Utilisation d'approches automatiques pour la reconnaissance des expériences de lecture.

François Vignale*[1], Guillaume Le Noé Bienvenu*[2], Guillaume Gravier[3], and Pascale Sébillot[4]

[1]Langues, Littératures, Linguistique des universités dÁngers et du Mans – Le Mans Université : EA4335, Université d'Angers : EA4335 – France
[2]irisa – CNRS : UMR6074 – France
[3]IRISA (IRISA) – CNRS : UMR6074 – Rennes, France
[4]Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Université Rennes I, INSA-Rennes – France

Résumé

Cette communication a pour but de présenter le rôle des techniques relevant de l'intelligence artificielle et du traitement du langage naturel dans la mise au point d'algorithmes de détection semi-automatique des expériences de lecture développés dans le cadre du projet READ-IT.
Au cours des dernières décennies, les connaissances sur l'histoire des pratiques de lecture ont considérablement augmenté au sujet des usages et des habitudes mais des questions fondamentales demeurent, telles que le "pourquoi" et le "comment" on lit. Grâce à l'exploration de sources numériques à la recherche de témoignages d'expériences de lecture, le projet READ-IT (Reading Europe Advanced Data Investigation Tool, https://readit-project.eu) vise à mieux comprendre ces phénomènes. Ce projet financé par le Joint Programming Initiative for Cultural Heritage (2018-2021) associe 5 partenaires de 4 pays (France, Royaume-Uni, Pays-Bas, République Tchèque).

En combinant différentes conceptions (Jauss 1982 ; Iser 1978) et en nous inscrivant dans une démarche fondée sur les sources, nous avons obtenu un modèle théorique et une ontologie (Reading Experiences Ontology, REO) proposant un description minimale où l'expérience de lecture est définie comme un phénomène temporel précédé de prémisses et suivis d'effets dans lesquels une personne interagit avec un contenu écrit par l'intermédiaire d'un médium (Antonini *et al.* 2019).

Pour répondre aux questions du " pourquoi " et du " comment " on lit, le projet READ-IT a fait apparaître des besoins importants en intelligence artificielle et plus particulièrement en traitement automatique du langage. Ces besoins entraînent entre autres la récupération en masse de données historiques et contemporaines ainsi que leur pré-annotation dans le but de détecter automatiquement dans les sources les passages contenant des témoignages de lecture ou des mentions d'œuvres d'art.

*Intervenant

Pour parvenir à ses objectifs, le projet READ-IT a mobilisé plusieurs technologies. Parmi elles, la reconnaissance des entités nommées (NER) qui est une tâche classique en traitement automatique des langues consistant à localiser et à associer les entités mentionnées présentes dans un texte dans des catégories prédéfinies telles que les noms de personnes, les organisations, les lieux, les œuvres d'arts... Des approches récentes, utilisées dans le cadre de READ-IT, permettent d'obtenir d'excellents résultats. Celles-ci se basent sur des modèles de langues pré-entraînés comme ELMo (Lample *et al.* 2016) ou BERT (Devlin *et al.* 2019)

La classification de textes a également été utilisée. Il s'agit du processus qui consiste à attribuer une catégorie à un texte en fonction de son contenu. Les approches en traitement automatique des langues pour cette tâche, comme pour les autres, se sont historiquement basées sur des méthodes à base de règles. Pour le projet READ-IT, les approches d'apprentissage automatique (*machine learning*) ainsi que les méthodes d'apprentissage profonds (*deep learning*) qui sont aujourd'hui considérées comme délivrant les meilleures performances ont été testées.

Après une série de campagnes d'annotation menées entre mars et septembre 2020, on peut livrer quelques résultats qui montrent la pertinence des approches retenues et qui permettent d'entrevoir des perspectives prometteuses.

