Génération et annotation de corpus pour l'entraînement et l'évaluation de modèles d'extraction de relations : utilisation de bibliothèques de génération de données et de LLMs

Ferial YAHIAOUI, Elias LIMOUNI
OPPSCIENCE
{fyahiaoui, elimouni}@oppscience.com

RESUME

Cette étude présente une méthode améliorée pour la génération et l'annotation de corpus destinés à l'entraînement et à l'évaluation de modèles d'extraction de relations. Nous identifions les limites des méthodes existantes et proposons des solutions pour surmonter ces défis. En utilisant des bibliothèques de génération de données comme Faker, des lexiques ou d'autres outils similaires ainsi que plusieurs modèles génératifs, notre solution permet de générer des données diversifiées et annotées automatiquement, offrant ainsi une alternative plus efficace et rentable à la création de données annotées manuellement.

ABSTRACT

Generation and Annotation of Corpora for Training and Evaluating Relation Extraction Models: Leveraging Generative Libraries and Models

This study presents an improved method for the generation and annotation of corpora for training and evaluating relation extraction models. We identify the limitations of existing methods and propose innovative solutions to overcome these challenges. By leveraging generative libraries such as Faker, lexicons, or other similar tools along with multiple generative models, our approach generates diverse and automatically annotated data, providing a more efficient and cost-effective alternative to manually annotated data.

MOTS-CLES : Génération de données, Annotation automatique, Modèles de langage, LLM.
KEYWORDS : Data generation, Automatic annotation, Language models, LLM.

Introduction

L'introduction des grands modèles de langage (LLMs) a révolutionné le domaine du traitement automatique des langues (TAL). Les LLMs comme GPT-3 et ses successeurs ont démontré des capacités impressionnantes dans une variété de tâches, y compris l'annotation automatique des entités nommées. Cependant, malgré ces avancées, plusieurs défis subsistent dans la génération et l'annotation des données pour l'extraction de relations, notamment le coût élevé et la répétitivité des entités générées.

Dans cette étude, nous présentons une méthode améliorée utilisant des bibliothèques de génération d'entités et plusieurs LLMs. Ces bibliothèques permettent de générer des entités nommées synthétiques variées avant de créer les textes correspondants avec un LLM. Nous avons sélectionné une solution existante adaptée à nos besoins pour éviter de coder notre propre librairie. Cependant, en raison du contexte multilingue de notre application, certaines langues comme

l'arabe et le russe ont posé des problèmes de translittération. Nous avons donc partiellement recréé le module pour ces langues afin de nous assurer d'obtenir une sortie en caractères latins. La langue de travail principale pour cette étude est l'anglais.

1. Problématique

L'extraction de relations est une tâche cruciale en TAL, permettant de déterminer les relations entre différentes entités dans un texte. Toutefois, la création de corpus annotés manuellement sur cette tâche est coûteuse et chronophage, et il est donc souhaitable de l'automatiser en tout ou partie.

Pour notre produit, un besoin concernant l'extraction de relations entre entités est remonté, et a donné naissance à ce projet de création de données. Cependant, les méthodes existantes de génération automatique de données souffrent souvent de manque de diversité et de répétitivité des entités. En effet, si nous travaillons sur des modèles d'extraction en parallèle, nous avions besoin de données annotées afin d'apprendre au modèle à extraire les relations entre entités.

Notre objectif est de combler ces lacunes en proposant une méthode qui combine les avantages de la bibliothèque Faker pour la génération de données synthétiques, avec la puissance des LLMs pour la génération de textes et leur annotation automatique.

2. Etat de l'art

L'utilisation des LLMs pour l'annotation automatique des entités nommées a été largement explorée dans la littérature scientifique récente. Voici quelques études marquantes qui soulignent les avancées et les défis dans ce domaine :

2.1. Approches initiales et avancées récentes

(Collins & Singer, 1999) explore l'utilisation de méthodes non supervisées pour améliorer les ensembles de données pour la reconnaissance d'entités nommées (NER). Cette étude utilise des règles contextuelles pour identifier et classer les entités nommées avec une intervention humaine minimale, démontrant l'importance des données non étiquetées pour améliorer les modèles.

