



HAL
open science

Language Portability Strategies for Open-domain Dialogue with Pre-trained Language Models from High to Low Resource Languages

Ahmed Njifenjou, Virgile Sual, Bassam Jabaian, Fabrice Lefèvre

► **To cite this version:**

Ahmed Njifenjou, Virgile Sual, Bassam Jabaian, Fabrice Lefèvre. Language Portability Strategies for Open-domain Dialogue with Pre-trained Language Models from High to Low Resource Languages. The 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS '23), Feb 2023, Los Angeles, United States. hal-04631021

HAL Id: hal-04631021

<https://hal.science/hal-04631021>

Submitted on 1 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Language Portability Strategies for Open-domain Dialogue with Pre-trained Language Models from High to Low Resource Languages

Ahmed Njifenjou, Virgile Sucas, Bassam Jabaian and Fabrice Lefèvre

Abstract In this paper we propose a study of linguistic portability strategies of large pre-trained language models (PLMs) used for open-domain dialogue systems in a high-resource language for this task. In particular the target low-resource language (L_T) will be simulated with French, as it lacks of task-specific resources and allows our human evaluation, when the source language (L_S) is English. For obvious reasons, recent works using such models for open-domain dialogue are mostly developed in English. Yet building specific PLMs for each possible target language supposes collecting new datasets and is costly. For this reason, trying to leverage all existing resources (PLMs and data) in both L_S and L_T , we wish to assess the performance achievable in L_T with different approaches. The first two approaches evaluate the usage of Neural Machine Translation (NMT) at different levels: TrainOnTarget where a L_S dataset is translated before fine-tuning in L_T and TestOnSource where a L_S model is coupled with NMT modules during inference. Then, the advent of BLOOM [2], the world first **open-access** multilingual large PLM, allow researchers to develop new approaches aiming to leverage not only the model’s full accessibility but also its multilingualism and translation abilities. In this context the task is learned in L_S first and adapted to L_T using the MAD-X Adapter architecture [16]. In the two sets of experiments models are evaluated in spoken dialogue conditions with human and the strategies can be compared in terms of perceived interaction quality.

Ahmed Njifenjou, Virgile Sucas, Bassam Jabaian, Fabrice Lefèvre
LIA/CERI, Avignon Université, France

{ahmed-ndouop.njifenjou, virgile.sucas, bassam.jabaian, fabrice.lefevre}@univ-avignon.fr

1 Introduction

Since the breakthrough of transformers [23], the field of Natural Language Processing (NLP) has been fuelled with many variants of PLMs. Auto-regressive transformers, those using the decoder part of the transformers, like GPT [17], BART [12] and derivatives helped to improve the state-of-the art of several generative tasks. Among these open-domain dialogue systems, aka chatbots, which as stated in [24] should develop some human abilities like empathy, personality and entertainment to socially engage with users. For this extent, specific crowd-sourced datasets have been created (such as PersonaChat [26], Empathetic Dialogues [19], Blended Skill Talk [22] etc.) on which PLMs are fine-tuned. Unfortunately, most of these resources are in English. Even French which is generally not considered as a low-resource language lacks of specific datasets for this task.

In this work we study portability strategies of chitchat models and datasets from a source language (L_S , English here) to a target language (L_T , French here), with the broad target of later application to truly low-resource language. Making the most of resources available in both L_S and L_T - NMT tools, datasets and PLMs models - we set up and conducted a human evaluation of different systems obtained by combining those resources in three different approaches, described hereafter, varying their usage of the resources: training on target, testing on source and training on source then adapting to target. All the models are compared to an overall reference model in L_S ; BlenderBot 1.0 [20] a much advanced system in terms of architecture and datasets is retained.

2 Related Work

In previous team works on domain and language portability of spoken dialogue systems [8, 11], we assessed language portability of Spoken Language Understanding module of a goal-oriented dialogue system. A Statistical Machine Translation (SMT) module, state-of-the-art back then, was used in two approaches efficient and low-cost for language portability. Here we assess these strategies with state-of-the-art NMT modules on transformers-based open domain dialogue systems.

