Jargon: A Suite of Language Models and Evaluation Tasks for French Specialized Domains

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
first.last@univ-ubs.fr
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
first.last@univ-grenoble-alpes.fr, [3]EPOS SAS [4]Shesmet
[5]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,
91400, Orsay, France mathilde.aguiar@lisn.fr
[6]Laboratoire d'Analyse et de Modélisation de Systèmes d'Aide à la Décision (LAMSADE)
[7]CENTAL, IL&C, UCLouvain, Belgium, first.last@uclouvain.be
[8]Faculty of Translation and Interpreting, University of Geneva, first.last@unige.ch
[9]MoDyCo, UPL, Univ Paris Nanterre, France, eshkoltaravella@parisnanterre.fr
[10]LIRMM, Univ Montpellier, mathieu.lafourcade@lirmm.fr
[11]Instituut voor de Nederlandse Taal, the Netherlands; [12]KU Leuven, Belgium
vincent.vandeghinste@kuleuven.be

Abstract

Pretrained 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 domains: 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.

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

1. Introduction

Pretrained masked language models (PLMs) form the basis of most state-of-the-art natural language processing (NLP) applications. The first proposals to reuse representations extracted from pretrained language models as general-purpose contextualized embeddings (Howard and Ruder, 2018; Peters et al., 2018) used directional language models. Devlin et al. (2019) introduced BERT, a self-attentive (Vaswani et al., 2017) architecture trained with a masked language modelling objective: given sequences of tokens where some tokens have been replaced by a [MASK] pseudo-token, the model is tasked with predicting the original tokens behind the masks. Since then, BERT-style models have been introduced for many languages, be they monolingual, e.g. Le et al. (2020) and Martin et al. (2020) for French, Antoun et al. (2020) for Arabic, Agerri et al. (2020) for Basque, de Vries et al. (2019) for Dutch; bilingual (generally to take advantage of large quantities of English-language data), e.g. Spanish/English (de la Iglesia et al., 2023), Chinese/English (Zeng et al., 2022); or multilingual (Conneau and Lample, 2019; Conneau et al., 2020). Many such PLMs for specialized applications have also been developed, such as legal BERT (Chalkidis et al., 2020, en), SciBERT (Beltagy et al., 2019, en), BioBERT (Lee et al., 2019, en), Juribert (Douka et al., 2021, fr), legal CamemBERT (Louis and Spanakis, 2022, fr), FlauBERT-oral (Hervé et al., 2022, fr), DrBERT (Labrak et al., 2023, fr), CamemBERT-bio-base

LREC-COLING 2024, pages 9463–9476
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

(Touchent et al., 2023, fr), to name a few for English and French. Constructing specialised models such as these involves either training a PLM from scratch on the target domain, or continuing the training of a general purpose PLM on target data (Chalkidis et al., 2020; El Boukkouri et al., 2022), with no overall clear-cut advantage for one method over the other.

In this paper, we introduce specialized French PLMs for three different NLP applications: speech transcriptions, medicine, and law, each of which is faced with specific domain shift issues, e.g. the absence of punctuation in speech transcriptions, or the highly specialised terminology and non-standard sentence construction found in legal and biomedical documents. We construct new pretraining datasets for these three domains, and release Jargon, a family of new PLMs. In contrast to prior work on French specialized domains, we use the LinFormer architecture (Wang et al., 2020a), which allows the model to treat as many as 4096 sub-tokens (whereas currently available models have a 512 subtoken limit). Moreover, we use the same architecture and training procedure for all three domains, allowing for cross-domain comparisons of the benefits of the architecture. Finally, we also train a multi-domain model to assess the cost-accuracy tradeoff between training many specialized models and training a single multi-purpose model.

We evaluate our proposal on a suite of 16 tasks (Section 2). In particular, on top of evaluating our models on existing datasets, we introduce a new French dataset for the legal domain: **ECtHR-FR**,[1] a corpus of French decisions from the European Court of Human Rights that is comparable to the existing dataset for English (Aletras et al., 2016).

Moreover, for the speech domain, we propose a new type of extrinsic evaluation: measuring the pretrained model through reranking.

Contributions:

- Construction of French pretraining datasets for three domains of applied NLP (biomedical, legal, transcribed speech);
- Pretraining and evaluation of French PLMs for the three above domains;
- A multi-domain evaluation benchmark that includes a new legal-domain dataset annotated for sequence classification.
- All code[2], models [3] and data [4] will be publicly released.

[1]The corpus is available at https://huggingface.co/datasets/audibeal/fr-echr and https://zenodo.org/uploads/10865547.

[2]https://github.com/PantagrueLLM/Jargon/

[3]https://huggingface.co/PantagrueLLM

[4]https://zenodo.org/uploads/10865547

2. Evaluation Benchmarks

2.1. Speech-related Tasks

Automatic Speech Recognition (ASR) Language models are important parts of neural ASR systems. However, language models trained on written text fail to adequately represent speech transcriptions due to speech-specific phenomena, in particular speech from spontaneous interactions, such as the lack of punctuation, hesitations (*hmmm*, *heu*) and repetitions. Hence, there is a need for language models that are better adapted to spoken language transcriptions. As far as we know, we are the first to evaluate French PLMs on the ASR task, since non-causal PLMs tend to be ill-suited for this task.

For ASR evaluation, we use CommonVoice (Ardila et al., 2019) version 10.0, a standard dataset for automatic recognition of read speech - see Table 1 for the descriptive statistics.

We evaluate ASR with two standard metrics: Character Error Rate (CER) and Word Error Rate (WER).

Dependency parsing Dependency parsing consists in assigning a labeled dependency tree to a natural language sentence. We evaluate our speech PLM with dependency parsing on two spoken treebanks: the CEFC-Orféo corpus (Benzitoun et al., 2016) and Paris Stories[5] (Nivre et al., 2020). The CEFC-Orféo corpus contains multiple subcorpora from different sources, with a diversity of interaction types (interviews, meetings, casual discussions, commercial interactions, etc.). The Orféo treebank contains 1,732,398 tokens (171,382 sentences) corresponding to 150 hours of recording.

