Extraction de connaissances à partir de données textuelles : application à la découverte de règles de changement d'usage des sols

Sarah Valentin*†, Arthur Crespin-Boucaud**†, Grisel Quispe Aramayo*,**, Christophe Peugeot**, Maguelonne Teisseire*

*TETIS (Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE).
Maison de la Télédétection 500, rue J.F. Breton 34090 Montpellier
sarah.valentin@cirad.frd
**HSM, IRD, Univ. Montpellier, CNRS, Montpellier, France.
arthur.crespin-boucaud@ird.fr
†Les auteurs ont contribué de manière équivalente à ce travail.

Résumé. Dans le cadre de la modélisation de dynamiques spatiales, les connaissances expertes permettent de définir des règles d'état ou de changement d'état. Ces connaissances contribuent à la cohérence des modèles mais sont fastidieuses à obtenir. Ce travail propose une approche d'extraction automatique de connaissances à partir de données textuelles, appliquée aux dynamiques de changement d'occupation et d'usage des sols en Afrique de l'Ouest. L'objectif principal est d'identifier des segments de texte pertinents à partir d'un corpus d'articles scientifiques en anglais, en mobilisant des méthodes de traitement automatique de la langue et d'apprentissage automatique. Nous comparons deux approches de classification supervisée avec une approche non supervisée reposant sur la proximité sémantique et montrons leur capacité à identifier l'information pertinente à partir de données non structurées.

1 Introduction

Les projections sur l'épuisement des ressources et les points de bascule (Lenton et al., 2008) soulignent l'importance d'utiliser des données scientifiques pour explorer les futurs possibles. Les approches combinant méthodes qualitatives et quantitatives (Hérivaux et al., 2021) requièrent des données plus diversifiées et complexes, couvrant des processus multiscalaires et multitemporels, pour aboutir à des modélisations précises (Elsawah et al., 2020). Dans le cadre de la modélisation des changements d'occupation des sols, ces connaissances incluent l'état actuel d'occupation, les transitions entre états et les influences externes (comme la déforestation ou l'étalement urbain). Ces connaissances, qualitatives ou quantitatives, sont ensuite transformées en règles et intégrées dans des modèles grâce à des méthodes spécifiques pour passer du qualitatif au quantitatif (Mallampalli et al., 2016). La sollicitation d'experts pour identifier ces connaissances est coûteuse en temps et souffre de biais (Mora et al., 2020). L'analyse de la littérature scientifique est une approche complémentaire, mais l'explosion du

volume de publications rend l'accès aux informations pertinentes fastidieux. L'extraction de connaissances vise à pallier ce problème en identifiant et en synthétisant automatiquement les informations clés présentes dans les documents à l'aide de méthodes de traitement automatique de la langage (TALN), dans des domaines variés tels que l'écologie (Nundloll et al., 2022) ou l'aménagement du territoire (Koptelov et al., 2023). L'extraction de connaissances à partir de documents scientifiques sur une thématique précise nécessite une expertise spécifique au domaine afin de garantir la précision dans la construction et la validation des données extraites. Si le domaine de l'agriculture s'est doté de nombreuses ressources sémantiques (Drury et Roche, 2019; FAO, 2024), peu de travaux ont exploré le traitement de données textuelles sur l'usage et l'occupation des sols, leurs dynamiques, ainsi que les facteurs qui les influencent.

Ce travail a pour objectif d'évaluer une approche automatique pour l'extraction de connaissances sur l'occupation et l'usage des sols en Afrique de l'Ouest à partir de sources textuelles, grâce à l'utilisation de techniques avancées de traitement automatique de la langue et apprentissage automatique. L'extraction automatique d'informations à partir de longs documents peut entraîner une perte de performance et générer du bruit, d'où l'importance d'identifier les segments de texte pertinents pour permettre des tâches d'extraction sémantique plus avancées (Hershowitz et al., 2024). Nous proposons une approche incluant : (1) la formalisation de la notion d'information pertinente sur la thématique de LULC (Lande Use and Land Cover), (2) la constitution d'un corpus annoté, (3) l'évaluation de méthodes pour l'identification automatique de segments pertinents (i.e. porteurs d'informations pertinentes).

