Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Fine-tuning de modèles de langues pour la veille épidémiologique multilingue avec peu de ressources

Stephen Mutuvi[1,2] Emanuela Boros[2] Antoine Doucet[2]
Adam Jatowt[3] Gaël Lejeune[2,4] Moses Odeo[1]
(1) Multimedia University of Kenya, Nairobi, Kenya
(2) L3I, La Rochelle Université, F-17000 La Rochelle, France
(3) University of Innsbruck, 6020 Innsbruck, Autriche
(4) STIH/CERES, Sorbonne Université, F-75006 Paris, France

RÉSUMÉ

Les modèles de langues pré-entraînés connaissent un très grand succès en TAL, en particulier dans les situations où l'on dispose de suffisamment de données d'entraînement. Cependant, il reste difficile d'obtenir des résultats similaires dans des environnements multilingues avec peu de données d'entraînement, en particulier dans des domaines spécialisés tels que la surveillance des épidémies. Dans cet article, nous explorons plusieurs hypothèses concernant les facteurs qui pourraient avoir une influence sur les performances d'un système d'extraction d'événements épidémiologiques dans un scénario multilingue à faibles ressources : le type de modèle pré-entraîné, la qualité du *tokenizer* ainsi que les caractéristiques des entités à extraire. Nous proposons une analyse exhaustive de ces facteurs et observons une corrélation importante, quoique variable ; entre ces caractéristiques et les performances observées sur la base d'une tâche de veille épidémiologique multilingue à faibles ressources. Nous proposons aussi d'adapter les modèles de langues à cette tâche en étendant le vocabulaire du *tokenizer* pré-entraîné avec les entités continues, qui sont des entités qui ont été divisées en plusieurs sous-mots. Suite à cette adaptation, nous observons une amélioration notable des performances pour la plupart des modèles et des langues évalués.

ABSTRACT

Fine-tuning Language Models for Low-resource Multilingual Epidemic Surveillance.

Pre-trained language models have been widely successful, particularly in settings with sufficient training data. However, it still remains challenging to achieve similar results for low-resource multilingual settings and specialized domains such as epidemic surveillance. In this paper, we make several hypotheses regarding the factors that could impact the performance of an epidemic event extraction system in a multilingual low-resource scenario : the type of pre-trained language model, the quality of the pre-trained tokenizer, and the characteristics of the entities. We perform an exhaustive analysis of these factors and observe a strong correlation between them and the observed performance on the basis of a low-resource multilingual epidemic surveillance task. Consequently, we propose to adapt the models to this task by extending the vocabulary of the tokenizer with the continued entities, which are entities that were split into multiple subwords. Following this adaptation, we observe notable performance improvements across most of the evaluated models and languages.

MOTS-CLÉS : extraction d'événements épidémiologiques, langues peu dotées, modèles de langues.

KEYWORDS: epidemic event extraction, low-resource languages, pre-trained language models.

Volume 1 : conférence principale, pages 345–354.

1 Introduction

Les dépêches sur les épidémies sont des données de plus en plus nombreuses qui permettent de concevoir des systèmes d'alerte précoce quant à ces événements imprévus. Si les méthodes de surveillance conventionnelles sont de plus en plus fiables, elles dépendent en partie de données bien définies produites manuellement par des établissements de santé (Dórea & Revie, 2021), ce qui limite la couverture géographique (Brownstein *et al.*, 2008; Lejeune *et al.*, 2015; Yangarber *et al.*, 2007, 2005). Dans une perspective d'automatisation du recueil des données, un frein se situe dans l'obtention d'annotations suffisantes et de haute qualité, nécessaires à l'entraînement et à l'évaluation des systèmes. C'est un processus laborieux et coûteux, car il nécessite l'intervention d'experts du domaine (Hedderich *et al.*, 2021; Neves & Leser, 2014). Le deuxième défi est lié à la nature multilingue des données en ligne, les informations décisives pouvant être émises dans un grand éventail de langues. Dans un tel contexte, les langues bien dotées en ressources d'analyse, comme l'anglais, sont favorisées, tandis que des langues peu dotées en ressources peuvent ne disposer que de données d'entraînement limitées, voire inexistantes (Ramesh & Sankaranarayanan, 2018; Lauscher *et al.*, 2020).

