Normalisation à base de règles: une stratégie efficiente pour l'extraction d'évènements fondée sur des LLMs

Luca Dini*, Pierre Jourlin **

*SATT Sud-Est, Marseille
luca.dini@univ-avignon.fr
** Laboratoire Informatique d'Avignon, Avignon
pierre.jourlin@univ-avignon.fr

Résumé. Dans cet article, nous explorons l'intégration d'un traitement symbolique des sorties d'un LLM pour obtenir une extraction d'événements à haute granularité. Nous montrons que la faiblesse des LLM dans la production d'informations structurées, souvent soulignée dans la littérature, peut être surmontée en concevant une fonction d'appariement (hybridation) adaptée au domaine. Afin d'appuyer cette affirmation, nous comparons les résultats d'une méthode d'apprentissage en contexte avec notre approche hybride et nous montrons que cette dernière permet d'obtenir des résultats supérieurs (+6,3 %) sur un nouvel ensemble de données, de triplets sujet-prédicat-objet dans le domaine médical (681 triplets pour 200 phrases). Ce résultat est obtenu en laissant le LLM (Llama-3) libre de générer les types de prédicats avec lesquels il est le plus familier, et en appliquant a posteriori une fonction de normalisation. Outre l'amélioration de l'explicabilité et de la contrôlabilité de la sortie, l'intervention d'une telle fonction (qui a été mise en œuvre en cinq jours) permet de réduire de moitié les émissions de gaz à effet de serre induites par le traitement du corpus.

1 Introduction

L'un des principaux domaines dans lesquels les LLMs [1] sont employés est la transformation d'informations non structurées (« textes libres ») en informations structurées (bases de données ou graphes). Dans cet article, nous explorons une nouvelle méthode pour réaliser cette transformation, qui s'appuie sur l'intégration de LLMs et d'un traitement symbolique (hybridation).

L'objectif fonctionnel de cette recherche est d'extraire des informations de haute qualité, par exemple, sous la forme de graphes de connaissance, à partir de rapports de cas médicaux. Pour ce faire, nous construisons d'abord un corpus de 500 extraits de rapports médicaux [2] dans le domaine de la cardiologie.

1. *Large Language Models*, en français : grands modèles de langue
2. « Un rapport détaillé du diagnostic, du traitement et du suivi d'un patient individuel. Les rapports de cas patient contiennent également des informations démographiques sur le patient (par exemple, l'âge, le sexe, l'origine ethnique). » Définition du NIH.

Dans un premier temps, chaque phrase du corpus est traitée par un LLM récent (Llama-3 de Meta) afin d'extraire les entités et les relations (ou associations) entre elles[3]. Un algorithme dédié analyse toutes les sorties afin de détecter « la façon préférée » dont le réseau tend à structurer l'information, élaborant ainsi une deuxième version de la demande qui « invite » le réseau à utiliser cette information synthétisée. La nouvelle sortie est ensuite traitée par un système à base de règles symboliques afin d'effectuer la normalisation nécessaire, détecter les incohérences et éventuellement enrichir l'information.

L'objectif non fonctionnel de cette recherche est de prouver la faisabilité de l'extraction d'information de haut niveau, d'une manière sûre, frugale et explicable. Par ailleurs, la sécurité impose l'utilisation de modèles locaux. La frugalité est obtenue par l'utilisation d'un modèle relativement petit (8B). L'explicabilité (et la flexibilité) est assurée par la couche symbolique, qui effectue une vérification supplémentaire de la sortie des LLMs.

Bien que nous nous concentrions dans cet article sur les « cas patients », la méthode est extensible à n'importe quel domaine, pourvu qu'une expertise humaine (même limitée) soit disponible, ainsi que des ontologies de domaine pour effectuer des vérifications de cohérence.

Nos principales contributions au domaine sont les suivantes :

— Un ensemble de données restreint, mais bien annoté, à utiliser pour tester les systèmes d'extraction de connaissances dans le domaine médical.
— Une méthodologie basée sur l'expertise pour augmenter la qualité des résultats des LLMs.
— Un cadre pour réaliser une extraction d'information sûre, frugale et explicable.

