Ariane: dispositif de fouille et de lecture synthétique de textes

Motasem Alrahabi

Ariane: dispositif de fouille et de lecture synthétique de textes

Motasem Alrahabi

OBVIL, Maison de la recherche, Sorbonne Université, 75005 Paris
motasem.alrahabi@gmail.com

Résumé. La croissance exponentielle des documents textuels en format numérique et le développement des moyens informatiques de plus en plus sophistiqués offrent de nouveaux moyens pour l'analyse sémantique et discursive des textes. Dans cette communication, nous présentons Ariane, une interface web basée sur des annotations sémantiques et permettant de fouiller des corpus textuels selon des modalités linguistiques: opinions, sentiments, émotions, perceptions, etc. L'annotation automatique est assurée par un outil à base de règles, et les ressources linguistiques sont manuellement crées et vérifiées. Un scénario d'utilisation concret est présenté pour montrer l'intérêt de l'application dans le domaine des humanités numériques.

1. Présentation

L'architecture générale d'Ariane[1] permet d'interroger des corpus textuels en combinant un modèle d'annotation sémantique et un modèle de lecture synthétique de textes. Les annotations sémantiques sont obtenues à l'aide d'un système d'annotation de patrons de surface. Le modèle de lecture synthétique permet la mise en œuvre de parcours de lecture allant d'une vision globale des modalités (au niveau du corpus entier) à une vision locale exprimée au niveau du texte, voire de la phrase[2].

L'utilisateur a le moyen de croiser ces informations avec une technique classique de recherche d'information: requêtes par mots clés et filtrage par métadonnées (auteur, date, titre…). À n'importe quel moment d'un parcours de fouille, l'utilisateur a le moyen d'afficher les statistiques en cours, de les visualiser par des diagrammes ou de les exporter sous forme de tableurs csv.

L'annotation automatique des textes indexés dans Ariane est assurée par Textolab, un outil à base de règles qui permet de manipuler les patrons linguistiques de surface et de créer les règles heuristiques.

[1] https://obvil.huma-num.fr/ariane/

[2] Ariane s'inspire directement de l'application e-quotes (Alrahabi, 2015).

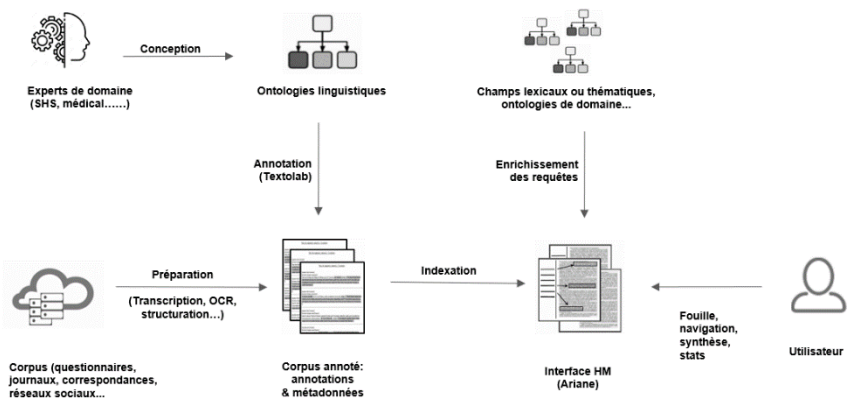

FIG. 1 – *Architecture générale du système*

Le processus d'étiquetage dans Textolab est incrémental et permet de localiser, grâce à des heuristiques, des patrons sous forme de chaînes de caractères, d'expressions régulières ou de métadonnées (annotations existantes). Les règles sont déclaratives et ordonnables selon la priorité. Chaque règle procède à la désambiguïsation des marqueurs identifiés, par des indices contextuels confirmatifs qui valident l'annotation, ou par des indices infirmatifs qui annulent l'annotation en cours. Lorsque toutes les conditions sont satisfaites, au moins une des étiquettes suivantes est attribuée au passage textuel concerné: modalité, polarité, intensité. Cette procédure prend en compte le traitement de la négation, de l'interrogation et du mode hypothétique.

