Segmentation multiple d'un flux de données textuelles pour la modélisation statistique du langage

Sopheap Seng (1, 2), Laurent Besacier (1), Brigitte Bigi (1), Eric Castelli (2)

(1)Laboratoire LIG/GETALP, Grenoble France
{Sopheap.Seng, Laurent.Besacier, Brigitte.Bigi}@imag.fr
(2)Laboratoire MICA, CNRS/UMI-2954, Hanoi Vietnam
Eric.Castelli@mica.edu.vn

Résumé Dans cet article, nous traitons du problème de la modélisation statistique du langage pour les langues peu dotées et sans segmentation entre les mots. Tandis que le manque de données textuelles a un impact sur la performance des modèles, les erreurs introduites par la segmentation automatique peuvent rendre ces données encore moins exploitables. Pour exploiter au mieux les données textuelles, nous proposons une méthode qui effectue des segmentations multiples sur le corpus d'apprentissage au lieu d'une segmentation unique. Cette méthode basée sur les automates d'état finis permet de retrouver les n-grammes non trouvés par la segmentation unique et de générer des nouveaux n-grammes pour l'apprentissage de modèle du langage. L'application de cette approche pour l'apprentissage des modèles de langage pour les systèmes de reconnaissance automatique de la parole en langue khmère et vietnamienne s'est montrée plus performante que la méthode par segmentation unique, à base de règles.

Abstract In this article we deal with the problem of statistical language modelling for under-resourced language with a writing system without word boundary delimiters. While the lack of text resources has an impact on the performance of language models, the errors produced by the word segmentation makes those data less usable. To better exploit the text resources, we propose a method to make multiples segmentations on the training corpus instead of a unique segmentation. This method based on finite state machine allows obtaining the n-grams not found by the unique segmentation and generate new n-grams. We use this approach to train the language models for automatic speech recognition systems of Khmer and Vietnamese languages and it proves better performance than the unique segmentation method.

Mots-clés : segmentation multiple, langue non segmentée, modélisation statistique du langage

Keywords: multiple segmentation, unsegmented language, statistical language modeling

1 Introduction

Un modèle statistique du langage est une distribution de probabilités sur des mots ou suites de mots. Il permet de classer les mots ou les phrases selon leur probabilité d'apparition. Son objectif est d'assigner relativement une grande probabilité aux séquences de mots fréquentes, significatives, grammaticalement correctes et une faible probabilité aux séquences de mots rares, insensées ou grammaticalement incorrectes. Les modèles de langage sont utilisés dans des applications telles que la reconnaissance automatique de la parole, la reconnaissance automatique de l'écriture manuscrite, la correction orthographique, la traduction automatique et toute autre application introduisant une composante linguistique. La nature statistique des approches utilisées dans la modélisation du langage par n-grammes, nécessite une grande quantité de données textuelles pour obtenir une estimation précise des probabilités. Ces données ne sont pas disponibles en grande quantité pour les langues dites peu dotées et le manque de données d'apprentissage a un impact direct sur les performances des modèles de langage.

Tandis que le mot est généralement l'unité de base dans la modélisation statistique du langage, l'identification de mots dans un texte n'est pas une tâche simple même pour les langues qui séparent les mots par un caractère (un espace en général). Pour les langues dites non segmentées qui possèdent un système d'écriture sans séparation évidente entre les mots, les n-grammes de mots sont estimés à partir de corpus d'apprentissage segmentés en mots en utilisant des méthodes automatiques. La segmentation automatique n'est pas une tâche triviale et introduit des erreurs à cause des ambiguïtés de la langue naturelle et la présence de mots inconnus dans le texte à segmenter. Alors que le manque de données textuelles a un impact sur la performance des modèles de langage, les erreurs introduites par la segmentation automatique peuvent rendre ces données encore moins exploitables. Une alternative possible consiste à calculer les probabilités à partir d'unités sous-lexicales. Parmi les travaux existants qui utilisent des unités sous-lexicales pour la modélisation du langage, nous pouvons citer (Kurimo, 2006), (Abdillahi, 2006) et (Afify, 2006) qui utilisent les morphèmes respectivement pour la modélisation de l'arabe, du finnois, et du somalien. Pour une langue non-segmentée comme le japonais, le caractère (idéogramme) est utilisé dans (Denoual, 2006). Dans un travail précédent sur la reconnaissance automatique de la parole en langue khmère[1] (Seng, 2008), nous avons exploité les différentes unités lexicales et sous-lexicales (mot, syllabe et groupe de caractères[2]) dans la modélisation du langage de cette langue peu dotée. Nous avons proposé des modèles de langage simples basés sur le mot, la syllabe, le groupe de caractères. Notre objectif était de comparer la performance de ces différentes unités et nous avons observé que le mot reste l'unité la plus performante.