2 Travaux antérieurs

Le présent travail s'inscrit dans le courant de recherche que Taillé et al. (2020) décrit comme « relation strict », c'est-à-dire une tâche axée sur l'identification du type de relation et des empans correspondants aux entités de tête et de queue. Notre travail entre spécifiquement dans la catégorie des approches reposant uniquement sur des techniques de demande (*prompt*) pour obtenir des résultats. La raison de ce choix est que nous visons à étendre notre approche aux cas où aucun ensemble de données d'entraînement n'est disponible. Dans le même ordre d'idées, Agrawal et al. (2022) présente une approche *zero-shot* et *one-shot* pour les notes cliniques. Parmi les tâches qu'ils décrivent, celle qui se rapproche le plus de notre travail est l'extraction d'attributs de médicaments, pour laquelle ils obtiennent des résultats raisonnables en utilisant GPT-3 (Open AI). Cependant, la tâche est beaucoup plus restreinte que celle décrite ici, car, par exemple, ils n'ont que 6 relations au lieu de nos 18 relations. Hanafi et al. (2022) utilisent également GPT-3 afin de prouver que l'intégration d'une approche humaine basée sur des règles augmente légèrement la performance d'une approche LLM pure. Dans un certain sens, nous arrivons à la même conclusion que ce travail, à ceci près que notre approche s'appuie sur les capacités des LLMs plutôt que sur l'introduction d'un tout nouveau système. De plus, la tâche qu'ils traitent est nettement plus simple que notre tâche d'extraction de triplets. La capacité de GPT-3 à extraire des informations structurées est également évaluée par Jimenez Gutierrez et al. (2022), qui compare les résultats de l'apprentissage en contexte

3. Comme il n'y a pas de différence théorique claire entre *événements* et *relations* (qui peuvent être considérés comme des événements de type « état »), nous utilisons les deux termes de manière interchangeable. Quant au terme *prédicat*, il est utilisé pour désigner le type d'une relation.

de GPT-3 avec 5 et 10 exemples pour affiner différents modèles basés sur BERT et pour arriver à la conclusion que les modèles basés sur BERT surpassent les techniques d'*ingénierie de demande* avec GPT-3 (voir également Taillé et al. (2020)). Plus récemment, Kwak et al. (2023) utilisent chat GPT-4 afin d'effectuer l'extraction de relations et d'événements à partir de textes testamentaires légaux. Ils rapportent des résultats raisonnables, qui ne sont pas, encore une fois, comparables aux nôtres, étant donné que les entités à associer sont insérées dans la demande.

Notre approche, avec un jeu ouvert d'entités, correspond davantage à celle de Chia et al. (2022) (*Zero-Shot Relation Triple Extraction*) où seule la phrase est donnée en entrée à un modèle BART entraîné sur un ensemble de données de relations synthétiques obtenues par interrogation de GPT-2. Il est évident que si l'objectif est similaire au nôtre, l'approche est complètement différente, car nous n'effectuons aucun réglage fin. La même double étape de création de données synthétiques et de réglage fin est adoptée dans Sun et al. (2024), qui effectue l'extraction de graphes de connaissances au niveau du document. Bien que nous partagions avec les auteurs l'objectif final de produire un graphe de connaissances au niveau du document, dans notre approche, pour des raisons d'explicabilité et de contrôle, nous préférons conserver une approche au niveau de la phrase et procéder à une phase de fusions successives.

3 Contexte

Le travail décrit dans cet article a été réalisé dans le cadre du projet BACOM, où la société Skilit (https ://www.skilit.io/) a pour objectif de produire un système de recherche de documents basé sur des connaissances efficaces contenues dans la base de documents. Le principal champ d'expérimentation est constitué par les rapports de cas de PubMed, mais la méthodologie doit être extensible à n'importe quel domaine.

Comme le système doit produire de la connaissance pure, indépendamment de sa réalisation linguistique, un champ de référence évident est représenté par la recherche sur l'extraction de graphes de connaissance. Dans ce domaine, des résultats remarquables ont été obtenus récemment en recourant aux capacités de généralisation des LLM, comme dans Bi et al. (2024). Cependant, contrairement aux approches standard, la méthodologie que nous proposons doit :

— Être extensible avec peu d'effort aux domaines où aucun ensemble d'entraînement n'est disponible.
— S'appuyer sur un modèle local, afin d'éviter la diffusion d'informations sensibles sur le réseau.
— Être durable en terme d'impact sur la consommation de carbone.
— Être explicable (identifier l'origine d'une information) et contrôlable (capacité à modifier rapidement les résultats).