2 Données et méthodes

La figure 1 illustre la chaîne de traitement mise en place, allant de la collecte d'articles jusqu'à l'identification automatique de segments pertinents traitant de l'occupation des sols.

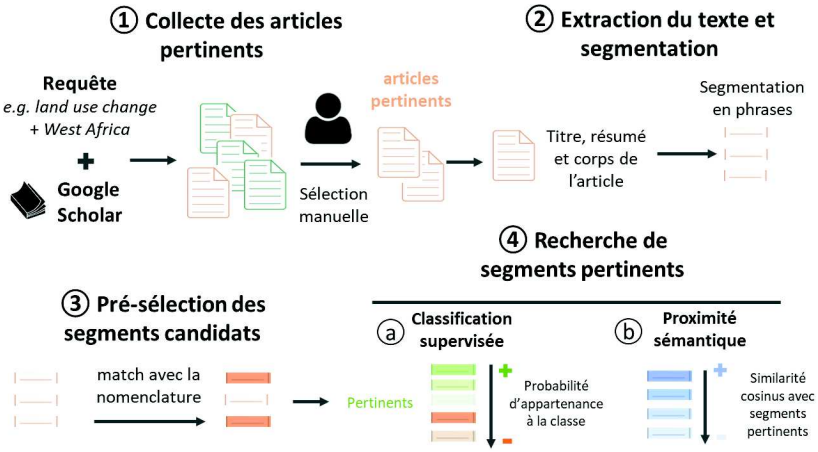

FIG. 1 – *Étapes de la chaîne de traitement.*

2.1 Corpus et nomenclature

Le corpus, en anglais, a été constitué à partir de requêtes dans la base de données de Google Scholar. Les requêtes ont combiné l'un des mots clés suivants : "land use change", "land cover change", "LULCC", "reforestation", "deforestation", "urbanization", "savannization", "agriculture change" avec la localisation "West Africa". Les résultats des dix premières pages ont été parcourus par un expert de la thématique afin de sélectionner les articles pertinents à partir de leur résumé. Le critère principal de sélection étant la cohérence thématique entre l'objectif de l'article et celui de notre étude (adéquation des termes, des orientations, méthodes et sources). Un total de 75 documents ont été identifiés à l'issue de ce processus. Le texte de chaque document a été extrait du fichier PDF à l'aide de l'outil GROBID (Lopez, 2009), en ne conservant que le titre, le résumé et le corps de l'article. Le corpus a été ensuite segmenté en phrases (segments) en utilisant la séquence "point suivi d'un espace" comme délimiteur.

Nous avons ensuite construit une nomenclature spécifique à notre thématique d'intérêt, en nous basant sur la nomenclature JECAM (JECAM, 2024) [1] qui regroupe des occupations et des usages des sols (e.g. céréales, forêt, jachère). Nous l'avons enrichie avec des termes désignant les processus de changements (e.g. savanisation, désertification). Pour étendre la nomenclature, nous avons recherché des termes apparentés et les synonymes dans AGROVOC (FAO, 2024). La nomenclature finale comprend 190 termes, i.e. 171 termes désignant des occupations et usages des sols et 19 termes désignant des processus de changement. Cette nomenclature a été utilisée pour pré-sélectionner les segments candidats pour la suite du processus de recherche : un segment est considéré comme candidat s'il contient au moins un terme de la nomenclature. L'ensemble des segments candidats est de 7 403 segments sur les 20 345 segments totaux, soit 36% de l'ensemble du corpus.