Coming as one of the rare works that tackled the issues of monolingual resources development for open domain chatbot, [13] is the closest work to our current study - to the best of our knowledge. Indeed they also worked with PersonaChat and provided its translated versions in French and five additional languages. Furthermore they assessed multilingual and cross-lingual approaches with a specific target on multilingual models. They also evaluated a model using translation before and after a L_S model inference, however with a model with a different architecture from the others. But automatic translation of PersonaChat were revised by native L_T speakers that are fluent in L_S , which is costly and cannot be afforded for some languages where bilingual speakers can be scarce. Besides, their cross-lingual model had poor results compared to others. Also our strategies implement another approach, already

used in [16] to perform cross-lingual transfer learning on a bunch of NLP tasks but not including generative tasks. They used a modular adapters framework, baptised MAD-X, combined to a multilingual PLM. The latter is frozen while its blocks are augmented with unfrozen language and task adapters. These adapters are then trained sequentially: first the language adapters which in turn are frozen before the task adapters training. This approach can be less expensive in terms of training cost when adding a new language support. Also with task adapters primarily trained on L_S original data rather than their noisy auto-translations counterparts, it may help improve performance.

For dialogue modeling, we used the same training framework as [25] on PersonaChat. Model inputs here are a concatenation of model-assigned personality traits, dialogue history and the “golden” reply it learns to mimic. In addition to what they did with a GPT [17] model, we tried the same approach in French and with the BLOOM [2] model which is an open access and multilingual model.

3 Portability Strategies Assessed

In this preliminary study of our ongoing work on French and other languages lacking specific data for open-domain dialogue, rather than focusing on intrinsic dialogue performance improvement, we assess how the data and models from L_S can be leveraged to develop shallow conversational models based on transformers in L_T .

3.1 *TestOnSource and TrainOnTarget*

Inspired from our previous work on cross-lingual Spoken Language Understanding [8, 11], these two approaches rely on the usage of NMT modules at different stages. Recent advances in this field provide us with ready to use high quality translation tools for L_S and L_T . The DeepL tool claimed to be the best performing¹ is only accessible through a paying (for more than 500K characters) API, hence we used Google Translate API which is the same API used in [13].

Test On Source

The large amount of resources including models for open-domain dialogue in L_S is a major asset. Consequently, it is interesting to evaluate how well these systems perform on inputs translated from L_T to L_S at inference. Hence the approach merely consists in using existing datasets, PLM and open-domain dialogue models available in L_S and combine them with NMT system during bot-human conversations in L_T as shown in Fig. 1.

¹ <https://www.deepl.com/en/quality.html>



Fig. 1: **Illustration of the TestOnSource approach:** an additional L_T to L_S (orange) and L_S to L_T (green) NMT modules around a L_S model fine-tuned for chitchat

Train On Target

While lacking open-domain dialogue specific datasets, the L_T in this study - French - has at its disposal a bunch of PLMs that can be used as a basis for a dialogue system. Train-OnTarget, illustrated in Fig. 2, consists in fine-tuning adequate L_T PLMs (bottom green) on chitchat task using an automatically translated dataset from L_S (top yellow). Despite being subject to noise injection in data, forward translation can still be effective for low-resource MT in some contexts [7] and we assume that the language specific abilities learned by the L_T PLMs can help handle the noisy NMT samples.

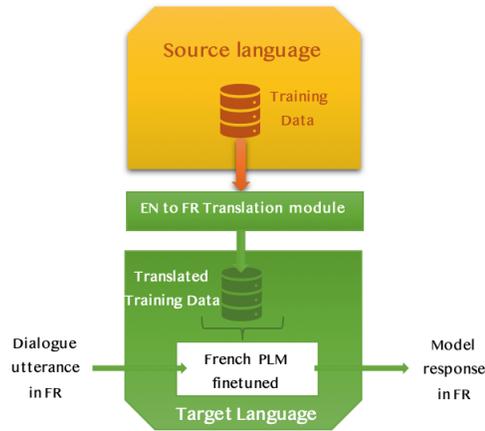


Fig. 2: **Illustration of the TrainOnTarget approach**

3.2 TrainOnSourceAdaptOnTarget: leverage multilingual PLMs

The two previous approaches rely heavily on the fact that outside of chitchat task L_T is not under-resource having available NMT tools and PLMs. Hence, the idea of using multilingual PLMs to not exclude the large majority of lower-resource languages.

