Outil de détection automatique des erreurs de coordination, développement et perspectives

Chunxiao YAN[1] Iris Eshkol-Taravella[1] Sarah De Vogué[1] Marianne Desmets[2]
(1) Laboratoire MoDyCo, 200 Av. de la République, 92000 Nanterre, France
(2) Laboratoire de linguistique formelle, 5 rue Thomas Mann,75013 Paris, France
chunxiao.y@parisnanterre.fr, ieshkolt@parisnanterre.fr

MOTS-CLES : Erreurs de coordination, apprentissage profond, amplification de données
KEYWORDS: Coordination errors, deep learning, data amplification

1 Introduction

1.1 Contexte

Notre recherche porte sur la détection automatique des erreurs de coordination dans les écrits des étudiants universitaires. Dans le cadre du projet écri+ (PIA n°ANR-17-NCUN-0015), Noreskal et al. (2023) ont développé un outil de détection automatique des erreurs de coordination. Cet outil identifie les phrases comportant des erreurs et fournit des informations plus spécifiques sur chaque type d'erreur rencontrée. Grâce à ces informations, les étudiants peuvent effectuer des révisions ciblées qui, à terme, amélioreront la qualité de leurs écrits académiques.

1.2 Erreurs de coordination

Les erreurs liées aux structures de coordination ne sont pas rares dans les écrits des étudiants de l'enseignement supérieur. Certaines peuvent être facilement localisées. Par exemple, dans la phrase (1), il s'agit d'une préposition mal choisie : *à* devrait être remplacé par *de*. L'exemple (2) présente un problème de ponctuation : la virgule après *positif* n'est pas nécessaire. Certaines erreurs, cependant, nécessitent une analyse plus fine. L'exemple (3) montre que la coordination entre les éléments de la phrase est complétement déstructurée. Il manque un verbe avant *d'aller*.

(1) **Ces contraintes font face au désir des auteurs de transposer le mythe à leurs époques et également* ***à*** *se conserver un espace de liberté*.

(2) **Il y a eu des effets positifs**,** et des effets négatifs.*
(3) **Enfin je conclurais en les remerciant et d'aller s'inscrire aux sports qu'ils veulent faire.*

2 Développement de l'outil

2.1 Corpus

Le corpus utilisé par Noreskal (2022) pour le développement de l'outil est composé de mémoires, d'exercices, de rapports et de devoirs d'étudiants. Ce corpus, une fois recueilli, a été annoté à la main, les phrases étant classées comme « correctes » ou « erronées ». Il comprend 2 240 phrases correctes et 1 137 phrases erronées. Ensuite, les phrases erronées ont été analysées et classées en neuf catégories d'erreurs :

- PREP REMP (remplacement d'une préposition par une autre)
- PREP ADD (préposition non-attendue)
- PREP ABS (absence de préposition)
- MASV (mauvais accord sujet-verbe)
- MCGS (mauvaise cohérence des groupes syntaxiques)
- SL (structure lourde)
- Dist Conj (grande distance entre conjoints)
- PB PONC (problème de ponctuation)
- Autres

2.2 Fonctionnement de l'outil

L'outil de détection automatique comprend des modèles de classification basés sur CamemBERT (Martin et al., 2019) et des règles syntaxiques.

Tout d'abord, un modèle binaire permet de prédire si une phrase comporte des erreurs de coordination. Dès qu'une phrase erronée est détectée, un modèle multi-label classe cette phrase selon un ou plusieurs types d'erreurs prédéfinis. Ensuite, en se basant sur des arbres de dépendance annotés automatiquement par *spaCy*[1] , des règles syntaxiques sont appliquées pour aider à détecter des erreurs ayant des régularités au niveau syntaxique, telles que les erreurs de l'accord entre sujet et verbe (MASV). L'avantage d'utiliser des règles syntaxiques pour identifier les erreurs est qu'elles peuvent fournir des indications plus précises pour les corrections nécessaires, comme l'illustre la Figure *1*.

[1] *https://spacy.io/*

```
Phrase 2: Il y a des pays d'Afrique qui sont riche et qui même fais
partie des pays important dans le monde.
***Erreur détectée : Les verbes coordonnés ne sont pas conjugués de la
même façon : sont est au pluriel et fais est au singulier. Vérifiez le
sujet des verbes.(règle2)***
***Erreur détectée : Les verbes ne sont pas conjugués à la même
personne.(règle2)***
```

Figure 1 Exemples de messages produits par l'outil

Le modèle binaire atteint une F-mesure de 0,82, avec une précision de 0,91 et un rappel de 0,74. Quant au modèle multi-label, le score de précision moyenne de rangs (ang. *label ranking average precision score*) est de 0,79. Concernant les règles syntaxiques, spécifiquement pour les accords entre le sujet et le verbe, le programme obtient une F-mesure de 0,6, avec une précision de 0,75 et un rappel de 0,5.

