550, rue Sherbrooke Ouest, bureau 100
Montréal (Québec) H3A 1B9
Tél. : 514 840-1234; Téléc. : 514 840-1244
Place de la Cité – Tour de la Cité
2600, boul. Laurier, bureau 625
Québec (Québec) G1V 4W1
Tél. : 418 648-8080; téléc. : 418 648-8141
http://www.crim.ca

CRIM - Documentation/Communications

Rapport technique

Intégration d'une grammaire de nombres dans les modèles de langage

CRIM-08/11-07

Maryse Boisvert
Agente de recherche
Équipe Reconnaissance de la parole

Novembre 2008

Collection scientifique et technique

ISBN-13 : 978-2-89522-115-9

Pour tout renseignement, communiquer avec :
CRIM Centre de documentation
CRIM
550, rue Sherbrooke Ouest, bureau 100
Montréal (Québec) H3A 1B9
http://www.crim.ca
Téléphone : (514) 840-1234
Télécopieur : (514) 840-1244

ISBN-13 : 978-2-89522-115-9
Dépôt légal - Bibliothèque et Archives nationales du Québec, 2008
Dépôt légal - Bibliothèque et Archives Canada, 2008

TABLE DES MATIÈRES

1. INTRODUCTION

Nous présentons ici des modèles de langage qui permettent de reconnaître tous les nombres possibles. Habituellement, les nombres sont considérés comme les autres mots, et sont inclus au vocabulaire du modèle de langage, s'ils sont assez fréquents. Ainsi, plusieurs chiffres échappent au modèle, car ils n'ont pas été vus souvent dans les textes d'entraînement, ce qui occasionne plusieurs erreurs.

Ces modèles sont basés sur une grammaire de nombres. Il y a une infinité de nombres, mais ils sont tous composés d'unités plus petites, en quantité finie. Avec quelques règles et prononciations, il est donc possible de générer les prononciations de tous les nombres possibles et donc, de les reconnaître dans la parole.

La classe de tous les nombres possibles est représentée dans le modèle de langage par le symbole *⟨nombre⟩* Le modèle de langage est composé avec la grammaire, c'est-à-dire que toutes les occurrences de *⟨nombre⟩* sont remplacées par une machine à états finis qui représente la grammaire.

Nous décrivons d'abord comment nous construisons un modèle de langage avec la classe de mots *⟨nombre⟩*. Puis, nous discutons de la performance de ces modèles sur des tâches où les chiffres sont beaucoup utilisés, soient le hockey et le football. Ensuite, nous comparons cette performance aux modèles de base sans grammaire de nombres.

2. NORMALISATION

Puisque les nombres sont représentés dans le modèle de langage par l'étiquette *⟨nombre⟩*, toutes les occurrences de chiffres dans les textes d'entraînement sont remplacées par cette étiquette. De plus, les séries de plus d'un *⟨nombre⟩* sont remplacées par un seul *⟨nombre⟩*, ainsi il ne peut y avoir de ngrams ($n > 1$) comportant plusieurs tags consécutifs. Ceci empêche les erreurs de reconnaissance où un nombre est remplacé par une série de plus petites unités dont la phonétique correspond, par exemple « 100-80-5 » au lieu de « 185 ». Ceci peut arriver notamment si le nombre n'est pas présent dans le vocabulaire, ou si les composants sont plus probables que l'union des composants. Ce type d'erreur de reconnaissance était fréquent avec les anciens modèles, de même qu'avec les modèles à grammaire de nombres ayant des bigrammes ou trigrammes de *⟨nombre⟩*.

Par exemple, dans l'ensemble de test de CFL, le modèle original n'utilisant pas la grammaire de nombres fait sept erreurs de ce type :

- 2209 au lieu de 229
- 2002169 au lieu de 2269
- 109 au lieu de 19
- 10 cette au lieu de 17
- 2001 au lieu de 201
- 10015 au lieu de 115
- 5167 au lieu de 567

Le modèle à grammaire de nombres empêchant ces transitions n'a fait aucune de ces erreurs.