2.2 Annotation du jeu de données

Une *connaissance* correspond à la description d'un état d'occupation ou d'utilisation des sols (LULC), d'un changement de LULC, de pratique pouvant entraîner des changements ou des évolutions de LULC ou de ses dynamiques. La définition précise de la notion de pertinence a été élaborée de manière itérative au cours de l'annotation du corpus, ce qui nous a permis de distinguer quatre thèmes d'intérêt à rechercher dans le corpus :

— **Occupation et usage des sols (LULC)** : l'occupation des sols désigne un ensemble d'éléments biophysiques (e.g. herbe, béton, eau) et l'usage des sols représente la manière dont sont utilisées les surfaces (e.g. bâtiment industriel, culture irriguée).
— **Changements d'occupation et d'usage des sols (LULCC : LULC change)** : ces changements désignent la transition d'un type d'occupation ou d'usage des sols vers un autre, ou un changement d'intensification dans l'usage.
— **Pratiques** : les pratiques incluent les choix faits par une personne, un groupe de personnes ou une institution pour appliquer des changements susceptibles de modifier la structure du paysage, l'occupation ou l'usage des sols.
— **Forçages** : les forçages désignent tout facteur anthropique ou biologique (ou un mélange des deux) qui conduit à un changement d'occupation ou d'usage des sols. Ils peuvent être directs ou indirects et peuvent entraîner des pratiques spécifiques.

1. http://jecam.org/wp-content/uploads/2018/10/JECAM_Guidelines_for_Field_Data_Collection_v1_0.pdf

Nous avons défini trois classes de pertinence. Les segments "très pertinents" contiennent un ou plusieurs thèmes d'intérêt tels que définis ci-dessus ainsi qu'au moins une information spatio-temporelle et/ou quantitative permettant de les contextualiser. Les segments "pertinents" contiennent un ou plusieurs thèmes d'intérêt mais peu ou pas d'information contextuelle. Les autres segments sont considérés comme non pertinents. Pour guider l'annotation, nous avons élaboré une grille de critères détaillés spécifiques à chaque classe. À partir de l'ensemble des segments candidats, 330 segments ont été sélectionnés aléatoirement puis annotés selon leur pertinence. Le groupe d'annotateurs est constitué d'un géographe expert en changement d'occupation et d'usage des sols (A3), d'une informaticienne experte en sécurité alimentaire (A2) et d'une informaticienne novice dans la thématique (A1). La labélisation de 161 segments par trois annotateurs (A1, A2 et A3) de manière indépendante s'est déroulée en trois étapes (50, 50, puis 61 segments) afin de mettre en commun leurs annotations, de trouver un accord sur le label final et d'affiner les définitions de chaque classe (Tableau 1). La principale difficulté résidait dans la distinction des segments pertinents et très pertinents, la quantification du niveau d'information reposant en partie sur une appréciation subjective.

	Classes pertinentes séparées		Classes pertinentes fusionnées	
	Kappa de Fleiss	% d'accord	Kappa de Fleiss	% d'accord
Ensemble (A1, A2, A3)	0.40	47.8	0.47	60.2
	Kappa	% d'accord	Kappa	% d'accord
Ensemble (A2, A3)	0.71	87.5	0.82	92.9

TAB. 1 – *Évaluation de l'accord global entre les annotateurs.*

Le reste de l'échantillon (169 segments) a ensuite été annoté de manière indépendante par A2 et A3, obtenant des Kappa de 0.71 (classes pertinentes séparées) et 0.82 (classes pertinentes fusionnées). Après évaluation de l'accord global, les désaccords ont été discutés afin de décider d'un label final pour chaque segment. Les 330 segments annotés sont utilisés pour les tâches d'apprentissage et d'évaluation.

2.3 Recherche de segments pertinents

Pour prendre en compte la composante subjective de la notion de pertinence, nous avons converti l'identification de segments pertinents en une tâche de recherche d'information. Autrement dit, nous cherchons à trier l'ensemble des segments par ordre décroissant de pertinence à partir d'un corpus de documents. Pour cela, nous comparons deux méthodes : une classification (supervisée) et la proximité sémantique (non-supervisée).

