Article : Text Summarization via Hidden Markov Models

Cet article présente une approche basé sur les chaînes de Markov Cachées (HMM, Hidden Markov Model) pour l'extraction d'extraits (phrases) de documents afin de produire des résumé automatiques.

Contexte et Problèmatique

Un article de 2001 de John M. Conroy et Dianne P.O'Leary qui se penchent sur le problème de la génération de texte automatique.
Le sujet n'est pas nouveau est a été étudié depuis 40 au moment de la parution de l'article, pour rappel il s'agit de choisir quelle phrases extraire d'un texte afin de représenter au mieux l'idée générale de celui dans un espace réduit. Les auteurs proposent ici une approche à l'aide de HMM, l'opposé avait été fait en 2000 par Jing & Keown, ils avaient mis en place une méthode pour passer d'un résumé humain au texte original.

Méthodes : HMM
Dans leur approche ils calculent la probabilité à postériori qu'une phrase soit incluse dans le résumé final. Par opposition au naive Bayes classique ou la probabilité que chacune des phrases soit dans le résumé est indépendante du fait que ses voisines soit dedans, ce n'est pas le cas ici. Pour chaque phrase ils considèrent trois attributs, sa position, le logarithme du nombre de mots de cette phrase plus un et le logarithme de la probabilité de la présence des mots de cette phrase sachant les mots du document.
On les notera respectivement $p,\ o1(i),\ o2(i)$.

La chaîne de Markov proposé possède $2s+1$ états, avec s états dans le résumé, et $s+1$ en dehors.

Ici on a **besoin** de donnée d'entrainement (training data) pour créer la matrice de transition. Elle est formée à partir du maximum de vraisemblance de chacune des transition dans les données d'entrainement.

Chaque état i est associé à une fonction $bi(O) = Pr(O|state\ i)$, O représente le vecteur d'attribut, le maximum de vraisemblance sera utilisé pour déterminer pour estimer la fonction (sa moyenne ainsi que sa matrice de covariance), ils ont estimé les 2s+1 moyenne mais fait l'hypothèse que chacune des fonctions partageait la même matrice de covariance

Pour chaque phrase on calcule la probabilité qu'elle appartienne à chacun des états i, si i est pair la probabilité correspond à la probabilité qu'elle soit dans le résumé, si i est impair, à la probabilité qu'elle n'y soit pas.
On somme ensuite sur les i pairs pour determiner la probabilité totale.

Résultats

L’expérimentation a été ménée sur 1304 documents résumés par la même personne. Ils ont découpé chacun de ces documents en deux parties (train & test). Ils ont ensuite comparé les deux résumés, ceux produits par l’homme, et ceux produits par la machine à l’aide de la métrique suivante :
$F_1 = 100 * \frac{2.r}{k_h + k_m}$
kh la longueur du résumé humain et km la longueur de celui de la machine. Et r le nombre de phrase en commun entre les deux résumés.

La méthode HMM apparait bien meilleur que la méthode DimSum

Pistes & Discussion

Rien de donné dans l’article