3. PÉNALITÉ DE L'ÉTIQUETTE *⟨NOMBRE⟩*

On estime la probabilité de *⟨nombre⟩* de la même façon que n'importe quel autre mot à partir des textes normalisés. Cependant, la probabilité résultante est un peu trop élevée, ce qui occasionne des erreurs de reconnaissance du type « 7 » au lieu de « cette ». On laisse quand mème inchangée la probabilité des ngrams avec *⟨nombre⟩*. Lors de la construction, on leur applique une pénalité.

Plus la pénalité est grande, moins l'étiquette *⟨nombre⟩* est probable donc, moins il y a de nombres dans la transcriptions résultante. Ainsi, au lieu d'avoir « 7 », on aura « cette », au lieu de « 1 », on aura « un », etc. À l'inverse, plus la pénalité est petite, plus on a de chiffres dans la transcription. Par exemple, au lieu de reconnaître « cette », on aura « 7 ». Il y aura donc plus d'insertions de nombres avec une petite pénalité et plus de suppressions avec une grande pénalité.

Les ensembles de test pour le hockey (Jeu et Entracte) et le football (NFL et CFL) sont composés de :

- Jeu : 7 926 mots, 120 nombres
- Entracte : 4 478 mots, 149 nombres
- NFL : 8 547 mots, 267 nombres
- CFL : 7 728 mots, 324 nombres

Les tableaux 1, 2, 3 et 4 démontrent l'effet de la pénalité sur le nombre d'erreurs reliées aux chiffres et le taux de reconnaissance en fonction du domaine. La première colonne indique la pénalité appliquée à l'étiquette *⟨nombre⟩*; la deuxième colonne indique la quantité de nombres reconnus; les colonnes 3 ,4, 5 et 6 donnent le détail des erreurs de reconnaissance reliées aux nombres (D est pour suppression, S pour substitution, I pour insertion et Err est le total de ces trois variables); la septième colonne donne le pourcentage de nombre d'erreurr de chiffres par rapport à la référence, soit le ratio entre la colonne 6 et la colonne 2.

Comme on le voit, en ne pénalisant pas l'étiquette *⟨nombre⟩* (pénalité = 0), on a plus d'erreurs reliées aux nombres et le taux de reconnaissance est moins bon. Cependant, en appliquant une trop grosse pénalité (3 +), le nombre de chiffres sortant est trop petit, ce qui occassionne plus d'erreurs. En général, il semble qu'on ait les meilleurs résultats entre 1 et 2.

JEU						
Pénalité	**# nombres**	**D**	**S**	**I**	**Err**	**% Err**
0	125	1	13	6	20	16,0
1	111	2	10	3	15	13,5
1,5	110	0	10	2	12	10,9
1,75	109	1	10	2	13	11,9
2	107	2	11	2	15	14,0
3	101	5	14	2	21	20,8
4	81	6	28	3	37	45,7

Tableau 1 – Effet de la pénalité sur les erreurs de reconnaissance pour le modèle Jeu

ENTRACTE						
Pénalité	**# nombres**	**D**	**S**	**I**	**Err**	**% Err**
0,0	169	2	20	8	30	17,8
1,0	155	2	12	5	19	12,3
1,5	151	2	11	4	17	11,3
1,75	149	3	13	4	20	13,4
2,0	148	3	13	3	19	12,8
3,0	138	3	14	1	18	13,0
4,0	119	1	29	1	31	26,1

Tableau 2 – Effet de la pénalité sur les erreurs de reconnaissance pour le modèle Entracte