(Tjong Kim Sang & De Meulder, 2003) introduit la tâche partagée de CoNLL-2003, qui se concentre sur la reconnaissance d'entités nommées indépendante de la langue dans des textes en anglais et en allemand. La tâche met en avant l'utilisation de diverses techniques d'apprentissage automatique et de ressources externes pour identifier des entités telles que des personnes, des organisations, des lieux et d'autres entités. L'étude a fourni des ensembles de données annotées pour l'entraînement et le développement, ainsi qu'un grand ensemble de données non annotées, encourageant ainsi l'intégration de ces dernières pour améliorer les performances des modèles. Les résultats ont montré que les systèmes les plus performants combinaient des techniques variées et utilisaient des données externes pour obtenir des scores de précision et de rappel élevés, synthétisés par le score F1.

2.2. Modèles pré-entrainés et techniques d'optimisation de la précision

(Peters et al., 2018) propose des techniques qui peuvent considérablement améliorer l'annotation automatique des corpus pour la reconnaissance des entités nommées en utilisant des représentations contextuelles profondes, augmentant ainsi la précision et la fiabilité des modèles NER. ELMo a été démontré comme étant efficace non seulement pour le NER, mais aussi pour d'autres tâches de TAL comme la résolution de coréférence et l'analyse de sentiments. Cette polyvalence suggère que les représentations ELMo peuvent enrichir la qualité des annotations automatiques à travers divers types de données et tâches.

(Devlin et al., 2019) introduit un modèle de langage pré-entraîné BERT qui a révolutionné de nombreuses tâches de NLP, y compris la reconnaissance des entités nommées. Utilisant des techniques de masquage et d'apprentissage bidirectionnel, BERT capture des relations contextuelles plus riches, améliorant ainsi la précision et les performances des modèles sur divers problèmes de traitement du langage naturel.

Cela souligne non seulement l'innovation technique apportée par BERT, mais aussi l'impact direct sur la précision des tâches NLP telles que la NER.

2.3. Reconnaissance et détection dans des contextes spécifiques

(Beltagy et al., 2019) présente SciBERT, un modèle pré-entraîné spécifiquement pour les textes scientifiques. Cette recherche démontre l'efficacité des modèles spécialisés comme SciBERT pour traiter les structures complexes des phrases et les termes techniques fréquents dans les écrits scientifiques, tout en améliorant les performances sur diverses tâches NLP telles que la reconnaissance des entités nommées, la classification de texte et l'extraction d'informations.

(Yan Hu et al., 2024) explore l'utilisation de LLMs, tels que GPT-2 et GPT-3, pour améliorer les systèmes de reconnaissance d'entités nommées cliniques. Les auteurs démontrent que GPT-3, en particulier, excelle dans l'identification et la classification des entités médicales grâce à sa capacité à comprendre des contextes complexes et à gérer des termes spécialisés. En outre, l'étude met en avant l'importance du prompt engineering, qui améliore la précision et la pertinence des entités extraites. Les résultats montrent une amélioration notable en termes de précision et de rappel par rapport aux méthodes traditionnelles, tout en soulignant les défis liés à la gestion des variations linguistiques et à la nécessité de grandes quantités de données annotées pour un entraînement optimal.

2.4. Augmentation de données avec les LLMs

(Ye et al., 2024) proposent LLM-DA, une méthode d'augmentation de données utilisant de grands modèles de langage pour la reconnaissance d'entités nommées en few-shot. Cette méthode utilise les capacités des LLMs pour générer des entités fictives qui peuvent être utilisées pour enrichir les ensembles de données existants, améliorant ainsi la performance des modèles d'extraction de relations. Leur approche démontre l'efficacité de l'utilisation des LLMs pour remplacer et générer des entités dans un cadre d'augmentation de données.