2.3.1 Classification supervisée

Dans le cadre supervisé, nous avons comparé deux approches classiquement utilisées pour la classification de texte : une représentation textuelle en sac-de-mots fréquentiste combinée à un algorithme de classification de type support vector machine (SVM) (Joachims, 1998), et

une représentation par un modèle de langue pré-entraîné, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Pour la première approche, nous avons utilisé la métrique de pondération TF-IDF (Term Frequency - Inverse Document Frequency) (Salton et Buckley, 1988) sur les 330 segments annotés, en excluant les mots vides (stopwords). Nous avons utilisé l'implémentation du SVM linéaire LinearSVC de la bibliothèque *scikit-learn*. Afin de prendre en compte le déséquilibre des classes dans les données d'apprentissage, nous avons ajusté les poids des classes de manière inversement proportionnelle à leur fréquence grâce au paramètre *class weight "balanced"*. La robustesse et la généralisation du modèle ont été évaluées en calculant les performances avec une validation croisée à 5 plis. Pour la deuxième approche, nous avons utilisé un des variants du modèle BERT, RoBERTa (Liu et al., 2019). Par défaut, RoBERTa utilise l'optimiseur AdamW, reconnu pour sa robustesse dans l'entraînement des modèles d'apprentissage profond. Nous avons chargé puis adapté le modèle pré-entraîné [2] sur une tâche de classification en utilisant les hyperparamètres suivants : longueur maximale de la séquence de 128 tokens, nombre d'*epochs* de 5, taille de *batch* de 8, et un ratio train-test de 80 :20. Le meilleur modèle, sur la base de l'exactitude globale (accuracy), a été conservé pour la suite des analyses.

Pour trier les segments par ordre décroissant de pertinence grâce à l'approche supervisée, nous avons utilisé le modèle entraîné et avons extrait la classe prédite ainsi que les probabilités d'appartenance aux classes. Les segments ont été triés en fonction de leur probabilité maximale d'appartenance, utilisée pour déterminer leur classe.

2.3.2 Proximité sémantique

Dans cette deuxième approche, les segments sont encodés à l'aide de deux modèles pré-entraînés : le modèle RoBERTa et un modèle d'encodage de phrases plus léger, MiniLM (Minimal Language Model), dans sa version all-MiniLM-L6-v2 [3]. Pour chaque segment non annoté, nous calculons sa proximité sémantique avec chacun des segments pertinents grâce à la similarité cosinus. La valeur de proximité finale correspond à la moyenne des similarités obtenues.

2.3.3 Évaluation

Les performances des approches de classification supervisée sur les données d'entraînement ont été calculées à l'aide des métriques classiques :

— la précision, qui est l'évaluation du taux de vrais positifs dans la classe prédite,
— le rappel, permet de calculer le taux d'éléments positifs correctement prédits par classe,
— le F-score, qui est la moyenne harmonique de la précision et du rappel.

Pour prendre en compte le déséquilibre entre les classes, les valeurs globales de ces métriques ont été calculées via une moyenne pondérée par la fréquence de chaque classe.

Pour l'évaluation de la recherche de segments pertinents sur l'ensemble du corpus, nous avons sélectionné aléatoirement 200 segments non annotés du corpus. Pour chacune des méthodes, nous avons manuellement évalué la pertinence des 50 premiers segments. Nous avons évalué la qualité du classement grâce à la precision@k, qui évalue la proportion d'éléments pertinents dans le top k des éléments triés. Nous avons évalué p@10 (top 10) et p@50 (top

2. https://huggingface.co/docs/transformers/model_doc/roberta
3. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

50), en considérant seulement les segments très pertinents (évaluation stricte) ou en prenant également en compte les segments pertinents.

Le code pour reproduire les résultats et le contenu de la nomenclature sont disponibles dans un dépôt git[4]. Le corpus est disponible après requête auprès des auteurs.