Approximately 5% of the total Orféo treebank have manually annotated (gold) syntactic trees, while the rest were automatically generated (Nasr et al., 2020). For this task, we use a mix of gold and automatically annotated data for the training set, while the validation and test sets contain gold data only. Since a subcorpus of Orféo (TCOF) is also included in the pretraining data, we took steps to ensure that there was no overlap with the test dataset.

The second treebank we evaluate on, Paris Stories, features interviews with people living in the Paris metropolitan area and contains 43,251 tokens.

For both corpora, we use standard metrics to evaluate parsing: part of speech tagging accuracy (POS), unlabeled attachment score (UAS), and labeled attachment score (LAS).

[5]https://universaldependencies.org/treebanks/fr_parisstories/

Task	Dataset	Domain	Source	Size			Classes
				Train	Dev	Test	
ASR	CommonVoice	speech	Ardila et al. (2019)	253,432s	15,479s	15,514s	-
Dependency Parsing	Orfeo	speech	Benzitoun et al. (2016)	169,685s	858s	839s	-
	Paris Stories	speech	Nivre et al. (2020)	1387s	692s	697s	-
NLU	Media	speech	Bonneau-Maynard et al. (2006)	12,908d	1,259d	3,005d	72
Sequence Classification	ECtHR-French	legal	**This paper**	7,756d	862d	957d	10
	OACS	legal	OACS GIP Justice Project[6]	3,570d	397d	441d	2
	Swiss-Judgement	legal	Niklaus et al. (2021)	21,179d	3,095d	6,820d	2
Sequence Classification	FrenchMedMCQA	biomedical	Labrak et al. (2022)	2,171d	312d	622d	31
	MQC	biomedical	Laleye et al. (2020)	2,161s	270s	270s	7
Token Classification	CAS-POS	biomedical	Grabar et al. (2018)	2,652s	569s	569s	31
	CAS-SG	biomedical	Grabar et al. (2018)	167d	54d	54d	15
	MEDLINE	biomedical	Névéol et al. (2014)	1,665s	-	833s	11
	EMEA	biomedical	Névéol et al. (2014)	1,036s	-	486s	11
	ESSAI-POS	biomedical	Dalloux et al. (2021)	5,072s	1,088s	1,087s	34
	E3C-NER	biomedical	Magnini et al. (2020)	168d	-	81d	3
Semantic Textual Similarity	CLISTER	biomedical	Hiebel et al. (2022)	1,080s	120s	800s	-

Table 1: Summary of all domain-specific downstream NLP tasks addressed in this paper. Size units: (s)entences, (d)ocuments.

Spoken Language Understanding (SLU) aims at extracting semantic representations from speech signals or speech transcriptions of utterances in natural language (De Mori, 1997).

We evaluate our PLMs on the French corpus MEDIA (Bonneau-Maynard et al., 2006). This corpus is made up of documents on the topic of hotel information and reservations in France, and is made up of 1,250 human-machine dialogues transcribed and annotated with 76 semantic concepts.

MEDIA has been used extensively in recent years for French SLU, both for statistical and neural models, and both in cascade systems, where an Automatic Speech Recognizer (ASR) feeds a Natural Language Understanding (NLU) module (Raymond et al., 2006; Dinarelli et al., 2009b,a; Quarteroni et al., 2009; Hahn et al., 2011; Caubrière et al., 2020; Ghannay et al., 2021), and *end-to-end* systems based on neural networks (Dupont et al., 2018; Serdyuk et al., 2018; Dinarelli et al., 2017; Lugosch et al., 2019; Caubrière et al., 2019; Dinarelli et al., 2020; Pelloin et al., 2021; Desot et al., 2022). In general, SLU focuses on extracting semantics from speech signals, while NLU addresses the problem of extracting semantics from text. Since in this work we assess the ability of SSL models to encode text, we perform semantic extraction from speech transcriptions, and henceforth refer to this task as NLU.

2.2. Legal Tasks

ECtHR-French: European Court of Human Rights We construct and release a dataset of legal judgements from the European Court of Human Rights in French. To do so, we follow the methodology of Aletras et al. (2016) and Chalkidis et al. (2019), who released a similar dataset in English. English and French are the two official languages of the court, even though claims can be submitted in any official language of a state of the Council of Europe. We extracted ∼10k judiciary decisions available in French on the ECtHR website.

A typical document contains: (i) a description of the facts and applicable national laws (ii) motivations for the decision (iii) the decision itself. A decision (iii) states whether an article or a protocol from the European Convention on Human Rights was violated. A document may have 0 (no violation was found) or many labels (several human rights violations were found). Following Chalkidis et al. (2019), we cast the task as a multilabel prediction task: predicting the decision (iii) from the description of the facts (i).

We construct training examples by recovering the structure of the documents using regular expressions (identifying i-ii-iii). We exclude labels that have fewer than 100 occurrences in the data, as well as documents that are too short (they often contain only references to other documents, typically appendices).

After these steps, the dataset contains 9,575 examples. We took the 10% most recent documents (2018 onwards) to form the test set. We randomly split the rest of the documents into a train set (81% of the total) and a development set (9% of the total, see Table 1 for details).

OACS: Identifying unfair clauses in contracts The OACS corpus[6] consists of 4,517 consumer contract clauses labelled as either 'unfair' or 'fair'. The corpus also includes some clause metadata such as the type of contract (vehicle rent, online service, conditions of use, etc.), and the legal basis grounding the labelling. The dataset has a

[6] https://www.jeuxdemots.org/OACS/oacs.php

Creative Commons 0 (CC0) license and was gathered by legal experts, who also constructed artificial examples by modifying real clauses to shift their labels. The task consists simply in predicting whether a clause is fair or unfair according to French law and jurisprudence (binary classification).

Swiss Judgement predictions We use the French part of the Swiss Judgement Prediction dataset introduced by Niklaus et al. (2021). This dataset contains 31k decisions from the Federal Supreme Court of Switzerland, the last level of appeal in Switzerland. The task consists in the binary classification of the facts of a case as either a dismissal or an approval.