Dans cet article, nous traitons du problème de la modélisation statistique du langage à base de mots pour les langues sans segmentation évidente entre les mots. Tandis que le manque de données textuelles a un impact sur la performance des modèles, les erreurs introduites par la segmentation automatique peuvent rendre ces données encore moins exploitables. Les n-

---

[1] Le khmer est la langue officielle du Cambodge

[2] En khmer, un groupe de caractères ou un cluster de caractères (CC) est une séquence de caractères inséparables et possède une structure bien définie. La segmentation d'un texte khmer en CC est triviale et peut se faire à bases des règles.

grammes de mots non trouvés dans le corpus d'apprentissage peuvent l'être à cause d'erreurs de segmentation mais aussi parce qu'une séquence de caractères peut avoir plusieurs segmentations correctes mais une seule segmentation a été considérée dans le corpus d'apprentissage. Dans un objectif consistant à mieux exploiter les données textuelles en utilisant les différentes vues sur les mêmes données, nous proposons une méthode qui effectue des segmentations multiples sur le corpus d'apprentissage au lieu d'une segmentation unique. Cette nouvelle méthode de segmentation basée sur des automates d'état finis permet de générer toutes les segmentations possibles à partir d'une séquence de caractères et nous pouvons ensuite en extraire les n-grammes. Elle permet de retrouver les n-grammes non trouvés par la segmentation unique et d'ajouter de nouveaux n-grammes dans le modèle de langage. L'application de cette approche pour l'apprentissage des modèles de langage pour les systèmes de reconnaissance automatique de la parole en langue khmère et vietnamienne s'est montrée plus performante que la méthode classique par segmentation unique. Dans les sections suivantes, nous allons d'abord faire un état de l'art sur les méthodes de segmentation automatique en mots avant de présenter notre méthode exploitant des segmentations multiples et les résultats d'expérimentations sur le khmer et le vietnamien.

2 Segmentation automatique en mots

2.1 Etat de l'art

La segmentation de textes est l'une des tâches fondamentales dans le traitement automatique des langues naturelles (TALN). Beaucoup d'applications de TALN nécessitent en entrée des textes segmentés en mots avant d'effectuer les autres traitements car le mot est considéré comme l'unité linguistique et sémantique de référence. Pour des langues comme le français et l'anglais, il est assez naturel de définir un mot comme une séquence de caractères séparés par des espaces. Cependant, pour les langues non segmentées, la segmentation en mots n'est pas un problème simple. A cause des ambiguïtés dans la langue naturelle, une séquence de caractères peut être segmentée de plusieurs façons. Cette ambiguïté ne pose pas vraiment de problème pour l'être humain, peut être à cause du fait qu'une segmentation incorrecte donne généralement une phrase incompréhensible. De plus, il peut exister des désaccords entre différentes personnes sur la segmentation d'une phrase donnée. Ce désaccord existe car il y a souvent différentes conventions de segmentation et la définition du mot dans une langue est souvent ambiguë.

La technique générale de segmentation en mots emploie un algorithme qui recherche dans un dictionnaire les mots correspondant à ceux du texte et qui, en cas d'ambiguïté, sélectionne celui qui optimise un paramètre dépendant de la stratégie choisie. Dans les stratégies les plus courantes, l'optimisation consiste à :

- maximiser la taille des mots, pris un par un de gauche à droite, avec retour arrière en cas d'échec (« plus longue chaîne d'abord » ou « longest matching »),
- minimiser le nombre de mots dans la phrase entière (« plus petit nombre de mots » ou « maximal matching »).

