Apprendre et mesurer la conflictualité avec le *deep learning* ?

Céline Poudat

Université Côte d'Azur, CNRS, BCL – celine.poudat@univ-cotedazur.fr

Abstract

This paper focuses on the linguistic expression of conflicts in Wikipedia. I am currently working on a pragmatic annotation task. I concentrate on the markers of disagreement and conflict (Poudat 2018, Poudat and Ho-Dac 2019) keeping in mind an objective of operative and partially automated description. In this context, I am naturally interested in the possibility of detecting disagreement and conflict in discussions. Deep learning and in particular the convolutional model (CNN) that my research team is currently implementing in Hyperbase Web (Vanni et al. 2018a&b) is very attractive for a number of reasons: on the one hand, it will allow me to assess conflict detection in discussion sequences and threads ; and on the other hand, to enrich the description and annotation of conflicts with new patterns and regularities using the TDS index.

Keywords: Conflits, Wikipedia as Corpus, Pragmatics, Interactions, Deep learning.

Résumé

Cette communication s'intéresse à l'expression linguistique du conflit dans Wikipédia. Sur le plan linguistique, je mène une réflexion sur les marqueurs du désaccord et du conflit (Poudat 2018, Poudat et Ho-Dac 2019) dans un objectif de description opératoire et partiellement automatisable. Dans ce cadre, je m'intéresse naturellement à la question de la détection des désaccords et des conflits dans les discussions. Le *deep learning* et en particulier le modèle convolutionnel (CNN) qu'implémente à l'heure actuelle mon équipe de recherche dans le logiciel Hyperbase (Vanni et al. 2018a et b) me semble intéressant à plus d'un titre : il me permettra d'abord d'évaluer la détection des conflits dans les séquences et les fils de discussion, annotés avec une variable catégorielle à trois modalités (désaccord / conflit / non conflictualité), et d'enrichir la description et l'annotation des conflits avec des motifs et des régularités potentiellement inédites au moyen de l'indice TDS.

Mots clés : conflits, Wikipédia comme corpus, pragmatique, interactions, *deep learning*

1. Introduction

Le présent article s'inscrit dans un cadre de recherche plus général qui s'intéresse à l'exploration des interactions conflictuelles dans les pages de discussion éditoriales autour des articles de l'encyclopédie collaborative Wikipédia. Je m'attache ainsi au traitement, à la structuration et à l'analyse des interactions conflictuelles de Wikipédia. Si mes premières recherches, qui s'intéressaient aux conflits dans l'encyclopédie du point de vue des sciences sociales (*e.g.* Auray, Hurault-Plantet, Poudat et Jacquemin 2009), m'ont permis de caractériser le dispositif et la situation communicative spécifique de Wikipédia en général et des pages de discussion en particulier, le développement du corpus *Wikiconflits* m'a permis d'engager un travail d'exploration linguistique des interactions conflictuelles.

Dans cette perspective, j'ai poursuivi deux pistes descriptives ces dernières années, au niveau des fils de discussion d'une part, et des marqueurs du désaccord et du conflit d'autre part : (i) une catégorisation des fils conflictuels au moyen d'une variable binaire (conflit *vs* harmonie), puis ternaire (conflit *vs* désaccord *vs* harmonie) a d'abord été réalisée (Poudat et Ho-Dac 2019) en suivant l'exemple de Denis *et al.* (2012) ; (ii) une annotation pragmatique du corpus a ensuite été mise au point afin de saisir le système linguistique du désaccord (Poudat 2018). Le désaccord est en effet caractéristique des pages de discussion Wikipédia et il est en lien direct avec le conflit. Tout désaccord ne mène pas au conflit, mais la plupart des conflits naissent d'un désaccord. Il s'agissait ainsi de développer un système de marqueurs qui rende compte du système de choix dont dispose le locuteur pour exprimer son désaccord tout en permettant une caractérisation et une comparaison des discussions sous ce prisme.

À l'heure actuelle, je dispose donc d'un corpus de travail annoté ; outre les fils de discussion, catégorisés selon les trois valeurs *conflit vs désaccord vs harmonie*, un travail d'annotation systématique des désaccords a été réalisé sur les messages, qui a été complété par une entreprise d'annotation des attaques. Les catégories sur lesquelles je travaille relèvent du niveau pragmatique : ce sont des annotations de haut niveau et le niveau pragmatique est bien connu pour résister aux traitements automatiques.

