Normalisation automatique de variables issues de bases de données en agroécologie

Oussama Mechhour, Sandrine Auzoux, Clément Jonquet, Mathieu Roche

Normalisation automatique de variables issues de bases de données en agroécologie

Oussama Mechhour[*,**,***], Sandrine Auzoux[*,**],
Clément Jonquet[****], Mathieu Roche[*,***]

[*]CIRAD, UPR AIDA & UMR TETIS, F-34398 Montpellier, France
prenom.nom@cirad.fr
[**]AIDA, Univ Montpellier, CIRAD, La Réunion, France
[***]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[****]MISTEA, Univ Montpellier, INRAE, Institut Agro, Montpellier, France
clement.jonquet@inrae.fr

Résumé. L'objectif de ces travaux est de proposer et évaluer des méthodes de mise en correspondance entre des variables sources [1] et des variables candidates du domaine de l'agroécologie [2] (en anglais). Le but de notre démarche est d'aider l'expert à normaliser les bases de données, ainsi qu'à lier les variables sources aux variables candidates des dictionnaires des modèles utilisés en agroécologie (AEGIS [4]).

1 Introduction

La mesure de similarité entre variables constitue un problème reconnu, en particulier dans notre cas d'étude impliquant des données textuelles. Cette complexité émerge de la diversité des langues, engendrant des problèmes tels que la synonymie et la prise en compte du contexte.

Dans cet article, nous nous concentrons sur la mesure de similarité des variables utilisées dans la culture de la canne à sucre, un domaine très spécifique et peu étudié en fouille de textes. L'objectif principal de ce travail est double : résoudre la problématique de l'hétérogénéité des variables utilisées par les chercheurs, appelée variables sources (chaque chercheur ayant sa propre méthode de nomination et de description de ces variables), grâce à l'aide d'experts pour normaliser les bases de données et lier ces variables aux variables candidates (dictionnaires des modèles utilisés en agroécologie, AEGIS). D'autre part, nous visons le développement d'une interface web permettant l'application des mesures décrites dans l'article (cf. Section 3) et d'autres fonctionnalités (cf. Section 3.4). Des efforts ont été déployés pour établir une similarité entre les variables sources et candidates stockées dans le système d'information AEGIS.

1. Ce sont des variables issues des travaux de recherche dans le domaine de l'agroécologie, et pour cet article, le travail s'est focalisé sur des données spécifiques liées à la culture de la canne à sucre.

2. Ces variables, composées de termes sémantiques issus de connaissances expertes et d'ontologies de référence, ont été spécifiquement définies dans le but de faciliter la comparaison et l'analyse des données, ainsi que d'établir des liens avec des modèles de culture tels que Modèle STICS. [3]

4. AEGIS, développé par le **CIRAD** (Auzoux, 2019), est une plateforme en ligne qui permet de stocker et exploiter des données provenant d'expérimentations en agroécologie menées dans les pays du Sud.

TAB. 1 présente des exemples des noms de ces deux types de variables, sachant que chaque nom d'une variable source ou candidate se compose de deux parties : son nom avant le dernier _ et son unité de mesure après le dernier _. Nos travaux vont consister à mettre en place des mesures de similarité pour la mise en lien des variables d'agroécologie. Pour illustrer l'importance de l'intégration de mesure de similarité des variables, prenons l'exemple du *products matching* (Tracz et al., 2020), qui permet de recommander automatiquement des produits similaires aux préférences des clients de manière efficace en termes de temps et de coût.

Cet article est structuré de la manière suivante. La section 2 présente une revue de littérature des différentes approches appliquées pour résoudre la problématique de la similarité entre chaînes de caractères, une question qui a été étudiée dans plusieurs travaux de recherche, y compris dans le domaine de l'agroécologie. La section 3 décrit les approches proposées (lexicale, contextuelle et combinaison) pour mesurer la similarité entre les variables et le développement d'une interface web pratique pour appliquer ces approches. La section 4 présente les méthodes de prétraitement des données appliquées, les résultats précédents et actuels, ainsi que la discussion montrant les meilleurs résultats pour chaque mesure (lexicale et contextuelle), ainsi que les améliorations des résultats. La section 5 conclut cet article et met en avant les perspectives de notre étude.