3 Réflexions sur les possibles améliorations

3.1 Amplification des données

Les performances de l'outil développé par (Noreskal 2022, Noreskal et al. 2023) sont bonnes mais pas satisfaisantes pour l'évaluation de la qualité des écrits académiques des étudiants. Une des pistes d'amélioration de résultats est d'avoir plus de données pour l'apprentissage (Banko & Brill, 2001). Dans notre cas, il s'agit de disposer d'un nombre important de phrases coordonnées erronées. C'est la raison pour laquelle nous nous tournons vers des méthodes d'amplification des données utilisés dans le domaine du TAL (Pellicer et al., 2023). Nous mettons de côté des approches par injection de bruit synthétique ou par substitution lexicale. La méthode choisie est fondée sur l'utilisation des grands modèles de langue qui génèrent systématiquement les phrases selon le type d'erreur que nous ciblons (Whitehouse et al., 2023).

Nous nous appuyons sur le modèle Llama-3.1-70B, interrogé via l'API Groq, pour générer des phrases comportant des erreurs spécifiques. Notre première approche a consisté en une instruction unique, incluant des exemples de phrases illustrant un type d'erreur ciblé, accompagnées de nos directives précises pour la génération d'exemples. Voici l'un des prompts que nous avons utilisés :

> *[sentences] based on these sentences, generate 30 sentences in French with syntax coordination errors involving prepositions. The sentences should vary in length from 10 to 30 words, cover different themes, and range in complexity. The errors should involve the absence of preposition in the second conjunct in coordinating phrases (e.g., '*

PREP conjunct et conjunct'). Sentences should mimic the style of a humanities student. Output the sentences in [...] format.

Voici deux exemples ainsi générés, avec absence de préposition dans le deuxième conjoint :

(4) **Ils cherchent à comparer et contraster les marqueurs psycholinguistiques pour les enfants avec des troubles spécifiques du langage.*
(5) **Tout cela renvoie donc aux définitions des mots et ce que cela implique dans notre société moderne.*

Sur les 30 phrases générées, 26 ont été considérées comme valides. Pour les 4 phrases restantes, le modèle n'a pas bien simulé le type d'erreur, ou la phrase était correcte. Nous avons trouvé que ces phrases générées étaient plutôt naturelles et qu'elles préservaient la variété des données.

3.2 Evaluation de l'outil

Le choix de mesure d'évaluation pour sélectionner le meilleur modèle est aussi important. Habituellement, la mesure F1 est considérée comme un bon indicateur d'évaluation, car elle équilibre la précision et le rappel. Mais lorsqu'il s'agit d'utiliser un outil de détection d'erreurs dans un objectif pédagogique, le fait que l'outil évalue une phrase correcte comme étant erronée est lourd de conséquences et doit être autant que possible évité. Nous aimerions que l'outil soit plus précis et qu'il essaie de ne pas reconnaître les phrases correctes comme erronées. À partir de là, nous pourrions envisager d'utiliser la mesure F0,5 qui donne plus de poids à la précision, pour sélectionner de meilleurs modèles entrainés.

Ces mises à jour peuvent améliorer la performance de l'outil et permettre son utilisation à grande échelle par les étudiants.

Références

BANKO, M., & BRILL, E. (2001, JULY). Scaling to very very large corpora for natural language disambiguation. *In Proceedings of the 39th annual meeting of the Association for Computational Linguistics* (pp. 26-33).

MARTIN, L., MULLER, B., SUAREZ, P. J. O., DUPONT, Y., ROMARY, L., DE LA CLERGERIE, É. V., ... & SAGOT, B. (2019). CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894.*

NORESKAL, L. (2022). *Erreurs dans les phrases coordonnées au sein des rédactions universitaires: typologie et détection* (Doctoral dissertation, Université Paris Nanterre).

NORESKAL, L., ESHKOL, I., & DESMETS, M. (2023). Détecter une erreur dans les phrases coordonnées au sein des rédactions universitaires. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1: travaux de recherche originaux--articles longs* (pp. 248-261).

PELLICER L. F. A. O., FERREIRA T. M., & COSTA A. H. R. (2023*)*. Data augmentation techniques in natural language processing. *Applied Soft Computing*, *132*, 109803.

WHITEHOUSE, C., CHOUDHURY, M., & AJI, A. F. (2023). LLM-powered data augmentation for enhanced cross-lingual performance. *arXiv preprint arXiv:2305.14288*.