Actes de JEP-TALN-RECITAL 2024. Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-Dac, Julie Mauclair, José G. Moreno, Julien Pinquier

35èmes Journées d'Études sur la Parole (JEP 2024)
31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024)
26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)

https://jep-taln2024.sciencesconf.org

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

Mathieu BALAGUER, Nihed BENDAHMAN, Lydia-Mai HO-DAC, Julie MAUCLAIR, Jose G MORENO, Julien PINQUIER (Éds.)

Toulouse, France, 8 au 12 juillet 2024

Avec le soutien de

Actes JEP-TALN-RECITAL Toulouse 2024 – ISBN : 978-2-917490-37-2
Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues – ISBN : 978-2-917490-40-2

Préface

La conférence RECITAL permet à des jeunes chercheuse et chercheurs (étudiant · e · s en master ou début doctorat) de présenter leurs travaux et d'échanger avec la communauté TAL sur leurs méthodes et problématiques. La soumission de travaux préliminaires, de projets de thèse, et de travaux résultant des premiers mois de recherche (état de l'art et positionnement, objectifs et premières pistes, etc.) est encouragée. Pour l'édition 2024, 7 soumissions sur 11 ont été acceptées. Toutes les soumissions acceptées donneront lieu à un poster et trois d'entre elles donneront également lieu à une communication orale.

Comités

Comité de programme :

— Nihed Bendahman (UT3 - IRIT)
— Lydia-Mai Ho-Dac (UT2J - CLLE)

Comité scientifique :

— Mathieu Balaguer (Université Paul Sabatier - Toulouse III, IRIT - IRIT)
— Nihed Bendahman (Université Paul Sabatier - Toulouse III, IRIT - IRIT/Berger-Levrault)
— Christophe Benzitoun (ATILF)
— Catherine Berrut (LIG, Université Joseph Fourier Grenoble I)
— Sandra Bringay (LIRMM)
— Sylvie Calabretto (LIRIS)
— Jérôme Farinas (Université Paul Sabatier - Toulouse III, IRIT)
— Lydia-Mai Ho-Dac (Université Toulouse Jean Jaurès - CLLE)
— Sylvain Kahane (Modyco, Université Paris Ouest Nanterre & CNRS)
— Mael Lesavourey (Université Paul Sabatier - Toulouse III, IRIT)
— Cédric Lopez (Emvista)
— Jesus Lovon-Melgarejo (Université Paul Sabatier, IRIT)
— Andon Tchechmedjiev (EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales)

Table des matières

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

An evaluation of current benchmarking strategies for French biomedical language models

Felix Herron[1,2]
(1) Laboratoire d'Informatique Grenoble, 700 Av. Centrale, 38401 Saint-Martin-d'Hères, France
(2) Laboratoire d'Analyse et de Modélisation de Systèmes d'Aide à la Décision, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France
felix.herron@univ-grenoble-alpes.fr

ABSTRACT

We describe the current state of benchmarking for French language biomedical natural language processing (NLP). We note two important criteria in biomedical benchmarking: first, that a biomedical benchmark clearly simulate a specific use cases, in order to offer a useful evaluation of a biomedical model's real life applicability. Second: that a biomedical benchmark be created in collaboration with biomedical professionals. We note that many biomedical benchmarks, particularly in French, do not adhere to these criteria; however, we highlight other biomedical benchmarks which adhere better to those criteria. Furthermore, we evaluate some of the most common French biomedical benchmarks on an array of models and empirically support the necessity of domain-specific and language-specific pre-training for natural language understanding (NLU) tasks. We show that some popular French biomedical language models perform poorly and/or inconsistently on important biomedical tasks. Finally, we advocate for an increase in publicly available, clinically targeted French biomedical NLU benchmarks.

RÉSUMÉ

Évaluation de benchmarking actuel pour des modèles de langage biomédicaux français

Nous présentons dans cet article une réflexion à propos des tâches d'évaluation en traitement automatique des langues (TAL) biomédical et clinique pour la langue française. Nous soulignons l'insuffisance de référentiels reflétant des scénarios d'utilisation réels et concrets, limitant ainsi la pertinence de leurs résultats pour les professionnels de santé. De plus, il est réputé que certains sont élaborés sans la participation active de spécialistes du domaine. Notre examen d'une sélection de référentiels biomédicaux français classiques soutient le besoin d'un préentraînement spécifique au domaine biomédical en français destiné plus particulièrement aux tâches de compréhension du langage naturel (NLU). Nous montrons également que certains modèles préentrainés pour les domaines biomédicaux français affichent des performances médiocres voire incohérentes lorsqu'ils sont testés sur des tâches biomédicales courantes dans la littérature biomédicale française. En conclusion, nous plaidons pour une augmentation des référentiels librement disponibles et focalisés sur des situations cliniques réelles.

KEYWORDS : Benchmarking, biomedical language modeling, deep learning.

MOTS-CLÉS: Benchmarking, modélisation de langage biomédicale, apprentissage profond.

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 1–16.

1 Introduction

Since the advent of the Transformer architecture and the subsequent rise of deep language models (LMs), the power of natural language processing (NLP) models has significantly improved (Vaswani *et al.*, 2017) (Devlin *et al.*, 2019). This improvement has led to the application of LMs in various domains, such as grading at universities (Fuchs, 2023), policing the internet for hate speech (Plaza-del Arco *et al.*, 2021), or helping doctors treat their patients (Agarwal *et al.*, 2018). However, with great power comes great responsibility; as these models become increasingly ubiquitous and their decisions increasingly relied upon, potential deployers must have a nuanced understanding of their abilities. One must know as precisely as possible how well an LM will perform on its assigned task in order to gauge the expected error in its calculations, and thus afford it adequate human supervision. A model ought not be deployed until it has been properly and thoroughly evaluated.

To perform this evaluation, the scientific community relies on benchmarks, which are series of tests designed to simulate real-life scenarios which an LM might encounter. In order for a new model to gain traction in the scientific community, it must perform well on certain benchmarks. For well established domains, these benchmarks have been studied for decades and have undergone multitudinous permutations and updates. At any given moment there are certain benchmarks that are understood by the community to be essential; a model not evaluated on these will not be taken seriously by the community, or reviewers at conferences or journals (Dehghani *et al.*, 2021)[1]. As the state-of-the-art (SOTA) improves, these benchmarks are continually updated or retired due to "degeneration", where human parity is reached (Dehghani *et al.*, 2021; Bowman & Dahl, 2021). However, domains in which there are not yet well-established benchmarks, such as French biomedical NLP, lack such self-regulation (Dehghani *et al.*, 2021). Therefore, the publishers of models in cutting edge domains must choose, without relying on significant precedent, on which benchmarks to evaluate their models. This freedom of choice can lead authors to primarily include benchmarks on which their models perform well compared with their competitors, a process referred to by Dehghani *et al.* (2021) as "rigging the lottery". This is counterproductive for a nascent domain, as it can motivate the reverse-engineering of evaluation systems to promote individual models, rather than the engineering of better models to solve known tasks.

In this paper, we show that French biomedical NLP benchmarking exhibits weaknesses consistent with an early stage domain as taxonimized by Dehghani *et al.* (2021). We consider challenges inherent in biomedical benchmark creation, and discuss ways in which benchmarks can be created more effectively. We then perform a review of benchmarks used in French biomedical NLP, and perform an independent evaluation of them using SOTA models. We show that more work towards benchmarking is necessary in order to better prepare French biomedical LMs for deployment.

2 Benchmarking biomedical LMs

2.1 Motivation

In machine learning, a series of tests on which a model can be evaluated. The purpose of a benchmark is to measure the quality of different models on identical input, both to rank the models amongst each

[1]For example, all of the English language masked language models (MLMs) published during the NLP boom promulgated by the release of the Transformer architecture, such as BERT, XLNet, RoBERTa, XLM-RoBERTa, were evaluated on GLUE and SQuAD (Wang *et al.*, 2018; Rajpurkar *et al.*, 2016)

other and to determine tractability of a problem using SOTA technology. Furthermore, a benchmark should mirror real life applications as closely as possible: the purpose for training and publishing biomedical LMs is for their eventual deployment to assist in some manner in the treatment of medical patients. Hence, when creating a biomedical benchmark, we should consider what real use cases exist for biomedical LMs. For example, Kanwal & Rizzo (2022) describe the task of summarizing dense clinical notes, Rabhi (2022) describes predicting patient outcomes based on previous visits in a multi-modal setting, and Carchiolo *et al.* (2019) the (semi)-automated prescription of medicines. Furthermore, Yang *et al.* (2023) identify three main phases of a patient's journey in which LMs could be applied.

1. Prior to formal medical care: screening without the input of human professionals, screening for potential medical conditions.
2. During medical care: diagnosing conditions based on written reports.
3. Post medical care: counseling patients, assisting in insurance billing.

In general, most perceived medical LM use cases involve automating a task that requires a nuanced understanding of medicine in general, and any individual patient likewise. Thus, we posit that most tasks envisioned for biomedical NLP fall under the umbrella of natural language understanding (NLU), which means a model's ability to parse texts semantically rather than merely syntactically[2]. Another important category of encoder LM benchmarks is named entity recognition (NER), which involves classifying individual words and phrases. In the biomedical domain, this could be useful for the extraction of keywords from long-form medical texts, and for text summarization based thereupon. However, according to the aforementioned clinical use cases for biomedical LMs, NER is in general of lesser significance than NLU tasks. Biomedical benchmarks in practice should reflect this proclivity towards NLU; however, in the following section we will discuss the challenges of creating biomedical NLU benchmarks.

2.2 Difficulties in biomedical NLP benchmarking

In order to create a biomedical benchmark, one must first have a biomedical corpus on which to build tasks. To the detriment of NLP scientists, access to and publication of medical data in general is heavily regulated in order to safeguard individuals' privacy ((European Parliament and Council, 2016; United States Department of Health and Human Services, 2013; Li & Qin, 2017)). In order to distribute data, patients' Protected Health Information (PHI) must be hidden from Electronic Health Records (EHR); however, PHI cannot simply be erased, as it is a critical piece of information in biomedical text analysis (Mamede *et al.*, 2016). Different anonymization standards and techniques exist for the automatic de-identifying of EHRs in order to facilitate data sharing, though there exists no industry gold standard (Sweeney, 2002; Machanavajjhala *et al.*, 2007; Li & Qin, 2017). For example, the most-frequently utilized English EHR corpus, MIMIC-III, uses a combination of regular expressions and dictionary lookups (Johnson *et al.*, 2016), though this system is continually updated and not guaranteed to completely de-identify all data. The difficulties of de-identifying data are exemplified in the `DrBERT` paper, which trains and evaluates a slew of models on private datasets,

[2]This is one substantial difference from traditional corpus linguistic use cases: in practical medical NLP, semantic understanding far outweighs syntactic precision. Classical general purpose LM use cases, such as grammar or spell checking, are superfluous for encoder biomedical LMs.

but these data remained siloed - i.e. private, accessible only to those with insider permissions (Labrak *et al.*, 2023; Lin *et al.*, 2022).

One technique to circumvent this problem is known as Federated Learning (FL), in which models are passed between secure data silos for on-site learning (Zhang *et al.*, 2021) as well as evaluation (Karargyris *et al.*, 2023). This way, no data must be transferred between institutions. Indeed, FL is gaining traction in many fields, including the biomedical one, as a means to avoid data leakage (Rieke *et al.*, 2020). Unfortunately, studies have shown that some models can be attacked to reveal training data, which defeats the purpose of privacy gains in private training in FL (Winograd, 2023). Furthermore, lack of data transparency further exacerbates the opacity inherent in highly parameterized LMs. FL is also expensive, as it requires a high degree of organizational cooperation, from thorough data inspection to functional model exchange platforms. While we advocate for this method in principle, its cost, both financially and administratively, renders its implementation challenging.

Another issue afflicting biomedical benchmarks is that they are often created by NLP scientists without significant input from biomedical professionals (Cardon *et al.*, 2020; Peng *et al.*, 2019; Carrino *et al.*, 2022). One solution to this problem is to collaborate directly with domain-specific experts. This is achieved in the Chinese and Russian biomedical benchmarks `CBLUE` and `RuMedBench` by working together with doctors (Zhang *et al.*, 2022; Blinov *et al.*, 2022). However, this collaboration can be challenging for any number of reasons, from pecuniary to bureaucratic to temporal. These challenges are particularly dire in the biomedical domain where, due to patient privacy concerns, there is an unusual abundance of administrative hurdles to clear in order to access, let alone share or publish data for potential benchmark usage. Thus, some benchmarks are created using sub-optimal corpora without the input of domain-specific experts, which can lead to self-professed ambiguity in quality of the resulting created benchmarks (Cardon *et al.*, 2020), further compounding the bias inherent in any human-based annotation (Schoch *et al.*, 2020). This results in benchmarks which are either insufficiently similar to real-life tasks, or potentially inaccurate. As noted in Cardon *et al.* (2020), where several common French biomedical benchmarks[3] were introduced: "...**the annotators' lack of medical training could diminish the annotation quality**" [4].

2.3 Evaluation of existing biomedical benchmarks

We compare the types of tasks in common biomedical NLP benchmarks (see Table 1). According to the two major criticisms interrogated in this paper (insufficient focus on NLU, non-medical annotators), some benchmarks are of higher quality than others. The Russian `RuMedBench`, for example, uses "clinician" annotators for each of their tasks, and focuses specifically on NLU tasks, introducing each with an explicit allusion to a clinical use case. For example, its `RuMedSymptomRec` symptom recommendation task helps users refine their (online) medical searches based on incomplete symptom lists. The Chinese `CBLUE` benchmark also highlights the medical credentials of its annotators ("doctors from class A tertiary hospitals"), and likewise is thorough in its motivation for each task. For example, in its `KUAKE-QIC` task, a biomedical LM must classify medically related search engine queries by category, such as diagnosis, treatment plan, or test result analysis. The English `BLURB` contains several tasks which were annotated by medical professionals. Like `CBLUE`, `BLURB` emphasizes the need for eclectically sourced corpora and a variety of different subtypes of tasks,

[3] `CAS-POS`, `CAS-SG`, and a semantic similarity task similar to `CLISTER` - see Section 2.4 for further details

[4] *Fr: l'absence de formation médicale des annotateurs peut également présenter un obstacle dans la qualité du travail d'annotation*

mainly of type NLU[5].

However, we find that not all biomedical benchmarks are as thorough as RuMedBench, CBLUE, and BLURB. For example, despite the greater importance of NLU tasks in biomedical NLP, both Bio-cli and CamemBERT-bio are evaluated on only NER tasks, as illustrated in Table 1. Furthermore, both jargon and DrBERT include part of speech (POS) tagging tasks as part of their principal analyses, despite little evidence for this being a useful clinical benchmark. In DrBERT, there is one particularly clinically relevant NLU task, aHF - the diagnosis of a heart condition based on a freeform text about a patient - but it is private, making it impractical for adoption by the community.

Despite compiling many tasks, of which some are NLU, neither Segonne *et al.* (2024) nor Labrak *et al.* (2023) discussed the clinical relevance that each task was trying to simulate, instead describing each task from a more technical NLP perspective. For example, they use the NLU task FrenchMedMCQA, which involves answering multiple choice questions from a real French pharmaceutical exam; the applicability of its results to a concrete use case are not immediately evident. However, its content was created by biomedical professionals, and thus the labels are as high quality as possible. To the contrary, CLISTER is a task based on judging the semantic similarity of pairs of sentences on a scale from 0 to 5. The clinical application of this is more immediately evident - for example, pairs of appointment summaries could be compared to determine whether a patient's health is changing. However, the four annotators of CLISTER were also the paper's four authors, none of whom has a background in medicine. Although they lay out a detailed annotation pipeline to ensure inter-annotator agreement, which emphasized "semantic similarity [of] medical concepts" (Hiebel *et al.*, 2022), given their lack of medical background, it appears they may be agreeing on potential shared medical misunderstanding. For example, consider this sample pair from the CLISTER corpus (similarity score 2.5):

Le reste de la vessie est strictement normal. *En.: The rest of the bladder was strictly normal*
Le reste du parenchyme rénal était normal. *En.: The rest of the renal parenchyma was normal*

To correctly annotate this pair, one must know what a renal parenchyma is (the author of this paper did not know what that was), as well as understand whether its normalcy is equivalent to bladder normalcy.

Table 1: Comparison of NLU focus for common biomedical benchmarks

Benchmark	DrBERT	CamemBERT-bio	jargon	BLURB	Bio-cli	CBLUE	RuMedBench
Citation	(Labrak *et al.*, 2023)	(Touchent *et al.*, 2023)	(Segonne *et al.*, 2024)	(Gu *et al.*, 2020)	(Carrino *et al.*, 2022)	(Zhang *et al.*, 2022)	(Blinov *et al.*, 2022)
Language	French	French	French	English	Spanish	Chinese	Russian
Grammar tasks	2	0	3	0	0	0	0
NER tasks	5[6]	5	4	6	3	2	1
NLU tasks	4[7]	0	3	7[8]	0	6	4

2.4 Benchmarks in our study

We will use a representative sample of six popular French biomedical benchmarks, as described in Table 2, on which we will evaluate several French biomedical LMs. Of the three publicly available

[5]Regarding English biomedical benchmarking: the ClinicalBERT paper uses clinic readmission from MIMIC-III longform clinical notes as a benchmark (Huang *et al.*, 2019) (Johnson *et al.*, 2016). This is a highly targeted use case! However, this benchmark has inexplicably not been reused in subsequent English biomedical literature.

[6]Of which two are private

[7]Of which two are private and one inaccessible

[8]Of which three are relation extraction tasks

French NLU benchmarks available, we chose two (CLISTER and FrenchMedMCQA), while leaving out the semantic similarity task from Cardon *et al.* (2020), given that CLISTER is basically its updated equivalent (Hiebel *et al.*, 2022). We are not aware of any other publicly available French biomedical NLU tasks at the time of writing.

Table 2: Statistics for each dataset included in this paper

	CAS-POS	ESSAI-POS	CAS-SG	QUAERO-MEDLINE	CLISTER	FrenchMedMCQA
Task	POS	POS	NER	NER	Semantic Similarity	Question Answering
Size (sentences)	3.8k	2.4k	4.5k	7.2k	1k	3.1k
DrBERT / CamemBERT-bio / jargon	✓✓✓	✓✗✓	✓✓✓	✓✗✓	✗✗✓	✓✗✓
Is task NLU?	✗	✗	✗	✗	✓	✓
Clinician annotated?	✗	✗	✗	✗	✗	✓

3 French biomedical LMs

In 2024, analysis of texts is accomplished using Masked Language Models (MLMs) based on the Transformer architecture (Devlin *et al.*, 2019). These models use fixed-length self-attention to process blocks of text and emit an encoded embedding for each sub-word of the input. MLMs are convenient because their pre-training is completely unsupervised, meaning it requires no labeled data. Given an input document composed of many tokens (syntactically selected sub-words), an MLM produces embeddings for each token, as well as a summarizing embedding which seeks to represent the document as a single unit. They can therefore be used for two main types of analysis: token-level analysis (using each token embedding) or document-level analysis (using the summarizing embedding). These embeddings can either be used out of the box or further refined by using end-to-end fine-tuning to create task-specific representations (Devlin *et al.*, 2019).

In this paper, we will examine three classes of French biomedical LMs. Each has in the order of 100M trainable parameters.

1. French bio-medical models (left portion of Table 3) - i.e. those pre-trained from scratch (or from general-purpose checkpoint) on French bio-medical corpora. We will test DrBERT-4 (Labrak *et al.*, 2023), CamemBERT-bio (Touchent *et al.*, 2023), Jargon-biomed and Jargon-gen-biomed (Segonne *et al.*, 2024). (The last was trained on a mixture of biomedical data and general data.)

2. General purpose French language models (middle portion of Table 3) - we seek to replicate the aforementioned necessity of domain-specific models for various biomedical tasks. We will be using CamemBERT (Martin *et al.*, 2020) trained on the CCNet corpus (Wenzek *et al.*, 2020), as well as FlauBERT-1 (Le *et al.*, 2020).

3. English biomedical models (right portion of Table 3) - i.e. those trained from scratch from English biomedical corpora. Because of the syntactical similarities of English and French, one strategy for creating French language biomedical LMs is simply to co-opt English language biomedical LMs, as was tested in Labrak *et al.* (2023); Touchent *et al.* (2023); Segonne *et al.* (2024). We will be testing ClinicalBERT (Huang *et al.*, 2019) and PubMedBERT (Gu *et al.*, 2020).

Table 3: Statistics for each model included in this paper

statistic	DrBERT-4	CamemBERT-bio	Jargon-biomed	Jargon-gen-biomed	CamemBERT-CCNet	FlauBERT-1	ClinicalBERT	PubMedBERT
Language	French	French	French	French	French	French	English	English
Domain	Bio-med	Bio-med	Bio-med	Bio-med	General	General	Bio-med	Bio-med
Train steps	80k	50k	50k	100k	240k	50k	200k	63k
Data size	4GB	2.7GB	5.4GB	24GB	4GB	71GB	5GB	21GB

4 Experimental setup

The goal of this empirical section is two-fold:

1. We seek to assess the effect of model type on benchmark type. We want to evaluate in which contexts domain-specific MLMs are useful, and consistently reliable. We measure utility by mean performance on a certain task, and reliability by low variance on replications of different splits of the data for each task as well as and different classifier initializations. We will achieve this by evaluating the six benchmarks in Table 2 on different classes of model and comparing their performances.

2. We are interested in how much information is stored in each model during only pre-training, to ascertain whether the models are useful out-of-the-box. Many applications involve employing pre-trained token embeddings from MLMs without fine-tuning them to a specific downstream task, so it is important to test whether these token-embeddings are useful in a specific downstream setting. To test this, we train the models in two settings: first, in conventional, "unfrozen", end-to-end training, in which all model parameters may be updated during fine-tuning; and second, in "frozen" fine-tuning, in which the model's pre-trained weights remain fixed during fine-tuning, and only the classification layer(s)[9] are updated.

We train each model on each of the six benchmarks. We cross-validate the learning rate for each model and dataset using a random 80/10/10 train/valid/test split, each for up to 2000 steps, stopping early given validation set convergence. We repeat this training for frozen and unfrozen model weights. We replicate each experiment twenty times to gauge each model's consistency.

4.1 Summary and discussion of results

1. Each experiment, frozen or otherwise, saw a French biomedical LM perform best, as illustrated in Figures 1 and 2, and tabulated explicitly in Appendix Tables 4 and 5. For all experiments apart from non-frozen POS tagging, that best performing model was `CamemBERT-bio`. However, the other French biomedical models all fall short on some of the NER and NLU tasks: `DrBERT` is nearly as strong as `CamemBERT-bio` on `CLISTER` for non-frozen fine-tuning, but much worse on `FrenchMedMCQA` and frozen `CLISTER`. The two `jargon` models are significantly inferior on almost all tasks, despite performing best on the non-frozen POS tagging tasks. So as to the question whether French biomedical LM training is worthwhile, the answer appears to be yes for the specific case of `CamemBERT-bio`, but should be studied further to determine why the other models are unable to replicate its performance.

[9] We use a single linear classification layer for all tasks except `FrenchMedMCQA`, where we use two, on the advice of the authors (Labrak *et al.*, 2022).

Furthermore, the model CamemBERT-CCNet performs reliably worse than CamemBERT-bio, though never by too huge of a margin, while the English biomedical models are inferior to CamemBERT-CCNet at almost every task. This motivates the usage of general purpose same-language LMs over English biomedical models LMs for languages without dedicated biomedical LMs. However, we caution that given the fact that these benchmarks were created without the input of medical professionals, this trend could be misleading. It is worth noting that the only benchmark in our study which was created by medical professionals - FrenchMedMCQA - resulted in an English language model, PubMedBERT, outperforming CamemBERT-CCNet.

Lastly, we note the lack of consistency on the NLU datasets. The standard deviation of test scores (red lines in Figures 1 and 2) for FrenchMedMCQA (and to a slightly lesser extent for CLISTER) are great, illustrating the unreliability of using a French biomedical LM on related tasks. Such a wide performance range renders clinical models much less useful.

2. We show that some models are usable without end-to-end fine-tuning, while others should not be. For example, as shown in Figure 3, CamemBERT-bio has a consistently small improvement (even negative for CLISTER) when model weights are unfrozen, while both the English biomedical LMs and Jargon-biomed tend to improve significantly with unfrozen weights. For English LMs, this is not surprising, given the model is adapting to a new language. For Jargon-biomed, this suggests pre-training that is somehow inferior compared to that of CamemBERT-bio.

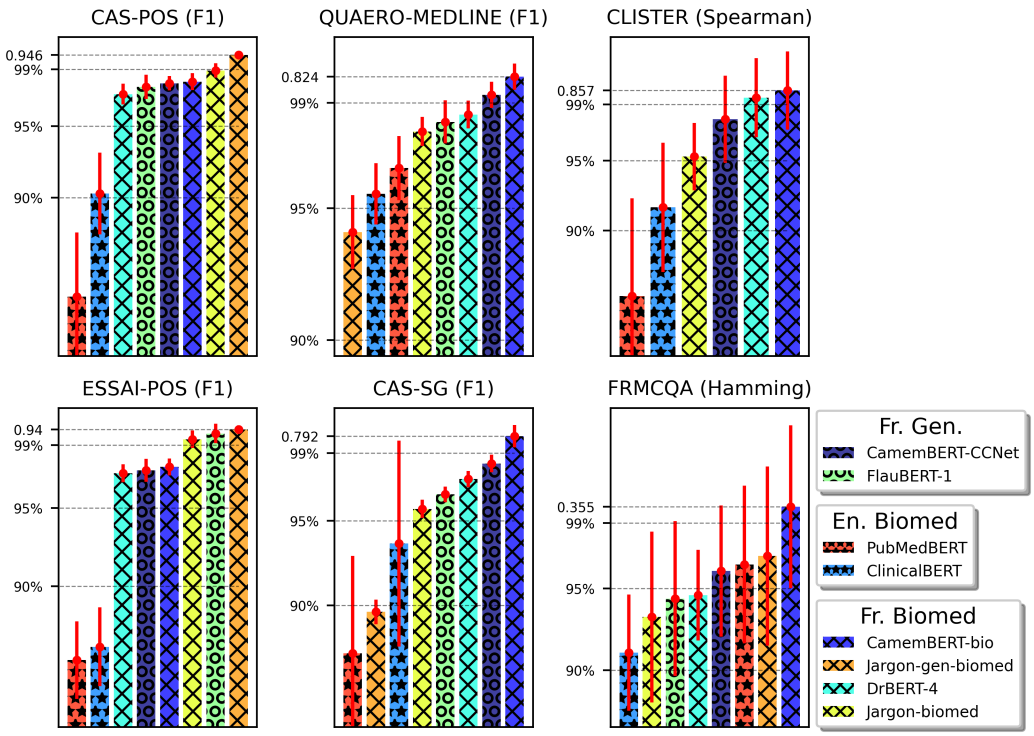


Figure 1: We compare the test-set scores of each model on each benchmark with **unfrozen** model weights. CamemBERT-bio is the best performer on all but the POS tasks. For scaling purposes, we left off models which performed significantly worse than the top model for each task.

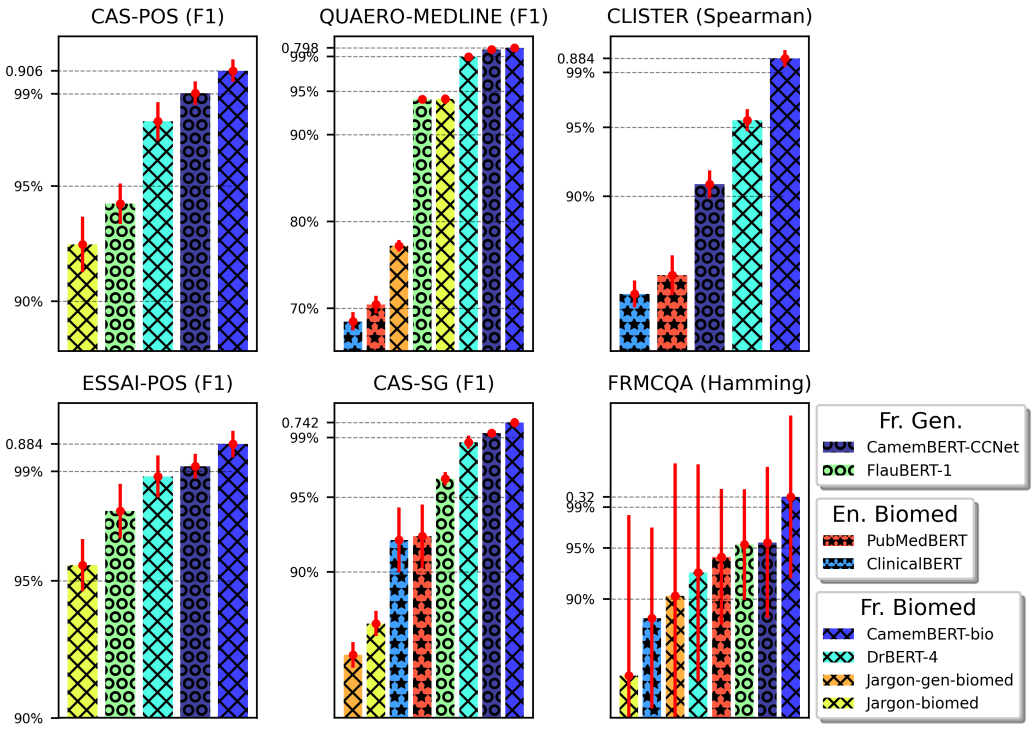


Figure 2: We compare the test-set scores of each model on each benchmark with **frozen** model weights. `CamemBERT-bio` is the best performing model on all tasks.

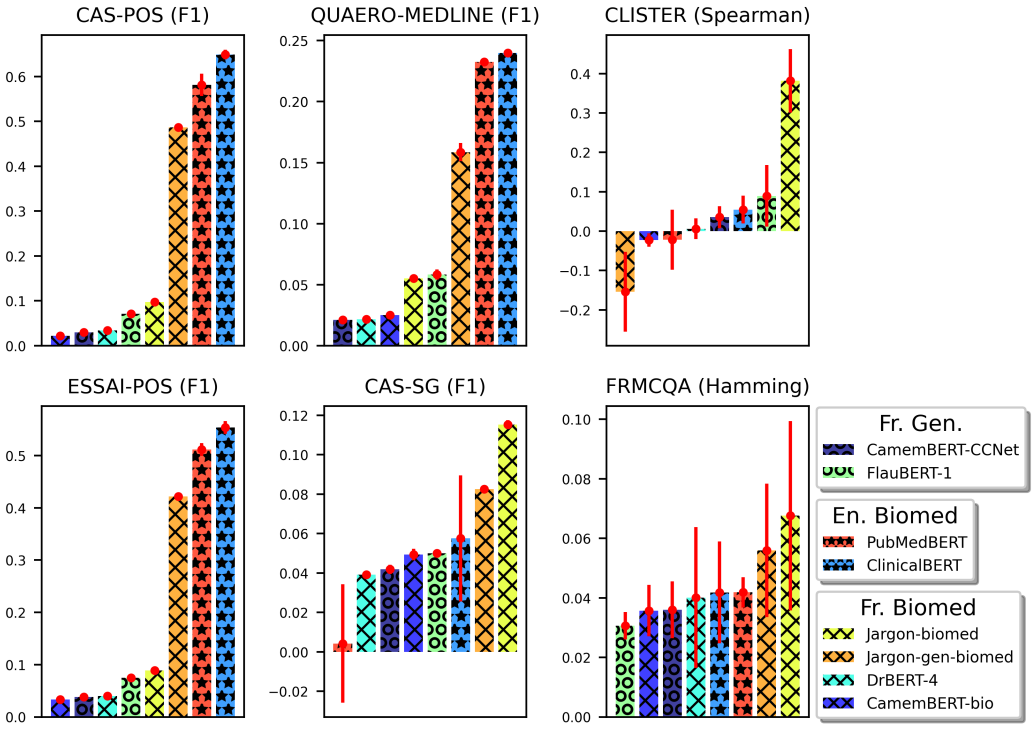


Figure 3: We calculate the difference between all pairs of tests (**frozen** and **unfrozen**) for each model and benchmark. A model that improves with unfrozen weights (as most do) has a positive score.

5 Conclusion and future work

We set out to study the state of benchmarking for French language biomedical LMs. We show that most clinical NLP tasks are best viewed through an NLU lens, and discuss the importance of benchmarks targeting specific use cases. Despite this, we show that the quantity and quality of biomedical NLU tasks is lacking in many languages, a trend particularly noticeable in French. With this in mind, we recommend that immediate future study of French biomedical NLP go towards improving benchmarking before it goes toward improving models. While benchmark creation may be less exciting than model development, it is essential to properly understand where our current models stand with respect to potential clinical application. We propose two criteria for benchmark design, inspired in large part by the excellent Russian and Chinese biomedical benchmarks described respectively in Blinov *et al.* (2022) and Zhang *et al.* (2022). Biomedical benchmarks should be:

1. constructed with a specific target use case in mind and in concert with biomedical professionals. These envisioned use cases should be briefly delineated in papers that apply them.

2. accompanied with a performance threshold above which a model could considered to be ready for some real life use. This will help users interpret the models' performances in an absolute sense, which is not currently the case for NLU benchmarks like `CLISTER` or `FrenchMedMCQA`.

Once this threshold for benchmark quality has been met, we can begin to pose more refined questions regarding a biomedical benchmark's quality, as has been done for domains with better established benchmarks (Bowman & Dahl, 2021; Dehghani *et al.*, 2021). For example, the AFLITE algorithm can be used to de-bias datasets for repetitiveness and prohibit models from picking up on spurious correlations (Sakaguchi *et al.*, 2021). However, given the nascent state of French biomedical NLP benchmarking, such sophisticated methods are not yet relevant.

Through experimentation, we observe that while all tasks benefit from domain-specific pre-training, the effect is most pronounced for NLU tasks[10]. While we identified one model which outperforms the others (`CamemBERT-bio`), even this model suffers from high variance under experimental replication. Therefore, we recommend further study of `CamemBERT-bio` and why it significantly outperforms its competitors. Are its training data higher quality, its architecture more effective, its pre-training strategy better? A brief analysis does not reveal any significant difference in construction and pretraining between any of the three French biomedical MLMs studied in this paper (Touchent *et al.*, 2023; Labrak *et al.*, 2023; Segonne *et al.*, 2024)[11].

Finally, we recommend a study into the rate at which LMs (French biomedical LMs included) are used without end-to-end finetuning. Barring a near-zero rate, we recommend regular frozen evaluation to complement end-to-end finetuning in subsequent publications.

[10]We note that our empirical conclusions were drawn based on results from two NLU benchmarks, a pittance when compared to the vast potential use cases for biomedical LMs. This conclusion should be re-evaluated once French biomedical benchmarking has advanced.

[11]The most notable exception is that `jargon` uses the Linformer architecture (Wang *et al.*, 2020), though studies have shown this architecture to perform like Transformer, and thus is unlikely to be the source of observed inferior performance.

References

AGARWAL A., BAECHLE C., BEHARA R. & ZHU X. (2018). A natural language processing framework for assessing hospital readmissions for patients with copd. *IEEE Journal of Biomedical and Health Informatics*, **22**(2), 588–596. DOI : 10.1109/JBHI.2017.2684121.

BLINOV P., RESHETNIKOVA A., NESTEROV A., ZUBKOVA G. & KOKH V. (2022). *RuMedBench: A Russian Medical Language Understanding Benchmark*, In *Lecture Notes in Computer Science*, p. 383–392. Springer International Publishing. DOI : 10.1007/978-3-031-09342-5_38.

BOWMAN S. R. & DAHL G. (2021). What will it take to fix benchmarking in natural language understanding? In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 4843–4855, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.385.

CARCHIOLO V., LONGHEU A., REITANO G. & ZAGARELLA L. (2019). Medical prescription classification: a nlp-based approach. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, p. 605–609. DOI : 10.15439/2019F197.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, p. 1–13, Nancy, France: ATALA et AFCP.

CARRINO C. P., LLOP J., PÀMIES M., GUTIÉRREZ-FANDIÑO A., ARMENGOL-ESTAPÉ J., SILVEIRA-OCAMPO J., VALENCIA A., GONZALEZ-AGIRRE A. & VILLEGAS M. (2022). Pre-trained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, p. 193–199, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.bionlp-1.19.

DEHGHANI M., TAY Y., GRITSENKO A. A., ZHAO Z., HOULSBY N., DIAZ F., METZLER D. & VINYALS O. (2021). The benchmark lottery. *ArXiv*, **abs/2107.07002**.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EUROPEAN PARLIAMENT AND COUNCIL (2016). Regulation (eu) 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

FUCHS K. (2023). Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse? In *Frontiers in Education*, volume 8, p. 1166682: Frontiers.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, **abs/2007.15779**.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2022). CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4306–4315, Marseille, France: European Language Resources Association.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, **abs/1904.05342**.

JOHNSON A. E. W., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, **3**(1), 160035.

KANWAL N. & RIZZO G. (2022). Attention-based clinical note summarization. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, p. 813–820, New York, NY, USA: Association for Computing Machinery. DOI : 10.1145/3477314.3507256.