We reproduce the MAD-X architecture [16] for dialogue using **BLOOM** which has translation abilities over a large set of low-resource languages. This is interesting as it can help translate subset of L_S datasets which will then be used for dialogue-task adapter’s few-shot tuning. The latter being firstly finetuned on human generated data in L_S rather than directly on NMT resulting data (as in 3.1) may also be a good asset. Fig. 3 shows the overall workflow of this approach.

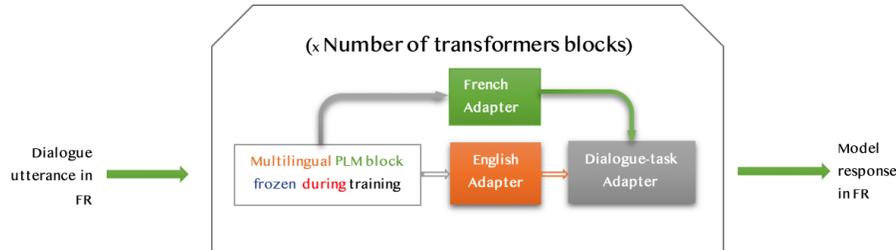


Fig. 3: A transformer block in **TrainOnSourceAdaptOnTarget** configuration. Empty arrows show the first fine-tuning stage of the task adapters (on L_S data) with L_S language adapters frozen. In the second stage (filled arrows) language adapters (still frozen) are switched from L_S to L_T and the same task adapters (assumed language agnostic [16]) are fine-tuned using few (or translated) L_T data. Prior to these stages, both L_S and L_T language adapters are trained independently on causal LM with transformers block frozen.

4 Experimental Setup

Having the previous portability approaches in mind, we trained and evaluated different PLMs from L_S and L_T and eventually multilingual with the recent advent of **BLOOM**. The target was first to assess the most efficient approach with the current resources but also to compare to the previous results on similar approaches.

4.1 PersonaChat Dataset

To train and evaluate the different models we chose the PersonaChat [26] dataset described in Table 1. It consists of a set of dialogues between two humans in English, each being assigned a personality defined by a few sentences. The aim of this dataset is to help build models with consistent persona throughout a conversation.

For French-based models, the dataset is automatically translated with the `Google Translate API`².

However, contrary to [13], no human re-annotation was performed after auto-translation as they reported a BLEU of **94.19** while assessing the difference between auto-translated and human corrected data in French. Also we didn't re-use their corrected dataset in order to mimic the case of a language where no native speakers fluent in English can be hired to correct translation errors.

Table 1: Statistics of the PersonaChat dataset used for models' finetuning

Language	Split	Dialogues	Utterances
English (L_S)	Train	14917*	109718
	Validation	100	773
	Test	1000	7801
French (L_T)	Train	14917	109759
	Validation	100	773
	Test	1000	7801

* During translation some problematic conversations were dropped, so for L_S we took the same training size as the L_T version to be fair to the corresponding trained models. The original dataset has more than 17K dialogs.

4.2 Models' Descriptions

We consider open-domain dialogue here as a generative task, hence we focused on leveraging auto-regressive models available in both L_S and L_T .

Monolingual-PLM based models For *TrainOnTarget* we fine-tuned the small version of **GPT-fr** [21] (124M parameters), a French version of GPT-2 [18] as it is close in terms of architecture and size to the GPT-based TransferTransfo [25] model (117M parameters) that will be used for monolingual *TestOnSource*.

Multilingual-PLM based models The model used here is **BLOOM** a brand new large Multilingual PLM accessible to researchers. We worked on its thinnest version available which has 560M parameters. Multilingual PLMs are the only to allow the third approach, hence we built a L_S and L_T models using MAD-X adapters architecture on BLOOM. The resulting models are reported as *madx-BLOOM* in Table 2. As the model is multilingual, we also built a L_S and L_T models for the first two approaches i-e without using adapters mixing across the languages which are reported as *BLOOM* (L_T) and *BLOOM* (L_S) in Table 2.