NFL						
Pénalité	**# nombres**	**D**	**S**	**I**	**Err**	**% Err**
0,0	289	1	20	16	37	12,8
1,0	262	0	17	4	21	8,0
1,5	256	0	17	2	19	7,4
1,75	255	0	18	2	20	7,8
2,0	253	0	19	1	20	7,9
3,0	233	2	28	1	31	13,3
4,0	204	4	51	3	58	28,4

Tableau 3 – Effet de la pénalité sur les erreurs de reconnaissance pour le modèle NFL

CFL						
Pénalité	**# nombres**	**D**	**S**	**I**	**Err**	**% Err**
0	326	2	18	10	30	9,2
1	317	2	13	6	21	6,6
2	312	2	14	5	21	6,7
3	299	3	16	2	21	7,0
4	266	6	42	2	50	18,8

Tableau 4 – Effet de la pénalité sur les erreurs de reconnaissance pour le modèle CFL

JEU			
P(⟨unk⟩)	**# insertions**	**# gobe-touts**	**Acc**
-3,33	179	35	94,25
-2,5	150	74	95,03
-2,0	122	134	95,66
-1,5	98	201	96,05

Tableau 5 – Effet de *P(⟨unk⟩)* sur les erreurs de reconnaissance poule modèle Jeu

De plus, on remarque que plus la pénalité est grande et moins on a de chiffres dans la transcription. Par exemple, pour le domaine NFL, on a 289 nombres avec une pénalité de 0, et 204 avec une pénalité de 4. On remarque qu'on a environ la même quantité de nombres reconnnus que dans la référence avec une pénalité se situant entre 1 et 2.

On constate que plus la pénalité est grande, plus il manque de chiffres (D plus élevé), et qu'à l'inverse, plus la pénalité est petite, plus on a des chiffres en trop (I élevé). Le bon équilibre entre délétions et insertions semble être entre 1 et 2.

4. PROBABILITÉ DE *⟨UNK⟩*

Le fait que tous les nombres sont connus du modèle affecte la probabilité de rencontrer un mot inconnu *P(⟨unk⟩)* (elle diminue). La probabilité de *⟨unk⟩* est celle que l'on donne aux « gobe-touts » (*fillers*). En variant *P(⟨unk⟩)*, le nombre de « gobe-touts » (*fillers*) et celui de mots insérés dans la transcription varient. Si la probabilité est trop faible, il y a plus d'insertions de petits mots, car les « gobe-touts » (*fillers*) ne sont pas assez probables.

Il faut trouver la probabilité optimale afin d'avoir un haut taux de reconnaissance et le moins d'insertions possibles. Les tableaux 5, 6, 7 et 8 démontrent l'effet de la probabilité de *⟨unk⟩* sur le taux de reconnaissance et le nombre d'insertions et de « gobe-touts » (*fillers*) dans la transcription.

ENTRACTE			
P(⟨unk⟩)	**# insertions**	**# gobe-touts**	**Acc**
-2,0	52	87	93,12
-1,65	44	119	93,36

Tableau 6 – Effet de *P(⟨unk⟩)* sur les erreurs de reconnaissance pour le modèle Entracte

NFL			
P(⟨unk⟩)	**# insertions**	**# gobe-touts**	**Acc**
-2,1	41	62	95,74
-1,9	41	72	95,76
-1,5	37	113	95,70

Tableau 7 – Effet de *P(⟨unk⟩)* sur les erreurs de reconnaissance pour le modèle NFL

CFL			
P(⟨unk⟩)	**# insertions**	**# gobe-touts**	**Acc**
-2,9	46	9	97,01
-2,0	41	28	97,05
-1,5	38	64	96,98

Tableau 8 – Effet de *P(⟨unk⟩)* sur les erreurs de reconnaissance pour le modèle CFL

JEU			
P(⟨unk⟩)	**# insertions**	**# gobe-touts**	**Acc**
-2,16	139	248	89,51
-1,96	149	230	92,43

Tableau 9 – Effet de *P(⟨unk⟩)* sur les erreurs de reconnaissance pour l'ancien modèle Jeu