KARARGYRIS A., UMETON R., SHELLER M. J., ARISTIZABAL A., GEORGE J., WUEST A., PATI S., KASSEM H., ZENK M., BAID U., NARAYANA MOORTHY P., CHOWDHURY A., GUO J., NALAWADE S., ROSENTHAL J., KANTER D., XENOCHRISTOU M., BEUTEL D. J., CHUNG V., BERGQUIST T., EDDY J., ABID A., TUNSTALL L., SANSEVIERO O., DIMITRIADIS D., QIAN Y., XU X., LIU Y., GOH R. S. M., BALA S., BITTORF V., PUCHALA S. R., RICCIUTI B., SAMINENI S., SENGUPTA E., CHAUDHARI A., COLEMAN C., DESINGHU B., DIAMOS G., DUTTA D., FEDDEMA D., FURSIN G., HUANG X., KASHYAP S., LANE N., MALLICK I., MASCAGNI P., MEHTA V., MORAES C. F., NATARAJAN V., NIKOLOV N., PADOY N., PEKHIMENKO G., REDDI V. J., REINA G. A., RIBALTA P., SINGH A., THIAGARAJAN J. J., ALBRECHT J., WOLF T., MILLER G., FU H., SHAH P., XU D., YADAV P., TALBY D., AWAD M. M., HOWARD J. P., ROSENTHAL M., MARCHIONNI L., LODA M., JOHNSON J. M., BAKAS S., MATTSON P., FETS CONSORTIUM, BRATS-2020 CONSORTIUM & AI4SAFECHOLE CONSORTIUM (2023). Federated benchmarking of medical artificial intelligence with MedPerf. *Nature Machine Intelligence*, **5**(7), 799–810. DOI : 10.1038/s42256-023-00652-2.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (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)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. DOI : 10.18653/v1/2022.louhi-1.5.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 16207–16221, Toronto, Canada: Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised language model pre-training for French).

In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 268–278, Nancy, France: ATALA et AFCP.

LI X.-B. & QIN J. (2017). Anonymizing and sharing medical text records. *Information Systems Research*, **28**, 332–352. DOI : 10.1287/isre.2016.0676.

LIN B. Y., HE C., ZE Z., WANG H., HUA Y., DUPUY C., GUPTA R., SOLTANOLKOTABI M., REN X. & AVESTIMEHR S. (2022). FedNLP: Benchmarking federated learning methods for natural language processing tasks. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Findings of the Association for Computational Linguistics: NAACL 2022*, p. 157–175, Seattle, United States: Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.13.

MACHANAVAJJHALA A., KIFER D., GEHRKE J. & VENKITASUBRAMANIAM M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **1**(1), 3–es.

MAMEDE N., BAPTISTA J. & DIAS F. (2016). Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, p. 1287–1294. DOI : 10.1109/CEC.2016.7743936.

MARTIN L., MULLER B., SUÁ REZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In D. DEMNER-FUSHMAN, K. B. COHEN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 58–65, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/W19-5006.

PLAZA-DEL ARCO F. M., MOLINA-GONZÁLEZ M. D., URENA-LÓPEZ L. A. & MARTÍN-VALDIVIA M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, **166**, 114120. DOI : 10.1016/j.eswa.2020.114120.

RABHI S. (2022). *Optimized deep learning-based multimodal method for irregular medical timestamped data*. Theses, Institut Polytechnique de Paris. HAL : tel-03600526.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In J. SU, K. DUH & X. CARRERAS, Éds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392, Austin, Texas: Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

RIEKE N., HANCOX J., LI W., MILLETARÌ F., ROTH H. R., ALBARQOUNI S., BAKAS S., GALTIER M. N., LANDMAN B. A., MAIER-HEIN K., OURSELIN S., SHELLER M., SUMMERS R. M., TRASK A., XU D., BAUST M. & CARDOSO M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, **3**(1), 119. DOI : 10.1038/s41746-020-00323-1.

SAKAGUCHI K., BRAS R. L., BHAGAVATULA C. & CHOI Y. (2021). Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, **64**(9), 99–106. DOI : 10.1145/3474381.

SCHOCH S., YANG D. & JI Y. (2020). "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In S. AGARWAL, O. DUŠEK, S. GEHRMANN, D. GKATZIA, I. KONSTAS, E. VAN MILTENBURG & S. SANTHANAM, Éds., *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, p. 10–16, Online (Dublin, Ireland): Association for Computational Linguistics.

SEGONNE V., MANNION A., CANUL L. C. A., AUDIBERT A., LIU X., MACAIRE C., PUPIER A., ZHOU Y., AGUIAR M., HERRON F., NORRÉ M., AMINI M. R., BOUILLON P., ESHKOL-TARAVELLA I., ESPERANÇA-RODIER E., FRANÇOIS T., GOEURIOT L., GOULIAN J., LAFOURCADE M., LECOUTEUX B., PORTET F., RINGEVAL F., VANDEGHINSTE V., COAVOUX M., DINARELLI M. & SCHWAB D. (2024). Jargon: A suite of language models and evaluation tasks for french specialized domains. In *Proceedings of the LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation [Forthcoming]*.

SWEENEY L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, **10**(05), 557–570.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In C. SERVAN & A. VILNAT, Éds., *18e Conférence en Recherche d'Information et Applications*
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 323–334, Paris, France: ATALA. HAL : hal-04130187.

UNITED STATES DEPARTMENT OF HEALTH AND HUMAN SERVICES (2013). 45 cfr parts 160 and 164.
https://www.govinfo.gov/content/pkg/FR-2013-01-25/pdf/2013-01073.pdf.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, NY, USA: Curran Associates Inc.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. LINZEN, G. CHRUPAŁA & A. ALISHAHI, Éds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium: Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WANG S., LI B. Z., KHABSA M., FANG H. & MA H. (2020). Linformer: Self-attention with linear complexity.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMÁN F., JOULIN A. & GRAVE E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4003–4012, Marseille, France: European Language Resources Association.

WINOGRAD A. (2023). Loose-lipped large language models spill your secrets: The privacy implications of large language models. *Harvard Journal of Law & Technology*, **36**(2).

YANG R., TAN T. F., LU W., THIRUNAVUKARASU A. J., TING D. S. W. & LIU N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, **2**(4), 255–263.

ZHANG C., XIE Y., BAI H., YU B., LI W. & GAO Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, **216**, 106775. DOI : https://doi.org/10.1016/j.knosys.2021.106775.

ZHANG N., CHEN M., BI Z., LIANG X., LI L., SHANG X., YIN K., TAN C., XU J., HUANG F., SI L., NI Y., XIE G., SUI Z., CHANG B., ZONG H., YUAN Z., LI L., YAN J., ZAN H., ZHANG K., TANG B. & CHEN Q. (2022). CBLUE: A Chinese biomedical language understanding evaluation benchmark. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 7888–7915, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.544.

6 Appendix

Table 4: Test set results for frozen models

Dataset	DrBERT-4	CamemBERT-bio	Jargon-biomed	Jargon-gen-biomed	CamemBERT-CCNet	FlauBERT-1	ClinicalBERT	PubMedBERT
CAS-POS	0.886 ± 0.008	0.906 ± 0.004	0.838 ± 0.011	0.459 ± 0.01	0.897 ± 0.005	0.854 ± 0.008	0.205 ± 0.019	0.204 ± 0.018
QUAERO-MEDLINE	0.79 ± 0.003	0.798 ± 0.003	0.751 ± 0.004	0.616 ± 0.005	0.797 ± 0.002	0.751 ± 0.003	0.547 ± 0.008	0.562 ± 0.008
CLISTER	0.845 ± 0.007	0.884 ± 0.005	0.435 ± 0.082	0.51 ± 0.078	0.804 ± 0.009	0.473 ± 0.018	0.733 ± 0.009	0.745 ± 0.013
ESSAI-POS	0.874 ± 0.007	0.884 ± 0.004	0.845 ± 0.008	0.517 ± 0.009	0.877 ± 0.004	0.862 ± 0.009	0.256 ± 0.03	0.29 ± 0.034
CAS-SG	0.733 ± 0.003	0.742 ± 0.002	0.643 ± 0.006	0.627 ± 0.006	0.737 ± 0.002	0.715 ± 0.003	0.684 ± 0.016	0.686 ± 0.016
FrenchMedMCQA	0.296 ± 0.034	0.32 ± 0.025	0.264 ± 0.05	0.289 ± 0.041	0.305 ± 0.024	0.305 ± 0.017	0.282 ± 0.028	0.301 ± 0.021

French bio-medical models are purple, French general-purpose models cyan, and English bio-medical models grey. POS and NER tasks are evaluated using F1 score; CLISTER is evaluated using the Spearman ranked correlation coefficient; FrenchMedMCQA is evaluated using either the Hamming distance between the (potentially) multiple correct answers and the answers chosen by the model.

Table 5: Test set results for non-frozen models

Dataset	DrBERT-4	CamemBERT-bio	Jargon-biomed	Jargon-gen-biomed	CamemBERT-CCNet	FlauBERT-1	ClinicalBERT	PubMedBERT
CAS-POS	0.92 ± 0.007	0.928 ± 0.006	0.936 ± 0.005	0.946 ± 0.002	0.927 ± 0.005	0.925 ± 0.008	0.854 ± 0.027	0.786 ± 0.042
QUAERO-MEDLINE	0.812 ± 0.004	0.824 ± 0.004	0.806 ± 0.005	0.775 ± 0.012	0.818 ± 0.004	0.809 ± 0.007	0.787 ± 0.01	0.795 ± 0.01
CLISTER	0.853 ± 0.024	0.857 ± 0.024	0.817 ± 0.021	0.367 ± 0.07	0.839 ± 0.027	0.563 ± 0.098	0.786 ± 0.04	0.731 ± 0.06
ESSAI-POS	0.913 ± 0.005	0.917 ± 0.005	0.934 ± 0.005	0.94 ± 0.003	0.915 ± 0.007	0.937 ± 0.006	0.809 ± 0.024	0.801 ± 0.023
CAS-SG	0.772 ± 0.004	0.792 ± 0.005	0.758 ± 0.004	0.71 ± 0.006	0.779 ± 0.004	0.765 ± 0.004	0.742 ± 0.048	0.69 ± 0.046
FrenchMedMCQA	0.336 ± 0.01	0.355 ± 0.018	0.331 ± 0.019	0.345 ± 0.019	0.341 ± 0.014	0.335 ± 0.017	0.324 ± 0.013	0.343 ± 0.017

Almost all models experienced performance improvement on all tasks when their weights were unfrozen.

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

Analyse sémantique du corpus des Cahiers citoyens

Sami GUEMBOUR[1]
(1) LASTIG, Univ Gustave Eiffel, ENSG, IGN, France
sami.guembour@ign.fr

RÉSUMÉ

Cet article présente une recherche originale qui se concentre sur une analyse sémantique du corpus des Cahiers citoyens, qui regroupe les contributions et les doléances des citoyens français déposées au niveau des mairies dans le cadre du Grand Débat National. L'article offre un état de l'art complet sur les divers travaux réalisés sur ce corpus et vise à obtenir une compréhension approfondie des thèmes émergents et des préoccupations citoyennes dans les différentes régions. Plusieurs hypothèses concernant ces travaux ont été émises, et différentes méthodes ont été proposées pour répondre à ces hypothèses, de la segmentation et du pré-traitement du corpus au calcul des vecteurs de plongement des phrases à l'aide de modèles de langues pré-entraînés, aboutissant au clustering de ces vecteurs pour construire des regroupements en fonction des problématiques abordées.

ABSTRACT

Semantic analysis of the "Cahiers citoyens" corpus

This article presents an original research focusing on a semantic analysis of the "Cahiers citoyens" (Citizen Notebooks) corpus, which compiles the contributions and grievances of French citizens submitted at the municipal level as part of the "Grand Débat National" (Grand National Debate). The article provides a comprehensive state-of-the-art review of various studies conducted on this corpus, aiming to achieve a profound understanding of emerging themes and citizen concerns in different regions. Multiple hypotheses regarding these studies have been formulated, and various methods have been proposed to address these hypotheses, ranging from corpus segmentation and pre-processing to calculating sentence embedding vectors using pre-trained language models. This culminates in clustering these vectors to construct groupings based on the addressed issues.

MOTS-CLÉS : Cahiers Citoyens - Grand débat National - Corpus - TAL - Modèle de langue - Vecteur de plongement - Classification.

KEYWORDS: Citizen Notebooks - Grand National Debate - Corpora - NLP - Language Model - Vector embedding - Clustering.

1 Introduction

La crise sociale déclenchée par le mouvement des Gilets Jaunes en France à l'automne 2018 a engendré une série de réponses institutionnelles, et plusieurs formes de participation citoyenne ont vu le jour. En décembre 2018, l'Association des Maires Ruraux de France (AMRF) [1] a lancé l'opération "Mairies Ouvertes". L'idée était de mettre des "Cahiers de doléances et de propositions" à disposition dans les mairies, offrant aux habitants une opportunité de s'exprimer librement. Ce qui devait être

1. https ://www.amrf.fr/

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 17–27.

une action courte du 8 au 15 décembre a été prolongé en raison de son succès inattendu.

En janvier 2019, le gouvernement français a lancé le Grand Débat National (GDN) [2], offrant à la fois une plateforme numérique dématérialisée et des supports matériels, les Cahiers citoyens, disponibles dans des lieux publics. Certains Cahiers de doléances et de propositions ont maintenu leur dénomination initiale, tandis que d'autres sont transformés en Cahiers citoyens. À la clôture de la période de contribution mi-mars 2019, les Cahiers de doléances et de propositions sont enrichis par ceux des Cahiers citoyens, créant ainsi une dualité entre les expressions en ligne des citoyens via la plate-forme officielle et les contributions des Cahiers citoyens.

Cet article s'inscrit dans le contexte d'une recherche qui entre dans le cadre d'une thèse. L'objectif de cette recherche est d'analyser les Cahiers citoyens de manière sémantique et spatiale en utilisant les méthodes et outils du Traitement Automatique des Langues. Dans ce contexte, l'analyse sémantique des Cahiers citoyens consiste à examiner le contenu textuel du corpus afin de comprendre les significations et les relations sémantiques entre les termes, les phrases, et les thèmes abordés. Quant à l'analyse spatiale du corpus, elle consiste à examiner comment les caractéristiques géographiques des citoyens, telles que leur lieu de résidence ou leur origine, sont liées aux thématiques abordées dans leurs contributions. Cette initiative tire ses fondements des résultats d'un travail antérieur (Chandora, 2023), au cours duquel des clusters regroupant des phrases abordant des thématiques similaires dans les contributions ont été construits. Cependant, les résultats n'ont pas atteint la satisfaction escomptée car 91 % des phrases n'ont pas été classées, incitant ainsi à entreprendre une nouvelle démarche de recherche plus approfondie visant à améliorer le regroupement de ces phrases.

Le plan de ce papier s'articule autour de plusieurs sections. La section 2 se consacre à la définition et à la présentation du corpus des Cahiers citoyens. La section 3 présente les travaux déjà entrepris en matière d'analyse des Cahiers citoyens. Dans la section 4, nous abordons la construction des hypothèses qui sous-tendent notre approche, ainsi que la définition des objectifs. La section 5 détaille la méthodologie que nous adopterons pour analyser le corpus et vérifier les hypothèses formulées. Enfin, la section 6 conclut l'article et évoque les principales attentes.

2 Définition du corpus de travail

Le corpus utilisé dans cette étude, désigné sous le nom de Cahiers citoyens (CC), rassemble des contributions provenant des habitants de diverses communes. Il s'agit des contenus rédigés par les citoyens et déposés au niveau des mairies pour exprimer leurs préoccupations. Elles ont été collectées à partir de divers supports d'expression fournis par les mairies participantes, incluant des carnets d'écoliers, des courriers électroniques, et des supports papiers avec des thèmes prédéfinis. Les Cahiers citoyens contiennent des contributions variées, allant de textes manuscrits ou dactylographiés à des courriers électroniques directement adressés aux mairies, des dossiers comportant parfois des pièces jointes, ainsi que des pétitions collectives dactylographiées. Plus de 16 000 communes ont participé à cette initiative. Ces contributions diverses sont associées à un code INSEE facilitant la localisation des communes. La consultation de ces Cahiers est soumise à une dérogation accordée par les Archives Nationales [3], accompagnée d'une clause de protection des données, en raison du caractère privé des informations qu'ils renferment.

2. https ://www.gouvernement.fr/le-grand-debat-national
3. https ://www.archives-nationales.culture.gouv.fr/

Le processus de construction du corpus a débuté par la collecte des Cahiers des différentes mairies, qui ont été transmis aux préfectures pour numérisation, générant ainsi un corpus de fichiers image. Ces fichiers images ont ensuite été envoyés à la Bibliothèque nationale de France (BnF), qui, en utilisant des outils d'OCR, a converti les fichiers en format texte, créant ainsi un corpus textuel où chaque fichier représente un ou plusieurs cahiers localisés. La BnF a également fait appel à trois prestataires pour vérifier la transcription automatique, la vérification portant notamment sur le découpage des cahiers en contributions, les métadonnées de chaque contribution, ainsi que sur leur contenu textuel. La concaténation de ces fichiers textuels a abouti à la création d'un fichier au format CSV, désormais appelé Corpus CC.

Le tableau 1 fournit des statistiques descriptives détaillées sur le corpus CC. La tokenisation des contributions s'est effectuée à l'aide de l'outil NLTK (Bird *et al.*, 2009).

TABLE 1 – Statistiques descriptives du corpus CC

Nombre total de contributions	225 224
Nombre total de tokens dans le corpus	55 838 490
Nombre de codes postaux uniques	5 551
Nombre moyen de tokens par code postal	10 059
Nombre de codes INSEE uniques	16 421
Nombre moyen de tokens par code INSEE	3 400
Nombre de dates de réception uniques	85

3 Travaux antérieurs

La recherche sur le Grand Débat National et les Cahiers citoyens a été marquée par des défis d'accès et des approches variées. Contrairement au GDN, dont le corpus est accessible en open data, l'accès aux Cahiers citoyens est complexe en raison du Règlement Général sur la Protection des Données (RGPD). De ce fait, plusieurs travaux se sont concentrés sur le corpus GDN et sur des débats alternatifs : Entendre la France [4] (une application Messenger inspirée des questions du GDN) et le Vrai Débat [5] (une plate-forme contestataire créée par des Gilets Jaunes), diversifiant les méthodes employées.

Dans (Ploux *et al.*, 2021), les auteurs ont opté pour une analyse sémantique des corpus GDN, Entendre la France et Vrai Débat en utilisant des réseaux lexicaux. Leur méthodologie, fondée sur l'identification de "cliques" via des calculs de co-occurrences, a révélé des variations thématiques en fonction de la taille des communes, apportant une perspective intéressante sur les dynamiques territoriales.

Le point de départ de l'analyse du corpus CC est la synthèse réalisée par (Berger *et al.*, 2019), commandée par le gouvernement, mais critiquée pour ses objectifs politiques et son opacité méthodologique. L'agence Cognito Consulting [6] a joué un rôle central dans cette analyse. La méthode, fondée sur une cartographie sémantique, a identifié des "clusters lexicaux", mais le manque de transparence quant à l'algorithme et la rapidité d'exécution soulèvent des interrogations sur la qualité des résultats.

4. https ://www.entendrelafrance.fr/
5. https ://levraidebat.org
6. https ://www.cognito.fr/

(Ray, 2023) a utilisé des modèles d'extraction de sujets tels que BERTopic (Grootendorst, 2022) et LDA (Latent Dirichlet Allocation) (Blei *et al.*, 2003) pour analyser les contributions du corpus des Cahiers citoyens, cherchant à comparer les résultats avec l'analyse précédente de Cognito et les catégories du corpus GDN. Cette approche offre une nouvelle perspective sur la diversité des thèmes abordés dans les contributions citoyennes.

(Monnier, 2023) a réalisé une étude approfondie sur la thématique éolienne à partir des Cahiers citoyens. Son travail s'inscrit dans une analyse transversale en sciences sociales, combinant des approches linguistiques et géographiques. L'analyse a ciblé trois départements où la thématique éolienne était prépondérante, permettant une analyse des contributions en fonction des caractéristiques naturelles et sociales spécifiques à chaque territoire. La production de cartes a été utilisée pour visualiser de manière spatialisée les extractions textuelles.

Dans son analyse exploratoire du corpus des Cahiers citoyens, présentée lors de la Journée d'étude "Cahiers citoyens 2019 : approches croisées" [7], Ploux a utilisé des méthodes fréquentielles qui s'appuient sur les variations de fréquence des mots par rapport à un corpus de référence. Ses observations ont révélé que les contributeurs abordaient des sujets différents en fonction de la taille de leur commune. Une analyse fréquentielle a également conduit à l'observation que la taille des communes était inversement corrélée au nombre des contributions de la commune. De plus, Ploux a formulé l'hypothèse que les sujets abordés dans les Cahiers citoyens seraient plus variés, voire différents, de ceux du corpus du GDN. Cette divergence s'expliquerait par le format libre des Cahiers citoyens, contrairement au GDN qui était structuré autour de quatre thèmes prédéfinis avec des questions associées, encadrant ainsi les productions des auteurs.

Le travail de (Bendinelli, 2023) se distingue par son approche fine et approfondie axée sur l'analyse d'un seul cahier provenant de la commune de Dole dans le Jura. Cette étude s'appuie sur une combinaison de disciplines, notamment la linguistique, les approches communicationnelles et sémiotiques de l'écrit, ainsi que l'analyse du discours outillée. Bien que ces méthodes offrent une compréhension approfondie du contenu et des nuances des contributions, il est important de noter qu'elles présentent un caractère peu automatisé. De ce fait, leur applicabilité à l'ensemble du corpus est limitée.

Enfin, (Chandora, 2023) a proposé une nouvelle approche dans l'exploration du corpus CC. Elle s'attache à une double perspective, alliant une analyse sémantique approfondie à une évaluation de la répartition géographique des préoccupations citoyennes. Son étude s'articule autour d'une analyse du vocabulaire complet et de la distribution géographique des contributions, révélant des thèmes et des caractéristiques propres au corpus CC. La fouille sémantique s'appuie sur deux méthodes distinctes. La première, le clustering à partir de plongements de phrases, identifie des propositions de contributeurs, tout en soulignant la pertinence de l'unité de la phrase pour l'exploration d'un corpus. Cependant, la qualité de la segmentation en phrases est impactée par la nature du corpus. La deuxième méthode, utilisant des automates à états finis, permet d'extraire un nombre plus important de séquences textuelles pour les propositions identifiées. La représentation spatiale des propositions citoyennes est explorée, combinant des techniques de TAL et des représentations cartographiques. Les résultats indiquent que les différences thématiques observées entre les propositions sont globalement peu marquées, ne semblant pas être spécifiques à des zones géographiques particulières en France. Cependant, ses résultats montrent que les citoyens les plus actifs, ayant le plus contribué aux Cahiers citoyens, résident plutôt dans de petites communes lorsqu'on rapporte les résultats au nombre d'habitants.

7. https ://geographie-cites.cnrs.fr/cahiers-citoyens-2019-approches-croisees/

4 Hypothèses et objectifs

L'hypothèse principale des travaux concernant les Cahiers citoyens est que les problématiques abordées dans les contributions dépendent de la localisation des contributeurs, et que les caractéristiques géographiques et socio-démographiques (telles que l'âge, le genre, le niveau d'éducation, l'appartenance sociale, etc.) de ces derniers peuvent influencer les thématiques qu'ils abordent. Comme explicité dans la section 1, cette recherche vise à améliorer les résultats obtenus par une étude antérieure (Chandora, 2023). Au cours de celle-ci, des groupes ont été formés en fonction des thèmes abordés dans les contributions à l'aide d'un clustering qui utilise les phrases comme unités d'étude. Les contributions ont été segmentées en phrases à l'aide de l'outil Unitex [8], ensuite les vecteurs de plongement de ces phrases ont été calculés avec le modèle de langue *sentence-camembert-base* [9] (Martin *et al.*, 2020; Nils Reimers, 2019). Pour former les groupes de phrases à partir de ces vecteurs, l'algorithme de *Fast Clustering* [10] a été choisi. Cependant, les résultats de ce clustering se sont avérés décevants, ne classant que 9% des phrases, ce qui ne permet pas de prétendre une analyse sémantique complète.

Dans le but d'atteindre les objectifs de cette recherche, des améliorations des résultats du clustering des phrases en augmentant le nombre de phrases classées ont été envisagées. Pour ce faire, des hypothèses remettant en question le choix des trois méthodes suivantes dans (Chandora, 2023) ont été émises :

- **Segmentation des phrases :** La première hypothèse s'oriente vers l'outil de segmentation choisi pour découper les contributions en phrases. Elle avance que Unitex n'est peut-être pas l'outil le mieux adapté au corpus des Cahiers citoyens, étant donné que le niveau de langue dans les contributions est varié et que leur typographie ainsi que leur syntaxe ne sont pas fiables, puisqu'elles proviennent de diverses catégories sociales . Cette non-conformité entraîne une segmentation imprécise et de qualité médiocre.
- **Modèle de langue :** Cette hypothèse se focalise sur le modèle *sentence-camembert-base* utilisé pour calculer les vecteurs de phrases, et estime qu'il ne garantit pas la meilleure représentation des phrases. Cela signifie que des phrases sémantiquement similaires ne sont pas suffisamment proches dans l'espace vectoriel, et que les distances qui les séparent ne sont pas minimales. Ceci pourrait conduire à des dispersions éloignées des phrases abordant des thématiques similaires dans cet espace vectoriel.
- **Algorithme de clustering :** La dernière hypothèse porte sur la sélection de l'algorithme de clustering et la configuration des hyperparamètres. Elle suggère que l'algorithme *Fast Clustering* ne facilite pas une agrégation efficace et complète des phrases abordant les mêmes thématiques. Il est également possible que les valeurs des hyperparamètres fixées ne soient pas appropriées, compromettant ainsi la capacité de l'algorithme à proposer un regroupement optimal, et entraînant par conséquent la dispersion de phrases abordant des thématiques similaires dans des clusters distincts.

8. https ://unitexgramlab.org/
9. https ://huggingface.co/dangvantuan/sentence-camembert-base
10. https ://www.sbert.net/examples/applications/clustering/README.html

5 Méthodologie

Les méthodes suivantes ont été proposées pour répondre aux hypothèses émises dans la section 4. L'objectif est d'obtenir des groupes de phrases regroupant, dans les mêmes clusters, celles traitant des mêmes sujets et thématiques, à travers l'application d'un nouveau processus de clustering, tout en proposant une nouvelle segmentation des contributions en phrases, un pré-traitement et un nettoyage des phrases résultantes, ainsi qu'un nouveau modèle pour calculer leurs vecteurs.

5.1 Nouvelle segmentation en phrases

La première hypothèse formulée remettait en question l'outil de segmentation en phrases utilisé et la qualité de son découpage. Afin d'améliorer la qualité du découpage en phrases, différentes approches de segmentation ont été explorées, impliquant le test de plusieurs méthodes alternatives. Parmi celles-ci, on trouve : NLTK (Bird *et al.*, 2009), SparkNLP (Kocaman & Talby, 2021) , et Spacy (Honnibal *et al.*, 2020) avec ses quatre modèles (*fr_core_news_sm*, *fr_core_news_md*, *fr_core_news_lg*, et *fr_dep_news_trf*).

Une évaluation manuelle des résultats de segmentation de ces méthodes est réalisée sur des extraits du corpus des Cahiers citoyens. À cette fin, ces extraits ont été segmentés manuellement, et les résultats des différentes méthodes ont été évalués en comparaison avec cette segmentation manuelle. Cette évaluation [11] a démontré que le modèle *fr_dep_news_trf* de Spacy, qui est fondé sur les transformeurs, permet d'obtenir la meilleure segmentation en phrases sur le corpus des Cahiers citoyens.

5.2 Pré-traitement et nettoyage des phrases

La segmentation des contributions en phrases a révélé que certaines d'entre elles sont formatées [12] (comme : "*Je vous prie d'agréer, Monsieur, mes sincères salutations*", "*Mardi 18 décembre 2018*"). Ainsi, un pré-traitement s'avère nécessaire pour identifier ces phrases formatées, afin de ne pas les inclure dans le processus de clustering. Cela permettra d'éviter une perte de temps et de ressources mémoire lors du calcul des vecteurs de phrases. De plus, ce pré-traitement permettra de réduire la consommation de mémoire et d'accélérer l'obtention des clusters via l'algorithme de clustering sélectionné.

La méthode préconisée pour le pré-traitement repose sur deux éléments. En premier lieu, il s'agit de mener une recherche syntaxique ciblant certaines structures linguistiques. En outre, une alternative consiste à mettre en œuvre un processus de clustering regroupant ces phrases formatées en fonction de leur similarité sémantique (par exemple, regrouper dans un même cluster toutes les dates et dans un autre cluster les formules de politesse).

11. La description de cette évaluation ne fait pas partie de cet article.
12. Dans cet article, nous appelons une phrase formatée une phrase qui ne porte aucune information sur les sujets et les problématiques abordés par les citoyens dans leurs contributions. Ces phrases comprennent des éléments tels que la date des contributions, les noms des contributeurs, le destinataire, les formules de politesse, etc.

5.3 Calcul des vecteurs de phrases

Afin de mettre à l'épreuve la deuxième hypothèse, qui suggère que le modèle *sentence-camembert-base*, utilisé pour calculer les vecteurs des phrases, ne garantit pas la meilleure représentation des phrases, une comparaison a été effectuée avec un autre modèle à architecture plus large, le *sentence-camembert-large* [13]. Le but de cette comparaison est de vérifier si le modèle avec l'architecture large permet d'obtenir une meilleure représentation des phrases que celui de base.

La comparaison entre les deux modèles repose sur l'évaluation des similarités cosinus [14] retournées par chacun d'eux sur trois catégories de phrases extraites du corpus, dont les sujets sont très abordés dans les Cahiers citoyens, et utilisées dans (Chandora, 2023) pour comparer le modèle de base à d'autres modèles multilingues. Ces trois catégories de phrases sont :

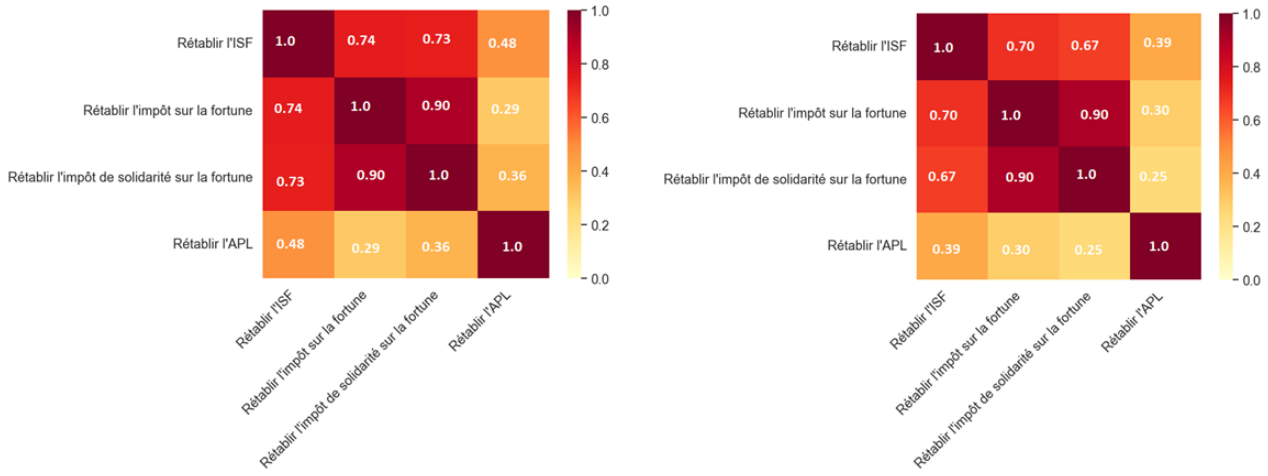
— Des phrases contenant à la fois des formes développées et des sigles d'appellation ("Rétablir l'ISF", "Rétablir l'impôt sur la fortune", "Rétablir l'impôt de solidarité sur la fortune" et "Rétablir l'APL").
— Des paraphrases ("Il faut diminuer le train de vie de l'État", "Il faut réduire le train de vie de l'État", "Il faut revoir à la baisse le train de vie de l'État", "Il faut diminuer les dépenses de l'État", et "Il faut revoir à la baisse les dépenses de l'État").
— Des phrases contradictoires sur la thématique et sur la polarité ("Vive les gilets jaunes", "Honte aux gilets jaunes", "Je soutiens les gilets jaunes", "Je suis contre les gilets jaunes" et "Je suis contre le glyphosate").

Pour les deux premières catégories de phrases, le modèle fournissant la plus grande similarité cosinus est considéré comme le meilleur, tandis que pour la dernière catégorie, étant donné qu'il s'agit de phrases contradictoires, le modèle renvoyant la plus faible similarité cosinus est considéré comme le meilleur.

La figure 1 illustre les scores de similarité cosinus retournés par les deux modèles entre les phrases de la première catégorie sous forme de matrices de confusion. Ces scores indiquent que le modèle *sentence-camembert-large* est plus efficace pour détecter la similarité entre une phrase contenant une forme développée et une autre contenant des sigles d'appellation, affichant des scores de similarité cosinus plus élevés que le modèle *sentence-camembert-base*. Les matrices de confusion de la figure 2 montrent également les scores de similarité cosinus retournés par les deux modèles mais entre les phrases de la deuxième catégorie. Ces comparaisons ont démontré que le modèle large surpasse le modèle de base dans l'identification de similarités entre des paraphrases, en renvoyant des scores plus élevés. Enfin, la figure 3 affiche les scores de similarité cosinus retournés par les deux modèles entre les phrases de la troisième catégorie. Ces scores sont plus bas pour le modèle large par rapport au modèle de base, suggérant que, même pour ce type de phrases, le modèle large offre de meilleures performances que le modèle de base.

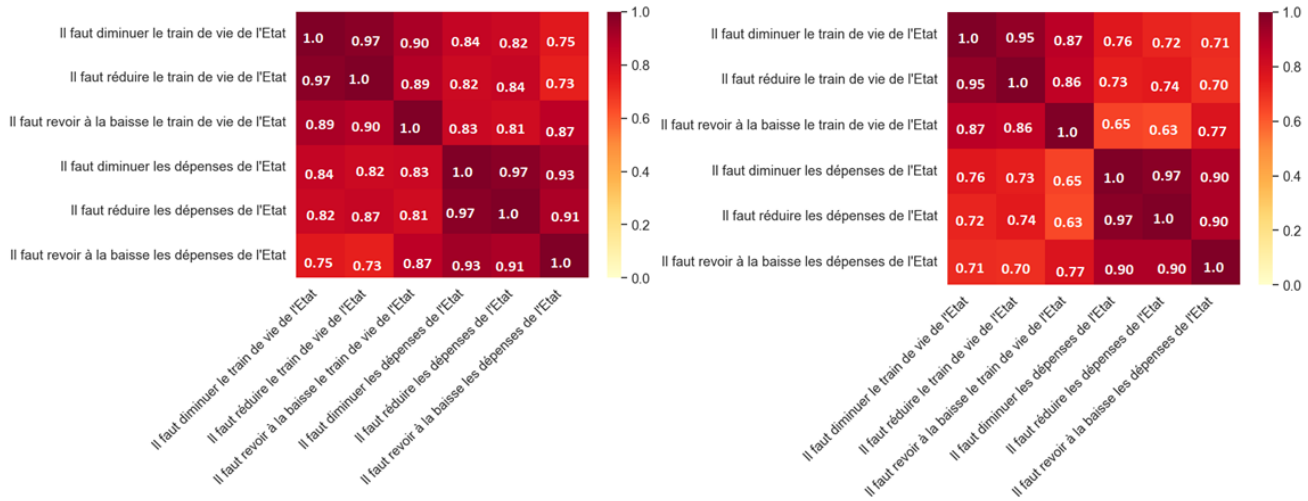
Ainsi, ces comparaisons confirment que le modèle *sentence-camembert-large* est plus adapté pour calculer les vecteurs des phrases du corpus des Cahiers citoyens, assurant une meilleure représentation, et garantissant que les phrases similaires soient plus étroitement projetées dans l'espace vectoriel, tandis que les phrases non similaires se trouvent à une plus grande distance dans cet espace.

13. https ://huggingface.co/dangvantuan/sentence-camembert-large
14. La similarité cosinus est une mesure statistique utilisée pour évaluer la similitude entre deux vecteurs dans un espace multidimensionnel. Dans le contexte de la similarité sémantique entre les phrases, elle mesure l'angle cosinus entre les vecteurs représentant ces phrases. Une similarité plus élevée indique une proximité sémantique accrue entre les phrases.



sentence-camembert-large *sentence-camembert-base*

FIGURE 1 – Matrices de confusion indiquant les similarités générées par chaque modèle entre des phrases contenant à la fois des formes développées et des sigles d'appellation



sentence-camembert-large *sentence-camembert-base*

FIGURE 2 – Matrices de confusion indiquant les similarités générées par chaque modèle entre des paraphrases

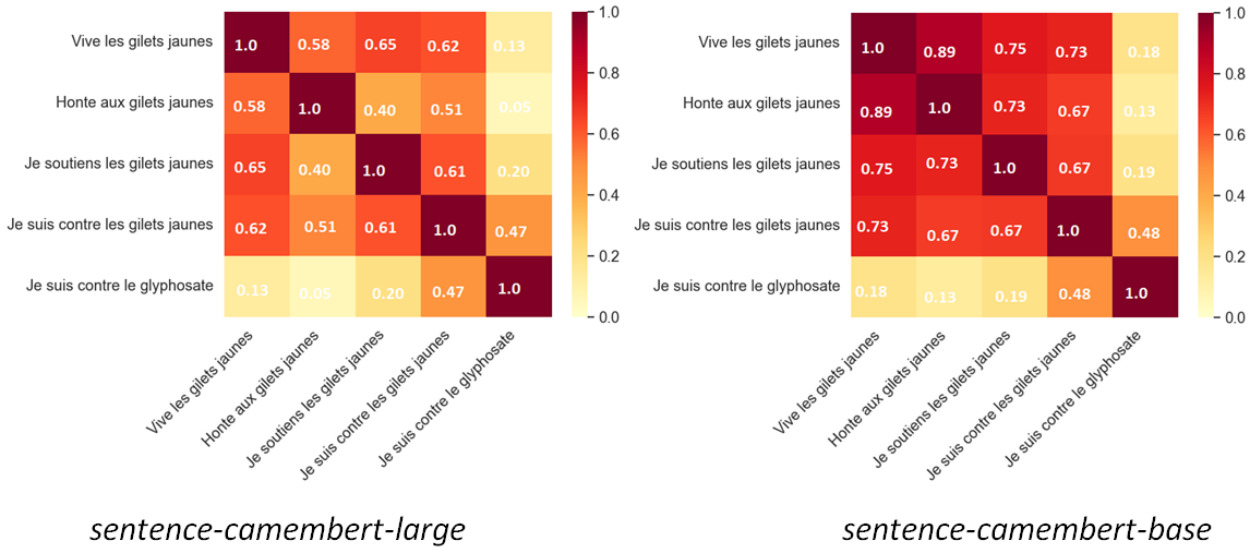


FIGURE 3 – Matrices de confusion indiquant les similarités générées par chaque modèle entre des phrases contradictoires

5.4 Clustering des phrases

La dernière hypothèse, qui permet d'interroger les résultats du clustering, se focalise sur le choix de l'algorithme et la configuration des hyperparamètres. Pour valider cette hypothèse, la méthode proposée s'emploie à explorer d'autres algorithmes de clustering, tels que K-means (Jin & Han, 2010), DBSCAN (Ester *et al.*, 1996), Classification Ascendante Hiérarchique (CAH) (Cecil C. Bridges, 1966), ainsi que des modèles de topic modeling, tels que Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) ou BERTopic (Grootendorst, 2022), avec différentes configurations d'hyperparamètres. Ces algorithmes seront appliqués sur les vecteurs de phrases calculés avec le modèle *sentence-camembert-large*, et celui qui permettra d'obtenir les meilleurs regroupements sera sélectionné [15]. Le critère principal pour évaluer les performances des algorithmes sera le nombre de phrases classées.

