Global Under-Resourced MEdia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D4.1 – Initial progress report on transfer learning

Nature	Report	**Work Package**	WP4
Due Date	30/6/2020	**Submission Date**	30/6/2020
Main authors	Antonio Valerio Miceli Barone (UEDIN)		
Co-authors	Felipe Sánchez-Martínez (UA), Mikel L. Forcada (UA), Bryan Eikema (UVA), Wilker Aziz (UVA), Rachel Bawden (UEDIN), Christos Baziotis (UEDIN), Arturo Oncevay (UEDIN), Ivan Titov (UEDIN)		
Reviewers	Juan Antonio Pérez-Ortiz (UA)		
Keywords	transfer learning, unsupervised, semi-supervised		
Version Control			
v0.1	**Status**	Draft	21/6/2020
v1.0	**Status**	Final	30/6/2020

Contents

1 Introduction **4**

2 Task 1: Learning from Multilingual Data **4**

2.1 Bridging Linguistic Typology and Multilingual Machine Translation with Multi-view Language Representations 5

2.2 Synthetic data generation through pivoting 6

2.3 Multilingual Neural Machine Translation and Zero-Shot Translation 7

3 Task 2: Learning from Monolingual Corpora **8**

3.1 Semi-supervised NMT (WMT English↔Gujarati, English↔Tamil) 9

3.1.1 English↔Gujarati 9

3.1.2 English↔Tamil 9

3.2 Language Model Prior for Low-Resource Neural Machine Translation 11

3.3 Cross-lingual word-embeddings induction 13

3.3.1 Cross-lingual embedding alignment with normalizing flows 14

3.3.2 Attention-based joint crosslingual embedding training 16

3.4 Probabilistic back-translation 17

3.5 Co-Training 19

4 Task 3: Learning from Lexical Resources **20**

4.1 Multi-source approach for exploiting linguistic resources 21

4.2 Exploiting bilingual lexicons in Neural Machine Translation 23

5 Publications **24**

6 Software **25**

7 Conclusion **25**

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 and our plans for the remainder 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. The little parallel data available often consists of government documents, religious texts or software manuals, which are out of the domain of news articles, or text scraped from web pages which is not only scarce but often of poor quality due to the difficulty of automatically extracting parallel data from the web.

This workpackage 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. The workpackage 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.
- **T1: Learning from Monolingual Corpora**
  We leverage corpora of monolingual text, which is much more abundant than parallel text, especially for the news domain.
- **T1: Learning from Lexical Resources**
  We make use of curated linguistic resources such as large bilingual dictionaries.

The approaches investigated in these tasks can be combined with each other, as we demonstrated for instance in our English-Gujarati system that combined multilingual and monolingual techniques to achieve state-of-the-art results in the WMT 2019 News Translation shared task (Barrault et al., 2019).

2 Task 1: Learning from Multilingual Data

Shared ancestry and cultural exchanges throughout history have resulted in extensive lexical and syntactical similarities between many languages. More speculatively, but plausibly given the available evidence, the shared aspects of the human psychology and the human experience result in deep structural and statistical similarities between all human languages. This is beneficial when designing a machine translation system for a low-resource language pair because it makes it possible to leverage resources for other languages, which help shape the *inductive bias* of the machine learning approaches that we use in a direction that improves their ability to learn from the limited parallel data available. This is the focus of this workpackage task.

In the GoURMET project initial proposal for task T4.1 we had mentioned baseline techniques such as pre-training a machine translation system on a high-resource language pair and then fine-tuning it on a low-resource pair, or jointly training on multiple language pairs using a language-indicator tag to select the output language. We have successfully applied and improved upon these techniques in the work done so far.

We had also discussed carrying out an analysis of the type of knowledge being transferred between languages (e.g. shallow transfer vs. interlingua semantic space), which we will reserve as further work for the next part of the project, while in the work done so far we have carried out an analysis of the quality of transfer between different languages in relation to their similarity.

2.1 Bridging Linguistic Typology and Multilingual Machine Translation with Multi-view Language Representations

Recent surveys consider linguistic typology as a potential source of knowledge to support multilingual natural language processing (NLP) tasks (O'Horan et al., 2016; Ponti et al., 2019). Linguistic typology studies language variation in terms of their functional processes (Comrie, 1989). Several typological knowledge bases (KB) have been crafted, from where we can extract categorical language features[1], such as in the lang2vec tool (Littell et al., 2017). Nevertheless, their sparsity and reduced coverage present a challenge for an end-to-end integration into NLP algorithms. For example, the World Atlas of Language Structure (WALS; Dryer and Haspelmath, 2013) encodes 143 features for 2,679 languages, but their mean coverage per language is barely around 14%.

Dense and data-driven language representations have emerged in response. They are computed from multilingual settings of language modelling (Östling and Tiedemann, 2017) and neural machine translation (NMT) (Malaviya et al., 2017). However, the language diversity in the corpus-based representations is limited. The language coverage could be broadened with other knowledge, such as that encoded in WALS, to distinguish even more language properties. Therefore, to obtain the best of both views (KB and task-learned) with minimal information loss, we project a shared space of discrete and continuous features using Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al., 2017).

Canonical correlation analysis (CCA) allows us to find a projection of two views for a given set of data. With CCA, we look for linear combinations that maximise the correlation of the two sources in each coordinate iteratively. CCA considers all dimensions of the two views as equally important. However, our sources are potentially redundant: KB features are mostly one-hot-encoded, whereas task-learned ones inherit the high dimensionality of the embedding layer. Moreover, few samples and sparsity could make the convergence harder. For the redundancy issue, singular value decomposition (SVD) is an appealing alternative. With SVD, we factorise the source data matrix to compute the principal components and singular values. The two-step transformation of SVD followed by CCA is called SVCCA (Raghu et al., 2017) in the context of understanding the representation learning throughout neural network layers. That being said, we use SVCCA to get language representations and not to inspect a neural architecture.

In the work by Oncevay et al. (2019), we fuse language-level embeddings from multilingual machine translation with syntactic features of WALS. We inspect how much typological knowledge is present by predicting features for new languages. Then, we infer language phylogenies (Rabinovich et al., 2017) and inspect whether specific relationships are induced from the task-learned vectors. Furthermore, to demonstrate that our approach has practical benefits in NLP, we apply our language vectors in multilingual NMT with language clustering (Tan et al., 2019) and adapt the ranking of related languages for multilingual transfer (Lin et al., 2019).

We list our key findings as follows:

[1] An example of a typological feature is a word order specification, like whether the adjective is predominately placed before or after the noun.

- SVCCA can fuse linguistic typology KB entries with NMT-learned embeddings without diminishing the originally encoded typological and genetic similarity of languages: By assessing an typological feature prediction task, we identified that the multi-view vectors outcome their single counterparts in most of the cases. Similarly, the SVCCA-based embeddings allow a better reconstruction of a phylogenetic tree of 17 Indo-European languages.

- Our representations provides a robust alternative to determine which related languages are suitable for multilingual transfer learning in clustering or ranking related languages. The notable advantage is that we do not need to pre-train MT systems from a specific dataset (e.g. like in LangRank (Lin et al., 2019)), and we can easily extend the coverage of languages without re-training a ranking model to consider new language entries, or pre-train multilingual language embeddings for a massive model.

- Factored language embeddings encodes more information to agglomerate related languages than an initial pseudo-token setting: With factors, the embedding of every input token was concatenated with the embedded pseudo-token that identifies the source language, whereas an initial pseudo-token setting only adds the source language at the beginning of every input sentence. The latter are not suitable for clustering, and they might only encode enough information to perform a classification task.

Furthermore, we are building an open-source tool to compute multi-view language representations using SVCCA. We enable the option to use any kind of language vectors from `lang2vec` (Syntax, Phonology or Phonetic Inventory) as a KB-source, and to upload new task-learned embeddings from different settings, such as many-to-one, one-to-many or many-to-many NMT and multilingual language modelling. Besides, given a list of languages to assess, our method will project new language representations when they are only available in the KB-view. Finally, we include the tasks of language clustering and ranking candidates, which could benefit multilingual NLP studies that involves massive datasets of hundreds of languages (Zhang et al., 2020).

2.2 Synthetic data generation through pivoting

When designing a machine translation system between a high-resource language such as English and a distant low-resource language, it is not uncommon for there to be a greater amount of parallel data available between another related high or medium resource language and English. This is the case when the ”pivot” language has a larger number of speakers or greater political importance than our low-resource language of interest to which is related by geographical or cultural proximity.

Our case study for this scenario is our English-Gujarati system (Bawden et al., 2019) where we were able to successfully use Hindi as a “pivot” medium-resource language.