2 État de l'art

Dans cette revue de littérature, nous abordons la problématique de la similarité entre chaînes de caractères, qui a été étudiée dans plusieurs travaux de recherche. Ces travaux se divisent généralement en deux approches distinctes : (1) approche ne prenant pas en compte de contexte (cf. Section 2.1), approche prenant en compte un contexte (cf. Section 2.2).

2.1 Approche ne prenant pas en compte de contexte

Dans cette approche, les travaux se concentrent principalement sur des mesures de similarité générales qui ne prennent pas en compte le contexte spécifique de l'agroécologie. Ils utilisent des méthodes telles que TF-IDF (Term Frequency, Inverse Document Frequency) (Jones, 1972), BM-25 (Okapi BM-25) (Robertson et al., 1994), la distance de Levenshtein (Levenshtein, 1966), la distance de Jaccard (Jaccard, 1901) et des modèles de descriptions des variables comme I-ADOPT framework [5] pour évaluer la similarité entre les variables. La distance de Levenshtein permet de calculer le nombre de changements nécessaires entre deux chaînes de caractères. Pour mieux appréhender le fonctionnement de la distance de Levenshtein, examinons l'exemple suivant : considérons les mots *agro-écologie* et *agroécologie*. La distance de Levenshtein entre ces deux mots est de 1, car il suffit de supprimer le tiret pour les transformer l'un en l'autre.

La mesure TF-IDF (Jones, 1972) est une méthode classique pour évaluer l'importance des termes dans un document par rapport à une collection de documents. Elle représente la fréquence du terme dans le document. Elle est calculée en divisant le nombre d'occurrences du terme par le nombre total de termes dans le document. La formule mathématique de TF pour un terme t dans un document d est donnée par :

5. https ://www.rd-alliance.org/group/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/wiki/i-adopt

$$\text{TF}(t, d) = \frac{\text{nombre d'occurrences de } t \text{ dans } d}{\text{nombre total de termes dans } d}$$

L'IDF mesure l'importance globale d'un terme dans la collection de documents. Elle est calculée en prenant le logarithme inverse de la proportion du nombre total de documents sur le nombre de documents contenant le terme. La formule mathématique de IDF pour un terme t dans une collection de documents est donnée par :

$$\text{IDF}(t) = \log\left(\frac{\text{nombre total de documents}}{\text{nombre de documents contenant } t}\right)$$

La mesure TF-IDF est obtenue en multipliant la valeur de TF par la valeur de IDF pour chaque terme. Ainsi, les termes qui sont fréquents dans un document particulier tout en étant rares dans l'ensemble de la collection auront une valeur TF-IDF plus élevée. Cependant, notons que ces approches peuvent présenter certaines limites pour capturer la similarité sémantique et de prendre en compte le contexte spécifique de l'agroécologie. Cette approche a été abordée dans un travail précédent (Ngaba, 2022), où la distance de Levenshtein a été utilisée comme mesure de similarité entre les noms des deux types de variables. De plus, le TF-IDF et la mesure cosinus (Manning et al., 2008) ont été utilisés respectivement pour la vectorisation des descriptions et la mesure de similarité entre ces vecteurs. Ces deux techniques ont ensuite été combinées dans une méthode appelée combinaison (cf. Section 3.3). Cependant, il est important de noter que l'utilisation du TF-IDF et Levenshtein dans cette approche ne tient pas compte du contexte spécifique de l'agroécologie, ce qui peut limiter les résultats obtenus.

Le framework I-ADOPT (InteroperAble Descriptions of Observable Property Terminology), faisant partie des modèles de description des variables scientifiques, de la RDA [6] vise à aborder le volet 'I' (interoperability) des principes FAIR [7]. Une variable est la combinaison de tous les composants descriptifs considérés comme nécessaires pour comprendre ce qui a réellement été observé, mesuré, simulé ou calculé. Elle décrit ce qui a été observé, mesuré, simulé ou calculé, indépendamment de L'endroit, de la manière, et du moment où l'acquisition des données a eu lieu. Par exemple, *Concentration of endosulfan sulfate in the flesh of Ostrea edulis expressed per unit wet weight*. La variable la plus simple requiert au moins un ObjectOfInterest et une Property. Cependant, une variable plus complexe impliquerait davantage d'entités ayant le rôle de Matrix et/ou ContextObject(s). Property représente la qualité d'un ObjectOfInterest, agissant comme un élément dépendant et une caractéristique de ObjectOfInterest. Les propriétés peuvent être exprimées en tant que types de quantité (QuantityTypes) ou types de qualité (QualityTypes) lorsqu'elles sont observées, mesurées, calculées ou simulées, par exemple : hauteur, couleur, vitesse, concentration, densité. ObjectOfInterest désigne l'entité qui joue le rôle de la cible d'observation pour laquelle la propriété est observée. Par exemple, dans *Concentration of endosulfan sulfate in the flesh of Ostrea edulis expressed per unit wet weight* 'Endosulfan sulfate' est l'ObjectOfInterest. ContextObject est l'entité qui joue un rôle descriptif pour fournir le contexte de ObjectOfInterest. Par exemple, dans *Concentration of endosulfan sulfate in the flesh of Ostrea edulis expressed per unit wet weight* 'Ostrea edulis' est le ContextObject. Matrix désigne l'entité qui joue le rôle du contexte matériel de