6 Conclusion

Dans cet article, une méthode permettant d'effectuer une analyse sémantique des contributions citoyennes du corpus des Cahiers citoyens a été proposée afin d'identifier les différents sujets abordés par les citoyens en vue de les exploiter pour réaliser une étude spatiale du corpus. Les résultats de (Dominguès & Jolivet, 2024) seront également exploités dans cette étude spatiale. La démarche énoncée dans ce papier présente les différentes études et analyses effectuées sur ce corpus, remettant en question certaines méthodes qui n'ont pas donné les résultats souhaités. Elle émet ainsi des hypothèses sur les raisons de ces échecs et propose des solutions alternatives.

La méthode proposée cherche à regrouper les contributions citoyennes en fonction des thématiques

15. Cette étape est en cours de réalisation et le critère du choix de l'algorithme offrant le meilleur regroupement est en cours d'étude.

abordées, en appliquant un clustering sur les phrases du corpus des Cahiers citoyens. Cet article décrit diverses démarches permettant d'obtenir les meilleurs regroupements, tout en proposant une méthode de segmentation des contributions en phrases, de pré-traitement et nettoyage des phrases, de calcul des vecteurs des phrases, et d'utilisation d'algorithmes de clustering.

Étant donné que ce travail est toujours en cours de réalisation, de nouvelles hypothèses et méthodes peuvent être envisagées en fonction des résultats obtenus et de l'évolution de cette recherche, qui, comme indiqué précédemment, s'inscrit dans le cadre d'une thèse.

Remerciements

Je tiens à exprimer ma profonde gratitude envers Catherine Dominguès pour son soutien et ses conseils précieux tout au long de la rédaction de cet article. Son expertise et ses suggestions ont grandement enrichi le contenu, et ses corrections attentives ont contribué à améliorer la clarté et la cohérence du texte. Je suis particulièrement reconnaissant pour ses critiques constructives qui ont permis d'affiner les idées et de renforcer l'argumentation.

De même, je souhaite exprimer mes sincères remerciements à Salomé Chandora pour sa générosité en mettant à disposition le code nécessaire à la génération des matrices de confusion illustrées dans les figures.

Enfin, je tiens à remercier Sabine Ploux pour ses différents conseils et orientations durant les réunions de ma thèse.

Références

BENDINELLI M. (2023). Sens et matérialités d'un cahier citoyen : le cas de la ville de Dole (Jura). *Mots. Les langages du politique*, **131**, 145–169.

BERGER R., BLUENOVE & COGNITO (2019). *Analyse des contributions libres : Cahiers citoyens, courriers et emails, comptes-rendus des réunions d'initiative locale*. Rapport interne.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

CECIL C. BRIDGES J. (1966). Hierarchical cluster analysis. *Psychological Reports*, **18**(3), 851–854. DOI : 10.2466/pr0.1966.18.3.851.

CHANDORA S. (2023). *Fouille sémantique et spatiale dans le corpus Cahiers citoyens : comparaison de méthodes symbolique et numérique*. Mémoire de master, Institut National des Langues et Civilisations Orientales, LASTIG - Univ Gustave Eiffel - ENSG - IGN, France.

DOMINGUÈS C. & JOLIVET L. (2024). Analyse textométrique et spatialisée des cahiers citoyens. In *JADT 2024 : 17th International Conference on Statistical Analysis of Textual Data*.

ESTER M., KRIEGEL H.-P., SANDER J., XU X. *et al.* (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, p. 226–231.

GROOTENDORST M. (2022). Bertopic : Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv :2203.05794*.

HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spacy : Industrial-strength natural language processing in python. *Journal of Open Source Software*, **5**(51), 2456. DOI : 10.5281/zenodo.1212303.

JIN X. & HAN J. (2010). *K-Means Clustering*, In C. SAMMUT & G. I. WEBB, Éds., *Encyclopedia of Machine Learning*, p. 563–564. Springer US : Boston, MA. DOI : 10.1007/978-0-387-30164-8_425.

KOCAMAN V. & TALBY D. (2021). Spark NLP : natural language understanding at scale. *CoRR*, **abs/2101.10848**.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MONNIER M. (2023). *L'analyse spatiale des Cahiers citoyens appliquée au thème de l'écologie*. Mémoire de master, École des hautes études en sciences sociales.

NILS REIMERS I. G. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *https ://arxiv.org/abs/1908.10084*.

PLOUX S., GENAY M. & PLOUX-CHILLÈS L. (2021). Les mots du Grand Débat national : les réseaux lexicaux des contributions déposées sur trois plateformes. *Humanités numériques*, (4). DOI : 10.4000/revuehn.2655, HAL : hal-03511103.

RAY M. (2023). *Analyse du corpus Cahiers citoyens*. Mémoire de master, Université Paris Cité.

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

Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats

Martina Barletta[1, 2]
(1) LIDILEM, Université Grenoble Alpes, 1086-1366 Avenue Centrale, 38400, Saint-Martin-d'Hères, France
(2) Dipartimento di Scienze Umane per l'Educazione "Riccardo Massa", Université Milan-Bicocca, 1 Piazza dell'Ateneo Nuovo, 20126, Milan, Italie
martina.barletta@univ-grenoble-alpes.fr

RÉSUMÉ

La recherche Scolinter s'intéresse à l'étude des compétences en écriture des élèves de l'école primaire en France, en Italie et en Espagne. Le corpus éponyme se présente comme un large corpus longitudinal d'écrits d'élèves comparables dans les trois langues (Ponton *et al.*, 2021). Il s'agit dans cette recherche de créer un outillage TAL applicable à ce type de corpus pour assister les chercheurs dans la description linguistique des phénomènes qui relèvent de la cohésion et de la cohérence textuelle, en particulier de la continuité référentielle. La première étape de cette recherche a consisté dans la conception d'un modèle et dans le choix d'un format d'annotation répondant à ces objectifs. Cette contribution fera tout d'abord un état des recherches sur l'annotation en anaphore, coréférence et continuité référentielle avant de présenter les spécificités du corpus Scolinter et de proposer des pistes méthodologiques pour la suite du travail.

ABSTRACT

Annotating referential continuity in a children's writing corpus - first results

The Scolinter research project investigates the writing proficiency of primary school students in France, Italy, and Spain. The eponymous corpus consists of a large longitudinal corpus of comparable children's writing in these three languages (Ponton *et al.*, 2021). The aim of this project is to create a NLP tool tailored to this corpus to assist researchers in the linguistic description of phenomena relating to cohesion and textual coherence, in particular referential continuity. To meet these goals, the first stage of this research consists in designing a model and choosing an annotation format meeting these goals. This paper reviews the research on anaphora, coreference and referential continuity, presents the specific features of the Scolinter corpus, suggesting methodological pathways for further work.

MOTS-CLÉS : corpus scolaires, TAL, continuité référentielle, annotation de corpus.

KEYWORDS: children's corpora, NLP, referential continuity, annotated corpora.

1 Linguistique de corpus, corpus scolaires et annotation de la continuité référentielle

La linguistique de corpus s'intéresse à l'identification de phénomènes linguistiques sur une grande quantité de données. Couplée aux outils du traitement automatique des langues (TAL), elle permet d'attester et vérifier de manière empirique des hypothèses quant au fonctionnement du langage, encore plus dans les cas où ces phénomènes concernent des exemples de langue non standard, comme

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 28–41.

dans l'analyse d'écrits scolaires. Dans le traitement de ce type de textes éloignés de la norme, les outils méthodologiques de la linguistique de corpus associés à l'outillage du TAL se heurtent à ces difficultés spécifiques (Wolfarth, 2019). En effet, l'utilisation d'outils développés pour de la langue standard s'avère délicate sur des corpus aussi fautifs. Toutefois, ces traitements automatiques sont nécessaires car ils permettent de traiter de vastes corpus fournissant ainsi des analyses fondées sur la réalité langagière (Jacques, 2005) ; analyses qui peuvent ensuite nourrir la réflexion didactique et la formation des enseignants (Elalouf, 2011; Elalouf & Perrin, 2019). Ceci est d'autant plus nécessaire quand les phénomènes analysés ont été rarement décrits de manière empirique, comme dans le cas du développement de la cohérence et de la cohésion textuelle dans les textes d'élèves de l'école primaire.

1.1 Les corpus Scoledit et Scolinter

Le corpus Scoledit représente actuellement un des seuls corpus longitudinaux d'écrits scolaires entièrement transcrits et librement accessibles en ligne (Ponton *et al.*, 2021). Ce corpus rassemble des textes narratifs sollicités auprès d'élèves de différentes écoles en France par les chercheurs du projet. Son but est de suivre de manière longitudinale le développement des compétences d'écriture des mêmes élèves, suivis du CP au CM2. Néanmoins, sur les 4 300 textes recueillis (Wolfarth, 2019, p. 101), 1 820 constituent la partie véritablement longitudinale du corpus, incluant 5 écrits par élève (un pour chaque année scolaire, du CP au CM2). Les productions écrites ont été recueillies par les chercheurs dans 38 écoles de 4 académies françaises (Bordeaux, Clermont-Ferrand, Grenoble et Lyon) sur la base de deux consignes, composées d'images et proposées aux élèves lors du recueil. La consigne utilisée en classe de CP propose de raconter l'histoire d'un petit chat à partir de 4 vignettes de type bande-dessinée. La consigne adoptée pour les classes de CE1 à CM2 propose aux élèves de choisir un ou des personnages présentés sur les vignettes (un chat, un loup, un robot et une sorcière), puis d'écrire une histoire les mettant en scène. L'utilisation de cette même consigne sur les quatre niveaux facilite la comparaison entre les productions produites au fil des années (Wolfarth, 2019). S'appuyant sur la même méthodologie de conception du corpus Scoledit, le projet Scolinter (Ponton *et al.*, 2021) vise la constitution d'un corpus de textes d'élèves du primaire en France, en Italie et en Espagne avec un suivi longitudinal des mêmes cohortes d'élèves. Les objectifs de ce corpus sont d'étudier les compétences en littéracie des élèves à chaque niveau, ainsi que l'évolution de ces compétences dans les trois langues tout au long de l'école primaire (Ponton *et al.*, 2021), sur la base de textes produits à partir d'une même consigne. Le corpus Scolinter comprend en l'état actuel : les 1 820 textes qui forment la partie longitudinale du corpus Scoledit ainsi que 1 333 textes en italien déjà traités, et 813 textes en espagnol. Le corpus français est complet sur les 5 niveaux de primaire alors qu'une partie des textes des corpus italien et espagnol est encore en phase de transcription et de normalisation, notamment pour les niveaux 2, 3 et 5 en Italie (équivalents aux niveaux CE1, CE2 et CM2), et pour les niveaux 3, 4 et 5 en Espagne (équivalents aux niveaux CE2, CM1 et CM2). Actuellement, nous exploitons ce corpus en nous intéressant plus particulièrement aux aspects cohérence/cohésion des textes à travers l'étude des chaines de continuité référentielle (Garcia-Debanc *et al.*, 2021) portant sur les personnages des histoires produites par les élèves. Dans cet objectif, nous annotons les mentions et les chaines de continuité référentielles qui font référence aux personnages induits par la consigne ou bien insérés par l'élève dans l'intrigue de la narration.

Le corpus Scolinter est composé des scans, des transcriptions ainsi que des normalisations des textes recueillis. Nous avons choisi d'annoter la version normalisée des textes des élèves, c'est-à-dire une version orthographiquement normée restant au plus près de la production initiale (Wolfarth, 2019). Elle permet à la fois des comparaisons avec la production de l'élève à différents niveaux (lexical, orthographique, morphologique...) et le recours aux outils TAL (Wolfarth *et al.*, 2018). Dans cette

normalisation, qui a été effectuée au format XML, des balises ont été introduites pour indiquer certains phénomènes spécifiques. Par exemple, une balise est proposée pour indiquer les tokens omis par l'enfant lors du processus d'écriture, là où le mot oublié influe sur la construction syntaxique de la phrase en question. Ceci constitue un obstacle pour notre travail, car ces omissions portent dans la plupart des cas sur des formes pronominales (51% des balises d'omission présentes du CE1 au CM2). L'absence de ces termes modifie l'analyse morphosyntaxique et en dépendance de la phrase, sur laquelle s'appuiera le processus d'annotation et d'analyse des chaines de continuité référentielle dans la suite de notre travail. Ces balises contiennent en l'état actuel seulement une proposition de catégorie grammaticale à laquelle le mot appartient, ce qui rend complexe la mise en œuvre d'une reconstruction automatisée de ces tokens dans les textes. Les textes ont été traités à l'aide du modèle transformeur bert-cased Flaubert (Devlin *et al.*, 2019; Le *et al.*, 2020) pour effectuer une substitution de ces balises qui indiquent une omission dans le texte avec le mot le plus probable selon le modèle.

2 Annotation de la continuité référentielle, des chaines de référence, de l'anaphore : état de l'art synthétique

La linguistique de corpus française s'intéresse, depuis des décennies, à l'étude et la description des phénomènes de construction de la textualité comme l'anaphore et les chaines de référence (Chastain, 1975; Corblin, 1985; Charolles, 1988; Corblin, 1995; Schnedecker, 1997, 2021). Bien que des corpus annotés en coréférence ou en anaphore existaient déjà avant la naissance de projets comme ANCOR (Muzerelle *et al.*, 2013) et DEMOCRAT (Landragin, 2016) en France, leurs caractéristiques ne les rendaient pas globalement représentatifs de la coréférence ou utilisables pour l'apprentissage profond (Grobol, 2020), soit parce que les annotations codent seulement certains types d'anaphore, comme dans le cas du corpus ARCADE (Tutin *et al.*, 2000), soit à cause de leur taille réduite, comme dans le cas du corpus DéDé (Gardent & Manuélian, 2005). En effet, même si ARCADE constitue un des premiers grands corpus annotés en anaphore pour le français avec son million de mots, seules les expressions anaphoriques et cataphoriques appartenant à des catégories fermées ont été retenues, peu importe leur antécédent. Les expressions anaphoriques et cataphoriques annotées sont : les pronoms personnels (à exception du pronom réfléchi), les pronoms et déterminants possessifs, les pronoms démonstratifs à l'exception des pronoms démonstratifs « neutres » (*ce, ça, cela, ceci*), les pronoms indéfinis et numéraux ; les adverbes anaphoriques comme *dedans, dessus* ; les expressions nominales anaphoriques ainsi que les « pointers » anaphoriques (*ce dernier, le premier*). Ces annotations encodent aussi la relation avec l'antécédent ainsi que la relation discursive et sémantique entre l'expression anaphorique et son antécédent (Tutin *et al.*, 2000). Cependant, dans ce corpus, les descriptions définies [1] ne sont pas annotées (Tutin *et al.*, 2000; Gardent & Manuélian, 2005).
Le corpus DéDé, pour sa part, cible l'annotation des descriptions définies, « c'est-à-dire les expressions de la forme *le/la/les N* » pour en rendre possible l'annotation automatique (Gardent & Manuélian, 2005, p. 3). Ce corpus est composé d'articles du journal Le Monde et il comprend 48 360 mots annotés au niveau morphosyntaxique, dont 4 910 descriptions annotées selon quatre catégories différentes : description autonome, description coréférentielle, description contextuelle, description non référentielle. Sa méthodologie d'annotation prévoyait l'utilisation d'outils de prétraitement de l'annotation (annotation morphosyntaxique) pour faciliter cette tâche, réalisée par des linguistes expérimentés. Le schéma d'annotation utilisé était affiné par plusieurs itérations après une première phase d'annotation, dont l'objectif était de résoudre les possibles désaccords entre annotateurs et d'intégrer les modifications nécessaires (Gardent & Manuélian, 2005), selon une stratégie déjà utilisée dans divers travaux d'annotation (Brants, 2000; Erk *et al.*, 2003; Gardent & Manuélian, 2005).

1. À savoir, les syntagmes du type article défini + nom.

Ces projets ont contribué à ouvrir la voie aux réflexions sur l'annotation de ces phénomènes sur des corpus de grande taille, ciblés dans certains cas pour l'entrainement de modèles *machine learning*. Le corpus Annodis (Péry-Woodley *et al.*, 2011), le corpus ANCOR (Muzerelle *et al.*, 2013) pour l'oral spontané, et le corpus DEMOCRAT (Landragin, 2016) pour la langue écrite représentent les trois principaux corpus qui encodent l'anaphore et/ou la coréférence à différents niveaux en français. Nous allons par la suite présenter les objectifs et la composition de ces corpus, ainsi que les méthodes d'annotations adoptées et les phénomènes qu'ils encodent. Nous aborderons également le corpus RésolCo (Garcia-Debanc *et al.*, 2019, 2021) qui représente le seul corpus français d'écrits scolaires de taille moyenne annoté en continuité référentielle, ce qui le rend similaire en objectifs et en méthodologie au corpus Scolinter.

2.1 Annodis

Le corpus Annodis (Péry-Woodley *et al.*, 2011) qui précède le corpus DEMOCRAT en élaboration et publication, tout en ayant une taille comparable à ce dernier, ne représente pas un corpus annoté en coréférence car il encode des chaines topicales, donc des segments caractérisés par l'apport d'informations au sujet d'un seul et même référent (Federzoni *et al.*, 2020). Il reste cependant un exemple de corpus de grande taille où des structures discursives ont été annotées manuellement. Le corpus Annodis est composé de textes variés et sélectionnés selon plusieurs critères, comme « le genre, la longueur et le type d'organisation discursive » (Péry-Woodley *et al.*, 2011, p. 72), et « il est le résultat de deux types d'annotations manuelles », une annotation qui part des unités élémentaires du discours pour reconstruire les relations rhétoriques entre unités du texte, dans une « démarche ascendante » qui vise à « construire la structure complète d'un discours » , et une deuxième annotation qui s'intéresse plutôt à la « mise en texte » et vise l'annotation sélective de structures discursives multi-échelles dont les structures énumératives et les chaines topicales (Péry-Woodley *et al.*, 2011, p. 72). La méthodologie et les questionnements en annotation qui ont marqué ce projet ont constitué les jalons du travail d'annotation fait sur le corpus RésolCo (Garcia-Debanc *et al.*, 2019, 2021), ce dernier étant lui-même inspiré des travaux effectués sur le corpus DEMOCRAT (Landragin, 2022).

2.2 ANCOR

Le corpus ANCOR représente « le premier corpus d'oral spontané d'envergure annoté en coréférence » et en anaphore pour la langue française et distribué librement (Muzerelle *et al.*, 2013). Il est composé de plusieurs corpus de parole spontanée transcrite (Accueil_USB, OTG et ESLO) et il compte 418 000 mots. Son objectif était de répondre au manque d'un corpus francophone en libre accès et « de taille suffisante pour entrainer un système de résolution de la coréférence efficace » (Muzerelle *et al.*, 2013, p. 2), à un moment où d'autres langues majoritaires étaient déjà dotées de tels corpus. Cependant, ce corpus était représentatif du français parlé conversationnel, alors qu'un corpus de taille similaire annoté pour la langue écrite était encore absent du panorama francophone. L'annotation a été réalisée de manière déportée sur ce corpus par deux annotateurs en quatre phases : repérage des entités nommées par un annotateur, consensus entre annotateurs par rapport à cette première annotation, repérage et caractérisation des relations anaphoriques par un annotateur et révision finale des relations caractérisées par un superviseur. Cette démarche a été adoptée pour éviter une « surcharge cognitive » des codeurs et pour favoriser l'accord interannotateur (Muzerelle *et al.*, 2013, p. 558).

2.3 DEMOCRAT

Le corpus DEMOCRAT répond à l'inexistence de grands corpus annotés représentatifs de l'écrit. Il constitue « le premier corpus de grande taille librement disponible pour le français écrit » (Landragin, 2021, p. 12) : 560 000 mots dont 198 000 expressions référentielles annotées et 20 000 chaines

de référence. Il est constitué de 58 textes appartenant à plusieurs genres, du roman aux articles journalistiques en passant par des textes littéraires historiques. L'objectif initial de ce projet était de constituer un corpus diachronique de textes, écrits entre le 12e et le 21e siècle, relevant de genres textuels variés, « et d'en autoriser des exploitations par des outils de traitement automatique des langues (TAL), plus précisément par des outils faisant appel à de l'apprentissage profond » (Landragin, 2022, p. 50). Ce corpus, sur lequel sont intervenues plus de 40 personnes, a fait l'objet de plusieurs expérimentations d'annotations pour vérifier la faisabilité du guide proposé. Il a également fait l'objet de séances d'annotation chronométrées pour calculer l'effort nécessaire pour réaliser l'annotation du corpus dans son entièreté. Après l'étape d'annotation manuelle, un script a été utilisé sur l'ensemble du corpus pour obtenir automatiquement la construction des chaines de référence à partir des annotations des différentes expressions référentielles (Landragin, 2021). Les choix méthodologiques fondamentaux du projet DEMOCRAT, qui le démarquent de projets existants similaires, sont le fait d'avoir annoté les chaines tout au long des textes inclus dans le corpus, ainsi que le fait d'avoir allié un travail d'annotation automatique à une génération automatique des chaines (Landragin, 2021, p. 20).

2.4 RésolCo

Dans le domaine des corpus d'écrits scolaires, dans le contexte du projet E-Calm, le corpus RésolCo (Garcia-Debanc *et al.*, 2017, 2021) a abordé l'annotation de la continuité référentielle sur des écrits de niveaux scolaires variés. Composé d'environ 400 textes, il a été récolté dans des classes de niveaux différents, du CE2 à l'université (Garcia-Debanc *et al.*, 2021). En s'appuyant sur l'expérience d'annotation faite lors de la conception d'Annodis ainsi que sur le guide et certains points méthodologiques du projet DEMOCRAT, ce corpus se donne le double objectif de (1) constituer une ressource annotée en continuité référentielle et de (2) élaborer une cartographie des formes linguistiques qui manifestent les compétences textuelles et discursives en cours de développement (Garcia-Debanc *et al.*, 2021). Les textes qui constituent ce corpus annoté ont été produits à partir de la même consigne, qui impose aux élèves la résolution de problèmes de cohésion textuelle (Garcia-Debanc & Bonnemaison, 2014; Garcia-Debanc & Bras, 2016). Cette consigne intègre trois phrases contenant des pronoms personnels (« ils » et « elle ») et des syntagmes nominaux introduits par un déterminant démonstratif (« cette maison », « ce grand bruit », « cette aventure »). La tâche impose aux élèves d'insérer ces trois phrases dans un texte narratif fictionnel (Garcia-Debanc *et al.*, 2021). Le phénomène annoté est défini « continuité coréférentielle » car les mentions annotées ne sont pas seulement celles qui représentent la coréférence stricte. Ces annotations sont aussi circonscrites « aux seuls référents présents dans l'ensemble des textes du corpus ; autrement dit, aux référents provoqués par la consigne RésolCo » (Garcia-Debanc *et al.*, 2021, p. 104). Le choix d'une annotation de ce type permet à l'annotateur de se concentrer sur les référents qui jouent un rôle de premier plan dans le récit, et permet aussi de créer des annotations comparables entre textes au même niveau ou tout au long du corpus.

Corpus	Année	Taille	Mentions annotées	Genre
ANCOR	2013	487 000 tokens	116 000 mentions, 51 000 relations anaphoriques	parole spontanée transcrite
DEMOCRAT	2019	58 textes 560 000 tokens,	198 000 mentions, 20 000 chaines de référence	écrits narratifs et autres genres variés
RésolCo	2021	385 textes, 72 873 tokens	12 261 mentions	écrits scolaires

TABLE 1 – Résumé des caractéristiques des corpus analysés

3 Annotation en continuité référentielle du corpus Scolinter

Bien que ces différents corpus s'intéressent tous aux phénomènes de cohérence et de cohésion textuelles, ils présentent des spécificités, à la fois liées aux genres textuels annotés et aux objectifs que les annotations contribuent à réaliser. Afin de mettre en place une annotation répondant au mieux à nos propres objectifs de recherche, nous avons comparé différents choix opérés, notamment au niveau méthodologique, dans les projets connexes, pour ensuite effectuer les choix les plus pertinents par rapport à nos objectifs d'annotation. Ce travail a abouti à la création d'une première version de notre guide d'annotation que nous avons ensuite testé sur notre corpus.

Afin de vérifier l'applicabilité du même guide sur deux des langues présentes dans le corpus, nous avons annoté 15 textes en français et 15 textes en italien. Cette première itération de test a été réalisée par plusieurs annotateurs experts et a comportée plusieurs sessions d'adjudications sur la base desquelles nous avons ultérieurement clarifié les descriptions des expressions linguistiques à annoter dans le guide. Ces premières annotations ont été utilisées ensuite pour produire les exemples montrés dans le guide. Certains des choix que nous avons faits à cette étape sont techniques et méthodologiques à la fois, comme par exemple le choix d'annoter les référents multiples à travers une superposition des étiquettes des référents indiqués, dans le tentative de résoudre partiellement le problème de l'annotation de l'anaphore discontinue.
Une deuxième itération d'application du guide a été effectuée sur des échantillons réduits de textes d'élèves du CE2, annotés par 22 annotateurs experts (étudiants en master de sciences du langage), à hauteur de 14 textes par binôme, ce qui nous a permis de vérifier la stabilité des lignes directrices décrites dans le guide. Les annotations obtenues par chaque binôme ont été analysées du point de vue qualitatif, selon les observations faites par les différents binômes dans leurs rapports finaux. Certaines des observations issues de ces annotations nous ont permis d'affiner ultérieurement les critères d'annotations, et d'apporter davantage d'exemples tirés du corpus dans le guide même, notamment en ce qui concerne la délimitation de certains types de mentions constituées par syntagmes nominaux. Cette itération a été cruciale pour la définition dans le guide du statut de l'annotation des pronoms dans le discours direct, et pour la première définition d'entité à annoter dans les textes en tant qu'entité animée et actante dans l'univers du texte analysé.
Suite à ces deux itérations de test du guide, nous avons mené une campagne d'annotation sur une partie restreinte du corpus français, sélectionnée pour être le plus représentative possible de la variété attestée dans le corpus longitudinal. Grâce à cette campagne nous avons obtenu un corpus de référence (corpus gold) qui nous a permis d'ébaucher une description des caractéristiques des chaines présentes dans des textes d'école primaire.

3.1 Le guide d'annotation

Le guide conçu pour l'annotation de la continuité référentielle s'inspire du travail effectué sur le corpus RésolCo, dans lequel sont annotées les chaines provoquées par les entités imposées par la consigne. En ce qui concerne Scoledit, nous avons décidé d'annoter les mentions (y compris singletons et anaphores) et les chaines relatives aux personnages présents dans la consigne ou ajoutés par l'élève. Dans le jeu d'étiquettes conçu, nous incluons les quatre personnages de la consigne (*cat*, *witch*, *robot*, *wolf*) ainsi que ces personnages « externes » qui interviennent de manière active dans l'intrigue du texte annoté à travers l'étiquette *extN*. Le *N* est rajouté à chaque nouvelle apparition d'un personnage. Dans le cas où seraient présents plusieurs personnages appartenant à la même catégorie, par exemple deux chats, les annotateurs rajoutent un chiffre en fin d'étiquette (*cat* et *cat1* si deux chats sont présents dans l'histoire, *cat2* si un troisième intervient et ainsi de suite). Concernant les passages à annoter, nous avons suivi les principes décrits dans le guide RésolCo, qui s'inspire lui-même du guide

d'annotation utilisé par le projet DEMOCRAT. Nous annotons : les syntagmes nominaux, les noms propres, les pronoms personnels, corrélés, objets et relatifs, les déterminants possessifs, l'anaphore zéro dans des phrases coordonnées ou dans les verbes à l'impératif en français et dans les phrases où le sujet n'est pas explicite en italien et en espagnol. Notre annotation tout comme celle de RésolCo ne relève pas de l'annotation de la coréférence stricte mais plutôt de la « continuité référentielle » au sens large, y compris les relations anaphoriques et les singletons présents dans les textes selon les critères définis auparavant. Nous n'annotons pas toutes les entités présentes dans les textes mais nous sélectionnons ces entités que nous considérons animées et récurrentes dans les textes pour permettre de comparer les résultats de manière longitudinale puis contrastive entre les trois langues du corpus.

3.2 Constitution et annotation du corpus de référence

Les textes à annoter ont été sélectionnés depuis le corpus longitudinal français, qui contient des textes des mêmes 337 élèves du CP au CM2. Le critère de sélection choisi est celui de la représentativité en termes de distribution par longueur des textes dans le corpus longitudinal sur chaque niveau scolaire. Après avoir effectué les traitement décrit dans 1.1 sur l'intégralité du corpus, nous avons réalisé des statistiques quant au nombre de tokens par texte. De cette manière, nous avons obtenu une image de la distribution des textes par nombre de tokens sur chaque niveau, pour pouvoir ensuite reproduire cette même distribution sur notre corpus de référence. Nous avons décidé d'exclure du corpus de référence les textes entre 1 et 20 tokens soit 21 textes au total, car nous les avons considérés trop courts pour contenir suffisamment d'informations quant à la cohérence textuelle. Nous avons partitionné le corpus par niveau et selon le nombre de tokens par texte (tranches 20-50 tokens, 51-100 tokens, 101-150 tokens etc.). Dans chaque tranche et pour chaque niveau, nous avons sélectionné aléatoirement le même pourcentage de textes que sur le corpus global pour conserver la même distribution. Nous avons décidé de rajouter davantage de textes en CE1 car la majorité des textes dans ce niveau sont assez courts et en CM2 pour enrichir l'échantillon de textes plus longues et donc susceptibles de contenir des chaines plus longues et complexes. Le corpus de référence est composé au final de 111 textes pour 16 838 tokens. Le corpus ainsi obtenu a été préalablement tokenisé automatiquement grâce à la librairie spacy-conll [2] de Spacy [3] et enregistré dans des fichiers au format CoNLL-U. Ces fichiers sont ensuite importés sur la plateforme INCEpTION (Klie *et al.*, 2018). Les textes ont été enfin annotés à l'aide de cette plateforme par deux annotateurs experts. Le schéma d'annotation s'appuie sur le layer d'annotation de la coréférence par défaut existant dans INCEpTION mais adapté à notre travail en proposant certaines étiquettes spécifiques comme *cat, witch, wolf*, et *robot*, qui permettent de suivre les personnages de la consigne et des étiquettes *ext1, ext2 et ext3* pour les autres personnages. Nous avons laissé les annotateurs libres de rajouter, si besoin, de nouvelles étiquettes sur la base de lignes directrices du guide (par exemple *ext4, wolf2*, etc.). Les analyses que nous présentons par la suite ont été effectuées à partir de l'export fourni par INCEpTION au format WebAanno TSV v3.3 grâce à des codes Python ciblés sur l'analyse des mentions et des chaines de référence annotées dans ce corpus.

4 Observation et résultats

Les annotations obtenues lors de cette première campagne sont en cours d'adjudication. Pour l'heure, l'adjudication a été opérée sur les textes présentant un écart important entre le nombre d'entités annotées et/ou le nombre de mentions annotées. Ceci a permis de faire ressortir les principales difficultés dans l'utilisation du guide d'annotation. Le processus d'adjudication en cours porte surtout sur des éléments tels que les personnages à annoter dans les textes (voir 4.2), sur l'annotation de

2. Disponible au lien suivant https://pypi.org/project/spacy-conll/. Consulté le 10/02/2024
3. Disponible au lien suivant https://spacy.io/. Consulté le 10/02/2024.

l'anaphore zéro (voir 4.3), ainsi que sur les limites des mentions constituées par des syntagmes nominaux, qui va nécessiter davantage d'explications surtout par rapport à la non inclusion des prépositions dans les syntagmes nominaux. Les résultats même partiels de l'adjudication nous permettent de faire ressortir certaines caractéristiques globales du corpus et des chaines annotées comme le nombre de maillons par texte, le nombre de référents annotés, la longueur des chaines, la présence des personnages issus de la consigne, etc. Ces premières mesures seront à confirmer une fois la phase d'adjudication achevée.
De nos annotations, nous avons exclu les singletons dans les calculs ici effectués. Nous avons retenu pour ces statistiques les chaines composées au moins de deux maillons, en raison de la longueur des textes présents dans notre corpus : certains textes sont assez courts (de 20 à 50 tokens) et on retrouve des chaines limitées à deux maillons par personnage. Ces chaines de deux maillons représentent 13% des chaines annotées dans nos textes. Les singletons concernent habituellement des personnages secondaires et non pas les quatre personnages de la consigne, par conséquent cette exclusion n'a pas eu d'impact sur les statistiques relatives à la présence des personnages de la consigne dans les textes.

4.1 Caractérisation du corpus de référence

Pour chaque texte du corpus de référence, nous avons calculé : le nombre de tokens, le nombre de maillons annotés (à l'exclusion des singletons), le nombre des chaines (à partir de deux maillons), la densité référentielle [4] et la longueur des chaines annotées. Nous avons pu observer que le nombre de tokens moyen par texte augmente avec le niveau scolaire, en même temps que le nombre d'entités présentes dans les textes et de maillons annotés. Comme remarqué par Landragin *et al.* (2024), en général, les textes narratifs présentent une densité référentielle plus importante que les textes appartenant à d'autres genres textuels [5], et cela confirme la densité calculée sur notre corpus (18,59% en moyenne sur tout le corpus de référence).
Cependant, une première comparaison quantitative entre les versions des deux annotateurs nous a permis d'observer un certain désaccord entre annotateurs, notamment en ce qui concerne le nombre de mentions ainsi que le nombre d'entités annotés dans les textes. Comme décrit dans 3.2, nous avons utilisé pour nos annotations le layer d'annotation de la coréférence proposé par défaut par INCEpTION. Même si la plateforme propose habituellement des fonctionnalités pour faciliter les étapes d'adjudication et du calcul de l'accord interannotateur, celles-ci n'ont pas été implémentées sur le niveau par défaut d'annotation de la coréférence. A terme, nous prévoyons de mettre en place des méthodes plus efficaces pour le calcul de l'accord interannotateur, ainsi que la création de notre propre layer d'annotation sur INCEpTION qui puisse nous permettre de mesurer cet accord interannotateur directement depuis la plateforme. Nous avons néanmoins pu observer des différences quant au nombre d'entités et de maillons annotés. Entre annotateurs, nous avons pu remarquer un écart type moyen de 2,06 sur le nombre de mentions annotées par texte et un écart type moyen de 0,55 sur le nombre d'entités annotées par texte. Parmi les différences observées dans une première étude qualitative, deux sont particulièrement saillantes, et en lien avec les observations quantitatives mentionnées. La première concerne l'annotation d'entités « externes » aux quatre personnages présents dans la consigne et la deuxième porte sur l'annotation de l'anaphore ou sujet zéro. Le tableau 2 résume les statistiques descriptives pour chaque niveau scolaire représenté dans le corpus de référence.

4. Nous définissons ici la densité référentielle comme le nombre des maillons divisé par le nombre de tokens.
5. Dans le corpus DEMOCRAT, la densité référentielle des textes narratifs du 16e siècle se situe à plus de 20,55%. (Landragin *et al.*, 2024)

Niveau	Nb textes	Nb tokens	Nb moyen tokens par texte	Maillons	Nb moyen de référents par texte	Densité référentielle	Longueur moyenne des chaines
CE1	32	2 388	74,63	461	2,22	19,30%	6,54
CE2	25	3 475	139	651	2,56	18,73%	10,68
CM1	25	4 541	168,19	868	3,11	19,11%	11,01
CM2	27	5 979	221,44	1 066	3,48	17,83%	11,6
Corpus	111	16 383	150,81	3 046	2,84	18,59%	9,96

TABLE 2 – Résumé des statistiques sur les chaines annotées dans le corpus de référence, inspiré de Landragin *et al.* (2024)

4.2 Ambiguïté du guide : l'identification des personnages « externes » à la consigne

Bien que la consigne de la tâche d'écriture cherche à imposer l'utilisation d'un ou deux personnages déterminés, cela n'a pas toujours été respecté par les enfants, et a donné lieu à un corpus riche en représentations de personnages différents. Dans notre guide, nous avons ciblé l'annotation de cette richesse, mais la définition donnée dans le guide de « personnage à annoter » s'est révélée floue par rapport à la réalité des textes auxquels nous avons été confrontés lors du processus d'annotation : si la notion de personnage animé tient à la confrontation avec la réalité des textes à annoter, la définition plus large donnée dans le guide d'« entités animées ou qui effectuent des actions utiles afin de suivre l'intrigue de la narration du texte »[6] n'était pas suffisamment claire et délimitée, ce qui a entrainé un désaccord sur le nombre d'entités annotées dans les textes.[7] Par exemple, dans le texte présenté dans la Figure 1, si on se limite à l'annotation des entités *cat* (présent dans la première partie du texte, qu'on ne reporte pas ici) et à la « protagoniste » *ext1*, on perd la présence des différents maillons qui indiquent les autres référents qui participent aux événements décrits dans le texte.