Il m'a donc semblé tout à fait intéressant de mener une entreprise d'apprentissage automatique sur mon corpus avec les méthodes actuelles, qui se sont largement développées ces dernières années, obtenant des taux de précision impressionnants et parvenant à classer des masses et des types de données selon des critères qui dépassent l'entendement humain. Parmi les techniques disponibles, le modèle convolutionnel (CNN) de *deep learning* qu'implémente à l'heure actuelle mon équipe de recherche dans le logiciel Hyperbase (Vanni et al. 2018), semble particulièrement concluant. Le présent article s'attachera donc à évaluer son intérêt en matière d'apprentissage et éventuellement de prédiction de passages conflictuels. Peut-on apprendre et mesurer le conflit ? Dans quelle mesure est-ce que le *deep learning* peut assister et éclairer ce genre de recherche ? Telles sont les questions que nous développerons dans ce texte.

La section 2 qui suit présentera le corpus de travail d'une part, et les annotations d'autre part tandis que la section 3 s'attachera à reporter les expériences menées en discutant les résultats obtenus.

2. Corpus de travail et annotations

2.1 Wikiconflits augmenté

Développé dans le cadre du projet CoMéRé (Chanier et al., 2014), le corpus *Wikiconflits* a été développé en collaboration avec Natalia Grabar et Camille Paloque (Poudat et al., 2017). Il regroupe l'ensemble des discussions éditoriales générées autour de sept articles collaboratifs relevant du champ des sciences et techniques ayant donné lieu à des conflits d'édition, soit 4 456 messages produits par 3 971 contributeurs distincts.

Ces dernières années, je me suis attachée à préciser les lieux et les marqueurs du conflit en réalisant en collaboration avec Mai Ho-Dac une tâche de catégorisation systématique des fils de discussion du corpus permettant de distinguer les fils conflictuels des fils non conflictuels (Poudat et Ho-Dac 2019). Après différentes expérimentations, j'avais finalement adopté une variable catégorielle à trois valeurs (**C2** signalant la présence d'un conflit sur le fil, **C1** celle d'un désaccord et **C0** une situation communicative qu'on pourrait dire harmonieuse) qui a été appliquée à l'ensemble des fils de discussion du corpus.

En parallèle, je me suis intéressée à l'expression linguistique du désaccord en développant un système de marqueurs du désaccord à partir d'une annotation systématique des messages du corpus, qui vient donc compléter la catégorisation des fils. L'ensemble des actes de langage exprimant un désaccord ont été balisés, le désaccord étant entendu comme un acte de langage réactif énonçant une réaction négative à une assertion préalable (Kerbrat-Orecchioni 2016). À titre de contraste, les accords explicites ont également été annotés (Poudat 2018). Il me semble important d'insister sur le fait que le système de marqueurs développé est d'abord fondé sur **l'observation du corpus** ; en d'autres termes, il ne s'agit pas d'un système développé hors de tout usage et projeté sur le corpus, mais d'un système spécifique aux discussions Wikipédia, qui nécessitera d'ailleurs à terme d'être évalué et adapté pour observer d'autres genres et d'autres usages.

Depuis 2018, le corpus a été étendu dans le cadre de collaborations avec le CLLE-ERSS d'une part, et l'Université et l'Institut de la langue allemande de Mannheim d'autre part. Trois pages de discussions supplémentaires ont d'abord été adjointes au corpus. Il s'agit des discussions autour des articles *Attentats du 11 septembre 2001* (1 397 messages), *Féminisme* (423 messages) et *Vladimir Poutine* (324 messages), qui ont fait l'objet d'un signalement dans Wikipédia du fait des guerres d'édition et des interactions antagoniques qu'elles ont connues. L'ajout de ces trois pages, qui ne relèvent pas du champ des sciences et techniques, améliore la valeur d'échantillon du corpus et sa représentativité des conflits dans l'encyclopédie collaborative.

2.2 Observer le conflit sur le plan linguistique: du désaccord à l'attaque

Si les trois nouvelles pages de discussion ont été intégrées au corpus et annotées suivant les principes que nous venons d'énoncer (catégorisation des fils de discussion, et annotation des accords et des désacords), un travail sur le développement d'un système de l'attaque dans les discussions Wikipédia a été initié. Dans cette perspective, j'ai procédé à une recension des travaux autour des questions de la violence verbale et de l'insulte : (i) la *violence fulgurante* (Laforest et Moïse 2013) semble ainsi pointer vers des marqueurs explicitement conflictuels. Elle englobe plusieurs phases telles que « des moments d'incompréhension, de négociation, de menaces voire d'insultes » relatives à une montée progressive de la tension entre les interlocuteurs, et elle est associée à différents phénomènes et marqueurs linguistiques, *i.e.*

« des effets langagiers (des ruptures dans la politesse, durcisseurs, par exemple), des actes de langage dépréciatifs directs (provocation, harcèlement, mépris, reproche, insulte), des procédés argumentatifs à visée de domination » *(ibid.)*. Si ces éléments sont éclairants, ils ne sont en revanche pas directement opératoires dans une tâche d'annotation dans la mesure où chaque marqueur convoqué requiert une définition stabilisée ; (ii) en lien direct avec le conflit, qu'elle signale ou qu'elle induit, l'insulte continue pour sa part de faire couler beaucoup d'encre. Son identification reste en effet problématique du fait qu'elle nécessite de prendre en compte des données contextuelles et pragmatiques (Lagorgette, 2004). Ainsi, dans Wikipédia, le rejet ou la violation des principes fondateurs de l'encyclopédie pourra provoquer des conflits et les normes admises par la communauté des contributeurs pourront générer des insultes très spécifiques (*e.g.* tu n'es pas NPOV-conforme, ou neutre).

