POPCORN : IA d'extraction d'information à partir de sources textuelles pour le renseignement militaire

Cédric Lopez, Sylvain Verdy, Guillaume Gadek, Maxime Prieur, Didier Schwab, Gilles Sérasset, Vuth Nakanyseth

POPCORN : IA d'extraction d'information à partir de sources textuelles pour le renseignement militaire

Cédric Lopez
Emvista
34830 Jacou, France
cedric.lopez@emvista.com

Sylvain Verdy
Emvista
34830 Jacou, France
sylvain.verdy@emvista.com

Guillaume Gadek
Airbus Defence & Space
78990 Elancourt, France
guillaume.gadek@airbus.com

Maxime Prieur
Airbus Defence & Space
78990 Elancourt, France
maxime.prieur@airbus.com

Didier Schwab
Univ. Grenoble Alpes, CNRS, LIG
38000 Grenoble, France
didier.schwab@univ-grenoble-alpes.fr

Gilles Sérasset
Univ. Grenoble Alpes, CNRS, LIG
38000 Grenoble, France
gilles.serasset@imag.fr

Nakanyseth Vuth
Univ. Grenoble Alpes, CNRS, LIG
38000 Grenoble, France
nakanyseth.vuth@univ-grenoble-alpes.fr

***Abstract*—Le projet de recherche collaboratif POPCORN (Peuplement OPérationnel de bases de COnnaissances et Réseaux de Neurones) a pour objectif de porter à maturité des technologies d'extraction d'informations contenues dans des documents textuels. L'article présente nos contributions autour de l'un de nos cas d'usage "renseignement militaire" concernant trois questions : Quelles données utiliser pour l'entraînement de nos modèles d'intelligence artificielle ? Quelles informations extraire ? Quels modèles adopter ? Nous exposons dans cet article les résultats obtenus sur deux tâches : l'extraction d'entités d'intérêt et l'extraction des relations entre ces entités.**

***Index Terms*—extraction d'information, renseignement militaire, traitement automatique du langage naturel, intelligence artificielle**

I. Introduction

Le projet de recherche collaboratif POPCORN a pour objectif de porter à maturité des technologies d'extraction d'informations contenues dans des documents textuels. Ce projet de 36 mois est subventionné par l'Agence de l'Innovation de Défense (AID) et implique les entreprises Emvista et Airbus (Defence and Space) ainsi que le Laboratoire d'Informatique de Grenoble (LIG). Le consortium ainsi formé est fort de spécialistes du domaine de recherche de l'Extraction d'Information (EI) et plus généralement des techniques de Traitement Automatique du Langage Naturel (TALN). Le projet touche à sa fin et nous présentons dans cet article nos contributions à l'état de l'art scientifique ainsi que quelques retours d'expériences et réflexions autour de l'un de nos cas d'usage « renseignement militaire ».

La section II décrit ce cas d'usage et explicite les contraintes métiers et techniques ainsi que les objectifs à atteindre. Un état de l'art scientifique relatif aux verrous à lever dans le cadre de ce cas d'usage est dressé en section III. Pour chacun des verrous identifiés, nous détaillons quelles approches ont été expérimentées pour contribuer à leur levée et les résultats obtenus. Enfin, nous présentons en section IV l'intégration de nos résultats dans différents démonstrateurs.

Le projet POPCORN duquel ces réflexions sont issues a bénéficié d'une subvention de l'Agence Innovation Défense (AID) et de l'accompagnement de la Direction Générale de l'Armement (DGA).

II. Cas d'usage « renseignement militaire »

Nous nous focaliserons dans cet article sur l'un des cas d'usage de POPCORN qui concerne la compréhension des menaces par un suivi des individus et des groupes subversifs, criminels ou terroristes et de leurs activités dans le domaine du renseignement de sécurité/défense. Pour tendre vers une maîtrise de ces menaces, POPCORN s'est concentré sur la capacité à extraire des informations à partir de bulletins de renseignements et autres sources de données textuelles. Il s'agit de se focaliser dans un premier temps sur l'analyse de bulletins en français car peu de travaux scientifiques traitent cette langue pour les tâches d'EI concernées par ce cas d'usage (cf. section III).

Dans ce cadre, il existe plusieurs systèmes (Palantir, IBM i2 Analyst's Notebook, DCGS, SIEM, etc.) qui disposent de capacités permettant à des utilisateurs de capitaliser les connaissances dans une base dédiée, mais ces fonctions sont aujourd'hui encore en grande partie quasi entièrement exécutées manuellement et demandent aux utilisateurs de s'approprier de grands ensembles documentaires afin de les structurer, ce qui est une charge lourde et coûteuse. Par la mise en œuvre de techniques automatisées d'extraction d'information, le projet POPCORN vise à simplifier le travail des utilisateurs en automatisant la transformation d'informations non structurées en informations structurées afin de peupler les bases de connaissances. Cette automatisation permettrait notamment d'augmenter la quantité d'informations en bases de connaissances et de réduire le temps nécessaire à l'intégration de ces connaissances dans lesdites bases.

Les avantages procurés par une base ainsi structurée à partir de milliers de documents consistent globalement en l'apport de connaissances aux opérationnels qui impactent leur compréhension de la situation. Par ailleurs, des alertes sur des entités critiques peuvent être levées automatiquement dans le

but d'augmenter la réactivité des opérationnels. Par exemple, en sécurité maritime, une alerte peut concerner un navire de transport devenu suspect en raison des relations entretenues par ses propriétaires.

III. État de la question

Dans le but de peupler une base de connaissances à partir d'informations extraites de textes tout venant (non structurés), il est nécessaire de considérer *a minima* les deux tâches suivantes :

- Extraction d'entités d'intérêt (EEI) : il s'agit d'extraire les éléments textuels qui ont un intérêt pour le cas d'usage considéré et de les classer, par exemple des termes désignant des exercices militaires ou des incidents diplomatiques (cf. Fig. 1) ;
- Extraction de relations d'intérêt (ER) : il s'agit de repérer et de classer les relations explicites et implicites qui existent entre deux entités d'intérêt. Par exemple les relations temporelles ou locatives des événements (cf. Fig. 1).

Les progrès récents en intelligence artificielle laissent penser que ces tâches appliquées à des cas d'usage sécurité et défense pourraient atteindre des résultats satisfaisants. Dans la suite, nous exposons les trois verrous traités dans POPCORN ainsi que nos contributions, respectivement sur la représentation des connaissances, l'accès aux données, et la performance des modèles.

