Détection et production de défigements linguistiques dans les réseaux sociaux assistées par les sciences participatives : fertilisation croisée entre traitement informatique et analyse linguistique

1 Contexte

L'entrée des outils informatiques dans les sciences du texte et la linguistique a produit des recherches fécondes dans des perspectives variées, élargies et enrichies ensuite par l'introduction des approches par apprentissage *(machine learning)*.

Si ces apports ont souvent produit des connaissances nouvelles, ils ont parfois contribué à souligner la frontière entre les deux disciplines, voire à creuser entre elles une séparation, qui se traduit dans la polarisation des positions épistémologiques : certains chercheurs perçoivent l'étude informatisée de la langue comme un prolongement de méthodes quantitatives soupçonnées de réductionnisme, d'autres, au contraire, reprochent aux approches qualitatives non instrumentales le caractère vague des inductions produites à partir de données insuffisantes et hétérogènes. Bien entendu, entre ces deux pôles, de nombreux chercheurs et chercheuses jouent le jeu de l'interdisciplinarité et consentent à adapter leurs objets de recherche au traitement par les outils et instruments informatiques.

C'est dans ce contexte que se place le sujet de thèse proposé ici. Il vise à interroger sur un plan méthodologique et épistémologique la collaboration entre sciences du langage et informatique à partir d'un cas d'étude précis, celui de la production et de la reconnaissance d'**expressions défigées** dans les **réseaux sociaux**. Les faits linguistiques traités sont des expressions telles que « Je trahis la rue et je vous trouve du travail », défigement d'une expression qui a circulé dans l'espace médiatique et que l'on peut assimiler à un mème langagier (Gautier & Siouffi 2016). Ces défigements peuvent avoir été recueillis automatiquement dans les réseaux sociaux, ou générés par apprentissage à partir d'un corpus d'un million de tweets constitué dans le cadre du projet Emergence MÈMES. Le projet doctoral pourra s'inscrire dans la continuité de ce travail.

2 Objectif scientifique

Le projet aura pour but d'interroger le **statut des données empiriques** extraites de réseaux sociaux (notamment Twitter) et utilisées pour le traitement (données brutes ou annotées). Il devra également déterminer les **conditions de compatibilité** de l'analyse linguistique et du traitement informatisé (ontologie et cadre théorique à employer, par exemple).

Dans l'examen du cas d'étude, le travail aura également pour tâche de qualifier la nature des **échanges entre informatique et sciences du langage** : peut-on imaginer que par un phénomène de cercle vertueux le traitement automatisé puisse affiner la description linguistique en même temps que celle-ci permette d'ajuster les outils informatiques ?

En effet, dans une approche en *machine learning*, la collaboration entre les deux disciplines se réduit parfois à une fourniture de données (notamment de données annotées) par les uns et de méthodes (notamment de classification) par les autres. Cela revient à valider une séparation étanche entre la définition de la tâche, la préparation des données puis l'application des méthodes. Or, notre proposition est au contraire d'explorer les possibilités offertes par des techniques telles que l'*active learning*, technique dans laquelle l'expert est impliqué dans le processus d'apprentissage afin d'affiner la catégorisation en la confrontant au réel. Ce procédé amène un renversement de certains paradigmes de l'apprentissage traditionnel, puisqu'il est alors admissible que les annotations puissent être contradictoires et que l'évaluation des sorties acceptées ou non par l'expert puisse évoluer dans le temps.

Un autre pan du projet consistera à mettre en place une interface de sciences participatives pour faire annoter des locuteurs (myriadisation ou crowdsourcing). La myriadisation a en effet déjà montré son efficacité pour faire annoter des unités polylexicales par le biais de la plateforme ludifiée *RigorMortis* (Fort *et al.* 2018 et 2020). L'interface prévue permettra d'une part de valider les extractions automatiques et d'autre part de les annoter (parties fixes, parties mobiles), ainsi que d'ajouter de nouveaux candidats. Elle pourra facilement être réalisée par le biais de la plateforme *Language Arc* du Linguistic Data Consortium (LDC) et ajoutée sur le portail *Science Ensemble*, de Sorbonne Université, afin de profiter de la publicité faite à ces plateformes.

3 Approches croisées

L'appréhension des défigements est d'abord d'ordre cognitif et linguistique chez l'humain. Il convient donc de proposer une approche qui rallie l'intuition humaine et les formalismes interprétables par la machine. En l'occurrence, les défigements et les mèmes ont la singularité d'être constitués de parties invariants et variantes. Les points de vue relatifs aux mèmes peuvent donc varier : si l'on se place du côté des invariants, on serait tenté de formaliser les variantes ; à l'inverse, si l'on se place du côté des variantes, on devrait définir formellement les invariants.

De ce fait, des études préliminaires croisées sont indispensables pour cerner notre objet d'étude. En effet, des travaux réalisés en cognition et en linguistique peuvent apporter des éclairages sur la conceptualisation et la modélisation. Nous citons à titre d'exemples la théorie de la Gestalt (Palmer *et al.* 1999), la théorie de l'émergence qui place la globalité d'un tout au-dessus de sa décomposabilité (Col 2004), l'étude sur l'appréhension du sens (Le Ny, 2005), ou les corrélats neuronaux de la reconnaissance des expressions figées et défigées (Blache *et al.* 2018).

