Correcteur d'orthographe

Proposition de projet M1 Informatique - Institut Galilée
Conduite et Gestion de Projets

La correction d'orthographe est une tâche importante en traitement de langage naturel (NLP) grâce à de nombreuses applications potentielles, en apprentissage de langues étrangères par exemple. Pour corriger les mots mal orthographié, un dictionnaire est souvent suffisant. Cependant, la détection et la correction des erreurs grammaticales sont plus compliquées. Deux approches différentes sont généralement utilisées : l'ingénierie de la connaissance et l'apprentissage statistique. Dans la première approche, un ensemble de règles conçues par un expert est utilisé ; par exemple, l'adjectif superlatif "meilleur" est précédé de l'article "le". En revanche, l'approche d'apprentissage statistique formule la tâche comme un problème de classification où les règles sont apprises automatiquement à partir d'un grand nombre d'exemple d'erreurs accompagnées de leurs corrections.

Nous proposons dans ce projet de développer un classifieur qui décide pour chaque mot dans une phrase s'il s'agit d'une erreur, et propose donc une correction. Le problème de cette approche est que les des décisions sont indépendantes alors que les erreurs ne le sont souvent pas. Par exemple, dans la phrase "le porte est ouvert", prendre les décisions de corrections d'une façon jointe pourrait améliorer leur cohérence. Pour résoudre ce problème nous imposons des contraintes de cohérence sur les décisions du classifieur et résolvons le problème d'inférence jointe obtenu en utilisant la programmation linéaire ou Integer Linear Programming (ILP).

Les composantes du logiciel

Toute réalisation de l'outil devrait avoir au moins les composantes suivantes :

- Prétraitement de données : ce qui implique le traitement d'un grand nombre de phrases accompagnées de leurs corrections pour obtenir les traits caractéristiques nécessaire pour entrainer le classifieur. Des techniques de traitement du langage naturel seront utilisés pour nettoyer et annoter les données collectées pour une utilisation ultérieure.
- Classifieur : entraîner et tester un classifieur supervisé, de type CRF (champs aléatoire conditionnel) à partir des données prétraitées.
- Développer et tester l'algorithme d'inférence jointe basées sur la programmation linéaire (ILP) dont le rôle est d'améliorer la cohérence entre des décisions du classifieur est incorporer des contraintes globales sur la structure de la phrase corrigée. L'utilisation d'un solveur ILP comme boîte noir est envisageable.
- Développer une interface graphique pour qui servira comme démonstration de l'outil.

Technologies

Aucune restrictions n'est imposée sur les technologies choisies pour la réalisation du logiciel. Cela peut inclure par example Python, PERL, C++, Java ou autre pour le traitement de texte et la mise en œuvre des algorithmes principaux ; et HTML/Javascript (jQuery, D3, etc.) pour l'interface et la visualisation.

Équipe recherchée

Un groupe de 4 à 6 étudiants motivés et bien organisés. Le projet peut être divisé en plusieurs sous tâches réalisées en parallèle.

Encadrant

Nadi Tomeh, LIPN-RCLN
Contact : nadi.tomeh@lipn.univ-paris13.fr