² <https://translate.google.fr>

4.3 Training Details

All the models were trained using a double-heads architecture as in [25]: a Causal Language Modeling head and a multi-choice head. The former, had a higher weight on the combined loss as we assumed the dialogue to be mainly a generative task.

The GPT-fr based model, as its L_S counterpart [25], was fine-tuned for one epoch with AdamW optimizer and a linear-decreasing learning rate of $6.25e-5$ on PersonaChat translated. Both BLOOM_fr and BLOOM_en were trained with that same learning rate for 5 epochs with evaluation performed every quarter of an epoch and the 5 best checkpoints with lowest perplexity kept. Then the best performing checkpoint on the whole test set was retained.

For the models with the adapter mixing architecture, we first trained the L_S and L_T language adapters on Wikipedia³ with a total batch size of 80 and a relatively high initial learning rate of $1e-4$ following [16] for 379K and 427k steps respectively (one week on five V100). As the validation perplexity was still decreasing, we used the last checkpoints as language adapters to train the task adapters. First they were trained on PersonaChat in L_S , then the language adapters were switched from L_S to L_T before the fine-tuning of the unchanged task adapters on the translated PersonaChat for the L_T model.

5 Evaluation

As assumed in [27] dialogue has a one-to-many structure which makes automatic metrics based on word overlaps often not correlate with human evaluations as they can dismiss good dialogue utterances that are different from the ground-truth [14, 6]. Hence human evaluation remains the most reliable but some automatic metrics like perplexity as shown in [1] can sometimes be somehow correlated to human judgements, so we add the information.

5.1 Automatic Metrics

We evaluated each of the models using the test set of PersonaChat in the relevant language. We computed perplexity for our models as it is available for all other state-of-the-art models. In addition Hits@1/3 was computed when possible (models trained, with a multi-choice head). This metric represents the accuracy of ranking the next gold utterance first in a set with two distractors.

We used greedy decoding to compute BLEU score [15] for comparison with other models when available, but also to show how it is at odd with actual dialogue model

³ The following pre-processed subsets of Wikipedia available on HuggingFace were used: 20220301.fr (L_T) and 20220301.en (L_S)

performance while perplexity is indicative. All the results of automatic metrics are reported in Table 2.

Table 2: Automatic evaluations of the different models grouped by strategy

Strategy	Models*	Perplexity	Hits@1**	BLEU
Train On Target	GPT-fr	10,82	0,88	N/A
	BLOOM	16,05	0,95	0,23
Test On Source	<i>GPT</i> [25]	18,49	0,84	N/A
	BLOOM	13,01	0,94	0,22
CrossLingual Training	<i>XNLG</i> *** (L_S)	54,74	N/A	2,25
	madx-BLOOM (L_S)	24,07	0,82	0,13
	<i>XNLG</i> (L_T)	640,33	N/A	0,09
	madx-BLOOM (L_T)	28,64	0,81	0,15

* Models in italics and metrics associated are state-of-the-art

** Available for models trained with double head

*** Both XNLG model metrics are from [13]

In [13], they reported extremely bad results on automatic metrics for models trained in cross-lingual framework. The MAD-X architecture on BLOOM improves these metrics with a significant gain especially on L_T models: from 640 to 28 on perplexity while we actually evaluate on the whole PersonaChat test set compared to a subset in their case. A lower perplexity can guarantee some generation capabilities, however the generated outputs may often be out of context in a dialogue framework yielding poor results in human evaluation.

5.2 Human Evaluation

We performed dialogue collection using the RASA-X [3] platform. It was done in two phases: in the first phase we deployed GPT-fr, TransferTransfo (GPT) and BlenderBot 1 and in the second all the four models based on BLOOM. We collected 140 conversations⁴ and evaluated them on three criteria selected based on those in [14, 9, 20]: coherence, engaging-ness and humanness. Detailed description of the dialogue collection and annotation process with an analysis of inter-annotators agreement are reported in Appendix.