Au fil de ces réflexions, je me suis finalement ralliée à certaines des distinctions que proposent Laforest et Moïse (2013), qui se concentrent sur l'*objet* du conflit, qui se négocie dans l'interaction :

La violence verbale naît en partie du passage d'un conflit sur un objet (« il y a trop de bruit et ça me dérange ») à un conflit sur les personnes (« vous faites trop de bruit » – condamnation du faire – ou « vous êtes trop bruyants » – condamnation de l'être). (ibid.)

Je me suis donc concentrée sur les attaques personnelles, en différenciant les actes de condamnation de l'être et du faire. Ainsi, le passage suivant contient un acte de condamnation direct de l'être, posant ainsi que la propriété *minable* est une caractéristique permanente de la personne.

Et **si vous n'étiez si minable**, vous vous seriez tu pour ne pas vous ridiculiser davantage. (page *Féminisme*)

À cette première distinction, ont été distinguées pour l'heure trois valeurs illocutoires qui apparaissent comme dominantes dans le corpus d'étude – l'insulte, ou la catégorisation disqualifiante directe étant rare dans le corpus (Poudat et Ho-Dac 2019) :

1. **Le reproche**, acte de langage évaluant négativement le comportement d'un individu et revendiquant son changement (Kerbrat-Orecchioni 1998)
    - **Ex.** Une fois de plus tu essaies de changer le sujet pour éviter de répondre à la question posée. (page *Chiropratique*)
2. **La menace**, acte de langage formulant un avertissement, une mise en garde qui sera suivie de sanctions si le co-énonciateur ne se plie pas à ce qui est demandé
    - **Ex.** Et pire, sans référence sérieuse, ce que tu as enlevé sera rétabli, car ces personnages sont des universitaires, donc leurs publications sont des sources entrant dans les critères de wp. (page *QI*)
3. **L'ironie**, dans un usage qui relève bien souvent de l'**humour vexatoire** (Laforest et Moïse 2013)
    - **Ex.** Ces théories sont moins connu que l'avis du Pape sur le préservatif qui ne figure pourtant pas dans l'intro. (page *Attentats du 11/09*)

L'ensemble du corpus de travail a été annoté selon ces catégories ; l'annotation des attaques des trois nouvelles pages a été réalisée par un groupe d'étudiants du Master 2 *Linguistique, traitements informatiques du texte et processus cognitifs* sous ma supervision.

Les attaques ont été annotées suivant des critères syntaxiques (annotation de la phrase ou de la proposition contenant l'attaque en cas de phrase complexe contenant différents types d'actes de langage), et dans le cas de séquences d'attaques de même type, l'ensemble du passage a été annoté. Les catégories adoptées posant évidemment des problèmes d'interprétation, certains problèmes demeurent d'une catégorie à l'autre, et il nous faudra préciser et affiner les marqueurs. Pour l'expérience que nous rapportons ici, ces difficultés typologiques ne sont pas tant problématiques puisque nous nous concentrons ici sur les classifications automatiques possibles et les différences entre désaccords, attaques et passages non conflictuels, indépendamment des sous-catégorisations linguistiques plus fines mises au point.

3. Apprendre et spécifier la conflictualité ?

Après avoir décrit rapidement le modèle convolutionnel de deep learning mobilisé (3.1), nous présentons le corpus d'entraînement que nous avons construit à partir du corpus de travail annoté présenté dans la section précédente ; étant données les particularités du corpus, le modèle d'apprentissage a demandé différents ajustements (3.2). Nous nous attacherons ensuite à évaluer le modèle obtenu, les motifs sous-jacents sur lesquels il fonde son modèle pour différencier chaque catégorie grâce au TDS (Vanni et al. 2018b) et ses capacités de prédiction de passages conflictuels (3.3).

