Improving Zero-Shot Multilingual Text Generation via Iterative Distillation

†Ernie Chang, *Alex Marin, †Vera Demberg
†Dept. of Language Science and Technology, Saarland University
*Microsoft Corporation, Redmond, WA
cychang@coli.uni-saarland.de

Abstract

The demand for multilingual dialogue systems often requires a costly labeling process, where human translators derive utterances in low resource languages from resource rich language annotation. To this end, we explore leveraging the inductive biases for target languages learned by numerous pretrained *teacher* models by transferring them to *student* models via sequence-level knowledge distillation. By assuming no target language text, both the teacher and student models need to learn from the target distribution in a few/zero-shot manner. On the MultiATIS++ benchmark, we explore the effectiveness of our proposed technique to derive the multilingual text for 6 languages, using only the monolingual English data and the pretrained models. We show that training on the synthetic multilingual generation outputs yields close performance to training on human annotations in both slot F1 and intent accuracy; the synthetic text also scores high in *naturalness* and *correctness* based on human evaluation.

1 Introduction

In multilingual dialogue systems, natural language generation is used to generate utterances in various languages, using as input semantic frames, which contain a representation of the user intent together with relevant information or entities related to said intent (Tur et al., 2010). Past works that generalize dialogue systems to multilingual settings often made two unrealistic assumptions about the data availability of any new dialogue domain (Liu et al., 2019; Xu et al., 2020; Schuster et al., 2019; Chang et al., 2020, 2022): (1) First assumption is that a large set of monolingual data has already been annotated. (2) Some in-domain text or annotated data in the target languages are already available for the purpose of transfer learning. Neither assumption holds in all cases (Upadhyay et al., 2018).

To overcome these challenges, we utilize knowledge-grounded pre-training (KGPT) (Chen et al., 2020) – a pretrained data-to-text generation model that was shown to be effective in overcoming data scarcity, as the model is capable of representing the inductive biases required to encode structured data such as the slot-value pairs (frames). In few-shot settings, we can exploit KGPT's pretrained knowledge to obtain in-domain text labels for a large set of unlabeled frame sequences. These text labels are then converted to multiple languages with the use of bilingual translation models as teacher models, as inspired by past works on sequence-level knowledge distillation (Wang et al., 2020; Kim and Rush, 2016; Gordon and Duh, 2019). In this way, we perform zero-shot cross-lingual transfer for all 6 languages. We use mBART (Liu et al., 2020) as the multilingual student model to acquire stronger bilingual knowledge from the translation teacher models from Tiedemann and Thottingal (2020).

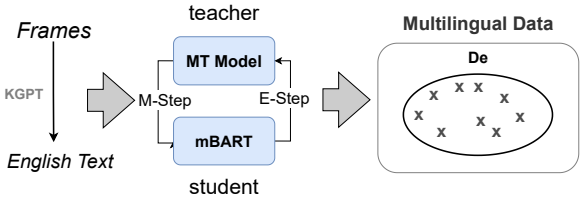

Figure 1: **Annotation scenario**: Each × represents a labeled data instance. The goal is to generalize from few-shot human-labeled instances in one language (left) to large synthetic multilingual data (right).

We leverage a two-step distillation process where we first derive a large synthetic English dialogue data from the English seed data, then generalize it to multilingual data by using the bilingual translation models to produce synthetic text labels. Finally, we perform rounds of iterative knowledge distillation following the process of the expectation-maximization algorithm for further improvements. This work makes the following contributions:

- We introduce a simple and effective technique in constructing a synthetic multilingual dia-

Proceedings of the 29th International Conference on Computational Linguistics, pages 5876–5881
October 12–17, 2022.

logue corpus using the *iterative knowledge distillation*.

- We demonstrate the efficacy of the technique by showing that its outputs display high *naturalness* and *correctness*.

2 Approach Summary

In our setting, we have (1) a seed English dataset S which consists of k labeled pairs, and (2) the full set of unlabeled frame sequences U where $|U| \gg k > 0$. The goal is to create labeled samples in all target languages consisting of the frame sequences ($\mathcal{X}$) with their corresponding texts ($\mathcal{Y}$).

Monolingual Text Generation. We first obtain the full synthetic English dataset $(\mathcal{X}_{En}, \mathcal{Y}_{En})$ from the k labeled pairs and the unlabeled semantic frame sequences. This is achieved by finetuning KGPT on the k samples and then labeling each semantic frame sequence with a corresponding English utterance.

Multilingual Text Generation. To create multilingual data, we perform the *iterative knowledge distillation* (see §3) to derive target language utterance from the source English utterance. Specifically, we update both the bilingual translation model (*teacher*) and mBART (*student*) iteratively following the expectation-maximization algorithm via likelihood maximization over parallel data $(\mathcal{X}, \mathcal{Y})$ and parameters ϕ and θ of the teacher and student models.

3 The Iterative Distillation Procedure

The iterative distillation procedure alternatively optimizes the student and teacher models until convergence. We generate the parallel synthetic data from pretrained bilingual models $p^{teacher}(y|x;\phi^1)$[1], and use the EM algorithm (Dempster et al., 1977) to optimize the process (see Figure 2). The intuition is that while the student model learns from the teacher, the teacher model also needs to discard some out-of-domain knowledge by learning from the student feedback. The iterative procedure alternates between the teacher model learning some knowledge from the target distribution, and then the student model is updated based on the new teacher model. In this way, both models are improved in training. As such, we use the following high-level strategy.

[1]Note that ϕ^1 is used to indicate the initial pretrained model at iteration 1.

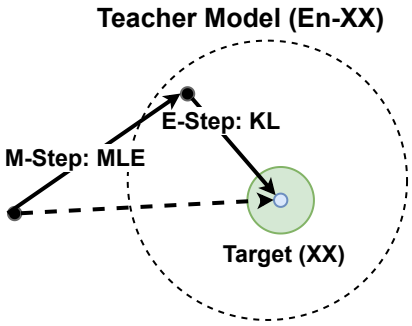

Figure 2: **Iterative Knowledge Distillation**: Each circle represents a labeled data instance from the *student* where **M-step** moves the output distribution towards that of the teacher model's; and **E-step** measures the distributional difference and makes further adjustment towards the *target language domain*.

In the E-step (multilingual labeling), we fix the $p^{student}$ and update the posterior distribution:

$$q^{t+1} = \text{argmin}\mathbb{KL}(\hat{\mathcal{Y}} \| p^{student}(\mathcal{Y}|\mathcal{X};\theta^t)),$$

Conversely, in the M-step (training), we fix $q(\mathcal{Y})$ and update θ to maximize the expected log-likelihood:

$$\theta^{t+1} = \text{argmax}_\theta \mathbb{E}_{q^{t+1}}[\log p^{student}(\mathcal{Y}|\mathcal{X};\theta)],$$

In what follows we introduce the details of the E-step and the M-step in our framework.

Expectation Step. The E-step aims to compute the posterior distribution $q(\mathcal{Y})$ that minimizes the KL divergence between $q(\mathcal{Y})$ and $p^{student}(\mathcal{Y}|\mathcal{X})$. This step basically brings the teacher model closer to the target distribution without having seen the distribution itself. Importantly, we also estimate the gradient of $\mathcal{L}(\cdot)$ w.r.t. teacher model's parameter ϕ by applying the REINFORCE algorithm (Williams, 1992) to compute the loss, which is then weighted by $\log \frac{p^{student}(y|x;\theta^t)}{p^{teacher}(y|x;\phi)}$ under the KL divergence equation. Overall, this constructs a weighted synthetic training dataset that intuitively *adjusts the outputs to be more in-domain*, as the original teacher model is general-domain.

