Évaluation de techniques non supervisées pour l'assistance à l'annotation manuelle de textes

Kévin Deturck, Hugo Lafayette, Bénédicte Parvaz Ahmad, Ilaine Wang, Afala Phaxay, Damien Nouvel

Évaluation de techniques non supervisées pour l'assistance à l'annotation manuelle de textes

Kévin Deturck[1] Hugo Lafayette[2] Bénédicte Diot-Parvaz Ahmad[1] Ilaine Wang[1]
Afala Phaxay[1] Damien Nouvel[1]
(1) Inalco – Ertim, 2, rue de Lille, 75007 Paris, France
(2) Kairntech, 29, chemin du vieux chêne, 38240 Meylan, France
kevin.deturck@inalco.fr, hugo.lafayette@kairntech.com,
benedicte.parvazahmad@inalco.fr, ilaine.wang@inalco.fr,
afala.phaxay@inalco.fr, damien.nouvel@inalco.fr

RÉSUMÉ

La production de données annotées est critique pour l'élaboration de nombreux systèmes de TAL, avec des problèmes liés à l'expertise des annotateurs et à l'organisation de leur travail dans l'optique d'obtenir des annotations de qualité. Nous évaluons l'impact de techniques d'assistance non supervisées pour la catégorisation manuelle de textes, lors de trois campagnes d'annotation d'articles journalistiques impliquant une vingtaine d'annotateurs à travers trois langues, l'hindi, le mandarin et l'arabe. Cette évaluation est quantitative, considérant le niveau de qualité ainsi que la vitesse de l'annotation, et qualitative en intégrant les retours d'expérience des annotateurs. Elle a permis la mise en œuvre innovante d'assistances basées sur des combinaisons d'approches standards et nouvelles ainsi que d'éprouver une méthode pour l'encadrement de campagnes d'annotation multi-annotateurs.

ABSTRACT

Evaluation of unsupervised techniques for assisted manual text categorisation
The production of annotated data is critical for the development of many NLP systems, with problems related to the expertise of the annotators and the organisation of their work in order to obtain quality annotations. We evaluate the impact of unsupervised techniques for assisted manual categorisation of journalistic articles by means of three annotation campaigns involving twenty annotators, each one being dedicated to one language among Hindi, Mandarin and Arabic. The evaluation is both quantitative, considering the level of quality as well as the speed of the annotation, and qualitative by integrating the annotators' feedback. It allowed for an innovative implementation of assistance tools based on standard and new approaches as well as the testing of a method for managing multi-annotator annotation campaigns.

MOTS-CLÉS : campagne d'annotation, classification de texte, TAL, zeroshot, bertopic
KEYWORDS: annotation campaign, NLP, text classification, zeroshot, bertopic

1 Introduction

La collecte et l'exploitation des données langagières sont devenues aujourd'hui des enjeux stratégiques dans la vie économique et sociale. Les apports de l'intelligence artificielle au TAL ont transféré les difficultés de mise au point d'un projet sur la production de données annotées de qualité, préalable indispensable à la création de modèles capables d'effectuer une tâche de TAL efficacement.

Le travail que nous présentons ici s'inscrit dans le cadre du projet VITAL (Valorisation de l'Innovation pour le Traitement Automatique des Langues), qui consiste à étudier de nouveaux outils pour produire des données annotées avec le meilleur compromis entre qualité, coût et temps de développement. Ce projet porte distinctement sur trois phases d'une campagne d'annotation : l'amorçage, la consolidation et la clôture.

C'est un partenariat industrie-recherche. Kairntech développe et commercialise depuis trois ans une plateforme de TAL qui permet, par une interface graphique simple d'utilisation, de créer des jeux de données annotées puis de les utiliser en implémentant, évaluant et en maintenant des modèles d'IA. Pour Vital, Kairntech développe des outils d'assistance à l'annotation manuelle sous forme de prototypes dans leur plateforme d'annotation. Le laboratoire de recherche ERTIM de l'Inalco, spécialiste en TAL multilingue, gère les campagnes d'annotation pour Vital. Le projet est financé par le dispositif RAPID, porté par l'agence de l'innovation de défense.