A. Quelle représentation adopter ?

Les deux tâches décrites ci-avant nécessitent l'existence d'une liste de classes d'intérêt prédéfinies (au moins pour une phase d'évaluation des systèmes) pertinentes pour le métier. Ces classes liées par les relations forment l'ontologie qui permet de représenter les connaissances d'intérêt pour un cas d'usage donné. Étant donné que le consortium de POPCORN n'a pas accès aux ontologies réellement utilisées par les services de renseignement, le consortium a expérimenté deux ontologies :

- **Ontologie POPCORN.** D'une part, le consortium s'est appuyé sur l'ontologie MIM (Multilateral Information Model) (cf. Fig. 3), développée dans le cadre du Multilateral Interoperability Programme (MIP), un organe de standardisation qui comprend 24 nations, l'agence européenne de défense et l'OTAN. Ce modèle adapté du MIM est structuré autour du pentagramme du renseignement (Joint C3 Information Exchange Data Model - JC3IEDM) et vise à fournir un standard d'interopérabilité pour les applications de C2 (Command and Control) et d'ISR (Intelligence Surveillance and Recognition). De cette ontologie qui représente 1200 classes d'intérêts, nous avons décliné une ontologie plus modeste, contenant une cinquantaine de classes d'intérêts, utilisables concrètement dans le cadre de nos expériences (cf. Fig. 2). La conception de cette ontologie a été guidée par le besoin d'Airbus Defence and Space.
- **Ontologie MR4AP.** En guise d'alternative à l'ontologie POPCORN et afin d'éviter le développement de solutions ayant une dépendance trop forte avec le métier, nous avons conçu et développé une ontologie qui est en mesure de représenter la totalité de l'information véhiculée dans le texte quel que soit le métier. Elle se distingue nettement de l'ontologie POPCORN par le fait qu'elle est centrée « événement ». Cette ontologie, nommée MR4AP (*Meaning Application For Application Purposes)* permet de structurer la totalité de l'information (i.e. pas uniquement les informations d'intérêt pour un cas d'usage) d'un document et est robuste au multilinguisme [7]. MR4AP est accessible librement[1].

Dans la suite, nous présentons les expérimentations de structuration d'information en utilisant de part et d'autre les deux ontologies.

B. Quelles données utiliser ?

En raison du caractère extrêmement sensible de la donnée opérationnelle, le consortium de POPCORN n'a pas accès aux données réelles du cas d'usage. Pourtant, les données sont nécessaires pour le développement d'intelligences artificielles, au moins pour leur évaluation.

Concernant la tâche d'EEI, en particulier d'entités nommées (noms de personnes, d'organisations, de lieux, etc.), de nombreux jeux de données en français annotés ont été publiés. Certains sont commercialisés (par exemple ESTER[2]), d'autres sont inaccessibles (par exemple DAWT [25]) ou distribués à usage non commercial uniquement (par exemple WikiNeural [34]). Enfin, certains sont annotés avec les URI DBpedia mais pas directement avec les types d'entités (par exemple [36]) et sont plutôt destinés à une tâche de liage d'entités. Notons que des approches permettent de générer des jeux de données annotés en entités nommées « à la volée » en fonction de certains critères (par exemple GeNER [52]).

Nous avons finalement recensé onze jeux de données en français qui sont à la fois accessibles et libres d'utilisation (cf. Tab. I). Dix des onze jeux de données sont annotés avec un nombre de classes inférieur à quinze. Sur cet aspect, le jeu de données Wikipedia-ner [40] se distingue des autres puisqu'il contient 41 classes bien qu'il soit limité en taille (21 855 tokens). Il apparaît ainsi qu'au lancement du projet, il n'existait aucun grand jeu de données en français annoté en entités et en relations selon des ontologies représentant plusieurs dizaines de classes et de relations d'intérêts pour le renseignement militaire.

Suite à ces constats, dans le cadre du projet POPCORN, nous avons construit les jeux de données suivants :

- **DWIE-FR** : jeu de données en français développé à partir d'un jeu de données en anglais nommé DWIE [37]. La totalité du jeu de données ainsi que la méthode suivie pour son développement ont été publiés [2]. Ce jeu de

[1]https://github.com/Emvista/MR4AP/tree/main

[2]http://catalog.elra.info/product_info.php?products_id=999

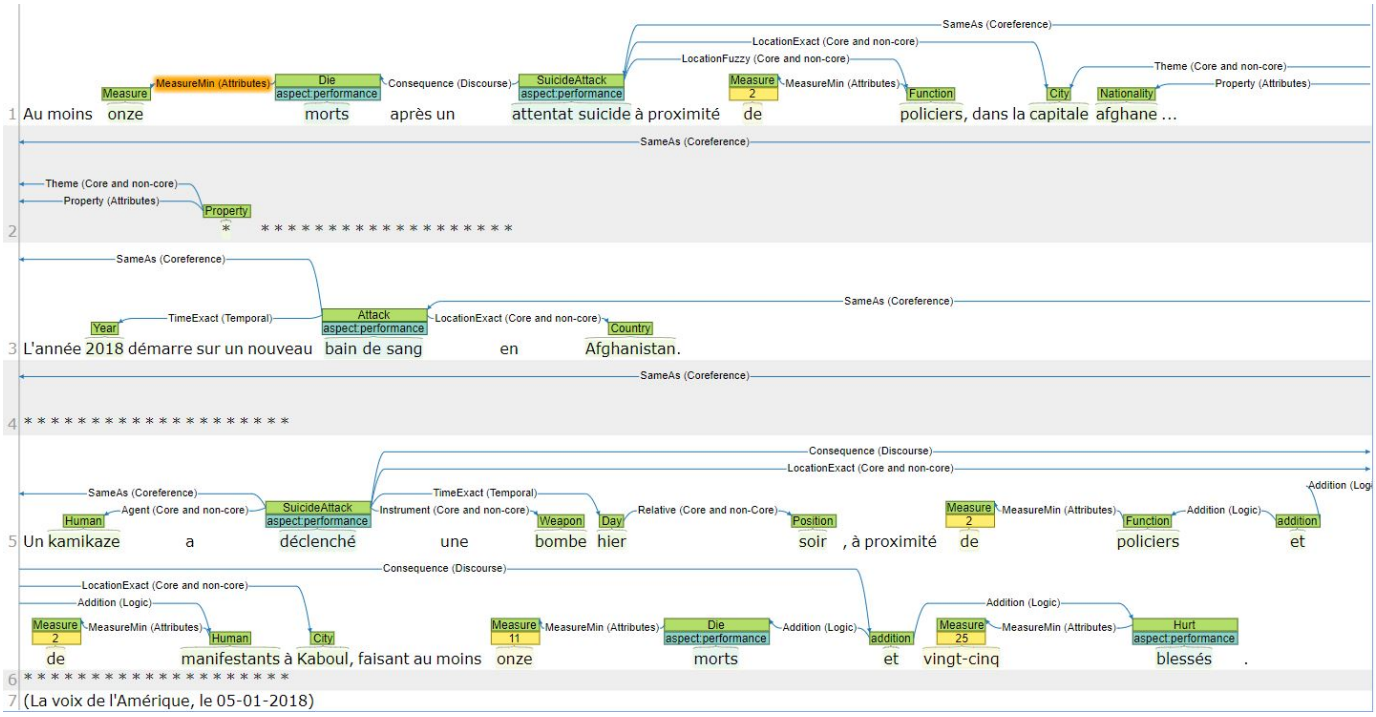

Fig. 1. Exemple d'un texte annoté manuellement avec l'ontologie MR4AP. Cette annotation comporte l'identification des entités d'intérêts (cadres verts) et des relations reliant ces entités (liens étiquetés). [7].

