CLAfICLe: Cross Lingual Adaptation for In-Context Learning

Giulio Starace
University of Amsterdam / Amsterdam, The Netherlands
giulio.starace@gmail.com

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

contributions

- successfully apply WECHSEL to GPT2 Large, release the checkpoints which did not exist
- More complete evaluation of WECHSEL in the CLM setting
- propose method for preserving FT when performing cross-lingual adaptation
- formalize concept of vessel adapters with targeted distillation, a form of post-hoc disentanglement

2 Related Work

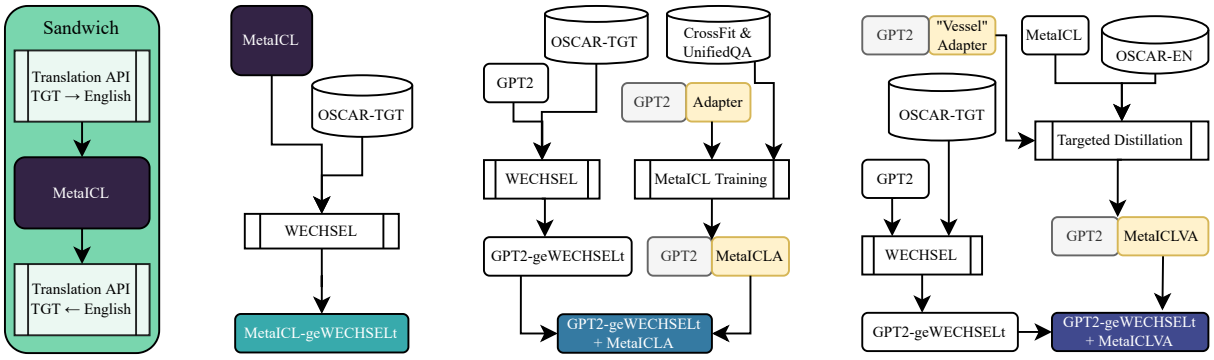
3 Method

4 Results and Discussion

Fig. 2 shows the performance of GPT2 after around 1k steps of training, evaluated intrinsically in terms of perplexity. For both French and German, we see perplexity decrease to sub-50 values, with the French model reaching a perplexity of ≈ 28. Both models are clearly underfit, still monotonically decreasing by the end of the training. These observations are roughly in-line with Minixhofer et al. (2022)'s findings for smaller variants of GPT2, although we train for much less time and hence are left with higher perplexities. While we believe our preliminary results suggest WECHSEL scales well to larger models in terms of intrinsic evaluation, future work may wish to investigate whether this holds for longer training times. The rest of our work considers, among other questions, the robustness of WECHSEL via extrinsic evaluation on downstream tasks performed by MetaICL.

Fig. 3 shows the performance on each dataset of our benchmark for the two baseline models, MetaICL and Sandwich. As summarized in Table 1, Sandwich performs roughly on-par with MetaICL on both target languages, respectively with scores of 0.317 and 0.322 in French and German compared to MetaICL's score of 0.327 in English. We note generally low scores across all tasks. This is particularly perplexing in the case of MetaICL, scoring around 0.1 points less than with the evaluation ensemble used by Min et al. (2022), where the same checkpoint was reported scoring 0.417 in the worst case (a 25 % decrease). While similar values are reached in certain tasks in our benchmark (e.g. most of XGLUE and WINO-X), it is unclear what the origin of this discrepancy is, whether due to differences in evaluation implementation or difficulty of the tasks. Given that Min et al. (2022) simply report macro-averaged scores, it is impossible to verify the latter. Nevertheless, our results suggest that Sandwich-like solutions may be satisfactory for transferring performance from English to other languages given the surprisingly closeness of the scores. The decision between using Sandwich or "properly" adapted models with the same capabilities then becomes an economic one in terms of the cost of API calls (for the former) vs the cost of inference plus training (for the latter).

Fig. 4 shows the difference in performance on each dataset of our benchmark between the proposed models and Sandwich. In general, we observe that the proposed models underperform across almost all tasks in both French and Ger-



(a) (b) (c) (d)

Figure 1: Overview of each of the models evaluated in one of the two TGT languages (French or German). The baseline Sandwich model (a) sandwiches MetaICL (Min et al., 2022) (which we separately evaluate only in English) between two complementary translation API calls. MetaICL-geWECHSELt (b) is the result of applying WECHSEL (Minixhofer et al., 2022) to MetaICL. GPT2-geWECHSELt+MetaICLA combines MetaICLA, an adapter trained on the MetaICL dataset and objective, with a TGT-language GPT2 base obtained via WECHSEL. GPT2-geWECHSELt+MetaICLVA does the same, except MetaICLVA is trained via targeted distillation with supervision provided by MetaICL. For more details, refer to section 3.