Ce travail concerne l'évaluation d'outils d'assistance à l'amorçage d'une tâche de classification manuelle de textes en hindi, mandarin et arabe, avec des techniques non supervisées. Dans la suite, nous positionnons ce travail par rapport à l'état de l'art en section 2, puis, nous décrivons les campagnes d'annotation organisées et les techniques d'assistance étudiées respectivement en sections 3 et 4, et enfin, nous présentons l'analyse et les résultats en section 5 avant de conclure en section 6.

2 État de l'art

2.1 Campagne d'annotation : organisation et analyse

Il a été démontré qu'une phase de formation des annotateurs était un levier fondamental pour maximiser la qualité de leur travail (Bayerl & Paul, 2011), avec en support, un guide d'annotation pour la cohérence des annotations (Alex et al., 2006). Ces composantes s'insèrent dans un processus d'annotation dont il existe deux grands types (Fort, 2012) : un dit « traditionnel », avec une seule itération des phases de préparation, d'annotation et d'analyse, et un autre, plus récent, dit « agile », avec plusieurs itérations progressives de ces phases. Ces révisions font suite aux retours d'expérience des annotateurs ou à l'analyse de leurs désaccords d'annotation. Ces désaccords peuvent être résolus par un expert de la tâche ou par les annotateurs eux-mêmes. Nos campagnes intègrent ces composantes pour donner un cadre standard à notre étude.

Pour évaluer la qualité des annotations produites, une approche classique est de mesurer l'accord inter-annotateur (Artstein & Poesio, 2008). Le critère de qualité sous-jacent est la reproductibilité du schéma d'annotation par plusieurs annotateurs (Krippendorff, 2004) ; un schéma d'annotation est d'autant plus fiable qu'il est reproduit par plusieurs annotateurs. Ce type d'approche a notamment l'avantage de ne pas nécessiter de référence, souvent indisponible, mais ne permet pas véritablement de dire si un ensemble d'annotations est correct, à l'inverse des métriques de comparaison à une référence, comme la F-mesure ou le « Slot error rate », moins représentées. Nous utiliserons la F-mesure parce que le protocole d'évaluation choisi nous a permis d'acquérir un corpus de référence.

2.2 Assistance à l'amorçage de la catégorisation manuelle de texte

« Zero-shot » est la dénomination d'un ensemble de techniques d'apprentissage se proposant de résoudre une tâche sans avoir d'exemple à disposition, puisqu'une description des différentes classes attendues suffit (Winata et al. 2021). Nous en avons choisi une implémentation à base de transformeurs et de modèles d'inférence en langage naturel, adaptée à notre cas d'usage de classification de documents. Les performances de modèles multilingues comme mBERT sont réputées bonnes sur des langues dotées, comme l'arabe et le mandarin (Wu & Dredze. 2020). BERTopic, qui exploite ces modèles multilingues, semble apporter des améliorations par rapport à l'état de l'art de la classification thématique de documents (Egger & Yu 2022). Il nous semble alors opportun de tester l'usage et les performances de BERTopic dans le cadre de VITAL.

Les outils de recherche par mots-clés ainsi que par facettes sont assez standards pour l'exploration de corpus, notamment dans le domaine des humanités numériques (Schnober & Gurevych, 2015). Cependant, l'étude de leur impact sur une tâche de classification manuelle n'est pas documentée à notre connaissance. Par ailleurs, certaines plateformes d'étiquetage manuel, comme Prodigy et TagTog, ne fournissent pas ou ne mettent pas en avant ces outils d'assistance. Nous souhaitons évaluer ce qu'ils apportent en pratique durant des campagnes d'annotation.

