Université de Paris

Segmentation phonémique du Yongning Na

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

mars 2020

Ce TP sera noté : le code source devra être envoyé ainsi qu'un court rapport (au format pdf) comportant la réponse aux différentes questions ainsi que la totalité du code source avant le 21 avril 8h.

1 Contexte

L'objectif de ce TP est d'implémenter un système capable de segmenter des transcriptions phonémiques en séquence de phonèmes pour le Yongning Na. Le Yongning Na est une langue tonale de la famille des langues sino-tibétaines parlée dans le centre sud de la province du Sichuan en Chine. C'est une langue aujourd'hui considérée en danger et développer des outils de TAL pour aider à sa documentation est un sujet de recherche particulièrement actif [1].

Le développement de système de transcription phonémique automatique constitue une première étape pour aider les linguistes de terrain dans leur travail de description et de documentation de la langue. Une première série d'expériences [2] a montré qu'il était possible, à l'aide de technique d'apprentissage automatique, de développer de tels systèmes à partir des annotations produites par des linguistes de terrain. Il faut toutefois, avant d'utiliser de tels systèmes, transformer les transcriptions en séquences de phonèmes (contenant uniquement les unités devant être prédites par le système). Par exemple la phrase :

si˧dzi˩-ʈʂʰɯ˩,␣|␣ʈʰææ˧␣|␣ʈʂʰɯ˧-bv̩˧-ʑ˧␣| dɑ˧-kv̩˥-mæ˩!␣|

devra être transformée en :

1. La vidéo « Versus | Faut-il sauver les langues en danger ? » explique l'intérêt et les enjeux de l'utilisation du TAL pour la documentation linguistique

2. Une description détaillée de ces expériences est faite dans l'article « Integrating automatic transcription into the language documentation workflow : Experiments with Na data and the Persephone toolkit »

s i ˦ dz i ˩ ʈʂʰ ɯ ˩ | ʈʰ æ æ̃ ˦ | ʈʂʰ ɯ ˦ b ɣ ˦ i ˦ | d ɑ ˦ k ɣ ˥ m æ ˩ |

L'objectif de ce TP est de développer un système capable de faire cette segmentation.

2 Segmentation du Yongning Na

La structure des phonèmes du Yongning Na est simple (d'un point de vue d'informaticien) : pour segmenter une transcription, il « suffit » :

1. de considérer la transcription d'un segment ;
2. d'identifier le plus grand phonème (en nombre de caractères) commençant celle-ci ;
3. d'ajouter celui-ci à la liste des phonèmes (il vient d'être reconnu !) ;
4. supprimer le préfixe du segment ;
5. recommencer la 1ère étape jusqu'à ce que le segment soit vide.

Ainsi, pour une langue ayant deux phonèmes `a` et `bc`, la segmentation de `abca` comportera trois étapes :

segment	phonème identifié
`abca`	`a`
`bca`	`bc`
`a`	`a`
∅	

1. Écrivez une fonction prenant en entrée une liste de phrases et une liste de phonèmes et qui réalise une segmentation selon le principe que nous venons de décrire.

3 Application à des données réelles

La liste des phonèmes du Yongning Na et des exemples de transcriptions sont disponibles sur le site du cours. Les transcriptions sont extraites de la collection Pangloss. Tous les fichiers sont (naturellement) encodés en UTF-8. En plus des phonèmes, les transcriptions contiennent des informations sur les tons. Il y a 7 tons en Yongning Na :

— ˩ (`U+02E9`) ;
— ˥ (`U+02E5`) ;
— ˦ (`U+02E7`) ;
— ˦ ˥ ;
— ˩ ˥ ;
— ˩ ˦ ;
— ˦ ˩.

2. Pourquoi les tons peuvent-ils être identifiés en utilisant l'algorithme décrit dans la section précédente ?

Les transcriptions sont organisées de la manière suivante
— chaque ligne du fichier correspond à la transcription d'un segment ;
— chaque ligne est composée de deux parties séparées par « @@@ » : la première partie comporte la transcription telle qu'elle est extraite de la collection Pangloss ; la seconde la segmentation de la transcription en phonèmes (c.-à-d. la sortie que vous devriez obtenir).

3. Implémentez l'algorithme décrit dans la section précédente pour segmenter les transcriptions.
4. Dans combien de cas arrivez-vous à retrouver la segmentation de référence ? Quel est l'intérêt de comparer les segmentations obtenues à des segmentations de référence ?
5. Quels sont les 10 phonèmes les plus fréquents ? les moins fréquents ? Y a-t-il des phonèmes qui ne sont pas représentés dans le corpus ?

La mise en œuvre de l'algorithme décrit dans la section précédente est compliqué par le fait que les transcriptions ne comportent pas que les phonèmes : il y a également des signes des ponctuations, des commentaires, ... qu'il faut traiter lors de la segmentation.

Lors de la mise en œuvre de l'algorithme décrit dans la section précédente, il faudra veillez à :
— le caractère ◊ (U+25CA) devra être transformé en | ;
— supprimer tous les caractères n'apparaissant pas des les phonèmes (notamment les ponctuations) à l'exception des marqueurs de groupes tonaux | ;
— tous les caractères entre crochets doivent être supprimés (y compris les crochets) [3]
— les chevrons (< et >) doivent être supprimés (mais le texte entre chevrons doit être conservé) ;
— les conventions de transcriptions doivent être normalisées : w̃æ doit être transformé en w̃æ et ṽ̰ (un v combiné avec un tilde (U+0303) et une ligne verticale supérieure U+030D), en ṽ̩ (la ligne verticale est maintenant sous le symbole et correspond au caractère U+0329)
— normalisé les hésitations : celle-ci sont indiqués par un ə ou un m répétés une fois ou plus et suivis de point de suspension et doivent être tous transformés pour que le symbole (m ou ə) soit répété exactement trois fois suivi de points de suspension [4] ;
— ne pas segmenter les énoncés comportant la chaine BEGAIEMENT (il faut retourner une chaine vide dans ce cas)

3. Ces caractères contiennent des « informations » sur l'énoncé et ne correspondent pas forcément à une partie audio. Un exemple d'une telle transcription est visilbe à l'url https://doi.org/10.24397/pangloss-0004558#S2.

4. Le schwa ə correspond au caractère U+0259, les points de suspensions … au caractère U+2026