Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Hyun Jung KANG : clinguist.hjkang@gmail.com

Titre : Regards croisés sur les avis en ligne : approches du TAL et de la linguistique de corpus

Mots-clés : fouille d'opinion, avis en ligne, évaluation, apprentissage de surface, apprentissage profond, traitement automatique des langues.

Title: *Crossed Perspectives on Online Reviews: NLP and Corpus Linguistics Approaches*

Keywords: *opinion mining, online reviews, evaluation, machine learning, deep learning, computational linguistics.*

Thèse de doctorat en sciences du langage, MoDyCo, Université Paris Nanterre, sous la direction de Iris Eshkol-Taravella (Pr, Université Paris Nanterre). Thèse soutenue le 29/01/2021.

Jury : Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, directrice), Mme Farah Benamara (MC HDR, Université Paul Sabatier, Toulouse, rapporteuse), M. Dominique Legallois (Pr, Université de Sorbonne-Nouvelle, rapporteur, président), Mme Caroline Brun (chercheuse, Naver Labs Europe, examinatrice), M. Guillaume Desagulier (MC HDR, Université Paris 8, examinateur), M. Jean-Luc Minel (Pr émérite, Université Paris Nanterre, examinateur).

Résumé : *La thèse se situe dans la lignée des recherches en traitement automatique des langues (TAL) sur la fouille d'opinions et propose la modélisation, l'analyse outillée et le traitement automatique des évaluations en ligne des restaurants. Le corpus sur lequel s'appuie l'étude est composé d'évaluations collectées sur un site web dédié aux restaurants. Afin d'étudier l'évaluation des expériences vécues dans les restaurants, un modèle d'évaluation est élaboré sur la base d'une observation manuelle du*

corpus. L'évaluation des restaurants visités ne se limite pas aux opinions positives ou négatives données par les clients, mais peut avoir d'autres fonctions : le visiteur laisse son avis pour donner son opinion (opinion), faire des suggestions (suggestion), exprimer ses intentions (intention) ou décrire son expérience (description). Ce modèle est la base du schéma d'annotation manuelle. Les tendances quantitatives à l'œuvre dans le corpus sont examinées grâce aux méthodes de linguistique de corpus. Chaque fonction est analysée et décrite à travers l'analyse lexicométrique (les mots-clés et les n-grammes), la distribution des parties du discours et sa position au sein d'un avis. L'analyse quantitative permet de décrire les évaluations sous un nouvel angle et de voir comment l'apprentissage automatique parvient à s'adapter à un modèle à quatre entrées de manière à dépasser la simple opposition entre opinion positive et opinion négative. Un module de détection automatique de chacune des fonctions d'évaluation est développé. La démarche choisie s'appuie sur des méthodes couramment employées en TAL telles que l'apprentissage supervisé (l'apprentissage de surface et l'apprentissage profond) et vise à obtenir de bonnes performances aussi bien sur les classes minoritaires que sur les classes majoritaires. Les résultats obtenus pour chaque fonction sont interprétés en tenant compte des spécificités du corpus traité. La généralisabilité du modèle développé a été testée et validée sur d'autres données : un corpus relevant d'un autre domaine (celui de l'hôtellerie) et un corpus écrit dans une autre langue (le coréen). Ces applications permettent en outre de marquer les différences entre les domaines qui n'ont pas les mêmes cibles d'une part et les deux langues (le coréen et le français) du point de vue linguistique et culturel d'autre part.

URL où le mémoire peut être téléchargé :
http://www.theses.fr/2021PA100037