3 Les campagnes d'annotation

3.1 La tâche d'annotation

La tâche d'annotation du lot 1 consiste à catégoriser en thème chacun des articles de journaux, tous écrits dans une même langue, en choisissant une seule étiquette de catégorie parmi un ensemble d'étiquettes imposé. Cela correspond à de la classification « multiclasse » (Rifkin, 2008) parce qu'il y a plusieurs étiquettes de catégorie possibles et une seule doit être choisie. Outre les étiquettes de catégorie, il y a aussi la possibilité de sélectionner une étiquette intitulée « Incertitude » quand aucune

étiquette de catégorie ne convient ou lorsqu'il y a un doute sur le choix d'une étiquette de catégorie, ceci afin de tracer les difficultés rencontrées.

La tâche d'annotation requiert de repérer le sujet majeur d'un document pour lui associer une catégorie. Ce travail est difficile parce que les articles journalistiques mêlent parfois plusieurs domaines de connaissance, ce qui laisse place à l'interprétation et donc à la subjectivité pour déterminer ce qui est majeur, compliquant l'obtention d'annotations cohérentes. La tâche est par ailleurs complexifiée par la nécessité de considérer l'ensemble du texte d'un article (par une attention soutenue) avant de prendre une décision pour l'annoter, ainsi que par le besoin de disposer d'éléments de connaissance sur les domaines abordés pour éclairer la décision.

3.2 Les étiquettes de catégorie

Hindi	Mandarin	Arabe
7 étiquettes	7 étiquettes	6 étiquettes
- Sport - Économie - Divertissements et médias - Société - Monde - Politique - Sciences et techniques	- Sport - Divertissements - Société - Monde - Culture - Santé et bien-être - Anti-corruption	- Sport - Économie - Politique internationale - Politique d'Oman - Art et culture - Religion

TABLE 1 : Les étiquettes de catégorie par campagne

Les journaux sources des corpus avaient leur catégorisation respective des documents. Cependant, nous avons choisi de ne pas reprendre les catégorisations originelles car nous avions observé qu'elles manquaient parfois de cohérence : par exemple, dans le corpus en hindi, un article sur le discours du Premier Ministre indien en Chine était catégorisé tantôt en « International », tantôt en « Politique ».

Pour définir les jeux d'étiquettes (cf. Table 1), nous avons fait appel à un « référent langue » par campagne : c'est une locutrice ou un locuteur distinct des annotateurs. Le travail du référent langue a été de faire une pré-annotation d'une petite partie du corpus (d'une cinquantaine à une centaine de documents), en utilisant au départ le jeu de catégories originel et en ayant la liberté de le modifier. Ce travail de pré-annotation par les référents langue fut à la base de la rédaction du guide d'annotation de chacune des campagnes.

3.3 Les corpus

Nous avons constitué un corpus d'articles journalistiques pour chacune des trois campagnes. Le corpus pour la campagne sur l'arabe provient d'un corpus pré-existant, « El-Watan 2004 » (Abbas et al., 2005), c'est le seul à être issu d'une source unique, le journal arabe « El Watan », en édition omanaise et en ligne. Nous avons extrait les deux autres corpus à partir de plusieurs journaux en ligne, sept pour le corpus en hindi et trois pour le corpus en mandarin.

Pour favoriser la comparabilité des campagnes, nous avons choisi de constituer des corpus de tailles similaires Le nombre de 900 documents a été initialement choisi afin de maximiser le nombre de documents annotés pendant chaque session d'annotation. Le corpus en arabe a un nombre un peu plus élevé de documents (954) parce que nous avions observé qu'il était judicieux de proposer un peu plus de documents par session d'annotation pour laisser plus de place à d'éventuels contrastes concernant la vitesse d'annotation entre les annotateurs, en particulier selon l'utilisation ou non d'une assistance.