6. https ://www.rd-alliance.org/

7. Les principes FAIR (Findable, Accessible, Interoperable, Reusable) décrivent comment les données doivent être organisées pour être plus facilement accessibles, comprises, échangeables et réutilisables.

ObjectOfInterest. Par exemple, dans *Concentration of endosulfan sulfate in the flesh of Ostrea edulis expressed per unit wet weight* 'Flesh' est la Matrix.

À notre connaissance, aucune recherche dans le domaine de l'agroécologie n'a encore mis en oeuvre la combinaison des approches ne prenant pas en compte de contexte et de celles prenant en compte un contexte (cf. Section 2.2). Par conséquent, notre travail vise à combler cette lacune en explorant la faisabilité et l'efficacité de cette méthode dans le contexte spécifique de l'agroécologie.

2.2 Approche prenant en compte un contexte

Cette approche utilise des techniques qui prennent en considération un contexte, comme FastText (Bojanowski et al., 2016), Word2Vec (Mikolov et al., 2013) et les modèles de langues (Jurafsky et Martin, 2019). Ces derniers jouent un rôle essentiel dans le domaine du traitement automatique du langage naturel. Ils visent à capturer les structures et les relations linguistiques dans un corpus de texte pour générer des prédictions précises. Dans cette sous-section, nous aborderons les différents types de modèles de langues, en commençant par les n-grammes (Jurafsky et Martin, 2019), puis en discutant des modèles basés sur les RNNs (Recurrent Neural Networks) (Yin et al., 2017) et les LSTM (Long Short-Term Memory) (Hochreiter et Schmidhuber, 1997), pour finalement présenter les modèles de langues basés sur les transformers, tels que BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018).

n-grammes : les modèles de langues (Jurafsky et Martin, 2019) basés sur les n-grammes sont parmi les plus simples. Ils reposent sur l'idée d'exploiter les fréquences d'apparition des n-grammes (séquences de mots) dans un corpus pour prédire le mot suivant. Par exemple, un modèle de langue basé sur les trigrammes considère les deux mots précédents pour prédire le prochain mot. Bien que les n-grammes soient faciles à mettre en oeuvre et puissent fournir des résultats acceptables pour des tâches de langage simples, ils présentent des limites majeures. L'une des principales faiblesses des n-grammes est leur incapacité à capturer les dépendances à long terme entre les mots, ce qui limite leur capacité à générer des séquences de mots cohérentes.

RNNs et LSTM : pour remédier aux limites des modèles de langues basés sur les n-grammes, les RNNs ont été introduits. Les RNNs sont conçus pour traiter des séquences de données, ce qui les rend bien adaptés pour modéliser les séquences de mots dans un texte. Cependant, les RNNs traditionnels souffrent du problème du *vanishing gradient* et ont du mal à capturer des dépendances à long terme. Les LSTM sont une extension des RNNs qui permettent de mieux gérer les dépendances à long terme. Grâce à l'utilisation de portes de mémoire, les LSTM peuvent décider de conserver, de modifier ou d'oublier certaines informations, ce qui améliore la capacité du modèle à capturer des dépendances à plus long terme. Cependant, même les LSTM ont leurs limites. Ils peuvent avoir du mal à capturer des relations complexes entre les mots et peuvent souffrir de problèmes de surapprentissage lorsqu'ils sont confrontés à de grands ensembles de données.