Maximization Step. In the M-step, we update the student model to be closer to the teacher model. To do so, we optimize the parameters θ^{t+1} with the parameterized posterior distribution $p^{teacher}(\mathcal{Y}|\mathcal{X};\phi^{t+1})$ so as to optimize the *student model* to generate target language text. We apply sequence-level knowledge distillation (Kim and Rush, 2016) and use the targets with maximum likelihood in the teacher model to train the student model.

Semantic frame sequence:
[B-fromloc.city_name] toronto, [B-toloc.city_name] newark, [B-round_trip] one, [I-round_trip] way, [B-depart_date.day_name] wednesday, [B-depart_time.period_of_day] evening, [B-depart_date.day_name] thursday, [B-depart_time.period_of_day] morning
English reference: i need a flight from toronto to newark one way leaving wednesday evening or thursday morning

English (En): I need a flight from Toronto to Newark, one way is to leave from Wednesday night or Thursday morning
German (De): Ich brauche einen Flug von Toronto nach Newark in eine Richtung ab Mittwochabend oder Donnerstagmorgen
Spanish (Es): Necesito un vuelo de Toronto a Newark solo ida y salida el miércoles por la noche o el jueves por la mañana
French (Fr): Besoin d'un vol de toronto à newark aller simple partant mercredi soir ou jeudi matin
Chinese (Zh): 需要一种从多伦多到纽瓦克的航班，从星期三晚上或星期四早上离开
Japanese (Ja): 水曜日の夕方または木曜日の朝を出して、トロントからニュアクへの片道のフライトが必要です
Portuguese (Pt): preciso de um voo de toronto para newark só de ida saindo na quarta à noite ou quinta de manhã

Figure 3: Table showing the labeled examples in all seven languages. The *upper* portion shows the monolingual (English) semantic frame sequence and utterance pair. The *bottom* region displays all seven languages.

Quality-Based Text Filtering. To ensure that only target data with high semantic correctness quality is used for training, we impose a filtering operation on the generated samples for quality control. In the process of *multilingual labeling*, we assume to only have the access to monolingual (English) frame sequence, and so we rely on the likelihood score of a pre-trained teacher model (bilingual MT models) as the quality metric $\mathrm{Q}_{x^i}(y^i) = \log p^{teacher}(y^i|x^i;\phi)$. where ϕ denotes the *initial* teacher model trained on the original ground-truth dataset. In practice, we use nuclear sampling (Holtzman et al., 2019) (which has a resizable beam size) as the heuristic sampling on $p^{teacher}(y|x;\phi^t)$, and then filter out the candidates which do not satisfy the condition $\mathrm{Q}_x(y) \geq b$, where b is set to be 0.5 based on our empirical findings. In this way, we control the quality of $p^{teacher}(y|x;\phi^{t+1})$ by manipulating the quality of its training data.

4 Experimental Settings

Training Configurations For mBART training, we use the same vocabulary of subword units as Liu et al. (2020); this vocabulary includes a sentencepiece model (Kudo and Richardson, 2018) with $250,000$ subword tokens. The mBART model consists of the standard sequence-to-sequence Transformer architecture with 6 encoder and 6 decoder layers; each layer consists of a 1024-dim model on 8 heads ($\sim$ 144M parameters altogether). Our model is trained on 256 Nvidia V100 GPUs (32GB). The final models are selected based on BLEU (Papineni et al., 2002) scores on the validation set.

Testing Scenarios We evaluate our technique on the MultiATIS++ corpus (Xu et al., 2020), which consists of re-annotated ATIS dataset in six additional languages: German (De), Spanish (Es), French (Fr), Chinese (Zh), Japanese (Ja), and Portuguese (Pt). The test sets are based on the released human-labeled set consisting of 893 instances. Particularly, we report the results on both intent classification and slot filling F1 scores for NLU inference; and evaluate the surface-level overlap with BLEU-4 for text generation. The reason for this is so that we could get a sense of the correlation between text quality and its usefulness for NLU inference. For our experiments, we *assume that semantic frames corresponding to all target languages are present*. For semantic frame sequence predictions, we employ Fairseq (Ott et al., 2019) and train it on the synthetic multilingual corpus for all targeted language pairs. We adopt several ways of generating the synthetic corpus (En-XX) from the English seed data consisting of **50-shot**, **100-shot**, and **all** English ATIS data:

MT: The baseline is the direct translation of the *seed English utterances* into target language utterance (XX), then training mBERT on the synthetic data consisting of target language utterance and its semantic frames.