Hindi is a widely spoken language, which, like Gujarati, belongs to the Indo-Aryan family. The two languages are closely related in terms of lexicon and syntax, and they are written using variants of the Devanagari script, which while encoded by different Unicode characters, can be easily transliterated into each other[2]. While we were able to access only a small amount of Hindi-Gujarati parallel data (approximately 8,000 sentence pairs from the Emille corpus[3]), we could exploit large amounts of monolingual data for both languages to train a Hi→Gu machine translation system

[2] Technically, transliteration from Hindi characters to Gujarati characters is a surjective function.
[3] https://www.lancaster.ac.uk/fass/projects/corpus/emille/

using the semi-supervised variant of XLM (Lample and Conneau, 2019), which turned out to be particularly effective given the similarity between the languages. We did not attempt fully unsupervised translation because we intended to exploit the parallel data in order to obtain the maximum possible translation quality. We used this system to translate approximately 1.1 million Hindi sentences from a Hindi-English parallel corpus into Gujarati. Since each of these sentences is aligned to an English sentence, this resulted in a parallel corpus consisting of 1.1 million natural English–synthetic Gujarati sentence pairs, which we used as backtranslations to train our Gu→En model and as forward translations to train our En→Gu model. Further description of this system is provided in Section 3.1 and in Deliverable 5.3 Section A.2.

We submitted our system to the 2019 Conference on Machine Translation (WMT19) news translation shared task, achieving the second-best position for Gu→En and first position for En→Gu among the constrained submissions.

We further developed synthetic data generation through pivoting as an online procedure applied during a fine-tuning stage of a massively multilingual machine translation system (Zhang et al., 2020), described in the next section.

2.3 Multilingual Neural Machine Translation and Zero-Shot Translation

In multilingual translation, a single NMT model is optimized for the translation of multiple language pairs (Firat et al., 2016a; Johnson et al., 2017b; Lu et al., 2018; Aharoni et al., 2019). Multilingual NMT eases model deployment and can encourage knowledge transfer among related language pairs (Lakew et al., 2018; Tan et al., 2019), improve low-resource translation (Ha et al., 2016; Arivazhagan et al., 2019), and enable zero-shot translation (i.e. direct translation between a language pair never seen in training) (Firat et al., 2016b; Johnson et al., 2017b). The last two goals are especially relevant in the GoURMET context.

Despite these potential benefits, multilingual NMT tends to underperform its bilingual counterparts (Johnson et al., 2017b; Arivazhagan et al., 2019) and results in considerably worse translation performance when many languages are accommodated (Aharoni et al., 2019). Since multilingual NMT must distribute its modeling capacity between different translation directions, we ascribe this deteriorated performance to the deficient capacity of single NMT models and seek solutions that are capable of overcoming this capacity bottleneck (as simply increasing model size might not scale to the quadratic number of language pairs and translation directions).

In Zhang et al. (2020), we proposed language-aware layer normalization and linear transformation to relax the representation constraint in multilingual NMT models. Language-aware layer normalization makes the gain and bias parameters of the layer normalization blocks of the neural network dependent on the language rather than being shared between all the language as the parameters of a multi-lingual NMT model usually are[4]. The language-aware linear transformation further extends this specialization by inserting a language-dependent matrix between the encoder and the decoder so as to facilitate the induction of language-specific translation correspondences. We also investigated very deep NMT architectures with specialized initialization schemes (Wang et al., 2019; Zhang et al., 2019) aiming at further reducing the performance gap with bilingual methods.

Another pitfall of massively multilingual NMT is its poor zero-shot performance, particularly compared to pivot-based models. Without access to parallel training data for zero-shot language pairs,

[4] This only slightly increases the total number of parameters of the system since gain and bias parameters have a size proportional to model width rather than the square of model width as with matrices.

multilingual models easily fall into the trap of *off-target translation* where a model ignores the given target information and translates into a wrong language. To avoid such a trap, in Zhang et al. (2020), we proposed the random online backtranslation (ROBt) algorithm. ROBt finetunes a pre-trained multilingual NMT model for unseen training language pairs with pseudo parallel batches generated by back-translating the target-side data during training, while language pairs are randomly sampled.

We performed backtranslation (Sennrich et al., 2016a) into randomly picked intermediate languages to ensure good coverage of ∼10,000 zero-shot directions. Although backtranslation has been successfully applied to zero-shot translation (Firat et al., 2016b; Gu et al., 2019; Lakew et al., 2019), whether it works in the massively multilingual set-up remained an open question and we investigated it in our work (Zhang et al., 2020).

For the experiments, we collected OPUS-100, a massively multilingual dataset sampled from OPUS (Tiedemann, 2012). OPUS-100 consists of 55M sentence pairs between English and 99 other languages. As far as we know, no similar dataset is publicly available. We have released OPUS-100 to facilitate future research.[5] We adopted the Transformer model (Vaswani et al., 2017a) and evaluated our approach under one-to-many and many-to-many translation settings. Our main findings are summarized as follows:

- Increasing the capacity of multilingual NMT yields large improvements and narrows the performance gap with bilingual models. Low-resource translation benefits more from the increased capacity.
- Language-specific modeling and deep NMT architectures can slightly improve zero-shot translation, but fail to alleviate the off-target translation issue.
- Finetuning multilingual NMT with ROBt substantially reduces the proportion of off-target translations (by ∼50%) and delivers an improvement of ∼10 BLEU points in zero-shot settings, approaching the conventional pivot-based method. We show that finetuning with ROBt converges within a few thousand steps.

3 Task 2: Learning from Monolingual Corpora

Monolingual text is usually much more readily available than parallel text for any language pair, which makes it an attractive resource to leverage when training a machine translation system, or really any natural language processing application (Devlin et al., 2018). Monolingual text is already beneficial for high-resource language pairs, as shown by Sennrich et al. (2016a) in their seminal work on *backtranslation*, but becomes of paramount importance for low-resource language pairs, where parallel data is scarce and usually out-of-domain and noisy.

In the GoURMET project proposal we had mentioned backtranslation and other standard semi-supervised baselines. We had also referenced cross-lingual word embedding induction (Miceli Barone, 2016) and the recently developed unsupervised translation techniques (Artetxe et al., 2018; Lample et al., 2018) which we planned to use in semi-supervised mode combining parallel and monolingual data. In the work carried out so far we applied these techniques and researched improvements and alternative methodologies, in synergy with the multilingual techniques described in section 2.

[5] https://github.com/EdinburghNLP/opus-100-corpus

3.1 Semi-supervised NMT (WMT English↔Gujarati, English↔Tamil)

3.1.1 English↔Gujarati

For our English↔Gujarati (Bawden et al., 2019) system we made ample use of semi-supervised machine translation techniques in order to leverage the available monolingual data.

We used the XLM cross-lingual pretraining approach of Lample and Conneau (2019), which consists of training a masked language model (much like BERT (Devlin et al., 2018)) on both monolingual data and parallel English and Gujarati data, and then intialising both the encoder and decoder of a multi-way English-Gujarati machine translation model with the pretrained parameters of the language model. The machine translation model and the additional translation-specific parameters are then fine-tuned using machine translation objectives. We chose to use semi-supervised machine translation learning for this step, which makes the most of both available parallel data but also monolingual data in both languages by alternating machine translation, auto-encoding and backtranslation steps.

We trained both an initial English↔Gujarati model to produce backtranslations (Sennrich et al., 2016a) and a (transliterated) Hindi↔Gujarati model to produce synthetic data through pivoting (Section 2.2). This synthetic data was added to the natural parallel data to train our final models, resulting in top-level translation quality. Specifically, in the WMT19 news translation shared task human evaluation, we obtain the best result in the English→Gujarati direction and the second best result in the Gujarati→English direction among the constrained system submissions. See also Deliverable D5.3 Section A.2 for a general description of this system.

3.1.2 English↔Tamil

Based on the large improvements brought by exploiting semi-supervised MT and backtranslation for English↔Gujarati at the WMT19 news shared task, we adopted a similar strategy for English↔Tamil at the WMT20 shared task. Other than the shared task submission, we also plan to conduct more extensive experiments looking at the best method of combining backtranslations produced by different systems of different quality. This presents ongoing work, and therefore results of our investigations will be presented at the end of the project.

