Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D4.2 – GoURMET Final Report on Transfer Learning

Nature	Report	**Work Package**	WP4
Due Date	31/03/2022	**Submission Date**	31/03/2022
Main authors	Alexandra Birch, Antonio Valerio Miceli Barone, Jindřich Helcl (UEdin)		
Co-authors	Felipe Sánchez-Martínez (UA)		
Reviewers	Felipe Sánchez-Martínez (UA)		
Keywords	transfer, research, resources		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

Contents

List of Figures

Abstract

In this deliverable for the GoURMET project we describe the work done in Workpackage 4: Transfer Learning, which focuses on improving news translation for low-resource languages by exploiting alternative data resources. The workpackage consists of three main tasks: *Learning from Multilingual Data*, *Learning from Monolingual Corpora* and *Learning from Lexical Resources*. We report on the work already carried out in the second half of the project.

1 Introduction

The GoURMET project aims to develop systems to automatically translate news articles between English and low-resource languages spoken in regions of the world that are of interest to international multi-lingual broadcasters such as BBC and Deutsche Welle. Corpora of parallel text are the most important resource used to build high-quality machine translation systems, but for low-resource languages, by definition, this data is scarce.

This work package aims at leveraging resources other than in-domain parallel text to improve the quality of our machine translation systems using techniques generally known as Transfer Learning.

In this document we report the research we carried out during the second and final part of Work Package 4 of the GoURMET project. The research performed in the first part of the project has been documented in the deliverable *D4.1 Initial progress report on transfer learning*.

1.1 Work package Tasks outline

As per the original project poposal, this work package is structured as three tasks, each focusing on a different type of data resource:

- **T1: Learning from Multilingual Data**
  We exploit the similarity between languages to improve translation quality for one language pair by leveraging data for related languages.
- **T2: Learning from Monolingual Corpora**
  We leverage corpora of monolingual text, which is much more abundant than parallel text, especially for the news domain. In the second half of the project, this task also focuses on using pre-trained models.
- **T3: Learning from Lexical Resources**
  We make use of curated linguistic resources such as large bilingual dictionaries.

In the second half of the project, we continued research in all tasks. In Task 2, which is aimed at using monolingual corpora to improve low resource translation, there has been a notable shift in focus in the field of machine tanslation towards using large pre-trained models. This is aligned with the recent advances in the field of natural language processing and the increased availability of these multilingual pre-trained models, which contain many of the low-resourced languages the GoURMET project is interested in.

2 Task 1: Learning from Multilingual Data

In this section we describe our contributions to the Task 1 of Work Package 4, focused exploiting multilingual resources for improving low-resource machine translation using transfer learning.

In the first part of this section (Section 2.1) we describe a novel approach for improving code-switching in context of creating multilingual corpora for pre-training of massively multilingual models such as mBART50 (Tang et al., 2020). This contribution is a work-in-progress and it is currently under review. In the second section (Section 2.2), we describe a multilingual fine-tuning approach applied in context of multilingual dialogue systems. Finally, Section 2.3 introduces an annotation effort to collect a multilingual dialogue dataset.

2.1 Contextual Code-switching for Improved Pretraining in Multilingual NMT

This section describes a work in progress focused on improving the quality of code-switched corpora which are used for multilingual pretraining.

Recent efforts for pretraining large multilingual NMT models have proposed training models to denoise artificially code-switched corpora (Yang et al., 2020; Lin et al., 2020), in an effort to enhance cross-lingual transfer learning. This has enabled the creation of powerful multilingual models such as mRASP (Lin et al., 2020) and its successor, mRASP2 (Pan et al., 2021), that have exhibited state-of-the-art (SOTA) performance for a variety of high, medium and low-resource languages across supervised, unsupervised and zero-shot translation scenarios. To create code-switched corpora, these works use lexicons, most commonly the MUSE dictionaries (Lample et al.,

2018), and randomly substitute words with their translations extensively. For example, the Aligned Augmentation (AA) algorithm used for training mRASP2 substitutes 90% of words in a source sentence with random translations sampled from MUSE dictionaries.

However, such techniques do raise certain significant concerns regarding the quality of the code-switched data. For instance, a principal weakness of such lexicon-based codeswitching is its inability to factor sentence-level context, which could result in many potential issues, including: a) violation of syntatic or grammatical rules and b) erroneous handling of polysemes and context-dependent synonyms. Secondly, the MUSE dictionaries only provide one-to-one word-level translations which cannot adequately scale to multi-word expressions, with the issues being potentially even more serious when encountering languages with higher synthesis. Lastly, the qualities of the dictionaries themselves have been shown to be quite dubious across a variety of languages (Kementchedjhieva et al., 2019). It is worth noting that the above-mentioned problems would not only affect performance for the specific scenarios and languages, but thanks to extensive code-switching, such errors could likely propagate – affecting the learned multilingual semantic representations, and potentially harming translation performance in general.