L'un des points centraux du projet est donc d'optimiser l'interaction entre d'une part la description et la modélisation théorique des observables, et d'autre part les raisonnements algorithmiques qui se focalisent sur la matérialité des signes, l'optimisation des tâches et la performance des modèles construits.

4 Difficultés techniques et méthodologiques

Nous identifions un certain nombre de verrous que ce soit du point de vue de l'identification des défigements, de leur classification ou encore de la génération.

Concernant l'identification des défigements, un premier verrou est la question du silence. S'agissant de données issues de réseaux sociaux, par définition foisonnantes, il est impossible de garantir l'exhaustivité. Néanmoins, il serait possible de suivre un échantillon de défigements préalablement identifiés comme intéressants. Si ces défigements sont repris, cela permettrait d'identifier certains des procédés à l'œuvre dans les retweets. S' ils ne le sont pas ou peu, cela permettra de mesurer la capacité du système de veille à repérer les phénomènes rares. Analyser des flux pose aussi la question du repérage de ce qui est étonnant, en liaison avec la faculté d'"oubli", corollaire de la notion de nouveauté (Zhang *et al.* 2015).

Concernant la classification, le verrou se situe du côté des critères de regroupement. En effet, en particulier dans un paradigme de classification non supervisée, l'humain qui observe les données classifiées a tendance à interpréter le résultat de la classification : pourquoi tel ou tel regroupement est effectué, quelle est la différence mesurée entre une instance et une autre ? Lorsque le jeu de caractéristiques exploité est homogène (par exemple des caractéristiques uniquement syntaxiques) l'interprétation est relativement facile, par contre il n'en est pas de même lorsque des indices de différents ordres sont exploités (lexique + syntaxe + phonétique par exemple) que ce soit dans le cadre d'une concaténation explicite des représentations (cadre classique) ou dans le cadre d'une classification multi-vues. Ceci rejoint des questionnements génériques de la communauté IA (Rudin 2019) ou plus spécifiquement de la communauté TAL (Bender 2020) sur le compromis entre efficacité et explicabilité.

Ensuite, pour ce qui est de la génération, le verrou principal pourrait se situer du côté de l'évaluation. En effet, le paradigme actuel du Traitement Automatique des Langues est fortement orienté du côté des benchmarks (évaluation supervisée) de sorte que valoriser ce travail du point de vue de la recherche en TAL impose de réfléchir à une évaluation qualitative, l'évaluation quantitative étant probablement d'intérêt limité. L' annotation a posteriori par myriadisation sera sans doute prometteuse à ce titre.

Enfin, les défigements et les mèmes étant dans notre esprit intimement liés, on pourra se poser la question d'exploiter les apports respectifs de l'image et du texte à la diffusion du mème : quels mèmes sont naturellement hybrides, plutôt fondés sur l'image ou sur le texte…

5 Profil recherché et encadrement

Compte tenu du caractère transdisciplinaire du projet, le profil recherché est assez large mais doit comporter une expérience en linguistique générale et/ou computationnelle. Une compétence en développement constitue un atout.

L'encadrement se répartira selon les champs disciplinaires entre le directeur, la co-directrice et le co-directeur.

Références citées

— Bender, E., Koller, A. Climbing towards NLU : On Meaning, Form, and Understanding in the Age of Data, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020*, en ligne.

— Blache, P., Rauzy, S., Bolger, D., Pattamadilok, C., Dufour, S. A Dataset for Studying Idiom Processing with EEG. *Linguistic and Neuro-Cognitive Resources (LiNCR), LREC 2018 Workshop, May 2018*, Miyazaki, Japan. pp.18-22.

— Col, G. Théories cognitives et l'hypothèse de l'émergence du sens. *Tropismes, CREA - Centre de Recherches Anglophones*, 2004, 12, pp.115-140.

— Ferrara, E., Jafari Asbagh, M., Varol, O., Qazvinian, V., Menczer, F. Flammini, A. Clustering Memes in Social Media, *Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 548-555.

— Fort, K., Guillaume, B., Pilatte, Y.-A., Constant, M., Lefèbvre, N. Rigor Mortis : Annotating MWEs with a Gamified Platform. *LREC 2020 - Language Resources and Evaluation Conference, May 2020, Marseille, France*.

— Fort, K., Guillaume, B., Pilatte, Y.-A., Constant, M., Lefèbvre, N. "Fingers in the Nose" : Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly GamifiedCrowdsourcing Platform. *Proceedings of the Joint Workshop on Linguistic Annotation, MultiwordExpressions and Constructions (LAW-MWE-CxG-2018), Aug 2018*, Santa Fe, United States. pp.207- 213.

— Gautier, A., Siouffi, G. Les mèmes langagiers : propagation, figement, déformation. *Travaux de linguistique*, 2016/2.

— Le Ny, J.-F. *Comment l'esprit produit du sens ?*, Odile Jacob, 2005.

— Palmer S. E., Fayard A.-L. Les théories contemporaines de la perception de Gestalt. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, n°28, 1999/1. Présences de la Gestalt, sous la direction de R. Casati. pp. 53-91.

— Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5) :206–215, 2019.

— Zhang, Y., Jatowt, A., Bhowmick, S., Tanaka ; K. Omnia Mutantur, Nihil Interit : Connecting Past with Present by Finding Corresponding Terms across Time. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.