Modèles de langues fondés sur les transformers : les modèles de langues basés sur les transformers (Vaswani et al., 2017), tels que BERT (Devlin et al., 2018), ont révolutionné le domaine du TALN (Ranjan et al., 2016) ces dernières années. Les transformers exploitent une

architecture d'attention (Vaswani et al., 2017 ; Niu et al., 2021) pour capturer les relations entre tous les mots d'une séquence, à la fois dans le contexte avant et après. Cette approche bidirectionnelle permet au modèle de comprendre plus efficacement les relations sémantiques complexes dans le texte. BERT a été largement reconnu pour ses performances exceptionnelles dans de nombreuses tâches de TALN (Ranjan et al., 2016), y compris la compréhension de texte, la traduction et le résumé automatique. Son architecture transformer lui permet de capturer de manière plus précise les dépendances à long terme par rapport aux modèles précédents. De plus, BERT peut être pré-entraîné sur de vastes corpus non supervisés, ce qui lui permet d'acquérir une connaissance linguistique riche et de s'adapter à diverses tâches spécifiques. BERT (Devlin et al., 2018) est un modèle de langues pré-entraîné sur de vastes corpus, tels que Wikipedia, qui utilise le mécanisme d'attention (Vaswani et al., 2017 ; Niu et al., 2021) pour comprendre le contexte des mots. Il prend en compte à la fois le contexte à gauche et à droite de chaque mot, ce qui lui permet de capturer les informations contextuelles de manière précise. Cette capacité à saisir le contexte permet à BERT de générer des représentations vectorielles de mots riches en sémantique. Il convient de noter qu'il existe deux types de modèles BERT couramment utilisés : BERT-base (Devlin et al., 2018) et BERT-large (Devlin et al., 2018).

3 Approches proposées

Dans le cadre de la problématique abordée dans l'introduction, qui est très spécifique et peu étudiée en fouille de textes, en plus de la création d'une interface web, différentes méthodes de fouille de texte sont utilisées pour proposer les variables candidates les mieux adaptées aux variables sources. Les méthodes suivantes sont mobilisées : (1) Des mesures lexicales, (2) Des mesures contextuelles et (3) La combinaison de ces deux dernières.

Dans un premier temps, notre démarche consiste à évaluer la similarité entre chaque variable source (au nombre de 84) et l'ensemble des 84 variables candidates, en les classant de la plus similaire à la moins similaire, en se basant uniquement sur les noms et les descriptions des variables. Par la suite, nous avons enrichi notre méthode en incorporant 15 articles en anglais représentatifs du domaine de l'agroécologie, constitués manuellement avec l'aide de 3 experts. Chaque article a une moyenne de 79 399 mots, contribuant ainsi à un corpus final (composé de 15 articles, des noms et des descriptions des variables sources et candidates) totalisant 5 620 198 mots. FIG. 1 illustre la similarité entre les variables sources et candidates en utilisant 15 articles en anglais comme informations supplémentaires. Pour la similarité sans contexte (se basant uniquement sur les noms et les descriptions des variables), le principe reste le même, mais les articles sont exclus. Pour l'évaluation des résultats, nous utilisons un fichier .csv qui contient le nom de chaque variable source et son vrai nom de sa variable candidate. TAB. 1 montre les 4 premières lignes de ce fichier à titre d'exemple, mais en totalité, nous avons 84 variables sources et 84 variables candidates.

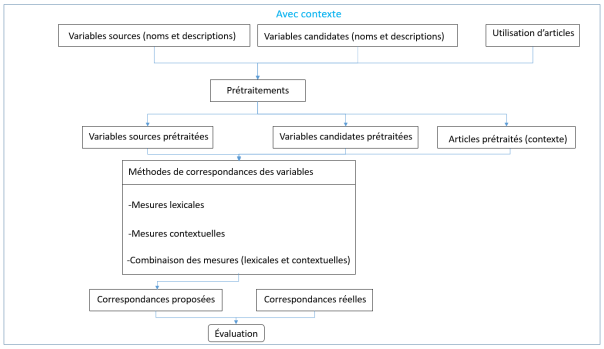

FIG. 1 – *Correspondance des variables avec enrichissement contextuel.*

Variable source	**Variable candidate**
Yield_CAS_t.ha-1	stem_crop_yield_fm_t.ha-1
Sugar_CAS_%	stem_sugar_fm_content_%
Rec_pds_R_%	plant_ground_cover_%
Rec_globale_plein_%	plant_ground_cover_%

TAB. 1 – *Exemples de noms de variables sources et candidates (en anglais).*

3.1 Mesure lexicale

Le but de l'approche lexicale est de comparer les variables à travers leurs noms. Dans le cadre de cette approche, la distance de Levenshtein (Levenshtein, 1966), la distance de Jaccard (Jaccard, 1901), TF-IDF (Jones, 1972), BM-25 (Robertson et al., 1994), FastText (Bojanowski et al., 2016), Word2Vec (Mikolov et al., 2013) ainsi que des modèles de langues tels que BERT-base (Devlin et al., 2018), BERT-large (Devlin et al., 2018), XLNet (Yang et al., 2019) et RoBERTa (Liu et al., 2019) ont été utilisés.

