Etudes et implantations de variantes d'un algorithme de désambiguïsation lexicale

Problématique, contexte théorique, but du travail

La désambiguïsation lexicale est un traitement qui consiste à trouver le sens particulier d'un mot dans un texte. Il s'agit d'une opération importante pour améliorer les applications du traitement automatique des langues (Traduction Automatique par exemple) ou de la recherche d'information. En TA, elle peut permettre de choisir la meilleure traduction (traduirait-on *devoir* par *homework* ou *duty* ?) et en RI de différencier les documents (pour une recherche de *chat*, différencier l'animal du moyen de communication).

L'un des algorithmes les plus simples et les plus efficaces pour la désambiguïsation lexicale est l'algorithme de Lesk. Dans sa version de base, il s'agit simplement de regarder les mots communs entre les définitions de deux mots et de sélectionner les sens qui ont le score le plus fort.

On peut imaginer de nombreuses variantes de cet algorithme ; celles que nous proposons d'étudier ici concernent essentiellement les extensions monolingues et multilingues des ressources (définitions) utilisées par l'algorithme. Le travail consistera à implanter ces variantes et à les comparer sur des corpus d'évaluation.

Description détaillée du travail à effectuer

Dans un premier temps, il s'agira d'étudier et de comprendre les travaux précédemment réalisés (à partir d'une ressource monolingue) afin d'utiliser une ressource alignée multilingue existante (MultiWordNet ou autre) et évaluer les performances de la désambiguïsation sur l'anglais,

a. si l'algorithme est mis en œuvre à partir d'une autre langue,
b. si l'algorithme considère une combinaison de scores issus de plusieurs langues.

Dans un second temps, il s'agira de construire une ressource lexicale plus importante et d'étudier son apport pour la désambiguïsation. Le travail consistera alors à enrichir la ressource monolingue existante sur l'anglais à partir de définitions anglaises extraites automatiquement du Web (Wiktionnary et Wikipédia). Il s'agira enfin de la compléter en étudiant l'alignement sur cette ressource de définitions d'autres langues extraites également du Web.

L'ensemble sera à réaliser en java ; certains programmes utiles (comme l'extraction automatique des définitions de Wiktionnary) étant déjà implantés dans ce langage. Les algorithmes seront testés sur des corpus d'évaluation déjà existants.

Enseignants Chercheurs encadrants à contacter : Didier SCHWAB et Jérôme GOULIAN

didier.schwab@imag.fr 04.76.63.56.54

jerome.goulian@imag.fr 04.76.51.43.69

Nom du laboratoire de recherche : Laboratoire d'Informatique de Grenoble, Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole