DWIE-FR : Un nouveau jeu de données en français annoté en entités nommées

Sylvain Verdy[1] Maxime Prieur[2, 3] Guillaume Gadek[3] Cédric Lopez[1]
(1) Emvista, 10 rue Louis Breguet, 34830 Jacou, France
(2) CNAM, 292 rue Saint-Martin, 75003 Paris, France
(3) Airbus Defence and Space, 1 Bd Jean Moulin, 78990 Elancourt, France
sylvain.verdy@emvista.com, maxime.prieur.auditeur@lecnam.net,
cedric.lopez@emvista.com, guillaume.gadek@airbus.com

RÉSUMÉ
Ces dernières années, les contributions majeures qui ont eu lieu en apprentissage automatique supervisé ont mis en evidence la nécessité de disposer de grands jeux de données annotés de haute qualité. Les recherches menées sur la tâche de reconnaissance d'entités nommées dans des textes en français font face à l'absence de jeux de données annotés "à grande échelle" et avec de nombreuses classes d'entités hiérarchisées. Dans cet article, nous proposons une approche pour obtenir un tel jeu de données qui s'appuie sur des étapes de traduction puis d'annotation des données textuelles en anglais vers une langue cible (ici au français). Nous évaluons la qualité de l'approche proposée et mesurons les performances de quelques modèles d'apprentissage automatique sur ces données.

ABSTRACT
DWIE-FR : A new French dataset annotated in named entities

In the recent years, major contributions have been made in the field of supervised machine learning, which increasingly empathize the need for large-scale high-quality annotated datasets. Research on the french named entity recognition task faces the lack of large-scale annotated datasets with many hierarchical entity classes. In this paper, we present an approach to obtain a dataset that relies on translation and annotation steps from English to a target language (French in our study). We evaluate the quality of this alignment and measure the performances obtained by machine learning models on an aligned dataset.

MOTS-CLÉS : TAL, reconnaissance d'entités nommées, jeu de données, traduction, alignement.

KEYWORDS: NLP, named entity recognition, dataset, translation, alignment.

1 Introduction

Les technologies d'apprentissage automatique ont connu une forte accélération et ont montré de nettes améliorations sur différentes tâches de compréhension des langues naturelles. Alors que l'effort est principalement porté sur les algorithmes d'apprentissage, les jeux de données demeurent rares pour le français. C'est notamment le cas pour la tâche de reconnaissance d'entités nommées (REN).

Les entités nommées, définies comme « toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus » (Ehrmann, 2008), et la reconnaissance de ces

entités ont fait l'objet de nombreuses études et de campagnes d'évaluation ((Galliano *et al.*, 2009), CLEF (Ehrmann *et al.*, 2020), ETAPE (Galibert *et al.*, 2014), QUAERO (Rosset *et al.*, 2011)). La tâche consiste généralement à repérer et à classer les tokens automatiquement selon une typologie prédéfinie. Les typologies proposées reprennent généralement le triptyque « Personne », « Lieu », « Organisation » ou sont spécifiques à un domaine (par exemple la santé) et ne participent pas au développement de systèmes d'analyse de texte tous domaines confondus. Les expériences réalisées sur cette tâche sont donc limitées par l'absence de jeux de données en français qui soient à la fois de grande taille et annotés avec une vaste typologie.

Des scores excellents (car comparables à un annotateur humain) sont désormais obtenus sur l'anglais (Ye *et al.* 2022 obtient un micro F1 à 91.4 sur OntoNotes 5.0, un jeu contenant 18 classes différentes), soutenus par la disponibilité de jeux de données annotés, tels que FewNerd (Ding *et al.*, 2021), Conll-2003/2005 (Sang & De Meulder, 2003) ou encore DWIE (Zaporojets *et al.*, 2021). Sur le français, les F1-scores atteignent des résultats de l'ordre de 85.7 (Bannour *et al.*, 2022) pour un nombre de classes limité. Créer de nouveaux jeux de données se différenciant des précédents sur un certain nombre de critères devrait permettre d'améliorer les modèles.