Dans le présent document, nous tentons de répondre à certaines de ces exigences. En particulier, nous comparons les résultats d'un LLM reposant fortement sur des techniques complexes d'élaboration de la demande (section 4.5) avec une méthode qui réduit largement la taille de la demande grâce à l'introduction d'une fonction de mise en correspondance conçue par un expert humain (section 4.6). Afin de vérifier nos résultats, nous décrivons un nouveau cor-

```
{"entities": [
  {
      "id": "he_33402219_4",
      "text": "he",
      "type": "person"
  },{
      "id": "refractory_hypocalcemia_33_4",
      "text": "refractory hypocalcemia",
      "type": "disease"
  },... ]
```

```
"relations": [
    {
      "type": "produces",
      "subject": "he_33402219_4",
      "object": "refractory_hypocalcemia_33_4"
    },
      ...
  ]
}
```


FIG. 1 – *Exemple d'annotation manuelle de la phrase « However, he developed refractory hypocalcemia with unstable vital signs . . . ».*

pus annoté (section 4) que nous mettons à la disposition de la communauté scientifique pour d'autres expériences. [4]

4 Le corpus

4.1 Pourquoi un autre corpus annoté ?

Le but ultime de cette recherche est de fournir un système capable d'extraire un ensemble complet de relations à partir de cas de rapports médicaux. Ainsi, même si nous n'effectuons aucun type d'apprentissage et de réglage fin, nous avons besoin d'une norme de référence pour vérifier l'efficacité de la méthodologie.

Les corpus *Open IE standard* dans le domaine médical (Wang et al. (2018)) ne correspondent pas à notre objectif, car ils n'utilisent pas un ensemble limité de prédicats sémantiquement fondés, mais adoptent plutôt la forme de surface linguistique en tant que relation entre deux entités ou plus. D'autres ensembles de données médicales, tels que le corpus ADE (Gurulingappa et al. (2012)), le corpus DDI (Herrero-Zazo et al. (2013)) et le corpus PGR (Gurulingappa et al. (2012)) ne correspondent pas à notre objectif.

Notre objectif d'annotation est proche de celui de Dumitrache et al. (2018), à la différence près qu'au lieu de nous concentrer sur un ensemble limité de relations, nous visons à inclure toutes les relations du graphe UMLS Metathesaurus, qui peuvent être utilisées pour obtenir une représentation sémantique raisonnable des phrases dans leur intégrité.

4.2 Description du corpus

Le corpus est composé de triplets associant un sujet et un objet via un prédicat typé. Les sujets et les objets sont des entités explicitement mentionnées dans une phrase et identifiées par leur empan. La relation a un type choisi parmi un ensemble fermé de 18 types de relations sémantiques du réseau sémantique UMLS et n'est pas caractérisée par une référence à un segment de texte particulier. La figure 1 fournit un exemple de phrase annotée.

4.3 Création du corpus annoté

Afin de disposer d'une base solide pour l'annotation, nous avons sélectionné 500 résumés de la revue Medical Case Reports sur Europe PMC, tous étant des réponses positives

4. Adresse :`https://github.com/morning-star-1789/case_reports_events_corpus`.

à la requête `cardiac` émise via l'API PMC standard. Après le nettoyage textuel habituel, nous avons procédé au découpage des phrases à l'aide du « tokeniseur » NLTK, ce qui nous a permis d'obtenir 2700 phrases qui représentent notre *corpus brut*. Avant de commencer la phase d'annotation, nous avons passé toutes les phrases à LLama-3 afin d'avoir une segmentation raisonnable des entités nommées (EN). Il est important de noter ici que la notion de « segmentation raisonnable » est floue, car, en définitive, la segmentation des entités nommées répond aux critères suivants :

— Être compatible avec le LLM en ce qui concerne l'EN qu'il considérerait normalement comme ayant un rôle dans une relation : par exemple, nous acceptons la décision de LLAMA-3 d'exclure le déterminant des EN.
— Être linguistiquement motivé : par exemple, LLAMA-3 décide parfois de remplacer de longs groupes nominaux tels que *Un homme caucasien de 29 ans* par `patient`. En outre, dans certains cas, une entité nommée n'est tout simplement pas reconnue, car elle ne fait pas partie d'une relation.
— Ne pas (encore) traiter les contextes intensionnels (Frege (1892)) : par exemple, dans la phrase *complaint of failure to thrive and short height*, nous nous concentrons sur *failure to thrive* et *short height*.

Le cœur de l'annotation consiste à relier les entités nommées par le biais de relations. Comme indiqué ci-dessus, nous utilisons les relations sémantiques du metathesaurus UMLS pour établir ces liens. Il est vrai que nous avons légèrement forcé certaines relations pour les adapter à la nature événementielle des rapports de cas, qui sont normalement structurés comme une séquence d'événements et de leurs résultats. Par exemple, la relation `carries_out` a été étendue à une relation plus générique « jouer un rôle dans un événement ». Deux relations non UMLS ont également été introduites, à savoir `goal` et `negates`.

200 phrases choisies au hasard dans le « corpus brut » ont été annotées par une personne ayant une certaine expérience des textes dans le domaine médical, sans être une praticienne de la médecine : ce choix est principalement dû au fait que nous voulions que les décisions d'appariement soient ancrées dans le texte et non pas inspirées par une connaissance linguistique supplémentaire qu'un professionnel du domaine de la santé pourrait avoir[5]. Pour la même raison, les phrases sont annotées en « isolement », c'est-à-dire sans référence au contexte précédent et suivant, toujours pour éviter l'introduction d'inférences contextuelles fallacieuses dans l'annotation.