3.4 Les annotatrices et annotateurs

Campagne	Hindi	Mandarin	Arabe
Nombre	6	5	10
Mode d'apprentissage de la langue	3 « natale » 2 « héritée » 1 « acquise »	5 « natale »	6 « natale » 4 « acquise »

TABLE 2 : Statistiques sur les annotateurs par campagne

Le mode d'apprentissage de la langue visée diffère entre les annotatrices et annotateurs (cf. Table 2). Pour beaucoup, c'est une langue « natale », apprise dès la naissance, pour d'autres, une langue « héritée », correspondant aussi à un apprentissage par l'environnement familial mais plus tardif, et pour d'autres encore, une langue « acquise » par la formation. Cette variété est positive pour notre champ d'application mais peut engendrer un biais de compétence à intégrer dans l'analyse.

Afin d'évaluer comparativement l'impact des outils d'assistance, nous avons conçu trois types de groupe d'annotateurs : un groupe témoin « Sans assistance » dont les annotateurs n'ont jamais accès aux outils d'assistance, un groupe « Avec assistance » avec les outils d'assistance constamment à disposition, et a un type de groupe entre-deux, c'est le groupe « Alternance », qui passe d'une session d'annotation avec les outils d'annotation à une session sans, ou inversement. Chaque type de groupe a été représenté par un à deux annotateurs.

3.5 Les techniques d'assistance étudiées

Nous avons étudié une aide à la sélection de catégorie, qui consiste en la suggestion d'une étiquette par deux systèmes non supervisés : un système standard basé sur une technique « Zero-shot » (Pourpanah & Abdar, 2022), et un système innovant qui associe le précédent à la technique BERTopic (Grootendorst, 2022). Nous avons aussi mis en œuvre une aide à la sélection de documents, regroupant des outils standards permettant de filtrer les documents (mots-clés, thèmes, …). Enfin, une assistance d'aide à la lecture est apparue comme un besoin récurrent durant les deux premières campagnes, et a été étudiée pour la dernière campagne, sur l'arabe. Elle consiste en la mise en exergue des phrases discriminantes et l'identification d'entités nommées dans les articles.

Les campagnes sur l'hindi et sur le mandarin incluent un seul type d'assistance, que nous nommons « assistance A » et qui correspond à la combinaison de l'aide à la sélection de catégorie et de l'aide à la sélection de documents. La campagne sur l'arabe a la particularité d'inclure un second type d'assistance, l'« assistance B », qui est l'aide à la lecture.

3.6 Le déroulement

La première séquence de nos campagnes d'annotation est dédiée à la présentation des enjeux, du déroulement et du guide d'annotation. Nous recommandons notamment aux annotatrices et annotateurs de privilégier la qualité à la quantité. Les deux autres séquences comprennent des sessions d'annotation, définies par une durée et un sous-corpus, dans différents formats que nous préciserons. Adoptant une organisation agile (cf. section 2.1), à la fin de chaque session d'annotation, nous demandons aux annotatrices et annotateurs de remplir un formulaire de retour d'expérience, puis nous les faisons participer à une réunion de réconciliation qui vise à résoudre les cas d'incertitude et à obtenir une annotation commune lorsqu'il y a un désaccord. Ces sessions duraient de 15 à 40 minutes selon le temps disponible et permettaient de traiter environ 5 à 20 documents. La fin du cycle est marquée par la révision éventuelle du guide d'annotation.

La première séquence d'annotation est dédiée à la formation et à l'évaluation initiale des annotatrices et annotateurs, qui sert à quantifier un éventuel biais de compétence. Cela se déroule pour toutes et tous dans chacun des modes d'annotation étudiés par campagne : sans assistance, avec l'assistance A pour les trois campagnes et, pour la campagne sur l'arabe, avec l'assistance B. La formation consiste en la prise en main de la plateforme d'annotation par des sessions d'entraînement de 10 minutes et une vingtaine de documents proposés ; exceptionnellement, les annotatrices et annotateurs était autorisés à échanger entre eux pendant l'annotation. Les sessions d'évaluation initiale durent 45 minutes avec environ 130 documents proposés. La seconde séquence d'annotation, comprend quatre sessions de 50 minutes, avec environ 140 documents proposés, embarquant différents modes par la répartition des annotateurs dans les groupes tels que décrits en section 3.4.