Dans cet article, nous montrons d'abord à travers le recensement des jeux de données français (section 2) qu'un jeu de données volumineux et annoté avec de nombreuses classes manque au panorama des jeux de données. Nous proposons une traduction automatisée et maîtrisée du corpus anglophone DWIE (Zaporojets *et al.*, 2021), qui préserve les types d'entités (section 3) dont la qualité est maîtrisée et rend possible l'entraînement de modèles pour la reconnaissance d'entités nommées (section 4). Nous montrons la capacité d'apprentissage de modèles sur ce corpus et fournissons des métriques de qualité. Nous espérons que la mise à disposition de ce jeu de données permettra d'améliorer la qualité des systèmes francophones à court terme.

2 Travaux antérieurs

De nombreux jeux de données en français annotés en entités nommées ont été publiés. Certains sont commercialisés (par exemple ESTER [1]), d'autres sont inaccessibles (par exemple DAWT (Spasojevic *et al.*, 2017)) ou distribués à usage non commercial uniquement (par exemple WikiNeural (Tedeschi *et al.*, 2021a)). Enfin, certains sont annotés avec les URI DBpedia mais pas directement avec les types d'entités (par exemple (Hellmann *et al.*, 2013)) et sont plutôt destinés à une tâche de liage d'entités. Notons finalement 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 (Kim *et al.*, 2021)). Nous avons finalement recensé neuf jeux de données en français qui sont à la fois accessibles et libres d'utilisation (cf. Tab. 1). Huit des neuf jeux de données sont annotés avec un nombre de classes compris entre trois et quinze. Sur cet aspect, le jeu de données Wikipedia-ner (Lopez *et al.*, 2019) se distingue des autres puisqu'il contient 41 classes bien qu'il soit limité en taille (21855 tokens). Il apparaît ainsi qu'il n'existe aucun grand jeu de données en français annoté avec plusieurs dizaines de classes alors que de tels jeux de données existent pour l'anglais notamment, par exemple FewNerd (Ding *et al.*, 2021) ou encore DWIE (Zaporojets *et al.*, 2021). L'objectif de ce travail est de proposer un tel jeu de données (cf. section 3).

1. cf. http://catalog.elra.info/product_info.php?products_id=999

Nom	Tokens	Annotations	Classes	Références
MultiNERD	4 300 000	279 300	15	(Tedeschi & Navigli, 2022)
WikiNeural	3 240 000	231 000	4	(Tedeschi *et al.*, 2021b)
Le tour du monde en 80 jours	84 972	6 076	12	(Lopez *et al.*, 2019)
WiNER-Fr	322 931	24 144	7	(Dupont, 2019)
Wikipedia-ner	21 855	6 132	41	(Lopez *et al.*, 2019)
CAP Twitter	env. 60 000	6 562	13	(Lopez *et al.*, 2017)
Europeana-newpapers-ner	207 000	13 860	3	(Neudecker, 2016)
Quaero French Medical Corpus	72 183	16 233	10	(Névéol *et al.*, 2014)
French TreeBank	350 931	11 636	7	(Sagot *et al.*, 2012)
WikiNer-Fr	3 499 695	420 061	4	(Nothman *et al.*, 2008)

TABLE 1 – Jeux de données annotés en entités nommées pour le français

3 Création du jeu de données

3.1 Choix du jeu de données anglais

Le choix du jeu de données a été réalisé selon plusieurs indicateurs. Ces indicateurs ont été identifiés pour répondre à plusieurs de nos contraintes, à savoir le nombre de classes, le nombre de tokens et d'entités suffisantes pour réaliser un apprentissage des systèmes. Nous avons également souhaité que le jeu de données soit annoté manuellement avec un score d'évaluation inter-annotateurs reconnu par la communauté scientifique tel que le Kappa de Cohen (Cohen, 1960). Nous avons ainsi identifié deux jeux de données anglophones susceptibles de nous intéresser, Few-Nerd (Ding *et al.*, 2021) et DWIE (Zaporojets *et al.*, 2021). Ces deux jeux de données présentent une ontologie sur plusieurs niveaux avec un nombre important de classes, de tokens et d'entités. À la suite de cette étape d'identification, nous avons retenu celui qui présentait le meilleur score d'accord inter-annotateur (Kappa de Cohen à 0.87) : DWIE. DWIE est annoté avec 169 classes organisées dans une taxonomie à 4 niveaux.