6. Le guide est actuellement disponible sur demande et pour les annotateurs participant au projet. Il sera publié lors de la publication du corpus.

7. Ce désaccord porte sur environ 50% des textes du corpus de référence.

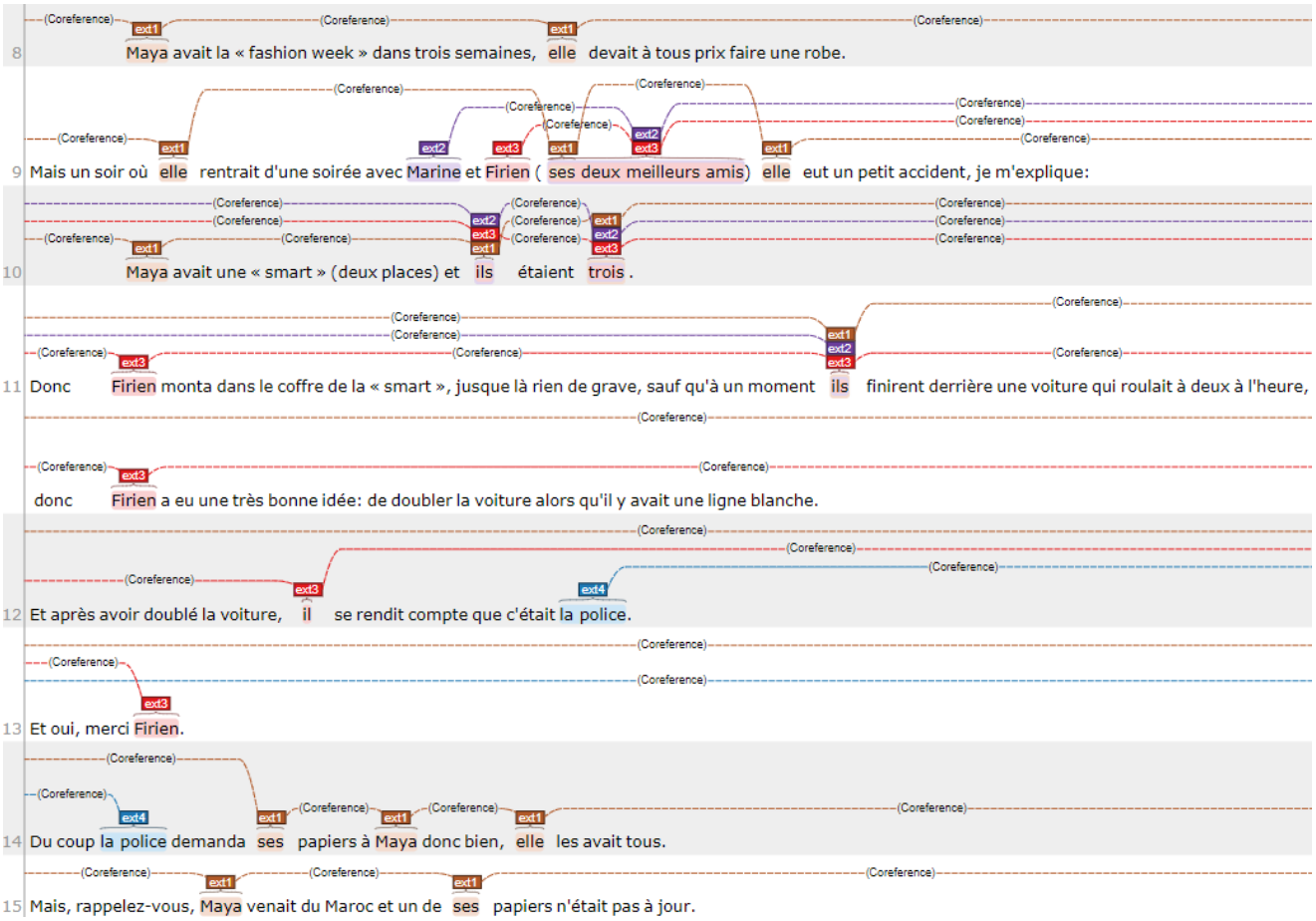


FIGURE 1 – Exemple de présence de personnages externes à la consigne dans les phrases 9 à 15 du texte NORM-EC-CM2-2018-14-D1-S1325. Annotation effectuée sur la plateforme INCEpTION.

Ces observations nous ont permis de faire évoluer la définition d'entité à annoter dans le texte à « entité animée dans le texte donné », en incluant des entités habituellement non animées mais qui prennent la forme de personnage animé dans le texte.

4.3 Anaphore zéro

Un autre point de difficulté que l'on a pu observer lors d'une analyse qualitative des annotations a porté sur l'annotation de l'anaphore zéro, ou sujet zéro. L'annotation de ce phénomène nous semble encore plus importante dans une perspective comparative car les deux autres langues du corpus, l'italien et l'espagnol, sont des langues où le sujet n'est pas obligatoirement exprimé à travers une marque pronominale, ce qui rend le phénomène de l'anaphore zéro très fréquent dans nos textes selon une première étude informelle. Toutefois, dans notre corpus de référence, ce phénomène linguistique n'est pas toujours annoté de manière cohérente. Si sa présence est toujours observable de manière fréquente dans les textes italiens et espagnols du corpus, son annotation pourrait être plus compliquée sur la partie française du corpus. Landragin *et al.* (2024) observent la différence dans l'annotation du sujet zéro entre français ancien et français contemporain, où les annotateurs du français moderne ont tendance à oublier ce phénomène, alors que sa présence et sa fréquence en ancien français « empêche tout oubli » (Landragin *et al.*, 2024, p. 15). De manière similaire entre le français et l'italien, notre hypothèse, suite aux tests effectués en parallèle sur des textes dans les deux langues, est que la

fréquence du phénomène en italien (ou dans la pratique d'annotation des annotateurs italophones) rend la présence de ce phénomène évidente et son annotation presque automatique, alors que cela ne l'est pas pour un annotateur francophone. Dans le texte d'un élève de CE1 (Figure 2), le verbe qui marque une anaphore zéro dans une phrase coordonnée (« voulait ») n'a pas été annoté par l'un des annotateurs. Cette difficulté rencontrée par les annotateurs sera abordée dans le guide, en fournissant davantage d'exemples de possibles formes verbales à annoter dans les textes en français car ils font effectivement partie des mentions à annoter.

FIGURE 2 – Exemple d'anaphore zéro dans la phrase 10 du texte NORM-EC-CE1-2015-130-D1-S1125. Annotation effectuée sur la plateforme INCEpTION.

5 Conclusion

Dans cette contribution, nous avons présenté les différentes étapes ayant mené à la constitution d'un corpus de référence annoté en continuité référentielle issu des textes du corpus Scolinter. Nous avons effectué plusieurs itérations d'annotation, suivies par des redéfinition et clarification du guide d'annotation. Celles-ci ont portées sur le plan technique, comme l'annotation empilée des référents multiples, ainsi que sur le plan méthodologique, comme l'annotation des pronoms dans le discours direct, ou l'éclaircissement des définitions d'anaphore zéro et de personnage à annoter dans les textes. Nous avons pu mener une campagne d'annotation conduite par deux annotateurs experts sur 111 textes, issus des productions d'élèves des niveaux scolaires du CE1 au CM2 du corpus français. Le corpus de référence a été sélectionné selon les critères de représentativité du corpus longitudinal en termes de distribution des textes par longueur sur chaque niveau scolaire retenu pour nos analyses. Cette première campagne d'annotation nous a permis d'effectuer une première description des anaphores et des chaines annotées dans notre corpus. Cette description semble confirmer que les textes présents dans le corpus correspondent à certains critères propres aux textes narratifs de scripteurs confirmés, ce qui confirme que le genre textuel influence profondément la construction des chaines de continuité référentielle. Les divergences que l'on a rencontrées dans les annotations nous ont permis d'identifier les « points faibles » de notre guide. Ceci nous permettra d'en proposer une nouvelle version, qui établira de manière plus claire les critères d'annotation des entités qui ne sont pas représentées dans la consigne et qui représente davantage d'exemples permettant aux annotateurs d'identifier de manière plus précise les occurrences d'anaphore zéro dans les textes du corpus français, et qui constituera la base de départ pour la rédaction d'un guide d'annotation spécifique pour la langue italienne.

Remerciements

Je remercie mon co-encadrant de thèse Claude Ponton pour les relectures de cette contribution ainsi que mes co-encadrantes de thèse Catherine Brissaud et Federica Da Milano pour leur aide.

Références

BRANTS T. (2000). Inter-annotator Agreement for a German Newspaper Corpus. In M. GAVRILIDOU, G. CARAYANNIS, S. MARKANTONATOU, S. PIPERIDIS & G. STAINHAUER, Éds., *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece : European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/333.pdf.

CHAROLLES M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, **57**(1), 3–13. DOI : 10.3406/prati.1988.1468.

CHASTAIN C. (1975). Reference and Context. *Language, mind, and knowledge*, **7**, 194–269.

CORBLIN F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectica. Revue de l'Association pour la Recherche Cognitive*, **1**(1), 123–143. DOI : 10.3406/intel.1985.851.

CORBLIN F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. HAL : ijn_00550962.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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.

ELALOUF M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? *Pratiques. Linguistique, littérature, didactique*, (149-150), 56–70. DOI : 10.4000/pratiques.1702.

ELALOUF M.-L. & PERRIN S. (2019). Entre recherche et formation, quels usages des corpus de textes scolaires ? In *Écrire et faire écrire dans l'enseignement postobligatoire Enjeux, modèles et pratiques innovantes*, p. 197–212. Presses universitaires du Septentrion. DOI : https ://doi.org/10.4000/books.septentrion.77013.

ERK K., KOWALSKI A., PADÓ S. & PINKAL M. (2003). Towards a Resource for Lexical Semantics : A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 537–544, Sapporo, Japan : Association for Computational Linguistics. DOI : 10.3115/1075096.1075164.

FEDERZONI S., HO-DAC L.-M. & REBEYROLLE J. (2020). Les chaînes topicales dans la ressource ANNODIS. *SHS Web of Conferences, Congrès Mondial de Linguistique Française CMLF 2020*, **78**, 11005. DOI : 10.1051/shsconf/20207811005, HAL : hal-02890989.

GARCIA-DEBANC C. & BONNEMAISON K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés. *SHS Web of Conferences, 4e Congrès Mondial de Linguistique Française*, **8**, 961–976. DOI : 10.1051/shsconf/20140801349.

GARCIA-DEBANC C. & BRAS M. (2016). Vers une cartographie des compétences de cohérence et de cohésion textuelle dans une tâche-problème de production écrite réalisée par des élèves de 9 -12 ans : indicateurs de maitrise et progressivité. **Recherches textuelles**(13). HAL : hal-01987031.

GARCIA-DEBANC C., HO-DAC L.-M., BRAS M. & REBEYROLLE J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, **16**, 157–184. DOI : 10.4000/corpus.2783, HAL : hal-01558836.

GARCIA-DEBANC C., HO-DAC L.-M., FEDERZONI S., BRAS M. & REBEYROLLE J. (2019). RésolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence. In *10èmes Journées Internationale de la Linguistique de Corpus*, Grenoble, France. HAL : hal-02877122.

GARCIA-DEBANC C., REBEYROLLE J. & HO-DAC L.-M. (2021). La continuité référentielle dans le corpus RésolCo : méthode d'annotation et premières analyses. *Langue francaise*, **211**(3), 99–114. DOI : 10.3917/lf.211.0099, HAL : hal-03559961.

GARDENT C. & MANUÉLIAN H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Revue TAL*, **46**(1), 115. HAL : halshs-00168567.

GROBOL L. (2020). *Coreference resolution for spoken French*. Thèse de doctorat, Université Sorbonne Nouvelle - Paris 3. HAL : tel-02928209.

JACQUES M.-P. (2005). Pourquoi une linguistique de corpus ? In G. WILLIAMS, Éd., *La linguistique de corpus*, Rivages Linguistiques, p. 21–30. Rennes, presses universitaires de rennes édition.

KLIE J.-C., BUGERT M., BOULLOSA B., ECKART DE CASTILHO R. & GUREVYCH I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. ZHAO, Éd., *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, p. 5–9, Santa Fe, New Mexico : Association for Computational Linguistics.

LANDRAGIN F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92), 11. HAL : hal-01347949.

LANDRAGIN F. (2021). Le corpus DEMOCRAT et son exploitation. Présentation. *Langages*, **224**(4), 11–24. DOI : 10.3917/lang.224.0011, HAL : hal-03474748.

LANDRAGIN F. (2022). Expressions référentielles et chaînes de référence en français : le projet Democrat et son exploration des rapports entre linguistique textuelle et linguistique de corpus. *Echo des études romanes*, **18**(1), 49–65. DOI : 10.32725/eer.2022.004, HAL : halshs-03876206.

LANDRAGIN F., GLIKMAN J., SCHNEDECKER C. & TODIRASCU A. (2024). Chaînes de référence dans le corpus Democrat : une analyse en diachronie longue. *Corpus*, (25). DOI : 10.4000/corpus.8581.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490 : European Language Resources Association.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA, Éd., *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 555–563, Les Sable d'Olonne, France. HAL : hal-01016562.

PONTON C., GUTIÉRREZ-CACERES R., TERUGGI L., FARINA E., BRISSAUD C. & WOLFARTH C. (2021). Scolinter : un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, **211**(3), 37–50. DOI : 10.3917/lf.211.0037, HAL : halshs-03384027.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL : traitement automatique des langues*, **52**(3), 71. HAL : halshs-00935201.

SCHNEDECKER C. (1997). *Nom propre et chaînes de référence*, volume 21 de *Recherches linguistiques*. Université de Metz : Librairie Klincksieck. HAL : hal-00808797.

SCHNEDECKER C. (2021). *Les chaînes de référence en français*. Collection l'Essentiel français. Paris : Éditions Ophrys.

TUTIN A., TROUILLEUX F., CLOUZOT C., GAUSSIER É., ZAENEN A., RAYOT S. & ANTONIADIS G. (2000). Annotating a large corpus with anaphoric links. In *Third International Conference*

on Discourse Anaphora and Anaphor Resolution (DAARC2000), p.2̃, United Kingdom. HAL : hal-00373327.

WOLFARTH C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal*. Thèse de doctorat, Université Grenoble Alpes. HAL : tel-02517320.

WOLFARTH C., BRISSAUD C. & PONTON C. (2018). Transcrire et normer un corpus scolaire : pour quelles analyses ? In C. BRISSAUD, M. DREYFUS & B. KERVYN, Éds., *Repenser l'écriture et son évaluation au primaire et au secondaire*, volume 36 de collection Diptyque, p. 121–145. Presses universitaires de Namur. HAL : hal-01883221.

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

État de l'art des méthodes de génération automatique de listes de lectures

Julien Aubert-Béduchaud[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
julien.aubert-beduchaud@univ-nantes.fr

RÉSUMÉ

L'augmentation croissante du volume d'articles scientifique rend difficile la montée en compétence des chercheurs sur un domaine de recherche ciblé. Pour faciliter l'accès à ces articles, diverses approches et tâches en recherche d'information ont été développées ces dernières années. Parmi elles, la tâche de génération automatique de listes de lecture a été étudiée dans la littérature. Elle consiste en la génération d'une liste ordonnée d'articles scientifiques couvrant un domaine de recherche spécifique. Plusieurs travaux ont exploré différents aspects de cette tâche, proposant des jeux de données et des méthodologies d'évaluation variées pour apporter des solutions à ce problème. Dans cet article, nous présentons un état de l'art des principales approches de génération de listes de lecture, incluant les données, méthodes, et métriques d'évaluation.

ABSTRACT

State-of-the-art of automatic reading lists generation methods

The growing increase in the volume of scientific articles makes it difficult for researchers to improve their skills in a targeted area of research. To facilitate access to these articles, various information retrieval approaches and tasks have been developed. Among them, the task of automatic reading lists generation has been studied in the literature. It consists of the generation of an ordered list of scientific articles covering a specific area of research. Several works have explored different aspects of this task, proposing various datasets and evaluation methodologies to provide solutions to this problem. In this article, we present a state-of-the-art of the main reading lists generation approaches, including evaluation data, methods, and metrics.

MOTS-CLÉS : Listes de lecture, recherche d'information, recommandation d'articles, génération automatique.

KEYWORDS: Reading lists, information retrieval, article recommandation, automatic generation.

1 Introduction

La fréquence de publication des articles scientifiques est de plus en plus conséquente (Larsen & von Ins, 2010; Thelwall & Sud, 2022). L'impossibilité de lire tous les articles parus ainsi que le manque de moyen pour déterminer quels sont les articles les plus pertinents et utiles à un domaine rendent difficile la montée en compétence et le fait de garder ses connaissances à jour. Une manière d'avoir un accès rapide aux connaissances d'un domaine est de passer par les revues de littérature associées. Tous les domaines ne proposent néanmoins pas d'article de revue ou peuvent ne plus être à jour à mesure des publications du domaine. Des approches se concentrant sur les connaissances et

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 42–56.

compétences attendues peuvent être privilégiées. Pour faciliter l'accès à ces informations, les experts d'un domaine peuvent proposer des listes de lecture (*reading lists* en anglais), listes de références académiques ordonnées permettant de couvrir de façon synthétique les notions importantes d'un domaine de recherche (Siddall & Rose, 2014). Un exemple de liste de lecture est proposé en Figure 1. En plus de permettre la montée en compétence d'un lecteur, ce type de liste permet de se positionner au sein d'un réseau d'articles scientifiques, les références à différents travaux permettant de tracer des liens entre différents aspects de la thématique recherchée.

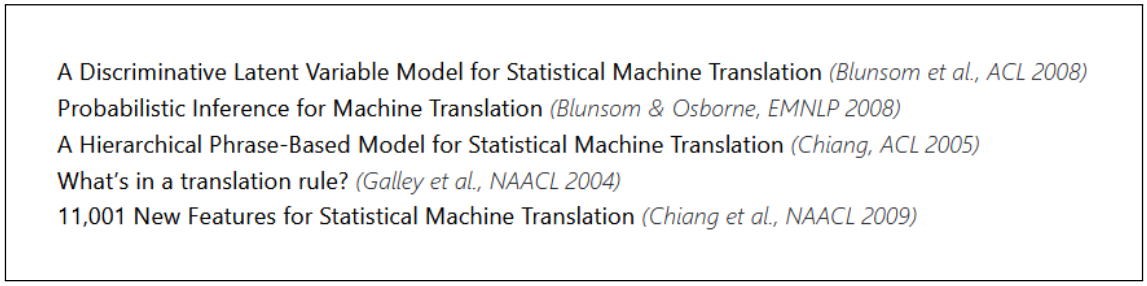


FIGURE 1 – Exemple de liste de lecture pour la thématique "statistical machine-translation models". (Jardine, 2014)

La génération de listes de lecture présentant un intérêt stratégique pour l'exploration d'un domaine scientifique, il est intéressant d'automatiser cette tâche. Soit un corpus d'articles scientifiques $C = \{a_1, a_2, ..., a_N\}$ et une requête R (mots-clés, articles de recherche, listes de notions, etc.), la tâche de génération de liste de lecture consiste à générer une liste ordonnée d'articles L telle que $L \subset C$ où les articles de L sont sélectionnés en fonction de la requête R. Également, la liste de lecture doit permettre d'obtenir une vision large d'un sujet de recherche, le but étant de faire monter rapidement en compétence un chercheur sur un domaine. La liste de lecture peut s'apparenter à un graphe d'articles dont les arcs représentent les ordres de lecture inter-articles. Aussi, l'ordre des références présentes dans les listes de lecture à une importance, l'ordre de lecture pouvant être différent selon le niveau de l'utilisateur.

La tâche de génération automatique de liste de lecture est à la frontière de deux tâches du domaine de la recherche d'informations, mais présente des différences notables :

- *La recherche d'articles scientifiques* a pour objectif de trouver les documents répondant à un besoin utilisateur exprimé sous la forme d'une requête. L'utilisateur précise ainsi directement son besoin. (Kaur *et al.*, 2010; Boudin *et al.*, 2020).
  Si cette approche propose des résultats ordonnés par pertinence, la liste de lecture doit également être ordonnée de façon à faciliter la compréhension d'une thématique.
- *La recommandation d'articles scientifiques* cherche à proposer des articles en fonction des intérêts d'un chercheur. Cette approche utilise une source de données initiale à partir de laquelle les recommandations sont déterminées selon la similarité entre l'utilisateur et les habitudes des autres utilisateurs. (Tang & Zhang, 2009; He *et al.*, 2010; Beel *et al.*, 2016).
  La tâche de recommandation d'articles scientifiques peut proposer des documents similaires et redondants, ce qui n'est pas pertinent pour une liste de lecture.

La génération de liste de lecture est à la fois précise et synthétique (Thompson *et al.*, 2004), cherchant un équilibre entre la pertinence des références et la couverture plus générale de la littérature existante. Une autre particularité de la liste de lecture académique est l'interchangeabilité des articles, les informations importantes d'un domaine pouvant être présentée par différents documents. La génération d'une liste de lecture a ainsi pour objectif de permettre une forme de navigation rapide au sein d'une

collection d'articles d'un domaine.

Cet article présente un état de l'art des travaux sur la génération de listes de lecture pour les collections d'articles scientifiques. La section 2 introduit les principales approches existantes, à commencer par les approches historiques issues du domaine de la recommandation, les modèles d'expertise et les approches utilisant de la modélisation de concepts. La section 3 s'intéresse aux jeux de données proposés pour cette tâche, à savoir les graphes de citations et les listes de référence. La section 4 présente les différentes métriques utilisées lors de l'évaluation des systèmes proposés ainsi qu'un aperçu des méthodes d'évaluation utilisateur mises en place. S'ensuivra une discussion (Section 5) quant aux perspectives de ce domaine de recherche ainsi qu'une conclusion.

2 Approches existantes

Plusieurs façons d'approcher la tâche de génération automatique de listes de lecture ont été explorées dans la littérature. Les approches historiques (Section 2.1) se sont initialement intéressées à la recommandation d'articles fondateurs (c'est-à-dire les articles initiaux ayant contribué de façon significative au domaine) par filtrage collaboratif (Tang, 2008; Ekstrand *et al.*, 2010). Ces méthodes nécessitent l'évaluation d'articles de la part d'autres utilisateurs, qu'elle soit directe (notes utilisateurs) ou indirecte (nombre de citations d'un article).

D'autres approches ont cherché à modéliser l'importance des articles à travers des notions marquant l'expertise d'un domaine (Section 2.2) en s'appuyant sur des informations liées aux articles (Jardine, 2014; Sesagiri Raamkumar *et al.*, 2017; Figueira *et al.*, 2019).

Enfin, une troisième famille d'approches s'intéresse aux dépendances nécessaires à la compréhension de certains sujets (Section 2.3). Ces méthodes s'intéressent notamment aux relations entre les concepts présents dans une collection d'articles dans le but de déterminer des chaînes de prérequis (Gordon *et al.*, 2017). Ces travaux ont conduit Ding *et al.* (2022) à s'intéresser à la construction de chemins de lecture à partir des concepts implicites d'un graphe de citation.

2.1 Approches liées à la recommandation d'articles scientifiques

Les approches historiques liées à la tâche de génération de listes de lecture sont issues du domaine de la recommandation. Plusieurs approches de recommandation traditionnelles ont été adaptées à la tâche, à savoir le filtrage collaboratif proposant d'utiliser les avis d'autres utilisateurs pour une recommandation de documents pertinents, le filtrage basé sur le contenu reposant sur la similarité entre les données des documents, ou encore les filtrages hybrides combinant les prédictions des techniques précédentes. Ces techniques sont utilisées à partir d'évaluation manuelle des articles de la collection (Section 2.1.1) ou par l'utilisation des citations inter-articles (Section 2.1.2).

2.1.1 Recommandation par notes utilisateurs

Tang (2008) a considéré la tâche de génération de liste de lecture comme une extension des tâches de recommandation traditionnelles. La génération de listes se distinguait ici par une recommandation plus large que les préférences du lecteur, les notions nécessaires à la compréhension d'un cours

ayant un impact sur les articles à recommander. Différentes techniques de filtrages ont été mises au point. À partir de notations d'articles de la part d'utilisateur du système, différents algorithmes cherchaient à répondre à des scénarios d'apprentissage (étudiant débutant un cours, se situant à la moitié, à la fin, nombre d'articles ou de notations suffisants ou non). Une approche par filtrage hybride, *PopCon2D*, fut identifiée comme répondant le mieux aux différents scénarios d'avancée du lecteur et de disponibilité de notations. Ce filtrage sélectionnait les articles à partir de la note moyenne, de la similarité entre les thèmes de l'article et ceux du champ d'étude que l'utilisateur avait au préalable déclaré connaître ainsi que les notes des utilisateurs ayant des habitudes semblables à l'utilisateur actif.

2.1.2 Recommandation par citations et co-références

Ekstrand *et al.* (2010) ont constaté que les techniques classiques de filtrage collaboratif utilisées dans le contexte d'articles scientifiques ne permettent que l'exploitation du contexte local d'un article, n'exploitant en effet que les coréférences d'articles et pas les relations ayant lieu au sein d'un domaine.

À partir d'un échantillons de documents d'un domaine, des techniques d'ordonnancement de sommets de graphe permettaient de mesurer l'importance des articles de l'échantillon au sein d'une collection plus large. L'impact individuel et l'influence sur d'autres travaux furent mesurés par *PageRank* (Page *et al.*, 1998), *HITS* (Kleinberg, 1999), *SALSA* (Lempel & Moran, 2001), *HITS with Priors* (White & Smyth, 2003) et *k-step Markov* (White & Smyth, 2003). La mesure d'importance obtenue était couplée à différentes approches de filtrage dérivées des modèles de recommandation. Parmi elles, l'approche privilégiée était un algorithme de filtrage collaboratif utilisant les citation inter-articles pour déterminer le contexte local et *PageRank* pour mesurer l'importance de l'article au sein du domaine.

2.2 Modélisation de l'expertise du domaine

Les techniques de recommandation étaient limitées par la nécessité de connaître les préférences de la communauté de chercheurs dans le processus de filtrage. Ces informations pouvant ne pas être disponibles, des travaux ont cherché à mesurer l'importance des articles en modélisant l'expertise d'un domaine à travers les thématiques sous-jacentes des articles (Section 2.2.1), les mots-clés spécifiés par les auteurs (Section 2.2.2), la probabilité de citation de la part d'un lecteur (Section 2.2.3) ou une agrégation de plusieurs critères (Section 2.2.4).

2.2.1 Couverture de thématiques latentes

Jardine (2014) a cherché à quantifier l'importance des articles selon leur correspondance aux thématiques scientifiques présentes dans une collection d'articles scientifiques. Dans ce but, une modification de l'algorithme PageRank, ThemedPageRank (TPR), était proposée par Jardine & Teufel (2014). Le processus identifiait tout d'abord un sac-de-termes-techniques à partir des titres d'articles, puis une phase de modélisation des thématiques sous-jacentes du corpus était réalisée. La modélisation de thématique pouvait être réalisée par allocation latente de Dirichlet (LDA) (Blei *et al.*, 2003) ou factorisation de matrice non-négative (NMF) (Lee & Seung, 2000). L'étape de classement de sommets de graphes de TPR permettait ensuite de quantifier l'expertises des articles sur chacune des

thématiques identifiées.

2.2.2 Couverture de thématiques explicites

Sesagiri Raamkumar *et al.* (2017) ont mis au point une mesure de couverture thématique et périphérique (*Topical and peripheral coverage*) (TPC). Celle-ci s'appuyait sur les mots-clés spécifiés par l'auteur d'un article dans le but d'identifier un ensemble d'articles possédant des mots-clés similaires et ayant un lien de citation avec l'article initial. Cette métrique cherchait à couvrir un ensemble de critères de contenu présents dans la liste : les articles *populaires ou fondateurs*, articles de *revue de littérature*, articles *récents* et articles *associés à un sous-domaine de recherche*.

La mesure de TPC fut utilisée au sein d'un processus de recherche basée sur les mots-clés spécifiés par l'auteur (*Author-specified Keywords based Retrieval*) (AKR). Le processus identifiait une liste d'articles similaires au sujet de recherche via Okapi BM25 (Robertson *et al.*, 1996) puis triait les résultats en fonction de la somme pondérée du nombre de citations, de références et du score de TPC d'un article. La liste de lecture obtenue était ainsi la liste des articles contribuant de façon importante dans leurs thématiques respectives, en termes de citations ou de références.

2.2.3 Probabilité de citation d'un article

Figueira *et al.* (2019) ont étendu les travaux de Sesagiri Raamkumar *et al.* (2017) en proposant un critère supplémentaire caractérisant l'intérêt des listes de lecture : la présence d'*articles ayant une forte probabilité d'être cité* par l'utilisateur de la liste de lecture.

Des méthodes par apprentissage supervisé furent proposées dans le but de prédire des classements de listes d'articles optimisés sur le critère de citation. Des méthodes *Learning to Rank* (Liu, 2009) permettant d'apprendre une fonction de classement à partir de données existantes ont ainsi été mises en place. Plusieurs méthodes furent exploitées pour cette approche, à savoir MART (*Multiple Additive Regression Trees*) (Friedman, 2001), *LambdaMART* (Burges, 2010), *ListNet* (Cao *et al.*, 2007), *RankNet* (Burges *et al.*, 2005), *AdaRank* (Xu & Li, 2007), *RankBoost* (Freund *et al.*, 2003), des forêts d'arbres décisionnels (Ho, 1995) et algorithmes d'ascension de coordonnées (Fessler *et al.*, 1997). Les modèles étaient entraînés avec des caractéristiques (*features*) tels que le score de TPC (Sesagiri Raamkumar *et al.*, 2017), l'âge des articles, le fait qu'un article soit un article de revue de littérature ou non, le nombre de références des articles associés aux mots-clés cherchés, des mesures de pondération de documents telles que BM25 ainsi qu'une mesure de similarité cosinus.

2.2.4 Agrégation de multiples critères

Figueira *et al.* (2019) ont également proposé des méthodes d'agrégation de classement pour la génération de listes de lecture, permettant de classer des listes sur la base des critères définis dans les sections 2.2.2 et 2.2.3.

Une phase de génération d'une liste d'articles en lien avec la requête était effectuée via Okapi BM25. La phase de classement était ensuite réalisée au moyen de différentes techniques d'agrégation de rangs : méthode Borda (de Borda, 1781), des techniques de comptage binaire ou classement par maximisation du gain cumulé. Les prédictions des méthodes Learning to Rank (Section 2.2.4) furent

utilisées dans certaines variantes afin d'estimer la probabilité d'un article d'être cité par le lecteur. Les scores de classement produits par LambdaMART furent utilisés pour estimer cette probabilité.

2.3 Utilisation de chaînes de prérequis

Les listes de lectures générées par les approches précédentes proposent un ordre de lecture dépendant de son impact sur une communauté d'utilisateurs (Section 2.1) ou de la correspondance à un critère spécifique (Section 2.2). Des travaux plus récents se sont intéressés à la construction d'un ordre de lecture s'appuyant sur les *concepts prérequis* pour la compréhension de ressources pédagogiques (Gordon *et al.*, 2016; Liu *et al.*, 2016; Pan *et al.*, 2017). Des approches se sont concentrées plus spécifiquement sur le domaine des articles scientifiques, cherchant à expliciter les concepts latents (Section 2.3.1) ou les induire par co-citations (Section 2.3.2).

2.3.1 Explicitation de concepts latents

Gordon *et al.* (2017) se sont intéressés à l'utilisation des *concepts*, notions nécessaire à la compréhension d'une notion pédagogique, pour la génération de listes de lecture. Cette approche proposait d'identifier les concepts présents dans les articles à partir des thématiques extraites d'un corpus par LDA à l'instar de Jardine & Teufel (2014). Des relations inter-concepts étaient ensuite calculées par similarité de Jaccard et méthode d'entropie croisée, l'association d'articles à des concepts permettant de déterminer des dépendances au sein d'une collection de documents scientifiques. La liste de lecture était ensuite générée en fonction de l'intérêt mesuré par chevauchement de lexique entre concepts et requête utilisateur. Les articles associés à un concept étaient identifiés par un parcours en profondeur du graphe. Cette approche calculait également un ordre de lecture via le niveau de familiarité de l'utilisateur avec les concepts des documents en fonction des réponses à un questionnaire ou des documents lus par l'utilisateur.

2.3.2 Identification de concepts par co-citations

Ding *et al.* (2022) se sont intéressés à la façon de trouver un chemin de lecture au sein d'un graphe de citations, à savoir un ordre de lecture spécifique dépendant des concepts pré-requis à la compréhension d'une thématique à partir des relations de co-citation des articles. Cette approche, proche de la génération de liste de lecture, s'intéressait à la génération d'une liste de références pertinente pour une requête contenant également les articles pré-requis aux références identifiées.

L'approche de Ding *et al.* (2022) proposait l'ordonnancement d'une liste d'articles scientifiques par relation de dépendance inter-articles, à partir d'une requête constituée de mots-clés. Une liste initiale d'articles était générée par Google Scholar [1] à partir des mots-clés d'une requête. Cette liste initiale, couplée à un graphe de citations de six millions d'articles du domaine de l'informatique, permettait de construire un sous-graphe de citations incluant les articles récupérés ainsi que leurs voisins proches. Les concepts prérequis étaient capturés implicitement à partir des informations de co-citation des articles de la liste initiale présents dans le graphe de citation permettant de détecter de nouveaux articles nécessaires à la compréhension. L'algorithme *Node-Edge Weighted Steiner Tree* (NEWST) (Segev, 1987) permettait de détecter un arbre couvrant minimal du sous-graphe pour

1. http://scholar.google.com

générer un chemin de lecture optimal. Le sens de lecture de la liste était ensuite déduit à partir des coréférences et dates de publication d'articles.

3 Jeux de données

Différents jeux de données ont été construits et utilisés pour les besoins de la tâche de génération de liste de lecture. Deux types d'ensembles se distinguent dans les travaux réalisés. Des collections d'articles scientifiques (Table 1) couvrent les articles d'un ou plusieurs domaines scientifiques et peuvent être utilisés pour la conception de graphes de citations, l'apprentissage, ou l'évaluation des modèles. Des corpus de référence (Table 2) permettent, quant à eux, l'évaluation des listes générées.

3.1 Collections d'articles scientifiques

Article	Source	Domaine	Plage	Nb Documents	Langue	Citations	Disponible
Tang (2008)	IEEE et ACM	Ingénierie informatique	1992-2004	21	en	✗	✓
Ekstrand *et al.* (2010)	ACM DL	Informatique	N.C.	201 145	en	✓	✗
Jardine (2014)	ANN	Traitement automatique des langues (TAL)	1960s - 2010	15 388	en	✓	✗
Gordon *et al.* (2017)	ANN, Wikipédia, ScienceDirect, Tutoriel Web	TAL	1960s - 2014	25 319	en	✓	✗
Sesagiri Raamkumar *et al.* (2017)	ACM DL	Informatique	1950-2011	103 739	en	✓	✗
Figueira *et al.* (2019)	Scopus	Informatique, Ingénierie	1970-2018	58 734	en	✓	✗
Ding *et al.* (2022)	S2ORC	Informatique	1913-2020	6 000 000	en	✓	✗

TABLE 1 – Résumé des collections d'articles scientifiques utilisées.

Les collections d'articles scientifiques présentés dans la littérature peuvent être des ensembles d'articles ou des réseaux de citations. Différentes sources ont été utilisées, telles l'*ACM Digital Library* (Ekstrand *et al.*, 2010; Sesagiri Raamkumar *et al.*, 2017), l'*ACL Anthology Network* (AAN) (Jardine, 2014; Gordon *et al.*, 2017), des journaux IEEE et ACM (Tang, 2008), Scopus (Figueira *et al.*, 2019) ou le corpus de *Semantic Scholar* (S2ORC) (Ding *et al.*, 2022). Si les collections utilisés sont de taille variable selon les approches, allant de la dizaine d'articles à 6 millions de documents, les tailles sont généralement de l'ordre de la dizaine à la centaine de milliers d'articles. Malgré différentes sources de données, les documents portent tous sur le domaine de l'informatique et présentent des articles scientifiques en langue anglaise seulement. Parmi les collections utilisées dans la littérature, seul celle de Tang (2008) est disponible, en annexe de son manuscrit de thèse. Les autres articles mentionnés proposent des indications permettant de construire des collections semblables à celles utilisées, mais ne sont pas reproductibles à l'identique

3.2 Corpus de référence

Les corpus de références sont utilisés à des fins d'évaluation de la tâche de génération automatique de listes de lecture. Ekstrand *et al.* (2010) ont utilisé un jeu de données de référence constitué à partir de 220 articles de l'*ACM Computing Survey* présentant au moins 15 références. Les données utilisées ne sont néanmoins pas disponibles. Jardine (2014) a proposé un corpus ayant été construit par des experts en TAL à partir de l'AAN. Huit listes de lecture y sont proposées, couvrant différentes

Article	Source	Domaine	Plage	Nb Documents	Nombre de références	Langue	Disponible
Ekstrand *et al.* (2010)	ACM Computing Surveys	Informatique	N.C.	220	15+	en	✗
Jardine (2014)	ANN (Annotations expertes)	TAL	1991-2009	8	∼12	en	✓
Figueira *et al.* (2019)	Scopus (Articles classifiés *Reviews* et *Short Surveys*)	Informatique, Ingénieurie	1970-2018	1648	N.C	en	✗
SurveyBank (Ding *et al.*, 2022)	S2ORC, Google Scholar	TAL, Machine Learning, Intelligence Artificielle, Apprentissage profond, Recherche d'informations	1913-2020	9 321	∼58	en	✓

TABLE 2 – Résumé des corpus utilisés comme référence aux listes de lecture.

thématiques de recherche. Ce jeu de données était utilisé par Gordon *et al.* (2017) pour une évaluation utilisateur. Parmi la collection de Figueira *et al.* (2019), 1648 articles sont considérés comme articles de revue, permettant d'utiliser leur listes de citations comme ensemble de références. Ding *et al.* (2022) ont proposé une collection d'articles de revue, SurveyBank [2]. Le jeu de données est constituée à partir de S2ORC et Google Scholar en utilisant les mots-clés du domaine de l'informatique issus de TutorialBank (Fabbri *et al.*, 2018) et LectureBank (Li *et al.*, 2019). Cette collection est initialement pensée pour la tâche de génération de chemin de lecture, mais peut être utilisée comme jeu de références pour la tâche de génération de listes de lecture.