⁴ For research purposes all collected data can be requested for by an e-mail to the first author.

5.2.1 Overall Ratings

Fig. 4 reports the overall ratings of the conversations for each model and by assessed quality⁵. For each conversation, we averaged 3 evaluators’ ratings per criterion.

Without a surprise BlenderBot 1 sets the reference. It’s a bigger model (~ 2.7 B parameters distilled into 400M) and it is trained on larger and varied datasets (Blended Skill Talk, Empathetic Dialogues, Wizard of Wikipedia [4] and also PersonaChat) with complex learning objectives. There is neither a comparable model in L_T , nor equivalent datasets. Hence this study of portability strategies focused on the other smaller models as an entry before generalizing to bigger and more complex models.

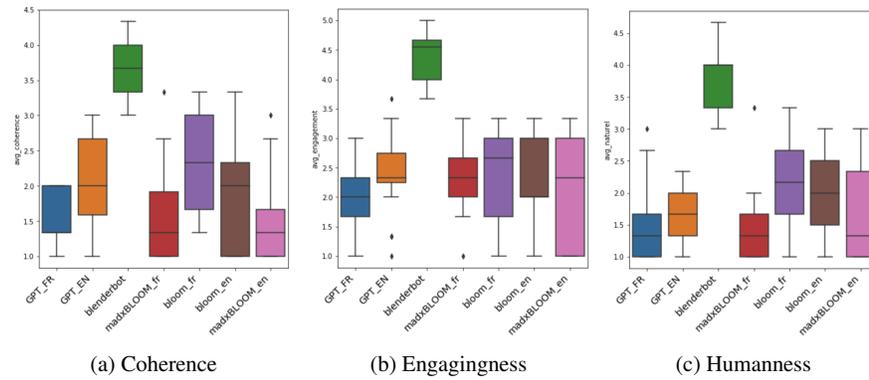


Fig. 4: Average overall ratings for the seven models deployed. In green is reference BlenderBot 1

Out of the six remaining models (with GPT_EN being from the state-of-the-art), BLOOM_fr emerges as the best in all three assessed categories in average: **0.25** on coherence, **0.07** on engagement and **0.3** on humanness from its closest runner-up GPT_EN. The latter has close ratings in average with BLOOM_en. The former has a marginal advantage on coherence (0.04) and engagingness (0.06) while the latter is slightly better on humanness (0.3). Then we have the last group composed by GPT_FR, madxBLOOM_fr, madxBLOOM_en in which the median average rating in all categories is below 1.5 meaning nearly half of the conversation generated with these models were rated to the lowest.

Table 3 show statistics on collected dialogues average number of utterances. They give another view of the models performances. Once again the BlenderBot sets the reference. We can notice that, the lowest rated models are also the one with the lowest average number of utterances with madxBLOOM_fr having the worst with 12.9

⁵ All are French (L_T) conversations generated and evaluated by French speakers even when the original conversational model is in English (L_T) (BlenderBot, GPT_EN, BLOOM_en and madxBLOOM_en in Fig. 4).

Table 3: Average number of utterances per model

Model	Blender	GPT_FR	GPT	xBLOOM_fr	xBLOOM_en	BLOOM_fr	BLOOM_en
#utterances	35.6	15.4	24.8	12.9	20.6	24.8	36.3

Where Blender stands for BlenderBot, xBLOOM for model with MAD-X architecture applied to bloom and GPT is for the English model based on GPT from [25].

utterances in average. Not only it explains the relatively low engagingness ratings of these models but also the coherence score. Indeed, early ending conversations with respect to our guidelines, indicate that some hallucinations and erratic behaviours like repetitiveness and dullness have been observed by the users (examples of these behaviour are provided in Table 4). The best models emerge with an average number of utterances per conversation higher ranging from **24.8** to **36.3**.

5.2.2 Impact of Additional Modules

In order to integrate the effect of orality, an additional, ready to use `Google API` module for ASR was involved. This module depending on the recording condition can produce erroneous transcriptions of the users inputs with a direct impact on the conversation flow as it often misleads the text-based model.