3.2 Mesure contextuelle

Le but de l'approche contextuelle est de comparer les variables à travers leurs descriptions. Dans cette approche, nous utilisons deux types de méthodes : sans contexte [8], comme TF-IDF (Jones, 1972) et BM-25, et avec contexte [9], en utilisant FastText (Bojanowski et al., 2016), Word2Vec (Mikolov et al., 2013) et des modèles de langues tels que BERT-base (Devlin et al., 2018), BERT-large (Devlin et al., 2018), XLNet (Yang et al., 2019) et RoBERTa (Liu et al., 2019).

8. Les méthodes qui ne prennent pas en considération un contexte se basent principalement sur des mesures de similarité ou de fréquence de termes.

9. Les méthodes qui prennent en compte de contexte utilisent souvent des modèles linguistiques avancés, tels que BERT.

3.3 Combinaison

La méthode que nous proposons permet de combiner les mesures lexicales et contextuelles afin d'améliorer la précision de la similarité des variables. Cette méthode est définie par la formule suivante :

$$\textbf{combinaison} = \alpha \cdot \mathrm{X} + (1 - \alpha) \cdot Y$$

où $\alpha \in]0, 1[$ est un facteur de pondération attribué à X et Y. Il est utilisé pour contrôler l'influence respective de X et Y (où X et Y représentent la mesure cosinus dans le travail actuel, mais dans le travail de (Ngaba, 2022), X représente le cosinus et Y représente la distance de Levenshtein) dans la combinaison finale. Après avoir effectué la vectorisation des noms et des descriptions des variables en utilisant les approches lexicales et contextuelles, nous appliquons la mesure cosinus pour évaluer la similarité entre les variables sources et les variables candidates. L'utilisation de ce paramètre (α) permet un réglage précis du degré d'importance attribué à X et Y dans la combinaison. Dans cette étude, nous avons exploré les valeurs de ce paramètre dans la plage de 0,1 à 0,9, avec un pas de 0,01.

3.4 Interface graphique

Au cours de ce travail, une interface web [10], a été développée pour fournir aux chercheurs un outil pratique. Elle prend en compte 4 fichiers texte : noms des variables sources, noms des variables candidates, descriptions des variables sources et descriptions des variables candidates. Elle leur permet d'appliquer les mesures de similarité entre les variables sources et les variables candidates expliquées dans la section 3.

4 Expérimentations

Après avoir expliqué les approches proposées, nous détaillons ici l'étape de préparation du jeu de données. Ensuite, nous montrerons nos résultats et les comparerons avec les résultats précédents (Ngaba, 2022).

4.1 Préparation du jeu de données

La préparation des données textuelles joue un rôle essentiel. Elle vise à transformer les noms et les descriptions des variables en une forme appropriée pour faciliter leur comparaison et leur similarité. Il existe plusieurs logiciels pour la préparation de données, tels qu'OpenRefine (Petrova-Antonova et Tancheva, 2020) et Trifacta (Petrova-Antonova et Tancheva, 2020). Cependant, pour notre travail, nous avons appliqué nos propres méthodes pour la préparation des variables (noms et descriptions). Voici l'ordre de leur application :

1. `clean_text()` : Cette fonction est utilisée pour nettoyer les variables sources et candidates. Elle élimine les nombres, les parenthèses [11] et leur contenu, et convertit les noms et les descriptions en minuscules. En normalisant les variables, cette étape facilite la comparaison et l'alignement ultérieur.

10. https ://drive.google.com/drive/folders/1NRUoG4LxAlXGnqMYsNz9QRQny1AVyWJl

11. La suppression des parenthèses et de leur contenu a amélioré les performances. Cependant, pour les travaux futurs, nous explorerons comment tirer parti de ces informations.