4 Méthodes d'évaluation

4.1 Comparaison de la liste générée à une référence

Différentes métriques d'évaluation de la capacité des modèles à générer une liste de lecture à partir d'une collection ont été proposées dans la littérature. Ces métriques cherchaient à comparer une liste générée à une liste de référence sur la base de différents critères. Un résumé des méthodes de comparaison de la liste générée à une référence est proposé dans la Table 3.

Ekstrand *et al.* (2010) ont cherché à mesurer parmi les algorithmes mis en place ceux présentant l'ordre le plus utile pour les utilisateurs, utilisant des données issues d'articles de *survey* comme référence. La métrique d'utilité de la demi-vie ***(half-life utility metric)*** (Breese *et al.*, 1998) permettait d'estimer la probabilité qu'un utilisateur ait besoin d'un article d'une liste ordonnée. L'utilité attendue de la liste de lecture générée se mesurait ainsi en tenant compte de la position des articles générés présents dans la liste de référence. L'utilité était pondérée en fonction du rang de l'article : plus l'article est généré tard dans la liste, moins son score est élevé. Cette métrique possédait le défaut de ne pas permettre la mesure des articles pouvant se substituer à d'autres, l'utilité étant définie par la présence ou non d'un article dans la liste générée.

Jardine (2014) a évalué la tâche sur un ensemble de méthodes historiques, et des métriques visant à combler leurs lacunes. Les métriques historiques de **précision, rappel et F1-score** (Rijsbergen, 1979) ainsi que la moyenne des précisions (***Mean Average Precision***) **(MAP)** ont été employées, mais ne permettaient pas de prendre en compte les documents interchangeables. Ding *et al.* (2022) ont également utilisé des mesures de **précision@K et F1-score@K** (Manning *et al.*, 2008) dans le but de comparer les listes de lecture générées par leur modèle NEWST aux résultats de plateformes de recherche d'articles en utilisant les références de leur jeu de données SurveyBank.

2. https://github.com/JiayuanDing100/Reading-Path-Generation

Métriques	Intuition	Avantages	Inconvénients
Précision, Rappel, F1score	Mesures de la pertinence des articles	✓ Métriques historiques	✗ Correspondance partielle non considérée ✗ Ne prends pas en compte l'ordre des documents
Half-life utility metric	Estimation de l'utilité des articles d'une liste pour un utilisateur	✓ Prise en compte de l'ordre des documents	✗ Articles substituables non considérés
MAP	Précision moyenne des éléments d'une liste pour un ensemble de requêtes	✓ Prise en compte de l'ordre des documents	✗ Correspondance partielle non considérée
RCP	Rapport entre le nombre d'articles citant à la fois les articles générés et les articles de référence et le nombre d'articles citant les articles de référence	✓ Prise en compte des correspondances partielles	✗ Impossibilité de mesurer les articles non-cités ✗ Pertinence partielle trop importante dans la mesure du score
CSC	Mesurer la distance de citation entre articles d'une liste générés et celle de référence	✓ Prise en compte des correspondances partielles ✓ Prise en compte des articles non-cités	✗ La co-citation seule ne permet pas de détecter la similarité

TABLE 3 – Résumé de métriques de comparaison de la liste générée à une référence.

Autre méthode explorée par Jardine (2014), la métrique de probabilité relative de co-citation ***(Relative co-cited probability)*** **(RCP)** (He *et al.*, 2011) permettait de déterminer des articles alternatifs à un article de référence cité par un auteur en mesurant le nombre de fois où deux articles (référence et générés) ont été co-cités dans un corpus par rapport au nombre de citations de la référence seule. Si cette métrique était utilisée par Jardine (2014) à titre d'information, un désavantage de cette méthode était l'impossibilité de mesurer la probabilité des articles n'ayant pas de citations, bien que le contenu puisse être similaire. Également, les articles référencés par plusieurs articles de la collection avaient un impact plus important dans l'équation, étant parfois plusieurs fois pris en compte.

Le coefficient de substitution de citation ***(Citation Substitution Coefficient)*** **(CSC)** était une métrique proposée par Jardine (2014) estimant la pertinence partielle entre deux articles d'un graphe de citations. Le nombre minimal de nœuds nécessaire pour relier deux articles au sein d'un graphe de citations permettait d'évaluer la distance de citations entre une liste d'articles de référence et une liste proposée par un système de génération de listes de lecture. Par rapport à la mesure de RPC, le CSC considérait les articles peu cités ayant des références aux articles pertinents pour le domaine. Une limite de cette approche résidait dans le fait que la co-citation d'articles ne marque pas systématiquement la similarité ou l'interchangeabilité.

4.2 Évaluation de critères

Un autre pan de recherche dans le domaine cherchait à évaluer la capacité des systèmes de génération de listes de lecture à répondre à un ensemble de critères définis dans le but de déterminer une liste pertinente. La liste des critères concernés est couverte par les sections 2.2.2 et 2.2.4.

La **méthode de l'entropie croisée** était proposée par Sesagiri Raamkumar *et al.* (2017) afin d'évaluer le système AKR. L'approche utilisée impliquait l'agrégation de classements calculés sur la base des prérequis proposés par l'article, la méthode satisfaisant le plus de prérequis étant celle sélectionnée.

Deux distances furent utilisées pour cette méthode, le ρ (rho) de Spearman (Spearman, 1904) mesurant la distance de déplacements entre deux éléments permutés et le τ (tau) de Kendall (Kendall, 1938) mesurant le nombre d'inversions dans un classement (Kumar & Vassilvitskii, 2010).

Figueira *et al.* (2019) ont utilisé la métrique de **précision@K**, afin d'évaluer le pourcentage d'articles pertinents dans un ensemble de candidats par rapport aux prérequis proposés par l'auteur. La mesure était utilisée pour évaluer la popularité, la présence d'articles de revue de littérature, la présence d'articles récents et la probabilité de citation. La diversité des listes de lecture était mesuré au moyen d'une métrique d'α**-NDCG@20** (Clarke *et al.*, 2008) (Vargas, 2014).

4.3 Évaluation humaine par les utilisateurs

Article	Mode d'évaluation	Critère évalué	Nb Participants	Conclusion de l'étude
Tang (2008)	Questionnaire à échelle d'évaluation	Évaluation de la qualité des articles proposés	25	Les approches utilisant un filtrage hybride à partir de la notre moyenne, de la similarité entre les thèmes de l'article et ceux du champ d'étude de l'utilisateur, ainsi que les notes des utilisateurs ayant des habitudes semblables à l'utilisateur actif permettent de bons résultats sur les scénarios de recommandation.
Ekstrand et al. *(2010)*	Questionnaire à échelle d'évaluation	Évaluation de la qualité des articles proposés	19	Les utilisateurs trouvent le système développé prometteur.
Sesagiri Raamkumar et al. *(2017)*	Questionnaire à échelle d'évaluation	Évaluation de la pertinence de la liste en fonction de critères spécifiques	132	La présence d'articles fondateurs à un impact direct sur la satisfaction des participants Les étudiants ont plus d'intérêt pour le système que les chercheurs confirmés
Gordon et al. *(2017)*	Questionnaire à échelle d'évaluation	Évaluation du potentiel pédagogique de la liste générée	33	Les listes générées présentent un potentiel pédagogique allant de modéré à élevé selon les domaines.
Gordon et al. *(2017)*	Modifications de listes générées et de références	Calcul de la distance de Levensthein entre les listes modifiées et les listes proposées	33	Les listes générées sont comparables aux listes expertes.
Ding et al. *(2022)*	Questionnaire à choix multiples	Évaluation de la pertinence, exhaustivité et de la présence de dépendances dans les listes générées	16	L'utilisation de chaînes de prérequis augmente la connaissance apportée par les listes générées

TABLE 4 – Résumé des méthodes d'évaluation utilisateur.

Il est difficile de tirer des conclusions sur l'utilisation des systèmes par un humain uniquement par des méthodes d'évaluation automatisées. Plusieurs types d'évaluations utilisateur ont ainsi pu être mis en place dans les travaux réalisés (Table 4).

Gordon *et al.* (2017) ont évalué les performances du système sur une expérience intégrant des utilisateurs. L'expérimentation implique 33 chercheurs en TAL chargés de modifier des listes générées ainsi que de référence afin qu'elles atteignent un niveau convenable selon eux. L'expérience cherche à comparer la liste originale et la correction au moyen de la distance de Levensthein (Levenshtein, 1966) afin de comparer le niveau de modification entre les deux. Les résultats obtenus indiquent que les listes générées sont comparables aux listes expertes

Des questionnaire à échelle d'évaluation (Likert, 1932) ont été proposés par plusieurs travaux. Tang (2008) a demandé à des 25 étudiants de noter les articles d'une liste afin de mesurer la pertinence des algorithmes de filtrage dans différents scénarios (nombre d'articles, avancement dans le cours). Ekstrand *et al.* (2010) ont posé des questions à 19 novices sur leur familiarité avec les articles de la liste générée, la pertinence des articles sélectionnés pour une thématique et l'importance des thématiques traitées. Les résultats indiquent que les utilisateurs trouvent le système prometteur. (Gordon *et al.*, 2017) ont évalué leur système sur la pertinence perçue de 33 chercheurs en TAL, indiquant un potentiel pédagogique modéré ou élevé selon les listes générées. Sesagiri Raamkumar

et al. (2017) ont évalué des listes générées en fonction des domaines de recherche de 132 participants (62 étudiants, 70 chercheurs), sur la pertinence de la liste en fonction de différents critères (familiarité, utilité, etc.). Les résultats indiquent que la présence d'articles fondateurs à un impact direct sur la satisfaction et que les étudiants ont plus d'intérêt pour le système que les chercheurs confirmés.

Ding *et al.* (2022) ont utilisé un questionnaire à choix multiple pour comparer le système NEWST à Google Scholar. 16 diplômés évaluent la pertinence des listes générées, leur exhaustivité et si les articles prérequis sont identifiables. Les résultats indiquent que les chaînes de prérequis augmentent la connaissance apportée par les listes générées.

5 Discussion

Les différentes approches de la littérature ont proposé des interprétations différentes des besoins de la tâche de génération de listes de lecture. Les approches historiques ont ainsi développé principalement des méthodes de recommandation améliorées par des techniques d'ordonnancement de sommets de graphe. Des approches plus modernes tentent de modéliser l'expertise du domaine pour l'ordonnancement des articles, d'autres explorent la modélisation de dépendances d'articles à des concepts. Ces différentes approches s'intéressent peu au contenu et informations sous-jacentes des articles scientifiques, des méthodes plus modernes de traitement automatique des langues pourrait permettre l'émergence de nouvelles méthodes de génération de listes de lecture. Certaines facettes de la tâche restent également peu explorées, notamment l'ordre de lecture des articles ou la minimisation de la redondance entre les articles retournés. L'introduction de mesure de sérendipité sur cette tâche (mesurant dans ce cas le fait de découvrir de façon fortuite un article pertinent) pourrait également être exploré afin de permettre la construction de listes favorisant les démarches de recherche scientifique.

Parmi les collections présentées, la majorité des corpus ne sont pas mis à disposition ou ne proposent pas un nombre d'échantillons suffisamment conséquent pour entraîner un modèle. Ce manque de données explique la quantité limitée d'approches supervisées et l'absence de méthodes neuronales dans le traitement de la tâche de génération de listes de lecture. Les avancées liées à la similarité sémantique pourraient notamment être exploitées pour la génération de listes de lecture. Les représentations vectorielles du contenu pourraient entre autres permettre de détecter plus efficacement les articles substituables d'une collection.

L'évaluation des systèmes de génération de listes de lecture a été réalisé de différentes façons dans la littérature. Néanmoins, le manque de méthodes d'évaluation couvrant plusieurs approches et les données non comparables ne permettent pas de comparer de façon exhaustive les travaux associés à cette tâche. L'évaluation de listes de lectures générées automatiquement reste donc un défi auxquels les futurs travaux devront s'intéresser.

Remerciements

Je tiens à remercier mon équipe encadrante, Florian Boudin, Richard Dufour et Béatrice Daille, pour leurs conseils et leur soutien. Merci également aux membres de l'équipe TALN du LS2N et aux relecteurs RECITAL pour leurs commentaires et suggestions. Ce travail est financé dans cadre du projet AID-CNRS NaviTerm (convention 2022 65 0079 CNRS Occitanie Ouest).

Références

BEEL J., GIPP B., LANGER S. & BREITINGER C. (2016). Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries*, **17**, 305–338. DOI : 10.1007/s00799-015-0156-0.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

BOUDIN F., GALLINA Y. & AIZAWA A. (2020). Keyphrase generation for scientific document retrieval. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1118–1126, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.105.

BREESE J. S., HECKERMAN D. & KADIE C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, p. 43–52, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

BURGES C., SHAKED T., RENSHAW E., LAZIER A., DEEDS M., HAMILTON N. & HULLENDER G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, p. 89–96, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1102351.1102363.

BURGES C. J. (2010). *From RankNet to LambdaRank to LambdaMART : An Overview*. Rapport interne MSR-TR-2010-82, Microsoft.

CAO Z., QIN T., LIU T.-Y., TSAI M.-F. & LI H. (2007). Learning to rank : from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, p. 129–136, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1273496.1273513.

CLARKE C. L., KOLLA M., CORMACK G. V., VECHTOMOVA O., ASHKAN A., BÜTTCHER S. & MACKINNON I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, p. 659–666, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1390334.1390446.

DE BORDA J. C. (1781). *Mémoire sur les élections au scrutin*, In *Histoire de l'Académie royale des sciences*, p. 657–665. Imprimerie royale de Paris.

DING J., XIANG T., OU Z., ZUO W., ZHAO R., LIN C., ZHENG Y. & LIU B. (2022). Tell me how to survey : Literature review made simple with automatic reading path generation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, p. 3426–3438. DOI : 10.1109/ICDE53745.2022.00322.

EKSTRAND M. D., KANNAN P., STEMPER J. A., BUTLER J. T., KONSTAN J. A. & RIEDL J. T. (2010). Automatically building research reading lists. In *Proceedings of the fourth ACM conference on Recommender systems*, p. 159–166, Barcelona Spain : ACM. DOI : 10.1145/1864708.1864740.

FABBRI A., LI I., TRAIRATVORAKUL P., HE Y., TING W., TUNG R., WESTERFIELD C. & RADEV D. (2018). TutorialBank : A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In I. GUREVYCH & Y. MIYAO, Éds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 611–620, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1057.

FESSLER J., FICARO E., CLINTHORNE N. & LANGE K. (1997). Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. *IEEE Transactions on Medical Imaging*, **16**(2), 166–175. DOI : 10.1109/42.563662.

FIGUEIRA P., BELEM F., ALMEIDA J. M. & GONCALVES M. A. (2019). Automatic Generation of Initial Reading Lists : Requirements and Solutions. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 1–10, Champaign, IL, USA : IEEE. DOI : 10.1109/JCDL.2019.00011.

FREUND Y., IYER R., SCHAPIRE R. E. & SINGER Y. (2003). An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, **4**, 933–969.

FRIEDMAN J. H. (2001). Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, **29**(5), 1189 – 1232. DOI : 10.1214/aos/1013203451.

GORDON J., AGUILAR S., SHENG E. & BURNS G. (2017). Structured Generation of Technical Reading Lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 261–270, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-5029.

GORDON J., ZHU L., GALSTYAN A., NATARAJAN P. & BURNS G. (2016). Modeling Concept Dependencies in a Scientific Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 866–875, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1082.

HE Q., KIFER D., PEI J., MITRA P. & GILES C. L. (2011). Citation recommendation without author supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, p. 755–764, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1935826.1935926.

HE Q., PEI J., KIFER D., MITRA P. & GILES L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 421–430, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1772690.1772734.

HO T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, p. 278–282 vol.1. DOI : 10.1109/ICDAR.1995.598994.

JARDINE J. & TEUFEL S. (2014). Topical PageRank : A model of scientific expertise for bibliographic search. In S. WINTNER, S. GOLDWATER & S. RIEZLER, Éds., *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 501–510, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.3115/v1/E14-1053.

JARDINE J. G. (2014). *Automatically generating reading lists*. Rapport interne UCAM-CL-TR-848, University of Cambridge, Computer Laboratory. DOI : 10.48456/tr-848.

KAUR J., YUSOF M., BOURSIER P. & OGIER J.-M. (2010). Automated scientific document retrieval. In *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, volume 5, p. 732–736. DOI : 10.1109/ICCAE.2010.5451344.

KENDALL M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1-2), 81–93. DOI : 10.1093/biomet/30.1-2.81.

KLEINBERG J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, **46**(5), 604–632. DOI : 10.1145/324133.324140.

KUMAR R. & VASSILVITSKII S. (2010). Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 571–580, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1772690.1772749.

LARSEN P. O. & VON INS M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, **84**(3), 575–603. DOI : 10.1007/s11192-010-0202-z.

LEE D. D. & SEUNG H. S. (2000). Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, p. 535–541, Cambridge, MA, USA : MIT Press.

LEMPEL R. & MORAN S. (2001). Salsa : the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, **19**(2), 131–160. DOI : 10.1145/382979.383041.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**, 707–710.

LI I., FABBRI A. R., TUNG R. R. & RADEV D. R. (2019). What should i learn first : Introducing lecturebank for nlp education and prerequisite chain learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 6674–6681. DOI : 10.1609/aaai.v33i01.33016674.

LIKERT R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, **22 140**, 55.

LIU H., MA W., YANG Y. & CARBONELL J. (2016). Learning concept graphs from online educational data. *J. Artif. Int. Res.*, **55**(1), 1059–1090.

LIU T.-Y. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, **3**(3), 225–331. DOI : 10.1561/1500000016.

MANNING C. D., RAGHAVAN P. & SCHÜTZE H. (2008). *Introduction to information retrieval*. Cambridge University Press.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1998). *The PageRank Citation Ranking : Bringing Order to the Web*. Rapport interne, Stanford Digital Library Technologies Project.

PAN L., LI C., LI J. & TANG J. (2017). Prerequisite relation learning for concepts in MOOCs. In R. BARZILAY & M.-Y. KAN, Éds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1447–1456, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1133.

RIJSBERGEN C. J. V. (1979). *Information Retrieval*. USA : Butterworth-Heinemann, 2nd édition.

ROBERTSON S., WALKER S., HANCOCK-BEAULIEU M. M., GATFORD M. & PAYNE A. (1996). Okapi at trec-4. In *The Fourth Text REtrieval Conference (TREC-4)*, p. 73–96 : Gaithersburg, MD : NIST.

SEGEV A. (1987). The node-weighted steiner tree problem. *Networks*, **17**(1), 1–17. DOI : https ://doi.org/10.1002/net.3230170102.

SESAGIRI RAAMKUMAR A., FOO S. & PANG N. (2017). Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing & Management*, **53**(3), 577–594. DOI : 10.1016/j.ipm.2016.12.006.

SIDDALL G. & ROSE H. (2014). Reading lists–time for a reality check ? an investigation into the use of reading lists as a pedagogical tool to support the development of information skills amongst foundation degree students. *Library and Information Research*, **38**(118), 52–73. DOI : 10.29173/lirg605.

SPEARMAN C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**(1), 72–101. DOI : 10.2307/1412159.

TANG J. & ZHANG J. (2009). A discriminative approach to topic-based citation recommendation. In T. THEERAMUNKONG, B. KIJSIRIKUL, N. CERCONE & T.-B. HO, Éds., *Advances in Knowledge Discovery and Data Mining*, p. 572–579, Berlin, Heidelberg : Springer Berlin Heidelberg. DOI : 10.1007/978-3-642-01307-2_55.

TANG Y. (2008). *The design and study of pedagogical paper recommendation*. Thèse de doctorat, University of Saskatchewan.

THELWALL M. & SUD P. (2022). Scopus 1900–2020 : Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies*, **3**(1), 37–50. DOI : 10.1162/qss_a_00177.

THOMPSON L., MAHON C. & THOMAS L. (2004). *Reading lists - how do you eat yours ?*, In *Learning and Teaching Projects 2003-2004*, p. 57–62. University of Wolverhampton.

VARGAS S. (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, p. 1281, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2600428.2610382.

WHITE S. & SMYTH P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, p. 266–275, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/956750.956782.

XU J. & LI H. (2007). Adarank : a boosting algorithm for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, p. 391–398, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1277741.1277809.

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

Évaluation de mesures d'accord sur des structures relationnelles par la dégradation contrôlée d'annotations

Antoine Boiteau[1]
(1) Normandie Université, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, FRANCE
antoine.boiteau@unicaen.fr

RÉSUMÉ

Les mesures d'accord inter-annotateurs sont essentielles pour évaluer la qualité des annotations humaines sur les corpus. Dans le cadre des structures relationnelles, la question de la qualité et de l'interprétabilité de ces mesures reste cependant ouverte. Cet article présente l'adaptation d'un outil déjà utilisé pour d'autres paradigmes d'annotation dont le but est de générer de manière contrôlée des annotations artificielles erronées. Les annotations obtenues sont fournies à des mesures d'accord adaptées aux structures relationnelles, permettant l'identification des comportements des mesures ainsi que les différences entre elles.

ABSTRACT

Controlled degradations on annotations for relational structures : utility and method

Inter-coder agreement measures are essential for assessing the quality of human annotations on corpora. In the context of relational structures, however, the question of the quality and interpretability of these measures remains open. This article presents the adaptation of a tool already used for other annotation paradigms. Its aim is to generate erroneous artificial annotations in a controlled manner. The resulting annotations are provided to agreement measures adapted to relational structures, allowing the identification of the behaviour of the measures as well as the differences between them.

MOTS-CLÉS : annotation de structures relationnelles, mesures d'accord inter-annotateurs, analyse de l'argumentation, évaluation.

KEYWORDS: annotation of relational structures, inter-coder agreement, argumentation analysis, evaluation.

1 Introduction

Les corpus textuels sont des jeux de données précieux pour la recherche scientifique. Déjà indispensables pour toute démarche expérimentale dans le domaine du TAL, leur intérêt a été encore grandissant ces dernières décennies avec les approches d'apprentissage machine qui ont révolutionné nos disciplines. Pour que ces jeux de données soient utilisables pour l'entraînement des systèmes d'apprentissage supervisés et semi-supervisés, et comme référence pour évaluer les performances de ces systèmes (Fort, 2016), il est nécessaire qu'ils fassent l'objet d'une étape supplémentaire d'annotation. Le procédé d'annotation consiste en l'identification et la mise en avant du phénomène recherché par la communauté. Sur ces corpus plusieurs types d'annotation, éventuellement complémentaires, sont possibles. Selon (Leech *et al.*, 1997), l'annotation peut être vue comme une information interprétative rajoutée à un jeu de données brutes, ou déjà précédemment annotées. Une annotation de référence est

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 57–68.

un jeu de données enrichi de connaissances que l'on juge fiables (ce jeu de données constitue alors une référence ou *gold standard*). La littérature présente de nombreux exemples de campagnes d'annotation sur des corpus textuels ou oraux retranscrits afin de mettre en lumière les phénomènes linguistiques. Si l'annotation de référence peut parfois s'obtenir au moyen d'un unique expert, dans de nombreux cas, notamment pour les tâches inédites ou difficiles, la méthode classique consiste à procéder à une annotation multiple du jeu de données. L'hypothèse sous-jacente à cette annotation multiple est que plus l'accord entre les annotateurs est important et plus l'annotation se rapproche de la vérité. Pour l'annotation multiple, différents annotateurs produisent chacun leur propre annotation, indépendante des autres. Les annotateurs peuvent lors de cette étape avoir des productions différentes. Une mesure de l'accord inter-annotateurs adaptée à la tâche d'annotation permet alors d'évaluer le degré d'accord de la multi-annotation. La qualité et le sens des mesures d'accord est une problématique étudiée depuis plusieurs décennies par la communauté. Ce questionnement s'illustre par la diversité des mesures d'accord inter-annotateurs, nombreuses pour la tâche de catégorisation (où l'annotateur doit typer des unités déjà identifiées), plus rares pour la tâche d'identification d'unités (où l'annotateur doit de surcroît identifier et positionner les unités dans le *continuum* textuel), tâche souvent dite d'*unitizing* (Krippendorff, 2019). Bien que l'état de l'art soit riche pour ces deux types de tâches d'annotation, les travaux sur l'accord pour d'autres tâches d'annotation telles que les structures relationnelles sont plus rares. C'est dans ce contexte que nos présents travaux se positionnent. Nous préciserons tout d'abord le cadre de l'annotation des structures relationnelles dans les corpus textuels. Nous reprendrons à notre compte le principe du Corpus Shuffling Tool (Mathet *et al.*, 2012), dédié initialement à l'identification d'unités, en l'appliquant à ce type de structures. Cet outil permet en effet d'évaluer et de mieux comprendre les comportements des différentes mesures et de pouvoir les comparer entre elles. Enfin, nous présenterons nos résultats expérimentaux suite à la mise à l'épreuve de mesures d'accord adaptées aux structures relationnelles grâce à notre outil de dégradation contrôlée.

2 Cadre des structures relationnelles

Nous nous intéressons ici à des structures relationnelles, c'est-à-dire des configurations au sein desquelles deux ou plusieurs unités d'intérêt sont mises en relation. Le TAL et la linguistique offrent de nombreux exemples de telles structures, parmi lesquelles on peut citer par exemple les structures syntaxiques, les structures rhétoriques ou encore les structures argumentatives.

2.1 Exemple de structure relationnelle

Par la suite, nous illustrerons nos propos en nous appuyant sur un modèle de structuration qui nous permettra de mettre en lumière certaines difficultés posées par la comparaison de structures relationnelles : les structures argumentatives telles que décrites par (Putra *et al.*, 2022). Cette modélisation est adaptée aux besoins du corpus ICNALE (Ishikawa, 2013, 2018). Pour l'annotation de cette structure, les unités de bases sont les phrases du corpus. Ce choix de granularité est déjà présent dans d'autres études de l'argumentation (Teufel *et al.*, 1999; Wachsmuth *et al.*, 2016). L'annotation de la structure se fait alors en deux étapes. Vient en premier lieu l'*identification des composants de l'argumentation*. Il s'agit ici de catégoriser chaque unité comme étant Composant de l'Argumentation (CA) ou détachée de la structure argumentative (non-CA). La seconde étape est celle de l'*identification de la structure argumentative* qui consiste à identifier et positionner les relations existant entre les CA. Dans ce

cas d'étude, les relations sont toutes étiquetées par l'un des trois types de relation binaire et dirigée proposés : *soutien* (*support*), *attaque* (*attack*) et *détail* (*detail*). En plus de ces trois types fréquemment rencontrés dans la littérature, sous ces appellations ou d'autres (Kirschner *et al.*, 2015), les auteurs ajoutent la relation de type *reformulation* (*restatement*) qui est bidirectionnelle et signale qu'un CA vient reprendre l'argument d'un autre CA pour le réitérer et l'évoquer stratégiquement à un moment différent du corpus (Skeppstedt *et al.*, 2018).

Dernière caractéristique de la structure de (Putra *et al.*, 2022), chaque unité CA a toujours une et unique relation sortante mais peut-être la cible de plusieurs relations entrantes, sauf pour le cas particulier des CA de type *proposition principale* (*main claim*) qui n'ont aucune relation sortante et désignent l'argument principal défendu ou attaqué par un ensemble de CA, voire par l'ensemble du texte. Ainsi pour un texte à annoter avec n CA et m CA de type *proposition principale*, il y a exactement $n - m$ relations dans l'annotation. Les structures d'argumentation résultantes peuvent être considérées comme des graphes (Park & Cardie, 2018). La structure sur laquelle nous portons notre attention est plus précisément une forêt d'arbres dont les racines sont les CA de type *proposition principale*.

2.2 Des mesures d'accord inter-annotateurs peu adaptées au cadre relationnelle

Certaines études actuelles continuent d'utiliser des mesures qui n'ont pas été spécifiquement pensées pour des structures de cet ordre, au risque de générer des biais. (Putra *et al.*, 2022), avant de présenter des mesures adaptées à des structures argumentatives, illustrent ce phénomène en reformulant le résultat de l'annotation par trois niveaux de catégorisation dans le but d'appliquer des mesures prévues pour ce paradigme d'annotation :

— la catégorisation binaire de chaque unité du texte, en l'espèce les phrases du corpus, comme étant, ou non, un composant de l'argumentation ;
— la catégorisation binaire de chaque paire d'unités non-identiques a et b telle qu'il existe, ou non, une relation entre a et b ;
— la catégorisation par l'étiquetage d'un type de relation pour chaque paire d'unités a et b où l'annotateur a annoté l'existence d'une relation entre a et b.

Cette approche permet aux auteurs d'utiliser deux mesures très connues pour la catégorisation, l'accord observé et le κ de Cohen (Cohen, 1960).

(Kirschner *et al.*, 2015) identifie certains inconvénients à l'utilisation de ces mesures dans le cadre des structures relationnelles : l'accord observé et le κ fonctionnent correctement lorsque l'annotation porte sur des objets indépendants les uns des autres et distribués uniformément dans le corpus ; or dans le cadre des structures relationnelles l'annotation d'un objet peut dépendre de l'annotation d'un autre objet du corpus, ceci pouvant entraîner des variations de la mesure d'accord et des scores erronés. De plus, dans les corpus présentant des structures relationnelles, les cas où les unités sont fortement connectées entre elles par des relations font figure d'exception. Ainsi, pour n unités on observe de l'ordre de n relations alors qu'étiqueter toutes les paires possibles d'unités distinctes revient à en annoter de l'ordre de n^2. Ainsi plus le corpus présente d'unités et plus on peut être amené à considérer un grand nombre de paires d'unités comme non-connectées, ce qui, selon les besoins de l'étude, apporte peu sur la compréhension du phénomène que l'on veut étudier. Ajoutons que plus deux unités sont éloignées dans l'ordre de lecture dans un même corpus et moins les chances qu'une relation entre ces unités existent sont grandes. (Kirschner *et al.*, 2015) propose de pallier cet effet

en pondérant l'importance des paires en fonction de la distance entre les unités de la paire, ainsi qu'en retirant les paires dont les unités sont trop éloignées l'une de l'autre. Ces réserves conduisent donc naturellement à s'interroger sur la pertinence d'encoder l'annotation de toutes les paires en deux catégories, porteuse ou non d'une relation. Cet exemple illustre les biais qui peuvent résulter de l'utilisation forcée d'une mesure pré-existante sans tenir compte des spécificités des objets annotés. Comme pour les autres domaines d'annotation, il apparaît donc nécessaire de disposer de mesures d'accord spécifiquement pensées pour les structures relationnelles. Un cadre formel commun est sous-jacent aux exemples présentés ci-dessus. Ces structures peuvent être représentées et analysées comme des graphes diversement contraints, ces contraintes pouvant s'exprimer sur les propriétés telles que l'orientation des arrêtes, la présence d'une arborescence ou la connectivité du graphe. Pour ce paradigme de représentation sous forme de graphe, il est dès lors nécessaire de disposer de mesures d'accords adaptées aux particularités de ces structures. C'est ce que fait (Kirschner *et al.*, 2015) en présentant ce qui est à notre connaissance la première mesure d'accord inter-annotateurs spécifiquement créée pour les graphes. Nous comparerons cette mesure aux variantes de *Mean Average Recall* (Putra *et al.*, 2022). Nous détaillerons le fonctionnement de ces mesures en 4.3.

3 Principe des dégradations contrôlées

Pour aider à la comparaison des mesures existantes et le cas échéant à l'élaboration de nouvelles mesures, nous proposons un environnement dans lequel on peut étudier le comportement des mesures en les confrontant à des données dont on contrôle la fiabilité.

3.1 Benchmark et interprétabilité des mesures d'accord inter-annotateurs

Nous l'avons vu plus tôt, les mesures d'accord inter-annotateurs sont des outils cruciaux pour l'établissement de corpus annotés de référence. La question de la qualité des mesures que nous employons est ainsi une problématique d'intérêt pour la communauté scientifique. (Artstein & Poesio, 2008) posent aussi la question de l'interprétabilité des mesures et du sens que portent les scores donnés par ces mesures. Prenant l'exemple de la famille des mesures κ, (Mathet *et al.*, 2012) posent quelques questions qui paraissent triviales en apparence mais qui sont loin de l'être : « Un score κ de 0,75 indique-t-il un bon résultat ? Un score κ de 0,8 est-il deux fois meilleur qu'un score de 0,4 ? Un score de 0,6 obtenu avec une première mesure est-il meilleur qu'un score de 0,5 obtenu avec une autre mesure, et pour quelle tâche d'annotation ? ». Pour répondre à ces questions, sur ces mesures et sur d'autres, les auteurs ont proposé un outil nommé le Corpus Shuffling Tool (CST) qui a pour but de comparer l'évolution des scores de différentes mesures d'accord inter-annotateurs face à des multi-annotations générées de manière contrôlée et paramétrable. Ces données d'annotations sont spécifiquement fabriquées par l'outil pour évaluer les réactions des mesures.

Décrivons étape par étape le fonctionnement du CST. Pour fonctionner, le CST nécessite tout d'abord un jeu d'annotations à dégrader. Nous appellerons ce jeu d'annotations *référence* pour la suite de cette section. Il n'est pas nécessaire que ce jeu soit réellement le *gold standard* issu d'une campagne d'annotation, il suffit qu'il soit représentatif des phénomènes qu'on retrouverait dans une telle annotation de référence. Pour démarrer son processus de dégradation, en plus d'une *référence*, le CST a besoin de deux autres paramètres : le nombre n d'annotateurs à simuler et une magnitude m dont la valeur est comprise entre 0 et 1 inclus. Lors de ce processus, le CST va, pour chacun des

n annotateurs, faire une copie de la *référence* et dégrader aléatoirement les annotations de la copie selon la magnitude m. L'ensemble des copies dégradées ainsi générées forment une multi-annotation que nous pouvons fournir aux mesures que l'on souhaite évaluer. Afin d'observer comment les mesures d'accord inter-annotateurs réagissent face à des multi-annotations de plus en plus dégradées, le CST exécute son processus de dégradation en faisant varier incrémentalement son paramètre de magnitude de 0 à 1 d'un pas paramétrable. Pour chaque valeur de la magnitude on obtient ainsi une multi-annotation composée de jeux d'annotations de plus en plus éloignés de la *référence*. Le désaccord entre les annotateurs simulés est donc théoriquement de plus en plus grand. Enfin, on calcule le score des mesures d'accord inter-annotateurs pour chaque multi-annotation et on trace le graphique résultant en prenant la magnitude m en abscisse du graphique et les scores en ordonnées. Pour que l'évaluation des mesures soit possible avec cet outil, il faut que les jeux d'annotations dégradés soient vraisemblables. Pour se faire, (Mathet *et al.*, 2012) présentent 5 types de dégradations différentes dont le but est de générer des erreurs qui sont observées dans les campagnes d'annotation.

3.2 Types de dégradation

Lorsque (Mathet *et al.*, 2012) présentent leur outil, celui-ci est restreint à deux tâches d'annotation : la catégorisation et l'identification d'unités. Cela les mène à présenter des types de dégradations spécifiques à ces tâches comme la fragmentation : le fait de séparer une unité de référence en plusieurs nouvelles unités plus petites. Dans le cadre de la structure relationnelle que nous étudions et que nous avons précisée en 2.1, plusieurs de ces types de dégradations ne sont pas adaptées à notre cas, et d'autres dégradations nouvelles et propres aux structures relationnelles méritent d'être explorées. Contrairement aux précédents travaux sur le CST, les dégradations proposées ici ne sont pas motivés par un recensement des erreurs rencontrées fréquemment dans les campagnes d'annotation. En effet, nos dégradations se reposent sur l'ensemble des paramètres qui définissent une relation de structure relationnelle, c'est-à-dire : l'origine de la relation, sa cible, son orientation et sa catégorie. Les dégradations ainsi proposées sont donc issues de considérations théoriques et peuvent ne pas pouvoir s'appliquer totalement à toutes les modélisations et campagnes d'annotation, ou alors nécessiter des adaptations substantielles pour être utilisées. L'objectif premier de cet article est de démontrer la possibilité de l'utilisation du CST pour des structures relationnelles et de fournir une *boîte d'outils* divers et indépendants de dégradations pour les futurs travaux s'intéressant à la pertinence des mesures d'accord inter-annotateurs employées. Pour illustrer ces outils, nous préciserons ci-après les types de dégradation que nous proposons dans le cadre particulier des structures argumentatives.