Entity Classes	Definition
Actor	Person or Organization
Organization	Administrative or functional structure. Remarks: An Organization is constituted to accomplish an aim, purpose, or mission
Government Organization	Organization controlled by a national or international government
Military Organization	Government Organization that is officially sanctioned and is trained and equipped to exert force
Non-Military Organization	Organization that controls and administers public policy either under a national or international mandate
Group of Individuals	Group of people gathered under a label for a specific purpose
Intergovernmental Organization	Organization conducted by two or more governments
Non-Governmental Organization	Organization that doesn't belong to the government
Person	Human being
Civilian	Person not in the armed services or the police force
Criminal	Person who violates the law or attempts to further their views by a system of coercive intimidation
Military	Person who belongs to a military force
Event	Routine, cyclical, planned, or spontaneous activities that significantly affect organizations, people, and military operations
Accident	Unfortunate event, especially one causing physical harm or damage, brought about unintentionally
CBRN Event	Event that involves chemical, biological, radiological or nuclear material individually or in combination, and are NOT attacks
Civil Unrest	Event that expresses dissatisfaction of citizens through disturbance and agitation, typically involving public demonstrations or disorder
Agitating Trouble Making	Stirring up of public interest on a matter of controversy, such as a political or social issue
Civil War Outbreak	Events related to a war among fellow citizens or within the limits of one community
Coup d'État	Violent or illegal seizure of power
Demonstration	Public meeting or march legally expressing protests, opinions or feelings towards a cause,

Fig. 2. Aperçu de l'ontologie POPCORN

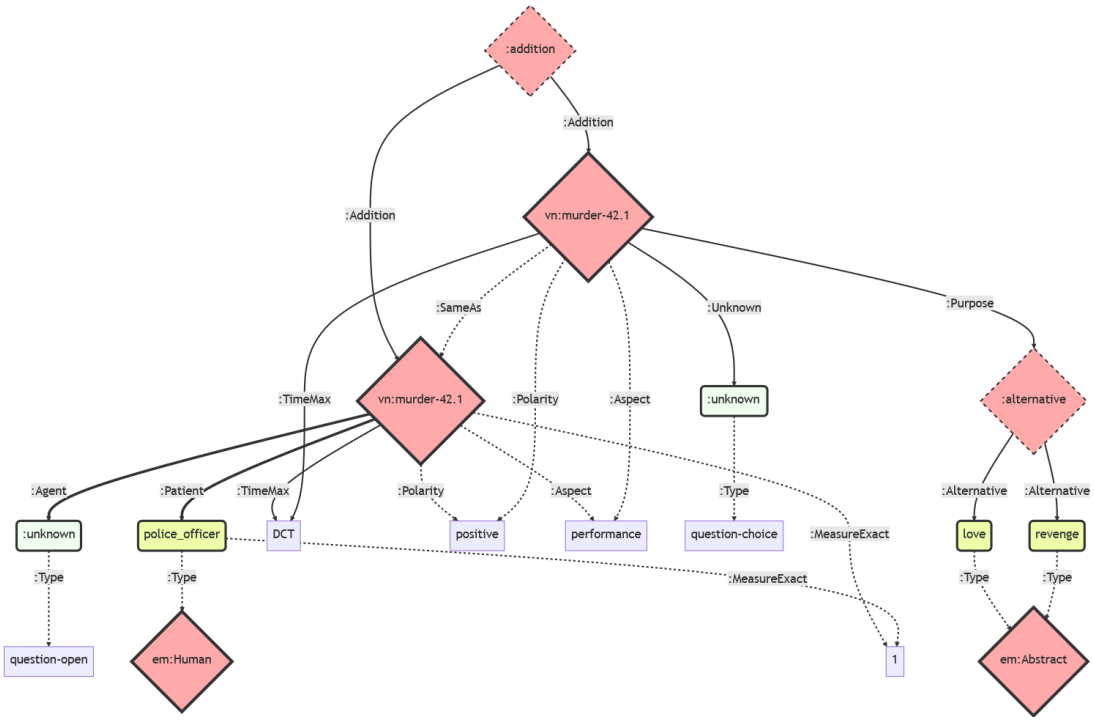

Fig. 3. Exemple de graphe MR4AP résultant de l'analyse du texte « Who committed the murder of that police officer, and was it for revenge or for love? » ; le même graphe MR4AP est obtenu pour le texte « Was this murder perpetrated out of revenge or out of love? And who killed that policeman? »

Jeu de données	Tokens	Annotations	Classes	Références
MultiNERD	4 300 000	279 300	15	[35]
WikiNeural	3 240 000	231 000	4	[34]
Le tour du monde en 80 jours	84 972	6 076	12	[40]
WiNER-Fr	322 931	24 144	7	[53]
Wikipedia-ner	21 855	6 132	41	[40]
CAP Twitter	env. 60 000	6 562	13	[38]
Europeana-newpapers-ner	207 000	13 860	3	[42]
Quaero French Medical Corpus	72 183	16 233	10	[43]
French TreeBank	350 931	11 636	7	[44]
WikiNer-Fr	3 499 695	420 061	4	[45]
UkraiNER	env. 400 000	43 623	8	[24]

TABLE I
JEUX DE DONNÉES EN FRANÇAIS, ANNOTÉS EN ENTITÉS ET ACCESSIBLES

données contient 589 394 tokens dont 60 292 annotés en entités d'intérêt selon 169 classes ;

- **POPCORN-data-manual** : rédigé et annoté en partie par un prestataire[3], ce jeu de données contient 2 000 textes simulant des bulletins d'information. Ce jeu de données a été annoté avec l'ontologie POPCORN. Une partie du jeu de données a été publié [26] ;
- **POPCORN-data-auto** : jeu de données annoté selon l'ontologie POPCORN en utilisant des IA génératives. Une partie du jeu de données a été publié [26] ;
- **Renseignor** : bulletins d'informations issus de Renseignor[4]. Ces données sont en cours d'annotation d'une part avec l'ontologie POPCORN et d'autre part avec l'ontologie MR4AP.

Concernant la tâche d'ER, il n'existe à notre connaissance aucun jeu de données en français accessible librement qui soit annoté dans un contexte sécurité défense. Parmi les jeux de données annotés en entités que nous avons produits, POPCORN-data-manual a également été annoté en relations, entièrement manuellement, et Renseignor est en cours d'annotation. Quelques centaines de textes déjà annotés permettent de publier les premiers résultats dans cet article (cf. section III-C).