Récemment, l'utilisation de modèles de langues pré-entraînés pour la veille épidémiologique a été explorée (Mutuvi *et al.*, 2020). Cependant, ces systèmes sont généralement monolingues ou difficiles à adapter à d'autres langues. Les récents modèles de langue multilingues pré-entraînés (par exemple, BERT multilingue (Devlin *et al.*, 2019a), XLM-RoBERTa (Conneau *et al.*, 2020)) ont démontré d'excellentes performances sur diverses tâches de transfert cross-lingue en *zero-shot* (Tian *et al.*, 2021; Faruqui & Kumar, 2015; Lin *et al.*, 2017; Zou *et al.*, 2018; Wang *et al.*, 2018), où un modèle est d'abord affiné sur la langue source, puis évalué directement sur de multiples langues cibles qui n'ont pas été vues lors de l'entraînement (Wu & Dredze, 2019; Wang *et al.*, 2019).

Bien que ces modèles pré-entraînés obtiennent des résultats remarquables, l'utilisation directe du vocabulaire pré-entraîné se fait au prix d'une sur-segmentation de termes spécifiques au domaine en plusieurs sous-mots, ce qui compromet l'efficacité de l'entraînement. Par exemple, Beltagy *et al.* (2019) ont proposé SciBERT qui utilise des mots et des sous-mots fréquemment observés dans un corpus d'articles scientifiques qui permettent une extension du vocabulaire du modèle initial pré-entraîné. L'adaptation du vocabulaire dans SciBERT a nécessité un coûteux entraînement du modèle à partir de zéro. Tai *et al.* (2020) ont proposé une extension de BERT au domaine biomédical, avec une étude similaire portant spécifiquement sur une tâche de reconnaissance d'entités nommées (Poerner *et al.*, 2020). De même, Hong *et al.* (2021) ont présenté une approche qui ne nécessiterait qu'un jeu de données spécifique en entrée pour adapter la classification de textes à différents domaines (biomédical, informatique, etc.), en identifiant un sous-ensemble du vocabulaire spécifique au domaine par des mesures d'importance relative des mots. Malgré ces exemples, les études portant véritablement sur des données multilingues avec des ressources limitées font défaut. Dans cet esprit, une analyse approfondie des performances du modèle qui tiendrait compte des attributs saillants des entités à extraire et du modèle de langage pré-entraîné ainsi que du choix du *tokenizer* semble nécessaire.

C'est pourquoi, dans cet article, nous proposons une analyse exhaustive des modèles de langage pré-entraînés multilingues et spécifiques à une langue, de leurs tokenizers et des caractéristiques des informations extraites par ces modèles. Nous observons qu'il existe une forte corrélation entre non seulement la qualité du tokenizer utilisé, mais aussi la casse des entités de noms de lieux et de maladies. De plus, nous montrons que l'adaptation des modèles pré-entraînés en exploitant des entités spécifiques au domaine entraîne des gains de performance, lorsqu'ils sont évalués sur des ensembles

de données multilingues d'événements épidémiologiques avec des langues à faibles ressources.

2 Extraction d'événements épidémiologiques

2.1 Jeu de données et verrous scientifiques

Pour notre étude, nous utilisons le corpus DANIEL [1] (Mutuvi *et al.*, 2021; Lejeune *et al.*, 2015), qui contient des articles dans plusieurs langues : anglais, grec, russe et polonais. Il comporte précisément 474 articles en anglais, 390 en grec, 352 en polonais et 426 en russe. La conception d'un système d'extraction d'événements épidémiologiques sur ce jeu de données implique la détection des occurrences de noms de maladie et de noms de lieux décrivant un événement. Nous prenons, pour exemple, l'extrait suivant d'un article anglais du jeu de données DANIEL faisant état d'une épidémie potentielle de norovirus à *Victoria*, au Canada : *[. . .] Dozens of people ill in a suspected outbreak of norovirus at a student journalism conference in Victoria are under voluntary quarantine in their hotel rooms. [. . .]* [2]. Un système d'extraction d'événements devrait détecter le nom de la maladie *norovirus* (DIS) ainsi que le lieu *Victoria* (LOC).

2.2 Hypothèses

Sur la base de l'analyse de la littérature sur le sujet, nous pensons qu'un système d'extraction d'événements épidémiologiques dans un scénario multilingue à faibles ressources pourrait être affecté par une série de facteurs que nous examinons dans les paragraphes suivants.