In an effort to address this, we propose a noising mechanism called Context-Aligned Substitution (CAS) that seeks to obtain contextual, many-to-many word translations for noising a corpus, leveraging sentence-level translations generated by a pretrained NMT system. Then, the word aligments of the source sentence with these translations are generated using a word-aligner. Given the translations and the alignments, the CAS algorithm extracts many-to-many aligned word pairs which are then used to create code-switching corpora for pretraining. The experiments in this work utilise mBART50 (Tang et al., 2020) and awesome-align (Dou and Neubig, 2021) as the respective translation and alignment models respectively. Experiments conducted on 3 different language families – the high-resourced Romance, the synthetic Uralic and the low-resourced Indo-Aryan – show CAS consistently outperforming the AA algorithm proposed by Pan et al. (2021), sometimes by as high as 5–6 BLEU points. In addition, on comparing with large pretrained models such as mRASP2 and mBART50, it can be observed that harnessing mBART50 for noising using the CAS methodology gives comparable or better performance than the aforementioned large models, despite using a fraction of the computational and data resources (often lower by 2 orders of magnitude). These results, which are shown to be vastly superior to traditional knowledge distillation baselines, suggest that this noising-based pretraining mechanism can also be used as a technique to distill large multilingual models more effectively.

We conduct ablation studies to examine some of the reasons behind the success of the CAS methodology. We show that pretraining models on contextually code-switched data generates more grammatically correct translations, and that the many-to-many substitutions yielded by CAS bring the multilingual semantic representations of multi-word expressions closer together, contributing to significant performance improvements. We also demonstrate how using a uniform replacement ratio of 0.9, as suggested by Pan et al. (2021), may not be the most optimal choice when building smaller models, and that it may be beneficial to consider different replacement ratios based on factors like agglutination and quality of substitutions available. Finally, we enlist some of the limitations of this work – namely cost, resource requirements, vulnerability to poor quality translations etc. – and discuss cheap but efficient ways to mitigate them.

The primary contributions of this work, therefore, are as follows:

1. We propose a code switching-based noising mechanism for NMT called CAS that suggests a greater focus on the quality of code-switching can consistently bring about significant

improvements in translation performance, across various languages.

2. Through ablation studies and other analyses, we aim to give a better understanding of the importance of different aspects when attempting to improve code-switching quality – context, many-to-many substitutions, replacement ratio etc. – and how this varies across language families

3. Leveraging large pretrained models, we demonstrate how the CAS mechanism can be used to train small but high-quality multilingual models, with a fraction of the training data or computation used for large models. We also show how these can perform substantially better than traditional knowledge distillation methods.

4. Finally, we discuss the limitations of such an approach, from cost and feasibility perspectives, propose efficient ways to mitigate them and lay down grounds for future work.

2.2 Cross-lingual Intermediate Fine-tuning for improving Multilingual Encoders

This work has been published at EMNLP 2021 (Moghe et al., 2021).

In recent years, task-oriented dialogue systems have achieved remarkable success by leveraging huge amounts of labelled data. This technology is thus limited to a handful of languages as collecting and annotating training dialogue data for different languages is expensive and requires supervision from native speakers (Chen et al., 2018).

To avoid having to create large annotated datasets for every new language, recent work focuses on transfer learning methods which use neural machine translation systems (Schuster et al., 2019), code-mixed data augmentation (Liu et al., 2020b; Qin et al., 2020) or large multilingual models (Lin and Chen, 2021). Neural machine translation models incur additional overhead of training on millions of parallel sentences that may not be available for all language pairs. In this work, we focus on transfer learning via large multilingual models, which will allow us to extend models to languages with limited labelled training data.

In techniques that use multilingual models, a task-specific architecture uses this pretrained model as one of its components and then is trained with task data from a high resource language (See Fig. 1). It is then evaluated directly or with some labelled examples in a different language. The use of intermediate fine-tuning, which is fine-tuning a large language model with a different but related data/or task and then fine-tuning it for the target task, has shown considerable improvements for both monolingual and cross-lingual natural language understanding tasks (Gururangan et al., 2020; Phang et al., 2020). But, it is relatively under-explored for multilingual dialogue systems.

In this work, we demonstrate the effectiveness of using cross-lingual intermediate fine-tuning of multilingual pretrained models to facilitate the development of multilingual conversation systems. Specifically, we look at cross-lingual dialogue state tracking tasks, as they are an indispensable part of task-oriented dialogue systems. In this task, a model needs to map the user's goals and intents in a given conversation to a set of slots and values - known as a "dialogue state" based on a pre-defined ontology. Our intermediate tasks are based on interaction between the source and target languages and interaction between the dialogue history and response. These tasks involve the prediction of missing words in different conversational settings. These include monolingual conversations, concatenated parallel bilingual conversations, and cross-lingual conversations. Further, we also introduce a task as a proxy for generating a response in a cross-lingual setup. Our