2.3. Biomedical Tasks

As detailed in Table 1, the biomedical evaluation benchmark we use in this work involves three different kinds of downstream task; sequence classification, token classification, and semantic textual similarity.

2.3.1. Sequence Classification

Biomedical sequence classification tasks involve a problem formulation whereby each element of a dataset has a single correct label associated with it. Our evaluation benchmark includes two distinct medicine-related tasks of this kind.

FrenchMedMCQA Multiple-Choice Question Answering involves choosing the correct answer from a list of available options. Automated question answering, particularly in the biomedical domain, requires advanced reading comprehension skills and the use of external sources of knowledge (Jin et al., 2022). FrenchMedMCQA (Labrak et al., 2022) is composed of 3,105 questions taken from the French medical specialization exams in pharmacy, with 2,025 multiple-answer questions and 1,080 single-answer questions. For each question, there are 5 different options to choose from (labelled from A to E), with at least one of the options being correct.

Medical Question Categorization (MQC) Labforsims (Laleye et al., 2020) is a corpus of French medical conversations annotated for a virtual patient dialogue system, including medical consultation interactions. We use this corpus to construct a sequence classification task that consists in classifying doctors' questions into one of seven categories: *Aim of Consultation, Personal Data, Medical History, Symptoms, Lifestyle, Treatments*, and *Unknown*. Laleye et al. (2020) used augmented datasets and reported the results of experiments using Convolutional Neural Networks and FastText (Bojanowski et al., 2017). We do not have access to this augmented dataset, and therefore used the publicly-available single-turn dataset, which contains 2,701 questions.

2.3.2. Token Classification

In token classification, the problem formulation associates a label with each token in a given sequence. Token classification often forms the backbone of many applied NLP tasks such as Named Entity Recognition (NER) and Word Sense Disambiguation (WSD).

CAS/ESSAI CAS (Grabar et al., 2018) and ESSAI (Dalloux et al., 2021) are corpora of clinical cases in French for which a subset is annotated with part-of-speech tags as well as semantic biomedical annotations (UMLS concepts, negation, and uncertainty). We evaluate our PLMs on three token-classification tasks from these corpora: CAS-POS and ESSAI-POS, which directly use the (non-biomedical) POS tags provided, and CAS-SG, which involves classifying each word in a document according to the most relevant UMLS semantic group.

QuaeroFrenchMed The QUAERO French Medical Corpus (Névéol et al., 2014) consists of a collection of biomedical documents annotated at the entity and concept levels for entity recognition and/or token classification tasks. This corpus is in fact made up of two distinct sub-corpora; a collection of 2,500 MEDLINE article titles and a collection of 1,520 medication descriptions from the European Medicines Agency (EMEA). These form the basis for two of the token classification tasks in our evaluation benchmark, referred to henceforth simply as MEDLINE and EMEA. We use the publicly-available annotations[7] of these corpora, which are labelled at the token level with ten different NER tags defined according to semantic types from the UMLS medical ontology (Bodenreider, 2004).

European Clinical Case Corpus (E3C) We also implement a token classification task based on the annotations provided as part of the European Clinical Case Corpus (Minard et al., 2021; Magnini et al., 2020). The E3C is divided into three 'layers'; layer 1 being manually annotated clinical cases, layer 2 containing automatic annotations according to the same schema, and layer 3 containing non-annotated documents. We make use of the

[7] https://huggingface.co/datasets/mnaguib/QuaeroFrenchMed

annotations in layers 1 and 2 indicating the presence of clinical entities in the text to construct a 3-class (B-I-O) token classification task, where the model is tasked with identifying which tokens form part of annotated clinical entities (once again defined according to the UMLS). Layers 2 and 1 are used as the train/validation and test partitions respectively. This token classification task is referred to as E3C-NER in our experiments. Layer 3 of this corpus is used for pretraining the biomedical models (see section 3.2).

2.3.3. Semantic Textual Similarity

The goal of Semantic Textual Similarity (STS) tasks is to accurately measure the extent to which pairs of text snippets are similar to one another in a conceptual/semantic way. In the clinical domain, STS can enable the detection and elimination of redundant information (Wang et al., 2020b).

CLISTER For STS evaluation, we use CLISTER (Hiebel et al., 2022), constructed based on the CAS corpus (see Section 2.3.2). It contains 1,000 sentence pairs manually annotated with a similarity score from 0 to 5.

3. Pretraining

3.1. Architecture

BERT base with Linformer We use a classical BERT base (Devlin et al., 2019) architecture, replacing the standard transformer layers with LinFormer (Wang et al., 2020a) layers and compressed the key-value initial layers into a 256-dimensional space. As recommended in the original Linformer paper, we also used parameter sharing between projections : headwise, key-value and layerwise sharing. We compared the efficiency of Linformer at inference-time against a standard attention layer following Wang et al. (2020a)'s protocol and observed an increase in speed (2.5x) and memory (3x) with sequences of 4096 tokens. However, we did not observe significant gains with sequences of length 512. All our experiments were performed on AMD MI250 GPUs.

Pretraining and models We use only the masked language modelling objective to train the models. As for the tokenizer we train specific BPE models (Sennrich et al., 2015) for each domain with a vocabulary size set to 50K tokens. We train one standard 512-token model from scratch for the legal and medical domains: Jargon-legal and Jargon-biomed, as well as one 4096-token model (Jargon-*-4096). Additionally, for control experiments, we train a multi-domain model, Jargon-multi-domain-base, on all the domain-specific data, and a generic model Jargon-general-base. For further comparison, we also trained biomedical models (Jargon-NACHOS, Jargon-NACHOS-4096) on the NACHOS corpus (Labrak et al., 2023). Given the small size of the speech transcription corpus (section 3.2), we do not train a specific model for this domain. Speech-related tasks are evaluated using the general and the multi-domain models, the latter including speech transcriptions in the training data. Finally, as for further pretraining, we continue the training of Jargon-general-base on specialized domain data and denote these models Jargon-general-*.