2. `remove_stopwords()` : Les stopwords sont des mots couramment utilisés dans la langue qui n'apportent pas de signification particulière dans le contexte de l'agroécologie. Cette fonction supprime ces mots fonctionnels, tels que les prépositions et les conjonctions. En éliminant les stopwords, nous nous concentrons sur les termes clés qui sont plus significatifs pour la mesure de similarité entre les variables.
3. `lemmatize()` : La lemmatisation est une technique linguistique qui consiste à ramener les termes à leur forme canonique ou à leur lemme. Elle permet de transformer les noms du pluriel au singulier et les verbes à leur forme infinitive. Par exemple, elle permet de ramener les termes *rédige*, *rédiges* et *rédigé* à leur forme de base *rédiger*. La lemmatisation facilite la mesure de similarité entre les termes similaires, améliorant ainsi la précision de l'alignement des variables.
4. `remove_punctuation()` : Cette fonction supprime la ponctuation des descriptions des variables. En éliminant les caractères spéciaux tels que les points, les virgules et les guillemets, nous évitons les interférences indésirables lors de la mesure de similarité entre les variables.
5. `replace_synonyms()` : Pour faciliter la mesure de similarité entre les variables, cette technique permet de remplacer certains mots par leurs synonymes. Par exemple, le mot *degré* peut être remplacé par *niveau*. En utilisant des termes équivalents, cette étape a amélioré la cohérence et la précision de l'alignement des variables.

Toutes les fonctions ont été appliquées à l'ensemble des descriptions des variables sources et candidates, tandis que seules les trois premières fonctions ont été utilisées pour les noms des variables. TAB. 2 présente trois exemples de noms de variables sources et leurs descriptions, tandis que TAB. 3 présente trois exemples de noms de variables candidates et leurs descriptions. Nous comptons un total de 84 variables sources et 84 variables candidates.

Nom	Description
Yield_CAS_t.ha-1	Cane yield (in fresh machinable stem)
Sugar_CAS_%	Sugar content of fresh stem mass
Rec_globale_plein_%	full weed and service plant coverage

TAB. 2 – *Exemples de noms et de descriptions des variables sources (en anglais).*

Nom	Description
stem_juice_crop_yield_l.ha-1	stem juice yield
stem_nonstalk_crop_yield_fm_t.ha-1	stem crop yield fresh mass
sugar_crop_yield_dm_t.ha-1	sugar solid in stem

TAB. 3 – *Exemples de noms et de descriptions des variables candidates (en anglais).*

4.2 Comparaison des résultats

Notre objectif est de calculer la similarité de chaque variable source avec toutes les autres variables candidates, en les classant de la plus proche à la moins proche. La méthode qui nous intéresse est la combinaison, et nous nous concentrons spécifiquement sur les précisions à des positions inférieures à 10 (p@10). Noter que p@i représente la probabilité que la solution pertinente, c'est-à-dire la variable candidate qui correspond réellement à la variable source, soit parmi les i premières variables candidates proposées.

4.2.1 Méthode sans utiliser d'informations contextuelles

(Ngaba, 2022) a utilisé la distance de Levenshtein entre les noms, TF-IDF pour la vectorisation des descriptions, et le cosinus pour mesurer la similarité entre ces vecteurs. Les résultats obtenus avec α = 0.3 sont présentés dans TAB. 4.

Pour le travail actuel, l'architecture BERT-base avec 2 couches cachées a été utilisée pour la vectorisation des noms et des descriptions des variables, ce qui a conduit à de meilleurs résultats avec α = 0.79 et sans utilisation d'informations contextuelles (15 articles scientifiques). La mesure de similarité du cosinus a été utilisée pour évaluer les similitudes. Les résultats ont été améliorés (TAB. 5).

Précision	Levenshtein (noms des variables)	TF-IDF + cosinus (descriptions des variables)	Combinaison
p@1	15.48%	33.33%	41.67%
p@3	19.05%	42.86%	51.19%
p@5	23.81%	51.19%	64.29%
p@10	42.86%	60.71%	73.81%

TAB. 4 – *Résultats précedents en termes de précision (84 variables).*

Précision	BERT-base + cosinus (noms des variables)	BERT-base + cosinus (descriptions des variables)	Combinaison
p@1	11.90%	33.33%	41.67%
p@3	28.57%	44.05%	60.71%
p@5	36.90%	52.38%	69.05%
p@10	53.57%	57.14%	79.76%