4.4 Données du corpus

Le corpus est composé de 200 phrases complètement annotées (pour un total de 2780 mots). Il contient globalement 681 triplets, entièrement annotés avec les rôles `subject` et `object`. Il existe au total 18 types de prédicats et 20 types d'entités : les dix premiers sont répertoriés dans le tableau 1.

L'objectif de cette recherche est de tester dans quelle mesure un post-traitement symbolique sur la sortie des LLM peut remplacer l'utilisation de techniques d'écriture de demande très complexes et consommatrices d'énergie. Afin de réaliser une évaluation équitable, nous avons utilisé la même architecture pour toutes les expériences.

5. Nous prévoyons de publier une deuxième version du corpus dans laquelle tous les triplets seront validés par un professionnel du domaine.

Prédicat	**Occurences**	**Type d'entité**	**Occurences**
treated_by	100	person	142
has_result	82	condition	107
temporally_related_to	80	disease	101
exhibits	70	symptom	89
affected_by	58	event	79
carries_out	50	treatment	69
has_feature	45	test	62
has_location	40	time	52
causes	38	medicine	49
consists_of	27	location	28

TAB. 1 – *Distribution des 10 types d'entités et de prédicats les plus fréquents*

En ce qui concerne la sélection du modèle le mieux adapté à nos expériences, nous avons testé plusieurs modèles locaux avec un ensemble limité de *demandes* afin d'évaluer leur capacité à produire des sorties structurées. Après avoir exclu ceux qui ne pouvaient pas produire un résultat structuré pour au moins 80 % des phrases du corpus brut, nous avons appliqué une procédure de sélection visant à choisir le modèle dont les résultats étaient les « plus stables ». La stabilité s'entend comme une tendance naturelle à produire, pour des phrases d'entrée différentes, des sorties ayant la même structure. À l'issue de cette évaluation, le modèle le plus performant a été le Lama-3 de Meta (`Llama-3-8B-Instruct`) avec un indice de stabilité de 0,97. Toutes les expériences de la section suivante utilisent invariablement ce modèle, qui est invoqué localement avec une température de 0,01.

En ce qui concerne l'évaluation, nous suivons un schéma traditionnel dans lequel la précision (P) est `VP/(VP+FP)`, le rappel (R) est `VP/(VP+FN)` et la mesure F (F) est la moyenne harmonique des deux [6]. En ce qui concerne le concept de correspondance, nous proposons deux types d'évaluation :

— **Évaluation des triplets** : un triplet de référence et un triplet prédit concordent ssi les valeurs du sujet, de l'objet et du type sont identiques du point de vue des *tokens*.

— **Évaluation des paires** : les triplets (références et prédits) sont divisés en paires subj-pred et pred-obj. Une paire correspond si les deux membres sont identiques. Si, d'un point de vue scientifique, cette évaluation peut être considérée comme peu intéressante, elle est importante dans un contexte d'application pour mesurer la qualité d'informations partiellement correctes.

Enfin, il convient de mentionner que la sortie de Llama-3 est traitée de manière plus approfondie afin de corriger les erreurs formelles, qui concernent principalement les entités qui apparaissent dans les triplets, mais qui ne sont pas répertoriées dans `entities`.

Dans la suite de cette section, nous comparons deux techniques pour forcer les LLMs à produire un résultat désiré. L'une est fortement fondée sur l'enrichissement de demande (4.5). L'autre repose quant à elle sur l'introduction d'une fonction de mise en correspondance, codée par une personne (4.6).

6. VP, FP, FN étant respectivement les nombres de vrais positifs, faux positifs et faux négatifs

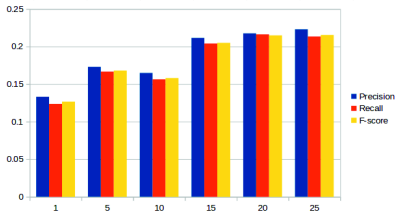

FIG. 2 – *Évolution de la qualité des triplets avec l'augmentation des exemples.*

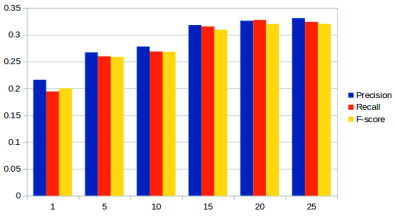

FIG. 3 – *Évolution de la qualité des paires avec l'augmentation des exemples.*

4.5 Fais ce que je veux

Dans la première série d'expériences, l'objectif était d'évaluer la quantité d'informations contextuelles à envoyer au LLM afin d'obtenir des résultats structurés valides et corrects. Nos expériences sont basées sur deux types d'exécution différents, qui ajoutent à la demande d'entrée du LLM :

— Un nombre croissant d'exemples aléatoires.
— Un exemple aléatoire pour *chaque* type de prédicat dans le jeu de données.