3.2.1 Dégradations élémentaires pour les structures argumentatives

La magnitude m est une valeur réelle telle que $0 \leq m \leq 1$ qui reflète un niveau de dégradation. Au niveau 0 aucune dégradation n'est subie. Au niveau 1 la dégradation est maximale, ce qui simule la perte par un annotateur de toute compétence. Sauf dans le cas particulier du faux positif, la magnitude correspond ici à la probabilité pour chaque relation de la *référence* de subir une dégradation.

Changement de cible : L'identification du rattachement d'un argument pour soutenir ou attaquer un autre peut être source d'erreur. Dans le cadre de cette dégradation l'annotateur simulé assigne comme cible de la relation une autre unité. Chaque unité a une probabilité pondérée d'être choisie comme la nouvelle cible. La structure argumentative peut être vue comme un arbre dont la racine est

la *proposition principale*. Les unités qui se trouvent sur le chemin entre la cible originale et la racine (incluse) ont une probabilité importante d'être sélectionnées. Les unités se trouvant entre l'origine de la relation et les feuilles de l'arbre ont un poids moyen. Enfin les autres unités sur cet arbre reçoivent une pondération faible.

Changement d'origine : Cette dégradation utilise le même principe que le changement de cible, mais c'est ici l'origine de la relation qui est modifiée. En considérant le graphe sous forme d'arbre, la distribution des pondérations pour la nouvelle origine est la suivante : les unités entre l'origine et les feuilles ont un poids élevé, les unités entre la cible et la racine (incluse) reçoivent une pondération moyenne et les autres unités ont un poids faible.

Permutation d'orientation : La permutation d'orientation est le fait pour l'annotateur simulé d'identifier la présence d'une relation entre deux unités en se trompant sur le sens de la relation.

Changement d'étiquetage : Le changement d'étiquetage intervient lorsque l'annotateur se trompe sur la catégorie à attribuer à une relation. Ce phénomène est particulièrement fréquent quand deux catégories de relation sont proches ou qu'une catégorie est rare dans le corpus. Lorsqu'il y a changement de catégorie, une nouvelle catégorie est attribuée au hasard tout en tenant compte la fréquence des catégories dans la *référence*.

Faux négatif : Le faux négatif consiste en le retrait d'une annotation présente dans la *référence*.

Faux positif : En miroir du faux négatif, le faux positif est l'ajout d'une annotation erronée absente de la *référence*. Pour a annotations dans la *référence* et une magnitude m, $a * m$ relations sont crées et ajoutées au jeu de données. Pour ne pas dénaturer substantiellement le sens de la *référence*, les nouvelles relations sont crées en reprenant les caractéristiques statistiques de la *référence*.

3.2.2 Combinaison de dégradations

Lors des campagnes d'annotation, les différents types d'erreurs d'annotation qui peuvent intervenir ne s'excluent pas forcément mutuellement. Pour restituer ce phénomène nous prévoyons de rendre possible la combinaison des dégradations précitées lors du processus de génération des multi-annotations. Dans le cas de la combinaison de dégradations, chacun des t types de dégradation sélectionnés est appliqué successivement sur les jeux d'annotations ; soit avec une magnitude t/m si l'on souhaite une distribution uniforme des dégradations, soit avec une pondération visant à reproduire une distribution observée dans des annotations réelles.

4 Expérimentations sur des structures argumentatives

4.1 Jeu d'annotations et instance étudiée

Pour évaluer les métriques de l'accord inter-annotateurs avec le CST il faut en premier lieu lui fournir une *référence annotée* que l'outil viendra dégrader. Afin de nous procurer une structure argumentative annotée que nous pourrions employer à ces fins, nous nous sommes tout d'abord tourné vers le corpus ICNALE (Ishikawa, 2013, 2018) qui est utilisé par (Putra *et al.*, 2022) pour introduire les mesures MAR. Néanmoins, ce corpus propose des structures d'argumentation avec peu d'unités argumentatives puisque nous y retrouvons environ 13,9 phrases par texte argumentatif, dont toutes ne sont pas CA. Or, nous pensons qu'apporter une simple dégradation aléatoire à des objets argumentatifs de si petite taille modifierait déjà substantiellement la structure argumentative et son sens. Pour opérer nos dégradations sur des objets dont le sens global n'est pas totalement altéré après quelques dégradations successives, et ainsi pouvoir comparer nos mesures sans craindre de trop grandes variations entre chaque itération des dégradations, tout en conservant un coût de calcul acceptable, nous proposons comme *référence* du CST une structure argumentative de taille suffisante avec 101 CA et 100 relations, représentée sur la figure 2. La structure argumentative de référence que nous fournissons à notre implémentation du CST représente l'exemple d'une structure que nous attendons d'une argumentation fournie avec de nombreux arguments et contre-arguments dirigés vers un argument principal. Sur l'illustration, les relations sont dépourvues d'étiquettes puisque les catégories des relations n'influent pas sur les mesures que nous évaluons dans notre étude (cf. 4.3). Chaque relation a néanmoins été annotée avec une étiquette lors de l'élaboration par nos soins de cette structure.

4.2 Dégradations retenues pour cette campagne

Dans la section 3.2 nous avons présenté les types de dégradations envisageables pour des objets tels que la structure argumentative. Pour nos expérimentations nous excluons de cette étude trois types de dégradations. Tout d'abord, nous écartons l'*étiquetage des relations* car les mesures que nous étudions ne prennent pas en compte ce paramètre. Nous n'incluons pas non plus le paradigme du *faux positif* car son implémentation concrète dans le cadre décrit par (Putra *et al.*, 2022) nous interroge encore. En effet, nous attendons $n-1$ relations *explicites* pour une structure à n CA ; or les exemples que nous traitons possèdent déjà $n-1$ relations. Nous ne pouvons donc pas rajouter de relations supplémentaires. Enfin, la *combinaison de dégradations* sera pour l'heure mise de côté car nous cherchons ici à étudier les comportements des mesures en fonction de dégradations précises. L'étude de l'impact d'une combinaison de facteurs serait intéressante mais dépasse notre objectif actuel.

4.3 Mesures d'accord inter-annotateurs implémentées

Comme nous l'avons vu, il existe des méthodes pour traduire les annotations relationnelles en plusieurs étapes successives de catégorisation (Putra *et al.*, 2022) afin de pouvoir utiliser des mesures comme le κ de Cohen sur des structures d'annotations qui ne s'y prêtent pas originellement. Pour les raisons que nous avons évoquées ci-dessus, nous pensons néanmoins qu'il faut privilégier des mesures spécifiquement adaptées à ces structures. Avec le CST, nous cherchons à comparer ici quatre mesures dédiées pour l'accord sur les structures argumentatives : la première *mesure d'accord*

inter-annotateurs basée sur les graphes de (Kirschner *et al.*, 2015) et les trois variantes de *mean average recall* (*MAR*) [1] de (Putra *et al.*, 2022).

Mesure de Kirschner : Soient deux structures argumentatives A et B. Cette mesure adaptée aux graphes dirigés évalue le taux d'inclusion du graphe A dans le graphe B, puis du graphe B dans le graphe A et donne finalement la moyenne (arithmétique ou harmonique) des deux taux.
Pour E_A l'ensemble des relations (x,y) dans le graphe A où x est l'origine de la relation et y la cible et $SP_B(x,y)$ la longueur du plus court chemin entre les nœuds x et y dans le graphe B, l'inclusion de A dans B est donnée par la formule 1 :

$$\text{Mesure de Kirschner} = \frac{1}{|E_A|} \sum_{(x,y)\in E_A} \frac{1}{SP_B(x,y)} \tag{1}$$

MAR^{link} : Cette mesure est la moyenne du rappel des relations de deux structures relationnelles A et B. Soient E_A l'ensemble des relations dans A et E_B l'ensemble des relations dans B, la moyenne des rappels de ces ensembles est donnée par la formule 2.

$$\text{MAR}^{link} = \frac{1}{2}\left(\frac{|E_A \cap E_B|}{|E_A|} + \frac{|E_A \cap E_B|}{|E_B|}\right) \tag{2}$$

MAR^{path} : Plutôt que de s'appuyer sur les relations elles-mêmes, cette variante cherche à mesurer l'accord sur les chemins créés par la mise bout à bout des relations dans la structure argumentative. Pour deux relations dirigées (a,b) et (b,c) où le premier élément est la source et le second l'origine, on obtient trois chemins : [a,b,c], [a,b] et [b,c].
On note respectivement P_A et P_B les ensembles des chemins pour les graphes A et B. L'accord entre ces ensembles est donné par la formule 3 :

$$\text{MAR}^{path} = \frac{1}{2}\left(\frac{|P_A \cap P_B|}{|P_A|} + \frac{|P_A \cap P_B|}{|P_B|}\right) \tag{3}$$

MAR^{dSet} : Cette variante mesure l'accord entre les ensembles de descendants de chaque unité dans la structure argumentative en partant de la racine (*i.e.* le CA *proposition principale*). L'ensemble des descendants d'une unité a est composé de l'unité a elle-même et de toutes les unités qui sont ses descendantes dans l'arbre (c'est-à-dire toutes les unités qui peuvent être atteintes en remontant en sens inverse les relations dirigées en partant de l'unité a). Soient deux jeux d'annotations A et B contenant respectivement les ensembles d'unités (a_1,a_2,...a_n) et (b_1,b_2,...b_n). Une fonction f est définie et prend en paramètre le jeu d'annotations A et retourne un vecteur comprenant les scores de correspondance entre les ensembles de descendants de a_i et de b_i pour $a_i \in$ A et $b_i \in$ B. N_A et N_B correspondent respectivement au nombre d'unités dans A et B. L'accord est donné par la formule 4 :

$$\text{MAR}^{dSet} = \frac{1}{2}\left(\frac{\sum f(A)}{|N_B|} + \frac{\sum f(B)}{|N_A|}\right) \tag{4}$$

1. Au sein même des variantes de MAR, on peut trouver des sous-variantes qui apprécient différemment les relations si elles ont pour trait d'être *explicit* ou *implicit*. Cette distinction, qui est engendrée par le cas particulier de la catégorie *reformulation* peut faire sens dans le cadre de la campagne ICNALE. Nous faisons cependant le choix d'écarter cette catégorie de relation si particulière à une étude pour privilégier la généralisation de nos résultats ; ainsi lorsque nous évoquons par la suite les variantes MAR, nous ne les décrivons que sous leur forme utilisant des relations uniquement *explicit*.

Le score de correspondance entre deux ensembles de descendants peut être *exact* (la valeur est de 1 si les deux ensembles de correspondants pour a_i et b_i sont identiques, 0 sinon) ou *partial* (la valeur est égale au nombre d'unités présentes dans les deux ensembles divisée par le nombre d'unités dans l'ensemble des descendants de b_i). Dans toutes les configurations possibles, le score de correspondance *partial* est toujours supérieur ou égal au score de correspondance *exact*. Nous implémentons dans nos expériences les versions *exact* et *partial* de MAR^{dSet} afin de les comparer.

4.4 Environnement expérimental

Il n'existe pas à notre connaissance d'outil en libre accès pour établir les scores des quatre mesures que nous tentons d'évaluer. Le CST tel que présenté par (Mathet *et al.*, 2012) n'est pas outillé pour des structures relationnelles. Face à ce double manque de logiciels à notre disposition nous avons fait le choix de créer nos propres outils de dégradation d'une référence et de calcul de mesures d'accord pour des structures relationnelles. L'ensemble de ces programmes est réalisé avec les langages *Python 3* et *DOT* et sera rendu accessible à la communauté [2].

4.5 Observations et commentaires

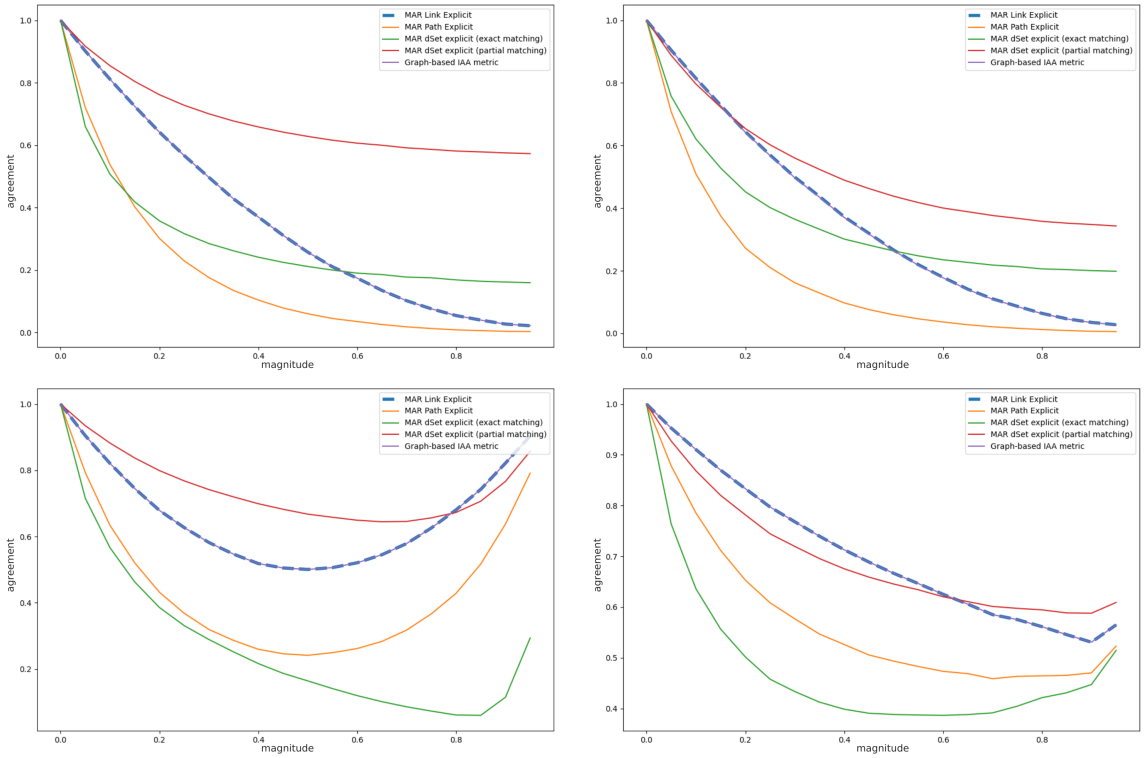


FIGURE 1 – *Changement de cible* (en haut à gauche), *changement d'origine* (en haut à droite), *permutation d'orientation* (en bas à gauche) et *faux négatif* (en bas à droite)

2. https://www.greyc.fr/equipes/codag/#ressources

Nous comparons ici les cinq mesures décrites ci-dessus : la mesure d'accord inter-annotateurs basée sur les graphes (*graph-based IAA metric* (GBM)) (Kirschner *et al.*, 2015), MAR^{link}, MAR^{path}, MAR^{dSet} *exact matching* et MAR^{dSet} *partial matching* (Putra *et al.*, 2022). La figure 1 présente le comportement de ces mesures pour 4 paradigmes de dégradation différents. Dans tous ces paradigmes on remarque que les scores de GBM et MAR^{link} sont toujours identiques alors que ces deux mesures ont des méthodes de calcul différentes. Nous avons remarqué lors de nos expérimentations que ces deux mesures donnent le même score lorsque deux jeux d'annotations comparés possèdent le même nombre de relations, comme c'est le cas dans nos dégradations contrôlées, mais que leurs valeurs divergent lorsque le nombre de relations des deux jeux de données diffère. Pour la suite de nos observations, ce que nous commenterons sur GBM s'appliquera donc aussi à MAR^{link}.
Pour le *changement de cible* et le *changement d'origine*, toutes les mesures suivent une évolution décroissante monotone, ce qui est attendu. Les deux variantes de MAR^{dSet} semblent être limitées chacune par une asymptote, tandis que GBM et MAR^{path} montrent une étendue complète de 1 (accord parfait) à 0 (absence d'accord).
Dans le paradigme de *permutation d'orientation*, aucune des mesures n'est monotone malgré l'augmentation des dégradations. MAR^{path} et GBM atteignent leur minimum lorsque la magnitude est à 0,5 (*i.e.* lorsque la moitié des relations sont inversées dans la structure argumentative) et sont symétriques à l'axe $x = 0.5$. Les variantes de MAR^{dSet} ne présentent pas cette symétrie. Cela s'explique car elles sont les seules mesure à avoir un *sens de lecture particulier* de la structure argumentative, de la racine vers les feuilles. Mécaniquement, leurs scores ne remontent que lorsque les ensembles de descendants se vident et deviennent de moins en moins différents les uns des autres.
Pour le *faux négatif*, aucune des mesures n'est pleinement monotone. Les mesures GBM, MAR^{path} et MAR^{dSet} *partial matching* sont décroissante monotone jusqu'à la magnitude 0,9 ; le score de MAR^{dSet} *exact matching* semble atteindre un plateau minimum dès la magnitude 0.45 avant de croître. Ce phénomène de non-monotonie des courbes pour le *faux négatif* après la magnitude $m = 0.9$ est déjà présent dans les premiers travaux sur le CST (Mathet *et al.*, 2012) pour des paradigmes d'annotation et des mesures différents. Cela est manifestement lié à l'implémentation du *faux négatif* qui est similaire dans les deux études. Nous pouvons l'expliquer simplement, moins il reste d'annotations à comparer dans les simulations générées par le CST et moins les mesures peuvent y trouver du désaccord.

5 Conclusion et perspectives

De manière similaire à (Mathet *et al.*, 2012), notre implémentation du CST pour des structures relationnelles présente des résultats exploitables pour les mesures d'accord inter-annotateurs et les paradigmes de dégradations étudiés. Les comparaisons offertes par cet outil mettent en évidence des similarités entre certaines méthodes, mais surtout les différences et les écueils qui caractérisent les mesures les plus récentes dans le domaine de l'annotation de l'argumentation.

Au rang des perspectives nous prévoyons dans de futurs travaux avec notre outil d'utiliser des références issues directement de campagnes d'annotation et d'améliorer la simulation des dégradations en tentant de nous rapprocher des erreurs commises par les annotateurs humains et de leur fréquence. De plus, l'implémentation du CST que nous proposons aujourd'hui est conçue et spécialisée pour la question de l'accord dans les structures argumentatives. Il serait intéressant de travailler sur d'autres paradigmes de structures relationnelles tels que celui de la coréférence qui possèdent un grand panel de mesures d'accord inter-annotateurs dédiées.

Remerciements

Nous remercions grandement Yann Mathet et Antoine Widlöcher pour leur aide précieuse en tant qu'encadrants, mais aussi pour leurs encouragements et leurs nombreux conseils tout au long de la réalisation de ce travail. Nous remercions également les trois relecteurs anonymes pour leurs commentaires très utiles à l'amélioration de ce document.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, **34**(4), 555–596. DOI : 10.1162/coli.07-034-R2.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46. Place : US Publisher : Sage Publications, DOI : 10.1177/001316446002000104.

FORT K. (2016). *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*. Wiley, 1 édition. DOI : 10.1002/9781119306696.

ISHIKAWA S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, **1**, 91–118.

ISHIKAWA S. (2018). The ICNALE Edited Essays : A Dataset for Analysis of L2 English Learner Essays Based on a New Integrative Viewpoint. *English Corpus Linguistics*, **25**, 1–14.

KIRSCHNER C., ECKLE-KOHLER J. & GUREVYCH I. (2015). Linking the Thoughts : Analysis of Argumentation Structures in Scientific Publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, p. 1–11, Denver, CO : Association for Computational Linguistics. DOI : 10.3115/v1/W15-0501.

KRIPPENDORFF K. (2019). *Content Analysis : An Introduction to Its Methodology*. SAGE Publications, Inc. DOI : 10.4135/9781071878781.

LEECH G. N., GARSIDE R. G. & MCENERY T. (1997). *Corpus annotation : linguistic information from computer text corpora / Roger Garside, Geoffrey Leech, Tony McEnery.* London : Longman.

MATHET Y., WIDLÖCHER A., FORT K., FRANÇOIS C., GALIBERT O., GROUIN C., KAHN J., ROSSET S. & ZWEIGENBAUM P. (2012). Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics. In M. KAY & C. BOITET, Éds., *Proceedings of COLING 2012 : Posters*, p. 809–818, Mumbai, India : COLING 2012 Organizing Committee.

PARK J. & CARDIE C. (2018). A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

PUTRA J. W. G., TEUFEL S. & TOKUNAGA T. (2022). Annotating argumentative structure in English-as-a-Foreign-Language learner essays. *Natural Language Engineering*, **28**(6), 797–823. Publisher : Cambridge University Press, DOI : 10.1017/S1351324921000218.

SKEPPSTEDT M., PELDSZUS A. & STEDE M. (2018). More or less controlled elicitation of argumentative text : Enlarging a microtext corpus via crowdsourcing. In N. SLONIM & R. AHARONOV, Éds., *Proceedings of the 5th Workshop on Argument Mining*, p. 155–163, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5218.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In H. S. THOMPSON & A. LASCARIDES, Éds., *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, p. 110–117, Bergen, Norway : Association for Computational Linguistics.

WACHSMUTH H., AL-KHATIB K. & STEIN B. (2016). Using Argument Mining to Assess the Argumentation Quality of Essays. In Y. MATSUMOTO & R. PRASAD, Éds., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1680–1691, Osaka, Japan : The COLING 2016 Organizing Committee.

A Annexes

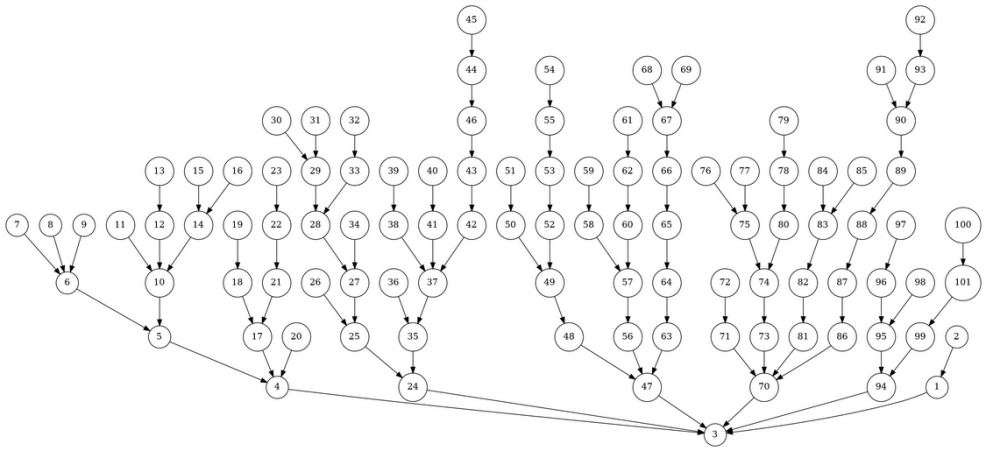


FIGURE 2 – Structure argumentative de référence fournie à notre implémentation du CST, l'AC 3 est de type *proposition principale*

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

Géométrie des vecteurs de tâches pour l'association et la combinaison de modèles

Loïc Fosse
Orange Innovation, Lannion, France
Aix Marseille Univ., CNRS, LIS, Marseille, France
loic.fosse@orange.com

RÉSUMÉ

Les adaptations de rang faible (LoRA, pour *Low-Rank Adaptation*) sont devenues un standard pour adapter des modèles à un faible coût. Elles sont de plus en plus utilisées que ce soit en traitement du langage ou des images. Plusieurs études utilisent ces adaptations et cherchent à les combiner *a posteriori* de manière à enrichir de manière additive les propriétés d'un modèle. Ces combinaisons suggèrent alors que nous pouvons associer les modèles dans l'espace des paramètres et que nous pouvons donner un sens à cela. Cette propriété n'est que très peu vérifiée dans la pratique et nous proposons ici plusieurs métriques visant à caractériser l'association entre les modèles dans l'espace des paramètres. Nous montrons finalement que nous pouvons corréler ces métriques avec les pertes de performance des modèles lorsque nous réalisons leurs combinaisons.

ABSTRACT

Task vectors geometry for models association and combination.

Low-Rank Adaptations (LoRA) have become a standard for adapting models at low cost. They are increasingly used in both language and image processing. Several studies use these adaptations and seek to combine them *a posteriori* so as to additively enrich the properties of a model. These combinations then suggest that we can associate models in parameter space and make sense of this. This property is only rarely verified in practice, and here we propose several metrics to characterize the association between models in parameter space. Finally, we show that we can correlate these metrics with losses in model performance when we combine them.

MOTS-CLÉS : transformers, adaptation de rangs faibles, similarité de modèles, distance de Grassmann, combinaisons de modèles.

KEYWORDS: transformers, low rank adaptation, models similarities, Grassmann distance, model combinations.

1 Introduction

L'accroissement de la taille des grands modèles de langue, basés sur l'architecture *transformers* (Vaswani *et al.*, 2017), rend de plus en plus complexe et coûteuse leur adaptation sur des tâches précises. Dans l'optique de dépasser les contraintes de ressources pour spécialiser des modèles, deux méthodes se distinguent : les méthodes d'affinages efficaces (*parameter efficient fine-tuning*) et la combinaison de modèles. Les méthodes d'affinage dites efficaces, comme la méthode LoRA (Hu *et al.*, 2021), le *prefix tuning* (Li & Liang, 2021), le *scaling and shifting* (Lian *et al.*, 2022) ou encore

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 69–84.

le IA3 (Liu *et al.*, 2022), qui sont des dérivées de la notion d'adapteur (Rebuffi *et al.*, 2017), visent à n'affiner qu'une fraction du modèle, basé sur l'hypothèse que pour se spécialiser dans une tâche, seules des modifications mineures du modèle sont nécessaires. Parmi ces méthodes efficaces, les adaptations de rangs faibles (LoRA : *Low-Rank Adaptation*) se détachent du lot, de par les résultats qu'elles permettent d'obtenir en traitement automatique des langues (TAL) (Aleem *et al.*, 2024; Dettmers *et al.*, 2024; Kaddour *et al.*, 2023) et en vision (Luo *et al.*, 2023; Gandikota *et al.*, 2023), ou encore les propriétés intéressantes qui se dégagent de ces méthodes (Fu *et al.*, 2023). La combinaison (ou interpolation) de modèles (Li *et al.*, 2023; Jin *et al.*, 2022; Matena & Raffel, 2022; Falissard *et al.*, 2023; Yu *et al.*, 2023), inspirée des méthodes ensemblistes (Dietterich, 2000), quant à elle, vise à considérer des modèles entraînés sur des tâches et à les combiner dans l'espace des paramètres de manière à créer un nouveau modèle potentiellement meilleur que les modèles précédents sur une tâche donnée (enrichissement), capable de faire plusieurs tâches (multi-tâches) (Yang *et al.*, 2023) ou encore à faire une nouvelle tâche encore jamais vue par les différents modèles (Pfeiffer *et al.*, 2023; Chronopoulou *et al.*, 2023) – dans la littérature, on parle de *Modular Learning* pour ce dernier cas.

Plus récemment, des études cherchent aussi à combiner les modèles adaptés avec la méthode LoRA en TAL. Par exemple, Zhang *et al.* (2024) proposent des méthodes de combinaisons simples, de manière à : supprimer les biais d'un modèle, faire du multi-tâche, supprimer des composantes d'un modèle (notamment la toxicité dans la génération de textes) ou encore faire du transfert de domaine. De la même manière, dans le domaine du traitement des images, certaines études apparaissent aussi sur la composition de modèles adaptés avec LoRA, de manière à additionner les propriétés dans la génération d'images (Shah *et al.*, 2023; Gu *et al.*, 2024).

La combinaison de modèles (affinés complètement ou avec des méthodes de faible rang) semble alors fournir des résultats intéressants, mais elle n'est pas sans poser de questions. En effet, Frankle *et al.* (2020) introduisent la notion de connectivité linéaire et montrent que la combinaison de modèles peut parfois mener à des modèles instables perdant toute notion des tâches sur lesquelles ils ont été entraînés, notamment à cause de problèmes d'interférences entre modèles, qui peuvent être évités en forçant les modèles à partager une partie de leur trajectoire d'entraînement. Ce problème d'interférence entre les modèles est évoqué aussi plus récemment par Yadav *et al.* (2023), et ils proposent un algorithme empirique, pour adresser ce problème, basé sur l'amplitude des poids entre les différents modèles : lors de la comparaison de deux modèles, celui ayant les poids de plus grande amplitude sera le plus important.

La combinaison de modèles soulève alors la question de l'associativité de modèles entraînés sur des tâches différentes, afin de comprendre et d'interpréter les possibles effets d'interférence entre ces modèles. Intuitivement, la combinaison de deux modèles qui sont « proches » résultera en un modèle similaire aux deux premiers (peu d'interférence) et, inversement, la combinaison de deux modèles qui sont « éloignés » résultera en un modèle différent des deux premiers (grande interférence). Deux questions viennent alors naturellement : (1) comment définir la notion de proximité entre les modèles dans l'espace des paramètres ; et (2) pouvons-nous interpréter la proximité entre les modèles par rapport à une proximité empirique des tâches sur lesquelles ils ont été affinés afin d'ainsi comprendre *a posteriori* le comportement de modèles combinés ?

Nous adressons ces questions dans cette étude, à travers le prisme des adaptations de rang faible ainsi que des adaptations complètes de modèles (*full fine-tuning*), sur des tâches classiques de classification en traitement automatique des langues. Dans la suite, la section 2 présente le formalisme et les tâches étudiées, puis la section 3 présente les expériences et les résultats.

2 Formalisme et tâches

Dans un premier temps, nous définissons le formalisme associé aux adaptations de rang faible, les différentes distances utilisées pour calculer les proximités entre modèles, ainsi que les différentes tâches sur lesquelles nous allons travailler.

2.1 Adaptations de rang faible et vecteurs de tâche

Les adaptations de rang faible (LoRA) consistent à re-paramétriser le gradient d'un modèle par une décomposition de rang faible. Lors de l'apprentissage d'un modèle (*full fine-tuning*) paramétré par une matrice $W \in \mathbb{R}^{d \times d}$ (où d représente généralement la dimension des états cachés), nous partons d'un point W_0 (en général, un modèle pré-entraîné) et nous mettons à jour ces poids par descente de gradient pour arriver finalement à $W_f = W_0 + \Delta W$. Cette mise à jour peut être lourde étant donné que ΔW est de la même taille que le modèle de base. L'idée est alors de poser un nouveau paramètre $r << d$ (que l'on choisit) et de considérer $A \in \mathbb{R}^{r \times d}$ et $B \in \mathbb{R}^{d \times r}$ tel que $\Delta W = BA$, puis nous cherchons simplement à apprendre les poids des matrices A et B en laissant intact W_0. Dans la pratique, ces adaptations se réalisent sur les couches linéaires d'un modèle, et nous définissons un couple (A, B) pour chaque couche linéaire que l'on souhaite adapter de cette manière. Typiquement, dans un modèle *transformer*, nous pouvons définir un couple (A, B), pour les différentes projections en requêtes, clés et/ou valeurs [1] (Ghojogh & Ghodsi, 2020).

Ces adaptations légères possèdent des propriétés intéressantes, notamment le fait de régulariser l'entraînement et donc d'éviter les effets de sur-apprentissage (Fu *et al.*, 2023).

De plus, ces adaptations sont étroitement liées avec la notion de vecteurs de tâches introduite par Ilharco *et al.* (2022). Dans cette étude, les auteurs définissent le *vecteur de tâche* comme la différence entre les poids du modèle pré-entraîné et le modèle affiné sur une tâche. Avec les définitions précédentes, le vecteur de tâche correspond donc exactement à la quantité ΔW définie plus haut et donc au produit BA si le modèle est affiné avec des adaptations de faible rang.

Nous décidons alors d'exploiter cette notion de vecteur de tâche et de regarder ce que ces vecteurs peuvent encoder dans leur géométrie. Nous définissons, dans la section suivante, les distances qui vont nous aider à extraire des informations de ces vecteurs de tâches.

2.2 Distance entre vecteurs de tâches

Afin d'estimer des similarités entre tâches, il faut pouvoir être capable de calculer des distances entre des modèles dédiées à ces tâches. Nous proposons ici trois distances (ou *pseudo*-distances), entre deux matrices de paramètres de même taille W_1 (modèle d'une tâche 1) et W_2 (modèle d'une tâche 2), de manière hiérarchique (de la plus contraignante à la moins contraignante) :

Distance L_2 : $||W_1 - W_2||_2$ Cette distance est la plus naturelle et nous donne des informations sur les positions absolues des modèles.

1. Nous pouvons le faire pour tout type de matrices de projection, du modèle sans se limiter aux blocs d'attention, mais une pratique courante est de se limiter aux projections en requêtes et valeurs, comme cela est fait dans (Hu *et al.*, 2021)

Distance du cosinus : $1 - cos(W_1, W_2)$ Cette distance [2, 3] nous donne la corrélation empirique entre les poids des différents modèles (Saporta, 2006), et nous renseigne donc sur les variations communes entre les poids des modèles. Cette distance a notamment été utilisée dans Ilharco *et al.* (2022) pour calculer des similarités entre des modèles de classification d'images.

Distance de Grassmann (Hamm & Lee, 2008) : $d_G(W_1, W_2)$ Cette distance particulière, travaille sur les espaces images des modèles. Elle évaluera la proximité entre les espaces vectoriels images $Im(W_1)$ et $Im(W_2)$. En d'autre termes, elle évaluera la proximité entre les représentations en sortie des modèles. Cette distance a déjà été utilisée dans (Hu *et al.*, 2021) (*cf.* annexe G) pour estimer des dimensions communes entre des matrices (A, B) issues de LoRA avec des rangs différents entraînés sur une même tâche.

La distance de Grassmann s'appuie sur la notion d'angles principaux entre des espaces vectoriels (Afriat, 1957; Miao & Ben-Israel, 1992). Si W_1 et W_2 sont deux matrices de rang r dont les colonnes sont orthonormées [4], nous pouvons alors considérer σ, l'ensemble des valeurs propres de $W_1^T W_2$. Un résultat théorique donné par Björck & Golub (1973) est que ces valeurs propres sont les cosinus des angles principaux entre les espaces images de W_1 et W_2 *i.e.* si on pose θ l'ensemble des angles principaux entre $Im(W_1)$ et $Im(W_2)$, alors $cos(\theta) = \sigma$. Basés sur cette information, nous avons alors :

$$d_G(W_1, W_2) = \sqrt{\sum_{i=1}^{r} (\theta_i)^2} \leq \sqrt{r}\frac{\pi}{2} \tag{1}$$

Une remarque supplémentaire sur cette distance de Grassmann, est que si l'on considère $(W_1, W_2) \in \mathbb{R}^{d \times d}$, deux matrices de rang maximal d, alors $Im(W_1) = Im(W_2) = \mathbb{R}^d$ impliquant donc $d_G(W_1, W_2) = 0$. Cette distance n'est donc intéressante que sur les matrices qui ne sont pas de rang maximal, la rendant particulièrement intéressante dans le cadre de LoRA.

Nous avons ainsi trois distances qui nous permettent d'avoir un niveau d'analyse en cascade : la distance L_2 donne les positions absolues et est donc très stricte ; la distance du cosinus, quant à elle, est plus souple et nous donne la corrélation empirique entre les poids en faisant abstraction de l'amplitude de ces derniers ; et enfin la distance de Grassmann nous donne les distances entre les espaces images et fait abstraction de l'amplitude et de l'orientation, comme on peut le voir sur la figure 1 qui illustre sur un cas simple nos différentes métriques.

2.3 Tâches et entraînements des modèles

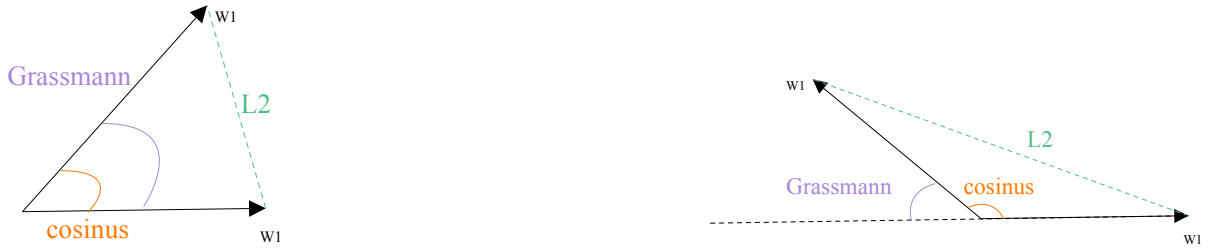
Nous décidons de nous focaliser ici sur des tâches de classification. Nous regardons les tâches suivantes appartenant au *benchmark* GLUE (Wang *et al.*, 2018) :

— COLA : tâche d'acceptabilité linguistique (classification binaire)
— MRPC : détection de paraphrase entre deux énoncés (classification binaire)
— RTE : inférence en langue naturelle (classification binaire)
— QQP : similarité / détection de paraphrase entre deux questions (classification binaire)

2. Le cosinus entre deux matrices est défini comme le cosinus entre les poids applatis de ces deux matrices. Nous considérons également la quantité $1 - cos()$ de manière à avoir une distance et non pas une similarité.

3. Mathématiquement il ne s'agit pas d'une distance. Nous utilisons le terme distance pour traduire le fait qu'une valeur élevée (faible) traduira un éloignement (une proximité).

4. si elles ne le sont pas, une simple décomposition en valeurs singulières des matrices peut nous permettre de le faire



(a) Grassmann = cos

(b) Grassmann ≠ cos

FIGURE 1 – Visualisation des différentes distances sur deux modèles ($W1$, $W2$) à 1 dimension (Vecteurs).