À la fin du projet, POPCORN aura produit quatre jeux de données en français annotés avec plusieurs dizaines de classes d'entités d'intérêts pour le renseignement militaire, dont deux jeux de données annotés en relations d'intérêts. Notons que, bien que le jeu de données UkraiNER indiqué

[3]https://fr.isahit.com/
[4]https://cf2r.org/publications/lettre/renseignor/

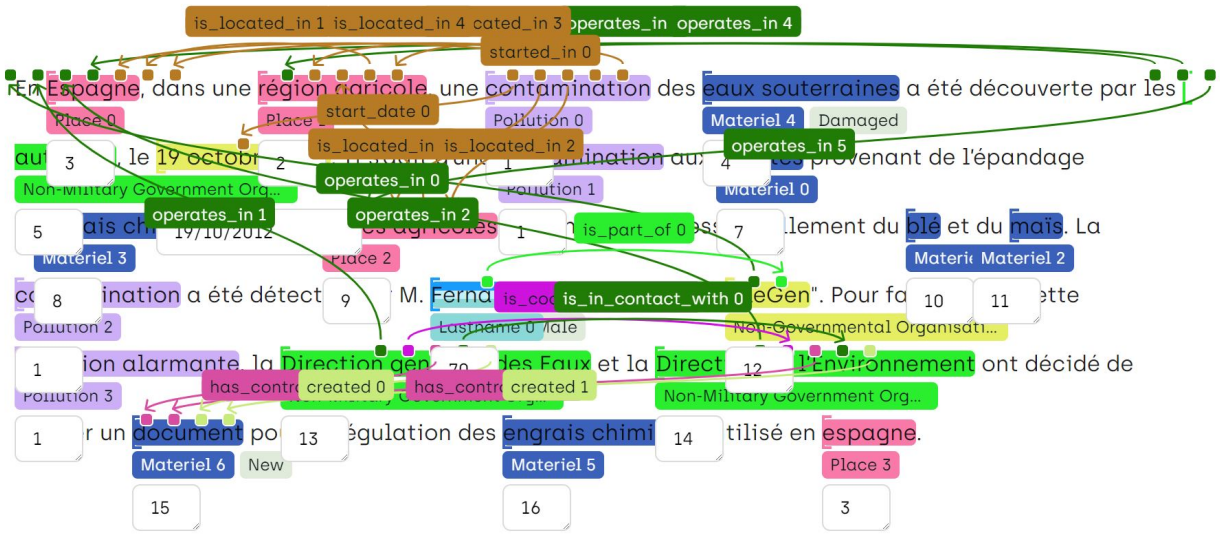

Fig. 4. Exemple d'un texte annoté manuellement issu de POPCORN-data-manual (texte 5_1470)

en section III soit annoté avec seulement 8 classes génériques (non spécifique à la sécurité / défense, la thématique du corpus en fait le quatrième jeu de données annoté connu jusqu'à maintenant (cf. II). À ce jour, les seuls jeux de données « sécurité/défense » accessibles qui sont annotés à la fois en entités et en relations sont les jeux de données POPCORN-data-manual et POPCORN-data-auto.

C. Quels modèles d'IA pour structurer et extraire l'information ?

Le troisième verrou qui doit être résolu dans POPCORN concerne les performances limitées des modèles supervisés pour l'extraction d'entités et de relations d'intérêt, en particulier lorsque l'on traite plusieurs dizaines de types d'entités et de relations. Le consortium de POPCORN a donc entreprit la reproduction et l'expérimentation de modèles à l'état de l'art, dans le but de les comprendre et de comparer leurs performances sur les jeux de données issus du consortium.

Les premières expérimentations ont porté sur des modèles d'EEI « classiques » reposant sur des modèles de langage francophones largement utilisés pour leur capacité à vectoriser l'information dans les documents, tels que CamemBERT [28] et FlauBERT [27], augmentés d'une couche linéaire de classification pour l'étiquetage de mots.

Les préoccupations récentes en extraction d'entités, telles que l'attention portée à l'extraction de fragments plutôt qu'à l'étiquetage de tokens, la détection d'entités imbriquées et la prédiction de multiples labels pour une même mention, nous ont conduit à explorer des modèles plus récents, notamment DeepSpanRepresentations [29], Biaffine-NER [30], Boundary Smoothing (BS) [8] et PromptNER [31], des modèles représentant efficacement les fragments textuels par des opérations bi-affine, des Bi-LSTM ou une meilleure gestion de l'ambiguïté liée aux frontières des mentions. Par ailleurs des modèles d'extraction de relations à l'état de l'art tels que DREEAM [48], ATLOP [51], PEMSCL [49] et KD-DocRE [50] ont aussi été expérimentés. En parallèle, nous avons également développé notre propre modèle résolvant les différentes modalités, le modèle unifié [23].

Le développement du jeu de données POPCORN-data-auto (cf. section III-B) a été rendu possible grâce à l'utilisation d'un grand modèle de langage génératif (LLM), Vigostral-7B[5], un modèle conversationnel en français librement accessible et affiné à partir de Mistral-7B [9].

Le développement du jeu de données POPCORN-data-auto (cf. section III-B) a nécessité l'expérimentation de processus sophistiqués impliquant des grands modèles de langage génératif (LLM) tels que Vigostral-7B[6], un modèle conversationnel en français affiné à partir de Mistral-7B [9].

Les expériences menées avec les modèles mentionnés, leurs résultats et les analyses qui en découlent sont présentés dans nos contributions [23] et [26]. Nous reportons ici quelques résultats obtenus sur l'ontologie POPCORN avec des modèles entraînés sur peu de données en comparaison avec les données dont dispose le consortium : 400 textes issus de POPCORN-data-manuel enrichis de 400 textes de POPCORN-data-auto (cf. Tableau III). Notons que ce rapport de 1 pour 1 (400 textes rédigés et annotés manuellement pour 400 produits de façon automatique) a apporté les meilleurs résultats à ce stade. Ces 800 textes sont librement accessibles afin d'assurer la reproductibilité de nos expériences par des tiers.