Tous les types d'expériences sont donc basés sur l'hypothèse qu'aucune intervention manuelle n'est nécessaire, à l'exception, bien sûr, de la constitution du corpus initial de 200 phrases qui est utilisé **seulement** pour l'évaluation et la fourniture d'exemples. La demande fournie au système utilisé dans ce lot d'expériences est assez simple :

```
You are an information extraction system. Output format { "entities":[], "relations":[]}. Relations have
an  attribute "type"  with valid values: {RELS_TYPES }
Relations must have subject and object ids which contain text in the input sentence. Do not invent ids.
Here are some examples:
```

où `RELS_TYPES` est remplacé par la liste cible de relations UMLS telles que `treated_by`, `has_result`, `temporally_related_to`, ...La chaîne `Here are some examples` est ensuite suivie de N exemples tirés au hasard dans l'ensemble de données de référence.

Nous réalisons des expériences avec un nombre d'échantillons de 1, 5, 10, 15 et 25. Comme les échantillons sont choisis au hasard et que leur qualité peut influencer les résultats, nous exécutons chaque expérience trois fois et faisons la moyenne des résultats. Les figures 2 et 3 montrent comment différentes mesures varient avec l'augmentation du nombre d'exemples. Comme nous pouvons le constater, nous observons une amélioration à mesure que le nombre d'exemples augmente, qui a tendance à se stabiliser autour de 20 exemples.

Une approche alternative consiste à fournir un exemple sélectionné au hasard pour chacun des prédicats. En particulier, pour chaque type de prédicat, nous fournissons comme exemple le « graphe sémantique » d'une phrase choisie au hasard contenant *au moins* une instance de ce prédicat. Les résultats (deux premières colonnes du tableau 3) sont comparables à ce que nous avons obtenu avec 15 exemples, et légèrement inférieurs à l'expérience des 25 exemples aléatoires.

4.6 Fais ce que tu veux

Dans cette expérience, nous avons adopté une approche opposée à celle décrite dans la section précédente. Plutôt que de choisir une approche sans travail humain et de jouer avec dif-

raw_sources	preferred
[{ 'HAS_SYMPTOM' : 47}, { 'HasSymptom' : 1}, { 'SYMPTOM' : 1}, { 'Symptom' : 25}, { 'had_symptom' : 4}, { 'has symptom' : 48}, { 'hasSymptom' : 3}, { 'has_symptom' : 196}, { 'has_symptoms' : 2}, { 'symptom' : 19}]	has_symptom
[{ 'TREATMENT' : 110}, { 'Treatment' : 40}, { 'co-treatment' : 2}, { 'has treatment' : 2}, { 'has_treatment' : 4}, { 'treatment' : 148}, { 'treatment_for' : 3}, { 'treatment_of' : 1}]	treatment
[{ 'CAUSE' : 85}, { 'Cause' : 15}, { 'Causes' : 13}, { 'cause' : 36}, { 'caused' : 19}, { 'causes' : 104}]	causes

TAB. 2 – *Normalisation pour les 3 prédicats les plus fréquents*

férentes demandes soumises au LLM, nous testons la possibilité d'obtenir des résultats comparables en intervenant sur la sortie des LLM. Ainsi, plutôt que de forcer le réseau à produire des données directement utilisables, nous essayons de restreindre la sortie du réseau à un ensemble de prédicats qui lui sont familiers, puis nous essayons de les projeter sur les prédicats cibles dans l'annotation de référence. L'idée sous-jacente est que, durant la phase d'apprentissage, le LLM a pu rencontrer un ensemble limité de phrases annotées avec des prédicats UMLS et pourrait ne pas être en mesure d'établir la connexion entre deux prédicats provenant de deux ontologies différentes. D'autre part, si nous le forçons à émettre uniquement les prédicats avec lesquels il est le plus familier, nous pourrions «capitaliser« sur un ensemble plus large d'exemples vus lors de l'entraînement, au détriment de l'écriture d'une fonction de mappage entre différentes ontologies.

