Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Jargon : Une suite de modèles de langues et de référentiels d'évaluation pour les domaines spécialisés du français (Accepté à LREC/Coling 2024)

Vincent Segonne[1] Aidan Mannion [2,3] Laura Cristina Alonzo Canul [2] Alexandre Audibert[2] Xingyu Liu[2,4] Cécile Macaire[2] Adrien Pupier[2] Yongxin Zhou[2] Mathilde Aguiar[5] Felix Herron[2,6] Magali Norré[7,8] Massih-Reza Amini[2] Pierrette Bouillon[8] Iris Eshkol-Taravella[9] Emmanuelle Esperança-Rodier[2] Thomas François[7] Lorraine Goeuriot[2] Jérôme Goulian[2] Mathieu Lafourcade[10] Benjamin Lecouteux[2] François Portet[2] Fabien Ringeval[2] Vincent Vandeghinste[11,12] Maximin Coavoux[2] Marco Dinarelli[2] Didier Schwab[2]

(1) Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
(2) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(3) EPOS SAS, France
(4) Shesmet, France
(5) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
(6) Laboratoire d'Analyse et de Modélisation de Systèmes d'Aide à la Décision (LAMSADE)
(7) CENTAL, IL&C, UCLouvain, Belgique
(8) Faculty of Translation and Interpreting, University of Geneva, Suisse
(9) MoDyCo, UPL, Univ Paris Nanterre, France
(10) LIRMM, Univ Montpellier, France
(11) Instituut voor de Nederlandse Taal, Pays-Bas
(12) KU Leuven, Belgique

Les modèles de langue préentraînés (PLM) constituent aujourd'hui *de facto* l'épine dorsale de la plupart des systèmes de traitement automatique des langues. Dans cet article, nous présentons Jargon, une famille de PLMs pour des domaines spécialisés du français, en nous focalisant sur trois domaines : la parole transcrite, le domaine clinique / biomédical, et le domaine juridique. Nous utilisons une architecture de transformeur basée sur des méthodes computationnellement efficaces (LinFormer) puisque ces domaines impliquent souvent le traitement de longs documents. Nous évaluons et comparons nos modèles à des modèles de l'état de l'art sur un ensemble varié de tâches et de corpus d'évaluation, dont certains sont introduits dans notre article. Nous rassemblons les jeux de données dans un nouveau référentiel d'évaluation en langue française pour ces trois domaines. Nous comparons également diverses configurations d'entraînement : préentraînement prolongé en apprentissage autosupervisé sur les données spécialisées, préentraînement à partir de zéro, ainsi que préentraînement mono et multi-domaines. Nos expérimentations approfondies dans des domaines spécialisés montrent qu'il est possible d'atteindre des performances compétitives en aval, même lors d'un préentraînement avec le mécanisme d'attention approximatif de LinFormer. Pour une reproductibilité totale, nous publions les modèles et les données de préentraînement, ainsi que les corpus utilisés.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 2 : traductions d'articles publiés, pages 9–10.

ABSTRACT

Jargon : A Suite of Language Models and Evaluation Tasks for French Specialized Domains (Accepted LREC/Coling 2024)

Pretrained Masked Language Models (PLMs) are the de facto backbone of most state-of-the-art NLP systems. In this paper, we introduce a family of domain-specific pretrained PLMs for French, focusing on three important applications : the domains of transcribed speech, medicine, and law. We use a transformer architecture based on efficient methods (LinFormer) to maximise their utility, since these domains often involve processing long documents. We evaluate and compare our models to state-of-the-art models on a diverse set of tasks and datasets, some of which are introduced in this paper. We gather the datasets into a new French-language evaluation benchmark for these three domains. We also compare various training configurations : continued pretraining, pretraining from scratch, as well as single- and multi-domain pretraining. Extensive domain-specific experiments show that it is possible to attain competitive downstream performance even when pre-training with the approximative LinFormer attention mechanism. For full reproducibility, we release the models and pretraining data, as well as contributed datasets.

MOTS-CLÉS : Autoapprentissage, modèles de langue préentraînés, référentiels d'évaluation, Traitement Automatique de la langue biomédicale et clinique, Traitement Automatique de documents légaux, transcription automatique.

KEYWORDS: Self-supervised learning, pretrained language models, evaluation benchmark, biomedical document processing, legal document processing, speech transcription.