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.

2 Expérience & Conclusion

Notre ensemble de données est composé de 148 hyperonymes, 200 méronymes, 179 relations causales et un « pot-pourri » de 182 relations sémantiques mixtes (synonymes, antonymes et co-hyponymes). La variation sur le nombre de relations pour chaque classe est due à un filtre qui exclut les relations ayant une cooccurrence pour leurs entités inférieure à 10000. Les précédentes études ont permis la découverte d'un grand nombre de motifs syntaxiques indicatifs de l'hyperonymie (Hearst, 1992), de la méronymie et de la causalité. Les motifs syntaxiques sélectionnés dans notre étude ainsi que la classe sémantique correspondante sont disponibles en annexe. Le réseau de neurones utilisé pour notre expérience est configuré comme suit : 15 neurones dans la couche d'entrée (autant que de motifs syntaxiques à analyser), 8 dans la couche cachée (calculé à partir des formules indicatives préscrites par (Tufféry, 2012)) et 4 dans la couche de sortie (autant que de classes recherchées). Enfin, nous avons choisi la fonction logistique car c'est la fonction de transfert la plus utilisée par les experts.

Pour estimer la qualité de notre méthodologie, nous avons construit huit modèles statistiques différents (en fonction de la présence ou non du score de conformité et du type de contexte) que nous avons pu évaluer via la méthode de validation « holdout ». Les résultats de notre expérience sont synthétisés en annexe. Les meilleurs résultats sont obtenus en ajoutant la page web comme contrainte pour le contexte, en incluant le score de conformité ainsi qu'en utilisant Web Dice avec 72% des relations sémantiques testées correctement identifiées.

Notre étude a confirmé l'efficacité des mesures de proximité sémantique pour l'identification de la classe de relations sémantiques : nous avons correctement identifié 72% de relations sémantiques dans un environnement fortement bruité. De futures améliorations sont possibles, notamment en perfectionnant le calcul du score de conformité. En effet, ce score gagnerait à être enrichi par les nombreux travaux qui portent sur l'acquisition de relations sémantiques.

Références

Costa, H., H. Oliveira, et P. Gomes (2011). Using the web to validate lexico-semantic relations. *EPIA 2011 4*, 597–609.

Harris, Z. (1954). Distributional structure. *Word 10*, 146–162.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*.

Tufféry, S. (2012). *Data Mining et Statistique Décisionelle. Quatrième édition*. Paris : TECHNIP.

Summary

Semantic relations are the core of a growing number of knowledge-intensive systems. The need to validate automatically such relations remains an up-to-date challenge. In this paper, we present a web-based method enabling the automatic identification of the class of a semantic relation. Using measures based on syntactic patterns as entry features for a learning algorithm, we are able to successfully identify 72% of semantic relations divided in 4 classes in a semantically rich environment.