3.2 Traduction du jeu de données en français

La première étape de notre approche pour la conversion d'un jeu de données consiste à traduire le texte source de DWIE. Nous avons effectué une comparaison de différents systèmes de traduction anglais-français en utilisant le jeu de données WMT14 EN-FR (Bojar *et al.*, 2014), en nous concentrant sur des modèles proches de l'état de l'art. Parmi ces modèles, DeepL [2] est l'un des plus performants mais ce dernier n'étant pas open-source, nous utilisons le modèle Ott *et al.* 2018 [3]. Ce modèle utilise une architecture Transformer dotée de 6 blocs encodeurs et décodeurs, et obtient un score BLEU (Papineni *et al.*, 2002) de 43.2 sur WMT14 EN-FR.

3.3 Annotation du jeu de données français par alignement

L'objectif de cette étape est d'annoter les entités du jeu de données français à partir du jeu de données anglais annoté DWIE. Nous avons expérimenté deux versions d'une approche d'alignement qui sera

2. https://www.deepl.com/translator
3. https://github.com/facebookresearch/fairseq/tree/main/examples/scaling_nmt

évaluée dans la section 4.

La première approche d'alignement est divisée en trois étapes. Tout d'abord, une identification des tokens par correspondance exacte des formes en anglais et en français permet d'annoter certaines entités avec un haut niveau de confiance. Pour les tokens restant à annoter, une mesure de distance sémantique (similarité cosinus) retournée par le modèle *Bert-base-multilingual-cased* (Devlin *et al.*, 2018) est utilisée pour déterminer le mot traduit le plus proche sémantiquement (« mot cible », dans la suite). Si la distance sémantique ou la distance lexicale entre le mot source et le mot cible est supérieure respectivement à 0,70 et 0,60 (seuils définis empiriquement) alors ce mot est annoté avec le type du mot source. Cette première approche a donné lieu à un jeu de données que l'on nommera dans la suite DWIE-FR-v1.

La seconde approche d'alignement est une extension à la première. Cette dernière est complétée par deux nouveaux modules. Le premier module consiste à établir une liste de plusieurs traductions candidates pour chaque entité de la phrase d'origine, puis à vérifier si l'une de ces traductions apparaît dans la phrase traduite. L'algorithme de recherche en faisceau utilisé pour la traduction par Vijayakumar *et al.* 2016 et réutilisé par Ott *et al.* 2019 génère justement des propositions de traduction. Ceci établit les correspondances avec les entités présentes dans la phrase traduite. Le module suivant met en œuvre des patrons d'annotation ; par exemple un token qui se situerait entre deux tokens de même type et qui est un article est annoté avec ledit type. Ces deux modules sont particulièrement utiles lorsque l'entité à traduire est composée de plusieurs tokens qui n'ont pas tous été annotés lors des étapes précédentes. Cette seconde approche résulte en un jeu de données que l'on nommera dans la suite DWIE-FR-v2.

Par ailleurs, nous utilisons l'aligneur *FastAlign* (Dyer *et al.*, 2013) afin de positionner les deux approches décrites vis à vis de cet aligneur très utilisé par la communauté scientifique. Le jeu de données obtenu avec cet aligneur est nommé "DWIE-FR FastAlign" dans la suite.

3.4 Protocole d'évaluation de DWIE-FR-v1 et DWIE-FR-v2

La qualité des jeux de données DWIE-FR-v1, DWIE-FR-v2 et DWIE-FR FastAlign obtenus par les approches présentées précédemment a été mesurée grâce à sept personnes qui ont manuellement validé ou invalidé les annotations sur des échantillons représentatifs. D'une part, ces évaluations mesurent l'impact de chaque méthode d'alignement sur le jeu de données produit ; l'objectif ici est d'évaluer à quel point les transferts de classes de la version anglaise vers la version traduite ont été correctement effectués. D'autre part, elles donnent une indication sur la qualité globale du jeu de données produit.

Pour chacun des trois jeux de données, un échantillon de 1000 tokens a été annoté par chacun des sept experts (chaque expert a annoté des échantillons différents) en respectant la typologie d'erreurs présentée dans le Tableau 3. Au total, 298 phrases ont été évaluées.

Analyse de l'évaluation Le Tableau 2 montre les résultats de cette évaluation en termes de précision, rappel et F1-score. Il apparaît que DWIE dans sa version originale obtient un F1-score de 99%, ce qui confirme la qualité de l'annotation réalisée par Zaporojets *et al.* 2021, dont le score inter-annotateurs annoncé par les auteurs est de 0.87 et conforte ce choix du jeu de données d'origine. Le rappel reste légèrement plus faible que la précision, ce qui est dû à certaines annotations manquantes notamment

pour la classe « rôle » (i.e. rôle des personnes, fonctions, métiers). Cette absence s'explique par la subjectivité d'un tel label.