3.1 Un modèle convolutionnel de deep learning

Le modèle de *deep learning* adopté est un modèle de réseaux de neurones artificiels de type convolutionnel (CNN). Le choix d'un modèle convolutionnel s'est avéré plus adapté que son pendant, le modèle récurrent, du fait de ses bonnes performances en classification de textes (Vanni et al. 2018a). Il s'agit d'un modèle inspiré du cortex visuel, qui parcourt les données textuelles d'apprentissage au moyen d'un ensemble de filtres qui permettent d'en construire une représentation abstraite et optimale pour la tâche d'apprentissage (Vanni et al. 2020).

Outre ses excellentes performances, ce modèle a également été retenu car Vanni et al. (2018a) ont mis au point un nouvel indice de spécificité, le TDS (*Text Deconvolution Saliency*), calculé à partir d'une couche supplémentaire de déconvolution interprétant les filtres du modèle. Implémentée dans Hyperbase, cette nouvelle mesure est prometteuse pour l'ADT, permettant de mettre en évidence des passages spécifiques qui pourraient nourrir le concept de motif textuel développé par Longrée, Mellet et Luong (2008).

Dans la présente étude, c'est donc ce modèle qui a été convoqué pour évaluer, distinguer et éventuellement identifier les passages conflictuels dans les discussions Wikipédia et dans cette perspective, nous avons d'abord dû adapter le corpus et ajuster le modèle.

3.2 Particularités du corpus d'entraînement et ajustement du modèle[1]

Le corpus présente des particularités qui ont nécessité différents pré-tests afin d'obtenir un modèle d'apprentissage acceptable, sinon satisfaisant.

[1] Nous remercions vivement Laurent Vanni pour son aide et ses conseils précieux.

Calibrage du corpus au niveau des messages

Le corpus d'entraînement se divise en trois parties *désaccords / conflits / neutre* (ou *catégorie non conflictuelle*). Les actes de langage annotés étant en proportions trop restreintes pour servir de corpus d'entraînement, nous avons décidé de partitionner le corpus au niveau du message ; ainsi, un message contenant un désaccord ou une attaque a été intégré à la catégorie correspondante tandis que les messages ne contenant ni désaccords ni attaques ont été versés dans la catégorie *neutre*. Certains messages contenaient des marqueurs de désaccord et d'attaque et après réflexion, nous avons fait le choix de les affecter à la catégorie *attaque*, ce qui réduit en conséquence la proportion des désaccords.

Cette agrégation au niveau du message nous semble pertinente dans la mesure où elle présente l'intérêt de conserver l'acte de langage annoté dans son contexte textuel. Les analyses qui suivent nous permettront ainsi de mettre en évidence des régularités tant linguistiques que cotextuelles.

Tableau 1 : Distribution des catégories dans le corpus d'entraînement

Catégorie	***Nombre de mots-formes***
Attaques	55 924
Désaccords	23 361
Neutre	391 849

Comme le montre le tableau 1, il s'agit d'abord d'un corpus qui a dans son état actuel une taille très limitée, en comparaison avec les *benchmarks* classiquement exploités.

Équilibrage du corpus d'entraînement

Nous obtenons donc un corpus d'entraînement sensiblement déséquilibré mais qui reflète pourtant ce que l'on observe dans les données, sélectionnées du fait de la présence d'un conflit sur la page : environ 80% des messages ne porteraient pas de traces d'attaques ni de désaccords, pourcentage qui semble cohérent. Environ 40% des fils de discussion du corpus sont en effet non conflictuels (C0) tandis qu'une interaction même très conflictuelle n'est jamais constituée de séquences d'attaques seules, mais d'échanges hétérogènes. Comme nous l'avions souligné dans la section précédente, rares sont les insultes directes, et encore plus inhabituelles sont les séquences d'attaques directes : s'il n'y a pas de modération des messages dans les fils de discussion comme dans les forums, une régulation est à l'œuvre avec des processus institutionnalisés de signalement et de médiation.

Cette distribution fidèle à la réalité des interactions pose néanmoins un problème d'équilibre du corpus d'entraînement et a nécessité un ajustement du modèle et de ses hyperparamètres.

Ajustement du modèle et réglage des hyperparamètres

Différents pré-tests ont donc été effectués, notamment pour éviter le sur-apprentissage en réduisant la taille des couches cachées, *i.e.* des réseaux du *deep learning*. Il s'agit en effet

d'une étape nécessaire pour améliorer les capacités d'identification des catégories du modèle, qui doit être capable de s'approprier d'autres données que celles du corpus d'entraînement. Nous avons construit deux modèles, sur le texte brut d'une part et sur le texte annoté avec TreeTagger d'autre part et dans les deux cas, il s'est avéré nécessaire de rendre les parties plus homogènes en taille, en réduisant donc le nombre de séquences neutres considérées. Un échantillon aléatoire de 100 000 occurrences a donc été constitué à cet effet. Par ailleurs, étant données les spécificités du corpus, et notamment le fait qu'il s'agit d'interactions constituées de messages souvent courts, la fenêtre contextuelle a été réduite à 20 mots-formes[2]. Au final, on a donc obtenu deux modèles, ayant un taux de précision de 71% pour le modèle construit sur les textes bruts et de 80% pour le modèle construit sur les textes munis d'une couche morphosyntaxique.