Like English–Gujarati, English–Tamil (EN-TA) is a low-resource language pair that presents challenges for standard supervised MT due to the lack of data but also the rich morphology of Tamil, which makes translation into Tamil particularly difficult. Using the available monolingual and parallel data, we compare a range of different MT models trained in different ways and using different types of data (either one for each direction or a joint model):

- **Moses baseline**: a phrase-based Moses model (Koehn et al., 2007) trained on parallel data segmented using SentencePiece (Kudo and Richardson, 2018) (vocabulary of size 20k). We also use 5-gram language models into English, lexicalised reordering and 5-gram operation sequence modelling.
- **Marian baseline**: transformer-based baselines trained on parallel data using Marian (Junczys-Dowmunt et al., 2018). The best subword segmentation vocabulary sizes are 7.5k for TA→EN and 2.5k for EN→TA, illustrative of the fact that more segmentation (smaller vocabulary) is

beneficial for more morphologically rich languages.[6]

- **Marian multilingual**: many-to-one and one-to-many multilingual Marian models (small transformers with two layers) on multilingual parallel Indian data. We use Telugu and Hindi as additional languages as they provide most gains in the multilingual setting. We preprocess using a 20k SentencePiece model for all language pairs transliterated into the Tamil script.
- **XLM pretraining**: As for En↔GU, an XLM model (Conneau and Lample, 2019), which involves pretraining a transformer-based model using a cross-lingual language modelling objective (with monolingual and parallel data),[7] before fine-tuning to MT. We choose to fine-tune using a semi-supervised MT objective (using both monolingual and parallel data). We preprocess using a SentencePiece vocabulary of 20k for both directions.
- **Marian mBART pretraining**: a second model using language model pretraining, this time trained used Marian and using the mBART objective (Liu et al., 2020), which differs from XLM in that it does fully sequence-to-sequence pretraining of the entire encoder-decoder model, whereas XLM performs masked language modelling and does not pretrain the whole model as one. We again use a SentencePiece vocabulary of 20k for both directions.
- **Marian DE–EN pretraining**: Marian transformer-based models pretrained on translations for an unrelated but higher resource language pair. We choose to pretrain on DE–EN translation, using a vocabulary that is shared between English, German and Tamil, in order for those parameters to be fine-tuned afterwards on English-Tamil translation.

The initial results of these individual models (all models fine-tuned on parallel EN–TA data), as measured using the standard automatic metric BLEU (Papineni et al., 2002) using the SacreBLEU toolkit (Post, 2018), are shown in Table 1. The NMT models outperform the phrase-based model (Moses base), and using additional resources, either in the form of multilingual data (Marian multilingual) or in pretraining (last three rows) helps in both directions. Pretraining produces the best models, with little difference between the three strategies tested.

In scenarios with little parallel data, but with some available monolingual data, it is common to use backtranslation to produce additional synthetic data. This brought a significant boost to our English-Gujarati models previously described. We therefore experiment with iterative backtranslation using our best models each time. This involves training models on the synthetically translated parallel data added to the genuine parallel and then fine-tuning on the genuine parallel data alone. The initial results (with one iteration of backtranslation) are presented in Table 2 for some of the models. This suggests that the scores can be greatly boosted by using backtranslation, especially when pretraining of the model is used as well.

[6] These correspond to relatively small vocabularies, reflecting the small amount of data these models are trained on, compared to the other models mentioned, which use monolingual and synthetic parallel data. The EN→TA direction especially benefits from a high degree of segmentation (i.e. a small vocabulary), as it is morphologically rich and the lack of data more heavily impacts how often different inflections of words are seen in the training data; a higher segmentation rate means that the model can better generalise.

[7] XLM involves training a transformer encoder to predict masked out words in raw sentences i.e. using a masked language modelling objective similar to BERT (Devlin et al., 2018). The difference is that as well as the model being trained on sentences from each language separately (from monolingual data), it is also trained on parallel examples (the concatenation of the sentence in both languages), from which several words can be masked out. The idea is that the information required to predict the masked out word could then be inferred from the translation of the sentence, thus teaching the model to recognise translation-like relations.

Model name	Dev BLEU EN↔TA	Dev BLEU TA↔EN
Moses base	3.0	6.2
Marian base	5.1	10.1
Marian multilingual	7.1	10.6
Marian mBART pretraining	7.4	14.0
XLM pretraining	7.4	13.4
Marian DE–EN pretraining	7.5	14.0

Table 1: BLEU scores on the development set of each model for EN–TA translation.

Model name	Backtranslation source	Dev BLEU EN↔TA	Dev BLEU TA↔EN
Marian base	XLM pretraining	9.9	16.5
Marian DE–EN pretraining	XLM pretraining	10.1	18.3
Marian mBART pretraining	Marian mBART pretraining	11	18.6

Table 2: BLEU scores on the development set of a selection of models are training on synthetic (back-translated) and parallel data for EN–TA translation. *Backtranslation source* refers to the model that was used to produce the backtranslations that are used as additional training data.

We plan to provide a set of systematic experiments exploring the usefulness of backtranslations produced by the different models described above, used to train different types of MT models. We hope that this will provide valuable insights into how to best exploit different models and what sort of backtranslated data is useful. We will specifically answer the following research questions:

- Is the quality of the model used to produce backtranslations (Table 1) really the defining factor when it comes to the usefulness of backtranslations?
- Can it be useful to use a mixture of backtranslations from different models despite differences in quality? And can this diversity in quality actually be exploited?
- What would the ideal combination of different models be to produce an optimal model after several rounds of backtranslation?

3.2 Language Model Prior for Low-Resource Neural Machine Translation

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015b; Vaswani et al., 2017b) relies heavily on large parallel corpora (Koehn and Knowles, 2017) and needs careful hyper-parameter tuning, in order to work in low-resource settings (Sennrich and Zhang, 2019). A popular approach for addressing data scarcity is to exploit abundant monolingual corpora via data augmentation techniques, such as back-translation (Sennrich et al., 2016a). Although back-translation usually leads to significant performance gains (Hoang et al., 2018), it requires training separate models and expensive translation of large amounts of monolingual data. However, when faced with lack of training data, a more principled approach is to consider exploiting prior information.

Language models (LM) trained on target-side monolingual data have been used for years as priors in statistical machine translation (SMT) (Brown et al., 1993) via the noisy channel model Shannon and Weaver (1949). Unlike maximum likelihood training that directly models $p(\boldsymbol{y}|\boldsymbol{x})$, noisy channel models the "reverse translation probability" $p(\boldsymbol{x}|\boldsymbol{y})$, by rewriting $p(\boldsymbol{y}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{y}) \times p(\boldsymbol{y})$. This approach has been adopted to NMT, with the neural noisy channel (Yu et al., 2017; Yee et al., 2019). However, neural noisy channel models face a computational challenge, because they require multiple passes over the source sentence x as they generate the target sentence y, or sophisticated architectures to reduce the passes.

LMs have also been used in NMT for re-weighting the predictions of translation models (TM), or as additional context, via LM-fusion (Gulcehre et al., 2015; Sriram et al., 2018; Stahlberg et al., 2018). But, as the LM is required during decoding, it adds a significant computational overhead. Another challenge is balancing the TM and the LM, whose ratio is either fixed (Stahlberg et al., 2018) or requires changing the model architecture (Gulcehre et al., 2015; Sriram et al., 2018).

In Baziotis et al. (2020), we propose to use a LM trained on target-side monolingual corpora as a weakly informative prior. We add a regularization term, which drives the output distributions of the TM to be probable under the distributions of the LM. Specifically,

$$\mathcal{L} = \sum_{t=1}^{N} -\log p_{\text{TM}}(\boldsymbol{y}_t|\boldsymbol{y}_{<t}, \boldsymbol{x}) + \lambda\, D_{\text{KL}}(p_{\text{LM}}(\boldsymbol{y}_t|\boldsymbol{y}_{<t}) \parallel p_{\text{TM}}(\boldsymbol{y}_t|\boldsymbol{y}_{<t}, \boldsymbol{x}))$$

The first term is the standard translation objective $\mathcal{L}_{\text{MT}}$ and the second is the regularization term $\mathcal{L}_{\text{KL}}$, which we interpret as a weakly informative prior over the TM's distributions p_{TM}, that expresses partial information about $\boldsymbol{y}$. $\mathcal{L}_{\text{KL}}$ is defined as the Kullback-Leibler divergence between the output distributions of the TM and the LM, weighted by λ.

This gives flexibility to the TM, by enabling it to deviate from the LM when needed, unlike fusion methods that change the decoder's distributions, which can introduce translation errors. The LM "teaches" the TM about the target language similarly to knowledge distillation (Bucila et al., 2006; Hinton et al., 2015). This method works by simply changing the training objective and does not require any changes to the model architecture. Importantly, the LM is separated from the TM, which means that it is needed only during training, therefore we can decode faster than fusion or neural noisy channel. We also note that this method is not intended as a replacement to other techniques that use monolingual data, such as back-translation, but is orthogonal to them.

Main Findings In our experiments we use two low-resource language pairs, the English-German (EN-DE) News Commentary v13 provided by WMT Bojar et al. (2018) [8] (~275K pairs) and the English-Turkish (EN-TR) WMT-2018 parallel data from the SETIMES2[9] corpus (~200K pairs). As monolingual data for English and German we use the News Crawls 2016 articles (Bojar et al., 2016) and for Turkish we concatenate *all* the available News Crawls data from 2010-2018, which contain 3M sentences. For English and German we subsample 3M sentences to match the Turkish data, as well as 30M to measure the effect of stronger LMs. In all experiments, we use the Transformer architecture for both the LMs and TMs.

