Identification de Classes Sémantiques Basée sur des Mesures de Proximité Sémantique

Jean Petit* Jean-Charles Risch*,***

*Université de Reims Champagne Ardenne, CReSTIC, Moulin de la Housse, 51100 Reims
jean.petit, jean-charles.risch@etudiant.univ-reims,
**Capgemini Technology Services, 7 rue Frédéric Clavel, 92287 Suresnes
***Université Technologique de Troyes, 12 rue Marie Curie, Tech-CICO, 10010 Troyes

1 Introduction & Méthode

L'acquisition de relations sémantiques tend à s'automatiser. Cependant, leur validation reste une tâche manuelle sujette aux erreurs. Les mesures de similarité basées sur l'hypothèse distributionnelle harissienne (Harris, 1954) permettent de suggérer l'existence d'une relation sémantique entre deux entités lexicales, mais n'apportent pas d'indication quant à leur éventuelle classe sémantique contrairement à la méthode d'Hearst (Hearst, 1992) basée sur des motifs syntaxiques. Cependant, cette dernière ne permet pas de juger la validité des relations extraites. Oliveira (Costa et al., 2011) croise les deux méthodes dans une approche basée sur le web permettant l'obtention de mesures de proximité sémantique indicatives d'une classe sémantique. Cependant, la méthode par seuil utilisée révèle des faiblesses pour distinguer entre elles des relations sémantiques correctes associées à différentes classes sémantiques.

Notre objectif principal est d'identifier automatiquement la classe de relations sémantiques. Pour ce faire nous étudions une méthode d'apprentissage de classes sémantiques en fonction de mesures de proximité sémantique associées à des motifs syntaxiques. Dans un premier temps nous présentons les grandes lignes de notre méthode, puis nous exposons l'expérience permettant d'évaluer cette dernière. Enfin, nous discutons des résultats obtenus et concluons en ouvrant sur des perspectives d'évolution.

La première étape de la méthode est l'acquisition de mesures de proximité sémantique. Afin d'obtenir ces mesures statistiques, nous avons suivi la méthodologie proposée par (Costa et al., 2011). Nous avons suivi deux pistes d'amélioration des mesures de proximité sémantique avec d'une part la mise en place de contraintes contextuelles (phrase, page web) dans l'expression de cette dernière et d'autre part une analyse de la cohérence entre la relation sémantique cherchée et celles présentes dans les résultats retournés se concrétisant dans un « score de conformité ».

La seconde étape de notre méthode configure un algorithme d'apprentissage supervisé (réseau de neurones) afin de permettre l'identification des classes sémantiques associées à des relations en travaillant sur les mesures de proximité créées. Nous avons fait le choix d'un perceptron monocouche pour sa simplicité d'utilisation, sa popularité et son accessibilité avec notamment le package R nnet.