3.3 Caractéristiques du désaccord et de l'attaque

Nous nous attachons ensuite à évaluer l'intérêt des modèles d'apprentissage obtenus en observant d'une part les motifs sous-jacents sur lesquels il fonde son modèle pour différencier chaque catégorie grâce au TDS (Vanni et al. 2018b) et ses capacités de prédiction de passages conflictuels d'autre part.

Passages clefs

Il faut d'abord souligner que le modèle construit sur les textes munis d'une annotation morphosyntaxique stimule davantage l'interprétation tout en permettant une meilleure prédiction des catégories.

La part très réduite des messages contenant une marque de désaccord a rendu leur identification difficile : sans annotation morphosyntaxique, les désaccords du corpus ne sont pas reconnus comme tels tandis que le second modèle en reconnait 44%. L'examen des passages clefs identifiés comme spécifiques avec le TDS montre que c'est essentiellement le niveau des parties du discours qui est sollicité par le système, faisant ressortir essentiellement des séquences de ce type :

" [...] ADJ PRP DET:ART NOM SENT PUN:cit puisque ADV PRP NOM ADJ PRP DET:ART matière PUN:cit PUN DET:ART NOM PRP PRO:REL [...] "

Légende : Parties du discours / Lemmes / Mots-formes

L'usage du connecteur de causalité rhétorique *puisque* nous a interpellée, et il s'avère qu'il est en effet sur-représenté dans les passages contenant des désaccords, comme le montre la Figure 1.

[2] Il s'agit d'une fenêtre contextuelle glissante, non bordée par des signes de ponctuation.

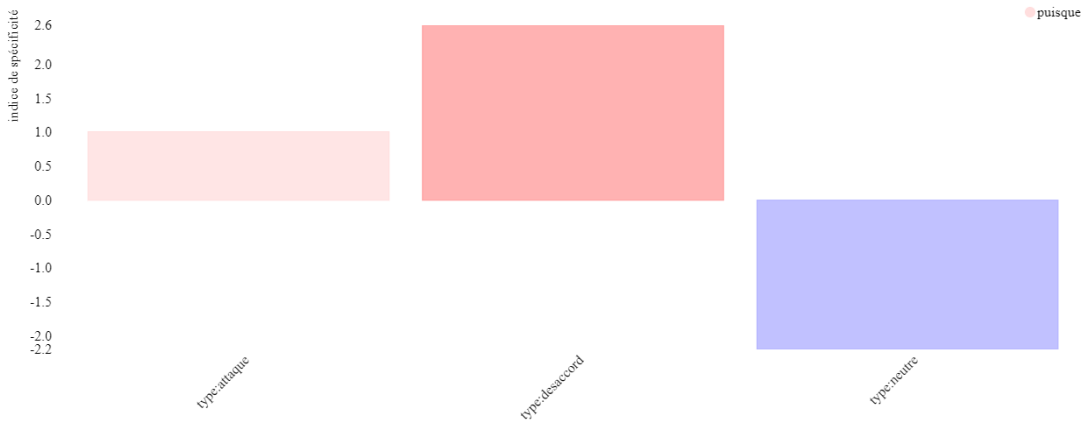

Figure 1 : Distribution spécifique de puisque d'une catégorie à l'autre

Le désaccord s'inscrit ainsi dans un processus argumentatif : le locuteur se confronte à l'autre, qu'il faut convaincre. L'attaque met quant à elle en évidence d'autres phénomènes, moins rhétoriques puisque l'échange d'arguments a échoué. C'est peut-être d'ailleurs l'une des raisons qui expliquent pourquoi les attaques posent moins de problèmes d'identification, comme le montre la figure 2.