D. Quelles performances ?

1) Avec l'ontologie POPCORN: Les modèles traitant de façon conjointe les deux tâches (noté « Modèle joint » dans le Tableau III) sont légèrement plus performants, ce qui a par

[5]https://huggingface.co/bofenghuang/vigostral-7b-chat
[6]https://huggingface.co/bofenghuang/vigostral-7b-chat

Jeu de données	Tokens	Entités annotées	Relations annotées	Classes d'entités	Classes de relations	Références
DWIE-FR	589 394	60 292	0	169	0	[2]
POPCORN	114 469	43 264	37 065	55	36	[26]
Renseignor	en cours	en cours	en cours	70	39	à paraître

TABLE II
JEUX DE DONNÉES ACCESSIBLES OU PROCHAINEMENT ACCESSIBLES, EN FRANÇAIS, ANNOTÉS AVEC DES CLASSES MÉTIERS « SÉCURITÉ/DÉFENSE ».

ailleurs encouragé l'annotation des jeux de données à la fois en entités et en relations. Il apparaît que les meilleurs résultats sont de l'ordre de 81% de F1 Micro pour la reconnaissance d'entités d'intérêt et de 59.01% pour la reconnaissance des événements. On remarque une difficulté à extraire les entités de type « événements » bien que la tâche semble très proche de la reconnaissance d'entités.

Dans le cadre de cette évaluation, les modèles d'extraction de relations ont pris en entrée le texte déjà annoté (manuellement) en entités d'intérêts ; les résultats présentés ici correspondent donc uniquement à la tâche de reconnaissance de relations et d'étiquetage. L'extraction de relations atteint un F1 Micro de 59.38% avec le modèle ATLOP contraint par les restrictions sémantiques sur les domaines (i.e. tête) et co-domaines (i.e. queue des relations).

Les évaluations menées dans le cadre de POPCORN ont généralement montré une faiblesse de ces modèles concernant la détection des frontières gauche et droite des entités d'intérêt. Une augmentation jusqu'à 10% sur le score F1 est observée lorsque l'on fait l'hypothèse d'une identification parfaite des frontières[7]. Dans ce sens, des expériences sont en cours : dans un premier temps, nous avons choisi de remplacer le module Bi-LSTM de l'architecture Biaffine-ner par un GCN (Graph Convolutional Networks). Nous aimerions savoir si l'ajout d'une représentation « graphe » est intéressante pour la détection de frontières.

2) Avec l'ontologie MR4AP: La tâche de structuration de l'information selon des formalismes de représentation sémantique tels que AMR [32] ou MR4AP [7] est complexe. De nombreuses recherches sont effectuées à ce sujet. Le premier modèle accessible permettant de transformer un texte en français en un graphe régi par un tel formalisme a été publié par l'un d'entre nous en 2023 [33]. Les résultats obtenus avec les méthodes à l'état de l'art (LLM, apprentissage par transfert, ou encore méta-apprentissage) ne permettent pas pour l'instant de dépasser des méthodes plus classiques fondées sur l'idée d'un système hybride (i.e. utilisant à la fois des techniques d'apprentissage et de règles linguistiques). La raison est certainement liée au manque de données annotées. Cela est illustré en tableau IV dans lequel le système hybride RLA (*Recursive Linguistic Analyzer*) dont les prémisses ont été publiées en 2019 [41], obtient les meilleurs résultats notamment en termes de précision probablement grâce aux règles linguistiques qui guident les modèles sous-jacents. À noter qu'aucune adaptation n'a été apportée au RLA dans le cadre de cette évaluation. L'apport de connaissances issues du jeu d'apprentissage (par exemple, des listes de noms d'organisations) améliorerait sans aucun doute les résultats. Il faut noter que les résultats mettent en évidence que la structuration selon MR4AP est plus complexe qu'avec l'ontologie métier POPCORN : sur des tâches identiques, les modèles communs appliqués aux deux ontologies obtiennent des résultats inférieurs au RLA d'au moins 13 points sur les scores F1 Macro et F1 Micro.

Dans le cadre de cette évaluation, comme dans le cadre des expériences avec l'ontologie POPCORN, les modèles d'extraction d'information ont pris en entrée le texte déjà annoté (manuellement) en entités d'intérêt ; au contraire le système RLA a directement traité le texte brut en entrée ce qui complique la tâche. Néanmoins, la précision haute (88.09%) permet d'envisager l'intégration de ce système qui a pour vocation de structurer 100% des informations véhiculées dans le texte dans une application où moins de 12% des résultats devraient être corrigés par un utilisateur. En revanche, le rappel est très faible (environ 25%) ce qui pourrait être rattrapé par la prise en compte des résultats d'autres modèles.

IV. INTÉGRATION DES RÉSULTATS

Dans cette section, nous présentons succinctement deux démonstrateurs, l'un utilisant les modèles d'EEI et d'ER selon l'ontologie POPCORN et l'autre selon l'ontologie MR4AP. Cette dernière n'étant pas une ontologie métier, nous donnons l'intuition du processus qui permet à un utilisateur de l'adapter rapidement à son cas d'usage.

A. Avec l'ontologie POPCORN

L'adaptation d'outils d'extraction d'information aux métiers du renseignement a montré l'importance de la sélection des connaissances capitalisées en base et du maintien de leur cohérence. Les résultats exposés précédemment montrent que l'amélioration de la qualité de l'extraction d'information à l'état de l'art ne suffit pas encore à garantir cette cohérence, notamment à cause de l'accumulation des erreurs en cas de capitalisation automatique sans vérification ni modification humaine.

Dans le cadre de POPCORN, un démonstrateur fonctionnel permet d'évaluer la difficulté du rôle de l'humain dans la chaîne globale de peuplement de bases de connaissances. Il est illustré en figure 5. Ici, l'humain est le validateur final des éléments extraits par l'IA. La réalisation du démonstrateur a requis de lister les modifications possibles par l'humain et de décider de la manière la plus pertinente de les lui proposer : au niveau des entités ou de leurs mentions, dans le texte ou dans le graphe. Certaines modifications n'ont

[7]Cette observation est d'importance lorsque l'on considère un module de liage des mentions du texte avec les entités de la base qui se situent en aval des modèles étudiés ici.

Tâche	Modèle	Précision	Rappel	F1 Macro	F1 Micro
Extraction des événements	Modèle joint	**52.87**	43.21	**44.79**	**59.38**
	Biaffine-NER + Boundary Smoothing	41.13	**49.16**	43.92	55.94
	Camembert-base(p.e. Wikikner) + BiLSTM + CRF	46.84	44.76	44.02	57.80
EEI sauf événements	Modèle joint + contraintes domaines	**76.23**	**67.01**	**69.23**	**82.31**
	Biaffine-NER + Boundary Smoothing	62.65	66.14	65.82	81.14
	Camembert-base(p.e. Wikikner) + BiLSTM + CRF	72.18	65.48	66.67	81.75
ER	Modèle joint + contraintes domaines	54.73	49.88	47.78	58.02
	Dreeam + contraintes domaines	**72.11**	47.65	53.48	59.24
	ATLOP + contraintes domaines	68.35	**50.27**	**54.33**	**59.38**

TABLE III
RÉSULTATS OBTENUS PAR LES MODÈLES EXPÉRIMENTÉS SUR L'ONTOLOGIE POPCORN ET UN JEU DE DONNÉES CONSTITUÉ DE 400 TEXTES ISSUS DE POPCORN-DATA-MANUEL ENRICHI DE 400 TEXTES SYNTHÉTIQUES ISSUS DE POPCORN-DATA-AUTO

