Projet MLDS proposé par M. Ballier pour 2019-2020 : **Stratégies de fouille de texte pour la détection automatique d'erreurs dans les corpus d'anglais non-natif**

Contexte scientifique : Depuis une quinzaine d'années, la communauté scientifique des linguistes de corpus d'apprenants a recours aux techniques d'analyse automatique du langage (TAL). Plusieurs compétitions ont été organisées dans la communauté du (TAL), au départ centrées sur un problème linguistique particulier (détection d'erreur sur les articles, puis les prépositions, puis toutes les erreurs possibles). Initialement conçu pour la détection d'erreur automatique des essais des non-natifs (Yannakoudakis et al. 2011), le domaine a essayé d'allier la correction automatique des erreurs et la prédiction automatique des notes (l'assignation d'un score) et s'est tourné ces derniers temps vers l'analyse par réseaux de neurones, afin de faire de prédictions à partir de modèles de langues..

Tâche à réaliser: Réalisation d'un système de *text mining* pour la détection automatique des erreurs sur la base des données de la compétition 2019.

Données : Extrait du corpus NUCLE réalisé à l'université de Singapour proposé dans le cadre de la compétition BEA2019 14th Workshop on Innovative Use of NLP for Building Educational Applications https://sig-edu.org/bea/current#shared-task-on-grammatical-error-correction
Détails de la compétition en https://www.cl.cam.ac.uk/research/nl/bea2019st/

Premières références :

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Doctoral Dissertation). Retrieved from http://scholarworks.gsu.edu/alesl_diss/35.
Meurers, Detmar (2015) Chapitre NLP and LCR in the Cambridge Handbook of Learner Corpus Research.
Diaz, Ballier & Thompson 2013
Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 793-805.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel R. Tetreault. 2013. *The CoNLL-2013 shared task on grammatical error correction*. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pages 1–12.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. *The CoNLL-2014 shared task on grammatical error correction*. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–14.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. *A new dataset and method for automatically grading ESOL texts*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 180–189.

Helen Yannakoudakis, Marek Rei, Øistein E. Andersen, and Zheng Yuan. 2017. Neural sequence-labelling models for grammatical error correction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2795–2806.