Applications Multilingues : Segmentation en mots du Japonais

Carl Goubeau

3 novembre 2014

Master 2 ATAL

Résumé

Ce rapport détaille des améliorations qui ont pu être apportées à la méthode de segmentation proposée par Constantine P. Papageorgiou[1] (*Japanese word segmentation by hidden markov model*)

Le japonais est une langue plus difficile à traiter que les autres en TALN. En effet, l'absence de délimiteurs entre les mots complexifie et rend ambigü le processus de segmentation en mots pour ce langage. Des erreurs de tokenisation sont donc encore fréquentes et ont un impact sur la précision des systèmes et outils qui nécessitent un segmenteur afin de traiter le japonais.

1 Segmenteur Initial

L'approche de réference de ce projet est une implémentation de la méthode proposée par Constantine P. Papageorgiou[1] (*Japanese word segmentation by hidden markov model*) qui utilise un modèle de Markov caché. Ce segmenteur se base sur des bigrammes et des probabilités de transition entre états de césure et de continuité entre les caractères. Cette méthode, qui donne déjà de bons résultats sur notre corpus de test (89% de f-mesure), fixe une probabilité non nulle pour les bigrammes non recontrés lors de l'entraînement, ce qui semble générer un nombre d'erreurs non négligeable.

2 Améliorations proposées

2.1 Modèle Trigramme

Prendre plus en compte le contexte afin de déterminer si l'on doit couper la chaîne à un endroit donné semble être une bonne idée. Il faut pour cela entrainer un modèle trigramme. Cependant, l'utilisation seule de trigrammes augmente certe la précision pour les éléments qui auront pu être trouvés lors de l'apprentissage, mais elle augmente aussi la probabilité de tomber sur une séquence de caractères non observée (car l'observation est plus longue), et donc d'assigner une valeur de césure "par défaut" à celle-ci.

Lors de l'implémentation, il a été choisi de regarder non pas "l'historique" des symboles rencontrés, mais les caractères à venir, c'est-à-dire à droite de la position actuellement regardée.

2.2 Backoff

Afin d'enrichir le modèle trigramme, l'utilisation d'une méthode de Backoff semble judicieuse. Ainsi l'utilisation de trigrammes, puis de bigrammes dans le cas d'une séquence non observée augmente la précision des résultats.

2.3 Les Alphabets

Le japonais utilise plusieurs alphabets :
— les Kanji
— les Hiragana
— les Katakana
— les Romanji

Les probabilités de succession de caractères des différents alphabets ne seront donc pas similaire, car certains enchaînements sont plus fréquents que d'autres, ou tout simplement impossibles. On peut donc construire un modèle bigramme pour la succession des alphabets, et s'en servir lorsque ni les trigrammes, ni les bigrammes sur les caractères, n'ont rencontrés la séquence à traiter. De plus, en distinguant les signes de ponctuations et autres caractères non classifiés (symboles tels qu'une note de musique, un smiley, ...), on assure de meilleurs résultats, car ces derniers segmentent déjà le texte en certains points.

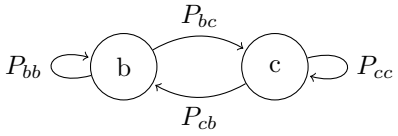
2.4 Séquences non observées

Donner une probabilité plus importante pour la césure d'une séquence dans le cas d'un "non observé" semble intéressant. En effet, l'observation sur quelques phrases, des cas non présents dans les modèles précédents, révèle qu'il faudrait donner une probabilité de césure plus importante par rapport à celle de la continuité. Cela s'averait plus remarquable lorsque le modèle construit avec la succession des alphabets n'était pas utilisé car ce dernier réduit le nombre de cas non observés par les trigrammes et bigrammes.

La probabilité utilisée pour les chaînes non vues était initialement de 0.1 pour la césure et la continuité du mot. La probabilité de couper la chaîne sera augmentée à 0.2 pour la rendre plus importante, sans pour autant différer trop de la probabilité de non césure (qui reste inchangée : 0.1) pour ne pas donner trop d'influence aux choix effectués lors d'une absence d'observation.

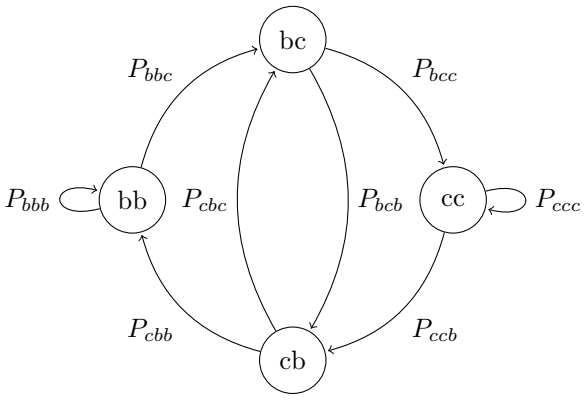
2.5 Transitions entre les états

La méthode de réference utilise des probabilités de passer d'un état à un autre : d'une coupure/non coupure à une coupure/non coupure. Cela respecte le schéma suivant :



(c : non coupure / b : coupure)

Afin d'augmenter la qualité du segmenteur, un état supplémentaire est pris en compte. On garde donc un historique de deux états.



3 Résultats

méthode	précision	rappel	f-mesure
baseline	0.90	0.88	0.89
approche présentée	0.95	0.93	0.94

Références

[1] Constantine P. Papageorgiou. Japanese word segmentation by hidden markov model. *Proceedings of the workshop on Human Language Technology*, HLT '94, 1994.