Ces techniques recourent intensivement à des dictionnaires, qu'il faut donc créer. Bien que cela puisse être fait automatiquement par apprentissage à partir d'un corpus, ces dictionnaires ont souvent été créés manuellement. Les travaux de recherche sur la segmentation

automatique en mots de la langue chinoise et thaïe sont très actifs. Parmi les travaux qui utilisent ces techniques, nous pouvons citer (Li, 1998) pour le chinois et (Haruechaiyasak, 2008) pour le thaï. La performance de ces méthodes est acceptable en général mais elle dépend fortement de la taille et de la qualité des dictionnaires utilisés pour la segmentation. La performance diminue en présence de cas d'ambiguïté et de mots inconnus (voir tableau 1 pour les résultats de la segmentation des textes khmers).

Il existe des méthodes plus élaborées qui utilisent des méthodes statistiques et/ou passent par une phase d'apprentissage. Dans (Wu, 2003), pour une phrase chinoise à segmenter, un treillis de tous les mots possibles est construit en fonction d'un vocabulaire. Ensuite, des méthodes statistiques sont appliquées pour décoder le chemin le plus probable sur le treillis. Une méthode statistique et linguistique de segmentation en mots est aussi proposée et implémentée sur la langue thaïe (Meknavin, 1997). Dans cette méthode, le contexte des mots est analysé linguistiquement pour déterminer la segmentation la plus probable.

Les méthodes de l'état de l'art utilisent la combinaison de dictionnaires avec les statistiques pour obtenir un meilleur résultat. Cependant, les méthodes statistiques nécessitent de disposer d'un grand corpus de texte segmenté au préalable manuellement. Les méthodes statistiques et les méthodes d'apprentissage complexes ne sont pas appropriées dans notre contexte des langues peu dotées car les ressources nécessaires pour implémenter ces méthodes n'existent pas. Pour une langue considérée, nous cherchons des méthodes de segmentation performantes, rapides, faciles à implémenter et qui tirent, au mieux, bénéfice des ressources limitées existantes pour la langue.

2.2 Segmentation automatique de la langue khmère

Pour illustrer l'impact des mots hors-vocabulaire sur la performance des méthodes de segmentation automatique à base de dictionnaire, nous développons les outils de segmentation automatique de textes khmers en utilisant les deux critères d'optimisation : « plus longue chaîne d'abord » (longest matching) et « plus petit nombre de mots » (maximal matching). Notre corpus de test contient 1000 phrases. Après la segmentation manuelle, nous obtenons 31042 mots et un dictionnaire de 4875 mots. Nous enlevons ensuite les mots les moins fréquents du dictionnaire de départ pour créer des dictionnaires avec taux de mots hors-vocabulaire croissants (de 5% à 50%) par rapport au corpus de test. Les performances de segmentation sont présentées dans le tableau 1.

Taux des mots hors vocabulaire	**Performance de la segmentation (%)**	
	Maximal Matching	**Longest Matching**
0%	91,6	91,7
5%	90,1	90,2
10%	90,2	90,3
20%	86,3	86,9
30%	82,6	83,5
40%	75,7	77,2
50%	68,8	72,4

Table 1 : Taux des mots corrects pour deux méthodes de segmentation à base de dictionnaire en fonction du taux de mots hors-vocabulaire

Nous observons que, dans le cas d'absence de mots hors vocabulaire, la performance est autour de 92% pour les deux méthodes mais la performance chute à 69% et 72% quand il y a 50% des mots hors vocabulaire dans le corpus à segmenter. Pour les langues peu dotées, il est difficile d'obtenir un dictionnaire avec un taux de mots hors-vocabulaire faible. Dans ce cas, on risque donc d'atteindre une mauvaise performance de segmentation automatique sur le corpus d'apprentissage et la performance du modèle du langage appris à partir de ce corpus mal segmenté sera alors mauvaise.

3 Segmentation multiple pour la modélisation statistique du langage

3.1 Pourquoi une segmentation multiple ?

Contrairement à la segmentation unique décrite dans la section précédente qui recherche dans une séquence de caractères la meilleure segmentation selon un critère d'optimisation, notre approche par segmentations multiples cherche à générer, à partir d'une séquence de caractères, toutes les séquences des mots valides (basant sur un dictionnaire). C'est à partir de toutes ces séquences de mots que des n-grammes seront comptés pour l'apprentissage du modèle de langage.