First, we use in all methods LMs trained on the *same* amount of monolingual data, which is 3M sentences. We used the total amount of available Turkish monolingual data (3M) as the lowest

[8] http://www.statmt.org/wmt18/translation-task.html
[9] http://opus.nlpl.eu/SETIMES2.php

common denominator. This is done to remove the effects of the size of monolingual data from the final performance of each method, across language-pairs and translation directions. Overall, adding the LM-prior consistently improves performance in all experiments. Specifically, it yields up to +1.8 BLEU score gains over the strongest baseline. This shows that the proposed approach yields clear improvements, even with limited monolingual data (3M).

Next, we lift the monolingual data constraint, in order to evaluate the impact of stronger LM-priors. Specifically, for English and German we use LMs trained on 30M sentences. We observe that the stronger LMs yield improvements only in the EN→DE direction. This could partially be explained by the fact that German has richer morphology than English. Therefore, it is harder for the decoder to avoid grammatical mistakes in low-resource settings while translating into German, and a stronger prior is more helpful for X→DE than X→EN.

We also conducted experiments that measure the effect of the LM-prior on different scales of parallel data. Specifically, we emulate more low-resource conditions, by training on subsets of the EN→DE parallel data. We observe that adding the LM-prior yields consistent improvements, even with as little as 10K parallel sentences.

Finally, we perform an analysis on the effects that different methods have on the output distributions of the TM. Specifically, we evaluate each model on the DE→EN test-set and for each target token we compute the entropy of each model's distribution. We find that the gains from the LM-prior cannot be explained just from smoothing the distributions of the TM, like what label smoothing (Szegedy et al., 2016) regularization does. This suggests that the propose technique indeed works by exploiting information from the LM. Next, we focus on LM-fusion techniques and show that, even though they might improve fluency, they can hurt translation quality in certain cases because the model cannot resolve large "disagreements" between the TM and LM, which leads to wrong predictions.

3.3 Cross-lingual word-embeddings induction

Cross-lingual word embeddings are word embedding vectors which are trained to be consistent between multiple languages: the embedding of a word in one language should be close to the embedding of its translation in a different language, possibly up to a know transformation. Cross-lingual word embeddings are useful to initialize multi-lingual models, such as unsupervised machine translation systems (Lample et al., 2018; Artetxe et al., 2018), which makes them attractive for the low-resource scenarios of the GoURMET project.

Cross-lingual word embeddings can be learned from parallel dictionaries or parallel text, but these resources might be hard to acquire for some of the very low-resource language pairs that we consider in this project. A related approach exploits the shared vocabulary that naturally occur in monolingual text: certain tokens such as numbers, dates, names of people, companies, products, etc. tend to have the same orthography and the same meaning between different languages, even from different language families. In some cases, this minimal shared vocabulary can provide enough signal to align embedding spaces between different languages, enabling cross-lingual word embedding induction (Artetxe et al., 2017). Nevertheless, this approach is not without limitations: languages written in different scripts tend to share less vocabulary and even after transliteration names might not match exactly; names might be also subject to declension in morphologically-rich languages such as the Turkic languages that we consider in Gourmet; moreover, number and date format might vary between languages: many languages such as English use West Arabic nu-

merals which descend from but are typographically different from the Indian numerals still in use in many Indic languages which we target in Gourmet. For these reasons, the naturally occurring shared vocabulary might not be always sufficient to provide a strong enough training signal. Therefore, we considered fully-unsupervised cross-lingual word embedding induction: in this setting we attempt to infer a correspondence between words in different languages, and hence their embeddings, based purely on their statistical properties in monolingual texts, based on an approximate distribution isomorphism assumption between languages (Miceli Barone, 2016).

A number of different methods to cross-lingual word-embedding induction have been proposed in the literature (Ruder et al., 2019). Broadly speaking, they can be divided in two main approaches: joint training of word embeddings in multiple languages and cross-lingual alignment between separately pretrained embeddings. In our work we investigated both approaches.

3.3.1 Cross-lingual embedding alignment with normalizing flows

Normalizing flows (Rezende and Mohamed, 2015; Papamakarios et al., 2019) are frameworks for estimating the density of continuous probability distributions by learning an invertible transformation between distributions. Zhou et al. (2019) applied normalizing flows to the problem of cross-lingual embedding alignment by defining a probability distribution over the embedding space of each language as an equal mixture of Gaussians with means corresponding to the pretrained monolingual embeddings and small spherical variance, then they learn linear flows to transfrom one of the distribution to the other one. Their approach, however, also uses the shared vocabulary between languages as an additional learning signal, hence it is not fully unsupervised, and fails completely when this signal is removed.

In this project we attempted to train an expressive non-linear flow based on the split-coupling architecture of Glow (Kingma and Dhariwal, 2018). Glow defines a parametric transformation implemented by a neural network which is trained to learn a mapping between two distributions. The invertibility of the neural network is provided by its architectural constraints, specifically being made of a sequence of linear layers with square matrices (invertible for almost all parameter values) and Feistel network-like parametric non-linear blocks.

Let

$$p_s(x) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}(x, w_s^{(i)}, \sigma^2 I)$$

$$p_t(y) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}(y, w_t^{(i)}, \sigma^2 I)$$

be the word embedding distributions for the s and t languages, each defined on top of N pretrained word embeddings $\left\{w_s^{(i)}\right\}_{i=1}^{N}$ and $\left\{w_t^{(i)}\right\}_{i=1}^{N}$. We seek to find an invertible parametrized transformation from $f_{s\to t}(x, \theta)$ that optimizes the normalizing flow objective

$$\text{argmax}_\theta \mathbb{E}_{x\sim p_s}\left[\log p_t(f_{s\to t}(x,\theta)) + \log\left|\det \frac{d}{dx} f_{s\to t}(x,\theta)\right|\right]$$

We discovered early on that without the shared vocabulary signal, our the model also struggles to train. We identified the problem with the shape of the mixture of Gaussians distribution that we

used as a prior: this distribution has as many modes as the number of word embeddings, resulting in an extremely non-convex loss function that is impossible to optimize by gradient descent[10]. We attempted to address this issue by augmenting the embedding spaces by concatenating them with zero-mean spherical Gaussian noise and by replacing the exact mixture of Gaussian prior with a learned neural network estimator trained by noise-contrastive estimation (Gutmann and Hyvärinen, 2010) or conditional noise-contrastive estimation (Ceylan and Gutmann, 2018). Unfortunately we were not able to train a satisfactory transformation between embedding spaces.

We also experimented with a different formulation of flows. Instead of trying to directly map the embedding distributions of different languages into each other, we learn two separate flows, one for each language, that map word embeddings into a shared latent space on which we impose a simple zero-mean spherical Gaussian distribution.

The optimization objective is

$$\begin{aligned}\text{argmax}_\theta \mathbb{E}_{x\sim p_s}\left[\log p_h(f_{s\to h}(x,\theta)) + \log\left|\det \frac{d}{dx} f_{s\to h}(x,\theta)\right|\right] + \\ \mathbb{E}_{y\sim p_t}\left[\log p_h(f_{t\to h}(y,\theta)) + \log\left|\det \frac{d}{dy} f_{t\to h}(y,\theta)\right|\right] \\ \text{where } p_h(z) = \mathcal{N}(z, 0, \sigma^2 I)\end{aligned}$$

Once the model is trained, a direct transformation between embedding spaces can be obtained by exploiting the fact that split-coupling flows enable closed-form efficient computation of inverse transformations; therefore

$$f_{s\to t}(x,\theta) = f^{-1}_{t\to h}(f_{s\to h}(x,\theta),\theta)$$

The main advantage of this approach is that the latent distribution is unimodal, enabling effective optimization as long as the flow transformations have sufficient expressive power. Unfortunately expressive transformations tend to align the embedding spaces in arbitrary ways that don't necessarily map words in one language to their translations in the other language. We tried to bias the model towards learning semantically consistent transformations by aggressively sharing parameters between the flows: they share all the layers except the first weight matrix which we keep language-specific:

$$\begin{aligned}f_{s\to h}(x,\theta) &= g(W_s \cdot x, \theta)\\ f_{t\to h}(y,\theta) &= g(W_t \cdot y, \theta)\end{aligned}$$

which results in

$$f_{s\to t}(x,\theta) = W_t^{-1} \cdot W_s \cdot x$$

We further experimented with constraining W_s and W_t to be orthogonal.

We found that this method can learn some degree of mapping between English-Spanish word embeddings, but generally underperforms the baselines, hence we did not pursue it any further.