La première étape consiste à identifier les prédicats avec lesquels le réseau est le plus familier. À cette fin, nous utilisons le corpus brut de 500 résumés provenant de Europe PMC, les divisons en phrases comme décrit dans la section 4.1, et les analysons selon la demande suivante, indépendante du domaine :

```
System: You are an information extraction system.
User: Extract noun phrases and temporal modifiers from "{sent}"format:JSON. Then extract the semantic
relations between the above noun phrases and temporal modifiers as a graph (format JSON). Be as
specific as possible in edge names.
```

Les objets JSON résultants sont normalisés, selon le type de propriété qui contient normalement le type de prédicat (`type, rel,` etc.) et ensuite toutes les valeurs possibles sont normalisées selon des modifications typographiques fixes (underscore vs. slash, casse, etc.) et regroupées en utilisant un ratio de proximité de Levenshtein supérieur à 0,8.

Cette phase de normalisation produit une liste de 1257 prédicats uniques. La figure 2 illustre les trois premiers éléments de la liste, avec la colonne `raw_sources` représentant la sortie du LLM et la colonne `preferred` représentant le prédicat préféré du LLM (celui que nous choisissons comme représentatif en raison de sa fréquence plus élevée).

Une fois que nous connaissons les types de prédicats préférés de notre LLM, nous pouvons reformuler la demande comme suit :

```
System:You are an information extraction system.  Output format: { "entities":[], "relations":[]}.
Relations have an attribute  "type" with valid values: { PREFERRED\_RELS}".
User:Extract a semantic graph from the sentence "{SENT}" (format: JSON)
```

Ici, le point crucial est représenté par la variable `PREFERRED_RELS` qui contient les 46 prédicats les plus fréquents sélectionnés selon l'étape précédente (par exemple, `has_symptom, treatment, causes, ...`)

Avec cette nouvelle demande, le LLM produit désormais un ensemble de types de prédicats beaucoup plus uniforme, qui doivent être projetés sur nos 18 relations UMLS.

La fonction de projection (voir aussi Agrawal et al. (2022) et leur module *resolver*) est constituée d'un ensemble d'instructions déclaratives qui s'appliquent séquentiellement pour

transformer la sortie du LLM en objets à soumettre à l'évaluation. Dans sa forme la plus simple, une déclaration est simplement une fonction de mappage de chaîne des prédicats préférés du LLM vers les types UMLS-SN. Cependant, elles peuvent être enrichies de conditions et d'actions supplémentaires telles que l'inversion sujet-objet et la vérification conditionnelle des types d'arguments. Par exemple, la règle suivante :

```
'performed': {
  'label':'carries_out',
  'id':'T141',
  'source':'https://evsexplore.semantics.cancer.gov/evsexplore/hierarchy/umlssemnet/T141',
  'role_actions':['invert', 'dummyfy_subj']
  }
```

indique qu'un prédicat `performed` du LLM est mappé sur UMLS comme `carries_out`, que le sujet original devient l'objet et qu'un sujet fictif est inséré. Le dépôt en ligne contient la liste des règles de mappage utilisées. Dans l'ensemble, nous appliquons un ensemble de 32 déclarations, qui ont été rédigées par un linguiste en environ cinq jours de travail. Il convient de noter que, bien que le linguiste ait eu accès aux directives d'annotation et au corpus brut analysé par le LLM, il n'avait pas accès à l'ensemble de données de référence.

Comme on peut le voir dans le tableau 3 (colonnes 5 et 6), le modèle de règles de mappage LLM-mixte produit des résultats comparables (et même légèrement meilleurs) à ceux produits par la meilleure demande de la section précédente. [7]

t_ou_p	m	micro_ou_macro	1 ex. / préd.	25 échantillons	Hybrid	Diff
paires	**F**	**macro**	0,3085	0,3173	**0,3320**	0,0147
		micro	0,3123	0,3204	**0,3320**	0,0147
triplets	**F**	**macro**	0,1900	0,2126	**0,2765**	0,0639
		micro	0,2002	0,2154	**0,2701**	0,0546

TAB. 3 – *Comparaison entre l'échantillonnage d'un exemple par type de prédicat, 25 exemples aléatoires et notre approche hybride.* F *est le F-score,* macro *est la moyenne des F par phrase,* micro *considère l'ensemble des triplets.*

5 Comparaison avec l'État de l'Art

Étant donné que nous avons introduit un nouveau jeu de données, il n'est pas possible de comparer nos résultats avec d'autres travaux dans le domaine. Malheureusement, aucun des jeux d'évaluation disponibles sur l'extraction de graphes de connaissances ne répond à nos besoins d'extraction d'événements typés à partir de textes médicaux, sans information fournie : en effet, ils supposent principalement que les entités sujet et objet sont données en entrée et que la tâche consiste à identifier le bon prédicat qui les relie (par exemple, Amin et al. (2022)).