En ce qui concerne la reconnaissance des entités nommées, la détection des mentions d'œuvres d'art fonctionne bien sur le plan qualitatif en utilisant les modèles BERT (Bidirectional Encoder Representations from Transformers) et plus particulièrement ontonotes pour l'anglais et Multilingual Cased pour les autres langues. Ces modèles de langues très complets permettent l'identification de mentions d'œuvre d'art dans une centaine de langages, en plus d'identifier 18 autres types d'entités (PERSON, NORP, FACILITY, ORGANIZATION, GPE, LOCATION, PRODUCT, EVENT, WORK OF ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL).

La mise au point de classifieurs pour la détection des témoignages de lectures dans les textes est une tache plus complexe car elle requiert des données annotées en très grand nombre. Des commentaires de livres issus de réseaux sociaux de lecture francophone et anglophone Babelio et Goodreads ont été annotés lors de plusieurs campagnes qui se sont déroulées de mars à septembre 2020. Elles ont permis de mettre en lumière un faible taux d'accord inter-annotateurs qui sert à mesurer la cohérence des annotations produites (kappa de Fleiss inférieur à 0,3). Ce constat est dû à la fois à la part considérable d'interprétation personnelle et la complexité des tâches qui consistaient à la fois en le balisage du début et de la fin des expériences de lecture et en l'identification des composantes des expériences, en respectant le modèle de données de READ-IT. Des approches d'apprentissage automatiques classiques ont été testées (td*idf/SVM, FastText) sur ces sources contemporaines, lesquelles ont donné des résultats très intéressants sur ce type de données avec un degré de précision très élevé mais un rappel (nombre de sources pertinentes) encore faible.

Cependant, la réutilisation de ces modèles sur des textes plus anciens annotés manuellement par le passé (correspondance de Joseph Conrad, Memories and Portraits de R. L. Stevenson) ont montré quelques faiblesses dans cette approche. En effet, les formes familières de commentaires issus d'un corpus web qu'un classifieur est en mesure d'apprendre sont très éloignées de celles que l'on retrouve dans des textes littéraires plus anciens et réciproquement, sans oublier le fait que l'accord inter-annotateur est probablement encore plus faible pour les sources historiques en raison d'une part encore plus importante laissée à l'interprétation personnelle.

Corpus

Nombre de sources

Nombre de sources contenant une expérience de lecture

Ratio

Babelio

87664

2713

3,1 %

Goodreads

10000

608

6,1 %

Conrad's Letters

273

2

0,7 %

Stevenson's Memories

326

2

0,6 %

Par conséquent, afin de détecter les témoignages de lecture présents dans tous les types de textes (article/commentaire web, correspondances, essai littéraire, roman, ...), nous pensons délivrer un modèle basé sur des règles (présence d'action de lecture, identification du lecteur, présence de médium ...), moins précis que des algorithmes entraînés sur des données spécifiques , mais plus robuste sur des données fortement hétérogènes. De plus, ces modèles permettent d'expliquer les caractéristiques qui ont permis la détection, lesquelles pourront être intégré dans une interface, et ne nécessiteront qu'une faible maintenance sur le long terme, facilitant ainsi la durabilité du projet et sa transposition à d'autres phénomènes expérientiels.

Références

Jauss, Hans Robert. 1982. *Toward an Aesthetic of Reception.* Minneapolis : University of Minnesota Press.

Iser, Wolfgang. 1978. *The act of reading: a theory of aesthetic response.* London : Routledge.

Antonini, Alessio, Vignale, François, Gravier, Guillaume et Ouvry-Vial, Brigitte. 2019. " The Model of Reading: Modelling principles, Definitions, Schema, Alignments ". https://hal-univ-lemans.archives-ouvertes.fr/hal-02301611.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami et Chris Dyer. 2016. " Neural Architectures for Named Entity Recognition ". Dans *Proceedings of*

the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 260-270. San Diego, California : Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1030.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee et Kristina Toutanova. 2019. " BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ". *arXiv:1810.04805 [cs]*, mai. http://arxiv.org/abs/1810.04805.