Modèles de langues Nous comparons deux modèles multilingues, BERT (Devlin *et al.*, 2019b) (avec deux variantes de prise en compte de la casse, *multilingual-cased* et *multilingual-uncased*) et XLM-RoBERTa (Conneau *et al.*, 2020), vis-à-vis de modèles monolingues existant pour les langues à faibles ressources de notre corpus. Pour les modèles monolingues, nous utilisons *bert-base-uncased*, *bert-base-greek-uncased-v1*, *rubert-base-cased* et *bert-base-polish-uncased-v1* [3].

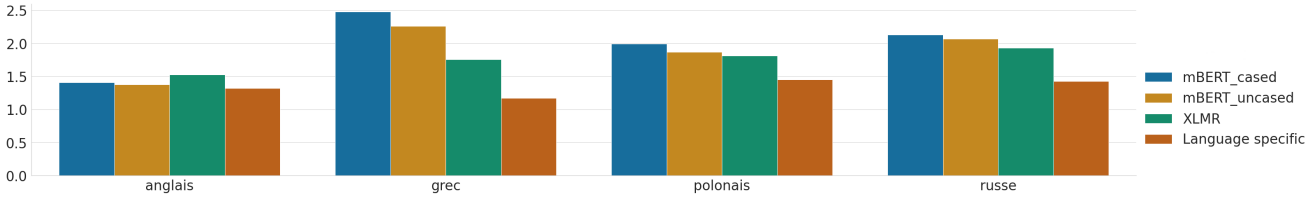

FIGURE 1 – Fertilité du *tokenizer*.

1. Le corpus est librement et publiquement disponible à l'adresse https://doi.org/10.5281/zenodo.6024726.
2. Extrait d'un article anglais du corpus DANIEL, daté du 15 janvier 2012 https://tinyurl.com/norovirus-outbreak-suspected.
3. Ces modèles sont disponibles sur https://huggingface.co/.

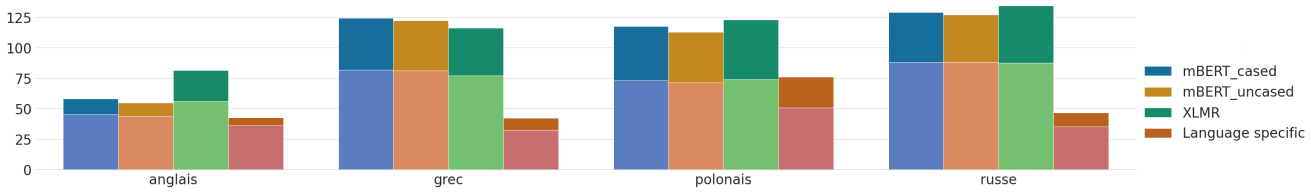

FIGURE 2 – Nombre de mots continués (en bas) et d'entités continuées (en haut) par langue et par modèle.

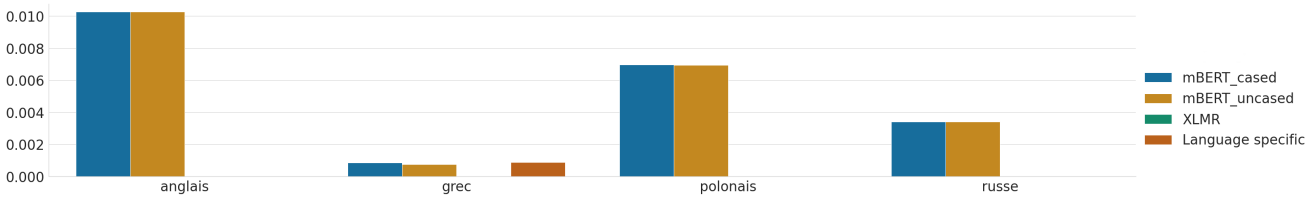

FIGURE 3 – Nombre de mots inconnus (OOV) par langue et par modèle.

Qualité du *tokenizer* pré-entraîné L'analyse de la qualité du *tokenizer* implique, une fois la segmentation réalisée, de calculer deux métriques proposées par Rust *et al.* (2021) : *fertilité* et *mots continués* (pour prendre une traduction littérale de *continued words*). La *Fertilité* mesure le nombre moyen de sous-mots générés par mot *tokenisé*, ce qui indique l'intensité avec laquelle un *tokenizer* va segmenter. Une fertilité minimale de 1 signifie que le vocabulaire du *tokenizer* contient simplement chaque mot du corpus. Les valeurs sont présentées dans la figure 1, on peut remarquer que les modèles monolingues ont les scores de fertilité les plus bas, ce qui semble naturel. Les *mots continués* sont les mots *tokenizés* en plusieurs sous-mots (indiqués par les symboles de continuation ##). Cette mesure indique la fréquence à laquelle un *tokenizer* divise les mots en moyenne et on recherche généralement un faible nombre de mots continués. Le nombre de mots continués est présenté dans la figure 2, où nous pouvons faire la même observation en ce qui concerne les modèles spécifiques à la langue. Enfin, nous ajoutons les ratios de *entités continuées* qui sont les lieux et les noms de maladies segmentés en plusieurs sous-mots. Nous observons, pour toutes les langues, un pourcentage élevé d'entités continuées.

Nous montrons également la quantité de mots hors vocabulaire (OOV) dans la figure 3. En général, la proportion devrait être extrêmement faible, c'est-à-dire que les *tokenizers* peuvent généralement diviser les OOV jusqu'à obtenir des sous-mots connus. Nous remarquons que *mBERT_uncased* et *mBERT_cased* semblent générer un nombre important de mots OOVs, en particulier pour l'anglais et le polonais. D'autre part, le *tokenizer* XLMR est capable de reconnaître et de diviser tous les *tokens* de notre jeu de données, sans avoir besoin du symbole <unk>.

Caractéristiques des entités à extraire Étant donné que certains des modèles pré-entraînés peuvent avoir deux configurations (*cased* pour le texte en l'état et *uncased* pour le texte en minuscules), nous considérons que cela pourrait avoir une influence sur la performance d'un système d'extraction d'événements épidémiologiques. Nous avons donc analysé les ratios d'entités (lieux et noms de maladies) qui comportent ou non des majuscules. Le tableau 1 révèle qu'il y a globalement un équilibre entre les deux, avec une valeur légèrement plus importante pour les entités en majuscules pour le grec, le russe et l'anglais. Au contraire, 54% des entités en polonais sont en minuscules, ce qui pourrait contribuer à rendre la plus-value du modèle *uncased* plus faible que pour les autres langues.

Langue	#Entités	Capitalisation	Entités (%)	DIS (%)	LOC (%)
English	563	majuscule	**55,60**	36,42	63,58
		minuscule	43,34	93,85	6,15
Greek	300	majuscule	**55,67**	22,16	77,84
		minuscule	44,33	96,99	3,01
Polish	638	majuscule	43,42	37,18	62,82
		minuscule	**54,70**	94,27	5,73
Russian	327	majuscule	**63,91**	53,59	46,41
		minuscule	36,09	94,92	5,08

TABLE 1 – Nombre d'entités en majuscules et en minuscules dans l'ensemble du corpus DANIEL.

2.3 Évaluation des modèles

Tout d'abord, nous affinons les modèles multilingues et monolingues sur le jeu de données d'extraction d'événements épidémiologiques [4]. Nous évaluons les modèles en faisant la moyenne de la F_1-mesure sur cinq exécutions réalisées avec différentes amorces aléatoires.

Langue	Métrique	mBERT_cased	mBERT_uncased	XLMR	Monolingue
Anglais	F_1	50,56±25,70	69,88±2,36	61,59±5,16	**72,30**±3,67
Grec	F_1	46,63±42,58	78,39±4,07	74,85±2,52	**80,96**±1,95
Polonais	F_1	88,61±2,53	88,60±3,05	87,14±2,49	**89,32**±2,54
Russe	F_1	59,05±12,40	**67,08**±0,76	54,47±3,55	58,14±1,75

TABLE 2 – Performance des différents modèles multilingues et monolingues, moyenne de la F_1-mesure sur 5 exécutions et écart-type.