Domaine	Nombre d'arcs	Err	Acc
Jeu	Ancien 2 300 K	73	90,01
	Nouveau 1 850 K	12	96,03
Entracte	Ancien 1 850 K	108	92,71
	Nouveau 1 850 K	17	93,44
CFL	Ancien 1 690 K	21	94,51
	Nouveau 1 650 K	18	97,15

Tableau 10 – Comparaison du nombre d'erreurs de chiffres et taux de reconnaissance entre les anciens et les nouveaux modèles

Dans nos expériences, se baser sur la probabilité de *⟨unk⟩* dans un modèle où il n'y a pas d'étiquette *⟨nombre⟩,* est une bonne méthode pour déterminer rapidement la valeur.

On note d'ailleurs le même phénoméne d'effritement de la probabilité de *⟨unk⟩* avec l'utilisation répétée de MDE sur plusieurs mois ou années, comme dans le tableau 9.

5. DISCUSSION

Nous comparons maintenant les résultats des modèles utilisant la grammaire de nombres aux modèles originaux. Les anciens modèles comportent à la fois des nombres écrits en lettres et des nombres écrits en chiffres, ce qui rend difficile l'évaluation du taux d'erreur. Nous comparons donc le nombre d'erreurs en comptant les nombres écrits en lettres au nombre d'erreurs en les ignorant.

Tout d'abord, afin de comparer les deux méthodes, il est nécessaire de réduire la taille (en arcs) des modèles originaux. En effet, les modèles de langage avec la classe de nombres doivent être réduits suffisamment afin que, lors de la composition avec la grammaire, il y ait assez de mémoire.

Le tableau 10 montre le nombre d'erreurs de nombres taux de reconnaissance du modèle d'origine en fonction du nombre d'arcs dans le modèle.

Le modèle Entracte original fait 10 erreurs, par rapport à 17 pour le modèle Entracte avec classe de chiffres de la même taille. Cependant, le modèle Entracte original a écrit seulement 29 nombres en chiffres. Les 98 autres nombres sont composés de plusieurs mots, par exemple « quatre vingt quatorze », ce qui porte à 108 le nombre d'erreurs. Finalement, le modèle n'a reconnu que 127 nombres, alors que la transcription en contient 149. Le modèle avec grammaire de nombres a reconnu 151 nombres, dont seulement 17 erreurs.

Le modèle Jeu original fait 37 erreurs de nombres. En comptant les 36 nombres écrits en lettres, on obtient 73 erreurs, par rapport à 12 pour le modèle avec classe de nombres. Finalement, le modèle CFL original fait 21 erreurs par rapport à 18 pour le modèle avec la grammaire. Dans tous les cas, il y a moins d'erreurs de nombres avec le modèle utilisant la grammaire de nombres. De plus, étant donné que tous les nombres peuvent être reconnus grâce à cette grammaire, il y aura nécessairement moins de mots hors-vocabulaire avec les nouveaux modèles, ce qui réduit également le taux d'erreurs global.

6. CONCLUSION

En résumé, la grammaire de nombres résout les problèmes de nombres remplacés par plusieurs chiffres consécutifs, à condition que les textes d'entraînement soient normalisés sans bigrammes de *⟨nombre⟩*. La perte de performance due au plus petit nombre d'arcs est compensée par les gains de performance sur les nombres.

De plus, on a vu qu'appliquer une pénalité aux transitions comportant l'étiquette *⟨nombre⟩* améliorait les taux de reconnaissance et diminuait le nombre d'insertions dans la transcription.

Finalement, la probabilité de *⟨unk⟩* doit être ajustée, de telle sorte que les *fillers* sortent facilement en reconnaissance.

En général, on regagne avec la grammaire de nombres ce que l'on perd en quantité d'arcs. À la fin, on a donc une meilleure performance même avec un modèle plus petit.