3 Résultats

Les segments annotés proviennent de 20 articles avec un maximum de 46 segments pour un article, un minimum de 3 et une moyenne de 16,2. Le corpus final contient 330 segments annotés, dont 218 (66,1%) sont non pertinents, 71 (21,5%) sont pertinents et 41 (12,4%) sont très pertinents. Les segments contiennent entre 6 et 645 tokens (correspondant à l'extraction des tableaux), avec une moyenne de 47 tokens par segment.

Le tableau 2 présente les résultats de classification obtenus avec les modèles TF-IDF/SVM et RoBERTa. L'approche TF-IDF/SVM obtient un F-score moyen de 0,64, et souffre d'un manque de précision pour les segments non pertinents (0,72) et un rappel des segments pertinents très faible (0,24). Le modèle RoBERTa obtient de meilleures performances sur l'ensemble des métriques, et notamment un rappel pour les segments très pertinents de 0,71 et une amélioration de la précision pour les segments non pertinents (0,84). La précision des segments non pertinents et les rappels des segments pertinents et très pertinents sont les métriques particulièrement importantes à optimiser, car l'objectif est de maximiser la détection d'information pertinente. Dans les deux approches, la majorité des confusions concerne des erreurs de classification entre segments pertinents et très pertinents, ce qui est acceptable. Le modèle RoBERTa semble avoir mieux capturé la sémantique des segments que la représentation TF-IDF, mais les résultats obtenus restent améliorables, notamment par une évaluation des hyperparamètres qui influencent les performances du modèle.

	SVM			RoBERTa		
	Précision	Rappel	F-score	Précision	Rappel	F-score
Non pertinents	0.72	**0.96**	0.83	**0.84**	0.90	0.87
Pertinents	0.52	0.20	0.29	0.62	**0.47**	0.53
Très pertinents	0.77	0.24	0.37	0.67	**0.71**	0.78
Moyenne	0.69	0.71	0.65	**0.77**	**0.78**	**0.77**

TAB. 2 – *Comparaison des résultats de classification supervisée.*

La recherche de segments pertinents avec le modèle issu de la classification RoBERTa est systématiquement plus performante que l'approche par proximité sémantique (Tableau 3). L'apprentissage supervisé permet au modèle de capturer des patterns spécifiques aux données d'entraînement et à la tâche de classification. A contrario, la proximité sémantique se base sur les représentations génériques des modèles de langue et se révèle moins performante dans ce cas d'étude. Le modèle MiniLM, spécifiquement adapté à la recherche sémantique et au calcul de similarité entre phrases, obtient des précisions légèrement supérieures à celles de RoBERTa.

4. https://github.com/SarahVal/KnowledgeExtraction_LULCC

	Classification (RoBERTa)	Proximité (MiniLM)	Proximité (RoBERTa)
p@10 (très pertinents)	0.70	0.60	0.20
p@10 (tous les pertinents)	0.90	0.70	0.70
p@50 (très pertinents)	0.26	0.24	0.14
p@50 (tous les pertinents)	0.78	0.58	0.52

TAB. 3 – *Comparaison des précisions@k (p@k), où k représente les k premiers résultats, des différentes méthodes de recherche de segments pertinents.*

4 Conclusion

Cet article constitue la première étape de travaux originaux sur l'extraction automatique de connaissances appliquée à la thématique de l'occupation et de l'usage des sols. A partir d'un faible nombre de données d'apprentissage, nous avons obtenu des résultats prometteurs pour la détection de segments pertinents. Ces résultats sont encourageants pour aboutir à la formulation de règles logiques qui pourront être intégrées dans des modélisations spatio-temporelles d'occupation des sols. De nombreuses perspectives s'ouvrent suite à ces premiers résultats. L'identification des thèmes associés aux segments pertinents nous semble une étape importante pour la représentation et l'utilisation des connaissances issues de ce segment. Il serait intéressant de développer une méthode de labellisation à une échelle plus large, englobant des groupes de phrases ou des paragraphes. Enfin, l'extraction d'entités nommées thématiques et spatio-temporelles pourra être combinée à la recherche de segments pertinents pour en améliorer la détection.