Comme le montre le tableau 2, les modèles monolingues obtiennent les meilleures performances globales, enregistrant les scores les plus élevés dans toutes les langues considérées, à l'exception du russe. Bien que la capitalisation soit considérée comme un facteur important pour les tâches de reconnaissance d'entités nommées (Mayhew *et al.*, 2019, 2020), les meilleures performances ont été obtenues en utilisant les modèles *uncased*. Ceci pourrait être dû à la présence d'un nombre relativement important de noms de maladies en minuscules dans notre jeu de données, comme le montre le tableau 1 ainsi qu'à une inconsistance de la casse. Nous constatons par ailleurs que mBERT_cased présente des écarts types plus élevés, supérieurs à 25 pour l'anglais et à 40 pour le grec, montrant une moindre robustesse de ce modèle.

Ensuite, dans les tableaux 3 et 4, nous présentons respectivement les corrélations de Pearson et de Spearman entre les F_1-mesure observées (tableau 2), la fertilité, ainsi que la proportion d'entités continuées et de mots OOV par type de modèle et par langue. Nous avons également ajouté les *p-value* entre parenthèses. Puisque toutes les entités de notre jeu de données ont été identifiées par les *tokenizers*, nous avons exploré l'analyse des mots OOV qui représentent les mots absents du vocabulaire pré-entraîné.

Le tableau 3 montre des corrélations relativement faibles entre les caractéristiques observées sur le lexique et la F_1-mesure. C'est le cas pour tous les modèles, avec une corrélation positive faible

4. Dans toutes les expériences, nous utilisons AdamW (Kingma & Ba, 2014) avec un taux d'apprentissage de $1e-5$ et 20 *epochs*. Nous avons défini la longueur maximale des phrases à 164 (Adelani *et al.*, 2021).

Modèle	Type de Corrélation	Entités continuées	Mots OOV	Fertilité
mBERT	Spearman	0,0000 (1,00)	**0,4000 (0,60)**	**-0,4000 (0,60)**
cased	Pearson	0,1194 (0,88)	0,2645 (0,74)	-0,0442 (0,96)
mBERT	Spearman	**-0,4000 (0,60)**	0,0000 (1,00)	0,0000 (1,00)
uncased	Pearson	0,0726 (0,93)	-0,0310 (0,97)	0,1989 (0,80)
XLMR	Spearman	-0,4000 (0,60)	nan (nan)	-0,2000 (0,80)
	Pearson	-0,0804 (0,92)	nan (nan)	0,0242 (0,98)
Modèles	Spearman	0,4000 (0,60)	-0,2108 (0,79)	0,2000 (0,80)
monolingues	Pearson	**0,6077 (0,39)**	0,2773 (0,72)	-0,1747 (0,83)

TABLE 3 – Corrélation (*p-value*) entre les F_1-mesures observées sur l'ensemble du corpus avec les modèles et différentes caractéristiques du lexique.

en moyenne pour mBERT_cased et plus forte lorsque l'on utilise les modèles monolingues (sauf pour la fertilité). Nous avons des corrélations faibles pour mBERT_uncased et XLMR. Lorsque le pourcentage d'entités continuées est élevé, la performance de ces deux modèles diminue. La fertilité du *tokenizer* semble être un facteur important pour mBERT_cased et XLMR. Cependant, la plupart des *p-values* sont élevées, ce qui indique une forte probabilité que nos résultats soient généralement non corrélés lorsqu'ils sont analysés par type de modèle. Le nombre de mots OOV pour XLMR montre *nan* (*Not a Number*) en raison du nombre de mots OOV (figure 3).

Modèle	Type de Corrélation	Entités continues	Mots OOV	Fertilité
Anglais	Spearman	**-0,8000 (0,20)**	-0,4472 (0,55)	**-0,8000 (0,20)**
	Pearson	-0,4569 (0,54)	-0,3955 (0,60)	-0,4782 (0,52)
Grec	Spearman	**-0,8000 (0,20)**	0,4000 (0,60)	**-0,8000 (0,20)**
	Pearson	-0,4900 (0,51)	-0,2198 (0,78)	-0,6880 (0,31)
Polonais	Spearman	**-0,8000 (0,20)**	0,1054 (0,89)	-0,2000 (0,80)
	Pearson	-0,7138 (0,29)	0,2374 (0,76)	-0,4378 (0,56)
Russe	Spearman	0,7379 (0,26)	**0,9487 (0,05)**	0,6000 (0,40)
	Pearson	0,2040 (0,80)	0,7354 (0,26)	0,3264 (0,67)

TABLE 4 – Corrélation (*p-value*) entre les scores F_1 et les caractéristiques du lexique de chaque *langue*.