Phrase	ព្រះពុទ្ធជាព្រះបរមគ្រូនៃយើង								3-grams Count
Segmentation 1	ព្រះពុទ្ធ w_1		ជា w_2	ព្រះ w_3	បរមគ្រូ w_4		នៃ w_5	យើង w_6	$w_1\ w_2\ w_3$ $w_2\ w_3\ w_4$ $w_3\ w_4\ w_5$ $w_4\ w_5\ w_6$
Segmentation 2	ព្រះពុទ្ធ w_1		ជា w_2	ព្រះ w_3	បរម w_7	គ្រូ w_8	នៃ w_5	យើង w_6	$w_2\ w_3\ w_7$ $w_3\ w_7\ w_8$ $w_7\ w_8\ w_5$
Segmentation 3	ព្រះ w_3	ពុទ្ធ w_9	ជា w_2	ព្រះ w_3	បរម w_7	គ្រូ w_8	នៃ w_5	យើង w_6	$w_3\ w_9\ w_2$ $w_9\ w_2\ w_3$
Traduction	Le bouddha est notre maître suprême								

Figure 1 : Exemple de la segmentation multiple d'une phrase en khmer

Figure 1 montre un exemple de la segmentation multiple d'une phrase en khmer. Nous montrons trois segmentations possibles d'une séquence de caractères en khmer. La segmentation 1 correspond bien à la segmentation unique de type « longest matching ». Dans le cas de la segmentation unique (segmentation 1), nous obtenons 4 tri-grammes. Si nous appliquons la segmentation multiple sur cette phrase, nous avons au total 9 tri-grammes. 5 nouveaux tri-grammes sont obtenus à partir des deux autres segmentations (segmentation 2 et 3). Il est à noter que nous ne comptons qu'une seule fois un tri-gramme qui se présente plusieurs fois dans les segmentations multiples d'un phrase.

Par rapport à la segmentation unique, la segmentation multiple permet d'obtenir plus de n-grammes. Nous pouvons diviser ces nouveaux n-grammes en trois différentes catégories :

1. des n-grammes de mots qui sont effectivement présents dans le corpus d'apprentissage d'origine, non segmenté, mais à cause d'erreurs introduites par la segmentation unique, ils ne sont pas retrouvés après la segmentation.

2. des n-grammes de mots qui sont effectivement présents dans le corpus d'apprentissage d'origine, non segmenté, mais comme une séquence de caractères peut avoir plusieurs segmentations correctes et qu'un seul choix est effectué lors de la segmentation unique, ils ne sont pas alors retrouvés après la segmentation.

3. des n-grammes de mots qui ne sont pas présents dans le corpus d'apprentissage même si la segmentation est parfaitement correcte. Dans ce cas, la segmentation multiple génère ces n-grammes parce qu'il est possible de segmenter entièrement une phrase en une séquence de mots valides (même si cela donne une phrase insensée) mais aussi parce que notre méthode de segmentation multiple permet également de générer localement les séquences de mots dans une phrase en marquant les parties restantes qui ne correspondent pas aux mots valides comme « mot inconnu ».

Les n-grammes de catégorie 1 et 2 sont des n-grammes potentiellement utiles pour la modélisation du langage car il s'agit de séquences de mots valides de la langue et ils sont effectivement présents dans le corpus d'apprentissage. Les n-grammes de catégorie 3 peuvent perturber la modélisation.

Nous développons un outil de segmentation multiple qui permet de sortir les N_{seg} meilleures segmentations à partir d'une séquence de caractères donnée en entrée. Nous allons décrire dans la section suivante comment la segmentation multiple est implémentée.