L'excellente précision de DWIE est héritée par DWIE-FR V1 mais la diminution de 13% du rappel indique que cette première approche d'alignement n'est pas assez couvrante. Ce rappel est augmenté grâce à la seconde approche d'alignement puisque DWIE-FR V2 obtient un rappel de 93.5% tout en conservant une précision très haute (98.6%). Cette perte de 3.2% de rappel par rapport à DWIE s'explique notamment par la difficulté d'aligner des tokens lorsque la traduction a généré plus de tokens que le texte source et par un score de similarité sémantique qui n'atteint pas le seuil fixé. La méthode non-supervisé FastAlign obtient un meilleur rappel (95.4%). Cependant, sa précision reste trop basse (93.4%).

Finalement, nous retenons DWIE-FR V2 que nous nommerons dans la suite DWIE-FR, version rendue libre et accessible [4]. Nous considérons que DWIE-FR est d'une qualité équivalente (à 3.2% près) à son homonyme anglais. Le jeu de données est composé de 589 394 tokens dont 60 292 annotés en entités nommées avec 169 classes.

Jeu de données (version)	Précision	Rappel	F1-Score
DWIE	99.6	98.5	99.0
DWIE-FR V1	99.0	85.0	91.4
DWIE-FR V2	98.6	93.4	95.8
DWIE-FR FastAlign	93.4	95.4	94.38

TABLE 2 – Évaluation de l'annotation des jeux de données

Label d'erreur	Type d'erreur
Ce n'est pas une entité nommée (l'erreur vient du jeu anglais)	Faux positif
Ce n'est pas une entité nommée (l'erreur vient de l'alignement)	Faux positif
L'entité possède la mauvaise étiquette (l'erreur vient du jeu anglais)	Faux positif
L'entité possède la mauvaise étiquette (l'erreur vient de l'alignement)	Faux positif
Le token devrait être annoté (l'erreur vient de l'alignement)	Faux négatif
Le token devrait être annoté (l'erreur vient du jeu anglais)	Faux négatif

TABLE 3 – Typologie d'erreurs utilisée lors de l'annotation en vue de l'évaluation de l'alignement

4 Expérimentations

Deux expériences principales ont été réalisées à partir de DWIE-FR : une spécialisation de FlauBERT (Le *et al.*, 2019) et une spécialisation de CamemBERT (Martin *et al.*, 2019) (section 4.1), ainsi que l'évaluation des modèles obtenus sur trois autres jeux de données (section 4.2). Les modèles ont été entraînés sur 50 *epochs* à l'aide du *framework* Flair (Akbik *et al.*, 2019), avec un *learning rate* de 1e−5, une taille de batch de 16 et la fonction de coût cross-entropique.

4. https://github.com/Emvista/DWIE-FR

4.1 Performance des modèles FR vs. EN

Dans un premier temps, nous avons spécialisé RoBERTa (Zhuang *et al.*, 2021) à partir de DWIE-EN afin de positionner les spécialisations de FlauBERT et CamemBERT à partir de DWIE-FR. Le tableau 4 indique les performances obtenues par les trois modèles pour chacun des quatre niveaux de la taxonomie.

Modèles	Niv. 1		Niv. 2		Niv. 3		Niv. 4	
	F1 Micro	F1 Macro	F1 Micro	F1 Macro	F1 Micro	F1 Macro	F1 Micro	F1 Macro
DWIE-EN								
RoBERTa	95.35	94.3	93.51	78.98	85.99	52.01	84.32	54.69
DWIE-FR								
CamemBERT-ner [5]	87.18	83.81	84.53	58.66	75.27	29.82	72.5	27.85
CamemBERT-base	87.06	83.68	84.05	59.14	73.48	24.52	71.12	24.95
FlauBERT-base-uncased	87.21	84.36	84.36	63.17	78.08	42.31	76.21	42.69

