Apprentissage automatique de relations d'équivalence sémantique à partir du Web

Florence Duclaye

Apprentissage automatique de relations d'équivalence sémantique à partir du Web

Cette thèse s'inscrit dans le contexte d'un système de Questions-Réponses, capable de trouver automatiquement sur le Web la réponse à des questions factuelles traitant de n'importe quel sujet. L'une des manières d'améliorer la qualité des réponses fournies consiste à augmenter le taux de rappel du système. Pour cela, il est nécessaire de pouvoir identifier les réponses sous de multiples formulations possibles. A titre illustratif, la réponse à la question « Quelle est la hauteur de la Tour Eiffel ? » peut non seulement être exprimée de la même manière que dans la question (« la hauteur de la Tour Eiffel est 300 mètres »), mais également sous d'autres formes lexico-syntaxiques (« la Tour Eiffel culmine à 300 mètres », « la Tour Eiffel fait 300 mètres de haut », etc). On parle alors de paraphrases de la réponse. Le recensement manuel de ces paraphrases étant un travail long et coûteux, l'objectif de cette thèse est de concevoir et développer un mécanisme capable d'apprendre de façon automatique et faiblement supervisée les paraphrases possibles d'une réponse. Inscrite dans le vaste domaine de l'acquisition automatique de connaissances sémantiques, la méthode d'apprentissage présentée fait du Web son corpus privilégié, en particulier par la redondance et la variété linguistique des informations qu'il contient. Considéré comme un gigantesque graphe biparti représenté, d'une part, par des formulations (expressions d'une relation sémantique, comme « culmine à » ou « fait ... de haut ») et d'autre part par des couples d'arguments (entités nommées régies par ces formulations, comme « Tour Eiffel - 300 mètres »), le Web s'avère propice à l'application de la citation de Firth, selon laquelle le sens d'un terme (respectivement d'une formulation, dans notre cas) est lié aux termes (respectivement aux arguments) avec lesquels il cooccurre. Ainsi, par un mécanisme itératif, le Web est échantillonné : les formulations (paraphrases potentielles) sont extraites par ancrage des arguments sur le Web et, inversement, de nouveaux arguments sont extraits par ancrages des formulations acquises. Afin de permettre à l'apprentissage de converger, une étape intermédiaire de classification statistique des données échantillonnées est nécessaire. Les résultats obtenus ont fait l'objet d'une évaluation empirique, ce qui permet en particulier de montrer la valeur ajoutée des paraphrases apprises sur le système de Questions-Réponses. De plus, ces résultats mettent en évidence quelques perspectives exploratoires qui permettront d'améliorer le processus d'apprentissage et de l'utiliser dans d'autres contextes applicatifs.

.

2003 E 044

Ecole Nationale Supérieure des Télécommunications

Groupe des Ecoles des Télécommunications - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - Fax + 33 (0) 1 45 89 79 06 - www.enst.fr

ENST 2003 E 044

Florence DUCLAYE

Spécialité : Informatique et Réseaux

Thèse de doctorat

Thèse

présentée pour obtenir le grade de docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Informatique et Réseaux**

Florence DUCLAYE

Apprentissage automatique de relations d'équivalence sémantique à partir du Web

Soutenue le 18 novembre 2003 devant le jury composé de :

Ludovic Lebart	Président
Béatrice Daille	Rapporteurs
Benoît Habert	
Olivier Collin	Examinateurs
Laurent Miclet	
François Yvon	Directeur de Thèse