Nos résultats sont plus comparables aux travaux réalisés dans le domaine de l'extraction de triplets en *zero-shot* (par exemple, Liao et al. (2024)), notamment sur des jeux de données tels que Wiki-ZSL et FewRel. Néanmoins, même dans ce cas, nous observons deux principales différences, qui rendraient la comparaison injuste : à savoir que notre objectif est de fournir

7. Étant donné que la méthodologie hybride implique un certain traitement humain, nous voulions nous assurer qu'aucun biais n'était introduit. Ainsi, à la fin de la rédaction de cet article, nous avons demandé à un annotateur différent d'annoter 51 phrases supplémentaires selon les directives. En appliquant notre méthode hybride sur cette section complètement inédite du corpus, on obtient 0,3677 F micro sur les paires et 0,2820 F micro sur les paires.

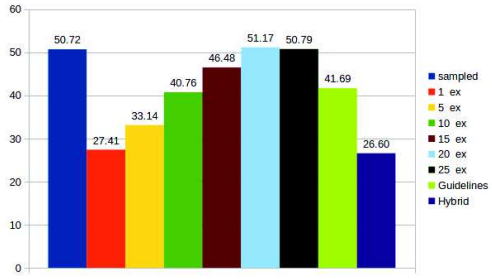

FIG. 4 – *Émissions de CO_2 (en kg.) pour le traitement du corpus dans diverses expériences.*

une représentation *complète* des événements dans une phrase, tandis que les jeux de données ci-dessus se concentrent sur un ensemble limité de prédicats. De plus, l'extraction de triplets est centrée autour de prédicats statiques tels que *père*, *propriétaire*, etc., tandis que notre jeu de données est de nature événementielle. Nous espérons qu'à l'avenir, notre jeu de données sera considéré pour des expérimentations avec différentes méthodes, établissant ainsi une base de référence plus consolidée.

6 Conclusions

Dans cet article, nous avons montré qu'une approche hybride peut atteindre des résultats comparables, voire meilleurs, que des approches entièrement automatiques reposant uniquement sur l'enrichissement de la demande. En conclusion, nous soutenons que, au-delà de la performance, des approches hybrides telles que celle introduite ici pourraient ouvrir la voie à une IA beaucoup plus durable.

La figure 4, qui rapporte la consommation de CO_2 de chaque expérience exprimée en kg. [8], montre que la consommation moyenne de notre meilleure expérience entièrement automatique avec 25 échantillons est presque le double de notre approche hybride (ratio : 0,51). Cela peut sembler un résultat marginal, mais nous croyons que lorsqu'il s'agit de jeux de données tels que PubMed, cela doit être pris en compte avec soin. Par exemple, si l'on devait analyser l'ensemble de PubMed (300 milliards de tokens contre nos 2780 tokens), l'émission de CO_2 selon la méthode entièrement automatique serait de 5,5 millions de tonnes contre 2,8 millions de tonnes. La différence équivaut à l'émission annuelle produite par 424 000 individus [9] ou aux GES de l'État de New York pendant 2,8 jours [10] Si l'on prend en compte le fait que l'intervention manuelle a nécessité environ une semaine de travail par un linguiste professionnel, l'impact environnemental et économique des approches hybrides est évident.

Du côté du contrôle et de l'explicabilité, notre approche hybride suppose une analyse attentive du jeu de données. Afin de mettre en œuvre les directives de mappage, il est en effet

8. Les détails du calcul sont disponibles dans le dépôt du projet.

9. Émissions mondiales de GES (gaz à effet de serre) par habitant en 2023 selon https ://edgar.jrc.ec.europa.eu/report_2023 : 6,76

10. "En 2021, les émissions brutes de GES à l'échelle de l'État étaient de 367,87 millions de tonnes métriques de dioxyde de carbone équivalent (mmt CO2e) selon la comptabilité CLCPA." Source : https ://dec.ny.gov/sites/default/files/2023-12/summaryreportnysghgemissionsreport2023.pdf

nécessaire d'avoir une gamme complète de sorties possibles des LLM. Cette phase d'analyse révèle également d'éventuelles déviations (par exemple, en termes de types de prédicats ou de structures de sortie formelles) qui peuvent être prises en compte dans la phase de mappage, prévenant ainsi, dans une certaine mesure, les hallucinations. De plus, la fonction de mappage elle-même constitue un contrôle de cohérence sur la sortie : seules les sorties mappées sont disponibles pour l'application en aval, empêchant ainsi des prédicats déraisonnables, ou pire : des prédicats nuisibles.

Références

Agrawal, M., S. Hegselmann, H. Lang, Y. Kim, et D. Sontag (2022). Large language models are few-shot clinical information extractors. In Y. Goldberg, Z. Kozareva, et Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, pp. 1998–2022. Association for Computational Linguistics, doi: 10.18653/v1/2022.emnlp-main.130.