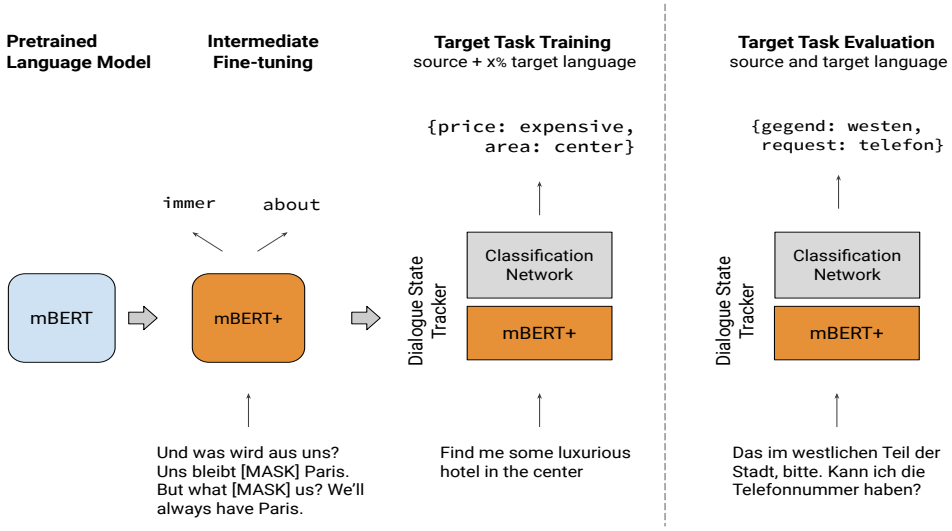

Figure 1: Pipeline of our work. A pretrained language model is fine-tuned with the task of predicting masked words on parallel movie subtitles data. A dialogue state tracker is then trained with this new multilingual model and evaluated for cross-lingual dialogue state tracking

intermediate tasks only use 200K lines of parallel data which is available for 1782 language pairs. Using parallel data for intermediate fine-tuning also becomes an important addition in the intermediate fine-tuning literature which has largely focused on related monolingual tasks. Please see 1 for an overview of the proposed setup. Our best method leads to an impressive performance on the standard benchmark of the Multilingual WoZ 2.0 dataset (Mrkšić et al., 2017) and the recently released parallel MultiWoZ 2.1 dataset (Gunasekara et al., 2020). The best method uses dialogue history and parallel conversational context confirming that our design principles based on conversation history and cross-lingual conversations indeed provide empirical gain. Our methods use 200k parallel movie subtitles (Lison and Tiedemann, 2016) for intermediate training and this data is already available for 1782 language pairs allowing extension to new language pairs.

Our contributions can be summarized as follows:

1. To the best of our knowledge, this is the first work to use parallel data for intermediate fine-tuning of multilingual models for multilingual dialogue tasks. We provide strong empirical evidence on four language directions in two datasets for low-resource and zero-shot data scenarios.

2. Our proposed intermediate fine-tuning techniques produce data-efficient target language dialogue state trackers. We achieve state-of-the-art results for the zero-shot Multilingual WoZ dataset for most of the metrics and obtain >20% improvement on joint goal accuracy with limited labelled data in the target language for the MultiWoZ dataset over the baseline.

3. We propose two new intermediate tasks: Cross-lingual dialogue modelling (XDM) and Response masking (RM) that can be extended to other cross-lingual dialogue tasks.

2.3 MultiChat: Benchmarking Cross-lingual Transfer Learning

This is an ongoing annotation effort.

The success of task-oriented dialogue is often evaluated only on a handful of languages (Razumovskaia et al., 2021). The two promising paradigms for extending conversational systems to multiple languages with limited training data are 1) the deployment of intermediate machine translation models and 2) cross-lingual transfer learning via pre-trained multilingual language models. However, the machine translation models and the pre-trained language models are often trained on news text, formal documents, and Wikipedia, which is different from conversational text leading to poor performance on the end conversational tasks. While there have been some efforts to build datasets for multilingual task-oriented dialogue (Razumovskaia et al., 2021), these often include synthetic translations and constrained setups. Synthetic translations include either post-editing of machine translations or instructing humans to provide translations, which still exhibit *translationese* that tend to overestimate the performance of the various conversational tasks (Majewska et al., 2022).

We are collecting a multilingual dialogue dataset to demonstrate the challenges involved in cross lingual transfer learning through the existing state-of-the-art academic machine translation systems and multilingual language models for various NLP applications. Specifically, we intend to develop a multilingual, multi-domain, multi-intent detection based dialogue dataset where every example contains a translation of the source dataset and a paraphrase of the translation. Intent detection involves classifying the goal of the user's utterance from several pre-defined classes (intents). We especially wish to consider low-resource languages – ones which have many speakers but few academic resources. We aim to collect the dataset in a two-stage process. In the first step, translators are asked to translate an English utterance into the target language and in the second step, a second worker is asked to produce a target language paraphrase of the translated utterance. This second step is intended as a way to address the issue of translationese in the dataset.

Our benchmark that addresses multilingual, multi-intent, multi-domain, and multi-reference aspects is the first of its kind. We expect several sub-communities within academic and industrial research to benefit from our work, especially those working on monolingual/multilingual dialogue, machine translation, and natural language evaluation. The modular design of intent detection will significantly improve the intent detection process in a commercial setup as well as let academic researchers work on novel ways to perform multi-label classification. As businesses start to cater to a global audience, reliance on automated multilingual customer services will also increase. We believe the presence of a multilingual benchmark can accelerate research in building such multilingual services. Beyond the development of multilingual multi-intent detection systems, the collection of parallel data specifically for dialogue is a useful benchmark to evaluate the adaptation of machine translation systems for the domain of dialogue. The presence of translation and its paraphrase can also help us determine whether existing machine translation quality estimation techniques can evaluate translations with different surface forms. Additionally, the multi-reference work can contribute to the understanding of different ways in which humans can communicate the same intent. By including geographically diverse languages (Spanish, Marathi, Amharic, and Turkish), we believe this work will also contribute to the democratization of language technologies.

3 Task 2: Learning from Monolingual Corpora

This section summarizes our contributions to Task 2 which focuses on leveraging monolingual data sources in NMT. Monolingual data has been successfully employed in generating synthetic

data either for backtranslation (Sennrich et al., 2016) or knowledge distillation (Kim and Rush, 2016). Another use for monolingual data has been found in large pre-trained multilingual models such as mBART50 (Tang et al., 2020). We focus on the latter approach adapted for low-resource scenarios in Sections 3.1 and 3.2.

Contributions related to this task are also described in other deliverables. A work focused on diversity in generated backtranslations is described in Section 4.3 in Deliverable 1.4. Another ongoing effort focused on improving translation of out-of-vocabulary words using lexicons in described in Section 4.2 of D1.4.

3.1 Exploitation of large pre-trained models for low-resource neural machine translation

The work described in this section has been submitted to the 29th International Conference on Computational Linguistics (COLING 2022).

Pre-trained or foundation models (Bommasani et al., 2021) have reshaped the landscape of natural language processing applications. These models are usually trained by following a self-supervised approach over large quantities of text. In addition to models pre-trained to obtain general-purpose neutral representations, there exist a number of multilingual encoder-decoder models specifically pre-trained to translate between many different language pairs. Well-known systems in this group include mBART50 (Tang et al., 2020), M2M-100 (Fan et al., 2021), CRISS (Tran et al., 2020), mT6 (Chi et al., 2021), or SixT+ (Chen et al., 2022). All these pre-trained models attain high translation quality (Tran et al., 2021) because they leverage information from multiple language pairs, thus becoming an interesting example of the possibilities of transfer learning.

As part of UA's work in task T4.2, we have worked on exploiting mBART50 for building NMT system for low-resource languages. mBART50 was obtained by additionally training mBART in a supervised manner to translate between English and 49 languages, and vice versa. We chose mBART50 because it is centred on English and because of the observation on its performance in the literature. Lee et al. (2022) compared mBART50 and mT5 and observed that mBART50 performed better in most translation directions and average BLEU. Liu et al. (2021, Table 1) reached a similar conclusion when comparing mBART and mT5 after fine-tuning them for NMT. The paper by Chen et al. (2022) includes, to our knowledge, the most recent comparison between NMT pre-trained models including mBART50, M2M-100, CRISS and SixT. Their evaluation for many-to-English NMT systems for 23 languages shows (Chen et al., 2022, Table 1) small average difference between M2M-100 and mBART50.

Approach. We propose a pipeline to tune mBART50 for the translation between English and a specific low-resource language and, afterwards, distil the knowledge in the fine-tuned mBART50 *teacher* model to build a lightweight *student* model that has a smaller number of parameters. In this regard, our pipeline considers mBART50 as an initial resource-hungry model which is conveniently exploited to generate synthetic parallel sentences that are conveniently filtered before training a smaller student NMT system that can then be run on edge computing devices such as desktop computers or smartphones.

This pipeline consists of two different stages: a first stage aimed at improving the pre-trained models by combining iterative back-translation, parallel corpus filtering and fine-tuning; and a

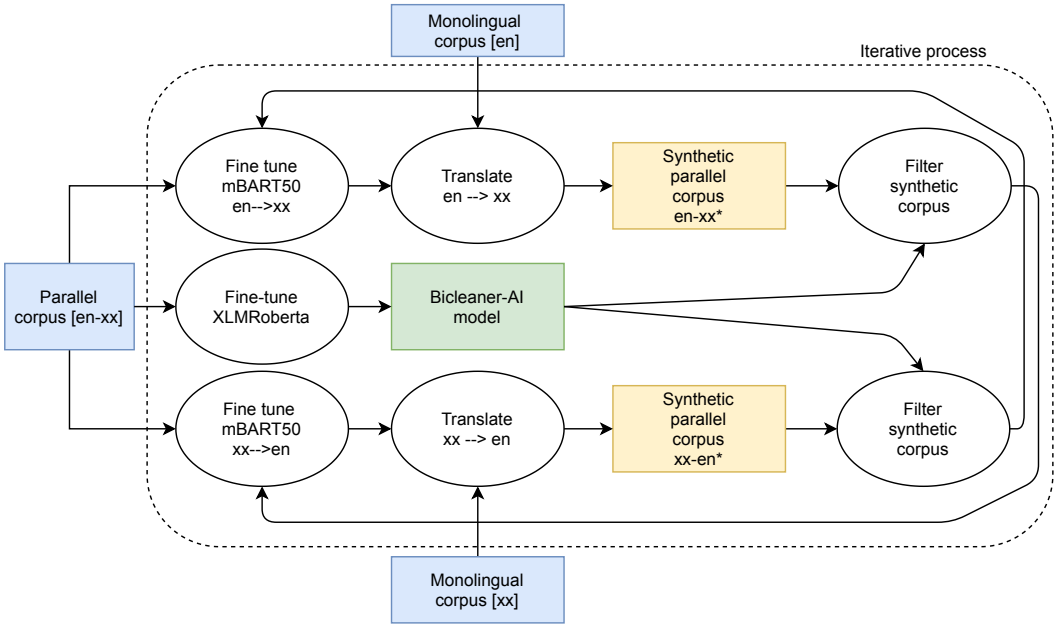

Figure 2: Fine-tuning of mBART50 to translate English (en) into a low-resource language (xx), and vice versa, using parallel and monolingual corpora.

second stage aimed at distilling the knowledge from the fine-tuned models to train a student model with far fewer parameters but comparable performance.

