Texte de développement : détermination des seuils des critères

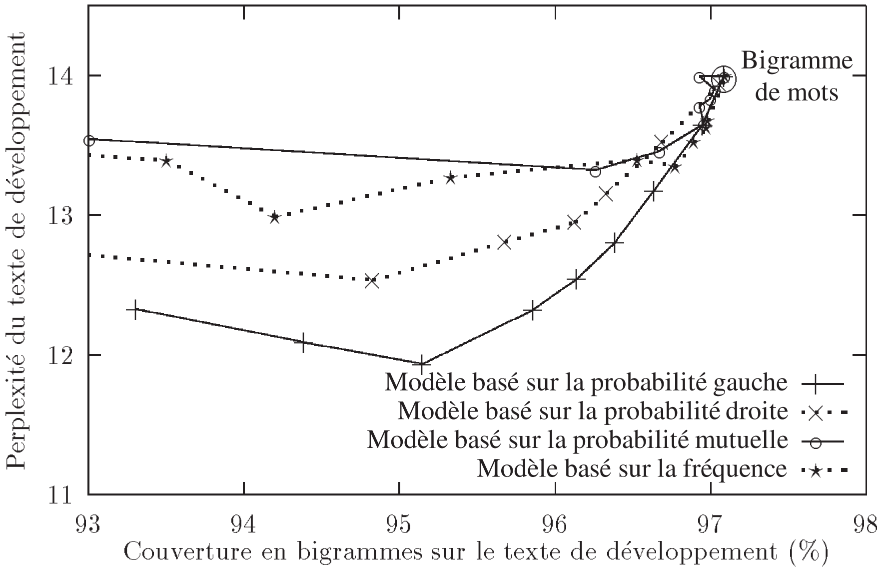

Figure 1:Perplexité en fonction de la couverture en bigrammes selon les seuils fixés, sur le texte de développement. Les seuils décroissent de la droite vers la gauche de 1 à 0 pour les probabilités et de 1500 à 500 pour les fréquences.

Modèles	Prob. gauche	Prob. droite	Prob. mutuelle	Fréq.	Big. de mots
Seuil	0,7	0,7	0,4	1250	—
Perplexité	**12,80**	12,95	13,46	13,63	13,99
Couverture (%)	96,4	96,1	96,7	97,0	**97,0**
Erreur mots (%)	**8,26**	8,57	8,41	8,56	8,62

Tableau 1: Perplexité, couverture en bigrammes et taux d'erreur sur les mots sur les données du développement

		Nombre de mots composés en fonction de leur taille							
Modèles		Prob. gauche		Prob. droite		Prob. mutuelle		Fréq.	
Seuil		0,7		0,7		0,4		1250	
Textes		App.	Dev.	App.	Dev.	App.	Dev.	App.	Dev.
Taille mots	2	111	46	136	77	45	37	3	3
	3	30	16	22	9	11	10	—	—
	4	9	6	2	1	3	2	—	—
	5	—	—	—	—	1	1	—	—

Tableau 2: Distribution des mots composés dans les textes d'apprentissage et de développement en fonction de leur taille

Evaluation

Modèles	Prob. gauche	Prob. droite	Prob. mutuelle	Fréq.	Big. de mots
Seuil	0,7	0,7	0,4	1250	—
Perplexité	13,70	**13,66**	14,03	14,18	14,59
Couverture (%)	95,7	95,5	96,0	96,3	**96,4**
Erreur mots (%)	**9,99**	10,47	10,03	10,08	10,15

Tableau 3: Perplexité, couverture en bigrammes et taux d'erreur sur les mots sur les données du test