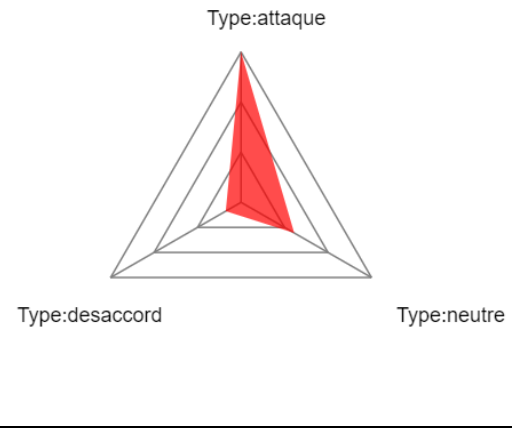	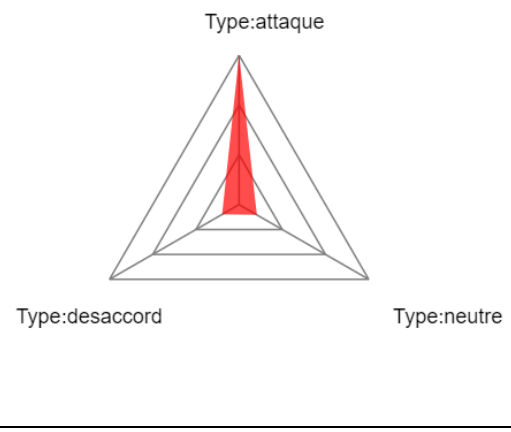
Modèle texte brut (75,24% prédiction)	**Modèle avec annotation morphosyntaxique (92,55% prédiction)**

Figure 2 : Prédiction des attaques avec les deux modèles

L'examen du TDS m'a permis d'observer deux particularités :

(i) le modèle fondé sur les mots-formes fait ressortir un motif contenant les traits sémantiques /insuffisance/ et /limitation/, caractéristique tant du reproche que de l'ironie vexatoire :

" [...] ? Sinon , tu t' es enfin décidé à corriger tes assertions sur Bentham ? C' est Calixte1 et pas [...] "

" [...] illuminés ? vous êtes à courts d' arguments donc ca y est on passe aux sophismes . ca m' a [...] "

(ii) le modèle construit sur le corpus annoté en morphosyntaxe met pour sa part en évidence un motif caractéristique du reproche, combinant l'usage répété du verbe *arrêter* et certains lieux conflictuels récurrents dans le corpus (*mauvaise foi, sans consensus, passage en force*) :

" [...] . arrêter de supposer le mauvais foi , arrêter de passer PRP force sans consensus , arrêter de ajo uter ton [...] "

Détection

Nous avons également procédé à un dernier test, en soumettant le passage suivant (Figure 3), qui nous semble nettement conflictuel aux deux modèles d'apprentissage.

> Tu n'es pas dans une cour de collège ou, pire, sur un forum New Age : tu ne duperas personne ici. Tes égosillements complotistes ne font que saper un peu plus ta crédibilité. Totodu74 (devesar…) 6 novembre 2019 à 21:01 (CET)
>
> > Totodu74, c'est toi qui es pris en flagrant délit de mensonge : dans l'étude de 1997, les chercheurs ne poursuivent pas comme tu l'écris, avec « pour un quelconque clinique » mais « dans tous les cas de figure ». Et tu ne réponds pas sur le fond de l'attitude des chercheurs qui, tout en reconnaissant le résultat positif de la méta-analyse, témoignent clairement leur souhait qu'on cesse de leur faire perdre du temps avec l'homéopathie. Le comble de la mauvaise foi a été atteint avec l'étude [référence] (citée dans l'article) où les chercheurs ont "oublié" trois travaux qui répondaient parfaitement aux critères exigibles, et dont les résultats, si on les incluait, penchaient clairement en faveur de l'efficacité de l'homéopathie. Je ne suis donc pas le seul à crier au complot, rejoignant en cela Ludtke et Rutten dans le **Journal of Clinical Epidemiology** (2008) (voir l'article homéopathie à ce sujet). On ne vit que deux fois Me contacter 7 novembre 2019 à 04:03 (CET)

Figure 3 : Extrait conflictuel de la page de discussion Homéopathie[3]

Les résultats sont intéressants et laissent entrevoir une exploitation possible du TDS comme aide à l'annotation des attaques : la figure 4 synthétise les passages identifiés comme conflictuels (en gras, rouge) et ceux identifiés comme neutres (surlignage gris) avec le modèle de *deep learning* entraîné sur le texte brut – qui évalue l'extrait comme constitué de 20% d'attaques et de 80% de passages neutres.

> **Tu n'es pas** dans une cour **de collège ou**, pire, sur **un forum New Age : tu ne duperas personne ici. Tes égosillements complotistes** ne font que saper un peu plus ta crédibilité. Totodu74 (devesar…) 6 novembre 2019 à 21:01 (CET)
>
> > Totodu74, c'est toi qui es pris en flagrant délit de mensonge : dans l'étude de 1997, les chercheurs ne poursuivent pas comme tu l'écris, avec « pour un quelconque clinique » mais « dans tous les cas de figure ». **Et tu ne réponds** pas sur le fond **de l'attitude** des chercheurs qui, tout en reconnaissant le résultat positif de la méta-analyse,