TABLE 4 – Résultats des modèles sur DWIE-EN et DWIE-FR

Le tableau 4 laisse apparaître que plus le niveau de la taxonomie est élevé et plus le score F1 Micro diminue. La F1 Micro diminue d'environ 10% à 15% alors que le F1 Macro diminue d'environ 40% à 45% entre le niveau 1 et le niveau 4. Ceci s'explique par le déséquilibre entre les classes de haut niveau. En effet, plus le nombre de classes augmente (ce qui va de pair avec la spécialisation des classes ; niveau 1 à 4), plus l'écart entre le F1 Micro et le F1 Macro se creuse. Cette observation est valable aussi bien pour la version anglaise que pour la version française, ce qui indique une difficulté à appréhender des taxonomies à plusieurs niveaux et justifie la publication de ce nouveau jeu de données afin d'encourager des recherches à ce sujet.

Par ailleurs, il apparait que Flaubert est plus robuste que Camembert lorsqu'ils sont confrontés aux niveaux les plus élevés de la taxonomie. La spécialisation de FlauBERT obtient globalement les meilleurs résultats et pourra servir de référence pour les prochaines évaluations.

4.2 Inférence sur d'autres jeux de données

La seconde expérience vise à étudier le comportement des modèles entrainés sur DWIE-FR sur trois autres jeux de données libres et accessibles : European Newspapers-FR, Jules Verne et Wikipedia NER (décrits en section 2). L'évaluation s'effectue uniquement sur les classes Personne, Lieu et Organisation qui sont les seules classes communes à ces jeux de données.

La reconnaissance d'entité nommée est évaluée de deux manières, la reconnaissance de l'entité dans son entièreté et une reconnaissance à l'échelle du mot. Nous évaluons les systèmes en calculant le F1-score Macro tel qu'indiqué dans l'équation 1, soit, la moyenne des F1-score obtenus pour chacune des trois classes.

$$F1_{Macro} = \frac{F1_{LOC} + F1_{PER} + F1_{ORG}}{3} \quad (1)$$

5. https://huggingface.co/Jean-Baptiste/camembert-ner

Modèles	European Newspapers		Wikipedia NER		Jules Verne	
	F1 (Tokens)	F1 (BIO)	F1 (Tokens)	F1 (BIO)	F1 (Tokens)	F1 (BIO)
DWIE-FR-v2						
CamemBERT-ner	**62.38**	**54.14**	**85.23**	**76.84**	**78.62**	**70.83**
CamemBERT-base	60.53	52.03	83.06	72.27	74.62	65.37
FlauBERT-base-uncased	51.40	44.56	81.03	71.13	66.38	55.19
DWIE-FR-FastAlign						
CamemBERT-ner	56.60	46.25	82.73	72.86	74.6	59.14
CamemBERT-base	55.72	44.95	82.64	69.88	70.83	58.02
Flaubert-base-uncased	57.35	46.05	78.82	68.89	72.81	60.95

TABLE 5 – Résultats des modèles sur European NewsPapers, Wikipedia-NER FR et Jules Verne

Les résultats obtenus et consignés dans le tableau 5 montrent que les modèles entraînés sur DWIE-FR sont assez robustes pour être utilisés en pré-entraînement sur des jeux de données différents. Nous avons remarqué que globalement FlauBERT obtient de moins bons résultats que CamemBERT sur les méthodes **DWIE-FR-v2** et **FastAlign**. L'hypothèse du sur-apprentissage semble être une piste à étudier pour expliquer ces résultats. Nous pouvons également voir que le pré-apprentissage de Camembert-ner sur Wikiner-fr dans une tâche de reconnaissance d'entité nommées, permet d'améliorer les performances sur ces corpus jusqu'à 5% du F1-score.

Camembert-ner a été entraîné sur des classes génériques avec le corpus Wikiner-fr. Ce pré-entraînement a pu permettre de capturer des informations sémantiques dans les représentations intermédiaires de Camembert. Nous pouvons émettre l'hypothèse que le pré-entraînement d'un modèle sur une tâche de reconnaissance d'enitités nommées aide à la généralisation dans d'autres corpus de REN. Cependant il faut considérer que chaque jeu de données possède une stratégie d'annotation différente impliquant un alignement d'annotation difficile.