3.2 Segmentation multiple utilisant les automates d'état fini

Notre outil de segmentation multiple est développé à l'aide d'automates d'état fini en utilisant la boîte à outils de AT&T *FSM toolkit* (Mohri, 2002). L'algorithme utilisé est inspiré des travaux sur la segmentation des mots arabes de (Zitouni, 2006) et (Lee, 2003). La segmentation multiple d'une séquence de caractères est faite à l'aide de la composition de trois automates. Le premier automate est un transducteur qui génère un treillis avec tous les segments possibles quand une séquence de caractères est donnée en entrée. Le deuxième automate peut être vu comme un dictionnaire sous forme de transducteur qui accepte les caractères et produit les séquences correspondant aux mots contenus dans le dictionnaire qui doit être disponible au début de l'algorithme. Le troisième automate est un modèle de langage qui peut assigner les scores à chaque séquence dans le treillis. Nous composons ces trois automates pour produire un treillis d'hypothèses de segmentation en mots, à partir d'une entrée en caractères (ou en syllabes pour le vietnamien). En parcourant ce treillis, nous pouvons générer les N_{seg} meilleures segmentations pour une entrée donnée. Les N_{seg} meilleures segmentations obtenues sont ensuite utilisées pour compter le nombre des n-grammes selon la méthode de comptage présentée dans figure 1.

4 Expérimentations

Les expérimentations sont menées sur deux langues peu dotées et non segmentées, le khmer et le vietnamien. Pour comparer les performances de la segmentation multiple et la segmentation unique à base de dictionnaire dans la modélisation statistique du langage, nous

apprenons des modèles de langage trigrammes à partir des corpus d'apprentissage segmentés en mots en utilisant ces deux approches de segmentation. Pour observer l'impact du nombre de segmentations multiples sur la performance des modèles de langage, nous effectuons plusieurs tests en faisant la segmentation multiple sur les corpus d'apprentissage en faisant varier le nombre N_{seg} de meilleures segmentations pour chaque phrase de 2 à 1000. A l'aide d'un corpus de développement, nous comparons la couverture en trigrammes (*trigram hits*) de ces modèles de langage et leur perplexité. Nous évaluons ensuite les performances de ces modèles de langage en les utilisant dans un système de reconnaissance automatique de la parole.

4.1 Expérimentations sur le khmer

Le khmer est la langue officielle du Cambodge parlée par plus de 15 millions de personnes dans le monde. Elle appartient au groupe des langues môn-khmères. Elle est classée comme une langue peu dotée car les ressources linguistiques et les services pour le traitement automatique de la langue ne sont pas encore bien développés. Au niveau de l'écriture, le khmer est écrit sans espaces entre les mots.

Notre corpus d'apprentissage de la langue khmère contient environ un demi million de phrases de type *news*. Un dictionnaire de 20k mots extraits du dictionnaire *Chuon Nath* de l'Institut Bouddhique du Cambodge est utilisé dans cette expérimentation. La segmentation unique à base de ce dictionnaire avec le critère d'optimisation « longest matching » donne un corpus de 15 millions de mots. Cinq autres corpus sont obtenus en effectuant les segmentations multiples avec le nombre de N_{seg} meilleures segmentations qui varie de 2 à 1000. Il est à noter que la segmentation multiple utilise le même dictionnaire que la segmentation unique. Le comptage des n-grammes est effectué sur ces corpus et les modèles de langage n-gramme sont ensuite appris en utilisant ce même dictionnaire de 20k mots.

Un corpus de développement (*dev*) de 370 phrases (11k mots après la segmentation unique) est utilisé pour évaluer la couverture en trigrammes (*trigram hits*) et la perplexité des modèles de langage du khmer. Nous présentons dans le tableau 2 le nombre de trigrammes dans les modèles de langage, la couverture en trigrammes de ces modèles, la perplexité et la performance du système de reconnaissance automatique de la parole en langue khmère (sur un corpus de test constitué de 160 phrases de type news et dont les transcriptions sont différentes de l'ensemble de *dev*) qui utilise ces modèles dans le décodage. Les détails sur le système de reconnaissance automatique en langue khmère (décodeur, modèle acoustique) sont donnés dans (Seng, 2008).

