Combinaison de fonctions de préférence par Boosting pour la recherche de passages dans les systèmes de question/réponse

Nicolas Usunier, Massih-Reza Amini, Patrick Gallinari

Laboratoire d'Informatique de Paris 6
8, rue du Capitaine Scott, 75015 Paris
{usunier, amini, gallinari}@poleia.lip6.fr

Résumé. Nous proposons une méthode d'apprentissage automatique pour la sélection de passages susceptibles de contenir la réponse à une question dans les systèmes de Question-Réponse (QR). Les systèmes de RI *ad hoc* ne sont pas adaptés à cette tâche car les passages recherchés ne doivent pas uniquement traiter du même sujet que la question mais en plus contenir sa réponse. Pour traiter ce problème les systèmes actuels ré-ordonnent les passages renvoyés par un moteur de recherche en considérant des critères sous forme d'une somme pondérée de fonctions de scores. Nous proposons d'apprendre automatiquement les poids de cette combinaison , grâce à un algorithme de réordonnencement défini dans le cadre du *Boosting*, qui sont habituellement déterminés manuelement. En plus du cadre d'apprentissage proposé, l'originalité de notre approche réside dans la définition des fonctions allouant des scores de pertinence aux passages. Nous validons notre travail sur la base de questions et de réponses de l'évaluation TREC-11 des systèmes de QR. Les résultats obtenus montrent une amélioration significative des performances en terme de rappel et de précision par rapport à un moteur de recherche standard et à une méthode d'apprentissage issue du cadre de la classification.

1 Introduction

Les systèmes de question/réponse (QR) ont pour objectif de trouver la réponse à une question formulée en langage naturel dans un grand corpus de documents. Nous nous intéressons ici aux systèmes de QR en domaine ouverts, développés dans le cadre des évaluations TREC[1]. Dans ces systèmes, le traitement d'une question s'effectue en trois étapes : **(1)** l'analyse la question, déterminant un type de réponse attendue et la structure syntaxique de la question, **(2)** la recherche d'information (RI), qui interroge un moteur de recherche pour sélectionner des passages susceptibles de contenir la réponse à la question. Selon les systèmes, les passages peuvent être des documents entiers (Monz 2003), des parties de documents de longueur fixe (Chalendar et al. 2002), ou des phrases consécutives d'un document (Prager et al. 2000). Enfin, **(3)** l'extraction et la sélection de la réponse dans les passages séléctionnés.

Dans la chaîne de traitement, le module de RI est crucial, car s'il échoue à renvoyer au moins un passage contenant la réponse dans sa sélection, le système ne peut pas répondre à la question. Par ailleurs, il pose de nouveaux problèmes de RI : la recherche qu'il doit effectuer est plus spécifique

[1] Text REtrieval Conference, http ://trec.nist.gov