3.2. Pretraining Data

General Data The Jargon-general-base was trained on a general corpus composed of French Wikipedia articles,[8] French literature from the Gutenberg project[9] and a 5GB sample of the French partition of the C4 multilingual corpus (Xue et al., 2020). This mixed corpus contains 8.5GB (after preprocessing) of textual data from encyclopedic, literature and general web sources.

Speech Data ASR benchmark datsets often contain read speech (or sometimes prepared speech), as opposed to speech arising from spontaneous interactions. The scarcity of spontaneous speech corpora greatly limits the overall size of the pretraining data; of the three applications addressed in this work, spontaneous speech faces the most acute paucity of freely-available data. For transcribed speech pretraining data, we use the following eight corpora: ESLO2 (Eshkol-taravella et al., 2011), EPAC2 (Estève et al., 2010), NCCFr (Torreira et al., 2010), MPF (Candea, 2018), TCOF (Canut et al., 2010), ESTER1 (Galliano et al., 2005), ESTER2 (Galliano et al., 2009) and CFPP (Branca-Rosoff and Lefeuvre, 2016). Most of these corpora were constructed for sociolinguistic studies and feature realistic interactions. In total, this gives us approximately 300 hours (25 MB) of transcribed speech.

Legal Data The bulk of our legal pretraining data come from open data repositories maintained by DILA,[10] a French governmental information mangagement agency. They contain several types of legal and metalegal data: decisions from judiciary institutions, parliamentary debates, and official directives.

[8] We use the official dump from 29/11/2022

[9] https://www.gutenberg.org/

[10] https://www.dila.premier-ministre.gouv.fr/repertoire-des-informations-publiques/les-donnees-juridiques

Other sources of data include the BSARD corpus (Louis and Spanakis, 2022), which contains 23k statutory articles from Belgian law, as well as the French version of two types of texts from the European Parliament's open data repository: DCEP and DGT-Translation Memory.[11] In total, we make use of 18GB of training data (DILA: 17GB, bsard: 20MB, European Parliament: 1GB).

Biomedical Data The biomedical training corpus used in this work was extracted from three different types of source documents:

- Scientific articles from the biomedical field, obtained via the open-access archives of French scientific articles provided by HAL[12] (600M tokens) and ISTEX[13] (190M tokens), as well as the French scientific articles provided as part of the BioWMT shared task parallel corpora (3M tokens).[14]
- Publicly-available clinical cases and medication descriptions, as compiled in the European Clinical Case Corpus (E3C Minard et al., 2021; Magnini et al., 2020, 63M tokens).
- General articles on health and medicine scraped from French Wikipedia (3M tokens).

Since a large proportion of the textual data scraped from HAL, ISTEX, and Wikipedia was extracted from PDF files and web pages, we implemented a relatively aggressive text-cleaning pipeline to remove references to figures, URLs, incomplete sentences, artifacts of the Optical Character Recognition process used, and non-French text passages. In total, the biomedical pretraining corpus contains 858M tokens (5.4GB).

4. Experiments

In this section, we present the systems and architectures we use for each task, as well as the results of our experiments. In all experiments, we fine-tune each PLM end-to-end using the Adam optimization algorithm for backpropagation, with PyTorch's default parameters.

4.1. Speech-related Tasks

4.1.1. Experimental Settings

Automatic speech recognition The speech recognition model is composed of a CRDNN (Sainath et al., 2015) as the encoder, taking mel filter bank as input. This CRDNN uses 3 CNN blocks, 5 LSTM layers and 2 fully-connected layers. The decoder is a single-layer GRU with attention, taking as input both the encoder (the signal representations) and the previous words. The acoustic model was trained with the Adadelta optimizer, using a batch size of 12 and a learning rate of 1. As an external non-causal language model, we implemented *masked language scoring* (Salazar et al., 2020). This algorithm allows a non-causal language model to give a score to a sentence through masked language modeling. We use the different language models to rescore the beam search once per sentence during decoding, thus getting the most probable sentence from the combination of the speech recognition model and the language model. In contrast to Salazar et al. (2020), we did not fine-tune each PLM on the downstream corpus.

Dependency parsing The dependency parsing model used for this task is a graph-based biaffine parser as defined by Dozat and Manning (2016). The downstream model is composed of a 3-layer bidirectional LSTM and 4 multi-layer perceptrons as in Dozat and Manning (2016). We use the pretrained models as feature extractors, where each word is represented by its last subword embedding. We fine-tune the pretrained models with a learning rate of 2e-5 for 30 epochs on the Orféo Corpus and 64 epochs on the Paris Stories corpus.

Spoken Language Understanding The neural architecture used for the SLU experiments is a multi-task architecture where the input to the encoder consists of manual transcriptions of spoken utterances. We use two outputs as tasks, one containing only BIO chunking; the other with semantic labels added. In preliminary experiments, we found that the first output aids with generalization and precision in the second output. The two tasks are learned jointly with a compound *NLL*-loss function. Both the architecture and the loss function are similar to (Gugliotta et al., 2020), except that in this work we use a transformer architecture instead of a LSTM as it showed higher performance on the NLU task.

Inspired by Martin et al. (2020), we use the mean of the last 4 hidden layers as input to our model's encoder.

4.1.2. Results

The relevant experimental results are reported in Table 2. We use classical evaluation measures for each task: Unlabelled Attachment Score (UAS), Labelled Attachment Score (LAS) and part-of-speech accuracy (POS) for dependency parsing, Word Error Rate (WER) and Character Error

[11] https://joint-research-centre.ec.europa.eu/language-technology-resources_en
[12] hal.archives-ouvertes.fr/
[13] https://www.istex.fr/
[14] https://github.com/biomedical-translation-corpora/corpora