[10] Neural networks can be often trained despite their optimization objective being generally non-convex as long as they are sufficiently overparametrized (Allen-Zhu et al., 2018); however, the loss function applied to the outputs must me convex in order for this to work.

3.3.2 Attention-based joint crosslingual embedding training

We investigated a method to jointly train cross-lingual word embeddings from pairs of monolingual corpora. Unlike the previous method where we attempted to align pretrained embeddings, here we seek to use a single training procedure to obtain monolingual embeddings for two different languages with the property that embeddings of words that are translations should be similar. Unlike most alignment methods, we don't seek to generate strict one-to-one alignments, since in general word translation is not strictly one-to-one, especially when one language is morphologically rich and the other one is not.

We extend the Skipgram approach (Mikolov et al., 2013). In the Skipgram approach a dictionary of the N most frequent word type is constructed associating each word type i to two vectors: a center embedding $w^{(i)}$, which will be returned as the final word embedding, and a context embedding $c^{(i)}$. The training procedure uses a contrastive objective that pushes the center embedding of words occurring in the training text close to the context embeddings of words occurring in a small window around it and far from randomly sampled words.

In our approach we assume a dictionary of N latent "concept" center embeddings $\left\{l_o^{(j)}\right\}_{j=1}^{N}$ and N corresponding latent concept context embeddings $\left\{l_c^{(j)}\right\}_{j=1}^{N}$ that are shared between the two languages. The main idea is that the actual center embeddings for the words in each language will be convex combinations of the latent concept center embeddings

$$w_s^{(i)} = \sum_{j=1}^{N} a_s^{(i,j)} \cdot l_o^{(j)}$$
$$w_t^{(i)} = \sum_{j=1}^{N} a_t^{(i,j)} \cdot l_o^{(j)}$$

and similarly for the context embeddings

$$c_s^{(i)} = \sum_{j=1}^{N} a_s^{(i,j)} \cdot l_c^{(j)}$$
$$c_t^{(i)} = \sum_{j=1}^{N} a_t^{(i,j)} \cdot l_c^{(j)}$$

The combination coefficient matrices a_s and a_t have non-negative elements with each row summing to one.

We parametrize the combination coefficient using an attention mechanism: we define N latent key vectors $\left\{l_k^{(j)}\right\}_{j=1}^{N}$ and for each language we define N query vectors $\left\{q_s^{(i)}\right\}_{i=1}^{N}$ and $\left\{q_t^{(i)}\right\}_{i=1}^{N}$, one for each word type in the vocabulary. Then

$$a_s^{(i,j)} = \text{softmax}_j(< q_s^{(i)}, l_k^{(j)} >)$$
$$a_t^{(i,j)} = \text{softmax}_j(< q_t^{(i)}, l_k^{(j)} >)$$

where

$$\text{softmax}_j(x_j) = \frac{\exp(x^{(j)})}{\sum_{j'} \exp(x^{(j')})}$$

and $< \cdot, \cdot >$ denotes the dot product between vectors.

All these parameters are trained using the skipgram method using batches of monolingual text in both languages.

As it is, this does not lead to representations that are consistent between each languages because the attention coefficients tend to be dense. Therefore we included sparsity regularization to promote each word to align to one or a few "concepts". We experimented with different sparsity losses such as those described by Soulos et al. (2019).

We evaluated our approach on the English-Spanish language pair training on Wikipedia dump monolingual corpora. We found that achieving both high sparsity and convergence in the skipgram objective is hard even when including a learning schedule in the loss weights. There appears to be a sharp "phase transition" between a dense attention regime where the skipgram objective improves and a sparse attention regime where the models fall into an inescapable local minimum. We also experimented with different parametrizations (e.g. including a MLP in the attention computation) and different normalization schemes (e.g. normalizing the embeddings to unit L2 norm, normalizing the attention weights to unit L2 norm instead of the usual unit L1 norm), but ultimately we were not able to achieve competitive results.

3.4 Probabilistic back-translation

One of the most successful techniques for integrating monolingual data into NMT is back-translation (Sennrich et al., 2016b), where a target monolingual corpus is translated using a pre-trained target-to-source NMT system creating a synthetic parallel corpus that can be used as additional training data. Whereas back-translation works very well in practice, it seems there is room for improvement in the way that the back-translations are obtained. The back-translation system is trained on its own separate objective, and not optimized towards the final goal (good translations in the forward direction). Furthermore, the back-translation system consists of an entire distribution over back-translations, yet we usually only use a single translation as additional data, limiting the capability to transfer knowledge. We formulate a probabilistic model of sentence pairs in which we address those issues. Here, back-translation will appear as an approximate inference procedure.

Let x denote a source sequence and y a target sequence. We formulate a joint model of sequence pairs as $p(x, y) = p(x)p(y|x)$, where we additionally include a (fixed) source language model on top of traditional NMT. Given both a parallel corpus $\mathcal{B}$ and a target monolingual corpus $\mathcal{M}$, maximum likelihood estimation tells us to optimize parameters θ by maximizing:

$$\mathcal{L}(\theta; \mathcal{B}, \mathcal{M}) = \sum_{i=1}^{|\mathcal{B}|} \log p(x^{(i)}, y^{(i)}|\theta) + \sum_{j=1}^{|\mathcal{M}|} \sum_{x' \in \mathcal{X}} \log p(x', y^{(j)}|\theta) \quad (1)$$

where $\mathcal{X}$ denotes the set of all possible source sequences. The marginalisation over source sentences here is clearly intractable. We can use variational inference (Jordan et al., 1999; Blei et al., 2017) to do approximate inference instead. This will introduce an approximate posterior distribution $q(x|y; \phi)$ that we will model as a sequence of categorical draws from the source vocabulary

parameterised in context in order to approximate the true posterior $p(x|y)$. $q(x|y)$ is an NMT system from target-to-source, i.e. a back-translation system. This yields the following lower-bound to Equation 1:

$$\mathcal{L}(\theta; \mathcal{B}, \mathcal{M}) \geq \sum_{i=1}^{|\mathcal{B}|} \log p(x^{(i)}, y^{(i)}|\theta) + \sum_{j=1}^{|\mathcal{M}|} \mathbb{E}_{q(x|y)}[\log p(y|x)] - \mathrm{KL}\left(q(x|y)||p(x)\right) \tag{2}$$

where KL denotes the Kullback–Leibler divergence.

Using Monte Carlo estimates we can now compute an estimate of the lower-bound. In order to obtain gradient estimates for ϕ we use the REINFORCE estimator (Williams, 1992) and standard variance reduction techniques (Ross, 2006). The procedure for target monolingual data using our formulation is as follows: *i)* sample a back-translation from $q(x|y)$; *ii)* update θ by treating the sample as a completion of the data and optimizing conditional likelihood; *iii)* update ϕ using REINFORCE gradients, maximizing forward likelihood on expectation and minimizing a KL divergence between the back-translation system and the source language model.

Our probabilistic formulation of back-translation differs from traditional back-translation in three ways: *i)* it samples back-translations rather than using beam search outputs, allowing for transfer of the entire distribution of the back-translation system rather than only its mode; *ii)* the back-translation system is trained on the same objective as the forward translation system is trained; *iii)* we are able to include a pre-trained source language model to encourage fluent back-translations.

We trained a probabilistic back-translation system on English-German Multi30k (Elliott et al., 2016) and Gourmet English-Turkish data. We note that it is necessary to initialize the back-translation system $q(x|y)$ using a traditional back-translation system as the REINFORCE gradients are too noisy otherwise. For comparison we also train an identical system trained on parallel data alone as well as a system including target monolingual data using traditional back-translation.

Our findings have shown that our probabilistic formulation is able to make use of monolingual data, improving over a baseline not including monolingual data. However, our probabilistic formulation is not able to match traditional back-translation in performance. We have boiled this down to two factors that we need to tackle independently.

The first problem is that of posterior collapse, a problem often occurring when using variational inference with strong decoders like the ones used in NMT (Bowman et al., 2016; Alemi et al., 2018). Specifically, the KL component in the lower-bound tends to push the back-translation system to generate fluent sentences that have semantically nothing in common with the target monolingual sentence. We have been able to combat this using commonly used techniques to remedy this such as KL annealing (Bowman et al., 2016) and KL free-bits (Kingma et al., 2016). This has improved result slightly, but not up to the point of traditional back-translation.

The second issue is that sampled back-translations work poorly on low-resource languages (Edunov et al., 2018) such as the ones used in the Gourmet project. We found that this problem is harder to tackle, as unbiased sampling is a core part of our method. Using techniques to obtain more likely samples did not give good results, likely due to biasing of the gradients. Therefore, we conclude that this is a problem that needs to be solved first independently before we are able to continue with any probabilistic formulation of back-translation in low-resource settings. This has motivated an investigation that takes a closer look at what probability distributions learned by NMT systems look like (Eikema and Aziz, 2020), also described in work package 3.