Regarding the usage of NMT, the translated dataset for L_T models seemed to have extremely impacted the monolingual PLM `GPT_FR` as its English counterpart performs much better. However, for the multilingual model `BLOOM`, it turns out it had a lower effect as the L_T *TrainOnTarget* model outperforms the others including its L_S counterparts. This may be due to the usage of NMT modules at inference with a model trained on high quality data. We believe that the multilinguality helped `BLOOM` during training on noisy data especially as it has translation abilities.

5.2.3 MAD-X Architecture for Dialogue

The two models implementing this architecture are the worst performing in this experiment. Although having strongly improved on automatics metrics compared to state-of-the-art models with similar approach. This finds explanation on the adapters being probably under-trained, especially the language adapters. Nevertheless, this approach is not to dismiss as its displayed glimpse of an ability to perform cross-lingual dialogue. Indeed, we notice that `madxBLOOM_fr` if given input in L_S was able to give a response in L_T keeping the context of the user’s L_S utterance. We believe that further training of models with this architecture should improve their performance and help build cross-lingual chitchat models. The latter is not further investigated in this study but can be of interest for future work.

Table 4: Examples of hallucinations and erratic behaviours in a dialogue

Bot Utterance	Translated utterance	User utterance
madxBLOOM_fr		
	<i>Hello, I may need your help</i>	Salut, j'aurais besoin de ton aide
bien je passe un peu de temps à enseigner à l'université de ma ville. tu?	<i>Well I spend sometimes teaching at my city's University. And you?</i>	
	<i>You teach Maths I hope</i>	Tu enseignes des maths j'espère
abétique à l'école. J'aimerais que vous êtes d'avantage avec lui	<i>"Abétique" at school. I want you to be even more with him</i>	
	<i>Well, I have some difficulties with my homeworks</i>	et bien j'ai des difficultés avec mes devoirs
Je porte souvent de promener dans la musique et à jouer de la musique en France	<i>I wear often wandering in the music and playing some music in France</i>	
GPT_fr		
Je suis un graphiste et un écrivain	<i>I am a graphist and a writer</i>	
c'est génial. J'aime lire	<i>Graphist and writer! it's awesome. I like reading</i>	Graphiste et écrivain !
	<i>What kind of book you say[read]</i>	quel genre de livre tu dis[lis]*
Je suis un écrivain créatif	<i>I'm a creative writer</i>	
	<i>Do you write?</i>	Est ce que tu écris?
Je suis un artiste	<i>I'm an artist</i>	
	<i>Me too</i>	Moi aussi
Je suis un écrivain	<i>I'm a writer</i>	
madxBLOOM_en		
	<i>Hi! I like climbing</i>	Salut! J'aime l'escalade
Moi je n'aime pas courir	<i>I don't like running</i>	
	<i>What sport do you practice?</i>	Quel sport pratiques-tu?,
J'aime aller à la gym et me détendre	<i>I like to go to the gym and relax</i>	
	<i>Are you a gymnast?</i>	Tu es gymnaste ?
Tu fais quoi comme boulot	<i>What do you do as a job?</i>	
	<i>Answer my question</i>	Réponds à ma question
j'ai travaillé dans un bar, mais je ne travaille pas	<i>I worked in a bar but I don't work</i>	
	<i>Why don't you work?</i>	Pourquoi tu ne travaille pas?
C'est ce que je fais dans la vie	<i>That's what I do for a living</i>	
	<i>Are you tired of life?</i>	Tu es fatigué de la vie ?
oui c'est un super metier	<i>Yes it's a great job</i>	

* Shows example of ASR error, "lis" is mistranslated into "dis" and it completely changes the sentence meaning and thus conversation flow.

In the first conversation we observe that the model never tried to answer the user's utterance and seems to generate random outputs. In the second conversation, we have a case of repetitiveness and dullness as the bot gets boring insisting on what is probably its personality traits. In the last conversation, the bot quickly stops following user's conversation and starts to be contradictory.

