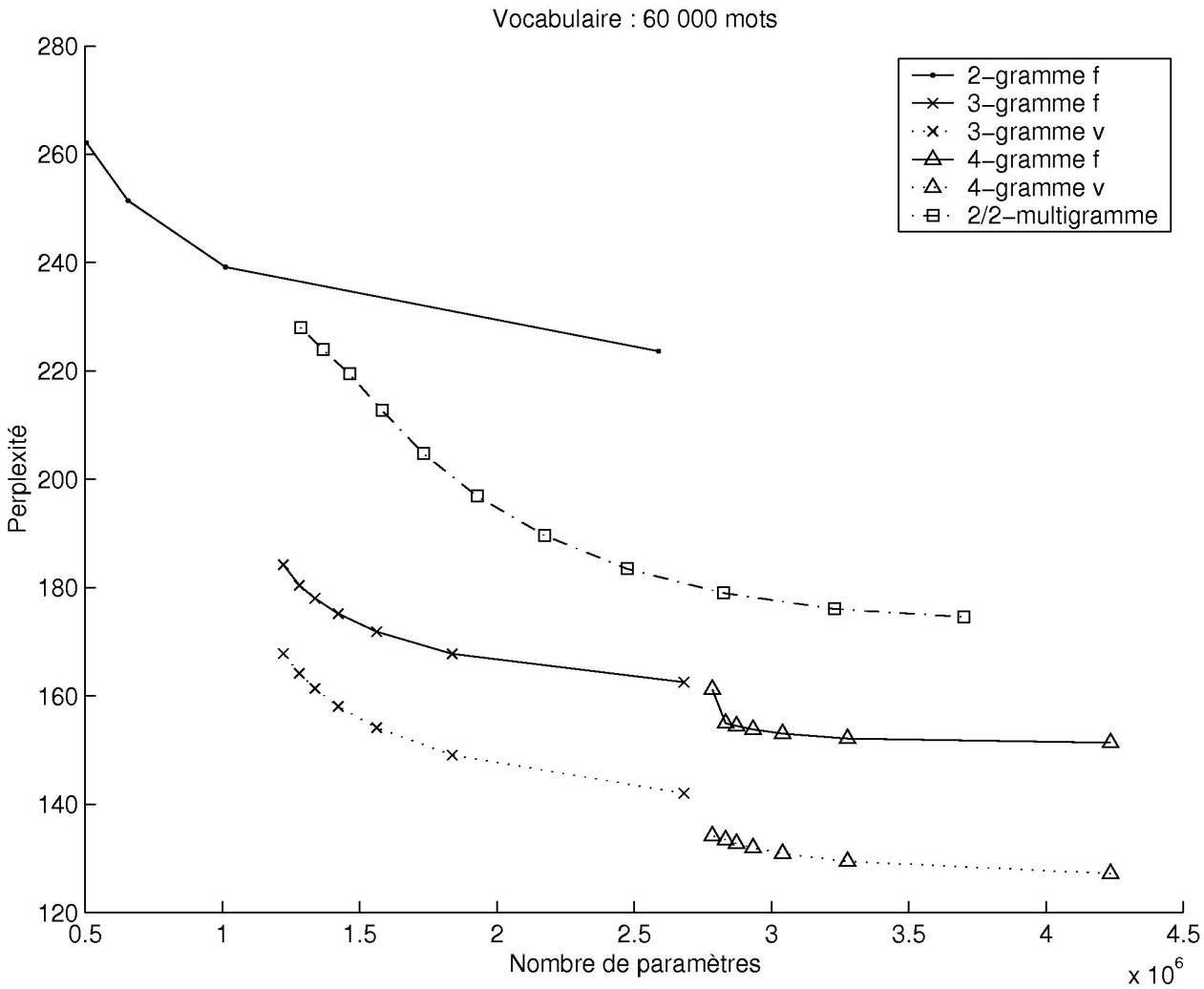


Résultats expérimentaux : vocabulaire de 3 000 et 60 000 mots



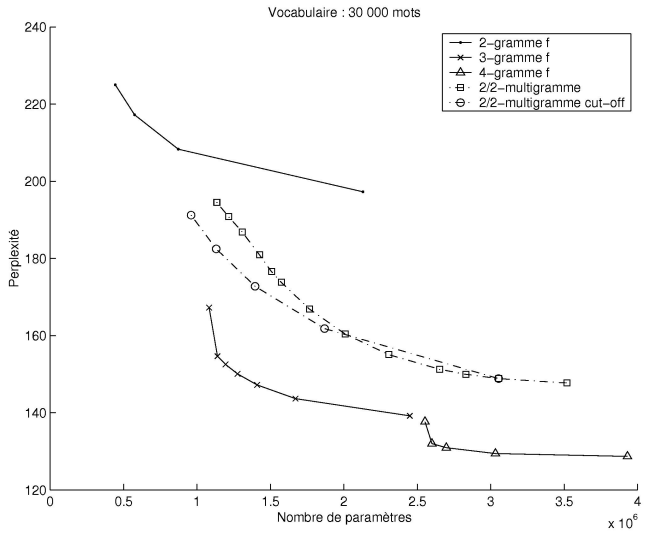


Résultats expérimentaux : vocabulaire de 3 000 et 60 000 mots

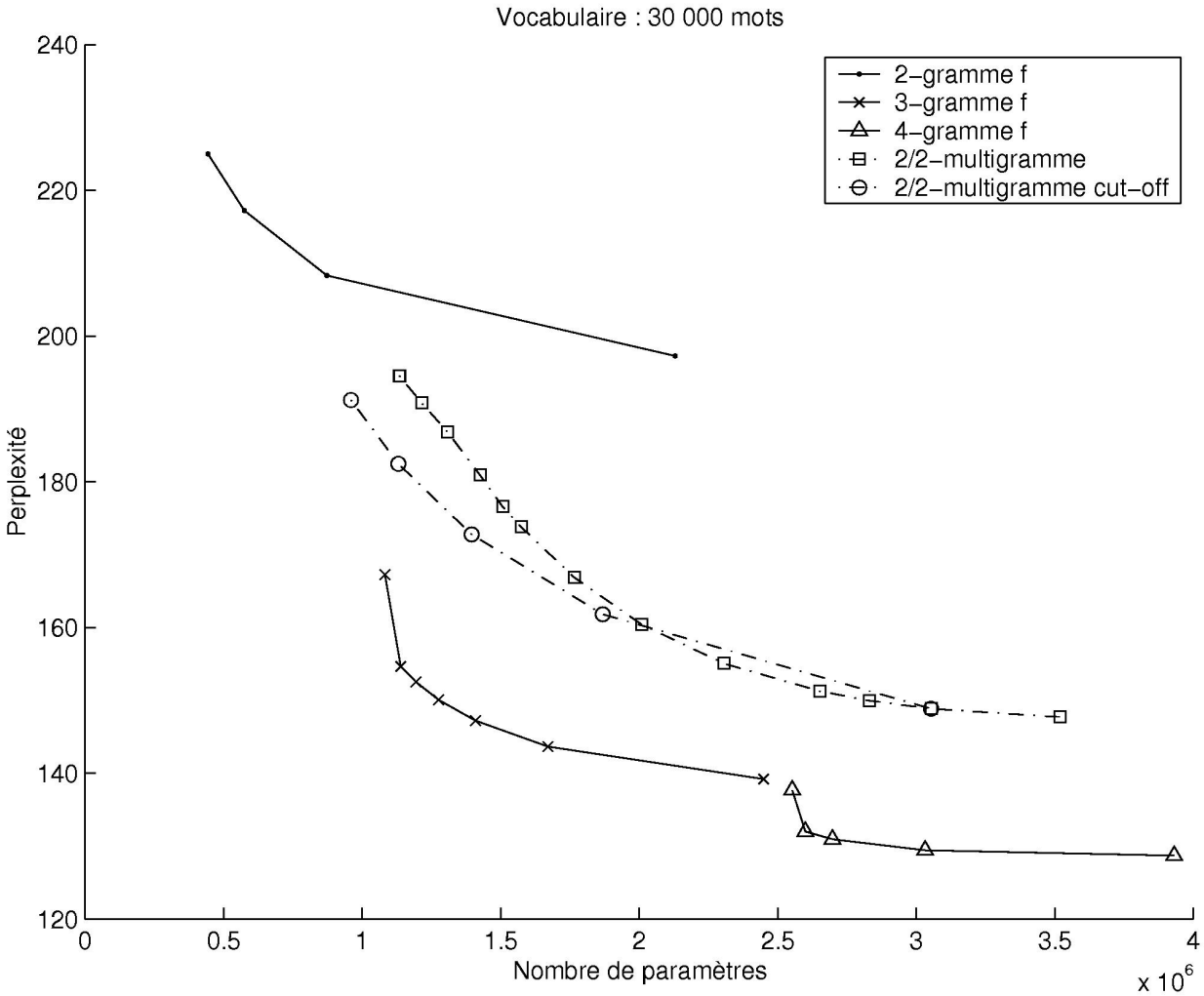


Résultats expérimentaux : Réduction du nombre de paramètres

La perplexité est meilleure si on augmente le cutoff plutôt que de diminuer le nombre de fragments reconnus.



Résultats expérimentaux : Réduction du nombre de paramètres



Travaux complémentaires

- Améliorer la perplexité en choisissant mieux les fragments ajoutés.
- Un fragment donne une meilleure perplexité que les 3-grammes correspondants si un certain 2-gramme est meilleur qu'un certain 3-gramme.

- Le modèle de n-gramme variable n'est meilleur que parceque construit pour faire baisser la perplexité.
- Utiliser une autre mesure que la perplexité.
- Comparer les modèles sur une statistique des rangs de prédiction.

Conclusion

- Hypothèses : Le modèle de multigrammes semble être bon car
  - on utilise un max sur les probabilités,
  - on évite des multiplications de probabilités (prédiction de plusieurs mots).
- Les expériences ont tendance à rejeter ces hypothèses.
- La perplexité n'est pas améliorée avec un modèle de bi-multigrammes.
- Complexité temporelle importante due à l'algorithme de Dijkstra.
- Effet sur la perplexité de la taille du vocabulaire reconnu.