Tâche	Modèle/Système	Précision	Rappel	F1 Macro	F1 Micro
Extraction des événements	Biaffine-NER + BS (p.e. WikiNer-FR)	11.18	10.49	9.90	34.45
	Camembert-base (p.e. WikiNer) + BiLSTM + CRF	13.79	12.35	12.07	41.80
	RLA	**53.98**	**55.41**	**57.05**	**59.69**
EEI sauf événements	Biaffine-NER + BS (p.e. WikiNer-FR)	20.87	21.30	19.95	**58.78**
	Camembert-base (p.e. WikiNer) + BiLSTM + CRF	28.74	25.97	25.19	**62.38**
	RLA	**69.99**	**58.11**	**55.17**	55.24
ER	ATLOP	19.41	14.25	15.84	**66.82**
	RLA	**88.09**	**25.75**	**34.90**	39.85

TABLE IV
RÉSULTATS OBTENUS PAR LES MODÈLES ET LE SYSTÈME EXPÉRIMENTÉS SUR LES DONNÉES RENSEIGNOR, WIKIPEDIA, WIKINEWS ANNOTÉES AVEC L'ONTOLOGIE MR4AP (EN COURS D'ANNOTATION ; EXPÉRIENCES MENÉES AVEC 19K *tokens* ANNOTÉS DONT 1835 MENTIONS D'ÉVÉNEMENTS ; P = PRÉCISION ; R = RAPPEL. L'ABBRÉVIATION "P.E." SIGNIFIE "PRÉ-ENTRAÎNÉ SUR".

pas nécessairement à être accessibles depuis chaque vue. Ainsi, l'utilisateur est capable de modifier toute information détectée (relations, attributs, mentions textuelles de ces entités et leurs coréférences, entités de référence en base), et d'ajouter de nouveaux éléments non détectés. Côté graphe, l'interface utilise les couleurs et icônes ainsi que les épaisseurs de traits pour expliciter les natures et provenances des informations. Une fois que les résultats de l'IA ont été revus par l'humain, la base de connaissances est mise à jour et les prochaines analyses bénéficieront de l'ensemble du contenu de la base.

B. Avec l'ontologie MR4AP

L'utilisation de l'ontologie MR4AP a pour intérêt de ne pas restreindre l'apprentissage de l'IA à un seul cas d'usage. Techniquement, MR4AP est une ontologie qui contient des centaines de concepts et seulement 44 relations qui permettent de représenter la totalité de l'information véhiculée dans un texte. Nous avons développé une interface qui permet à un utilisateur d'aligner les concepts et relations génériques de MR4AP avec les siens. Ainsi les informations visualisées sur l'interface adoptent le vocabulaire métier nécessaire aux opérationnels.

La figure 6 montre un bulletin d'information à partir duquel un événement MR4AP de type « Murder » a été détecté. Des informations relatives à cet événement ont été détectées : la date (normalisée afin d'être intégrée en base de données), l'agent (celui qui fait l'action), le patient (celui qui subit l'action), le lieu. Ces informations sont liées à une base de données (DBpedia pour ce démonstrateur) lorsque cela est rendu possible. Dans l'exemple, la fiche contenant des informations sur le lieu de l'événement peut être visualisée. Toutes les informations peuvent être éditées avant leur intégration dans la base de connaissances.

V. CONCLUSION

Le projet POPCORN a contribué aux méthodes scientifiques de structuration d'information pour des cas d'usage sécurité/défense. De nombreuses expériences ont été effectuées pour extraire les entités d'intérêts et les relations qu'elles entretiennent entre elles. Les résultats obtenus permettent d'envisager l'insertion de certains types d'entités et relations en bases de données avec leurs indices de confiance et les sources desquelles elles proviennent afin d'être vérifiées par un analyste qui en aurait l'utilité. Pour un cas d'usage où il s'agirait de peupler les bases de données avec les informations extraites, il est aujourd'hui encore nécessaire que l'humain valide certaines informations extraites automatiquement par l'IA, notamment pour les types d'informations dont les résultats sont trop faibles.

Le système global répondant au cas d'usage est en cours de développement : à partir d'une évaluation plus fine que celle présentée ici, nous avons pu déterminer quel type d'information est le mieux inféré pour chaque modèle. Même si de façon globale les types obtiennent des résultats similaires, certains modèles sont plus robustes aux faibles quantités d'exemples présents dans leur jeu d'entraînement.

L'étude des données synthétiques est prometteuse pour la suite de nos recherches. Ces données résultent en une augmentation significative des performances qualitatives des modèles. La qualité des données générées automatiquement

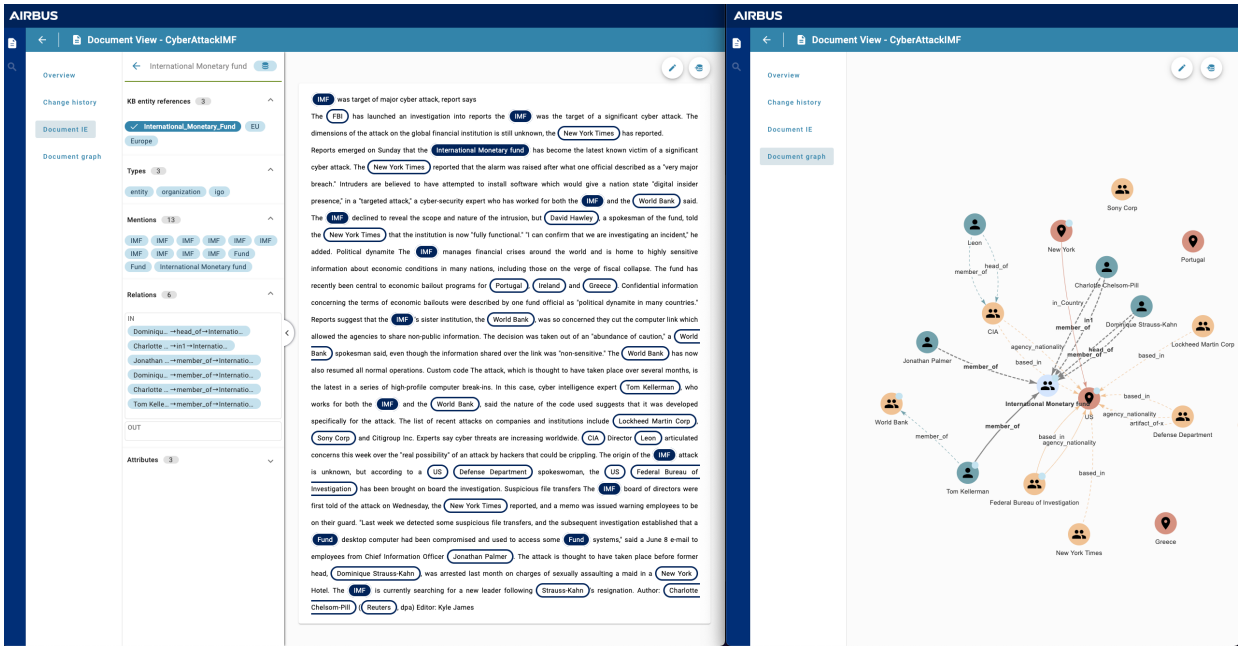

Fig. 5. Aperçu du démonstrateur POPCORN : vues synchronisées du texte et des connaissances extraites.