[3] https://fr.wikipedia.org/wiki/Discussion:Hom%C3%A9opathie

témoignent clairement leur souhait qu'on cesse de leur faire perdre du temps avec l'homéopathie. Le comble de la **mauvaise foi a été atteint avec** l'étude [référence] (citée dans l'article) où les chercheurs ont "oublié" trois travaux qui répondaient parfaitement aux critères exigibles, et dont les résultats, si on les incluait, penchaient clairement en faveur de l'efficacité de l'homéopathie. Je ne suis donc pas le seul à crier au complot, rejoignant en cela Ludtke et Rutten dans le **Journal of Clinical Epidemiology** (2008) (voir l'article homéopathie à ce sujet). On ne vit que deux fois Me contacter 7 novembre 2019 à 04:03 (CET)

Figure 4 : Passages identifiés comme neutres vs conflictuels avec le TDS texte brut

Le modèle entraîné sur le texte annoté en morphosyntaxe reconnaît davantage d'attaques, à hauteur de 48% (vs 52% de séquences jugées neutres). Les passages les plus significatifs ramenés se situent dans les mêmes zones textuelles que celles que nous avons reportées dans la figure 4 mais privilégie dans certains cas la catégorie morphosyntaxique ou le lemme. Par exemple, le passage qui suit pointe également sur la séquence « Tu ne duperas personne ici », qui est une attaque nette. En revanche, c'est le futur encadré des deux adverbes de négation qui est mis en évidence, ce qui me semble tout à fait pertinent : on sait en effet que le futur assorti d'une négation renvoie dans une très grande majorité des cas à l'interdiction ou au refus net, ce qui signale un conflit dans le contexte.

" [...] ADV VER:futu ADV ici . ton égosillement ne faire KON VER:infi DET:ART peu plus ton crédibilit é . ([...] "

4. Conclusion

Nous avons proposé quelques pistes de réflexion autour de l'utilisation du *deep learning* dans le cadre d'une tâche d'annotation pragmatique de corpus. L'expérience menée m'a semblée particulièrement intéressante, ouvrant des perspectives pour l'avenir et posant des questions pertinentes tant sur mon corpus de travail que sur les marqueurs du conflit.

Concernant d'abord le corpus de travail, il faut souligner que la taille réduite des données a limité l'expérience présentée, les systèmes d'apprentissage informatique nécessitant des données volumineuses. Le corpus et les annotations sont encore en cours de développement, et les résultats devraient être plus significatifs dans les années à venir. Le bilan de l'expérience est néanmoins satisfaisant et les séquences identifiées au moyen du TDS, outre leur intérêt descriptif, nous laissent entrevoir différentes applications, de l'exploitation des motifs multi-niveaux signalés pour procéder à une annotation semi-automatique du corpus avec TXM par exemple (concordance sur index pour annoter) à l'ajustement possible dans Hyperbase Web d'un système d'aide aux annotations fines (d'ordre sémantique et pragmatique) fondé sur le TDS.

Il me semble également important de souligner que notre démarche méthodologique participe au renouvellement de l'ADT sur au moins deux points : (i) la mobilisation du deep learning pourrait être *ex nihilo* considérée comme hors cadre, mais le développement de l'indice TDS et son potentiel de nouvel indice de spécificité et d'identification de motifs textuels en fait au contraire un instrument qui s'inscrit sans aucun doute dans une démarche d'exploration de corpus ; (ii) l'exploitation de données sémantiquement annotées se situe également sur la brèche, mais il me semble que le cadre que nous avons mis en œuvre reste dans le champ : la

question du choix linguistique (quels marqueurs adoptés parmi les choix possibles dans l'usage considéré, *e.g.* pour exprimer le désaccord ou attaquer la face d'autrui), qui est une vraie question linguistique, peut parfaitement être explorée contextuellement dans un cadre ADT. Ainsi, en va-t-il par exemple du fait que nous n'ayons pas considéré les marqueurs annotés seuls mais le cadre de leur message d'origine.

Enfin, je terminerai cet article en soulignant une particularité du corpus d'entraînement qui m'a interpellée ; le fait de devoir réduire la proportion de séquences non conflictuelles m'a interrogée, sur un plan peut-être plutôt philosophique. En effet, comme je l'ai souligné dans ce texte à maintes reprises, les séquences entièrement conflictuelles sont des événements rares dans les discussions Wikipédia (et dans la vie réelle), et cette proportion de 80% de passages non conflictuels me paraissait tout à fait pertinente. Puisque nous sommes dans un cadre d'apprentissage automatique, je me demande comment intégrer ce facteur événementiel au processus. Car la catégorisation adoptée, qui distingue entre *désaccords*, *attaques* et *harmonie*, est particulière en ce sens qu'elle est sélective. Si j'ai développé un système de marqueurs linguistiques permettant de spécifier et d'identifier les désaccords et les attaques, je ne travaille pas sur l'harmonie, qui est une catégorie « autre ». Une interaction non conflictuelle entre contributeurs dans Wikipédia aurait-elle des particularités ? L'examen des particularités des passages neutres avec le TDS pourrait là encore nous apporter quelques éléments de réponse, entre humour, objectivité et politesse…