	Les modèles de langage issus des différentes segmentations							
	M_Unique	**M_2**	**M_5**	**M_10**	**M_50**	**M_100**	**M_500**	**M_1000**
Nombre de trigrammes dans le modèle de langage (million)	5,67	7,34	8.95	10,17	12,52	13,31	14,85	15,41
Nombre de *trigram hits* sur *dev*	3404	3744	3799	3867	4020	4065	4162	4204
% trigram hits sur *dev*	31%	34,1%	34,6%	35,2%	36,6%	37%	37,9%	38,3%
Perplexité sur *dev*	394,9	**322,5**	348,8	361.8	373,9	374,7	378	378
Taux d'erreur Reco. sur *test*	22%	21.7%	20.8%	**20.5%**	20.6%	20.7 %	20.9%	21%

Table 2 : Les résultats des expérimentations en langue khmère

4.2 Expérimentations sur le vietnamien

Le vietnamien est la langue officielle du Vietnam. Elle est parlée par environ 70 millions de personnes dans le monde. Son origine est toujours sujette à débat parmi les linguistes. Il est cependant généralement admis qu'elle a des racines communes et fortes avec le môn-khmer qui fait partie de la branche austro asiatique. L'orthographe est latine depuis le XVIIè siècle, avec des caractères accentués pour les tons. Le vietnamien est écrit avec les espaces entre les syllabes mais ces espaces ne marquent pas les frontières entre les mots dans une phrase car un mot peut se composer d'une ou plusieurs syllabes. La figure 2 donne un exemple d'une phrase de la langue vietnamienne.

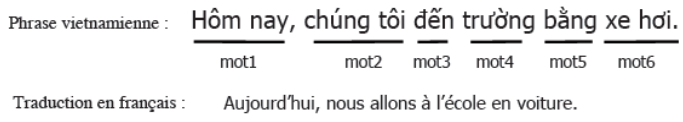

Figure 2 : Exemple d'une phrase vietnamienne

Le corpus d'apprentissage du vietnamien contient 3 millions de phrases soit plus de 56 millions de syllabes. Un dictionnaire de 30k mots extraits à partir d'un dictionnaire bilingue Vietnamien-Français est utilisé dans cette expérimentation. Après la segmentation unique automatique à base de ce dictionnaire avec le critère d'optimisation « longest matching », nous obtenons un corpus de 46 millions de mots. Les segmentations multiples sont effectuées avec les nombres de N_{seg} variant de 2 à 1000. Les modèles de langage de trigrammes sont ensuite appris à partir de ces corpus en utilisant un dictionnaire de 30k mots (cf expérimentation sur le khmer).

Un corpus de développement (*dev*) de 1000 phrases (44k mots après la segmentation unique) est utilisé pour évaluer la couverture en trigramme et la perplexité des modèles de langage. Les performances de reconnaissance de la parole sont estimées sur un corpus de test de 400 phrases de type *news* (dont les transcriptions sont différentes de l'ensemble de *dev*). Les détails sur le système de reconnaissance automatique en langue vietnamienne sont donnés dans (Le, 2008). Les résultats des expérimentations sur le vietnamien sont dans le tableau 3.

	Les modèles issus des différentes segmentations							
	M_Unique	**M_2**	**M_5**	**M_10**	**M_50**	**M_100**	**M_500**	**M_1000**
Nombre de trigrammes dans le modèle de langage (million)	20.32	24,06	28,92	32,82	34,2	34,9	35.83	36.8
Nombre de *trigram hits* sur le *dev*	15901	16190	16384	16458	16547	16569	16593	16614
% de trigram hits sur le *dev*	47,7%	48,6%	49,2%	49,4%	49,7%	49,7%	49,8%	49,9%
Perplexité sur le *dev*	118,9	**118,1**	125,9	129	133,4	134,8	136,9	137,6
Taux d'erreur de Reco sur le *test*	36,5%	**35,5**%	36%	36,1%	36,1%	36,2%	36,5%	36,5%

Table 3 : Les résultats d'expérimentation sur la langue vietnamienne

4.3 Discussion

A travers les résultats d'expérimentations sur le khmer et le vietnamien, nous pouvons constater que l'approche par segmentations multiples permet de générer des nouveaux trigrammes par rapport à la segmentation unique, quand le nombre de N_{seg} meilleures segmentations est augmenté Cette augmentation de nombre de trigrammes dans le model du