Task →	Dependency Parsing CEFC-ORFÉO			Dependency Parsing Paris Stories			ASR		SLU
Model ↓	POS↑	UAS↑	LAS↑	POS↑	UAS↑	LAS↑	WER↓	CER ↓	SER ↓
ASR (No LM)	-	-	-	-	-	-	14.1±0.17	5.6±0.07	-
FlauBERT-base	98.3±0.09	89.9±0.20	87.4±0.19	97.2±0.12	**80.2±0.2**	**76.5±0.33**	13.9±0.16	5.5±0.07	9.77
FlauBERT-large	**98.4±0.05**	**89.9±0.14**	**87.4±0.14**	**97.3±0.03**	79.8±0.11	76.3±0.18	>100±0.00	>100±0.00	10.05
CamemBERT-base	98.4±0.05	89.9±0.20	87.4±0.20	96.7±0.21	79.4±0.17	75.3±0.21	13.3±0.17	5.3±0.07	**8.54**
CamemBERT-large	98.4±0.11	89.9±0.15	87.4±0.14	96.9±0.09	79.0±0.25	75.1±0.25	13.3±0.16	5.3±0.07	8.99
FlauBERT-oral	98.3±0.04	89.8±0.13	87.2±0.15	96.4±0.07	78.3±0.12	74.3±0.12	>100±0.00	>100±0.00	9.67
Jargon-multi-domain-base	98.2±0.04	87.4±0.47	84.7±0.46	97.0±0.07	78.5±0.11	74.6±0.14	13.1±0.15	5.3±0.05	11.26
Jargon-general-base	98.2±0.06	87.7±0.26	85.1±0.29	97.1±0.08	78.9±0.16	75.1±0.17	**12.4±0.14**	**5.1±0.06**	10.35

Table 2: Results on the speech-related tasks.

Rate (CER) for ASR, Concept Error Rate (reported as Semantic Error Rate or SER to avoid the confusion with CER) for SLU.

Dependency parsing: CEFC-ORFEO On the CEFC-ORFEO dataset, the majority of the models seem to hit a ceiling of approximately 87 for the LAS metric. One possible explanation for this is the silver data used to train the different model, which may force models to simply learn to copy the model used to annotate the silver data, thus not being able to reach higher score on this corpus. An interesting finding is that the model FlauBERT-oral trained from scratch on a massive (automatically transcribed) prepared speech corpus does not reach a higher score than other language models after fine-tuning. This may be explained by the mismatch between spontaneous speech contained in the corpus, compared to the prepared speech (mostly transcribed TV shows) used to pretrain FlauBERT-oral.

Dependency parsing: Paris Stories The Paris Stories experiments show more variation in results; the Flaubert models perform best, followed by the Camembert model and finally the Jargon models. Overall, the performance of the Jargon models lag slightly behind the state-of-the-art for dependency parsing on transcribed speech.

Speech recognition In contrast to dependency parsing, the Jargon models clearly outperform both FlauBERT and Camembert on the speech recognition task. The Jargon-general-base shows a relative decrease of 12% in WER, the largest in the experiment. Another interesting point is the two failed experiments, using FlauBERT-large or FlauBERT-oral cause the sequence to sequence network of the speech recognition model to degenerate, reaching WER as high as 250. This is especially surprising in the case of the FlauBERT-oral model since it was trained on speech data and thus would be expected to produce representations closer to the target domain.

Spoken Language Understanding First, we note that we have a strong baseline, though it is not a state-of-the-art model (Dinarelli and Grobol, 2019). Among SSL models taken from the literature, surprisingly, CamemBERT-base shows the highest performance on the test data. Both Jargon models substantially outperform our baseline on the test data. However, they lag behind CamemBERT. We hypothesize that this is due to (i) the very small amount of transcribed data used to train our SSL models and (ii) the use of Linformer attention in the Jargon models.

4.2. Legal Tasks

Given that all of our legal-domain benchmarks are sequence classification tasks, we use the standard classification fine-tuning architecture for all datasets. This involves feeding the vector corresponding to the [CLS] representation into a single projection head (linear layer) for multi-label and multi-class classification. All experiments are run in mixed precision with a batch size of 32, learning rate warmup over 10% of the training steps, and a linear learning rate decay. Due dataset imbalanced, we selected the best checkpoints for each run based on macro-F1 scores on the validation set. The results reported for these tasks are the average of five runs initialized with varying random seeds.

We compare our models to existing French pretrained models for the legal domain, namely Camembert-Legal (Louis and Spanakis, 2022) and Juribert (Douka et al., 2021).

We present all results in Table 3. For the ECtHR-FR dataset and OACS, surprisingly enough, general purpose models outperform all legal models trained from scratch (CamemBERT-Legal used further pretraining from Camembert), and the larger models (FlauBERT-large and CamemBERT-large) obtain the best scores. Our Jargon models are slightly better than Juribert on these tasks. For the Swiss Judgment Prediction (SJP) dataset, the pattern is similar except for the Jargon-4096 model that is able to take the longer context into account, and outperforms every model by a large margin.