Fine-tuning of pre-trained models. This process, depicted in Figure 2, combines fine-tuning of the pre-trained models with back-translation (Hoang et al., 2018) and synthetic parallel corpus filtering via a fine-tuned XLM-R model (Conneau et al., 2020). For our English-centric scenario and a particular low-resource language, this consists of the following steps:

1. Use the available parallel corpora[1] and a 1:10 positive to negative ratio[2] to fine-tune XLM-R using Bicleaner-AI (Zaragoza-Bernabeu et al., 2022). Bicleaner-AI learns a classifier on top of XLM-R that predicts whether a pair of input sentences are mutual translation or not. Positive training samples are obtained from the aligned sentences in the bilingual corpus, whereas negative samples are obtained by randomly selecting non-aligned sentences.
2. Fine-tune both the English-to-many and the many-to-English mBART50 models with the original parallel corpora.
3. Perform iterative back-translation starting with one million monolingual sentences in each language (if such amount of sentences is not available, use the amount of monolingual corpora available):
   (a) Translate the English monolingual corpora into the low-resource language, and vice versa, using the last fine-tuned mBART50 models.
   (b) Filter the synthetic corpora using the XLM-R model trained in step 1.

[1] In our experiments, we have used no more than 600,000 parallel sentences due to time constraints.
[2] According to the Bicleaner-AI web page, these values are inspired on the winner approach of the WMT 2020 parallel corpus cleaning task (Açarçiçek et al., 2020).

(c) Use the filtered synthetic corpora together with the available parallel corpora to update the last fine-tuned mBART50 models translating to and from English.

(d) Evaluate the performance of the two resulting models on a development set. If none of them gets improved, stop the iterative process. Otherwise, add 1 million sentences in each language (if available) to the monolingual corpora and jump to step 3(a).

In order to filter the synthetic corpora generated in each iteration, a threshold in the interval [0,1] is used to discretize the output of the Bicleaner-AI classifier. This threshold is set in the first iteration of the back-translation process —step 3(b)— by exploring all the thresholds in the interval [0.2,0.9] at steps of 0.1. The threshold for the remaining iterations is the one that produces the synthetic corpus that leads to the best mBART50 models according to the development set.

Training of student models. Knowledge distillation is usually implemented in NLP at token level (Tan et al., 2019; Shleifer and Rush, 2020), but in tasks like NMT performing it at sequence level (Kim and Rush, 2016) is usually equivalent and easier to implement: the *student* is trained on a synthetic corpus obtained by translating the source segments of the original training parallel corpus with the *teacher*. Knowledge distillation is therefore usually carried out by exploiting the same training corpus used when training the teacher. However, in the case of third-party-developed pre-trained models, this corpus is not necessarily available. In its absence, as well as for languages never seen by pre-trained models, we can generate synthetic training samples by translating monolingual data with the teacher model and then filtering the synthetic data generated to discard low-quality or noisy sentence pairs.

Once the pre-trained models have been properly fine-tuned, we train a student model by performing standard sentence-level knowledge distillation (Kim and Rush, 2016). To this end, monolingual English data is automatically translated into the low-resource language with the best fine-tuned English-to-many mBART50 system and the resulting synthetic bilingual corpus (opportunely cleaned with the same Bicleaner-AI model) together with the true bilingual corpus are used to train the student model translating the low-resource language into English. Conversely, monolingual data available for the low-resource language is automatically translated into English with the best fine-tuned many-to-English mBART50 model and the resulting cleaned corpus together with the bilingual corpus are used to train the system translating from English into the low-resource language.

Summary of results. Our pipeline is evaluated on eight translation tasks involving four low-resource languages of interest to the project, and English: Swahili, Kyrgyz, Burmese and Macedonian. In order to evaluate the transferability of the pre-trained model to unseen languages, two of our languages (Swahili and Kyrgyz) were not considered during mBART50's pre-training. Languages were chosen so that each one belongs to a different linguistic family.

The results show two different trends, depending on whether English is the source or the target language. When English is the target language, the difference in performance between the students trained using both forward and backward translations generated with the teacher model, and the teacher is notably larger than when English is the source language. This is clearly motivated by the fact mBART50 is an English-centric model pre-trained on a large quantity of English texts and thus excelling in English over the other languages. On the contrary, when English is the

source language, the student models outperform the teacher model by a small margin or perform comparably.

The best student models consistently improve the results of the bilingual baselines by a wide margin thus confirming the appropriateness of considering large pre-trained models as the seed for NMT models and the effectiveness of our pipeline. A comparison between our models and two prominent multilingual models, namely M2M-124 (Goyal et al., 2021; Wenzek et al., 2021) and DeltaLM+Zcode (Yang et al., 2021), the baseline and winner system at WMT 2021, respectively, show that the student models perform considerably better than DeltaM+Zcode when the target language is not English, except for English–Macedonian. When the target language is English, DeltaM+Zcode clearly outperforms the teacher and student models. Our students are noticeably smaller, but note that both M2M-124 and DeltaLM+Zcode are one-size-fits-all models which have not been bilingually fine-tuned.