" [...] passerait à la télé et reprocherait à certains téléspectateurs de l' avoir regardée et critiquée . C a ferait rigoler tout [...] "

" [...] doit également être objectif et les sources doivent l' être aussi , même critiques . Je ne suis pas contre [...] "

" [...] . En conséquence , j' ai déplacé l' intervention qui précède ici . Bref , j' aimerais savoir quel est [...] "

Références

Auray, N., Hurault-Plantet, M., Poudat, C., & Jacquemin, B. (2009). La négociation des points de vue : une cartographie sociale des conflits et des querelles dans le Wikipédia francophone. In *Réseaux* 2/2009, n° 154: 15-50.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Longhi, J. et Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In Beißwenger, M., Oostdijk, N., Storrer, A. et van den Heuvel, H. (ed.), Special Issue of *Journal of Language Technology and Computational Linguistics* "Building and annotating corpora of computer-mediated discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics", pp. 1-30.

Denis, A., Quignard, M., Fréard, D., Détienne, F., Baker, M., & Barcellini, F. (2012, juin 4). *Détection de conflits dans les communautés épistémiques en ligne*. TALN - Actes de la Conférence sur le Traitement Automatique des Langues Naturelles - 2012.

Ho-Dac, M., Laippala, V., Poudat, C. et Tanguy, L. (2017). « Exploring Wikipedia Talk Pages for Conflict Detection » in Fiser, D. and Beisswenger, M. (ed.), Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World. Book series Translation Studies and Applied Linguistics. Ljubljana University Press, Ljubljana, pp. 146-172.

Kerbrat-Orecchioni, C. (1998). *Les interactions verbales. Tome 1* (3ème édition). Armand Colin.

Kerbrat-Orecchioni, C. (2016). Le désaccord, réaction « non préférée » ? Le cas des débats présidentiels. *Cahiers de praxématique, 67*.

Laforest, M. et Moïse, C. (2013). Entre reproche et insulte, comment définir les actes de condamnation ? In B. Fracchiolla, C. Moïse, & C. R. et N. Auger (Éd.), *Violences verbales. Analyses, enjeux et perspectives* (p. 85-105). Presses universitaires de Rennes.

Lagorgette, D. (2004). Les insultes : approches sémantiques et pragmatiques. In *Langue Française* 144, Lagorgette D. et P. Larrivée eds

Lebart, Ludovic, Pincemin, Bénédicte et Poudat, Céline (2019). *Analyse des données textuelles*. Presses de l'Université du Québec.

Longrée, D., Mellet, S., et Luong, X. (2008). Les motifs : Un outil pour la caractérisation topologique des textes. In *Actes des JADT 2008*, vol. *2*, pp. 733-744.

Poudat, C. (2018). Explorer les désaccords dans les fils de discussion du Wikipédia francophone. In Iezzi D. F., Celardo, L. et Misuraca, M., *Actes des JADT 2018*, Rome, vol. 2, pp. 602-610.

Poudat, Céline, Grabar, Natalia, Paloque-Berges, Camille, Chanier, Thierry et Kun, Jin (2017). « Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone » *in* Wigham, C.R & Ledegen, G., *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Collection Humanités numériques. Paris : L'Harmattan, pp. 19-36.

Poudat, C. et Ho-Dac, L.-M. (2019). Désaccords et conflits dans le Wikipédia francophone. In Col, G. et Hanote, S. « Accord et désaccord », *Travaux linguistiques du Cerlico*, n°29, pp. 155-176.

Poudat, C. et Landragin, F. (2017). *Explorer un corpus textuel. Méthodes – Pratiques – Outils*. Collection Champs linguistiques, De Boeck, Louvain-la-Neuve.

Poudat, C., Vanni, L. et Grabar, N. (2016). How to explore conflicts in Wikipedia talk pages? In Mayaffre, D, Poudat, C., Vanni, L., Magri, V. et Follette, P., *Actes des JADT 2016*, Nice, vol. 2, pp. 645-656.

Vanni L., Ducoffe M., Mayaffre D., Precioso F., Longrée D. et al. (2018a). Textual Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis In Proceedings Of The 56th Annual Meeting of the Association for Computational Linguistics. ACL 2018, Melbourne, Volume 1 p. 548-557.

Vanni L., Mayaffre D., Longrée D. (2018b). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. 14es Journées internationales d'Analyse statistique des Données Textuelles. JADT 2018, Rome, p. 459-466.