6 Conclusion

The development of French open-domain dialogue models is still far behind its English or even Chinese counterparts nowadays. So is the case for many other languages. The main reason being the lack of specialized datasets. However, the availability of PLMs in this language and advanced NMT tools are assets that can be leveraged to exploit the state-of-the-art from a higher resource language for this task. In this line we evaluated three different approaches and compared the models obtained and also to an English reference model. The TrainOnTarget strategy with a multilingual model, here BLOOM, yielded the best results on human evaluation. It opens the way for future work on more automatically translated data with multilingual models like BLOOM which implicitly possesses translation abilities. Improvement on the learning objectives then may help catch the reference models in high resource language for this task despite the obstacle of language specific dataset scarcity. Finally, the fact that, outside of chitchat dialog, French is high-resource is not totally limiting on these approaches. Indeed our best model was based on the TrainOnTarget approach with a multilingual model in particular BLOOM which includes some of the lowest-resources languages and is an open-access resource.

Acknowledgements We thank the coworkers in the Laboratoire Informatique d’Avignon (LIA) in Avignon University who volunteered during the evaluation process of our models. The work in this study is supported by μ DialBot project funded by the French National Research Agency (*Agence Nationale de Recherche, ANR*) under the grant ANR-20-CE33-0008.

References

1. Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q.V.: Towards a human-like open-domain chatbot (2020). URL <https://arxiv.org/abs/2001.09977>
2. BigScience Workshop: Bloom: A 176b-parameter open-access multilingual language model (2022). URL <https://arxiv.org/abs/2211.05100>
3. Bocklisch, T., Faulkner, J., Pawlowski, N., Nichol, A.: Rasa: Open source language understanding and dialogue management. CoRR **abs/1712.05181** (2017). URL <http://arxiv.org/abs/1712.05181>
4. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: Knowledge-powered conversational agents. CoRR **abs/1811.01241** (2018). URL <http://arxiv.org/abs/1811.01241>
5. Faloutico, R., Quatto, P.: Fleiss’ kappa statistic without paradoxes. *Quality & Quantity* **49**, 463–470 (2015)
6. Gupta, P., Mehri, S., Zhao, T., Pavel, A., Eskénazi, M., Bigham, J.P.: Investigating evaluation of open-domain dialogue systems with human generated multiple references (2019)
7. Haddow, B., Bawden, R., Miceli Barone, A.V., Helcl, J., Birch, A.: Survey of Low-Resource Machine Translation (2022)
8. Jabaian, B., Besacier, L., Lefèvre, F.: Comparison and combination of lightly supervised approaches for language portability of a language understanding system. *IEEE Transactions on Audio, Speech and Language Processing* **21**(3), 636–648 (2013)

9. Ji, T., Graham, Y., Jones, G.J.F., Lyu, C., Liu, Q.: Achieving reliable human assessment of open-domain dialogue systems (2022). DOI 10.48550/ARXIV.2203.05899
10. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of Hallucination in Natural Language Generation. arXiv.org (2022). URL <https://doi.org/>
11. Lefèvre, F., Mairesse, F., Young, S.J.: Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In: *Inter-speech 2010, 11th Annual Conference of the International Speech Communication Association*, pp. 78–81. Chiba, Japan (2010)
12. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)
13. Lin, Z., Liu, Z., Winata, G.I., Cahyawijaya, S., Madotto, A., Bang, Y., Ishii, E., Fung, P.: Xpersona: Evaluating multilingual personalized chatbot (2020). URL <https://arxiv.org/abs/2003.07568>
14. Mehri, S., Eskénazi, M.: Unsupervised evaluation of interactive dialog with DialoGPT (2020)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). DOI 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>
16. Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: MAD-X: An adapter-based framework for multi-task cross-lingual transfer (2020)
17. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
19. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: Towards empathetic open-domain conversation models: a new benchmark and dataset. In: *ACL* (2019)
20. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E.M., Boureau, Y.L., Weston, J.: Recipes for building an open-domain chatbot (2020)
21. Simoulin, A., Crabbé, B.: Un modèle Transformer Génératif Pré-entraîné pour le français. In: P. Denis, N. Grabar, A. Fraisse, R. Cardon, B. Jacquemin, E. Kergosien, A. Balvet (eds.) *Traitement Automatique des Langues Naturelles*, pp. 246–255. ATALA, Lille, France (2021). URL <https://hal.archives-ouvertes.fr/hal-03265900>
22. Smith, E.M., Williamson, M., Shuster, K., Weston, J., Boureau, Y.: Can you put it all together: Evaluating conversational agents’ ability to blend skills (2020)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2017)
24. Walker, M.A., Harmon, C., Graupera, J., Harrison, D., Whittaker, S.: Modeling performance in open-domain dialogue with PARADISE (2021)
25. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: Transfertransfo: A transfer learning approach for neural network based conversational agents (2019). URL <http://arxiv.org/abs/1901.08149>
26. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing Dialogue Agents: I have a dog, do you have pets too? arXiv.org (2018)
27. Zhao, T., Zhao, R., Eskenazi, M.: Learning discourse-level diversity for neural dialog models using conditional variational autoencoders (2017). DOI 10.48550/ARXIV.1703.10960