As regards the use of Bicleaner-AI for filtering noisy synthetic parallel segments, its use has demonstrate to improve translation quality, as the teacher model may be producing a larger degree of hallucinations when translating into English than the other way round; something that we plan to further investigate in the future.

Finally, the student models are much faster than the teacher models: on one GPU NVIDIA A100, the students are 61% faster than the teachers, whereas on an Intel i5 CPU at 2.9 GHz, the students are 92% faster than the teachers. Note that the student models have 13 times fewer parameters that the teacher models.

3.2 Exploring Unsupervised Pretraining Objectives for Machine Translation

This work has been published in Findings of ACL 2021 (Baziotis et al., 2021).

Neural machine translation (NMT) is notoriously data-hungry (Koehn and Knowles, 2017). To learn a strong model it requires large, high-quality and in-domain parallel data, which exist only for a few language-pairs. The most successful approach for improving low-resource NMT is back-translation (Sennrich et al., 2016), that exploits abundant monolingual corpora to augment the parallel with synthetic data. However, in low-resource settings, it may fail to improve or even degrade translation quality if the initial model is not strong enough (Imankulova et al., 2017; Burlot and Yvon, 2018).

Unsupervised pretraining is a complementary technique, that has revolutionized many natural language understanding (NLU) tasks (Wang et al., 2019). The dominant approach is to train a (large) model on a lot of unlabeled data using the masked language modeling (MLM; Devlin et al. (2019)) objective and then finetune it on a downstream task. Besides improving generalization, good initialization drastically reduces the need for labelled data. This paradigm has been applied recently to NMT yielding impressive results in low-resource settings, with models such as XLM (Conneau and Lample, 2019), MASS (Song et al., 2019) and BART/mBART (Lewis et al., 2020; Liu et al., 2020a), that adapt MLM to sequence-to-sequence architectures. Although pretraining alone is not enough to outperform backtranslation, it helps the initial model to produce synthetic data of sufficient quality, and combining them yields further improvements.

Most prior work in pretraining has focused on optimizing the masking strategy (Rogers et al., 2021). Similarly, MASS and mBART consider slightly different masking strategies. However, due to differences in their experimental setup (i.e., capacity or training data) and lack of analysis that goes beyond evaluation on downstream tasks, it is unclear if there is a meaningful difference

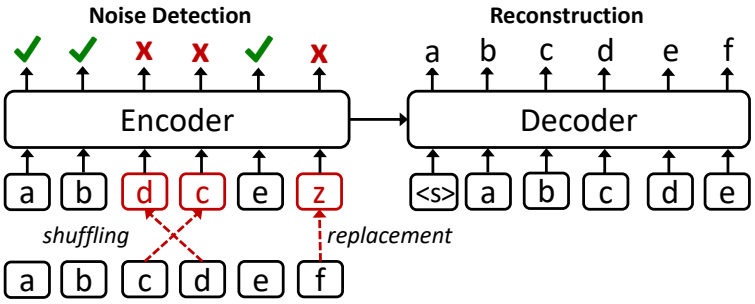

Figure 3: We consider noising methods that produce inputs which resemble real sentences, unlike masking.

between them, as far as NMT is concerned. They also suffer from a pretraining-finetuning discrepancy (Yang et al., 2019), in which a model is pretrained on masked inputs, but finetuned on full sentences.

In this work, we explore different objectives to masking for unsupervised cross-lingual pretraining. We inject noise that creates examples (Fig. 3), similar to those encountered in finetuning, unlike masking. This includes, randomly replacing input words based on their context using a cross-lingual generator, inspired by Clark et al. (2020), and locally reordering input words, which prevents the cross-attention from naively (monotonically) attending over the source. We also explore auxiliary losses over the encoder to improve its representations.

First, we pretrain models with different configurations, on English-German, English-Nepali and English-Sinhala monolingual data. Then, we *systematically* compare them on the downstream tasks of supervised, semi-supervised and unsupervised NMT. In (semi-) supervised NMT, we observe that models yield surprisingly similar results, although some methods are better than others. We find that even pretraining with shuffled inputs leads to significant improvements over random initialization, similar to the concurrent work of Sinha et al. (2021) on pretrained encoders for NLU. Unsupervised NMT, however, reveals large (up to 9 BLEU points) differences, and against our expectations, masking achieves the best performance. To understand these results, unlike prior work, we thoroughly analyze the pretrained models using a series of probes, and discover that each objective drives the models to encode and use information in unique ways.

Based on our findings, we conclude that each finetuning process is sensitive to specific properties of pretrained models, similar to Artetxe et al. (2020). We hypothesize that (semi-) supervised NMT is mostly sensitive to the LM abilities of pretrained models, as the source→target mappings can be learnt from the parallel data. Unsupervised NMT requires models to also rely on their own word-translation abilities. Our contributions are:

1. We *systematically* compare many pretraining methods, including alternatives to masking, in three NMT tasks and for three language-pairs.

2. We discover that (semi-) supervised NMT is not sensitive to the pretraining strategies. Our ablation suggests that a strong decoder is the most important factor, while differences in the encoder don't affect the results.

3. Unsupervised setting is much more sensitive to the pretraining objective, and masking methods are the most effective. We hypothesise that learning to copy is important here as is cross-lingual encoding.