	COLA	MRPC	RTE	QNLI	QQP	MNLI	SST2	SNLI	YELP	IMDB
train	8.55k	3.67k	2.49k	105k	364k	393k	67.3k	550k	560k	25k
dev	1.04k	408	277	5.46k	40.4k	19.65k	872	10k	-	-
test	-	-	-	-	-	-	-	10k	38k	25k

TABLE 1 – Cardinalité des différents jeux de données utilisés

— QNLI : question-réponse sous forme d'inférence en langue naturelle (classification binaire)
— MNLI : inférence en langue naturelle (classification à trois classes)
— SST2 : classification en polarité (classification binaire)

Nous ajoutons la tâche SNLI (Inférence en Langue Naturelle) (Bowman *et al.*, 2015) en complément de la tâche MNLI ainsi que YELP (Asghar, 2016) et IMDB (Maas *et al.*, 2011) en complément de la tâche SST2. Cet ajout est fait pour avoir des jeux de données plus volumineux, contenant des biais différents (notamment pour SNLI), afin d'enrichir nos observations. L'ensemble de ces corpus sont anglais et nous donnons dans la table 1 les tailles respectives des différents jeux de données.

Pour les jeux de données du *benchmark* GLUE, nous n'avons pas à notre disposition un jeu de test annoté [5], les performances sont donc données sur le jeu de validation. Le jeu d'entraînement est lui coupé en deux (90-10) pour faire un jeu d'entraînement et un jeu de validation. C'est cette même stratégie qui est utilisée dans (Zhang *et al.*, 2024).

Tous nos modèles sont affinés à partir du modèle pré-entraîné `roberta-base` (Liu *et al.*, 2019). Pour les affinages complets, nous utilisons pour chaque jeu de données un taux d'apprentissage de 2×10^{-5} avec des lots de taille 20 pendant 20 époques avec une stratégie d'arrêt prématuré fondée sur la fonction de perte de validation avec une patience de 5. Pour les affinages `via` LoRA, nous utilisons un rang $r = 8$ avec un coefficient régularisateur $\alpha = 16$ (2 fois le rang) et un taux d'apprentissage de 10^{-4} (5 fois plus grand que pour les affinages complets), sur le même nombre d'époques et la même stratégie d'arrêt que pour l'affinage complet. Ces adaptations sont réalisées sur les projections de requêtes et de valeurs dans chaque module d'attention. Nous reportons les performances des différents modèles dans le tableau 2. L'ensemble de nos modèles a été entraîné avec les bibliothèques `transformers` et `peft` (pour la méthode LoRA).

5. Les jeux de données tests sont donnés simplement avec le texte et nous devons envoyer nos prédictions aux auteurs pour qu'ils réalisent l'évaluation.

	COLA	MRPC	RTE	QNLI	QQP	MNLI	SST2	SNLI	YELP	IMDB
f-FT	56.36	85.78	69.66	91.87	86.36	85.75	93.69	90.61	97.73	93.95
LoRa	54.98	86.52	71.84	92.28	86.02	86.73	93.92	90.61	98.01	95.34

TABLE 2 – Performances des modèles sur chacune des tâches avec une distinction entre *full-finetuning* (*f-FT*) et adaptations de rangs faibles (LoRa). Les performances sont données pour chaque jeu de données avec la métrique officielle GLUE donnée dans (Wang *et al.*, 2018) et calculée avec la librairie evaluate. Le jeu de données SNLI est évalué avec le F1 de MNLI tandis que YELP et IMDB sont évalués avec la métrique de SST2. Nous ne pouvons pas ici comparer les colonnes entre elles (métrique différente pour chaque jeu de données)

Dans l'ensemble de nos corpus, il est important de noter que les tâches d'inférence (lien logique entre deux énoncés) et les tâches de détection de paraphrase (similarité sémantique entre deux énoncés), peuvent être très similaires dans leurs définitions. La différence majeure entre ces deux étant que la dernière est une tâche symétrique par rapport aux énoncés, contrairement à la première.

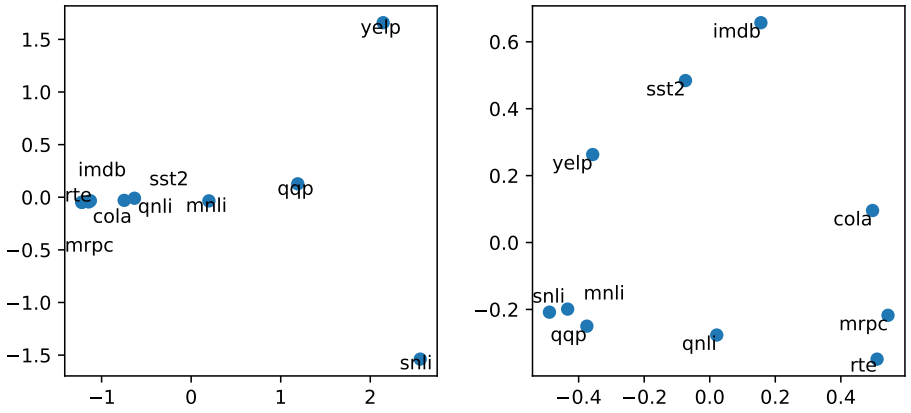
3 Expériences et résultats

Dans cette section, nous utilisons les différentes métriques introduites pour montrer que les distances entre les modèles sont cohérentes par rapport aux tâches sur lesquelles ils sont entraînés. Enfin, nous illustrons le fait que combiner des couples de matrices LoRA (A, B) éloignées, au sens de nos métriques, dégrade les performances de nos modèles allant ainsi dans le sens de nos hypothèses sur les interférences entre ces derniers : combiner des modèles éloignés produira un modèle aux performances dégradées sur les tâches initiales de chaque modèle combiné.

3.1 Distance entre vecteurs de tâches

Pour chacune des distances d évoquées précédemment, nous construisons la matrice 3D T_d telle que $T_d(i, j, k)$ représente la distance d entre le vecteur de tâche des paramètres de la couche k des modèles respectivement entraînés sur la tâche i et j. Pour faciliter les analyses de cet objet, nous considérerons la réduction de T_d par la moyenne sur les couches *i.e.* $\bar{T}_d(i, j) = \frac{1}{|K|} \sum_{k=1}^{|K|} T_d(i, j, k)$. Nous obtenons finalement une matrice symétrique $\bar{T}_d(i, j)$ qui rend compte des distances entre les modèles [6]. Un moyen simple de visualiser cette matrice est alors de réaliser une réduction de dimension par analyse en composantes principales (ACP) (F.R.S., 1901) afin de visualiser les modèles qui sont proches sur un plan. Nous reportons sur la figure 2 cette ACP sur les vecteurs de tâches des modèles affinés complètement. Nous ne présentons sur cette figure que les distances du cosinus et L_2, en effet pour les modèles affinés complètement les rangs des vecteurs de tâches sont très proches des rangs maximaux (rang variant entre 766 et 768 pour un rang maximal de 768) rendant peu informative et très coûteuse la distance de Grassmann (*cf.* section 2.2). Sur la figure 2b, nous observons des comportements intéressants. Les modèles affinés sur les tâches de classification en polarité se retrouvent groupés ensemble (YELP, SST2, IMDB), les modèles affinés sur RTE et MRPC

6. Nous notons que la moyenne à travers les couches peut masquer les comportements différents de chaque couches et nous adressons ceci plus loin dans l'étude

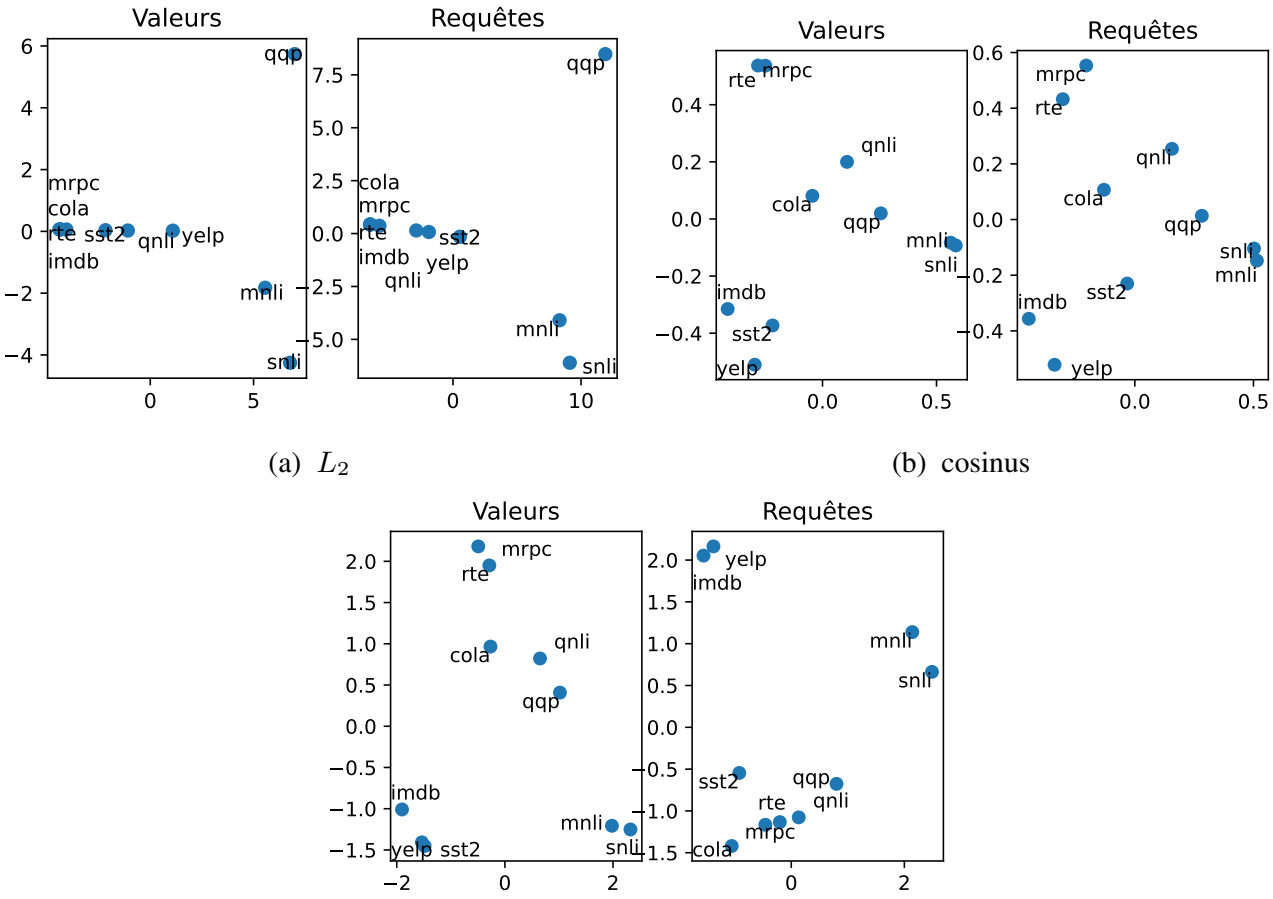


(a) L_2 (b) cosinus

FIGURE 2 – Projection deux dimension par PCA des matrices de distances $\bar{T}_d$, pour les vecteurs de tâches des modèles affinés complètement.

sont aussi groupés ensembles ce qui peut s'expliquer par le fait que les tâches d'inférence et de détection de paraphrases sont très proches dans leurs définitions. Les tâches d'inférence comme MNLI, SNLI et QNLI dans une moindre mesure sont aussi groupées. La tâche QQP, qui semble s'apparenter à la tâche MRPC (détection de paraphrase), se range plutôt du côté des tâches d'inférence (MNLI, SNLI). Enfin, la tâche COLA, seule tâche d'acceptabilité linguistique, ne semble pas montrer d'association particulière avec une autre tâche. La séparation entre le groupe MNLI, SNLI et RTE, MRPC (qui sont 4 tâches très proches dans leurs définitions) peut s'expliquer par les tailles des jeux d'entraînements (*c.f.* table 1) : pour MNLI et SNLI, nous avons des jeux d'entraînements très volumineux (plusieurs centaines de milliers de phrases) tandis que ceux de RTE et MRPC sont plus modestes (quelques milliers). L'entraînement sur les jeux de données volumineux présente donc plus d'étapes d'entraînements (plus de mises à jour du modèle par descente de gradient), pouvant donc expliquer les différences observées. La distance du cosinus nous permet ainsi d'identifier des liens entre les modèles, qui semblent cohérents par rapport à leurs jeux de données d'entraînement. En revanche, les groupements réalisés sur la figure 2a ne donnent lieu à aucune interprétation intéressante traduisant le fait que la distance L_2 est trop restrictive pour bien évaluer des distances entre des modèles en comparaison à la distance du cosinus.

Nous reprenons ainsi cette analyse sur les modèles affinés avec des adaptations de faible rang. Pour rappel, dans le cadre de la méthode LoRA, le vecteur de tâche nous est directement donné par les produits BA. Comme évoqué précédemment, avec cette méthode, nous faisons une adaptation sur les projections en requêtes et en valeurs. Pour l'analyse de ces modèles, nous décidons donc de calculer T_d^r et T_d^v respectivement pour les requêtes et les valeurs (que l'on moyenne sur les couches). Nous pouvons retrouver les résultats de projection sur la figure 3. Nous ajoutons en plus la distance de Grassmann étant donné que cette fois-ci les vecteurs de tâches sont de faible rang (ce sont les produits BA), donnant ainsi un intérêt à cette distance. Sur la figure 3a, nous faisons la même constatation que sur la figure 2a : la distance L_2 est une distance trop restrictive pour apprécier les différences entre les modèles et nous ne distinguons pas de groupements particuliers. Sur la figure 3b, représentant les distance du cosinus entre produits BA, nous faisons des interprétations sensiblement similaires à celles sur la figure 2b avec, cette fois, la tâche COLA qui semble se rapprocher du groupement MNLI-SNLI. On observe en plus, très peu de différences entre les requêtes et les valeurs (excepté un léger déplacement de la tâche SST2), traduisant donc que, vis à vis de la distance du cosinus,



(a) L_2

(b) cosinus

(c) Grassmann

FIGURE 3 – Projection deux dimension par PCA des matrices de distances $\bar{T}_d$, pour les différentes distances.

les modules de requêtes et de valeurs dans les blocs d'attention donnent les mêmes informations sur les différentes tâches. En revanche, sur la figure 3c, représentant les distances de Grassmann, nous remarquons désormais des différences entre les requêtes et les valeurs. L'interprétation sur les valeurs est identique à celle sur les cosinus, avec, en plus, des groupements qui sont mieux définis et mieux espacés sur les axes principaux (*cf.* échelle sur les axes). En revanche, celle sur les requêtes est légèrement différente, le groupement des tâches de polarité n'est plus clairement formé, avec la tâche SST2 qui se rapproche de tâches comme MRPC, RTE ou COLA. Il semblerait alors que, dans les projections de requêtes avec la distance de Grassmann, nous perdons certaines informations relatives à nos différentes tâches. La distance de Grassmann permet ainsi d'identifier, géométriquement, le fait que nous perdons certaines informations relatives aux différentes tâches sur les projections en requêtes dans le module d'attention. Ceci corrobore les travaux de Ethayarajh (2019); Fosse *et al.* (2023, 2022) qui montrent que, dans les modèles *transformers*, c'est la direction prise par les plongements en sortie des modules d'auto-attention qui donne le plus d'informations. Or, comme cela est rappelé dans (Fosse *et al.*, 2023), dans les blocs d'attention, c'est principalement le module de valeurs qui influe sur la direction donnée aux plongements.

	L_2	cos	d_G	cos^{5+}	d_G^{5+}
Requêtes	0.45 (0.21)	0.65 (0.16)	0.65 (0.25)	0.65 (0.16)	0.61 (0.25)
Valeurs	0.48 (0.23)	0.68 (0.21)	0.64 (0.19)	0.74 (0.20)	**0.77 (0.14)**

TABLE 3 – Moyenne (écart type) sur les tâches, des coefficients r de Spearman entre $\delta(i,.)$ et $\bar{T_d}(i,.)$.

3.2 Combinaison de modèles

Nous décidons de poursuivre cette étude sur les distances entre les modèles en réalisant des combinaisons entre ces derniers afin d'inspecter le comportement des modèles combinés en fonction de leur distance dans le but de mieux comprendre les effets d'interférence. Dans cette partie, nous nous concentrons sur les adaptations de rang faible et adoptons la démarche utilisée par Zhang *et al.* (2024). Si on considère les deux couples (A_1, B_1) et (A_2, B_2) de matrices LoRA dont le rang est identique, alors le couple résultant de leur combinaison sera défini par $(\frac{1}{2}(A_1 + A_2), \frac{1}{2}(B_1 + B_2))$ [7]. Ainsi, pour l'ensemble des tâches que nous avons introduites, nous définissons δ tel que $\delta(i,j)$, représente la différence de performance sur la tâche i entre (1) le modèle entraîné sur la tâche i et (2) le modèle entraîné sur la tâche i et combiné avec la tâche j. Pour toutes les tâches i, nous calculons la corrélation de Spearman (Sedgwick, 2014) entre $\delta(i,.)$ et $\bar{T_d}(i,.)$. Nous reportons les résultats de corrélations dans la table 3 (colonnes L_2, cos et d_G). Nous pouvons tout d'abord y voir que l'ensemble des corrélations sont non nulles, ce qui va dans le sens de notre hypothèse de départ, à savoir à nouveau que l'association entre deux modèles éloignés résultera en un modèle moins bon sur les tâches initiales (une grande distance semble être associée à un grand δ et donc une grande interférence). Nous pouvons également voir que les valeurs de corrélations pour la distance L_2 sont plus faibles que pour les autres distances avec aussi des écart-types associés relativement grands. Ceci est à mettre en lien avec nos observations sur la distance L_2 qui était trop stricte et ne nous permettait pas de visualiser des associations intéressantes entre les tâches. Enfin, nous constatons que les corrélations calculées pour les distances de Grassmann et du cosinus sont plus fortes, traduisant le fait que ces métriques semblent plus intéressantes pour estimer les effets d'interférence entre modèles (et aussi les distances entre modèles *cf.* section 3.1).

Les observations de Rogers *et al.* (2021) suggèrent cependant que les différentes couches d'un modèle *transformer* encodent des informations différentes qui sont de plus en plus liées à la tâche à mesure que l'on monte dans les couches. Nous décidons alors de recalculer les corrélations précédentes entre $\delta(i,.)$ et $\bar{T_d^k}(i,.)$, où $\bar{T_d^k}(i,.)$ représente les distances moyennées sur les k *dernières* couches. Dans cette dernière expérience, nous supprimons la distance L_2 qui n'a pas semblé fournir de résultats intéressants jusqu'ici. Nous testons dans un premier temps pour $k = 12$ (c.-à-d. la dernière couche dans un modèle `roberta-base`) et observons une chute des coefficients de corrélation, traduisant le fait que la dernière couche d'un modèle ne semble pas suffire pour estimer les effets d'interférence entre ces derniers. Nous reportons finalement les résultats pour $k = 5$ (qui semble nous fournir les meilleurs résultats) sur la table 3 (colonnes cos^{5+} et d_G^{5+}). Cette fois-ci, pour les modules de valeurs dans les blocs d'attention, nous observons une hausse des coefficients de corrélations et une baisse de la variance associée. La distance de Grassmann semble en particulier donner des corrélations plus fortes. Les modules de requêtes ne montrent pas d'évolution particulière, ce qui est à mettre en lien avec les précédentes observations de la section 3.1 (les modules de requêtes influent peu sur la

7. Nous travaillons avec des modèles de type encodeur, la tête de classification n'est pas modifiée lors de la combinaison de deux modèles

direction des plongements). Il semblerait alors que les couches profondes des modèles *transformers* soient plus intéressantes pour caractériser les tâches et les possibles effets d'interférence entre les modèles, les couches se situant aux extrémités étant trop ou pas assez spécialisées sur les données ce qui semble correspondre à nouveau avec les observations de Rogers *et al.* (2021).

3.3 Interprétation géométrique

À ce stade, nous pouvons faire une interprétation géométrique sur la distance de Grassmann. Durant les expériences précédentes sur les adaptations de rang faible (distance et corrélation suite aux combinaisons), la distance de Grassmann nous a donné des résultats très similaires à la distance du cosinus. Cependant, lorsque nous calculons cette distance entre deux couples (A_1, B_1), (A_2, B_2) LoRA, nous avons l'égalité suivante [8] :

$$d_G(B_1A_1, B_2A_2) = d_G(B_1, B_2)\ . \tag{2}$$

Cette égalité est la conséquence directe du fait que pour tout couple (A, B), nous avons $Im(BA) \subset Im(B)$ ainsi que du fait que A est surjective (vérification empirique). Ainsi, le fait que la distance du cosinus et la distance de Grassmann amènent aux mêmes interprétations suggère que la majeure partie des informations portées par une adaptation LoRA réside dans leurs matrices B. Ceci peut s'expliquer de par le fait que les matrices A projettent des représentations de très grandes dimensions (ici 768) dans un espace de très faible dimension (ici 8). Cette forte compression induit nécessairement une forte distortion et donc une grande perte d'informations dans les représentations. C'est ensuite la matrice B qui se charge de re-projeter les représentations dans un espace de grande dimension et de donner ainsi à chaque représentation sa direction.

4 Conclusion et perspectives

À travers cette étude, nous reprenons différentes métriques afin de calculer l'association entre les modèles dans l'espace des paramètres, basé sur la notion de vecteurs de tâches. Nous montrons que la distance L_2 est trop restrictive et ne permet pas de bien apprécier les différences entre les modèles. Il faut alors se tourner vers des métriques plus vectorielles (sensibles à la direction et non à la norme), comme la distance du cosinus ou celle de Grassmann, lorsque cette dernière fait du sens, qui permettent de mettre en évidence des liens cohérents entre les tâches sur lesquelles les modèles ont été entraînés. Cette association entre les tâches est notamment mise en avant en combinant des modèles adaptés avec LoRA. Nous montrons que combiner des modèles éloignés dans l'espace des paramètres semble perturber ces derniers en les rendant moins bons sur les tâches initiales, répondant ainsi à Ilharco *et al.* (2022), qui évoquaient dans leur discussion que les combinaisons entre modèles orthogonaux, au sens du cosinus, aboutiraient à moins d'interférences entre les modèles. De manière plus précise, les récents travaux de Ortiz-Jimenez *et al.* (2024) suggèrent que cette perte de performance suite à la combinaison des modèles est due au fait que le modèle ne respecte pas la propriété dite d'*arithmétique des tâches*, ce qui semble donc traduire que le régime d'entraînement du modèle n'est pas linéaire. Les auteurs proposent alors une méthode pour rendre linéaire le régime d'entraînement des modèles, afin de faciliter la combinaison des paramètres de ces derniers.

8. Nous justifions théoriquement cette égalité dans l'annexe A

Notre étude possède plusieurs limites. Tout d'abord, à une époque où le paradigme des modèles génératifs est dominant, nous n'avons traité que des tâches de classification relativement simples (3 classes au maximum) avec un modèle de type encodeur. Nous pouvons observer la simplicité des tâches traitées lors des observations de la section 3.1. Dans cette section, les interprétations sur les distances entre les tâches sont relativement similaires entre les modèles affinés complètement et les modèles affinés avec LoRA, suggérant donc que, avec un objet dont la dimension intrinsèque est faible, nous capturons essentiellement les mêmes informations qu'avec un vecteur de tâche en très grande dimension. Cet énoncé n'est pas à prendre pour résultat mais plutôt à mettre en regard avec la simplicité des tâches que nous avons considérée ici. En effet, nous supposons que l'on peut établir une corrélation entre le nombre de dimensions minimum qu'il faut donner à la méthode LoRA pour encoder les mêmes informations qu'un vecteur de tâche complet et la difficulté des tâches traitées : plus il faut de dimensions, plus la tâche est complexe. Une extension possible est alors de reprendre ces travaux avec des modèles de type encodeur/décodeur (Raffel *et al.*, 2020; Lewis *et al.*, 2020) ou décodeur pur (Radford *et al.*, 2019) et sur des tâches génératives (résumé, question réponse, *etc* ...) plus complexes qui demanderaient *a priori* plus de dimensions pour être modélisées correctement. Une seconde limite réside dans notre expérience sur les combinaisons des modèles. Nous nous limitons ici à un cadre très simple qui ouvre certaines pistes pour comprendre la combinaison dans un but de multi-tâches. Nous n'avons cependant pas évoqué le cadre des entraînements modulaires (combiner des modèles pour réaliser une nouvelle tâche). Dans ce cas, combiner des modèles éloignés peut effectivement avoir un intérêt. Il nous faut alors vérifier que le modèle sur la nouvelle tâche peut s'exprimer comme une combinaison linéaire des modèles initiaux. Dans un cas d'affinage complet, cela ne semble pas poser de problèmes étant donné que les vecteurs de tâches sont de très grandes dimensions et donc la probabilité de couvrir tout l'espace avec les combinaisons de ces vecteurs est grande. Cela devient moins évident dans un cas d'adaptation de rang faible où les vecteurs de tâches sont de très faible dimension (relativement à leur taille) et donc, couvrir tout l'espace avec de simples combinaisons linéaires devient moins probable. Avoir des vecteurs de tâches qui sont orthogonaux dans ce dernier cas permettrait d'augmenter cette dernière probabilité. Cette remarque suggère alors que la notion d'interférence entre les modèles lors de leur combinaison dépend du contexte dans lequel on se place. Dans notre étude, nous avons supposé que les interférences se mesuraient *via* la perte de performance dans une tâche sur laquelle il a été entraîné suite à une combinaison avec un autre modèle. Cette dernière ne semble pas en adéquation avec un cadre d'entraînement modulaire où on cherche à adresser une nouvelle tâche.

Références

AFRIAT S. N. (1957). Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. In *Mathematical proceedings of the Cambridge philosophical society*, volume 53, p. 800–816 : Cambridge University Press.

ALEEM S., DIETLMEIER J., ARAZO E. & LITTLE S. (2024). Convlora and adabn based domain adaptation via self-training. *arXiv preprint arXiv :2402.04964*.

ASGHAR N. (2016). Yelp dataset challenge : Review rating prediction. *arXiv preprint arXiv :1605.05362*.

BJÖRCK A. & GOLUB G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, **27**(123), 579–594.

BOWMAN S. R., ANGELI G., POTTS C. & MANNING C. D. (2015). A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, p. 632–642 : Association for Computational Linguistics (ACL).

CHRONOPOULOU A., PFEIFFER J., MAYNEZ J., WANG X., RUDER S. & AGRAWAL P. (2023). Language and task arithmetic with parameter-efficient layers for zero-shot summarization. *arXiv preprint arXiv :2311.09344*.

DETTMERS T., PAGNONI A., HOLTZMAN A. & ZETTLEMOYER L. (2024). Qlora : Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, **36**.

DIETTERICH T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, p. 1–15 : Springer.

ETHAYARAJH K. (2019). How contextual are contextualized word representations ? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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)* : Association for Computational Linguistics.

FALISSARD L., GUIGUE V. & SOULIER L. (2023). Apprentissage de sous-espaces de préfixes. In *18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 59–73 : ATALA.

FOSSE L., NGUYEN D.-H., SÉBILLOT P. & GRAVIER G. (2022). Une étude statistique des plongements dans les modèles transformers pour le français. In *Traitement Automatique des Langues Naturelles (TALN 2022)*, p. 247–256 : ATALA.

FOSSE L., NGUYEN D. H., SÉBILLOT P. & GRAVIER G. (2023). Géométrie de l'auto-attention en classification : quand la géométrie remplace l'attention. In *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, p. 137–150 : ATALA.

FRANKLE J., DZIUGAITE G. K., ROY D. & CARBIN M. (2020). Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, p. 3259–3269 : PMLR.

F.R.S. K. P. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572. DOI : 10.1080/14786440109462720.

FU Z., YANG H., SO A. M.-C., LAM W., BING L. & COLLIER N. (2023). On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, p. 12799–12807.

GANDIKOTA R., MATERZYNSKA J., ZHOU T., TORRALBA A. & BAU D. (2023). Concept sliders : Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv :2311.12092*.

GHOJOGH B. & GHODSI A. (2020). Attention mechanism, transformers, bert, and gpt : tutorial and survey.

GU Y., WANG X., WU J. Z., SHI Y., CHEN Y., FAN Z., XIAO W., ZHAO R., CHANG S., WU W. *et al.* (2024). Mix-of-show : Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, **36**.

HAMM J. & LEE D. D. (2008). Grassmann discriminant analysis : a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, p. 376–383.

HU E. J., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2021). Lora : Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

ILHARCO G., RIBEIRO M. T., WORTSMAN M., SCHMIDT L., HAJISHIRZI H. & FARHADI A. (2022). Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

JIN X., REN X., PREOTIUC-PIETRO D. & CHENG P. (2022). Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv :2212.09849*.

KADDOUR J., HARRIS J., MOZES M., BRADLEY H., RAILEANU R. & MCHARDY R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv :2307.10169*.

KOPICZKO D. J., BLANKEVOORT T. & ASANO Y. M. (2023). Vera : Vector-based random matrix adaptation. *arXiv preprint arXiv :2310.11454*.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880.

LI W., PENG Y., ZHANG M., DING L., HU H. & SHEN L. (2023). Deep model fusion : A survey. *arXiv preprint arXiv :2309.15698*.

LI X. L. & LIANG P. (2021). Prefix-tuning : Optimizing continuous prompts for generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 4582–4597.

LIAN D., ZHOU D., FENG J. & WANG X. (2022). Scaling & shifting your features : A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, **35**, 109–123.

LIU H., TAM D., MUQEETH M., MOHTA J., HUANG T., BANSAL M. & RAFFEL C. A. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, **35**, 1950–1965.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LUO S., TAN Y., PATIL S., GU D., VON PLATEN P., PASSOS A., HUANG L., LI J. & ZHAO H. (2023). Lcm-lora : A universal stable-diffusion acceleration module. *arXiv preprint arXiv :2311.05556*.

MAAS A., DALY R. E., PHAM P. T., HUANG D., NG A. Y. & POTTS C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies*, p. 142–150.

MATENA M. S. & RAFFEL C. A. (2022). Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, **35**, 17703–17716.

MIAO J. & BEN-ISRAEL A. (1992). On principal angles between subspaces in rn. *Linear algebra and its applications*, **171**, 81–98.

ORTIZ-JIMENEZ G., FAVERO A. & FROSSARD P. (2024). Task arithmetic in the tangent space : Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, **36**.

PFEIFFER J., RUDER S., VULIĆ I. & PONTI E. M. (2023). Modular deep learning. *arXiv preprint arXiv :2302.11529*.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). Language models are unsupervised multitask learners.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

REBUFFI S.-A., BILEN H. & VEDALDI A. (2017). Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, **30**.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2021). A primer in bertology : What we know about how bert works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866.

SAPORTA G. (2006). *Probabilités, analyse des données et statistique*. Editions technip.

SEDGWICK P. (2014). Spearman's rank correlation coefficient. *Bmj*, **349**.

SHAH V., RUIZ N., COLE F., LU E., LAZEBNIK S., LI Y. & JAMPANI V. (2023). Ziplora : Any subject in any style by effectively merging loras. *arXiv preprint arXiv :2311.13600*.

SUN Y., CHEN Q., HE X., WANG J., FENG H., HAN J., DING E., CHENG J., LI Z. & WANG J. (2022). Singular value fine-tuning : Few-shot segmentation requires few-parameters fine-tuning. *Advances in Neural Information Processing Systems*, **35**, 37484–37496.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. R. (2018). Glue : A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

YADAV P., TAM D., CHOSHEN L., RAFFEL C. & BANSAL M. (2023). Resolving interference when merging models. *arXiv preprint arXiv :2306.01708*.

YANG E., WANG Z., SHEN L., LIU S., GUO G., WANG X. & TAO D. (2023). Adamerging : Adaptive model merging for multi-task learning. *arXiv preprint arXiv :2310.02575*.

YU L., YU B., YU H., HUANG F. & LI Y. (2023). Language models are super mario : Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv :2311.03099*.

ZHANG J., LIU J., HE J. *et al.* (2024). Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, **36**.

A Retour sur la distance de Grassmann

Dans cette section, nous ajoutons des éléments de justifications au fait que si nous considérons deux couples de matrices LoRA (A_1, B_1) et (A_2, B_2), alors nous avons $d_g(B_1A_1, B_2A_2) = d_G(B_1, B_2)$. Comme énoncé dans le corps de l'étude, cette propriété découle directement du fait que dans un premier temps, nous avons $Im(B_iA_i) \subset Im(B_i)$ par définition, puis du fait que les matrices A_i sont de rangs plein [9], induisant ainsi que $Im(B_i) \subset Im(B_iA_i)$ (utilisation du théorème du rang) amenant donc à l'égalité $Im(B_iA_i) = Im(B_i)$.

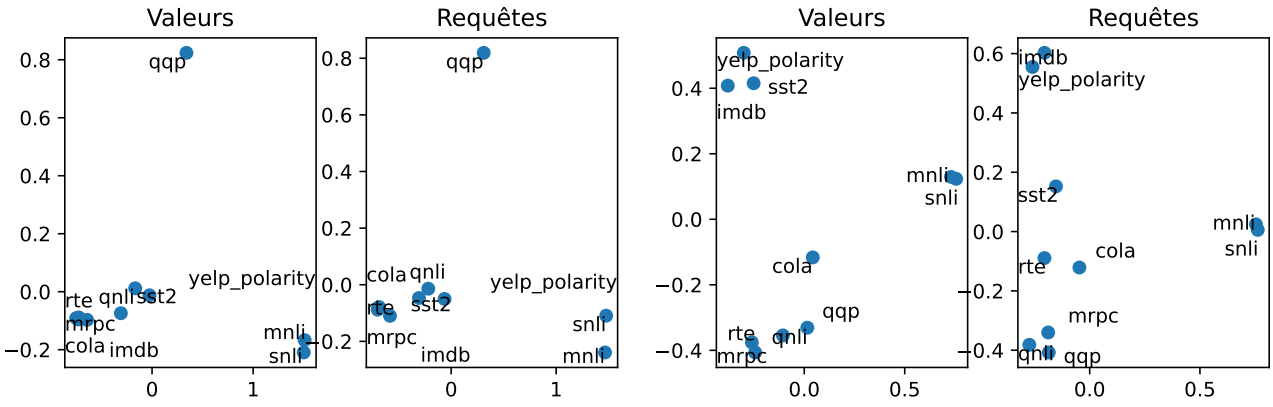
En effet, sachant cette égalité, pour calculer $d_G(B_1A_1, B_2A_2)$ nous devons dans un premier temps trouver, W_1 et W_2, des matrices dont les colonnes sont une base orthonormée de respectivement $Im(B_1A_1)$ et $Im(B_2A_2)$ (*c.f.* équation 1). Or d'après l'égalité ensembliste précédente trouver une base orthonormée de $Im(B_1A_1)$ (*resp.* $Im(B_2A_2)$) est équivalent à trouver une base orthonormée de $Im(B_1)$ (*resp.* $Im(B_2)$), montrant finalement ainsi que :

$$d_g(B_1A_1, B_2A_2) = d_G(B_1, B_2)$$

De cette égalité ainsi que de la proximité des résultats entre les distances du cosinus et de Grassmann, nous en déduisons que c'est principalement la matrice B qui encode des informations relatives aux tâches et non la matrice A dans les décompositions de faible rang. Afin de supporter ceci nous réalisons à nouveau l'opération de regroupement des modèles *via* la distance du cosinus, en séparant cette fois-ci l'analyse entre les matrices A et B, nous reportons les résultats sur la figure 4. Sur cette figure, nous remarquons, effectivement, que le regroupement effectué simplement avec les matrices B est relativement proche de celui observé avec le produit complet sur la figure 3b, tandis que celui observé sur les matrices A semble relativement mauvais pour l'association des différentes tâches. Appuyant ainsi empiriquement notre argument théorique.

Ce dernier point semble intéressant pour l'étude des vecteurs de tâches. En effet, une des difficultés de l'étude de ces vecteurs, est la dimension de ces derniers qui est très grande. Le fait de pouvoir supprimer une matrice réduit considérablement le nombre de paramètres à prendre en compte et réduit ainsi fortement la dimensionnalité. La réduction de dimension des vecteurs de tâches semble un point clé pour l'étude de ces objets. De plus en plus d'études apparaissent pour réduire le nombre de paramètres à mettre à jour dans les décompositions de faible rang (Kopiczko *et al.*, 2023; Sun *et al.*, 2022), permettant ainsi de venir diminuer fortement la dimension du vecteur de tâche. Une réduction de la dimension impliquant nécessairement une meilleure compréhension empirique de l'objet, et donc possiblement une meilleure combinaison de vecteurs de tâches *a posteriori*.

9. vérification empirique



(a) A (b) B

FIGURE 4 – Regroupement *via* la distance du cosinus en séparant l'analyse entre les matrices A et B

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

TAL et analyse de l'activité en ergonomie : extraction d'informations spécialisées dans des transcriptions d'entretiens

Andréa Blivet[1]
(1) SNCF DTIPG, 1-3 avenue François Mitterrand, 93210 Saint-Denis, France
andrea.blivet@gmail.com

RÉSUMÉ

L'ergonomie du travail est une discipline qui étudie les conditions de travail des individus. Son application se traduit, entre autres, par la réalisation d'entretiens d'analyse de l'activité qui ont pour objectif de faire émerger les impacts négatifs de la situation de travail sur la santé physique et morale des travailleurs. Notre étude consiste en la mise en place d'un système d'extraction automatique des informations relatives à ces impacts dans les transcriptions des entretiens réalisés. Ce système se fonde sur une approche hybride, associant ressources terminologiques et calcul de similarité contextuelle.

ABSTRACT

NLP and Activity Analysis in Ergonomics : Extraction of Specialized Information from Interview Transcriptions

The field of occupational ergonomics examines the working conditions of individuals, often through interviews analyzing their activities to identify negative impacts of the work environment on both physical and mental well-being. Our study focuses on implementing an automated system for extracting information pretaining to these impacts from interview transcriptions. The system adopts a hybrid approach, combining terminological resources with contextual similarity calculations.

MOTS-CLÉS : Extraction d'information, Marqueur, Ergonomie du travail, Corpus Spécialisé, Oral.