Appendix

Rather than performing self-chat pairwise human comparison like in [13] we chose to involve human during dialogues generation and annotation. In addition, we believe pairwise evaluation exclude the eventuality to conclude that the two compared models are poor as evaluators have to chose one over the other.

Dialogues collection During the first phase, three models were deployed using a modified version of the RASA-X [3] platform: `GPT-fr`, `TransferTransfo (GPT)` and `BlenderBot 1` - the reference model in this experiment. The two previous in L_S were combined with a `Google Translate` NMT module at input and output. 4 volunteers were asked to each do a minimum of 15 conversations (to have at least 5 conversations per model per person) without knowing which model they were interacting with. To avoid redundant conversations, testers were asked to start conversation with a different sentence. Also some personas from the PersonaChat dataset were randomly assigned to the testers in the case they wanted some inspiration to start a conversation. Each conversation should last at least 20 back-and-forths unless the model starts to hallucinate⁶ [10]; in this case it was asked to add 2 more inputs before closing the conversation if it didn't get better. This resulted in 60 conversations.

In the second phase, the exact same guidelines were given and this time we deployed the four models based on BLOOM. Again with 4 testers (different from the previous), we collected another 80 conversations. We gathered a total of 140 conversations generated by L_T native speakers in L_T .

Dialogues annotations Dialog-level evaluation was performed in order to evaluate each conversation individually and completely. Each conversation was evaluated by three annotators from a different batch of volunteers (10), rating from 1 to 5 each of the following criteria based on [14, 9, 20]:

- **Coherence:** are there hallucinations?, the quality of the bot's expression, coherent answer even if not factual, is the personality the same from the beginning to the end of the dialogue?, does it tend to change the subject too often?
- **Engagingness:** does the bot tend to be engaged in the conversation?, does it give constructive and not too vague answers ("okay", "yes", "maybe", "?" etc.)?, is it willing to restart the conversation when it stalls?
- **Humanness:** how much does it feel like a human to human discussion?, is the system repetitive?

Inter-annotator agreement: We can see in Table 5 that we have a fair to moderate agreement overall in each category with engagingness the lowest. However the trends seem to be slightly different among given model's conversations, with for instance a low agreement in coherence for the reference model BlenderBot. This means that evaluators struggles to agree on what a good coherence is worth in rating

⁶ A definition of hallucination was provided to the volunteers in the user guide

which is the opposite for the madxBLOOM models established as the worst were there is a stronger agreement across the abilities.

Table 5: Fleiss- κ [5] per ability for each model and overall

Model	Coherence- κ	Engagingness- κ	Humannes- κ
BlenderBot	0.187	0.313	0.417
GPT_FR	0.521	0.229	0.375
GPT_EN	0.292	0.292	0.521
madxBLOOM_fr	0.423	0.487	0.423
madxBLOOM_en	0.405	0.286	0.524
BLOOM_fr	0.219	0.271	0.193
BLOOM_en	0.528	0.278	0.167
Overall	0.361	0.301	0.379