3. Méthode

3.1. Génération d'entités

Notre approche repose sur plusieurs modules de génération d'entités, chacun conçu pour créer des types spécifiques de données. Ces types incluent :

- Numéros de téléphone : Génération de numéros de téléphone synthétiques avec une diversité de formats. Par exemple : *05-33-94-75-24*, *+34 982 04 08 09*.
- Entités financières : Création d'entités financières fictives telles que des IBAN, des BIC, des cartes de crédit et des comptes bancaires. Par exemple : *DE39223442542375764825*, *SOGEFRPP*, *5431 7756 2519 3799*, 49855779.
- Numéros d'identité : Génération de numéros d'identité fictifs pour différents systèmes d'identification. Par exemple : *020378A275*, *418615182413489*.
- Entités liées aux véhicules : Production d'entités liées aux véhicules, y compris des marques, des modèles, des couleurs et des plaques d'immatriculation. Par exemple : *Une Land Rover Evoque noire immatriculée JC-910-WK*.

3.2. Modèle de relations

Le modèle de relations est en cours de développement et fera l'objet d'une publication ultérieure. Ce modèle utilise les *embeddings* des entités, que nous transformons pour mieux discriminer les types de relations. Les modèles de relations sont de type *few-shot*, avec 50 instances par relation pour l'entraînement et 100 pour l'évaluation. Cette approche permet de capturer les nuances des relations complexes entre les entités et d'améliorer la précision de l'extraction des relations.

La coordination de l'ensemble du processus de génération d'entités assure l'intégration harmonieuse des différents modules, permettant ainsi de produire un ensemble cohérent et varié de données synthétiques. Cette approche garantit une couverture complète des types d'entités requis, tout en maintenant la diversité et la précision des données générées.

3.3. Scénarios et prompts

Les scénarios dirigent notre processus de génération de données, assurant la pertinence et la richesse des prompts générés. Cette fonctionnalité permet à l'utilisateur de décrire les textes qu'il veut générer (voir Figure 1 ci-dessous) dans un format structuré aisément compréhensible par l'utilisateur (ici du YAML), puis va se charger de générer les entités via Faker ou autre dictionnaire par exemple, puis générer le prompt afin de l'envoyer au LLM pour génération. Cette étape permet un contrôle et une validation des prompts afin de respecter un standard défini durant la phase d'étude.

Exemple de prompt généré à partir d'un scénario :

```
Can you generate 5 different texts in English with direct style that mention:
- "Jean Michel" of type "person"
- "Gabriel Durand" of type "person"
- "10/04/2001" of type "date"
- "Madrid" of type "location"
- "ES6313409413296697853329" of type "iban"

and links them as follows:
- "Jean Michel" and "Gabriel Durand" by a synonym of "brother, sister, parent, family or sibling" relation
- "Jean Michel" and "10/04/2001" by a synonym of "birthdate" relation
- "Jean Michel" and "Madrid" by a synonym of "address or residence" relation
- "Jean Michel" and "ES6313409413296697853329" by a synonym of "has an IBAN" relation

Each text should contain all the entities and all the relations. Respect the given order to mention the entities.
Do not explicitly mention the label of the entities. Ensure that the format and content of the entities remain unchanged,
they all should match exactly the given text.
Provide each text between <text> and </text> tags, like <text> This is a text. It's AI generated. </text>.
```

FIGURE 1. Exemple de prompt généré à partir d'un scénario

Cet exemple de prompt illustre comment notre solution peut générer un ensemble de textes contenant les entités et les relations définies dans un scénario donné. Cette interface simple à comprendre permet à tout utilisateur d'utiliser aisément la solution afin de produire des textes répondant à ses besoins. Si tous les textes de ce scénario possèdent les mêmes types d'entités et de relations, notre solution permet d'obtenir une variété importante d'entités et de manières de mentionner la relation, sans ajout d'effort pour l'utilisateur final.

En raison de contraintes de propriété intellectuelle, nous ne pouvons pas partager certains détails internes du module de génération. Toutefois, l'exemple ci-dessus illustre clairement la méthode de création des prompts et montre comment les entités et les relations sont définies et utilisées pour générer des textes pertinents.

4. Evaluation

4.1. Génération d'entités et scénarios

Les résultats qualitatifs observés montrent que l'utilisation des modules de génération d'entités a permis de créer des jeux de données synthétiques diversifiés et réalistes. L'utilisation de ces scénarios a contribué à la production de prompts cohérents et pertinents pour l'extraction de relations.