Task $\rightarrow$	**ECtHR-FR**		**OACS**		**SJP**	
Model $\downarrow$	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Juribert-small	49.0 $\pm$ 1.2	45.5 $\pm$ 1.3	42.2 $\pm$ 2.9	56.1 $\pm$ 1.9	26.5 $\pm$ 3.7	56.5 $\pm$ 1.7
Juribert-base	51.1 $\pm$ 1.3	46.3 $\pm$ 0.4	38.8 $\pm$ 4.0	53.4 $\pm$ 2.5	23.5 $\pm$ 1.6	55.1 $\pm$ 0.8
legal-CamemBERT	54.3 $\pm$ 2.1	49.2 $\pm$ 2.9	51.1 $\pm$ 1.9	61.2 $\pm$ 0.9	30.2 $\pm$ 1.0	57.9 $\pm$ 0.9
FlauBERT-base	54.4 $\pm$ 2.2	50.5 $\pm$ 1.8	48.7 $\pm$ 2.3	59.6 $\pm$ 1.9	29.4 $\pm$ 7.4	58.2 $\pm$ 3.0
FlauBERT-large	58.0 $\pm$ 1.7	54.7 $\pm$ 1.8	**51.9** $\pm$ 4.0	61.6 $\pm$ 0.9	**32.7** $\pm$ 2.8	60.4 $\pm$ 1.4
CamemBERT-base	55.0 $\pm$ 0.9	50.6 $\pm$ 1.0	51.0 $\pm$ 2.8	**61.7** $\pm$ 1.2	32.1 $\pm$ 2.8	59.1 $\pm$ 1.4
CamemBERT-large	**60.3** $\pm$ 0.9	**58.0** $\pm$ 1.2	51.2 $\pm$ 2.4	60.9 $\pm$ 2.0	32.5 $\pm$ 2.2	59.7 $\pm$ 0.9
Jargon-general-base	52.5 $\pm$ 1.2	42.9 $\pm$ 2.3	35.1 $\pm$ 4.0	50.8 $\pm$ 1.2	24.4 $\pm$ 2.4	55.1 $\pm$ 0.9
Jargon-multi-domain-base	53.5 $\pm$ 0.9	44.5 $\pm$ 0.8	41.8 $\pm$ 5.2	55.6 $\pm$ 3.3	29.3 $\pm$ 2.1	58.1 $\pm$ 0.9
Jargon-general-legal	50.5 $\pm$ 1.4	43.1 $\pm$ 2.5	34.5 $\pm$ 24.1	49.9 $\pm$ 0.5	22.8 $\pm$ 3.8	54.5 $\pm$ 1.4
Jargon-legal	51.6 $\pm$ 0.8	44.6 $\pm$ 0.9	40.6 $\pm$ 1.9	51.6 $\pm$ 1.9	27.6 $\pm$ 1.7	56.7 $\pm$ 0.7
Jargon-legal-4096	52.6 $\pm$ 0.7	45.9 $\pm$ 2.4	40.5 $\pm$ 2.4	54.1 $\pm$ 1.5	47.5 $\pm$ 1.5	**68.2** $\pm$ 0.5

Table 3: Results on legal tasks (test sets).

4.3. Biomedical Tasks

In these experiments, we compare the Jargon PLM family with 13 others, involving a mixture of models trained specifically for the medical domain in both French and English, as well as general-domain French models. Table 4 contains a summary of the most salient results from these experiments.

4.3.1. Sequence Classification

For the medical-domain sequence classification tasks, we use the same architecture as for our legal-domain benchmarks (see Section 4.2).

FrenchMedMCQA We build the sequence classification input sequences for MCQA using the following format: *[CLS] <question> [SEP] (A) <answer.a> [SEP] (B) <answer.b> [SEP] (C) <answer.c> [SEP] (D) <answer.d> [SEP] (E) <answer.e> [EOS]*, following Labrak et al. (2022)'s approach. We finetuned all models for 10 epochs, using an effective batch size of 32 and a learning rate of 2e-5. We used the Exact Match Ratio (EMR), which corresponds to the proportion of exact correct answers, and the Hamming score, which is similar to multi-label accuracy.

Medical Question Categorization (MQC) Formalizing question categorization as a text classification task, we fine-tuned all the biomedical models listed for 20 epochs, with early stopping, meaning that training stops when the accuracy score deteriorates for 2 consecutive epochs. We used the same batch size and learning rate as for FrenchMedMCQA.

4.3.2. Token classification

For the six token classification tasks – CAS-POS, ESSAI-POS, CAS-SG, MEDLINE, EMEA, and E3C-NER – we carry out fine-tuning by adding a linear classification layer to the BERT model output that projects the embeddings associated with labelled input tokens into a n-dimensional vector, where n is the number of classes for the task in question. Each model was fine-tuned for 2,000 update steps on each train dataset, with learning rate 2e-5 and a batch size of 16.

CLISTER We use the SentenceTransformers framework (Reimers and Gurevych, 2019) to fine-tune sentence embedding methods for the CLISTER STS task. The sentence-transformer architecture consists of two layers: a pretrained transformer model and a mean-pooling layer. We fine-tuned all models for 10 epochs, with a batch size of 16 and a learning rate of 2e-5. Following Hiebel et al. (2022), we used Spearman correlation as the evaluation metric.

For all the above-described tasks, the results reported in Table 4 are the average of five independent runs initialized with varying random seeds.

4.3.3. Results

FrenchMedMCQA In terms of exact-matching evaluation, we see that specialized biomedical PLMs, notably Jargon-NACHOS-4096 and CamemBERT-bio-base, hold a distinct advantage on this task. The Hamming measure, which takes into account partially correct answers, shows more mixed results among general-domain and specialized models.

MQC For the MQC task, the performance of Jargon-NACHOS-4096 is in line with or better