langage améliore la couverture en trigrammes et la perplexité. Cette amélioration montre que les nouveaux trigrammes générés par la segmentation multiple sont pertinents pour la modélisation statistique du langage. Dans le cas du khmer, la meilleur taux d'erreurs du système de reconnaissance automatique de la parole est obtenue avec le model du langage *M_10* et la performance drops si nous continuons à augmenter le nombre de N_{seg} meilleures segmentations. Cela montre qu'à partir d'un certain niveau de segmentation, quand on augmente encore N_{seg}, on ajoute beaucoup de mauvais trigrammes et cela perturbe la bonne répartition des probabilités dans le modèle du langage. Ce phénomène peut être observé clairement dans le cas de la langue vietnamienne : la couverture en trigramme n'augmente que de 0,2% quand on augmente le nombre de N_{seg} meilleures segmentations de 50 à 1000 mais on ajoute plus de 2,5 millions de nouveaux trigrammes dans le modèle. La meilleur taux d'erreurs du système de reconnaissance automatique de la parole dans le cas de vietnamien est obtenue avec le nombre de segmentation N_{seg}= 2. Avec une analyse plus détaillée sur le corpus d'apprentissage vietnamien, nous avons constaté que près de 80% des mots dans le corpus sont les mots monosyllabiques et seulement 20% qui sont multi-syllabiques. Cela veut dire qu'il n'y pas beaucoup de bonne segmentations possibles que l'on peut générer comparant à la langue khmère.

5 Conclusion

Nous proposons dans cet article une approche qui consiste à effectuer des segmentations multiples sur le corpus d'apprentissage pour la modélisation statistique du langage dans le contexte des langues peut dotées et non segmentées. Cette approche permet de retrouver les n-grammes non trouvés par la segmentation unique et de générer de nouveaux n-grammes dans les modèles. L'application de cette méthode pour l'apprentissage des modèles de langage pour les systèmes de reconnaissance automatique de la parole en langue khmère et vietnamienne s'est montrée plus performante (en perplexité et en taux d'erreur de reconnaissance) que la méthode par segmentation unique.

Références

Abdillahi N. et al. (2006). Automatic transcription of Somali language. Interspeech'06. 289-292. Pittsburgh, PA

Afify M. et al. (2006) On the use of morphological analysis for dialectal Arabic Speech Recognition. Interspeech'06, 277-280. Pittsburgh, PA

Denoual E., Lepage Y. (2006). The character as an appropriate unit of processing for non-segmenting languages. NLP Annual Meeting. 731-734, Tokyo Japan

Haruechaiyasak C., Kongyoung S., et Dailey M.N. (2008). A Comparative Study on Thai Word Segmentation Approaches. In Proceedings of ECTI-CON. 125-128. Thailand

Kurimo M. et al. (2006). Unsupervised segmentation of words into morphemes - Morpho Challenge 2005: Application to Automatic Speech Recognition. Interspeech'06. 1021-1024. Pittsburgh, PA

Le V.B., Besacier L., Seng S., Bigi B., DO T.N.D. (2008). Recent Advances in Automatic Speech Recognition for Vietnamese. International Workshop on Spoken Languages Technologies for Under-Ressourced Languages. SLTU'08 Hanoi Vietnam

Lee, Y., Papineni, K., Roukos, S., Emam, O., et Hassan, H. (2003). Language model based arabic word segmentation. In Proceedings of the 41st Annual Meeting on Association For Computational Linguistics - Volume 1 399-406. Sapporo. Japan.

Li H., Yuan B. (1998). Chinese word segmentation. Proceedings of the 12th Paci Asia Conference on Language, Information and Computation. PACLIC-12. Singapore

Meknavin S., Charoenpornsawat P., Kijsirikul B. (1997). Feature-based Thai Word Segmentation. NLPRS'97. Phuket. Thailand

Mohri M., Pereira F., et Riley M. (2002). Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language. 16(1) 69-88

Seng S., Sam S., Le V. B., Besacier L. et Bigi B. (2008). Which Units for Acoustic and Language Modelling for Khmer Automatic Speech Recognition? SLTU'08. 33-38, Hanoi Vietnam

Wu A. (2003) Chinese word segmentation in MSR-NLP, SIGHAN Workshop on Chinese Language Processing. Sapporo. Japan

Zitouni I. (2006). Finite state based Arabic word segmentation. ArabTEXtest for Ali Farghaly. CSLI Publication.