Classifications de mots non étiquetés par des méthodes statistiques

Clustering unlabelled words with statistical methods

Christel Beaujard et Michèle Jardino

Édition électronique

URL : <http://journals.openedition.org/msh/2795>

DOI : 10.4000/msh.2795

ISSN : 1950-6821

Éditeur

Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée

Date de publication : 1 septembre 1999

ISSN : 0987-6936

Référence électronique

Christel Beaujard et Michèle Jardino, « Classifications de mots non étiquetés par des méthodes statistiques », *Mathématiques et sciences humaines* [En ligne], 147 | Automne 1999, mis en ligne le 10 février 2006, consulté le 23 juillet 2020. URL : <http://journals.openedition.org/msh/2795> ; DOI : <https://doi.org/10.4000/msh.2795>

Ce document a été généré automatiquement le 23 juillet 2020.

© École des hautes études en sciences sociales

Classifications de mots non étiquetés par des méthodes statistiques

Clustering unlabelled words with statistical methods

Christel Beaujard et Michèle Jardino

RÉSUMÉS

Notre thématique de recherche est le développement de modèles de langage robustes pour la reconnaissance de la parole. Ces modèles doivent prédire un mot connaissant les mots qui le précèdent. Malgré le nombre croissant de données textuelles électroniques, toutes les possibilités de la langue ne sont pas présentes dans ces données, un moyen de les obtenir est de généraliser la représentation textuelle en regroupant les mots dans des classes. Les modèles de langage fondés sur des classes présentent alors une plus large couverture de la langue avec un nombre réduit de paramètres permettant une reconnaissance plus rapide des mots par les systèmes de reconnaissance de la parole dans lesquels ils sont introduits. Nous décrivons deux types de classification automatique de mots, appris statistiquement sur des textes écrits de journaux et de transcriptions de parole. Ces classifications ne nécessitent pas d'étiquetage des mots, elles sont réalisées suivant les contextes locaux dans lesquels les mots sont observés. L'une est basée sur la distance de Kullback-Leibler et répartit tous les mots dans un nombre de classes fixé à l'avance. La seconde regroupe les mots considérés comme similaires dans un nombre de classes non prédéfini. Cette étude a été réalisée sur les données d'apprentissage en français de domaines, de taille et de vocabulaire différents.

Our goal is to develop robust language models for speech recognition. These models have to predict a word knowing its history. Although the increasing size of electronic text data, all the possible word sequences of a language cannot be observed. A way to generate these non encountered word sequences is to map words in classes. The class-based language models have a better coverage of the language with a reduced number of parameters, a situation which is favourable to speed up the speech recognition systems. Two types of automatic word classification are described. They are trained on word statistics estimated on texts derived from newspapers and transcribed speech. These classifications do not require any tagging, words are classified according to the local context in which they occur. The first one is a mapping of the vocabulary words in a fixed number of classes according to a Kullback-Leibler measure. In the second one, similar words are clustered in classes whose number is not fixed in advance. This work has been performed with French training data coming from two domains, both different in size and vocabulary.

INDEX

Keywords : classification, language modeling, mapping, optimization

Thèmes : classification - partition, distances, linguistique, optimisation

Mots-clés : distance, modèle de langages probabilistes, optimisation, partitionnement, statistique