KEYWORDS: Information Extraction, Marker, Occupational Ergonomics, Specialized Corpus, Oral.

1 Extraction d'information en corpus spécialisé

En ergonomie du travail, les analyses de l'activité s'appuient en partie sur des interactions langagières, issues d'entretiens de terrain. Dans cette discipline, des approches linguistiques telles que l'analyse des pronoms personnels (Kebir *et al.*, 2021) ou encore l'analyse discursive des structures fonctionnelles des entretiens (Bonneau, 2008) ont déjà été utilisées pour proposer aux ergonomes de nouvelles manières de traiter les données d'entretien. Celles-ci nécessitent l'intervention de linguistes, et laissent aux ergonomes peu d'autonomie dans la réalisation de leurs analyses. Dans notre étude, nous proposons d'explorer les apports du TAL pour automatiser l'extraction des informations considérées comme pertinentes dans les données d'entretien tout en laissant la maîtrise de leurs analyses aux ergonomes. Les informations ciblées doivent être représentatives de la situation de travail de l'individu interrogé et s'inscrivent dans les thématiques suivantes (inspirées de la méthode ITMaMi [1]) : l'individu,

1. https ://bionet.scenari-community.org/Methodes_outils_PSE/co/ITMaMi.html

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 85–97.

les tâches qu'il effectue, le matériel qu'il utilise, l'organisation de son travail (soit l'organisation « humaine » et la planification des tâches), et son environnement physique. Définies de concert avec un ergonome expert, ces thématiques permettent à la fois de guider, structurer et filtrer l'extraction des informations qui y sont liées.

L'extraction d'information (EI) est une tâche classique du domaine du TAL, qui vise à dégager des segments textuels saillants pour un besoin donné en utilisant des méthodes de *text-mining* (règles lexico-syntaxiques, apprentissage automatique) ou des méthodes plus récentes en apprentissage profond (approches neuronales). Traditionnellement, les systèmes d'EI s'intéressent à la détection d'entités nommées (Nouvel *et al.*, 2015), à l'identification d'événements (Kodelja *et al.*, 2019), aux relations sémantiques entre les items (Wang, 2013). L'EI est utilisée dans divers domaines comme le domaine judiciaire pour l'analyse de dossiers d'enquêtes (Gianola, 2020) ou le domaine médical pour le traitement de compte-rendus médicaux (Lemaitre *et al.*, 2020), l'extraction de connaissances spécialisées (Trzmielewski & Gnoli, 2022) ou l'aide au diagnostic (Amato *et al.*, 2013).

Dans ces travaux, la délimitation des segments textuels considérés comme pertinents est définie en fonction du besoin informationnel. Les ressources (lexicales, ontologiques, etc.) et les outils (analyses syntaxiques, sémantiques, etc.) choisis pour leur extraction dépendent quant à eux de la nature des données dont sont extraits les segments. En effet, les données issues de l'oral sont particulièrement marquées par le caractère spontané de la situation d'énonciation, ne pouvant pas répondre à des normes aussi strictes qu'à l'écrit. L'oral spontané a ses propres spécificités syntaxiques, qui se traduisent par des structures moins rigides et moins explicitées que dans la forme écrite (Kurdi, 2003; Flamein, 2019). Le discours oral présente également des *extragrammaticalités*, comme des disfluences, des pauses ou encore des faux-départs, qui influent sur la prédictibilité du discours (Blanche-Benveniste *et al.*, 1990; Crystal, 2001; Kurdi, 2003) et en conséquence, sur le choix des outils et méthodes utilisées pour son analyse (Nasr & Béchet, 2009; Lacheret *et al.*, 2014). (Even, 2005) utilise par exemple des ontologies de termes spécifiques pour l'extraction d'information depuis des notes de l'oral quand (Tambellini, 2007) s'appuie sur des indices et des fenêtres contextuelles pour traiter des transcriptions d'entretiens d'embauche. Plus récemment, des approches neuronales ont été utilisées par (Parcollet *et al.*, 2019) pour l'identification de thèmes dans des conversations téléphoniques.

Dans notre étude, les informations ciblées par la tâche d'extraction sont dites « spécialisées » et se définissent comme des informations professionnelles (Chaudiron, 2000). Elles réfèrent à la fois au domaine ferroviaire et au domaine de l'ergonomie du travail, discipline guidant les analyses de l'activité. Les modèles d'extraction d'information classiques sur ce type de données, c'est-à-dire intégrant du vocabulaire technique ou des contraintes spécifiques à un domaine, présentent des lacunes avec une absence de données similaires à celles ciblées dans les données d'entraînement (Ramponi & Plank, 2020).

Sur la base des travaux mentionnés précédemment et des caractéristiques de nos données, nous proposons de décomposer la tâche d'extraction d'information en mettant au point un système capable de :

— **localiser** dans un texte les informations pertinentes à partir de ressources terminologiques,

— **délimiter** les informations afin de les rendre autonomes,

— **catégoriser** les informations selon les thématiques établies en amont,

— **suggérer** des informations qui pourraient être intéressantes pour l'utilisateur.

2 Corpus

2.1 Données

Les données que nous utilisons sont des entretiens d'analyse de l'activité réalisés en interne entre 2021 et 2023. Le corpus à disposition se compose de 52 entretiens oraux (soit 35h47 d'enregistrement) réunissant entre 2 et 5 locuteurs. Ces entretiens sont considérés comme relevant d'un genre et d'un domaine spécifiques pour plusieurs raisons. Tout d'abord, leur contexte de production correspond à des conversations dispensées à l'oral avec des intentions de communication spécifiques. Un locuteur (l'enquêteur) cherche à obtenir des productions langagières du type informationnel de la part d'un autre locuteur (l'agent) : les participants n'ont pas les mêmes intentions et l'échange peut être considéré comme asymétrique. Dans la typologie des échanges oraux proposée par (Walter, 1996), ces entretiens peuvent être considérés comme du « dialogue d'enquête ». L'autre spécificité concerne le contenu des entretiens : ces derniers contiennent des termes qui relèvent à la fois d'un domaine (le ferroviaire), et d'une discipline scientifique (l'ergonomie du travail). Les questions révèlent les enjeux de l'ergonomie : identifier des douleurs, des contraintes organisationnelles ou physiques, etc. Les réponses éclairent à la fois sur ces enjeux, et apportent également des éléments techniques sur le domaine ferroviaire, en décrivant des processus, ou des tâches par le biais d'outils utilisés, etc.

2.2 Pré-traitement des données

La première étape consiste en la transcription orthographique des enregistrements d'entretiens. Pour cela, les enregistrements sont d'abord segmentés en tours de parole à l'aide du système de reconnaissance de locuteur PyAnnote (Bredin *et al.*, 2020). Les tours de parole sont ensuite transcrits à l'aide du modèle Whisper [2] (Radford *et al.*, 2023), qui propose un premier niveau de normalisation en écartant de la transcription une partie des disfluences et en rétablissant certaines formes contractées (« aprem » devient « après-midi »). Enfin, un classifieur est entraîné pour reconnaître le rôle de chaque locuteur à savoir, enquêteur ou agent, à partir de descripteurs linguistiques (fréquence des pronoms, longueur des tours de parole, etc.). La Figure 1 représente l'ensemble des processus impliqués dans le pré-traitement et les données obtenues qui servent d'entrée au système d'extraction.

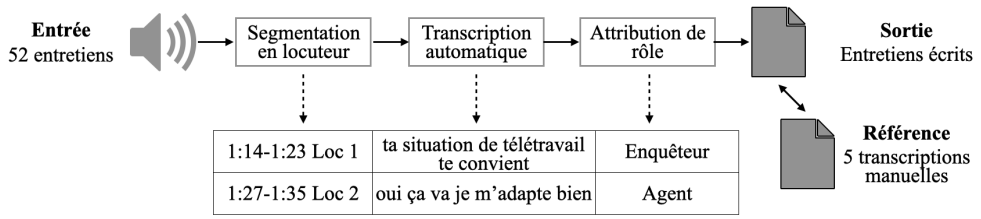


FIGURE 1 – Chaîne de traitement pour la transformation des données orales

En parallèle, 5 entretiens impliquant 2 locuteurs ont été transcrits manuellement pour servir de référence, représentant ainsi 41 minutes d'enregistrement et comportant 8280 tokens (dont 1166 différents) répartis en 418 tours de parole (soit 2% du corpus). Ainsi, la segmentation en locuteurs

2. Modèle Large-V2 https ://huggingface.co/openai/whisper-large-v2

présente 85,41% d'exactitude, le *Word Error Rate* obtenu pour la transcription est de 11,3% et enfin, le classifieur pour l'attribution des rôles des locuteurs obtient un F-Score de 0,94.

3 Méthodologie : mise en œuvre d'une méthode hybride pour optimiser l'extraction d'information

La localisation, la délimitation et la catégorisation des informations sont des tâches réalisées à l'aide d'une approche supervisée, s'appuyant sur les connaissances transmises par les experts ergonomes internes. Ainsi, le système extrait dans un premier temps des segments textuels considérés *a priori* comme pertinents. Dans un second temps, un système de suggestions, s'appuyant sur une méthode non supervisée, a été mis en place, afin de pallier le manque de connaissances incluses dans les ressources externes.

3.1 Approche supervisée : utilisation de ressources lexicales

L'approche supervisée consiste à extraire les informations en s'appuyant sur les connaissances transmises par les experts, capitalisées dans des ressources externes, selon une méthode collaborative déjà réalisée dans des tâches d'extraction sur des données médicales (Lemaitre *et al.*, 2020). Les ressources utilisées ici sont des lexiques thématiques de marqueurs lexicaux présentés en Table 1. Ils se composent de termes lemmatisés et identifiés par des experts comme des pointeurs de la mention d'une thématique dans l'entretien. Par exemple, lorsque l'agent interrogé mentionne son épaule, le terme « épaule » est considéré comme marqueur d'une information portant sur la thématique de l'*Individu* et est ensuite répertorié.

Thématiques	**Nombre de marqueurs**	**Exemples de marqueurs**
Individu	42	pénibilité, nuque, position, posture
Tâche	30	méthode, manutentionner, cercler, procédure
Matériel	74	casque, chaise, outil, tournevis
Organisation	52	pause, mécanicien, aiguilleur, polyvalent
Environnement	37	sol, température, bruit, place
Total	235	

TABLE 1 – Répartition et composition des lexiques par thématique

La présence d'un marqueur permet de localiser une information dans le texte, leur localisation repose sur une recherche de correspondance sur une version tokenisée et lemmatisée [3] de l'entretien. L'étape suivante consiste à délimiter le segment textuel qui correspond à l'information. D'après (Dupont *et al.*, 2002), une information est définie comme étant la combinaison d'entités ou d'événements avec leurs caractéristiques et relations. L'EI permet alors à formaliser les informations à partir d'analyses linguistiques comme l'analyse morphosyntaxique ou encore l'analyse sémantique compositionnelle. La stratégie que nous avons mise en place (cf.3 Partie A) repose sur cette hypothèse. Elle consiste dans un premier temps à récupérer les dépendants directs et indirects des marqueurs à partir d'une analyse

3. https ://spacy.io/models/fr - modèle 'fr_core_news_lg'

syntaxique proposée par Stanza [4] (Cao *et al.*, 2020). Concrètement, les marqueurs nominaux sont récupérés avec leurs dépendants directs et indirects (cf. Exemple a)), les adjectifs, quant à eux, sont mis en contexte en récupérant leur tête (cf. Exemple b)), enfin les arguments directs des marqueurs verbaux sont récupérés, soit les sujets et les compléments directs (cf. Exemple c)). Dans un second temps, les segments textuels extraits sont organisés selon les thématiques mentionnées en introduction. La thématique affectée à une information correspond à la thématique du marqueur ayant permis la localisation de cette même information.

a) ta **situation** de télétravail te convient ?
b) souvent on déplace des charges **lourdes**
c) je **ressentais** des douleurs pendant la journée

Cette approche se nourrit uniquement des ressources constituées par les experts. Celles-ci ne peuvent cependant pas être exhaustives. Une approche complémentaire doit donc être intégrée.

3.2 Approche à base de modèles non-supervisés : proximité distributionnelle

Une approche non-supervisée, basée sur la proximité distributionnelle, a été utilisée pour compléter l'extraction et pallier aux limites de l'approche supervisée en proposant une alternative aux ressources lexicales.

La proximité distributionnelle est un principe qui considère que deux termes avec des contextes similaires peuvent être considérés comme sémantiquement proches, indépendamment de leur relation lexicale (Fabre, 2015; Firth, 1957). Pour accéder aux contextes des termes, nous utilisons des plongements lexicaux (pré-entraînés de SpaCy [5]) qui sont des représentations vectorielles des termes calculés, entre autres, à partir de leur environnement linguistique. La représentation par plongement permet de comparer la similarité des contextes entre deux termes pour définir leur niveau de proximité sémantique en calculant leur cosinus. Pour déterminer si deux termes sont considérés comme proches, il faut définir un seuil qui correspond au cosinus minimal attendu. Pour cela, un clustering hiérarchique a été effectué pour regrouper les marqueurs proches dans chaque lexique thématique. Les scores de silhouette des classes obtenus ont permis de confirmer ou non leur homogénéité. Ensuite, la moyenne du cosinus de similarité entre tous les marqueurs d'une même classe est considérée comme le seuil minimal de proximité nécessaire pour considérer un terme nouveau comme sémantiquement lié aux autres.

La Figure 2 illustre le processus de sélection, opérant sur l'ensemble des noms, verbes et adjectifs, à partir de l'exemple du terme *scoliose*, absent des ressources lexicales. Le cosinus de similarité est calculé entre ce terme et tous les autres marqueurs. Si le cosinus maximal est entre *scoliose* et *séquelle*, le terme est considéré comme candidat marqueur pour le système de suggestion d'informations puisque le cosinus est supérieur au seuil déterminé pour la thématique Individu dont est issu le terme *séquelle*. En revanche, si le cosinus maximal observé est entre *scoliose* et *motrice*, le terme *scoliose* sera exclu de la suggestion car le cosinus est inférieur au seuil pour la thématique à laquelle appartient le terme *motrice*.

4. https ://stanfordnlp.github.io/stanza/models.html - modèle 'fr'
5. https ://spacy.io/models/fr - modèle 'fr_core_news_lg'

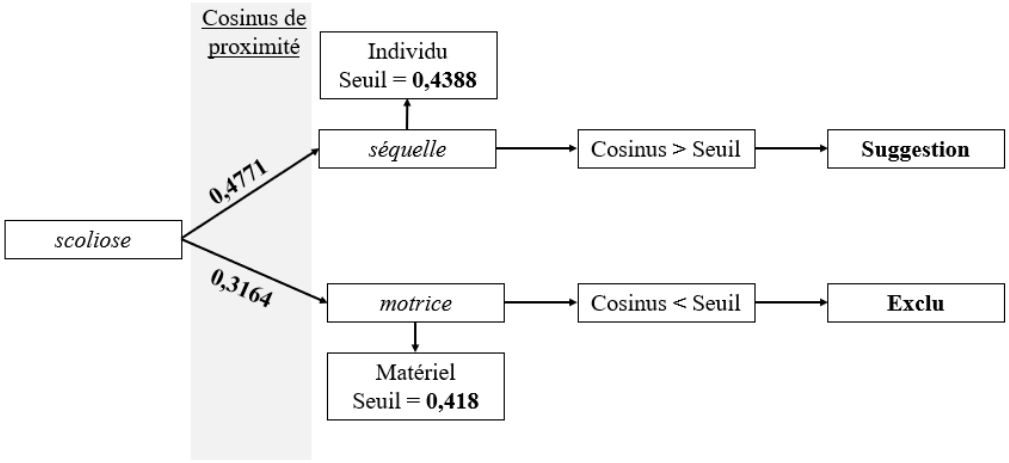


FIGURE 2 – Processus de sélection des marqueurs potentiels pour le système de suggestion

Les termes avec un score de proximité supérieur au seuil sont donc considérés comme des marqueurs potentiels. Le traitement appliqué est le même que pour les marqueurs avérés avec une délimitation des contextes à l'aide de l'analyse syntaxique (Figure 3 - Partie B). En revanche, contrairement aux informations obtenues à partir des ressources lexicales, un système de suggestion interactif est proposé à l'utilisateur pour qu'il puisse valider ou non la suggestion. Lui sont présentés, le marqueur, son contexte et la thématique associée (définie à partir de la thématique du marqueur avéré avec la proximité la plus élevée). L'utilisateur peut ensuite décider de valider l'information, de la rejeter ou bien d'attribuer une autre thématique. A la suite de cela, il est possible d'envisager un enrichissement de la liste des marqueurs en proposant un système d'amélioration continue, tel que présenté dans (Belkacem & Teissèdre, 2021). Ainsi, les marqueurs issus du système de suggestion qui auront été validés par les utilisateurs pourront être intégrés aux ressources.

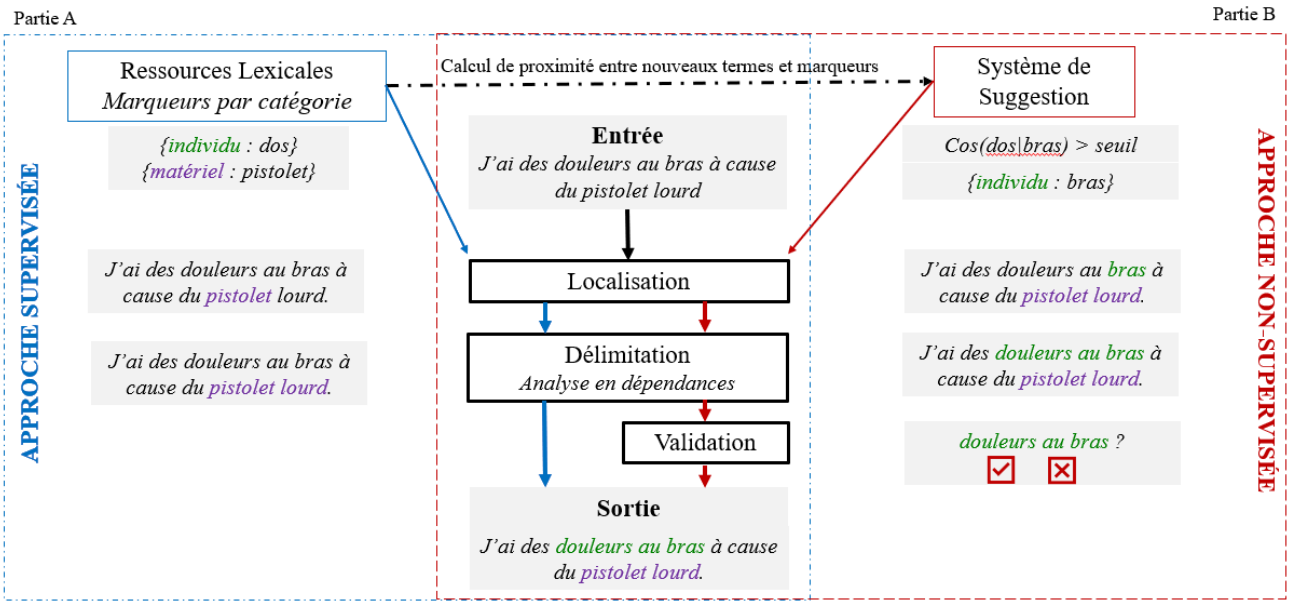


FIGURE 3 – Chaîne de traitement hybride pour l'extraction d'informations spécialisées - Partie A : Chaîne de traitement pour l'EI supervisée Partie B : Apport de l'approche non supervisée dans l'EI

4 Évaluation

L'objectif de l'évaluation est de déterminer les capacités du système à répondre aux attentes des utilisateurs et de mesurer l'exhaustivité de la double approche avec une confrontation des résultats de chacune.

Cinq experts SNCF ont été mobilisés pour participer à la phase d'évaluation. Les participants sont tous expérimentés et habitués des entretiens d'analyse de l'activité, ils ont également des connaissances approfondies sur l'organisation et les activités du groupe ferroviaire. L'évaluation est réalisée sur deux jeux de données. Le premier, illustré en Figure 4, se compose de cinq extraits issus de la transcription d'un même entretien. Ces cinq extraits ont ensuite été répartis entre les participants. Ils représentent au total 2901 tokens sur 124 tours de parole.

EN T'as dit que t'étais pas un grand sportif dans l'âme. Est-ce que tu aurais des restrictions d'aptitude ?

AG Non, pas de restriction.

EN Une limite de port de charge ?

AG Non, rien du tout.

EN T'as pas d'antécédent au niveau du dos ?

AG légère scoliose, voilà, pas une spécificité.

FIGURE 4 – Extrait de transcription à annoter

Le second jeu de données soumis aux évaluateurs se compose de 44 sorties du système de suggestions, correspondant à des informations potentielles détectées par le système, présentées avec leur contexte d'apparition et leur thématique assignée automatiquement (cf. Figure 5).

Suggestion	Contexte	Thématique
Douleur musculaire	je ressentais une petite douleur musculaire en bas du dos.	Individu

FIGURE 5 – Exemple de suggestion à évaluer

Dans un premier temps, les évaluateurs devaient relever dans un extrait de transcription d'entretien (premier jeu de données) les segments qui, selon eux, correspondent à des informations pertinentes pour représenter la situation de travail de l'individu interrogé. Ils devaient également assigner une ou plusieurs thématiques aux informations relevées. Dans un second temps, 20 suggestions issues de l'approche non-supervisée leur ont été présentées (second jeu de données). Pour chacune d'entre elles, les participants devaient : i) juger si elles renvoyaient à une information pertinente, le cas échéant ils devaient indiquer si la délimitation de cette dernière était correcte, en la corrigeant au besoin ; et ii) valider ou rectifier la thématique associée automatiquement à la suggestion.

L'évaluation prend ainsi en compte les sorties de l'approche supervisée et également l'ensemble de celles de l'approche non-supervisée, validant par défaut toutes les suggestions du système.

5 Résultats

Les résultats sont décomposés selon les tâches réalisées par le système, à savoir, localiser, délimiter, catégoriser et suggérer les informations, afin de les évaluer indépendamment. Cependant, la tâche de localisation a une influence sur la performance des autres tâches, puisque ces dernières sont réalisées à partir de la localisation. De plus, les approches mises en oeuvre ont un impact direct sur la localisation des informations. Ainsi, la Table 2 présente les résultats obtenus pour la localisation d'informations uniquement avec l'approche supervisée, puis combinée avec l'approche non-supervisée.

	Supervisée	**Supervisée + non-supervisée**
Précision	0,55	0,52
Rappel	0,32	0,67
F-score	0,40	0,58

TABLE 2 – Comparatif des résultats obtenus entre l'approche supervisée seule et combinée à l'approche non-supervisée

La différence notable dans ces résultats concerne le rappel, pour lequel l'alternative aux ressources lexicales de l'approche non-supervisée permet de le doubler, passant de **0,32** à **0,67**. A l'inverse, la précision est légèrement impactée par l'intégration de l'approche non-supervisée avec un score de **0,52** contre **0,55** avec uniquement l'approche supervisée. Dans la globalité, le *F-score* de **0,40**, obtenu uniquement avec l'approche supervisée, passe à **0,58** avec la combinaison des deux approches. Pour ce qui est de la délimitation, de la catégorisation et de la suggestion des informations, les résultats sont présentés en Table 3.

Tâches	**Taux d'exactitude**
Délimitation	5,82%
Catégorisation	55,28%
Suggestion	59,56%

TABLE 3 – Taux d'exactitude obtenus sur les différentes sous-tâches du système

Le taux d'exactitude pour la délimitation est de **5,82%**. Ce score très bas s'explique en partie par la méthode d'évaluation qui s'appuie sur une correspondance stricte, entre l'annotation experte et la sortie du système. Sur la totalité des erreurs de délimitation des informations, le placement de la frontière initiale de l'information est celle qui présente le plus de difficulté avec **56,47%** d'erreurs. A titre de comparaison, le placement de la frontière finale engendre une erreur dans **8,01%** des cas. Enfin, **20%** des erreurs de délimitation sont dues à un placement incorrect des deux frontières et prenant la forme d'informations enchâssées.

La catégorisation des informations, qui consiste à attribuer une thématique à une information retenue (indépendamment de la délimitation effectuée en amont), obtient un taux d'exactitude de **55,28%**. Le système assigne uniquement une seule thématique possible pour chaque information, tandis que les experts sont libres d'attribuer plusieurs thématiques. Les résultats ont permis de constater que seulement **3,01%** des désaccords s'expliquaient par cette différence de fonctionnement, c'est-à-dire les situations où l'expert attribue une thématique supplémentaire à celle sur laquelle il y a déjà un accord en l'expert et le système.

Enfin, la tâche de suggestion, uniquement réalisée par l'approche non-supervisée, présente un taux

d'exactitude de **59,56%**. Ce résultat signifie que presque 60% des suggestions proposées par le système sont considérées comme pertinentes par l'expert. Cette fonctionnalité participe donc largement à l'amélioration du rappel présenté plus haut.

Les annotations expertes permettent d'apporter des éléments de définition sur la notion d'information. En premier lieu, il a été observé que les informations se composent majoritairement de six tokens maximum. Seulement, **32%** des informations relevées comportent plus de six tokens. Ensuite, près de **90%** des informations expertes se terminent par un nom (*grosse semaine*), un verbe (*je me lève*) ou un adjectif (*charges lourdes*). En revanche, il est nécessaire de prendre en compte huit classes grammaticales pour représenter **90%** les frontières initiales d'informations (pronom, préposition, adverbe, déterminant, nom, verbe, conjonction et adjectif). Il apparaît donc que le début d'une information est soumise à une plus grande variabilité que sa fin. Enfin, les analyses montrent que quasiment l'intégralité des segments textuels considérés comme informations sont continus et que la plupart correspondent à des syntagmes nominaux ou a des triplets sujet-verbe-prédicat, comme « *une grande baie vitrée* » ou encore « *j'allonge ma pause repas* ». Cela confirme la stratégie de récupération des contextes basée sur l'analyse en dépendance.

6 Conclusion

Nous avons utilisé des méthodes d'extraction d'information supervisée (basée sur des ressources terminologiques expertes) et non-supervisée (avec de la similarité terminologique) dans l'objectif de proposer une structuration des informations abordées lors des entretiens d'analyse de l'activité. L'approche hybride que nous proposons permet d'obtenir des premiers résultats qui traduisent une certaine cohérence entre la méthode adoptée et les objectifs pré-définis. Cependant, ces résultats doivent être améliorés pour proposer par la suite un système utilisable et fiable.

6.1 Discussion

Les résultats de la délimitation des informations montrent que la méthode d'évaluation de cette tâche peut être repensée. En effet, une information est actuellement considérée comme validée par l'expert si la délimitation qu'il en fait est identique à celle du système au niveau des caractères. Cependant, cette méthode est trop contraignante et dévalue les capacités du système. En effet, l'information « *charges lourdes* » relevée par le système entraîne une erreur de délimitation quand l'expert a relevé « *des charges lourdes* ». Il apparaît pourtant dans cet exemple que l'intégration ou non du déterminant « *les* » influence peu, voire pas, la compréhension et l'interprétation sémantique de cette information. Pour améliorer la représentativité de l'évaluation sur le système, il est possible de proposer d'autres stratégies de comparaison entre deux segments, avec par exemple une prise en compte de la similarité à l'aide de coefficients tels que Dice (Cao *et al.*, 2020) ou Jaccard (Tapi Nzali, 2020), des métriques déjà utilisées en extraction d'information.

Toujours à propos de l'évaluation, certains extraits ont été annotés par différents experts. Il serait donc intéressant de prendre en compte le recouvrement de leurs annotations afin d'extraire un taux d'accord. En revanche, dans les cas de désaccords entre experts, il nous semble impossible d'être en mesure d'adjudiquer. Ainsi, l'ensemble des annotations seront tout de même considérées comme avérées, avec certaines qui pourront être qualifiées comme fiables lorsqu'elles font l'unanimité entre les experts.

Par ailleurs, l'approche non-supervisée permet d'améliorer les résultats et donc d'accroître les performances du système. De cette façon, le rappel est largement renforcé, passant de **0,32** à **0,67** sans pour autant impacter nettement la précision, qui perd uniquement 0,02 points. Ce rééquilibrage des scores est favorable aux attentes des utilisateurs qui ne souhaitent pas passer à côté des informations importantes (visant davantage le rappel), même si cela implique par la suite de trier certaines informations (impactées par la précision). Malgré la présentation des atouts de cette approche, il est important de prendre en compte certaines limites déjà identifiées. Dans un premier temps, l'approche non-supervisée doit être adaptée à la spécificité des données, en proposant, par exemple, un ré-entraînement des plongements sur des données spécifiques, afin de prendre en compte la terminologie du domaine ferroviaire et de l'ergonomie, une étape mise de côté par le manque de données. Dans un second temps, les seuils établis pour l'identification des marqueurs potentiels doivent être retravaillés pour être plus discriminants et limiter le nombre de suggestions à présenter. Sur le même principe, les marqueurs issus des ressources lexicales peuvent être pondérés selon leur productivité, c'est-à-dire leur capacité à identifier une information réellement pertinente. En effet, certains marqueurs issus des ressources lexicales ne sont pas toujours intégrés dans les informations relevées par les experts. Ainsi, il semble que certains d'entre eux aient davantage un statut d'indice. Nous envisageons donc de distinguer les marqueurs (productivité forte, les marqueurs suffisent, seuls, à identifier une information) des indices (productivité moyenne, les indices doivent être associés à d'autres éléments lexico-syntaxiques) et espérons que cela puisse favoriser la précision du système.

D'autre part, la collaboration et les interactions avec les experts ont guidé et influencé positivement le développement du système et de la méthodologie. Les experts ont manifesté leur intérêt pour intégrer un tel système dans leur pratique professionnelle. L'explicabilité du système et le rôle de l'utilisateur en tant que juge final ont été des déterminants de la confiance accordée au système proposé. Le système ne se place pas en remplaçant de l'enquêteur mais propose une réelle collaboration pour mettre à profit les capacités d'interprétations para-verbales et les habilités à mener un entretien de l'enquêteur.

6.2 Limites et perspectives

Les ressources lexicales contiennent au total 235 marqueurs répartis en 5 thématiques. Enrichir ces ressources lexicales serait une première piste à explorer pour garantir la fiabilité du système et améliorer le calcul des seuils du système de suggestion. Cependant, faire appel aux experts pour obtenir de nouveaux marqueurs est un processus coûteux et compliqué à mettre en place. Ainsi, nous envisageons d'explorer les patrons syntaxiques autour des marqueurs identifiés pour faire émerger les structures lexico-syntaxiques propices à la présence de marqueurs.

Actuellement, nous nous sommes principalement concentrés sur le contenu textuel des données. Pour la suite, la spécificité concernant le contexte de production (l'oral) doit être envisagée. Par exemple, les structures grammaticales spécifiques de l'oral pourraient être prises en compte pour la délimitation des informations qui se base sur une analyse syntaxique des structures standards de l'écrit. Sur ce même principe, une étape de normalisation des formes contractées ou spécifiques de l'oral peut être intégrée au moment du pré-traitement des données orales afin d'accroître la possibilité de correspondance entre ces formes et leur représentation vectorielle dans les plongements lexicaux pré-entraînés. La modalité orale permet également de laisser la possibilité d'analyser les disfluences ou les hésitations langagières afin de proposer, dans le futur, des indicateurs sur le doute ou la confiance exprimés sur des informations énoncées.

Pour poursuivre ce premier travail, nous souhaitons proposer un système qui soit en mesure de s'adapter aux ressources lexicales et aux thématiques. Nous visons ainsi la mise en place d'un système qui soit utilisable dans d'autres cas d'usage ou d'autres disciplines, ayant recours à des entretiens semi-directifs.

Remerciements

Je tiens à remercier Hèlene FLAMEIN et Luce LEFEUVRE pour leur accompagnement et encadrement tout au long de ces travaux. Je remercie également Yonnel GIOVANELLI, expert ergonome à la SNCF à l'initiative de ce projet, ainsi que tous les experts en analyse de l'activité qui ont apporté leur contribution.

Références

AMATO F., LÓPEZ A., PEÑA-MÉNDEZ E. M., VAŇHARA P., HAMPL A. & HAVEL J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, **11**(2), 47–58. DOI : https ://doi.org/10.2478/v10136-012-0031-x.

BELKACEM T. & TEISSÈDRE C. (2021). Outil interactif et évolutif pour l'extraction d'information dans des documents techniques. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3 : Démonstrations*, p. 12–14, Lille, France : Association pour le Traitement Automatique des Langues. HAL : hal-03265912.

BLANCHE-BENVENISTE C., BILGER M., ROUGET C., VAN DEN EYNDE K., MERTENS P. & WILLEMS D. (1990). *Le français parlé (études grammaticales)*. Sciences du Langage.

BONNEAU J. (2008). Outils d'aide à l'exploitation d'entretiens semi-directifs : Etude de l'interaction entre intervieweur et interviewés sur un corpus ethnoécologique. In B. PINCEMIN & S. HEIDEN, Éds., *JADT 2008*, p. 219–232, ENS Lettres et sciences humaines, France : Presses Universitaires de Lyon. HAL : hal-01362690.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). Pyannote.audio : neural building blocks for speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain. HAL : hal-02995345.

CAO D., BENAMAR A., BOUMGHAR M., BOTHUA M., OULD OUALI L. & SUIGNARD P. (2020). Participation d'EDF R&D à DEFT 2020. In R. CARDON, N. GRABAR, C. GROUIN & T. HAMON, Éds., *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, p. 26–35, Nancy, France : ATALA. HAL : hal-02784739.

CHAUDIRON S. (2000). Nouveaux modes d'intermédiation de l'information spécialisée. In C. E. S. P. D. NITITENKO, Éd., *La publication en ligne*, p. 153–165. Hermès. HAL : hal-02568788.

CRYSTAL D. (2001). *Language and the Internet*. Cambridge University Press. DOI : 10.1017/CBO9781139164771.

DUPONT M., VUILLAUME J.-M., VICTORRI B., ENJALBERT P., MATHET Y. & MALANDAIN N. (2002). Nouvelles perspectives en extraction d'information. *Revue des Sciences et Technologies*

de l'Information - Série TSI : Technique et Science Informatiques, **1**(21), 37–63. HAL : halshs-00009485.

EVEN F. (2005). *Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale*. Theses, Université de Nantes. HAL : tel-00109400.

FABRE C. (2015). Sémantique distributionnelle automatique : la proximité distributionnelle comme mode d'accès au sens. *Études de linguistique appliquée : revue de didactologie des langues-cultures*, (4), 395–405. HAL : hal-02074874.

FIRTH J. R. (1957). Applications of general linguistics. *Transactions of the Philological Society*, **56**(1), 1–14. DOI : https ://doi.org/10.1111/j.1467-968X.1957.tb00568.x.

FLAMEIN H. (2019). *Etude de la perception d'une ville : Repérage automatique, analyse et visualisation*. Theses, Université d'Orléans. HAL : tel-04429419.

GIANOLA L. (2020). *Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique*. Theses, Université de Cergy-Pontoise. HAL : tel-02522680.

KEBIR Y., DELSART A., ARFAOUI S., AURIAC-SLUSARCZYK E. & SAINT-DIZIER DE ALMEIDA V. (2021). L'apport de la linguistique à l'ergonomie pour enrichir l'analyse de l'activité de consultation de suivi médicale. In *55ème Congrès de la SELF, L'activité et ses frontières. Penser et agir sur les transformations de nos sociétés*, p. 74–79, Paris, France. HAL : hal-02541962.

KODELJA D., BESANÇON R. & FERRET O. (2019). Modèles neuronaux pour l'extraction supervisée d'événements : état de l'art [neural models for supervised event extraction : state of the art]. *Traitement Automatique des Langues*, **60**(1), 13–37.

KURDI M.-Z. (2003). *Contribution a l'analyse du langage oral spontané*. Theses, Université Joseph-Fourier - Grenoble I. HAL : tel-00005071.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. **8**, 2675–2689. DOI : 10.1051/shsconf/20140801305, HAL : halshs-01061368.

LEMAITRE T., GOSSET C., LAFOURCADE M., PATEL N. & MAYORAL G. (2020). Deft 2020 - extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance. In *Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes*, p. 55–65, Nancy, France : Association pour le Traitement Automatique des Langues. HAL : hal-02784742.

NASR A. & BÉCHET F. (2009). Analyse syntaxique en dépendances de l'oral spontané. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 21–30, Senlis, France : ATALA.

NOUVEL D., EHRMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE Group.

PARCOLLET T., MORCHID M. & LINARÈS G. (2019). Réseaux de neurones convolutifs de quaternions pour l'identification de thèmes de conversations téléphoniques. In *Conférence en Recherche d'Infomations et Applications*.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, p. 28492–28518, Honolulu, Hawaii, USA : JMLR.org.

RAMPONI A. & PLANK B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on*

Computational Linguistics, p. 6838–6855, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.603.

TAMBELLINI C. (2007). *Un système de recherche d'information adapté aux données incertaines : adaptation du modèle de langue*. Thèse de doctorat. Thèse de doctorat dirigée par Berrut, Catherine Informatique Grenoble 1 2007.

TAPI NZALI M. (2020). DEFT 2020 : détection de similarité entre phrases et extraction d'information. In R. CARDON, N. GRABAR, C. GROUIN & T. HAMON, Éds., *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, p. 91–96, Nancy, France : ATALA. HAL : hal-02784745.

TRZMIELEWSKI M. & GNOLI C. (2022). Organisation des connaissances médicales : principes et recherche récentes. *Pratiques d'information et connaissances en santé*, p. 173–210.

WALTER H. (1996). Dialogue : unités lexicales et analyse du sens. *Onomázein : Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, (1), 167–179.

WANG W. (2013). *Extraction d'Information Non Supervisée à Partir de Textes – Extraction et Regroupement de Relations entre Entités*. Theses, Université Paris Sud - Paris XI. HAL : tel-00998391.