Task / Metric →	FrenchMedMCQA		MQC	CAS-POS	ESSAI-POS	CAS-SG	MEDLINE	EMEA	E3C-NER	CLISTER
Model ↓	EMR	Hamming	Accuracy	Macro F1	Macro F1	Weighted F1	Weighted F1	Weighted F1	Weighted F1	Spearman
BioBERT†[15]	15.2±1.9	34.9±1.9	93.5±1.0	96.0	95.4	73.7	82.6	96.1	93.1	79.2
PubMedBERT† (Gu et al., 2020)	15.6±1.6	34.5±0.8	92.7±1.8	94.8	95.4	74.6	85.3	95.9	92.8	80.6
ClinicalBERT† (Wang et al., 2023)	13.7±0.2	34.0±0.7	92.2±1.2	95.5	95.7	72.4	83.8	96.2	92.7	84.0
BioClinicalBERT† (Alsentzer et al., 2019)	16.2±2.4	35.3±2.1	93.6±1.0	94.9	95.5	73.8	83.9	95.7	93.0	78.8
SapBERT-XL*† (Liu et al., 2021)	15.3±1.3	34.5±1.3	**95.3**±0.7	<u>96.9</u>	**96.6**	74.2	84.8	96.0	93.3	86.8
DrBERT-7GB* † (Labrak et al., 2023)	17.4±0.8	36.1±1.1	94.6±0.4	96.5	<u>96.5</u>	76.2	83.9	96.4	93.4	**88.1**
DrBERT-4GB* † (Labrak et al., 2023)	14.9±1.0	34.8±1.5	93.5±1.1	96.7	**96.6**	76.1	84.9	96.5	93.7	<u>87.6</u>
CamemBERT-bio-base* † (Touchent et al., 2023)	<u>17.5</u>±2.7	<u>36.8</u>±1.6	93.4±1.2	<u>96.9</u>	**96.6**	**76.9**	**86.4**	96.5	<u>94.0</u>	87.1
FlauBERT-base* (Le et al., 2020)	15.3±2.0	34.1±1.9	90.4±5.4	96.7	95.6	67.4	83.7	84.6	93.6	83.6
FlauBERT-large* (Le et al., 2020)	14.6±1.4	33.9±1.3	91.7±5.4	96.5	96.2	67.2	83.6	85.3	93.1	75.0
CamemBERT-base* (Martin et al., 2020)	14.0±0.8	34.7±1.2	93.3±1.6	**97.0**	**96.6**	<u>76.4</u>	85.8	<u>96.7</u>	93.9	86.0
CamemBERT-oscar-4gb* (Martin et al., 2020)	14.4±1.3	34.0±1.5	94.1±0.7	96.9	96.4	75.7	85.7	96.6	93.8	84.5
CamemBERT-ccnet-4gb* (Martin et al., 2020)	16.5±1.1	**37.0**±1.2	<u>95.2±1.0</u>	96.8	**96.6**	75.9	<u>85.9</u>	94.2	**94.2**	86.5
Jargon-general-base*	12.9±0.8	32.6±1.3	76.7±6.3	96.6	96.0	69.4	81.7	96.5	91.9	78.0
Jargon-biomed*†	15.3±1.2	34.5±1.1	91.1±1.4	96.5	95.6	75.1	83.7	96.5	93.5	74.6
Jargon-biomed-4096*†	14.4±1.1	33.8±1.8	78.9±18.6	95.6	95.9	73.3	82.3	96.3	92.5	65.3
Jargon-general-biomed*†	16.1±1.0	34.8±0.8	69.7±3.1	95.1	95.1	67.8	78.2	96.6	91.3	59.7
Jargon-multi-domain-base*†	14.9±1.9	34.2±1.5	86.9±3.5	96.3	96.0	70.6	82.4	96.6	92.6	74.8
Jargon-NACHOS*†	13.3±0.1	32.7±0.8	90.7±7.5	96.3	96.2	75.0	83.4	**96.8**	93.1	70.9
Jargon-NACHOS-4096*†	**18.4**±1.4	36.2±1.4	93.2±1.5	96.2	95.9	74.9	83.8	**96.8**	93.2	74.9

Table 4: Test set results on the biomedical tasks. * denotes models pretrained on French-language data, † those that were trained on biomedical corpora (fr/en). We exclude the standard deviation for tasks for which less than half of the values were above 0.005.

than that of most previous models, while it is 1–2 points lower than that of the best model, SapBERT-XL. The performance of Jargon-general-base, Jargon-general-biomed is below 80%, while Jargon-biomed-4096 showed a large performance difference in one of the five runs, presented as a standard deviation of 18.61.

Token classification We report either macro-F_1 or weighted F_1 for the token classification tasks. For NER, the F_1 is computed at the token level and not at the mention level. The results on our six token classification tasks shows a slight advantage for the CamemBERT-bio-base on average, although overall general and specific models performed similarly. Among the Jargon models, we see that the biomedical-only Jargon-NACHOS-4096 and Jargon-biomed tend to give the highest F1 scores.

CLISTER For the Spearman correlation coefficient, biomedical models performed the best, though the difference between them and the best performing French general model (CamemBERT-ccnet-4gb) and the best English biomedical model (SapBERT-XL) is small. Furthermore, none of the Jargon models performed very well on this task, even when trained on the same corpus, NACHOS, as the best-performing model, DrBERT-7GB. It is unclear therefore what explains this discrepancy, as Jargon models were competitive for most other tasks.

5. Discussion and Conclusion

In recent years, the interest in the development of pretrained language models on specialized domains, especially in the biomedical and legal domains, has greatly increased due to the widening scope of potential applications. Thus, many specialized models have now been publicly released along with domain specific corpus and tasks. These models were either trained from scratch or had their pretraining continued on in-domain data, which requires additional human and computing resources. Now, when taking a look at the overall results of our experiments, one might observe that the gain is quite humble. Indeed, out of the 16 tasks that we evaluated in this work, only 11 are led by specialized models. Furthermore, the average gain over all tasks is less than 1% meaning that even when specialized domain outperform general trained ones, the gain is rather small. This may highlight a limitation to the widely adopted methodology of (1) collecting more domain-specific data and (2) training new PLMs on these data.

In conclusion, this work investigates the application of large language models in specialized domains: biomedical, legal, and spontaneous speech. Our first contribution is the introduction of novel models employing an efficient attention computation architecture (Linformer), allowing us to extend the context size up to 4096 tokens. Additionally, we investigated and experimented multiple training configurations: further pretraining versus training from scratch, single-domain versus multi-domain training. Our second main contribution is the evaluation of state-of-the-art models across a wide range of tasks, including newly introduced ones, from the three domains gathered into a unified benchmark. All of our models, data, and code are made publicly available for the purpose of reproducibility.

Limitations and Ethics Statement

As this work aims to prioritise the breadth of our evaluation benchmarks, we constructed our experiments as comparitive evaluations between models rather than optimisation problems for practical application. Therefore, in the interest of limiting the computational cost of the experiments, we restricted ourselves to a very limited set of hyperparameters that were applied to all models in our comparitive experiments. Consequently, it is likely that some models may achieve better results with more refined parameter settings.

Regarding the speech domain, the experiments were constrained by complications associated with data acquisition, primarily due to copyright and GPDR policies.

Acknowledgements

This work was performed using the Jean Zay and Adastra clusters from GENCI-IDRIS (Grant 2022 A0131013801). This work was partially funded by the French National Research Agency, through grants Pantagruel (ANR-23-IAS1-0001), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), PROPICTO (ANR-20-CE93-0005), Lawbot (ANR-20-CE38-0013), and the Swiss National Science Foundation (grant PROPICTO N°197864).