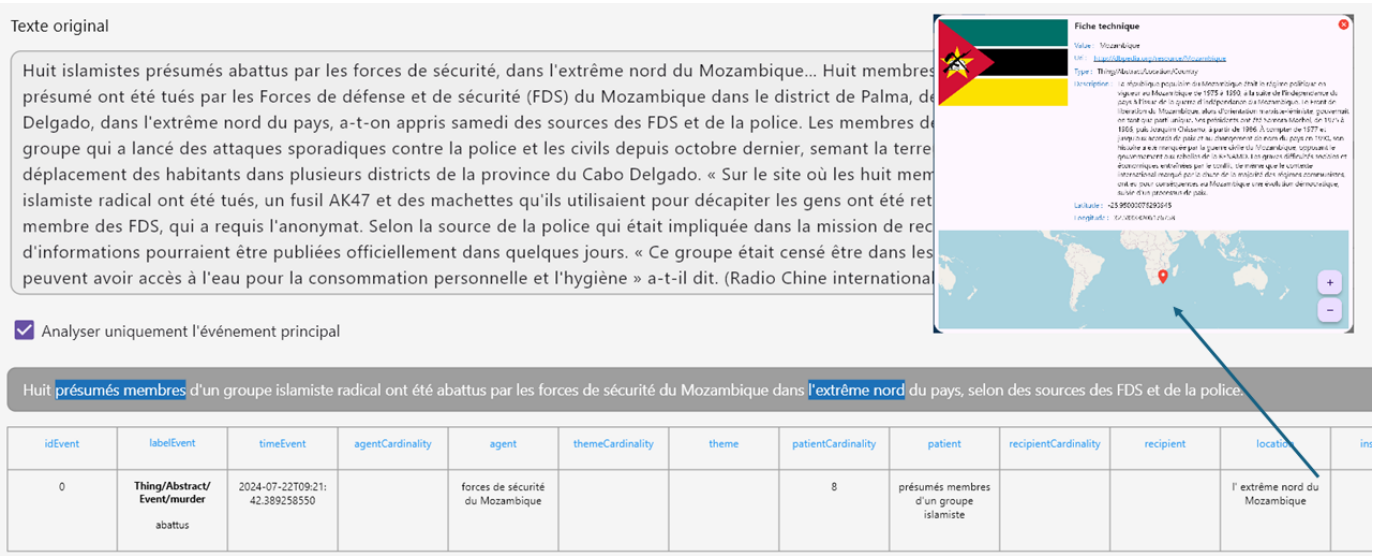

Fig. 6. Exemple d'un texte annoté automatiquement issu de Renseignor. Image avant l'étape de configuration "métier". Le tableau contient une ligne générée automatiquement qui représente un événement "Murder", sa date, les personnes impliquées, son lieu. La fiche technique contient des informations de la base de connaissance associées au lieu identifié.

fait actuellement l'objet de recherches et des données de meilleure qualité à venir dans les prochains mois permettront de produire de nouvelles versions des modèles.

L'apparition des LLM conversationnels au cours du projet POPCORN a permis de les considérer sérieusement comme une alternative aux modèles déterministes avec l'avantage qu'ils ne nécessitent pas (ou peu) de données annotées. Néanmoins, les LLM à eux seuls s'avèrent peu performants pour structurer l'information selon des ontologies complexes (cf. EvalLLM[8]). Une voie de prolongement de ces travaux consisterait à tester des modèles de taille bien supérieure (d'un facteur 10 ou 100) et/ou des modèles plus centrés sur le français afin de déterminer si nos méthodes de création de données synthétiques résultent en un gain qualitatif dans ces contextes.

[8]Atelier sur l'évaluation des modèles génératifs (LLM) et challenge d'extraction d'information few-shot : https://evalllm2024.sciencesconf.org/program?lang=fr

REFERENCES

[1] Zaporojets, Klim, Deleu, Johannes, Develder, Chris and Demeester, Thomas. (2021) "DWIE: An entity-centric dataset for multi-task document-level information extraction". *Information Processing and Management*. **58**(4): 102563

[2] Verdy, S., Prieur, M., Gadek, G. & Lopez, C. DWIE-FR : Un nouveau jeu de données en français annoté en entités nommées. *Actes De CORIA-TALN 2023. Actes De La 30e Conférence Sur Le Traitement Automatique Des Langues Naturelles, TALN 2023 - Volume 2 : Travaux De Recherche Originaux - Articles Courts, Paris, France, June 5-9, 2023*. pp. 63-72 (2023)

[3] Sang, E. and Meulder, F. (2003) "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". *Proceedings Of The Seventh Conference On Natural Language Learning, CoNLL 2003, Held In Cooperation With HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*. pp. 142-147.

[4] Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S. & Grilheres, B. Extraction et agrégation automatique d'événements pour la veille en sources ouvertes: du texte à la connaissance. (2013) *IC-24èmes Journées Francophones D'Ingénierie Des Connaissances*.

[5] Prieur, M., Mouza, C., Gadek, G. & Grilhères, B. Evaluating and Improving End-to-End Systems for Knowledge Base Population. *Proceedings Of The 15th International Conference On Agents And Artificial Intelligence, ICAART 2023, Volume 3, Lisbon, Portugal, February 22-24, 2023*. pp. 641-649 (2023)

[6] Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J. & Sun, M. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. *Proceedings Of The 57th Conference Of The Association For Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. pp. 764-777 (2019)

[7] Giordano, B. & Lopez, C. (2023). MR4AP: Meaning representation for application purposes. In Proceedings of the Fourth International Workshop on Designing Meaning Representations, pp. 110-121.

[8] Zhu, Li (2022) "Boundary Smoothing for Named Entity Recognition" *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*

[9] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E. (2023). Mistral 7B. ArXiv, abs/2310.06825.

[10] Sosuke Kobayashi. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 452–457

[11] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6382–6388

[12] Zhang, Danqing & Li, Tao & Zhang, Haiyang & Yin, Bing. (2020). On Data Augmentation for Extreme Multi-label Classification.

[13] George A. Miller. 1994. WordNet: A Lexical Database for English. In Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

[14] Sérasset, Gilles. 'DBnary: Wiktionary as a Lemon-based Multilingual Lexical Resource in RDF'. 1 Jan. 2015 : 355 – 361.