4 Analyse

4.1 Méthodologie

L'analyse d'impact des techniques d'assistance a pour objectif de déterminer si celles-ci améliorent, détériorent ou n'ont pas d'effet notable sur le travail des annotatrices et annotateurs à travers les trois campagnes. Elle porte sur deux critères de performance : la qualité et la vitesse d'annotation.

Dans un premier temps, nous quantifions l'éventuelle différence de compétence des annotatrices et annotateurs entre les modes. Pour ce faire, nous calculons la différence entre les résultats des évaluations initiales reliés aux modes suivant la proportion de représentation des groupes lors des sessions en embarquant plusieurs. Par exemple, les annotatrices et annotateurs « Avec assistance A » lors des sessions embarquant différents modes affichent une moyenne d'évaluations initiales à 0,85 de F-Mesure, tandis que celles et ceux ayant pratiqué le mode « Sans assistance » produisent une F-Mesure moyenne à 0,82. Le biais de compétence en faveur du mode « Avec A » est égal à 0,85 - 0,82, soit 3 points de F-Mesure, dont nous tiendrons compte dans le calcul d'impact de l'assistance A.

$$\text{Impact(Avec)} = \text{Résultat(Avec)} - \text{Résuiltat(Sans)} - \text{Biais(Avec)}$$

ÉQUATION 1 : Formule de calcul d'impact d'une assistance

Dans un second temps, nous comparons les moyennes des résultats de vitesse et de qualité selon les modes à l'échelle des sessions en embarquant plusieurs, avec le mode « Sans assistance » servant de mode témoin. Nous intégrons à cette comparaison le biais de compétence décrit précédemment afin que les différences mesurées reflètent essentiellement l'impact des outils d'assistance (cf. Équation 1).

Pour évaluer la qualité des annotations, nous les comparons à une « référence », c'est-à-dire un ensemble d'annotations que nous considérons comme suivant précisément le guide d'annotation. Pour constituer cette référence, nous utilisons d'une part les annotations issues de la réconciliation et d'autres part celles correspondant à un certain niveau d'accord. Pour l'hindi et le mandarin, l'accord minium requis est d'au moins trois annotatrices ou annotateurs, tout le monde s'étant exprimé (environ 400 documents). Pour l'arabe, dont la campagne a davantage d'annotatrices et annotateurs (cf. section 3.4), le niveau minimum d'accord est d'au moins quatre annotatrices ou annotateurs et au maximum une divergence (environ 120 documents).

Nous utilisons la formule de la F-mesure par annotatrice ou annotateur G sur un sous-ensemble de documents qui est l'intersection de l'ensemble des documents de la référence et de l'ensemble des documents annotés par G. Autrement dit, nous n'évaluons la qualité qu'à partir des documents sur lesquels il y a une annotation de référence et une annotation de G. En ce qui concerne la vitesse

d'annotation, nous utilisons le volume de documents annotés, en pourcentage du volume de documents proposés par session (cf. section 3.6), et la durée d'annotation.

4.2 Résultats

Métrique	F-mesure	Volume (%)	Durée
Assistance A	+3,3pts	+2,43pts	0min
Assistance B	+2pts	-1,5pt	-2min

TABLE 3 : Statistiques sur les annotateurs par campagne

L'analyse des sessions embarquant différents modes à travers les trois campagnes semble montrer que l'assistance A aide à gagner en qualité d'annotation, avec une moyenne à 3,3 points supplémentaires de F-Mesure. L'impact de l'assistance A sur la vitesse d'annotation semble globalement négligeable, avec cependant 2,43 points d'impact positif en moyenne sur le pourcentage de volume d'annotation. Les retours d'expérience disent que la suggestion de catégorie, par un niveau de qualité variable, a parfois eu tendance à créer davantage de remises en question voire de confusion chez les annotateurs, induisant une vitesse d'annotation amoindrie.