TAB. 5 – *Résultats actuels en termes de précision (84 variables).*

4.2.2 Méthode utilisant le contexte

Pour le travail précédent, un corpus de documents a été utilisé dans l'étude menée par (Ngaba, 2022), dans le but d'améliorer le contexte des descriptions des variables et d'augmenter les résultats. Ce corpus est composé d'articles scientifiques, de chapitres d'ouvrages, de rapports et de thèses, se concentrant principalement sur les thèmes de la canne à sucre, de la fertilisation des sols, des plantes de service et des adventices. Au total, 122 documents, tous rédigés en anglais, ont été utilisés, variant en longueur de 5 à 160 pages. Pour plus d'informations, veuillez consulter le travail (Ngaba, 2022). TAB. 6 présente les résultats obtenus avec α = 0.3.

Pour le travail actuel, 15 articles scientifiques en anglais ont été utilisés. Ces articles ont été prétraités par les fonctions décrites dans la section 4.1, puis regroupés avec les noms et les descriptions des variables candidates et sources, tous prétraités et représentés dans un corpus unique (sous forme d'une liste de chaînes de caractères en Python), également appelé contexte. Ces données ont ensuite été pondérées avec TF-IDF pour constituer le vocabulaire. Pour la vectorisation des noms, des descriptions des variables et la mesure de similarité, l'architecture BERT-base avec 2 couches cachées, TF-IDF (utilisant 15 articles) et le cosinus ont été respectivement employés. Les résultats obtenus (TAB. 7) reposent sur α = 0.25.

Précision	Levenshtein (noms des variables)	TF-IDF + cosinus (descriptions des variables)	Combinaison
p@1	15.48%	33.33%	44.05%
p@3	19.05%	42.86%	55.95%
p@5	23.81%	51.19%	64.29%
p@10	42.86%	60.71%	73.81%

TAB. 6 – *Résultats précedents en termes de précision (84 variables).*

Précision	BERT-base + cosinus (noms des variables)	TF-IDF + cosinus (descriptions des variables)	Combinaison
p@1	11.90%	29.76%	52.38%
p@3	28.57%	42.86%	66.67%
p@5	36.90%	51.19%	71.43%
p@10	53.57%	64.29%	80.95%

TAB. 7 – *Résultats actuels en termes de précision (84 variables).*

4.3 Discussion

Différentes méthodes ont été utilisées pour la vectorisation des noms de variables (cf. Section 3.1). Parmi ces méthodes, le modèle BERT-base avec 2 couches cachées a donné les meilleurs résultats (TAB. 8). De même, pour la vectorisation des descriptions des variables, plusieurs méthodes ont été explorées (cf. Section 3.2). Cependant, les meilleurs résultats ont été obtenus avec l'utilisation de TF-IDF (avec l'utilisation de 15 articles) (TAB. 9). À cet effet, TF-IDF a été entraîné sur un corpus composé des noms, des descriptions des variables et d'un contexte supplémentaire issu de 15 articles.

Précision	BERT-base + cosinus	BERT-large + cosinus	XLNet + cosinus	RoBERTa + cosinus
p@1	11.90%	11.90%	9.52%	7.14%
p@3	28.57%	17.86%	23.81%	11.90%
p@5	36.90%	25.00%	27.38%	19.05%
p@10	53.57%	27.38%	34.52%	28.57%

TAB. 8 – *Les résultats en termes de précision obtenus pour la vectorisation des **noms** des variables (84 variables).*

Précision	BERT-large + cosinus	XLNet + cosinus	RoBERTa + cosinus	TF-IDF + cosinus
p@1	21.43%	11.90%	14.29%	41.67%
p@3	32.14%	21.43%	21.43%	54.76%
p@5	38.10%	25.00%	22.62%	60.71%
p@10	50.00%	30.95%	23.81%	69.05%

TAB. 9 – *Les résultats en termes de précision obtenus pour la vectorisation des **descriptions** des variables (84 variables).*

Plusieurs combinaisons de méthodes ont été utilisées pour la vectorisation des variables. Les meilleurs résultats obtenus avec ces combinaisons sont tout à fait satisfaisants, avec une

précision allant de 52.38% à 80.95%, en considérant respectivement le premier jusqu'aux 10 premiers éléments retournés. En outre, nos résultats ont largement dépassé ceux de (Ngaba, 2022). Pour la méthode de combinaison, les résultats ont augmenté de 9.52%, 4.76% et 5.95% pour p@3, p@5 et p@10 respectivement. Pour la similarité (i) entre les noms, les résultats ont augmenté de 9.52%, 1.09% et 10.71% pour p@3, p@5 et p@10 respectivement, (ii) entre les descriptions, les résultats ont augmenté de 1.19% pour p@3 et p@5.

