Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Correction automatique de textes d'apprenants et certification linguistique en français : évaluation de la généralisabilité, de l'accord et de la validité

Rodrigo Wilkens[1] Rémi Cardon[2] Vincent Folny[3] Thomas François[4]
(1) University of Exeter, Exeter, Royaume-Uni
(2) Computer Science and Engineering Department, Universidad Carlos III de Madrid, Espagne
(3) France Éducation international, Paris, France
(4) Cental, IL&C, UCLouvain, Louvain-la-Neuve, Belgique
r.wilkens@exeter.ac.uk, rcardon@inf.uc3m.es
Folny@france-education-international.fr, thomas.francois@uclouvain.be

RÉSUMÉ

Dans le domaine de la correction automatique de textes (CAT), les pratiques d'évaluation comparatives ont favorisé des approches minimalistes, en contraste avec les recommandations de protocoles d'évaluation tels que le cadre de validation basé sur l'argumentation (ABV). Celui-ci préconise une évaluation multidimensionnelle des systèmes, notamment dans le contexte des tests linguistiques à forts enjeux. Dans cet article, nous présentons une version améliorée et plus concrète du cadre ABV, intégrant une analyse de l'équité, des corrélations avec des caractéristiques linguistiques, une évaluation des erreurs de prédiction et l'accord entre les modèles et les correcteurs humains. En appliquant ce cadre à la CAT en français, nous comparons 8 architectures de modèles sur un corpus de 27 000 rédactions d'examen (deux correcteurs chacune) et un corpus de généralisation de 961 rédactions (au moins neuf correcteurs chacune). Nos analyses illustrent les avantages de l'application du cadre ABV pour mieux comprendre les capacités et les limites des modèles de CAT, tout en faisant progresser l'état de l'art pour la CAT en français.

ABSTRACT

Automated Essay Scoring and Language Certification : Assessing Generalizability, Agreement and Validity for French

In Automated Essay Scoring (AES), benchmarking practices have fostered minimalist evaluation practices, in contrast with the broader-view recommendations of evaluation frameworks, such as the argument-based validation framework (ABV), which argued in favor of a multidimensional assessment of systems, especially in the context of high-stakes language tests. In this paper, we introduce an enhanced and more practical version of the ABV framework, incorporating fairness analysis, correlations with linguistic features, prediction error evaluation, and model agreement compared with human raters. Applying this framework to French AES, we compare 8 model architectures on a corpus of 27k exam essays (2 raters each) and a generalization corpus of 961 essays (at least nine raters each). Our analyses illustrate the benefits of applying the ABV framework to better understand the capabilities and pitfalls of AES models, while also advancing the state-of-the-art for French AES.

MOTS-CLÉS : correction automatique de textes, certification en langues, cadre de validation basé sur l'argumentation, français, TAL.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 7–8.

KEYWORDS: automated essay scoring, language certification, argument-based validation framework, French, NLP.

ARTICLE ACCEPTÉ À : Transactions of the Association for Computational Linguistics (à paraître).
URL : https://transacl.org/index.php/tacl