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.

Évaluation des patrons générés. Nous n'évaluons que les patrons apparus au moins pour deux couples de termes différents. Nous réalisons ensuite une recherche d'information pour chaque couple d'évaluation, en passant en paramètres le patron, ainsi que le premier terme de la relation. Nous récupérons alors tous les segments correspondant à ce patron, et comparons la véritable valeur du second terme du couple avec le mot rencontré dans le patron. Nous comptabilisons ensuite le nombre total de correspondances correctes ainsi que le nombre d'exemples de couples de termes ayant correspondu de façon juste au moins une fois avec ce patron. Ces valeurs sont utilisées pour estimer la précision et le rappel du patron, lesquels sont combinés dans une F-mesure (1) qui correspond au score final du patron.

La précision d'un patron est obtenue en faisant le rapport du *nombre de correspondances correctes* sur le *nombre total de correspondances* de ce patron, tous exemples confondus. Le rappel est obtenu en faisant le rapport entre le *nombre d'exemples ayant donné une correspondance correcte au moins une fois*, et le *nombre total d'exemples*. La F-mesure vaut zéro si la précision ou le rappel du patron sont nuls. Dans le cas contraire, on la calcule en appliquant la formule suivante, avec β=1 mais paramétrable pour les tests futurs :

$$\text{F-mesure} = \frac{(1+\beta^2) * précision * rappel}{(\beta^2 * précision + rappel)} \quad (1)$$

Une fois les poids des patrons calculés, nous trions les patrons en fonction de leur F-mesure. Ainsi, les meilleurs patrons sont utilisés, par la suite, pour l'enrichissement de la base de connaissance, ou le seront pour le moteur de questions-réponses.

Enrichissement semi-supervisé de la liste de patrons. Nous utilisons une liste de quelques couples pour lesquels nous n'avons qu'un des termes de la relation. Le système utilise les meilleurs patrons afin de trouver un (ou des) second(s) terme(s) correspondant à chaque terme de la liste fournie. Avec ces couples potentiels, il tente de générer une nouvelle batterie de patrons, sur le même principe. Les patrons ainsi générés sont évalués avec les couples d'amorce et les meilleurs d'entres eux viennent s'ajouter à la liste déjà présente.

2 Résultats préliminaires

Des résultats préliminaires, obtenus sur 30 couples de test pour chaque relation, donnent une précision allant de 13 à 95%, un rappel compris entre 13 et 77% et une F-mesure moyenne de 0,43. Les meilleurs scores sont obtenus pour la relation **ANNEE_MORT**, qui fait correspondre à une personne l'année de sa mort. Les plus difficiles à traiter comme **COMMENT_MORT** ou **INVENTEUR** ont les plus faibles scores. La difficulté de ces relations résident dans la variabilité de l'expressions des informations comme la façon dont la personne est morte ou le nom de son invention.

3 Références

Brin S. (1999). Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, Valencia, Spain, pages 172-183

Riloff E., Jones R. (1999). Learning dictionaries for information extraction by multilevel bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 474-479, Orlando, Florida, July 1999