Amin, S., P. Minervini, D. Chang, P. Stenetorp, et G. Neumann (2022). MedDistant19 : Towards an accurate benchmark for broad-coverage biomedical relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, pp. 2259–2277. International Committee on Computational Linguistics.

Bi, Z., J. Chen, Y. Jiang, F. Xiong, W. Guo, H. Chen, et N. Zhang (2024). Codekgc : Code language model for generative knowledge graph construction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23*(3), doi: 10.1145/3641850.

Chia, Y. K., L. Bing, S. Poria, et L. Si (2022). RelationPrompt : Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In S. Muresan, P. Nakov, et A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics : ACL 2022*, Dublin, Ireland, pp. 45–57. Association for Computational Linguistics, doi: 10.18653/v1/2022.findings-acl.5.

Dumitrache, A., L. Aroyo, et C. Welty (2018). Crowdsourcing ground truth for medical relation extraction. *ACM Trans. Interact. Intell. Syst. 8*(2), doi: 10.1145/3152889.

Frege, G. (1892). Uber sinn und bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik 100*(1), 25–50.

Gurulingappa, H., A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, et L. Toldo (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics 45*(5), 885 – 892, doi: https://doi.org/10.1016/j.jbi.2012.04.008. Text Mining and Natural Language Processing in Pharmacogenomics.

Hanafi, M., Y. Katsis, I. Jindal, et L. Popa (2022). A comparative analysis between human-in-the-loop systems and large language models for pattern extraction tasks. In E. Dragut, Y. Li, L. Popa, S. Vucetic, et S. Srivastava (Eds.), *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, Abu Dhabi, United Arab Emirates (Hybrid), pp. 43–50. Association for Computational Linguistics.

Herrero-Zazo, M., I. Segura-Bedmar, P. Martínez, et T. Declerck (2013). The ddi corpus : An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of*

Biomedical Informatics 46(5), 914–920, doi: https://doi.org/10.1016/j.jbi.2013.07.011.

Jimenez Gutierrez, B., N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, et Y. Su (2022). Thinking about GPT-3 in-context learning for biomedical IE ? think again. In Y. Goldberg, Z. Kozareva, et Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics : EMNLP 2022*, Abu Dhabi, United Arab Emirates, pp. 4497–4512. Association for Computational Linguistics, doi: 10.18653/v1/2022.findings-emnlp.329.

Kwak, A., C. Jeong, G. Forte, D. Bambauer, C. Morrison, et M. Surdeanu (2023). Information extraction from legal wills : How well does GPT-4 do ? In H. Bouamor, J. Pino, et K. Bali (Eds.), *Findings of the Association for Computational Linguistics : EMNLP 2023*, Singapore, pp. 4336–4353. Association for Computational Linguistics, doi: 10.18653/v1/2023.findings-emnlp.287.

Liao, W., Z. Liu, Y. Zhang, X. Huang, N. Liu, T. Liu, Q. Li, X. Li, et H. Cai (2024). Zero-shot relation triplet extraction as next-sentence prediction. *Knowledge-Based Systems*, 112507, doi: https://doi.org/10.1016/j.knosys.2024.112507.

Sun, Q., K. Huang, X. Yang, R. Tong, K. Zhang, et S. Poria (2024). Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, New York, NY, USA, pp. 4407–4416. Association for Computing Machinery, doi: 10.1145/3589334.3645678.

Taillé, B., V. Guigue, G. Scoutheeten, et P. Gallinari (2020). Let's Stop Incorrect Comparisons in End-to-end Relation Extraction ! In B. Webber, T. Cohn, Y. He, et Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 3689–3701. Association for Computational Linguistics, doi: 10.18653/v1/2020.emnlp-main.301.

Wang, X., Y. Zhang, Q. Li, Y. Chen, et J. Han (2018). Open information extraction with meta-pattern discovery in biomedical literature. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, New York, NY, USA, pp. 291–300. Association for Computing Machinery, doi: 10.1145/3233547.3233594.

Summary

In this paper, we explore the integration of LLMs with symbolic processing for achieving high granularity event extraction. We will show that the weakness of LLMs in producing structured information, often pointed out in the literature, can be overcomed by designing a domain tailored mapping function (hybridization). In order to support this claim, we compare the results of an in-context learning method with our hybrid methodology and we show that we can achieve superior results (+6.3 %) on a new dataset of subject-predicate-object triples in the medical domain (681 triples for 200 sentences). This result is achieved by leaving the LLM (Llama-3) free to generate the predicate types it is more familiar with, and then applying a mapping function. Besides improving explainability and controllability of the output, the intervention of such a function (which was implemented in five days), causes about a half reduction of GHG emissions produced when processing the corpus.