2. Scénario d'utilisation

En traitement automatique du langage et fouille de textes, les travaux relatifs aux modalités recoupent souvent avec l'analyse des opinions, des sentiments et des émotions (Liu, 2015).

Différents parcours de fouille ont été expérimentés avec Ariane dans le domaine des humanités numériques et médicales (Riguet et Alrahabi, 2020 ; Alrahabi et al., 2020 ; Alrahabi et Bordry, 2020). Nous nous proposons ici d'explorer l'interface d'Ariane à travers une étude de cas qui porte sur le jugement critique des écrivaines au XIXe siècle (Riguet et Alrahabi, 2020). Notre hypothèse était de savoir si les jugements des critiques du XIX^{e} siècle étaient formulés de la même façon pour les écrivaines que pour les écrivains. Quels registres sont convoqués, et surtout quels types de valeurs littéraires (esthétique, éthique, intellectuelle, psychoaffective…) leur sont particulièrement attachés ? L'étude du discours critique de la seconde moitié du XIX^{e} siècle, formulé par des auteurs exclusivement masculins, parle en majorité d'hommes écrivains. Cependant, et dans une perspective d'étude du genre, il nous a semblé intéressant de nous pencher sur le traitement particulier de certaines femmes qui font exception dans cette disparité proportionnelle, comme George Sand, Germaine de Staël ou Madame de Lafayette.

À partir de la bibliothèque numérique du labex OBVIL[3], nous avons isolé 134 ouvrages de critique littéraire publiés entre 1850 et 1914. Le corpus, composé de 487 fichiers (environ 33000 phrases), a été annoté par Textolab à l'aide d'une ontologie des modalités linguistiques (Alrahabi, 2010 ; Alrahabi et Riguet, 2017) qui comprend des classes comme l'opinion, le jugement, le désaccord, l'appréciation, la louange, l'indignation, etc.

À partir des résultats obtenus[4], deux constats ont été observés. D'une part, les proportions entre jugements positifs et jugements négatifs semblent très proches, qu'ils portent sur les hommes ou les femmes: avec 71% (hommes) et 65% (femmes) de catégories globalement connotées positives ; et 29% (hommes) et 34% (femmes) de catégories globalement connotées négatives.

D'autre part, si les grands types de jugement se répartissent de façon similaire, c'est davantage la nature des *valeurs* associées qui se distinguent. En effet, la lecture à l'échelle de la phrase rendue possible par Ariane montre que, contrairement à la variété des valeurs d'appréciation (esthétique, intellectuelle, éthique, référentielle, etc.) repérées dans le corpus des hommes, les appréciations du corpus des femmes mettent nettement en avant la dominance de l'émotionnel, du psycho-affectif et de l'esthétique sur la dimension intellectuelle.