3.5 Co-Training

NMT systems require vast amounts of labelled data, i.e. bilingual sentence pairs, to be trained effectively. To meet this requirement, we typically combine different sources of data in one large (or as large as possible) mix. For example, we combine data from different domains (e.g., religion, politics, news, subtitling, manuals), synthetic data produced via back-translation (Sennrich et al., 2016b), copied data (Currey et al., 2017), and even data involving related languages (Johnson et al., 2017a). Translation direction, original language, and quality of translation are some of the many factors that we typically cannot or choose not to control for (due to lack of information or simply for convenience). To better deal with mixed-distribution datasets, we propose (Eikema and Aziz, 2018) to model sentence pairs under a shared latent space in the framework of variational auto-encoders (Kingma and Welling, 2014). We show improvements across various testing conditions, in particular, when learning from synthetic data in addition to parallel data.

In probabilistic back-translation (Section 3.4), we see that a back-translation component arises naturally in a joint generative model trained semi-supervisedly via variational inference (Jordan et al., 1999). The back-translation component essentially corresponds to an NMT model operating in target-to-source direction. Unfortunately, difficulties with variance of unbiased gradient estimators, necessary for variational inference, capped the potential of the approach. Here we side-step the need for unbiased gradients by working in the framework of co-training (Blum and Mitchell, 1998). Essentially, we train two models, one in source-to-target and another in target-to-source direction, and let these models complete monolingual data to one another. In particular, we have these models be Auto-Encoding Variational Neural Machine Translation (AEVNMT) models (Eikema and Aziz, 2018).

Model. Define an AEVNMT model working in the source to target direction:

$$\text{ELBO}_1 = \mathbb{E}_{q(z|x,y)}\left[\log \frac{p(z)p(x|z,\theta_1)p(y|z,x,\theta_1)}{q(z|x,y)}\right] \tag{3}$$

where we make the parametric choice $q(z|x,y) \propto q(z|x,\lambda_1)q(z|y,\lambda_2)$. That is, $q(z|x,y)$ is proportional to a product of Gaussians each of which individually depends only on either the source or target sentence. In addition, define another AEVNMT model, this one operating in target to source direction:

$$\text{ELBO}_2 = \mathbb{E}_{q(z|x,y)}\left[\log \frac{p(z)P(y|z,\theta_2)p(x|z,y,\theta_2)}{q(z|x,y)}\right] \tag{4}$$

where we share $q(z|x,y) \propto q(z|x,\lambda_1)q(z|y,\lambda_2)$ with the source-to-target model. For test-time predictions we use either $q(z|x,\lambda_1)$ or $q(z|y,\lambda_1)$, depending on the translation direction. We follow Eikema and Aziz (2020) and use beam search conditioned on the mean of the variational posterior.

Parameter estimation. We train both models on the same bilingual data while promoting their inference models to agree by minimising the following loss

$$\mathcal{L}(\theta_1,\lambda_1,\theta_2,\lambda_2) = -\text{ELBO}_1(\theta_1,\lambda_1) - \text{ELBO}_2(\theta_2,\lambda_2) + \text{JS}(q(z|x,\lambda_1)\|q(z|y,\lambda_2))\,, \tag{5}$$

which aggregates both negative ELBOs and a Jensen-Shannon penalty to promote agreement between inference models.

Semi-supervised training. Consider the case where we train the generative model $p(z, x, y|\theta_1)$, which operates from x to y, on target monolingual data. For that, we need to give latent treatment to both z and x. In variational inference, we derive an ELBO with respect to a variational approximation $q(x, z|y) = q(x|y)q(z|x, y)$:

$$\mathbb{E}_{q(x|y)}\left[\mathbb{E}_{q(z|x,y)}\left[\log \frac{p(z)p(x|z,\theta_1)p(y|z,x,\theta_1)}{q(z|y)q(x|z,y)}\right]\right] \tag{6}$$

At this point, we make a very convenient parametric choice:

$$q(x|y) \triangleq p(x|y,\theta_2) \tag{7a}$$

where we employ the second model to perform inferences about missing source sentences. The inference model for latent z remains the same, namely, $q(z|x, y) \propto q(z|x, \lambda_1)q(z|y, \lambda_2)$, but note that in this case x is synthetic (generated). Note that to sample x from the AEVNMT model operating in backward direction, we sample from $q(z|y, \lambda_2)p(x|z, y, \theta_2)$. Importantly, we have not violated any of the requirements of variational inference, we have simply defined different inference networks for different data points depending on their nature (bilingual or monolingual). Because we borrow components of a model to define the inference component of another, we like to think of this as a form of *co-training* (Blum and Mitchell, 1998). We further embrace the analogy of co-training and keep the inference model $q(x|y)$ fixed in a target-monolingual update, which circumvents the need for high-variance gradient estimation with respect to the parameters of that component.

Summary of findings. We experimented with English-Turkish data using a subset of the corpora used to build our systems in deliverable D5.3, namely, the SETIMES (Tyers and Alperen, 2010) parallel corpus ($\sim 200,000$ sentence pairs) and a subset of its monolingual part ($\sim 500,000$ sentences in each language). Overall, we observed gains from 0.5 to 1.0 BLEU compared to conditional NMT with back-translation. Though modest, the gains were consistent across independent training runs and were more pronounced as we reduced the amount of bilingual supervision. We noticed that the output of co-trained AEVNMT models showed unigram statistics that were closer to that of the training data, and that its latent code was predictive of the source of the data (namely, parallel or synthetic). A comparison based on all of the available data for that language pair showed 0.2 and 0.4 BLEU improvement into Turkish and into English (see D5.3, section 2.3.3). The modest improvements at this point might not seem to justify a more complex formulation, but in fact, matching back-translation with a latent variable model is an important step for the project given the importance of this class of models across work-packages. A complete report and experiments involving additional GoURMET language pairs is in preparation.

4 Task 3: Learning from Lexical Resources

Bilingual dictionaries exist for most pairs of written languages. These dictionaries are curated by linguists, hence they are usually of higher quality compared to parallel data automatically or semi-automatically extracted from large collections of documents. Their general coverage may be limited for low-resource languages, but it is often possible to create small domain-specific dictionaries for specific terminology, which can highly benefit the quality of translation if properly leveraged.

The most obvious technique to exploit bilingual dictionaries is to simply add them to the parallel data (which we mentioned as a baseline technique in the GoURMET project proposal), in addition to auxiliary word prediction objectives for neural machine translation systems such as Nguyen and Chiang (2017). We explored the data augmentation technique and additional techniques based on multi-source translation (post-editing) based on a bilingual dictionary and additional morphological transfer rules. We plan to investigate additional training objectives based on bilingual dictionaries in the second part of the project.

4.1 Multi-source approach for exploiting linguistic resources

Even though the languages of interest to the GoURMET project are under-resourced, for some of them there are linguistic resources available, such as morphological analysers, bilingual dictionaries and translation rules. In particular the rule-based MT platform Apertium (Forcada et al., 2011) provides linguistic resources to some of the languages of interest to the project (see deliverable *D1.1 Survey of relevant low-resource languages*).

In order to determine the potential of the linguistic resources available in Apertium we have conducted experiments with the Breton–French language pair in Apertium (Tyers, 2010). This language pair was selected for the following reasons:

- Breton is an under-resourced language. Bilingual corpora available amount to about 400,000 sentence pairs, of which about 80% is software localization text.
- The quality of the French it generates is not suitable for publishing; although it may be used to get a rough idea of the meaning of a Breton text.
- Breton and French show some interesting grammatical contrasts. Both have morphological richness but it is distributed differently. The translation between them poses a challenge similar to that of some of the languages of interest to GoURMET:
  - While French has a rich verb morphology, Breton's is more reduced. Breton resorts more often than French to periphrastic (compound) structures.
  - French prepositions are mostly written separately from the following word (except for a few contractions such as *du*, *des* or *aux*, and the use of apostrophes before a vowel); in contrast, Breton prepositions join tightly to the next word when it is a pronoun, making them similar to a verb conjugation: (*ganin* 'with me', *ganit* 'with you', *ganto* 'with them').
  - French nouns and adjectives inflect at the end (number, gender). Breton nouns and adjectives may additionally change at the beginning (in a process called initial consonant mutation), which adds some additional sparseness to the vocabulary: *tad* 'father' but *da dad* 'your father'; *ar vro vihan* 'the small country' (lit. 'the country small') but *ar broioù bihan* 'the small countries' (lit. 'the countries small').[11]
  - The neutral word order in a Breton sentence is verb–subject–object (as in most Celtic languages) while in French it is subject–verb–object.