Notons enfin que les expérimentations réalisées avec la mesure de Jaccard ont montré une augmentation de p@1 (p@1 = 58.33%), p@3 (70.24%) et p@5 (73.81%), mais en réduisant la précision pour p@10 (78.57%). Okapi BM-25, Fastext, GloVe et Word2Vec ont donné des résultats peu satisfaisants qu'il faudra approfondir dans les futurs travaux.

5 Conclusion et perspectives

Dans nos travaux, plusieurs approches ont été testées pour mesurer la similarité entre les variables sources et candidates, notamment l'utilisation du modèle BERT-base qui a donné de meilleurs résultats pour la vectorisation des noms de variables. Pour les descriptions des variables, la méthode TF-IDF avec l'enrichissement contextuel de 15 articles scientifiques en anglais a obtenu les meilleurs résultats. Des travaux complémentaires pourraient consister à explorer l'application de modèles génératifs tels que Chat-GPT pour générer un contexte de variables. Enfin, nous envisageons d'appliquer des modèles de description de variables scientifiques, tels que I-ADOPT de la RDA. Notre approche consistera à utiliser ce framework pour décrire nos variables, puis à évaluer la similarité entre les variables sources et candidates.

Remerciements : Ce travail est financé par l'Agence Nationale de la Recherche (ANR) au titre de France 2030 portant la référence ANR-16-CONV-0004 (#DigitAg) et par le programme de recherche et d'innovation Horizon Europe dans le cadre de la convention 101081973 - IntercropValuES. Ce travail a bénéficié du soutien du Conseil régional de La Réunion, du ministère français de l'Agriculture et de l'Alimentation, de l'Union européenne (programme Feader, subvention AG/974/DAAF/2016-00096 et programme Feder, subvention GURTDI 20151501-0000735).

Références

Auzoux, S. (2019). Aegis, an extended information system to support agroecological transition for sugarcane industries. In *ISSCT*.

Bojanowski, P., E. Grave, A. Joulin, et T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*.

Devlin, J., M. W. Chang, K. Lee, et K. Toutanova (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the ACL : Human Language Technologies*, pp. 4171–4186.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 17351780.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles 37*, 241–72.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jurafsky, D. et J. H. Martin (2019). *Speech and Language Processing (3rd Edition)*. Pearson.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707–710.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et V. Stoyanov (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

Manning, C. D., P. Raghavan, et H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space.

Ngaba, B. (2022). Rapport de stage : Couplage d'un modèle de culture avec une plateforme de capitalisation des données issues d'agroécosystèmes à la réunion.

Niu, Z., G. Zhong, et H. Yu (2021). A review on the attention mechanism of deep learning. *Neurocomputing 452*, 4862.

Petrova-Antonova, D. et R. Tancheva (2020). Data cleaning : A case study with openrefine and trifacta wrangler. In *Int. Conf. on the Quality of Inf. and Com. Technology*, pp. 32–40.

Ranjan, N., K. Mundada, K. Phaltane, et S. Ahmad (2016). A survey on techniques in nlp. *International Journal of Computer Applications 134*(8), 69.

Robertson, S., S. Walker, S. Jones, M. Hancock-Beaulieu, et M. Gatford (1994). Okapi at trec-3. pp. 0–.

Tracz, J., P. I. Wójcik, K. Jasinska-Kobus, R. Belluzzo, R. Mroczkowski, et I. Gawlik (2020). BERT-based similarity learning for product matching. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pp. 66–75.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems 30*.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, et Q. V. Le (2019). Xlnet : Generalized autoregressive pretraining for language understanding. In *arXiv preprint arXiv :1906.08237*.

Yin, W., K. Kann, M. Yu, et H. Schütze (2017). Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv :1702.01923*.

Summary

The objective of this work is to propose and evaluate methods of matching between source variables and candidate variables from the agroecology domain (in English). The aim of our approach is to support the experts to standardize the databases, as well as to link the source variables to the candidate variables of the dictionaries of the models used in agroecology (i.e. AEGIS).