Nous observons dans le Tableau 4 de fortes corrélations négatives du côté des entités continuées à l'exception du russe. Ainsi, pour l'anglais, le grec et le polonais, les modèles sont susceptibles d'être moins performants du fait des proportions d'entités continuées et de mots OOV. Il est intéressant de noter que pour le russe, il y a une forte corrélation positive entre les entités continues, les mots OOV et la F_1. Sur la base de ces observations, si nous segmentons correctement les entités continuées, nous devrions nous attendre à ce que les performances de mBERT_uncased et XLMR augmentent pour toutes les langues.

2.4 Correction des entités continuées

Nous cherchons à évaluer l'influence de l'augmentation de la capacité du modèle en ajoutant des termes du domaine dans le vocabulaire du *tokenizer*. Les entités invisibles de l'ensemble d'appren-

Langue	Modèle	mBERT_cased	mBERT_uncased	XLMR	Monolingue
Anglais	Défaut	50,56±25,70	**69,88**±2,36	**61,59**±5,16	**72,30**±3,67
	E-Model	**56,34**±5,55	50,11±2,30	55,30±1,73	56,79±2,20
Grec	Défaut	46,63±42,58	**78,39**±4,07	74,85±2,52	80,96±1,95
	E-Model	**63,53**±35,53	76,47±2,61	**79,88** ±2,73	**83,95**±4,08
Polonais	Défaut	88,61±2,53	88,60±3,05	87,14±2,49	89,42±2,54
	E-Model	**90,02**±1,89	**91,43**±2,90	**87,67**±2,92	**91,76**±2,39
Russe	Défaut	**59,05**±12,40	**67,08**±0,76	54,47±3,55	58,14±1,75
	E-Model	58,38±2,02	60,40±0,09	**58,77**±9,90	**61,81**±3,22

TABLE 5 – Comparaison entre le modèle pré-entraîné par défaut et le modèle étendu (E-Model). L'E-Model est obtenu en enrichissant le vocabulaire du *tokenizer*.

tissage, qui ont été divisées en plusieurs sous-mots par le *tokenizer*, sont utilisées pour étendre le vocabulaire pré-entraîné. Les performances s'améliorent pour les langues et les modèles à faibles ressources, comme le montre le tableau 5 (E-model). Une amélioration notable des performances a été observée pour le polonais sur tous les modèles, tandis que pour le grec, seule la version étendue BERT_uncased multilingue n'a pas enregistré de gain de performance. Une baisse significative de la performance est observée pour l'anglais, baisse qui peut être attribuée soit à l'influence négative des mots rares, soit au fait que le vocabulaire des modèles pré-entraînés et ajustés devient trop éloigné et que le modèle perd les connaissances apprises précédemment, ce qui entrave la généralisation. Ce phénomène est appelé l'oubli catastrophique (*catastrophic forgetting*) (Kirkpatrick *et al.*, 2017; Chen *et al.*, 2019).

3 Conclusion

Dans cet article, nous avons réalisé une analyse systématique de la qualité de la sortie du *tokenizer*, en mesurant la proportion de mots inconnus (OOV), de mots et d'entités continuées, ainsi que la fertilité du *tokenizer* pour la veille épidémiologique multilingue. Sur la base de cette analyse, les entités inconnues liées au domaine de l'épidémiologie sont identifiées et exploitées pour étendre le vocabulaire de modèles de langues pré-entraînés. Les résultats indiquent que le vocabulaire du domaine joue un rôle important dans l'adaptation des modèles pré-entraînés à un domaine spécifique et est particulièrement efficace dans les configurations à faibles ressources.

Remerciements

Ce travail a été soutenu par les projets ANNA et Termitrad financés par la Région Nouvelle-Aquitaine.

Références

ADELANI D. I., ABBOTT J., NEUBIG G., D'SOUZA D., KREUTZER J., LIGNOS C., PALEN-MICHEL C., BUZAABA H., RIJHWANI S., RUDER S. *et al.* (2021). Masakhaner : Named entity

recognition for african languages. *arXiv preprint arXiv :2103.11811*.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : Pretrained language model for scientific text. In *EMNLP*.

BROWNSTEIN J. S., FREIFELD C. C., REIS B. Y. & MANDL K. D. (2008). Surveillance sans frontieres : Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, **5**(7), e151.