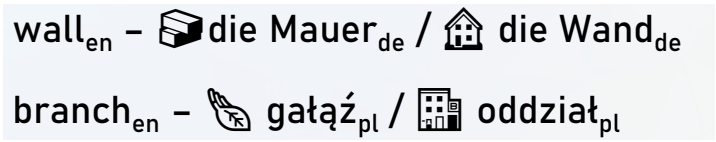

Figure 4: Word sense disambiguation example where one English word can have different translations depending on the context in German or Polish

4. We analyze the pretrained models with a series of probes, and show noticeable differences in how they encode and use information, offering valuable insights.

4 Task 3: Learning from Lexical Resources

In this section, we describe our effort to incorporate lexical resources into NMT. These resources may be especially useful in low-resource scenarios, where they can supplement the low amount of availabel parallel data. In the following section, we describe an ongoing study of using knowledge graps for English–German translation.

Apart from the research described here, Section 8.3 of D5.5 describes a method for incorporating terminology lists into English–Turkish translation.

4.1 Knowledge Graphs in Neural Machine Translation

This is ongoing work between Edinburgh and Mateusz Klimaszewski from the Warsaw University of Technology.

NMT methods require large scale datasets with parallel sentences in a source and target languages. However, NMT quality suffers when the data is unavailable, or the translation is out-of-domain. Following the successful work in the text generation field (Yu et al., 2022), we aim to include Knowledge Graphs to improve the mentioned shortcomings of NMT. Knowledge Graphs are specific knowledge bases intended to extract and structure human knowledge. Formed as a graph, Knowledge Graphs allow AI systems to perform complex reasoning leveraging organised data. In the experiments, we studied the impact of the shallow representation of a KG, Knowledge Graph Embeddings (Bordes et al., 2013; Trouillon et al., 2016), on a task which requires in-depth natural language understanding - Word Sense Disambiguation. Our preliminary study demonstrates improvements in English to German translation on out-of-domain datasets.

The problem of Word Sense Disambiguation requires in-depth natural language understanding. The task is defined as choosing the correct meaning (in our case – translation) given the context – usually as a sentence. We derive possible translations based on the Knowledge Graph neighbours, while the gold translation is extracted from parallel corpora.

Our solution uses a pre-trained language model alongside Knowledge Graph Embeddings (KGEs) to determine which of the extracted translations from the KG should be picked. Afterwards, the chosen translation is incorporated into the source sentence with additional tags or source factors. The system is described in Figure 5.

We can see in Table 1 results from evaluating our English-to-German translation model, including two out-of-domain datasets: medical – Himl and Reddit posts from Common Voices'. The automatic evaluation showcased improvement in 4 out of 6 cases.

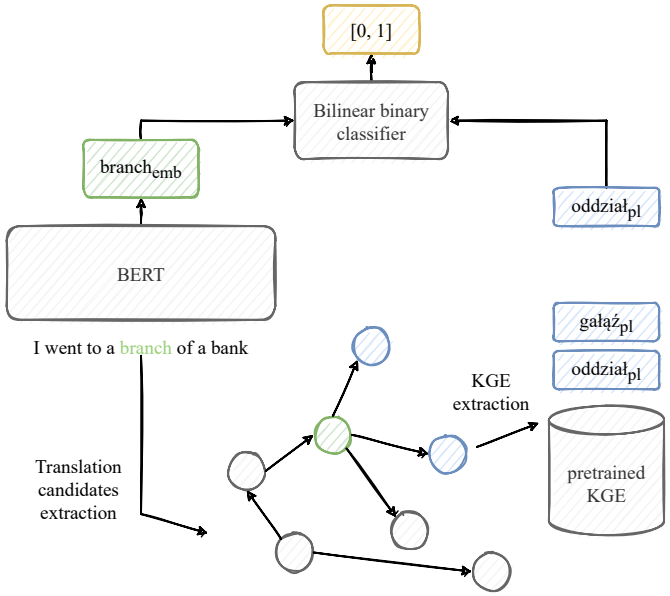

Figure 5: Architecture of our KG enhanced NMT system

Model	BLEU	chrF	COMET
		Himl	
Baseline	29.23	0.575	0.3452
Ours	29.07	0.578	0.3894
		Common Voices	
Baseline	25.63	0.527	0.3464
Ours	25.28	0.528	0.3667

Table 1: Results for two out-of-domain datasets: HimL and Common Voices

5 Publications

These papers are the result of research done in transfer learning in the second half of the GoURMET project.

- Nikita Moghe, Mark Steedman, and Alexandra Birch. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.87. URL https://aclanthology.org/2021.emnlp-main.87
- Aarón Galiano, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, and Juan Antonio Pérez-Ortiz. Exploiting large pre-trained models for low-resource neural machine translation. Submitted to *The 29th International Conference on Computational Linguistics (COLING 2022)*.
- Christos Baziotis, Ivan Titov, Alexandra Birch, and Barry Haddow. Exploring unsupervised pretraining objectives for machine translation. In *Findings of the Association for*

Computational Linguistics: ACL-IJCNLP 2021, pages 2956–2971, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.261. URL https://aclanthology.org/2021.findings-acl.261

6 Software

Here is the list of released software as part of the output of this workpackage.

