Sujet de stage de Master 2 SIF
2018-2019

Prédiction de recommandations d'âge pour l'accès par des enfants à des textes

Encadrants :

- Dr. Gwénolé Lecorvé , IRISA, Lannion, gwenole.lecorve@irisa.fr
- Pr. Delphine Battistelli, MoDyCo, Nanterre, del.battistelli@gmail.com

Site : Lannion (IRISA) ou Nanterre (MoDyCo)

Mots-clés : Traitement automatiques des langues, machine learning, intelligence artificielle.

Objectifs applicatifs

Le sujet traite du langage enfantin et vise à produire une méthode d'estimation de l'âge qu'un enfant doit avoir pour être capable de comprendre un texte. Pour ce faire, ce travail s'appuiera sur différents corpus textuels, chacun reflétant le langage adéquat pour une tranche d'âge donnée, qu'il s'agira d'analyser et confronter. Au sein de cet objectif général, le projet ambitionne, d'une part, d'aboutir à un outil d'estimation via du machine learning et, d'autre part, de fournir des éléments d'explications quant aux notions linguistiques qui portent principalement l'estimation.

Contexte sociétal

La maîtrise de plus en plus précoce par les enfants des outils informatiques et d'Internet pose des questions quant à leur capacité à comprendre certains contenus et crée la nécessité sous-jacente d'en contrôler l'accès. Les enfants ont notamment une maîtrise de la langue moindre en comparaison à celle des adultes (lexique et syntaxe plus limités, appréhension différente des connexions logiques…) et peuvent donc se retrouver en difficulté face à certains textes.

Contexte scientifique

Le sujet se place au cœur du domaine de l'intelligence artificielle et pose un problème de *machine learning* non encore résolu ni même étudié. Le domaine de la linguistique fournit néanmoins des pistes claires pour aborder celui-ci et d'autres problèmes qui sont, eux, déjà connus en informatique partagent des similitudes avec le sujet.

En linguistique, des travaux se sont intéressés au langage adressé par des adultes à des enfants de moins de 3 ans (Saint-Georges et coll., 2013). Sur des tranches d'âge plus élevées, il a été mis en évidence que la maturation cérébrale de l'enfant, en particulier en terme de mémoire et d'attention, a des conséquences sur son langage et la manière dont il peut par exemple raconter une histoire (Gathercole, 1999) ou appréhende la notion de temps (Tartas, 2001 ; Vion et coll., 1999).

En traitement automatique des langues (TAL), peu de travaux s'intéressent cependant au langage enfantin, et encore moins au langage pour les enfants (par opposition au langage produit par les enfants). Les quelques travaux existants s'intéresse à des tâches variées comme l'adéquation de contenus textuels à des enfants (Eickhoff et coll., 2010), sur l'adaptation aux enfants du processus de recherche d'information sur Internet – y compris sur les aspects liés au filtrage ou au réordonnancement de résultats (Gossen et Nürnberger, 2013) – ou encore la simplification de textes (De Beldet et Moens, 2010 ; Barlacchi et Tonelli, 2013 ; Nunes et coll., 2013).

Environnement

L'équipe Expression dispose déjà d'un corpus de texte étiquetés en tranches d'âge et associée à des annotations linguistiques présumées d'intérêt. Un premier travail d'exploration du domaine a également été effectué sur le plan linguistique pour identifier les spécificités du langage enfantin. De même, de premiers résultats d'estimation des âges recommandés ont été produits récemment. Le stage s'appuiera sur ces premiers éléments.

Missions

Les grandes tâches du travail à réaliser sont les suivantes :

- Produire des modèles de prédiction par des techniques d'apprentissage automatique (par exemple, réseaux de neurones, régression logistique, arbres de décision, SVM) ;
- Déduire des modèles les caractéristiques linguistiques prépondérantes dans la construction de la prédiction et les confronter aux règles expertes ;
- Formuler des règles expertes de prédiction ;
- Évaluer et comparer les techniques proposées ;
- Éventuellement, compléter les données fournies par d'autres.

Compétences attendues (acquises ou vues en Master 2) :

- Machine learning ;
- Traitement automatique des langues.

Bibliographie

Barlacchi, G., & Tonelli, S. (2013, March). ERNESTA: A sentence simplification tool for children's stories in Italian. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 476-487). Springer, Berlin, Heidelberg.

https://www.researchgate.net/profile/Gianni_Barlacchi/publication/262250828_ERNESTA_A_Sentence_Simplification_Tool_for_Children%27s_Stories_in_Italian/links/5921c2c4458515e3d4076834/ERNESTA-A-Sentence-Simplification-Tool-for-Childrens-Stories-in-Italian.pdf

De Belder & Moens (2010). Text simplification for children. In Proc. of the SIGIR worksh. on accessible search systems. https://lirias.kuleuven.be/bitstream/123456789/276005/1/beldersigir-as.pdf

Eickhoff, Serdyukov & de Vries (2010). Web page classification on child suitability. In Proc. of the ACM international conference on Information and knowledge management. http://dmirlab.tudelft.nl/sites/default/files/cikm331s-eickhoff.pdf

Gathercole (1999). Cognitive approaches to the development of short-term memory. Trends in Cognitive Sciences. https://faculty.biu.ac.il/~armonls/924/NWR/gathercole%2520(1999).pdf

Gossen & Nürnberger (2013). Specifics of information retrieval for young users: A survey. Information Processing & Management. http://wwwiti.cs.uni-magdeburg.de/iti_dke/Pdf/GossenIPM.pdf

Nunes, B. P., Kawase, R., Siehndel, P., Casanova, M. A., & Dietze, S. (2013). As simple as it gets-a sentence simplifier for different learning levels and contexts. In Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on (pp. 128-132). IEEE. https://www.researchgate.net/profile/Bernardo_Pereira_Nunes/publication/261301359_As_Simple_as_It_Gets_-_A_Sentence_Simplifier_for_Different_Learning_Levels_and_Contexts/links/54df400f0cf2953c22b17a04.pdf

Ma, S., & Sun, X. (2017). A semantic relevance based neural network for text summarization and text simplification. ACL. https://arxiv.org/pdf/1710.02318.pdf

Saint-Georges, Chetouani, Cassel, Apicella, Mahdhaoui, Muratori, Laznik, Cohen (2013). Motherese in Interaction: At the Cross-Road of Emotion and Cognition? (A Systematic Review). PLoS ONE. https://pdfs.semanticscholar.org/a37f/8fc857d4e7c435e6d645bc3a37ddd517c308.pdf

Tartas (2010). « Le développement de notions temporelles par l'enfant », Développements. https://www.cairn.info/article.php?ID_ARTICLE=DEVEL_004_0017

Tomasello (2003), Constructing a language, a usage-based theory of language acquisition. http://journals.lww.com/jonmd/Citation/2005/06000/Constructing_a_Language__A_Usage_Based_Theory_of.12.aspx

Vion, Colas (1999). L'emploi des connecteurs en français : contraintes cognitives et développement des compétences narratives (le cas de la narration de séquences arbitraires d'événements). Prof. of Conference of the International Association for the Study of Child Language. https://hal.archives-ouvertes.fr/hal-00241527/document