CHEN X., WANG S., FU B., LONG M. & WANG J. (2019). Catastrophic forgetting meets negative transfer : Batch spectral shrinkage for safe transfer learning. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 32 : Curran Associates, Inc.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 8440–8451 : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019a). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019b). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DÓREA F. C. & REVIE C. W. (2021). Data-driven surveillance : Effective collection, integration and interpretation of data to support decision-making. *Frontiers in Veterinary Science*, **8**, 225.

FARUQUI M. & KUMAR S. (2015). Multilingual open relation extraction using cross-lingual projection. *arXiv preprint arXiv :1503.06450*.

HEDDERICH M. A., LANGE L., ADEL H., STRÖTGEN J. & KLAKOW D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2545–2568, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.201.

HONG J., KIM T., LIM H. & CHOO J. (2021). AVocaDo : Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4692–4700, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.385.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

KIRKPATRICK J., PASCANU R., RABINOWITZ N., VENESS J., DESJARDINS G., RUSU A. A., MILAN K., QUAN J., RAMALHO T., GRABSKA-BARWINSKA A. *et al.* (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, **114**(13), 3521–3526.

LAUSCHER A., RAVISHANKAR V., VULIĆ I. & GLAVAŠ G. (2020). From zero to hero : On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv :2005.00633*.

LEJEUNE G., BRIXTEL R., DOUCET A. & LUCAS N. (2015). Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, **65**(2), 131–143.

LIN Y., LIU Z. & SUN M. (2017). Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 34–43.

MAYHEW S., GUPTA N. & ROTH D. (2020). Robust named entity recognition with truecasing pretraining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, p. 8480–8487 : AAAI Press.

MAYHEW S., TSYGANKOVA T. & ROTH D. (2019). ner and pos when nothing is capitalized. *arXiv preprint arXiv :1903.11222*.

MUTUVI S., BOROS E., DOUCET A., LEJEUNE G., JATOWT A. & ODEO M. (2020). Multilingual epidemiological text classification : A comparative study. In *COLING, International Conference on Computational Linguistics*.

MUTUVI S., BOROS E., DOUCET A., LEJEUNE G., JATOWT A. & ODEO M. (2021). Token-level multilingual epidemic dataset for event extraction. In *International Conference on Theory and Practice of Digital Libraries*, p. 55–59 : Springer.

NEVES M. & LESER U. (2014). A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics*, **15**(2), 327–340.

POERNER N., WALTINGER U. & SCHÜTZE H. (2020). Inexpensive domain adaptation of pretrained language models : case studies on biomedical ner and covid-19 qa. *arXiv preprint arXiv :2004.03354*.

RAMESH S. H. & SANKARANARAYANAN K. P. (2018). Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 112–119, New Orleans, Louisiana, USA : Association for Computational Linguistics. DOI : 10.18653/v1/N18-4016.

RUST P., PFEIFFER J., VULIĆ I., RUDER S. & GUREVYCH I. (2021). How good is your tokenizer ? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 3118–3135, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.243.

TAI W., KUNG H., DONG X. L., COMITER M. & KUO C.-F. (2020). exbert : Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : Findings*, p. 1433–1439.

TIAN L., ZHANG X. & LAU J. H. (2021). Rumour detection via zero-shot cross-lingual transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 603–618 : Springer.

WANG X., HAN X., LIN Y., LIU Z. & SUN M. (2018). Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1156–1166.

WANG Z., MAYHEW S., ROTH D. *et al.* (2019). Cross-lingual ability of multilingual bert : An empirical study. *arXiv preprint arXiv :1912.07840*.

WU S. & DREDZE M. (2019). Beto, bentz, becas : The surprising cross-lingual effectiveness of BERT. 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)*, p. 833–844, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1077.

YANGARBER R., BEST C., VON ETTER P., FUART F., HORBY D. & STEINBERGER R. (2007). Combining information about epidemic threats from multiple sources. In *Proceedings of the MMIES Workshop, International Conference on Recent Advances in Natural Language Processing (RANLP 2007)* : Citeseer.

YANGARBER R., JOKIPII L., RAURAMO A. & HUTTUNEN S. (2005). Extracting information about outbreaks of infectious epidemics. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, p. 22–23.

ZOU B., XU Z., HONG Y. & ZHOU G. (2018). Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 437–448.