5 Conclusion

Dans cet article, nous avons constaté un manque de jeux de données en français qui soient accessibles, libres, et annotés avec plusieurs dizaines de classes d'entités nommées. Une approche constituée d'une étape de traduction et d'une étape d'alignement d'étiquettes a été expérimentée. Les évaluations permettent de considérer que le jeu de données DWIE-FR obtenu en appliquant cette approche est d'aussi haute qualité que le jeu source en anglais DWIE à 3,2% près. Des premiers apprentissages ont eu lieu avec ce nouveaux jeu de données sur la base des modèles de langues français CammemBERT et FlauBERT.

Dans la suite, nous étudierons comment maintenir des scores élevés quel que soit le niveau de l'ontologie considéré. Nous étudierons également la différence remarquée entre FlauBERT et CamemBERT qui s'accentue lorsque l'on tend vers les niveaux les plus spécifiques de la taxonomie.

Références

AKBIK A., BERGMANN T., BLYTHE D., RASUL K., SCHWETER S. & VOLLGRAF R. (2019). FLAIR : An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p. 54–59.

BANNOUR N., WAJSBÜRT P., RANCE B., TANNIER X. & NÉVÉOL A. (2022). Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français. *Actes de la journée d'étude sur la robustesse des systemes de TAL*, p.1̃2.

BOJAR O., BUCK C., FEDERMANN C., HADDOW B., KOEHN P., LEVELING J., MONZ C., PECINA P., POST M., SAINT-AMAND H., SORICUT R., SPECIA L. & TAMCHYNA A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 12–58, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-3302.

COHEN J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

DING N., XU G., CHEN Y., WANG X., HAN X., XIE P., ZHENG H.-T. & LIU Z. (2021). Few-nerd : A few-shot named entity recognition dataset. DOI : 10.48550/ARXIV.2105.07464.

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

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *North American Chapter of the Association for Computational Linguistics*.

EHRMANN M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Paris Diderot University.

EHRMANN M., ROMANELLO M., BIRCHER S. & CLEMATIDE S. (2020). Introducing the clef 2020 hipe shared task : Named entity recognition and linking on historical newspapers. In *Advances in Information Retrieval : 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, p. 524–532 : Springer.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The etape speech processing evaluation. In *LREC*, p. 3995–3999.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

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*, p. 98–113 : Springer.

KIM H., YOO J., YOON S., LEE J. & KANG J. (2021). Simple questions generate named entity recognition datasets.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. *CoRR*, **abs/1912.05372**.

LOPEZ C., MEKAOUI M., AUBRY K., BORT J. & GARNIER P. (2019). 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*, volume 79, p. 81–92.

LOPEZ C., PARTALAS I., BALIKAS G., DERBAS N., MARTIN A., REUTENAUER C., SEGOND F. & AMINI M.-R. (2017). Cap 2017 challenge : Twitter named entity recognition. *arXiv preprint arXiv :1707.07568*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

NEUDECKER C. (2016). An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4348–4352, Portorož, Slovenia : European Language Resources Association (ELRA).

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, p. 24–30.

NOTHMAN J., CURRAN J. R. & MURPHY T. (2008). Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, p. 124–132.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv :1904.01038*.

OTT M., EDUNOV S., GRANGIER D. & AULI M. (2018). Scaling neural machine translation. *CoRR*, **abs/1806.00187**.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique.

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)*, volume 2.

SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

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*, WWW '17 Companion, p. 1655–1662, Republic and Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee. DOI : 10.1145/3041021.3053367.

TEDESCHI S., MAIORCA V., CAMPOLUNGO N., CECCONI F. & NAVIGLI R. (2021a). Wikineural : Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 2521–2533.

TEDESCHI S., MAIORCA V., CAMPOLUNGO N., CECCONI F. & NAVIGLI R. (2021b). Wiki-NEuRal : Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 2521–2533, Punta Cana, Dominican Republic : Association for Computational Linguistics.

TEDESCHI S. & NAVIGLI R. (2022). 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*, p. 801–812.

VIJAYAKUMAR A. K., COGSWELL M., SELVARAJU R. R., SUN Q., LEE S., CRANDALL D. & BATRA D. (2016). Diverse beam search : Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv :1610.02424*.

YE D., LIN Y., LI P. & SUN M. (2022). Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4904–4917.

ZAPOROJETS K., DELEU J., DEVELDER C. & DEMEESTER T. (2021). Dwie : An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, **58**(4), 102563.

ZHUANG L., WAYNE L., YA S. & JUN Z. (2021). A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, p. 1218–1227.