Références

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv :1810.04805 [cs]*.

Drury, B. et M. Roche (2019). A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture 163*, 104864, doi: 10.1016/j.compag.2019.104864.

Elsawah, S., S. H. Hamilton, A. J. Jakeman, D. Rothman, V. Schweizer, E. Trutnevyte, H. Carlsen, C. Drakes, B. Frame, B. Fu, C. Guivarch, M. Haasnoot, E. Kemp-Benedict, K. Kok, H. Kosow, M. Ryan, et H. van Delden (2020). Scenario processes for socio-environmental systems analysis of futures : A review of recent efforts and a salient research agenda for supporting decision making. *Science of The Total Environment 729*, 138393, doi: 10.1016/j.scitotenv.2020.138393.

FAO (2024). AGROVOC Multilingual Thesaurus.

Hershowitz, B., M. Hodkiewicz, T. Bikaun, M. Stewart, et W. Liu (2024). Causal knowledge extraction from long text maintenance documents. *Computers in Industry 161*, 104110, doi: 10.1016/j.compind.2024.104110.

Hérivaux, C., F. Vinatier, M. Sabir, F. Guillot, et J. D. Rinaudo (2021). Combining narrative scenarios, local knowledge and land-use change modelling for integrating soil erosion in a global perspective. *Land Use Policy 105*, 105406, doi: 10.1016/j.landusepol.2021.105406.

JECAM (2024). Jecam - joint experiment for crop assessment and monitoring. Accessed : 2024-08-20.

Joachims, T. (1998). Text categorization with Support Vector Machines : Learning with many relevant features. In *Machine Learning : ECML-98*, Volume 1398, pp. 137–142. Springer, doi: 10.1007/BFb0026683.

Koptelov, M., M. Holveck, B. Cremilleux, J. Reynaud, M. Roche, et M. Teisseire (2023). A manually annotated corpus in French for the study of urbanization and the natural risk prevention. *Scientific Data 10*(1), 818, doi: 10.1038/s41597-023-02705-y.

Lenton, T. M., H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, et H. J. Schellnhuber (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences 105*(6), 1786–1793, doi: 10.1073/pnas.0705414105.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach.

Lopez, P. (2009). Grobid : Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *13th European Conference in Digital Libraries, ECDL 2009, Corfu, Greece, Sept 27-Oct 2, 2009*, pp. 473–474.

Mallampalli, V. R., G. Mavrommati, J. Thompson, M. Duveneck, S. Meyer, A. Ligmann-Zielinska, C. G. Druschke, K. Hychka, M. A. Kenney, K. Kok, et M. E. Borsuk (2016). Methods for translating narrative scenarios into quantitative assessments of land use change. *Environmental Modelling & Software 82*, 7–20, doi: 10.1016/j.envsoft.2016.04.011.

Mora, O., C. L. Mouël, M. d. Lattre-Gasquet, C. Donnars, P. Dumas, O. Réchauchère, T. Brunelle, S. Manceron, E. Marajo-Petitzon, C. Moreau, M. Barzman, A. Forslund, et P. Marty (2020). Exploring the future of land use and food security : A new set of global scenarios. *PLOS ONE 15*(7), e0235597, doi: 10.1371/journal.pone.0235597.

Nundloll, V., R. Smail, C. Stevens, et G. Blair (2022). Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. *Heliyon 8*(10), e10710, doi: 10.1016/j.heliyon.2022.e10710.

Salton, G. et C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24*(5), 513–523, doi: 10.1016/0306-4573(88)90021-0.

Summary

This work presents an approach to automatically extract expert knowledge from textual data, focusing on land use change dynamics in West Africa. The main objective is to identify relevant text segments from a corpus of scientific articles in English, using automatic language processing and machine learning methods. We compare two supervised classification approaches with an unsupervised approach based on semantic proximity and show their ability to identify relevant information from unstructured data.