[11] Examples taken from https://en.wikipedia.org/wiki/Breton_mutations

The idea was to combine knowledge extracted from the parallel corpus and knowledge explicitly coded in the Apertium linguistic resources. For that, we used multi-source NMT and formalise the problem of combining both sources of knowledge as an automatic post-editing (Chatterjee et al., 2018) problem. In this way, we were able to explore different ways of generating the Apertium output, using different resources, to study which resources are more useful for the hybrid approach.

The experiments conducted compared the performance of several baseline systems and the performance of two multi-source NMT architectures —a recurrent attentional encoder–decoder (Bahdanau et al., 2015a) and a Transformer encoder–decoder (Vaswani et al., 2017a)— when using different resources. The baseline systems used were:

- A baseline NMT system trained solely on the training corpus.
- A baseline NMT system trained on a concatenation of the training corpus and the entries in the Breton–French bilingual dictionary of Apertium.
- Apertium with hand-crafted rules: the full Apertium system. The linguistic resources used by this system are: morphological analyser for Breton, morphological generator for French, part-of-speech tagger of Breton, Breton–French bilingual dictionary of lemmas and shallow structural transfer rules.
- Apertium with automatically-inferred rules. Same as above but using the shallow structural transfer rules automatically inferred by Sánchez-Cartagena et al. (2015), instead of using hand-crafted rules.
- Apertium with no structural transfer rules; i.e. without applying any structural transfer rule to make the output more grammatical.

As regards the different ways of exploiting the linguistic resources in Apertium, we generated the additional input translation provided to the multi-source NMT system with the same Apertium configurations used as reference systems as well as a word-for-word translation obtained using exactly the same bilingual dictionary we used for the baseline NMT system trained on bilingual entries.

The results allow us to conclude that the use of Apertium resources improves translation quality. The best improvement —1.34 BLEU points and 2.11 chrF2++ points— was obtained when the additional input to the multi-source NMT system was generated without structural transfer rules. However, the performance of the Apertium baseline without structural transfer rules was worse than that of Apertium using hand-crafted rules and automatically inferred rules. This results suggest that Apertium may help the NMT system to perform a better lexical selection, since the improvement in the grammaticality of the Apertium output provided by the shallow-transfer rules had no effect on the quality of the final translation. In any case, the use of a morphological analyser and part-of-speech tagger for Breton had a positive effect on the translation quality of the multi-source NMT system. The addition of the bilingual dictionary to the training corpus seems to have no effect on translation quality. Finally, the best results were obtained with a recurrent attentional encoder–decoder; the Transformer seems to perform worse when the amount of training corpora is scarce.

A detailed description of the approach and the experimental results obtained can be found in the paper by Sánchez-Cartagena et al. (2020). The paper also includes the results of an automatic error analysis that reveals that the multi-source NMT system using no transfer rules make fewer

lexical errors, which account for most of the errors produced by the systems, but more reordering and inflection errors. It seems that the multi-source system is able to make a better use of the translations from the bilingual dictionary when they are sequentially placed in the additional input rather than when they have been processed by transfer rules.

Future work. We plan to apply the multi-source approach described here to the development of the translation models for Macedonian (one of the language selected for the third round of languages), for which there are resources in Apertium. We will also study other architectures for the inclusion of linguistic lexical resources in NMT.

4.2 Exploiting bilingual lexicons in Neural Machine Translation

Looking more specifically at the extra knowledge that can be gained from bilingual lexicons (parallel lists of words or phrases), this ongoing work seeks to compare different approaches of exploiting bilingual lexicons (parallel lists of words or phrases) to improve NMT for low resource languages. The main advantage of using these resources is to provide extra vocabulary coverage for previously unseen or rarely seen words in the training data.

In comparison to work exploiting monolingual data to increase vocabulary coverage, where parallel data is scarce (Sennrich et al., 2016b; Edunov et al., 2018; Artetxe et al., 2018; Lample et al., 2018; Lample and Conneau, 2019), integrating information from bilingual lexicons has the potential to provide a light-weight solution to integrating words and phrases that would otherwise be unseen in the parallel training data. The reason for this is that the linguistic information present within them is already structured, whereas models have to learn to induce parallel correspondences when using monolingual data. The use of these resources could therefore also be complementary to these monolingual-data-heavy approaches.

Different approaches have already been proposed to exploit bilingual lexicons in NMT, of which the following are representative of the different methods tried:

- Arthur et al. (2016) use a probabilistic phrase table as a form of bilingual lexicon, from which they externally calculate lexicon probabilities. These lexicon probabilities are then combined with the NMT probabilities for each time step using a fixed gating parameter.
- Feng et al. (2017) also exploit phrase tables, but integrate how they are queried into the NMT model itself, by using an additional attention mechanism to query the possible target terms that correspond to the matching source terms in the source sentence.
- Constrained decoding (Hasler et al., 2018; Post and Vilar, 2018) has been proposed to force outputs to contain phrases that are matched in a lexicon. This method is particularly adapted for the use of terminologies in unambiguous and controlled language settings.
- An alternative approach was proposed by Dinu et al. (2019), which consists in training an NMT model on data in which the source sentences are annotated inline for possible translations selected from a lexicon (specifying the origin of each word using source factors).

The last two approaches in particular gave performance gains to MT quality because of their ability to heavily bias if not force the model to produce certain words in the translation. However, the methods have not been compared in equal settings; the last two approaches were evaluated in gold

settings, whereby you assume that you know in advance which source terms should definitely be translated by which target terms.

The aim of this ongoing work is therefore to compare the different approaches in a generic translation setting (rather than a controlled language one), to see whether these approaches can give gains for low resource languages using the latest NMT architectures.

The comparison cannot be based purely on automatic metric scores such as BLEU (Papineni et al., 2002), as they offer little insight into whether the modification of terms leads to good alternative translations, especially if they differ from the human reference translations. We will therefore study more in depth how to evaluate lexicon-guided MT, asking what it means for (i) a translation to be good, and (ii) what a good exploitation of the lexicon means. What we observe is a juggling act between exploiting the lexicon as much as possible, without resorting to overly literal and word-by-word translation.

We also propose a new approach to exploiting bilingual lexicons, combining the various advantages of the approaches cited. The details of this model are currently being refined, and experimental results will be available in the coming months. The method, inspired by the use of the lexicon as an external memory is defined by the following characteristics:

- Effective use of the fact that the lexicon is structured as source and target items that are aligned. The association between source and target terms provided in the lexicon acts as useful supervision (i.e. it should not be entirely left to the NMT model to decide which target words to look at most).
- The NMT model is able to query the lexicon flexibly depending on how confident the model is about its own decision and whether or not it requires the additional support from the lexicon
- Subwords units is a built-in feature of the model rather than a detail that must be handled heuristically. The use of segmentation into subword units (Sennrich et al., 2016c) has become standard in NMT and is essential for generalising vocabulary coverage particularly in low resource scenarios. Therefore the representation of the lexicon is compatible with such a representation, and with the fact that both source and target sentences are represented by subword units rather than whole words.
- Finally, ambiguity should be allowed in terms of which lexicon items are possible targets, and the model is able to decide between these variants based on the current context.

5 Publications

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

- Arturo Oncevay, Barry Haddow, and Alexandra Birch. Towards a multi-view language representation:a shared space of discrete and continuous language features. In *the First Workshop on Typology for Polyglot NLP*, Florence, Italy, 2019 **Best Paper Award**
- Arturo Oncevay, Barry Haddow, and Alexandra Birch. Bridging linguistic typology and multilingual machine translation with multi-view language representations. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2004.14923

- Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. The University of Edinburgh's Submissions to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy, 2019
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. Association for Computational Linguistics (to appear), 2020
- Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2004.14928
- Bryan Eikema and Wilker Aziz. Auto-encoding variational neural machine translation. *arXiv preprint arXiv:1807.10564*, 2018
- Bryan Eikema and Wilker Aziz. Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation. *arXiv:2005.10283 [cs]*, May 2020. URL http://arxiv.org/abs/2005.10283. arXiv: 2005.10283
- Víctor M. Sánchez-Cartagena, Mikel L. Forcada, and Felipe Sánchez-Martínez. A multi-source approach for Breton–French hybrid machine translation. In *Proceedings of the 22th Annual Conference of the European Association for Machine Translation*, pages 61–70, Online Conference, November 2020

6 Software

Another research output from this workpackage is software.

- WMT19 Gujarati system models and scripts http://data.statmt.org/wmt19_systems/
- Tool for fusing, extending and using language representations github.com/aoncevay/multiview-langrep
- Code for the improving massively multilingual NMT work https://github.com/bzhangGo/zero
- Code for the language model prior work github.com/cbaziotis/lm-prior-for-nmt
- Code for the auto-encoding variational NMT work github.com/Roxot/AEVNMT

7 Conclusion

In this document we have reported the progress and future plans for Workpackage 4 Transfer Learning of the GoURMET project.