KGPT+MT: We use KGPT to create the full synthetic English corpus, then perform **MT**.

mBART: On top of **KGPT+MT**, we finetune mBART on the synthetic En-XX corpus, then create (En-XX) via translation.

mBART+EM: Building on top of **mBART**, we perform the proposed EM algorithm and allowing both the bilingual model (teacher) and mBART (student) to be updated.

	En		Fr		De		Zh		Es		Ja		Pt	
NLU (Slot %F-1 Intent %Acc)														
50-shot														
MT	67.15	72.34	66.61	71.66	63.66	78.54	58.41	74.42	58.83	71.53	70.55	68.83	66.72	75.51
KGPT+MT	70.19	77.28	70.99	75.37	66.73	82.88	61.56	80.65	61.92	76.83	75.22	73.18	70.20	78.22
mBART	72.43	79.32	72.68	79.66	68.27	84.78	66.29	83.21	64.19	77.52	75.61	77.43	**72.01**	**80.33**
mBART+EM (Ours)	**75.48**	**82.15**	**75.37**	**71.24**	**71.75**	**85.34**	**67.31**	**84.20**	**64.88**	**77.63**	**76.38**	**78.72**	71.28	80.29
100-shot														
MT	65.24	71.73	56.53	72.48	64.62	79.51	63.55	74.48	60.37	72.61	71.47	71.52	68.48	75.41
KGPT+MT	74.37	85.54	65.84	84.17	80.33	83.43	76.32	83.11	63.53	78.54	76.57	79.34	77.63	80.74
mBART	78.52	87.82	66.38	83.43	72.28	85.80	77.81	84.33	67.42	81.37	79.16	80.26	77.13	82.61
mBART+EM (Ours)	**82.22**	**88.14**	**67.44**	**83.82**	**74.29**	**88.32**	**77.92**	**85.32**	**68.39**	**81.98**	**79.89**	**81.12**	**77.45**	**84.50**
All														
MT	85.15	89.88	70.33	87.42	75.72	91.42	77.72	92.26	72.41	84.35	81.71	83.25	80.35	87.74
mBART	87.42	89.73	81.74	88.33	76.62	92.93	78.73	92.15	**74.42**	**84.63**	81.74	83.83	**80.62**	**87.12**
mBART+EM (Ours)	**88.97**	**90.10**	**82.35**	**90.02**	**76.93**	**93.66**	**79.01**	**92.89**	74.25	84.19	**82.55**	**84.30**	80.60	87.28
NLG (BLEU-4)														
MT	9.22		7.58		9.71		7.82		8.63		9.33		8.88	
Ours (50-shot)	10.37		8.29		10.21		7.67		8.45		9.72		8.23	
Ours (100-shot)	11.23		9.38		11.87		**9.44**		**9.90**		10.91		**9.58**	
Ours (All)	**12.67**		**10.32**		**12.43**		9.33		9.80		**10.99**		9.11	

Table 1: Benchmark comparison on all seven languages reporting both NLU (**slot(%) intent(%)**) and NLG (**BLEU-4**). KGPT (Chen et al., 2020) is the pretrained data-to-text generation model; and MT refers to the use of Helsinki bilingual translation model (Tiedemann and Thottingal, 2020).