[15] Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

[16] Jeffrey Pennington, Richard Socher, and Christopher Manning. (2014) GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543

[17] Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., & Ng, A. Y. (2019). Data noising as smoothing in neural network language models. Paper presented at 5th International Conference on Learning Representations, ICLR 2017, Toulon, France.

[18] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 704–717

[19] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances In Neural Information Processing Systems 35: Annual Conference On Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. (2022)

[20] Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2015. Improving distant supervision using inference learning. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 273–278, Beijing, China. ACL.

[21] Xiang Deng and Huan Sun. 2019. Leveraging 2-hop Distant Supervision from Table Entity Pairs for Relation Extraction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 410–420, Hong Kong, China. Association for Computational Linguistics.

[22] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E.H., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. ArXiv, abs/2203.11171.

[23] Prieur, M., Mouza, C., Gadek, G. & Grilhères, B. Shadowfax: Harnessing Textual Knowledge Base Population. *Proceedings Of The 47th International ACM SIGIR Conference On Research And Development In Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*. pp. 2796-2800 (2024), https://doi.org/10.1145/3626772.3657666

[24] Lauriane Aufrant, Lucie Chasseur (2024). UkraiNER: A New Corpus and Annotation Scheme towards Comprehensive Entity Recognition. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) pp. 16941-16952.

[25] Spasojevic, N., Bhargava, P., & Hu, G. (2017). Dawt: Densely annotated wikipedia texts across multiple languages. In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1655-1662.

[26] Bastien Giordano, Maxime Prieur, Nakanyseth Vuth, Sylvain Verdy, Kévin Cousot, Gilles Sérasset, Guillaume Gadek, Didier Schwab, Cédric Lopez (2024) POPCORN: Fictional and Synthetic Intelligence Reports for Named Entity Recognition and Relation Extraction Tasks. In Proceedings of KES, Sevilla, Spain, to appear.

[27] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., ... & Schwab, D. (2020, June). FlauBERT: des modèles de langue contextualisés pré-entraînés pour le français. In 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles (pp. 268-278). ATALA; AFCP.

[28] Martin, L., Muller, B., Suárez, P., Dupont, Y., Romary, L., Clergerie, Seddah, D. & Sagot, B. CamemBERT: a Tasty French Language Model. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. pp. 7203-7219 (2020), https://doi.org/10.18653/v1/2020.acl-main.645

[29] Zhu, E., Liu, Y., & Li, J. (2022). Deep span representations for named entity recognition. arXiv preprint arXiv:2210.04182.

[30] Yu, J., Bohnet, B. & Poesio, M. Named Entity Recognition as Dependency Parsing. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. pp. 6470-6476 (2020), https://doi.org/10.18653/v1/2020.acl-main.577

[31] Shen, Y., Tan, Z., Wu, S., Zhang, W., Zhang, R., Xi, Y., Lu, W. & Zhuang, Y. PromptNER: Prompt Locating and Typing for Named Entity Recognition. *Proceedings Of The 61st Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. pp. 12492-12507 (2023), https://doi.org/10.18653/v1/2023.acl-long.698

[32] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... & Schneider, N. (2013, August). Abstract meaning representation for sembanking. In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse (pp. 178-186).

[33] Kang, J., Coavoux, M., Lopez, C., & Schwab, D. (2023). Analyse sémantique AMR pour le français par transfert translingue. In 18e

Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 55-62). ATALA.

[34] Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., & Navigli, R. (2021, November). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the association for computational linguistics: EMNLP 2021 (pp. 2521-2533).

[35] Tedeschi, S., & Navigli, R. (2022, July). MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 801-812).

[36] Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating NLP using linked data. In The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12 (pp. 98-113). Springer Berlin Heidelberg.

[37] Zaporojets, K., Deleu, J., Develder, C. & Demeester, T. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*. **58**, 102563 (2021)

[38] Lopez, C., Partalas, I., Balikas, G., Derbas, N., Martin, A., Reutenauer, C., ... & Amini, M. R. (2017). Cap 2017 challenge: Twitter named entity recognition. arXiv preprint arXiv:1707.07568.

[39] Dupont, Y. (2019). Un corpus libre, évolutif et versionné en entités nommées du français. In TALN 2019-Traitement Automatique des Langues Naturelles.

[40] Lopez, C., Mekaoui, M., Aubry, K., Bort, J., & Garnier, P. (2019, January). Reconnaissance d'entités nommées itérative sur une structure en dépendances syntaxiques avec l'ontologie nerd. In Extraction et Gestion des Connaissances: Actes de la conférence EGC (Vol. 79, pp. 81-92).

[41] Lopez, C., Mekaoui, M., Aubry, K., Bort, J., & Garnier, P. (2019, January). Reconnaissance d'entités nommées itérative sur une structure en dépendances syntaxiques avec l'ontologie nerd. In Extraction et Gestion des Connaissances: Actes de la conférence EGC (Vol. 79, pp. 81-92).

[42] Neudecker, C. (2016, May). An open corpus for named entity recognition in historic newspapers. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4348-4352).

[43] Grouin, A. N. C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014). The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.

[44] Sagot, B., Richard, M., & Stern, R. (2012). Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In Traitement Automatique des Langues Naturelles (TALN) (Vol. 2).

[45] Nothman, J., Curran, J. R., & Murphy, T. (2008, December). Transforming Wikipedia into named entity training data. In Proceedings of the Australasian Language Technology Association Workshop 2008 (pp. 124-132).

[46] Serrano, L. Vers une capitalisation des connaissances orientée utilisateur: extraction et structuration automatiques de l'information issue de sources ouvertes. (Universté de Caen,2014)

[47] Gerz, M., Mulikita, M., Bau, N. & Gökgöz, F. The MIP information model-a semantic reference for command & control. *2015 International Conference On Military Communications And Information Systems (ICMCIS)*. pp. 1-11 (2015)

[48] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering.

[49] Guo, J., Kok, S., & Bing, L. (2023). Towards integration of discriminability and robustness for document-level relation extraction. arXiv preprint arXiv:2304.00824.

[50] Tan, Q., He, R., Bing, L., & Ng, H. T. (2022). Document-level relation extraction with adaptive focal loss and knowledge distillation. arXiv preprint arXiv:2203.10900.

[51] Zhou, W., Huang, K., Ma, T., & Huang, J. (2021). Document-level relation extraction with adaptive thresholding and localized context pooling. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 16, pp. 14612-14620).

[52] Kim, H., Yoo, J., Yoon, S., Lee, J., & Kang, J. (2021). Simple questions generate named entity recognition datasets. arXiv preprint arXiv:2112.08808.

[53] Dupont, Y. (2019). Un corpus libre, évolutif et versionné en entités nommées du français. In TALN 2019 - Traitement Automatique des Langues Naturelles.