

L'évaluation en Induction de Lexique Bilingue

Contexte

L'Induction de Lexique Bilingue est une tâche qui consiste à produire des lexiques bilingues à l'aide de corpus eux aussi bilingues et parallèles ou comparables. Cette tâche attire de plus en plus de chercheurs, dont beaucoup venant du big data. Cela a pour conséquence que les résultats et la quantité de données traitées se retrouvent privilégiées vis à vis de la qualité de ces données (data scientist vs linguiste). Les données d'évaluations ont donc semble-t-il perdu en qualité et en sens. L'objectif de ce TER sera de récupérer certaines listes les plus mises en avant dans la recherche actuelle, pour pouvoir les évaluer, regarder leur contenu et finalement pouvoir mettre en place un protocole ou des règles qui permettraient d'obtenir des listes plus cohérentes.

Objectifs :

- Etudier des listes d'évaluation d'induction de lexique bilingue en évaluant la qualité de l'induction. Il faudra mettre en place une chaîne d'évaluation d'induction de lexique bilingue.
- Regarder ces listes plus en détails pour en retirer le surplus et l'incohérent, fixer un protocole de lissage.
- Evaluer de nouveau les résultats et analyser les différences.

Nb étudiants : 1/2

Encadrants : Emmanuel Morin & Martin Laville

Lieu : LS2N