We have performed research on the three workpackage tasks identified in the project proposal, achieving substantial advances in the state of the art in the field of transfer learning for low-resource

machine translation. Our research was reported in several papers published at top-level academic conferences, achieving a best paper award and a top position in a scientific evaluation for news translation system quality.

However, our work is not merely of academic interest. We have applied our techniques to produce machine translation systems for multiple low-resource language pairs that we have delivered to the project user partners (BBC and Deutsche Welle). These systems, described in deliverable D5.3, are currently deployed and are being integrated in the workflow of the user partners.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://www.aclweb.org/anthology/N19-1388.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR. URL http://proceedings.mlr.press/v80/alemi18a.html.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019, 2019. URL http://arxiv.org/abs/1907.05019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised Neural Machine Translation. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR'18, Vancouver, Canada, 2018.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, 2016. doi: 10.18653/v1/D16-1162. URL https://www.aclweb.org/anthology/D16-1162.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015a.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, CA, USA, 2015b. URL http://arxiv.org/abs/1409.0473.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W19-5301.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. The University of Edinburgh's Submissions to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy, 2019.

Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2004.14928.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1285773.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. Citeseer, 1998.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the Conference on Machine Translation*, pages 131–198, Berlin, Germany, August 2016. doi: 10.18653/v1/W16-2301. URL https://www.aclweb.org/anthology/W16-2301.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the conference on machine translation (WMT). In *Proceedings of the Conference on Machine Translation*, pages 272–303, Belgium, Brussels, October 2018. doi: 10.18653/v1/W18-6401. URL https://www.aclweb.org/anthology/W18-6401.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL https://www.aclweb.org/anthology/K16-1002.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19 (2):263–311, 1993. URL https://www.aclweb.org/anthology/J93-2003.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541, Philadelphia, PA, USA, 2006. URL https://doi.org/10.1145/1150402.1150464.

Ciwan Ceylan and Michael U Gutmann. Conditional noise-contrastive estimation of unnormalised models. *arXiv preprint arXiv:1806.03664*, 2018.

R. Chatterjee, M. Negri, R. Rubino, and M. Turchi. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 723–738, Belgium, Brussels, October 2018.

Bernard Comrie. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press, 1989.

Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc., 2019.

Anna Currey, Antonio Valerio, and Kenneth Heafield. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation*, WMT'17, pages 148–156, Copenhagen, Denmark, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, 2019. doi: 10.18653/v1/P19-1294. URL https://www.aclweb.org/anthology/P19-1294.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP'18, pages 489–500, Brussels, Belgium, 2018.

Bryan Eikema and Wilker Aziz. Auto-encoding variational neural machine translation. *arXiv preprint arXiv:1807.10564*, 2018.

Bryan Eikema and Wilker Aziz. Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation. *arXiv:2005.10283 [cs]*, May 2020. URL http://arxiv.org/abs/2005.10283. arXiv: 2005.10283.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-3210. URL http://www.aclweb.org/anthology/W16-3210. event-place: Berlin, Germany.

Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. Memory-augmented neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark, 2017. doi: 10.18653/v1/D17-1146. URL https://www.aclweb.org/anthology/D17-1146.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1101. URL https://www.aclweb.org/anthology/N16-1101.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas, November 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1026. URL https://www.aclweb.org/anthology/D16-1026.

M.L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J.A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F.M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144, 2011.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1121. URL https://www.aclweb.org/anthology/P19-1121.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015. URL http://arxiv.org/abs/1503.03535.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, USA, December 8-9 2016.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana, 2018. doi: 10.18653/v1/N18-2081. URL https://www.aclweb.org/anthology/N18-2081.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. doi: 10.18653/v1/W18-2703. URL https://www.aclweb.org/anthology/W18-2703.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351, 2017a. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1081.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017b. doi: 10.1162/tacl_a_00065. URL https://www.aclweb.org/anthology/Q17-1024.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, November 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR, 2014*, Banff, Canada, 2014.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved Variational Inference with Inverse Autoregressive Flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4743–4751. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6581-improved-variational-inference-with-inverse-autoregressive-flow.pdf.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007. URL https://www.aclweb.org/anthology/P07-2045.

Taku Kudo and John Richardson. 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, 2018.

Surafel M. Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. Multilingual Neural Machine Translation for Zero-Resource Languages. *arXiv e-prints*, art. arXiv:1909.07342, Sep 2019.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1054.

Guillaume Lample and Alexis Conneau. Cross-lingual Language Model Pretraining. In *arXiv:1901.07291*, 2019.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised Machine Translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR'18, Vancouver, Canada, 2018.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1301. URL https://www.aclweb.org/anthology/P19-1301.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2002.

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, 2020.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6309. URL https://www.aclweb.org/anthology/W18-6309.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1268. URL https://www.aclweb.org/anthology/D17-1268.

Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv preprint arXiv:1608.02996*, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Toan Q Nguyen and David Chiang. Improving lexical choice in neural machine translation. *arXiv preprint arXiv:1710.01329*, 2017.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1123.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. Towards a multi-view language representation:a shared space of discrete and continuous language features. In *the First Workshop on Typology for Polyglot NLP*, Florence, Italy, 2019.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. Bridging linguistic typology and multilingual machine translation with multi-view language representations. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2004.14923.

Robert Östling and Jörg Tiedemann. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2102.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3): 559–601, 2019. doi: 10.1162/coli_a_00357. URL https://doi.org/10.1162/coli_a_00357.

Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, 2018.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, 2018. doi: 10.18653/v1/N18-1119. URL https://www.aclweb.org/anthology/N18-1119.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1049. URL https://www.aclweb.org/anthology/P17-1049.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-learning-dynamics-and-interpretability.pdf.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Sheldon M. Ross. *Simulation, Fourth Edition*. Academic Press, Inc., Orlando, FL, USA, 2006. ISBN 0-12-598063-9.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

V. M. Sánchez-Cartagena, J. A. Pérez-Ortiz, and F. Sánchez-Martínez. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46 – 90, 2015. ISSN 0885-2308.

Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July 2019. doi: 10.18653/v1/P19-1021. URL https://www.aclweb.org/anthology/P19-1021.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany, August 2016a. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

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*, ACL'16, pages 86–96, Berlin, Germany, 2016b.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL'16, pages 1715–1725, Berlin, Germany, 2016c.

Claude E Shannon and Warren Weaver. The mathematical theory of communication. *Urbana*, 117, 1949.

Paul Soulos, Tom McCoy, Tal Linzen, and Paul Smolensky. Discovering the compositional structure of vector representations with role learning networks, 2019.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. Cold fusion: Training seq2seq models together with language models. In *Proceedings of Interspeech*, pages 387–391, 2018. doi: 10.21437/Interspeech.2018-1392. URL http://dx.doi.org/10.21437/Interspeech.2018-1392.

Felix Stahlberg, James Cross, and Veselin Stoyanov. Simple fusion: Return of the language model. In *Proceedings of the Conference on Machine Translation*, pages 204–211, Belgium, Brussels, October 2018. doi: 10.18653/v1/W18-6321. URL https://www.aclweb.org/anthology/W18-6321.pdf.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Víctor M. Sánchez-Cartagena, Mikel L. Forcada, and Felipe Sánchez-Martínez. A multi-source approach for Breton–French hybrid machine translation. In *Proceedings of the 22th Annual Conference of the European Association for Machine Translation*, pages 61–70, Online Conference, November 2020.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao QIN, and Tie-Yan Liu. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1089. URL https://www.aclweb.org/anthology/D19-1089.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

F.M. Tyers. Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, pages 174–181, Saint-Raphaël, France, May 2010.

Francis M Tyers and Murat Serdar Alperen. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, 2010.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017a.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA, 2017b.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1176. URL https://www.aclweb.org/anthology/P19-1176.

Ronald J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3-4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696. Place: Hingham, USA.

Kyra Yee, Yann Dauphin, and Michael Auli. Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 5700–5705, Hong Kong, China, November 2019. doi: 10.18653/v1/D19-1571. URL https://www.aclweb.org/anthology/D19-1571.

Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomás Kociský. The neural noisy channel. 2017. URL https://openreview.net/forum?id=SJ25-B5eg.

Biao Zhang, Ivan Titov, and Rico Sennrich. Improving deep transformer with depth-scaled initialization and merged attention. 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 898–909, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1083. URL https://www.aclweb.org/anthology/D19-1083.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. Association for Computational Linguistics (to appear), 2020.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. Density matching for bilingual word embedding. *arXiv preprint arXiv:1904.02343*, 2019.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D4.1 Initial progress report on transfer learning