Enrique Henestroza Anguiano

Mesures d'association lexicale pour le traitement d'attachement des groupes prépositionnels dans l'analyse syntaxique statistique du français.

Si l'analyse syntaxique (parsing) statistique a l'avantage de fournir une technique de désambiguïsation, elle a le désavantage d'être tributaire de la couverture des phénomènes linguistiques présents dans le corpus d'apprentissage du parseur. Dans ce travail, nous visons à améliorer la performance du parsing statistique en atténuant des lacunes de couverture de phénomènes linguistiques avec des données exogènes lexicales extraits automatiquement d'un corpus de grand taille. Plus précisément, on définit un modèle qui remet en question les attachements obtenus par un parseur statistique en dépendances, ce modèle étant spécialement conçu pour permettre l'injection de scores d'association lexicale.

On travaille dans un premier temps sur l'attachement des groupes prépositionnels (GP), qui constituent un tiers des erreurs constatées lors de l'evaluation d'un parseur en dépendances état-de-l'art pour le français. Les GP ont souvent plusieurs sites d'attachement (gouverneurs) dans une phrase qui sont grammaticalement corrects, contrairement à d'autres types d'attachement qui sont moins ambigus structurellement (e.g. les sujets de verbe). Par conséquent, un parseur doit s'appuyer sur des informations lexicales syntactico-sémantiques pour trouver l'attachement qui est sémantiquement acceptable: ces informations comprennent les cadres de sous-catégorisation du gouverneur et les associations lexicales entre le gouverneur, la préposition, et l'objet de la préposition (ainsi que d'autres informations contextuelles que nous n'examinons pas, comme le contexte discursif). Si théoriquement ces informations lexicales sont modélisées implicitement par le processus d'apprentissage d'un parseur, la taille limitée du corpus d'apprentissage ne rend effective cette modélisation que pour les gouverneurs et les GP les plus fréquents.

Pour améliorer le traitement d'attachement des GP, on applique un modèle de post-traitement du parsing qui remet en question le site d'attachement de chaque GP en utilisant le score d'association lexicale entre le GP et chaque gouverneur plausible. Ces scores d'association lexicale sont approximés automatiquement avec des méthodes distributionnelles sur un corpus de grand taille, en utilisant des mesures de collocation classiques comme l'information mutuelle et le rapport de vraisemblance.