Rubrique préparée par Anne Condamines (ERSS-CNRS)

Thierry Poibeau

Extraction d'information à base de connaissances hybrides.

Thèse d'informatique, Université Paris 13

LIPN et Thales Recherche et Technologie.

Jury : D. Kayser (directeur), A. Nazarenko (Co-directeur), C. Jacquemin, P. St Dizier, Y. Wilks (Rapporteurs). C. Fluhr, C. Sedogbo (examinateurs)

Notre travail se situe dans le domaine de l'extraction d'information. Ce terme désigne l'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle. La mise au point des ressources d'un système d'extraction est une tâche longue et fastidieuse, qui demande le plus souvent une expertise du domaine abordé et des connaissances en linguistique informatique. Ce point est bien connu et les concepteurs de systèmes mentionnent tous des temps prohibitifs passés à développer des ressources.

Comment rendre un système d'extraction plus adaptable ? Cette question est au centre de notre travail de thèse. Afin de bien cerner le problème, nous nous sommes attaché à prendre en compte une grande variété d'applications pour définir de manière précise les besoins opérationnels en matière d'extraction. Nous avons réalisé un système complet d'extraction appelé SEMTEX, qui a pu être appliqué à des tâches et à des domaines variés. Dans ce contexte, nous avons étudié différentes méthodes d'acquisition de connaissances afin de voir quelle stratégie était la mieux adaptée. Nous avons montré que les méthodes endogènes et exogènes pour l'acquisition de connaissances propres à un domaine étaient complémentaires. Il faut donc que le système d'extraction mis au point soit souple, capable d'exploiter des connaissances hybrides, provenant de différentes sources.

La thèse présente un état de l'art des travaux passés en extraction d'information et en l'acquisition de ressources. Nous montrons que la plupart des systèmes d'acquisition de ressources existants sont inadaptés à nos besoins, dans la mesure où ils reposent soit sur de grands corpus, soit sur des corpus annotés. Dans les contextes d'utilisation étudiés, la présence de corpus de taille réduite, variés mais généralement non annotés, nous a poussé à utiliser des connaissances hybrides.

Le module de repérage des entités nommées est décrit en détail. Il repose sur une étude préalable de différentes méthodes et de différents systèmes, fondés soit sur des règles soit sur des techniques d'apprentissage. La solution proposée dans la thèse repose sur une intégration fine de différentes techniques et une mesure permet d'évaluer la gain potentiel dû à l'apprentissage en fonction du corpus à analyser. L'utilisateur peut donc adapter la stratégie d'analyse en fonction de ses attentes et de ses besoins.

Une fois les entités repérées, il importe de les mettre en relation à travers des patrons d'extraction, puis de généraliser ces patrons au moyen de classes sémantiques. L'acquisition automatique de classes sémantique par apprentissage a pu être testé grâce au système ASIUM ; cette approche nécessite un important travail de révision et un domaine d'application stable. L'acquisition à partir d'un réseau sémantique général comme le DICTIONNAIRE INTÉGRAL permet de bien couvrir un domaine donné mais des éléments clés peuvent ne pas être repérés. Les deux approches se complètent naturellement et la stratégie d'acquisition doit être adaptée en fonction de la situation d'utilisation.

Enfin, le mécanisme d'acquisition de patrons d'extraction mis en place adopte une approche hybride, en confrontant le corpus à un réseau sémantique général. Seules des séquences attestées en corpus sont retenues par le système, mais c'est le réseau sémantique (le DICTIONNAIRE INTEGRAL) qui permet de calculer le sens de séquences visées. L'utilisation de tables de contraintes et d'automates patron implantés dans le système de gestion d'automate INTEX permet ensuite de gérer la variation des séquences visées sur le plan morpho-syntaxique. La solution mise au point a été testée en milieu opérationnel, avec un bilan globalement satisfaisant en termes de performances et d'utilisabilité.

Thierry.Poibeau@thalesgroup.com