- The code used for Cross-Lingual Intermediate Fine-Tuning for Dialogue State Tracking is available at https://github.com/nikitacs16/xlift_dst. We provide the script to train new intermediate models as well as list the links to our trained intermediate models that are hosted on HuggingFace repository. It also provides the code for training the dialogue state trackers.
- The code used for the experiments on the exploitation of large pre-trained models is available at https://github.com/transducens/tune-n-distill.

7 Conclusion

In this deliverable we describe our contributions to the three tasks in Work Package 4 addressing transfer learning methods for low-resource MT. In the first task focused on learning from multilingual data, we introduced a method to enhance the pre-training of multilingual models with code-switching, and a fine-tuning method applied to multilingual dialogue systems. The goal of the second task is to develop methods to leverage monolingual data. We contribute to this task by exploring fine-tuning of large pre-trained language and translation models, such as mBART50 or M2M-100. The third task focuses on using lexical resources. In this task, we described a method of incorporating knowledge graphs into neural machine translation.

References

Haluk Açarçiçek, Talha Çolakoğlu, Pınar Ece Aktan Hatipoğlu, Chong Hsuan Huang, and Wei Peng. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.105.

Mikel Artetxe, Gorka Labaka, Noe Casas, and Eneko Agirre. Do all roads lead to rome? understanding the role of initialization in iterative back-translation. *Knowledge-Based Systems*, 206:106401, October 2020. ISSN 0950-7051. doi: 10.1016/j.knosys.2020.106401. URL http://dx.doi.org/10.1016/j.knosys.2020.106401.

Christos Baziotis, Ivan Titov, Alexandra Birch, and Barry Haddow. Exploring unsupervised pretraining objectives for machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2956–2971, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.261. URL https://aclanthology.org/2021.findings-acl.261.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

Franck Burlot and François Yvon. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6315. URL https://aclanthology.org/W18-6315.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. Towards making the most of multilingual pretraining for zero-shot neural machine translation. In *Proceedings of ACL*, 2022.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. XL-NBT: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.125. URL https://aclanthology.org/2021.emnlp-main.125.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.181. URL https://aclanthology.org/2021.eacl-main.181.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021. URL http://jmlr.org/papers/v22/20-1307.html.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193, 2021. URL https://arxiv.org/abs/2106.03193.

R. Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David R. Traum, Maxine Eskénazi, Ahmad Beirami, Eunjoon Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. Overview of the ninth dialog system technology challenge: DSTC9. *CoRR*, abs/2011.06486, 2020. URL https://arxiv.org/abs/2011.06486.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. Association

for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://aclanthology.org/W18-2703.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/W17-5704.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. 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 3336–3341, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1328. URL https://aclanthology.org/D19-1328.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://www.aclweb.org/anthology/W17-3204.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?, 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Yen-Ting Lin and Yun-Nung Chen. An empirical study of cross-lingual transferability in generative dialogue state tracker, 2021.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.210. URL https://aclanthology.org/2020.emnlp-main.210.

P. Lison and J. Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*, 2016.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020a. doi: 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8433–8440. AAAI Press, 2020b. URL https://aaai.org/ojs/index.php/AAAI/article/view/6362.

Zihan Liu, Genta Indra Winata, and Pascale Fung. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.239. URL https://aclanthology.org/2021.findings-acl.239.

Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. Cross-lingual dialogue dataset creation via outline-based generation, 2022. URL https://arxiv.org/abs/2201.13405.

Nikita Moghe, Mark Steedman, and Alexandra Birch. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.87. URL https://aclanthology.org/2021.emnlp-main.87.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017. doi: 10.1162/tacl_a_00063. URL https://aclanthology.org/Q17-1022.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.21. URL https://aclanthology.org/2021.acl-long.21.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on*

Natural Language Processing, pages 557–575, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.56.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860. ijcai.org, 2020. doi: 10.24963/ijcai.2020/533. URL https://doi.org/10.24963/ijcai.2020/533.

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo M. Ponti, Anna Korhonen, and Ivan Vulić. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems, 2021.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8(0):842–866, 2021. ISSN 2307-387X. URL https://transacl.org/index.php/tacl/article/view/2257.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1380. URL https://aclanthology.org/N19-1380.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Sam Shleifer and Alexander M. Rush. Pre-trained summarization distillation. *CoRR*, abs/2010.13002, 2020. URL https://arxiv.org/abs/2010.13002.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, J. Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *ArXiv*, abs/2104.06644, 2021.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 5926–5936, Long Beach, California, USA, 2019. PMLR. URL http://proceedings.mlr.press/v97/song19d.html.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *Seventh International Conference on Learning Representations*, New Orleans, USA, May 2019. URL https://openreview.net/forum?id=S1gUsoR9YX.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401, 2020. URL https://arxiv.org/abs/2008.00401.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. Cross-lingual retrieval for iterative self-supervised training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc., 2020.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook AI WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages "205—-215", 2021.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2071–2080. JMLR.org, 2016.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.2.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.54.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.208. URL https://aclanthology.org/2020.emnlp-main.208.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5753–5763, 2019.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.*, jan 2022. ISSN 0360-0300. doi: 10.1145/3512467. URL https://doi.org/10.1145/3512467. Just Accepted.

Jaume Zaragoza-Bernabeu, Marta Bañón, Gema Ramírez-Sánchez, and Sergio Ortiz-Rojas. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2022.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D4.2 GoURMET Final Report on Transfer Learning