4.2. Annotation automatique

L'annotation automatique réalisée à l'aide de plusieurs LLMs a montré une bonne précision et a réduit le temps nécessaire à l'annotation des corpus, observations faites lors de l'analyse qualitative de textes générés.

Pour ce qui est de l'évaluation quantitative, le meilleur moyen était de prendre jeu de données que nous avons généré, d'apprendre le modèle d'extraction de relations, puis de l'évaluer sur des données réelles annotées manuellement et contenant les relations du dataset généré. Le Tableau 1 ci-dessous montre des résultats obtenus sur un jeu de données généré.

Relation	Précision	Rappel	F1-Mesure	Support
BIRTHDATE	0,74	0,83	0,79	200
BIRTHPLACE	0,93	0,93	0,93	200
CURRENT_RESIDENCE	0,96	0,93	0,94	200
HAS_USES_PHONE_NUMBER	0,94	0,9	0,92	180
NOT_BIRTHDATE	0,81	0,69	0,75	200
NOT_CURRENT_LOCATION_BIRTHPLACE	0,98	0,96	0,97	200
NOT_HAS_USES_PHONE_NUMBER	0,95	0,93	0,94	200
Accuracy			0,88	1380
Macro avg	0,79	0,77	0,78	1380
Weighted avg	0,9	0,88	0,89	1380

TABLEAU 1. Tableau des résultats de l'évaluation des modules de génération de jeux de données d'entités nommées automatiquement annotées

Les résultats avec un f1-score pondéré moyen de 0,9 indique que notre méthode offre une diversité et une pertinence qui semblent prometteuses par rapport aux approches existantes. En effet, elle reproduit suffisamment la distribution des données de productions pour répondre au problème client posé.

5. Métriques utilisées pour estimer la qualité des textes

5.1. Métriques de respect des prompts : le taux de rejet

Afin d'estimer quels LLMs sont les plus efficaces pour nos cas d'usages, aussi bien au niveau du volume généré que du coût de revient, nous avons opté pour une métrique simple : le taux de rejet.

5.2. Acceptation ou rejet d'un texte généré

Pour déterminer si un texte généré est accepté ou rejeté, nous utilisons la recherche de la présence de chaque entité demandée dans le texte. Comme ces entités sont spécifiées dans le prompt, une recherche par correspondance exacte (« exact match ») de chaque entité permet de vérifier leur présence. Cette hypothèse est cruciale car toute altération des entités rendrait très compliquée l'annotation automatique du texte en trouvant les indices des entités à annoter.

Taux de rejet

Le taux de rejet est défini comme le rapport entre le nombre de textes rejetés et le nombre total de textes générés. Le taux de rejet T_r est ainsi donné par :

$$T_r = \frac{N_r}{N_t}$$

où :

- N_r est le nombre de textes rejetés.
- N_t est le nombre total de textes générés.

5.3. Métriques de diversité

Nous avons opté pour des métriques couramment utilisées pour estimer la variabilité des textes : le taux de répétition des n-grammes. En effet, cette métrique permet de détecter les patterns répétés fréquemment dans un texte, et ainsi déterminer si le LLM se répète lors de la génération de textes.

Taux de répétition des n-grammes

Pour évaluer la répétitivité des textes générés, nous calculons le taux de répétition des n-grammes, défini comme le rapport entre le nombre de n-grammes uniques (U) et le nombre total de textes (N_t) générés à partir d'un prompt. Le taux de répétition (R) en base 100 est ainsi donné par :

$$R = 100 * \frac{U}{N_t}$$

où :

- U est le nombre de n-grammes uniques.
- N_t est le nombre total de textes générés avec un prompt.

Un fort taux de répétition indique une forte répétitivité, ce qui peut suggérer une faible diversité dans les textes générés. Par exemple, si ce dernier vaut 100, cela signifie que le n-gramme étudié est présent autant de fois qu'il y a de textes.

Utilisation des n-grammes

Dans notre cas d'usage, cette métrique permettait notamment de détecter les séquences trop fréquemment générées par les LLMs. Par exemple, lors de la demande de génération d'un texte contenant des dates de naissances durant nos premiers essais, le tri-gramme (est, né, le) était présent dans la totalité des textes générés. Cette métrique nous a permis aisément de nous en rendre compte afin d'itérer et améliorer le prompt de génération. Cela permettait également de détecter d'éventuelles régressions lors des modifications faites par la suite sur les prompts.