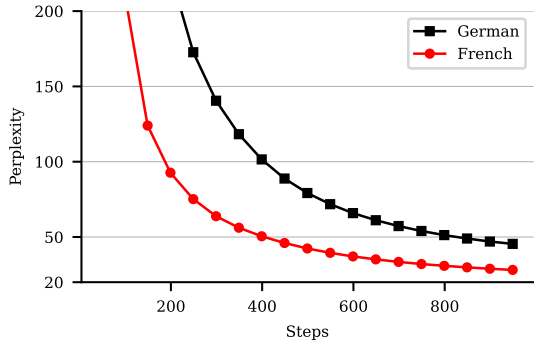


Figure 2: Perplexity on the held out set when performing the recommended CLM training after WECHSEL language-adaptation of GPT2. A step corresponds to an optimizer update. We evaluate every 50 steps.

man, with the trends aligning at a task-level (e.g. all models underperform on QAM, by roughly the same amount). As reported in Table 1, the best of our proposed models is MetaICL-geWECHSELt, which underperformed Sandwich by roughly 0.02-0.03 points. This undermines the motivation for the other two models, which were designed to avoid catastrophic forgetting by separating language and ICL capabilities via adapters. The results suggest that the tradeoff between catastrophic forgetting and needing to train ICL-adapters leans in favour of the former in this compute regime. In this sense, we can conclude that WECHSEL does not suffer tremendously due to catastrophic forgetting when adapting fine-tuned causal language models such as the MetaICL variant of GPT2.

Future work should

- todo

Table 1: Average performance (max is 1) across the datasets from our multi-task benchmark for the models considered in this work. We use "W" as a shorthand for "geWECHSELt". We report average difference in performance for each proposed alternative to Sandwich. Negative values indicate underperformance compared to Sandwich.

	en	fr	de
MetaICL	0.327	-	-
Sandwich	-	0.317	0.322
Difference in Performance w.r.t. Sandwich			
MetaICL-W	-	-0.020	-0.026
GPT2-W+MetaICLA	-	-0.041	-0.042
GPT2-W+MetaICLVA	-	-0.036	-0.045

5 Conclusion

References

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to Learn In Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of*

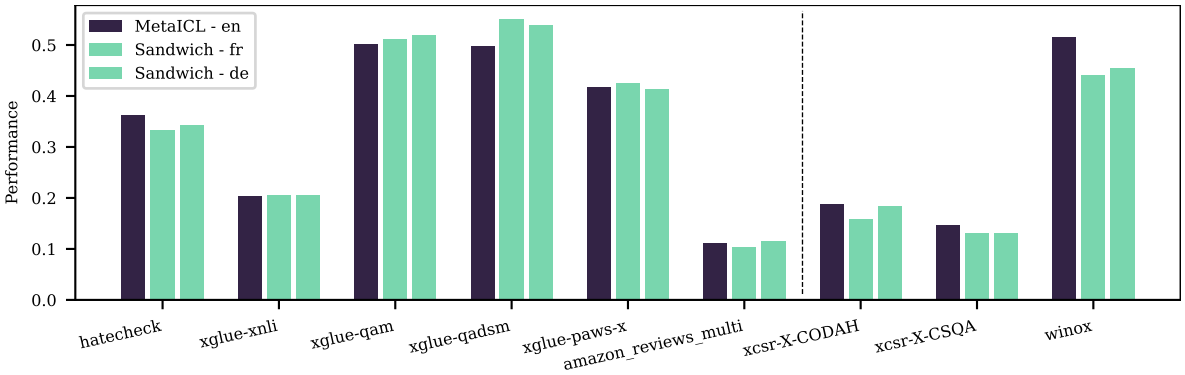


Figure 3: Performance (max is 1) on a particular language dimension of our multi-task benchmark of our two baseline models, MetaICL and Sandwich. The dashed line separates whether a given task uses accuracy (left) or F1-score (right) as the performance metric.

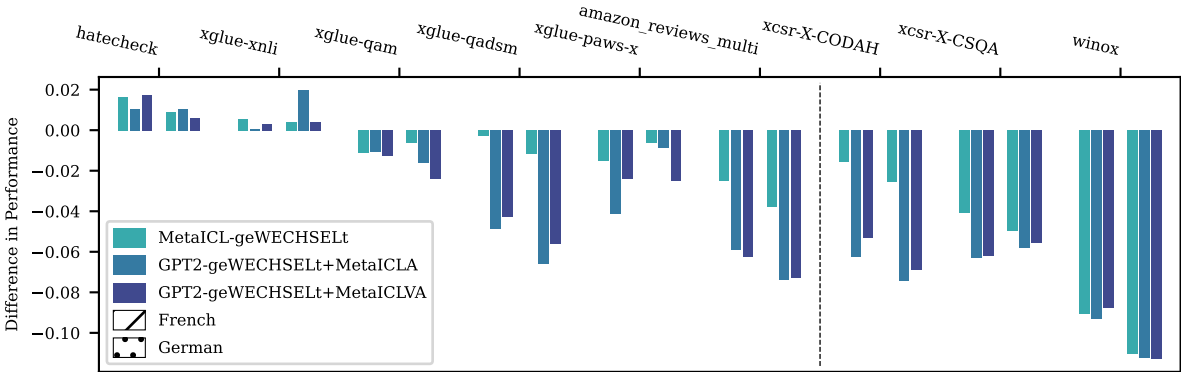


Figure 4: Performance gap on our multi-task benchmark between each of the language-adapted models and the "Sandwich" baseline. Positive values indicate that the adapted models are outperforming the baseline, while negative values indicate the reverse. The dashed line separates whether a given task uses accuracy (left) or F1-score (right) as the performance metric.

the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

6 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

A Example Appendix

This is an appendix.