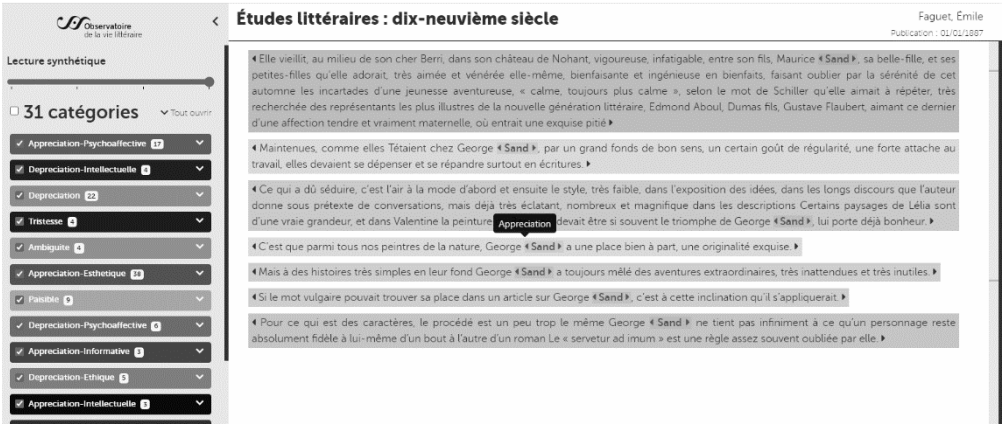

FIG. 2 – *Capture d'écran d'un texte annoté sur Ariane*

À l'aide d'Ariane, il est possible d'interroger les résultats par sous-corpus (métadonnées), par polarité ou par modalité. L'utilisateur a le moyen d'effectuer des requêtes par mots clés au sein des passages annotés, de visualiser les résultats document par document, ou par concordance, avec la possibilité d'accéder au contexte d'origine de chaque mots clé. Des fonctionnalités de lecture synthétique permettent de réduire le texte autour des passages annotés.

Cette approche sémantique nous permet de rendre apparents, dans le discours, la part que les écrivaines occupent et les types de traitements dont elles font l'objet ; elle nous permet de comparer les modalités et les valeurs potentiellement distinctes par le biais desquels les critiques évaluent les œuvres selon le genre de l'auteur.

[3] https://obvil.sorbonne-universite.fr/corpus/critique/
[4] https://obvil.huma-num.fr/ariane/etudeGenre/search

3. Conclusion

En combinant les annotations sémantiques et la lecture synthétique à des échelles différentes, du corpus au texte, l'objectif final d'Ariane est d'être à la fois un outil d'exploration et un support à l'interprétation des textes dans le domaine des humanités. Cette application permet de mettre en lumière des agencements linguistiques et participe à l'analyse du discours en aidant l'utilisateur à formuler des hypothèses, susceptibles par la suite d'être précisées ou infirmées au regard des résultats d'annotation.

Notre prochain objectif est de permettre aux utilisateurs de charger leurs propres textes dans le système, de les annoter et de les exploiter via l'interface.

Références

Alrahabi, Motasem. 2010. « Excom-2 : plateforme d'annotation automatique de catégories sémantiques : conception, modélisation et réalisation informatique : applications à la catégorisation des citations en arabe et en français ». Thèse de doctorat, Paris 4. http://www.theses.fr/2010PA040005.

—. 2015. « E-Quotes: Enunciative Modalities Analysis Tool for Direct Reported Speech in Arabic ». Dans *Computational Linguistics and Intelligent Text Processing*, édité par Alexander Gelbukh, 479-490. Lecture Notes in Computer Science. Cham : Springer International Publishing. https://doi.org/10.1007/978-3-319-18111-0_36.

Alrahabi, Motasem et Marguerite Bordry. 2020. « L'ironie dans la critique littéraire : quelques pistes pour un traitement automatique ». Dans *Humanités numériques et Digital Studies*. Montpellier.

Alrahabi, Motasem, Pauline Flepp et Camille Koskas. 2020. « Polémiques dans le rituel épistolaire : les cas de la correspondance Ponge et Paulhan ». *Revue Épistolaire 46, Honoré Champion.*

Liu, Bing. 2015. *Sentiment Analysis: mining sentiments, opinions, and emotions.* Cambridge University Press, 2015. doi:10.1017/CBO9781139084789.

Riguet, Marine et Motasem Alrahabi. 2017. « Pour une analyse automatique du Jugement Critique: les citations modalisées dans le discours littéraire du XIXe siècle ». Dans *DHQ: Digital Humanities Quarterly*.

—. 2020. « Analyse automatique pour une étude du genre : quels jugements des écrivaines au XIXe siècle ? » Dans *Digital Humanities Conference*. Ottawa, Canada.

Summary

The exponential growth in the digitization of texts and the development of increasingly sophisticated computing tools offer new opportunities for the semantic and discursive analysis of texts. In this talk, we present Ariane, a web interface based on semantic annotations and allowing to mine textual corpora according to linguistic modalities (opinions, sentiments, emotions, perceptions…). Automatic annotation is provided by a rule-based tool, and linguistic resources are manually created and verified. A concrete use case is presented to show the potential of the application in the field of Digital Humanities.