5.4. Résultats obtenus sur nos jeux de données générés

Le taux de rejet nous a permis de détecter les prompts et LLMs qui avaient du mal à générer les textes demandés, et ainsi d'accélérer grandement notre vitesse de développement de l'outil de génération. Cela nous a également permis de détecter des régressions lors de changements majeurs des prompts.

Nous avons décidé de nous limiter à l'étude des n-grammes avec n allant de 2 à 4. Nous aurions pu étendre l'espace d'étude à 5 ou 6, mais après divers essais nous avons jugé cette échelle suffisante pour notre cas d'usage. Voici un exemple de résultats obtenus lors de la génération d'un document, donnant pour ordre au LLM de mentionner la possession par une personne d'un véhicule d'une marque donnée. On y remarque notamment l'usage prononcé d'un pattern (*if*, *you're*).

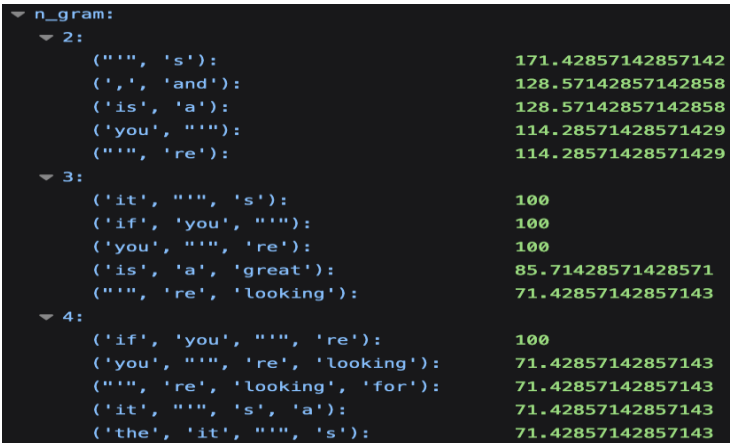

FIGURE 2. Statistiques N-grammes pour N allant de 2 à 4 inclus sur un corpus généré en anglais

6. Discussion

La méthode proposée offre une solution efficace pour la génération et l'annotation de corpus, particulièrement utile pour la génération de documents annotés, que ce soit pour les cas d'extraction d'entités nommées ou de relations. Les résultats obtenus confirment que la combinaison de la bibliothèque Faker et des LLMs permet de surmonter les limitations des méthodes actuelles. Toutefois, certaines limitations subsistent, notamment en ce qui concerne la gestion des entités rares et des contextes complexes, qui nécessitent une validation plus rigoureuse.

Conclusion

En conclusion, cette étude présente une méthode pour la génération et l'annotation de corpus, en combinant la bibliothèque Faker avec plusieurs LLMs. Notre approche améliore la diversité et la qualité des données, offrant une alternative prometteuse aux méthodes d'annotation manuelle. Les travaux futurs se concentreront sur l'amélioration de la gestion des entités rares et des contextes complexes pour encore affiner la précision des annotations.

Références

BELTAGY I., LO K., & COHAN A. (2019). SciBERT: A pretrained language model for scientific text. In EMNLP. Association for Computational Linguistics.

COLLINS M. & SINGER Y. (1999). Unsupervised models for named entity classification. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), pages 100-110, College Park, MD, USA.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, pages 4171-4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT 2018, pages 2227-2237, New Orleans, Louisiana. Association for Computational Linguistics.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142-147, Edmonton, Canada. Association for Computational Linguistics.

YAN HU Q., CHEN J., DU J., PENG X., KUTTICHI KELOTH V., ZUO X., ZHOU Y., LI Z., JIANG X., LU Z., ROBERTS K. & XU H. (2024). Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association, ocad259. https://doi.org/10,1093/jamia/ocad259

YE J., XU N., WANG Y., ZHOU J., ZHANG Q., GUI T., & HUANG X. (2024). LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. arXiv preprint arXiv:2402.14568.