6. Bibliographical References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Antoine Caubrière, Sahar Ghannay, and et al. 2020. Error analysis applied to end-to end spoken language understanding. In *ICASSP*, Barcelona, Spain.

Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève. 2019. Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *Proc. Interspeech 2019*, pages 1198–1202.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Iker de la Iglesia, Aitziber Atutxa, Koldo Gojenola, and Ander Barrena. 2023. Eriberta: A bilingual pre-trained language model for clinical natural language processing.

Renato De Mori. 1997. *Spoken Dialogues with Computers*. Academic Press, Inc., Orlando, FL, USA.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model.

Thierry Desot, François Portet, and Michel Vacher. 2022. End-to-End Spoken Language Understanding: Performance analyses of a voice command task in a low resource setting. *Computer Speech and Language*, 75:101369.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Dinarelli and Loïc Grobol. 2019. Hybrid neural models for sequence modelling: The best of three worlds. In *arXiv Technical Report*.

Marco Dinarelli, Nikita Kapoor, Bassam Jabaian, and Laurent Besacier. 2020. A data efficient end-to-end spoken language understanding architecture. In *ICASSP*.

Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009a. Concept segmentation and labeling for conversational speech. In *Interspeech*, Brighton, U.K.

Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009b. Re-ranking models based-on small training data for spoken language understanding. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1076–1085, Singapore. Association for Computational Linguistics.

Marco Dinarelli, Vedran Vukotic, and Christian Raymond. 2017. Label-dependency coding in simple recurrent networks for spoken language understanding. In *Interspeech*, Stockholm, Sweden.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Yoann Dupont, Marco Dinarelli, and Isabelle Tellier. 2018. Label-dependencies aware recurrent neural networks. In *Computational Linguistics and Intelligent Text Processing*, pages 44–66, Cham. Springer International Publishing.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or train from scratch? comparing pre-training strategies of BERT in the medical domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2626–2633, Marseille, France. European Language Resources Association.

Sahar Ghannay, Antoine Caubrière, and et al. 2021. Where are we in semantic concept extraction for spoken language understanding? $\star$. In *In SPECOM 2021*, Saint Petersburg, Russia.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. 2020. Multi-task sequence prediction for Tunisian Arabizi multi-level annotation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 178–191, Barcelona, Spain (Online). Association for Computational Linguistics.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehnen, Renato de Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. 2011. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6):1569–1583.

Nicolas Hervé, Valentin Pelloin, Benoit Favre, Franck Dary, Antoine Laurent, Sylvain Meignier, and Laurent Besacier. 2022. Using ASR-generated text for spoken language modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 17–25, virtual+Dublin. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Comput. Surv.*, 55(2).

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In *Proc. Interspeech 2019*, pages 814–818.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Alexis Nasr, Franck Dary, Frederic Béchet, and Benoit Fabre. 2020. Annotation syntaxique automatique de la partie orale du Orféo. In *Langages*.

Valentin Pelloin, Nathalie Camelin, Antoine Laurent, Renato De Mori, Antoine Caubrière, Yannick Estève, and Sylvain Meignier. 2021. End2end acoustic to semantic transduction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7448–7452. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Silvia Quarteroni, Giuseppe Riccardi, and Marco Dinarelli. 2009. What's in an ontology for spoken language understanding. In *Proc. Interspeech 2009*, pages 1023–1026.

Christian Raymond, Frédéric Béchet, Renato De Mori, and Géraldine Damnati. 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48(3-4):288–304.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Rian Touchent, Laurent Romary, and Eric De La Clergerie. 2023. Camembert-bio : Un modèle de langue français savoureux et meilleur pour la santé. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 323–334, Paris, France. Association pour le Traitement Automatique des Langues.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020a. Linformer: Self-attention with linear complexity.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020b. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. Glm-130b: An open bilingual pre-trained model.

7. Language Resource References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. In *International Conference on Language Resources and Evaluation*.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15).

H. Bonneau-Maynard, C. Ayache, F. Bechet, A. Denis, A. Kuhn, F. Lefevre, D. Mostefa, M. Quignard, S. Rosset, C. Servan, and J. Villaneau. 2006. Results of the French evalda-media evaluation campaign for literal understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Sonia Branca-Rosoff and Florence Lefeuvre. 2016. Le corpus de français parlé parisien des années 2000 : constitution, outils et analyses. le cas des interrogatives indirectes. *Corpus*, 15.

Maria Candea. 2018. Françoise gadet (dir.) les parlers jeunes dans l'Île-de-france multiculturelle paris, ophrys, 2017, 174 p. *Langage et société*, 163:192.

Emmanuelle Canut, Virginie André, and Bertrand Gaiffe. 2010. Mise à disposition de corpus oraux interactifs : le projet TCOF (traitement des corpus oraux en français). *Pratiques : théorie, pratique, pédagogie*, Interactions et Corpus Oraux:147–148.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Iris Eshkol-taravella, Olivier Baude, Denis Maurel, Linda Hriba, Celine Dugua, and Isabelle Tellier. 2011. Un grand corpus oral disponible : le corpus d'orléans 1968-2012 [a large available oral corpus: Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, 52(3):17–46.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas.

2010. The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier. 2005. The ESTER phase ii evaluation campaign for the rich transcription of french broadcast news. In *Proc. Interspeech 2005*, pages 1149–1152.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Proc. Interspeech 2009*, pages 2583–2586.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. Cas: French corpus with clinical cases. In *LOUHI 2018-The Ninth International Workshop on Health Text Mining and Information Analysis*, pages 1–7.

Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2022. Clister: A corpus for semantic textual similarity in french clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4306–4315.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fréjus A. A. Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020. A French medical conversations corpus annotated for a virtual patient dialogue system. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 574–580, Marseille, France. European Language Resources Association.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases.

Anne-Lyse Minard, Roberto Zanoli, Begoña Altuna, Manuela Speranza, Bernardo Magnini, and Alberto Lavelli. 2021. European clinical case corpus. Bruno Kessler Foundation.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. 2010. The Nijmegen corpus of casual french. *Speech Communication*, 52(3):201.