L'assistance B, qui ne concerne quant à elle que la campagne sur l'arabe, semble avoir produit un gain en qualité évalué à 2 points de F-Mesure. Concernant la vitesse d'annotation, nous avons relevé globalement un impact négligeable, avec une perte en moyenne de 1,3 point en pourcentage de volume annoté mais 2 minutes en moins de durée d'annotation. D'après les retours d'expérience, l'assistance B a aidé à prendre connaissance du contenu des documents plus rapidement, par la mise en exergue des passages clés ainsi que des entités nommées. Cependant, la pré-catégorisation des passages était parfois fausse, engendrant un temps de réflexion supplémentaire.

5 Conclusion et perspectives

Grâce à ce partenariat, nous avons pu étudier avec succès l'apport de certaines assistances durant la phase d'amorçage d'une campagne d'annotation, et constaté des gains mesurables lorsqu'elles sont de bonnes qualités. Si la démarche semble donc validée, il n'en reste pas moins que la qualité des systèmes est une piste d'amélioration déterminante pour fournir une assistance efficace et aboutir à des outils qui soient présentés à des clients de Kairntech.

Dans la suite du projet Vital, nous allons étudier la consolidation des données produites, avec des annotations disponibles ouvrant la voie à des assistances supervisées ou semi-supervisée. Nous étudierons aussi la clôture d'une campagne d'annotation, qui consiste à finaliser le jeu de données en revisitant ses points faibles et en déterminant le point d'arrêt optimal.

Références

ABBAS, M., & SMAILI, K. (2005). Comparison of topic identification methods for arabic language. In Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP (pp. 14-17).

ALEX, B., NISSIM, M., & GROVER, C. (2006, May). The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text. In *LREC* (pp. 595-600).

ARTSTEIN, R., & POESIO, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, *34*(4), 555-596. DOI : 10.1162/coli.07-034-R2

BAYERL, P. S., & PAUL, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699-725. DOI : 10.1162/COLI_a_00074

EGGER, R., & YU, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in sociology, 7, 886498. DOI : 10.3389/fsoc.2022.886498

FAZAKIS, N., KANAS, V. G., ARIDAS, C. K., KARLOS, S., & KOTSIANTIS, S. (2019). Combination of active learning and semi-supervised learning under a self-training scheme. *Entropy*, 21(10), 988. DOI : 10.3390/e21100988

FORT, K. (2012). Les ressources annotées, un enjeu pour l'analyse de contenu: vers une méthodologie de l'annotation manuelle de corpus (Doctoral dissertation, Université Paris-Nord-Paris XIII). HAL : hal-00797760.

GROOTENDORST, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. DOI : 10.48550/arXiv.2203.05794

KRIPPENDORFF, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and quantity*, *38*, 787-800. DOI : 10.1007/s11135-004-8107-7

MATHET, Y. & WIDLÖCHER, A. (2016). Évaluation des annotations : ses principes et ses pièges. *Revue TAL*, 57-2, 73-98. HAL : hal-01712282

POURPANAH, F., ABDAR, M., LUO, Y., *et al* (2022). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI : 10.1109/TPAMI.2022.3191696

RIFKIN, R. (2008). Multiclass classification. *Lecture Notes, Spring08. MIT, USA*, 59.

SCHNOBER, C., & GUREVYCH, I. (2015). Combining topic models for corpus exploration: applying LDA for complex corpus research tasks in a digital humanities project. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications* (pp. 11-20). DOI : 10.1145/2809936.2809939

WINATA, G. I., MADOTTO, A., LIN, Z., LIU, R., YOSINSKI, J., & FUNG, P. (2021). Language models are few-shot multilingual learners. DOI : 10.48550/arXiv.2109.07684

WU, S., & DREDZE, M. (2020). Are all languages created equal in multilingual BERT? DOI : 10.48550/arXiv.2005.09093