Model	Spanish (Es)			Chinese (Zh)		
	Naturalness	Miss	Wrong	Naturalness	Miss	Wrong
Reference	4.00	-	-	5.00	-	-
MT	**3.66**	57	24	3.33	28	33
KGPT+MT	3.33	45	22	**4.33**	24	32
mBART	3.33	37	17	3.33	23	28
Ours	**3.66**	**23**	**12**	**4.33**	**18**	**21**

Table 2: Human Evaluation on the sampled outputs (100 instances) for all models on the 100-shot scenario. Three annotators were asked to evaluate the *Naturalness* (0-5), *Miss* (i.e. # missed slots), and *Wrong* (i.e. # hallucinated slots).

5 Results and Analysis

Here we first present two forms of analysis for both monolingual and multilingual data, then analyze the synthetic data with human evaluation.

Analysis of Monolingual Data. In table 1, we first observe that the use of KGPT (*KGPT+MT*) in the few-shot settings helps to produce high quality synthetic English (En) data that allows the mBART models to achieve decent NLU performance. We also see that the difference between 50-shot and 100-shot is minor, which we think is highly dependent on the random sampling process of the few-shot data. The improvement for *Ours* is slightly more drastic when all data is used (*All*), where the performance (82.22 to 91.97) approaches that of the system using the real full English data.

Analysis of Multilingual Data. In the multilingual setting, we observe that finetuning mBART on the translated data (*mBART*) brings about noticeable improvements generating high quality text over some language pairs (e.g. En-Fr, En-De); the improvement is limited for some languages (e.g. Ja). We attribute this to the cross-lingual similarity that allows some language pairs to obtain more useful inductive biases than in cases of more dissimilar language pairs. Further, we observe consistent improvements that *mBART+EM* has over the base models, suggesting that the iterative knowledge distillation process is crucial to draw both the *teacher* and *student* models to the in-domain region. This can be seen across most languages. As such, we conclude that the proposed technique does indeed help to create useful synthetic data, even in zero- or few-shot cross-lingual settings.

On Generation Quality. In Table 1, we also notice the *limited BLEU-4 scores*, which means that the multilingual human annotation has rephrased the utterance quite drastically different from the original English text. To examine further, we perform human evaluation (See Table 2) on the Spanish and Chinese generation outputs based on 100-shot data, to look for evidence of naturalness and high fidelity. We observe that the human evaluation is consistent with that of the intent classification and slot filling scores, while having high naturalness and fidelity as defined by *Miss* and *Wrong*.

6 Limitations

We also recognize that the approach gradually loses its effectiveness as the size of the data increases.

Moreover, in some language pairs such as Portuguese and Chinese, the improvement with EM steps remain largely limited. We attribute this to the linguistic gap across language pairs, which is the biggest limitation of our approach, since the approach's effectiveness hinges upon the proximity between source-side high resource language (i.e. English) and the target-side languages. Therefore we postulate that the approach would be very limited for extremely low resource languages such as many of the African languages.

7 Conclusion and Future Work

In this paper, we show the effectiveness of the proposed technique in constructing synthetic multilingual data from few-shot monolingual samples. Surprisingly, training on the synthetic outputs yields decent performance in terms of slot F1, intent accuracy, and human evaluation. We hope to extend the work in the future to low resource languages – applying it to additional tasks beyond NLU such as coreference resolution.

Acknowledgements

This research was funded in part by the German Research Foundation (DFG) as part of SFB 248 "Foundations of Perspicuous Software Systems". We sincerely thank the anonymous reviewers for their insightful comments that helped us to improve this paper.

References

Ernie Chang, David Adelani, Xiaoyu Shen, and Vera Demberg. 2020. Unsupervised pidgin text generation by pivoting english data and self-training. In *In Proceedings of Workshop at ICLR*.

Ernie Chang, Jesujoba Alabi, David Adelani, and Vera Demberg. 2022. Dialogue pidgin text adaptation via contrastive fine-tuning. In *AfricaNLP @ ICLR 2022*.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. Association for Computational Linguistics.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Mitchell A Gordon and Kevin Duh. 2019. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *arXiv preprint arXiv:1912.03334*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in ATIS? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020. Structure-level knowledge distillation for multilingual sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3317–3330.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063.