Génération et enrichissement automatique de listes de patrons de phrases pour les moteurs de questions-réponses

Co-financé par l'Association Nationale de la Recherche Technologique

Cédric Vidrequin*, Juan-Manuel Torres-Moreno*
Jean-Jacques Schneider**, Marc El-Beze*
* Laboratoire Informatique d'Avignon, Agroparc BP1228, 84911 Avignon CEDEX 9, France
{cedric.vidrequin, marc.elbeze, juan-manuel.torres}@univ-avignon.fr
** Société SEMANTIA
30 avenue du château de Jouques, Parc d'activité de Gémenos, 13420 Gémenos, France
jjschneider@semantia.com

Résumé. Nous utilisons un algorithme d'amorce mutuelle (Riloff et Jones 99), entre des couples de termes d'une relation et des patrons de phrase. À partir de couples d'amorce, le système génère des listes de patrons qui sont ensuite enrichies de façon semi-supervisée, puis utilisées pour trouver de nouveaux couples. Ces couples sont à leur tour réutilisés pour générer, par itérations successives, de nouveaux patrons. L'originalité de l'étude réside dans l'interprétation du rappel, estimé comme la couverture d'un patron sur l'ensemble des exemples auxquels il s'applique.

Summary. We use a mutual bootstrapping algorithm (Riloff & Jones 99), between couples of terms of a relation and pattern phrases. Starting from bootstrap couples, the system generates lists of patterns, which are then enriched in a semi-supervised way and used to find new couples. These couples are used iterativly to find new patterns. The originality of the study lies in the interpretation of recall, estimated as the overlap of the pattern with the set of examples to which it applies.

1 Méthode

Constitution de l'amorce. Actuellement, nous construisons manuellement l'amorce sous la forme d'une dizaine de couples de termes pour lesquels nous sommes sûrs de leur lien à travers la relation qui les unit (Brin 99). Mais cette amorce peut également se trouver dans des mini bases de connaissances ou dans toute table de base de données disponible.

Génération de patrons. Tout d'abord, nous sélectionnons les termes de la base de connaissance qui seront utilisés pour la génération des patrons. Dans le but d'en générer le plus possible de nouveaux, nous utilisons les termes a) générés lors de la dernière itération ou lors des précédentes ; b) de l'amorce : choisis en dernier lieu ou pour la première itération. Nous réalisons ensuite la recherche d'information qui renvoie les données textuelles parmi lesquelles nous recherchons les plus petits segments contenant les deux termes de la relation. Ces patrons de base sont étendus à gauche et à droite, en gardant l'ensemble des patrons intermédiaires. Afin d'en améliorer la couverture, tout en essayant de ne pas diminuer leur précision, nous factorisons si possible les nouveaux patrons avec des patrons déjà existants, si et seulement si ceux-ci ne diffèrent que d'un seul mot.