A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation

Surafel M. Lakew
University of Trento
Fondazione Bruno Kessler
lakew@fbk.eu

Mauro Cettolo
Fondazione Bruno Kessler
cettolo@fbk.eu

Marcello Federico
MMT Srl, Trento
Fondazione Bruno Kessler
federico@fbk.eu

Abstract

Recently, neural machine translation (NMT) has been extended to multilinguality, that is to handle more than one translation direction with a single system. Multilingual NMT showed competitive performance against pure bilingual systems. Notably, in low-resource settings, it proved to work effectively and efficiently, thanks to shared representation space that is forced across languages and induces a sort of transfer-learning. Furthermore, multilingual NMT enables so-called zero-shot inference across language pairs never seen at training time. Despite the increasing interest in this framework, an in-depth analysis of what a multilingual NMT model is capable of and what it is not is still missing. Motivated by this, our work (i) provides a quantitative and comparative analysis of the translations produced by bilingual, multilingual and zero-shot systems; (ii) investigates the translation quality of two of the currently dominant neural architectures in MT, which are the Recurrent and the Transformer ones; and (iii) quantitatively explores how the closeness between languages influences the zero-shot translation. Our analysis leverages multiple professional post-edits of automatic translations by several different systems and focuses both on automatic standard metrics (BLEU and TER) and on widely used error categories, which are lexical, morphology, and word order errors.

1 Introduction

As witnessed by recent machine translation evaluation campaigns (IWSLT (Cettolo et al., 2017), WMT (Bojar et al., 2017)), in the past few years several model variants and training procedures have been proposed and tested in neural machine translation (NMT). NMT models were mostly employed in conventional single language-pair settings, where the training process exploits a parallel corpus from a source language to a target language, and the inference involves only those two languages in the same direction. However, there have also been attempts to incorporate multiple languages in the source (Luong et al., 2015a; Zoph and Knight, 2016; Lee et al., 2016), in the target (Dong et al., 2015), or in both sides like Firat et al. (2016) which combines a shared attention mechanism and multiple encoder-decoder layers. Regardless, the simple approach proposed in Johnson et al. (2016) and Ha et al. (2016) remains outstandingly effective: it relies on single "universal" encoder, decoder and attention modules, and manages multilinguality by introducing an artificial token at the beginning of the input sentence to specify the requested target language.

The current NMT state-of-the-art includes the use of recurrent neural networks, initially introduced in Sutskever et al. (2014; Cho et al. (2014), convolutional neural networks, proposed by Gehring et al. (2017), and so-called transformer neural networks, recently proposed by Vaswani et al. (2017). All of them implement an encoder-decoder architecture, suitable for sequence-to-sequence tasks like machine translation, and an attention mechanism (Bahdanau et al., 2014).

Besides specific studies focusing on new architectures and modules, like Luong et al. (2015b) that empirically evaluates different implementations of the attention mechanism, the comprehension of what a model can learn and the errors it makes has been drawing much attention of the research community, as

Proceedings of the 27th International Conference on Computational Linguistics, pages 641–652
Santa Fe, New Mexico, USA, August 20-26, 2018.

evidenced by the number of recent publications aiming at comparing the behavior of neural vs. phrase-based systems (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Bentivogli et al., 2018). However, understanding the capability of multilingual NMT models in general and zero-shot translation, in particular, has not been thoroughly analyzed yet. By taking the bilingual model as the reference, this work quantitatively analyzes the translation outputs of multilingual and zero-shot models, aiming at answering the following research questions:

- How do bilingual, multilingual, and zero-shot systems compare in terms of general translation quality? Is there any translation aspect better modeled by each specific system?
- How do Recurrent and Transformer architectures compare in terms of general translation quality? Is there any translation aspect better modeled by each specific system?
- What is the impact of using related languages data in training a zero-shot translation system for a given language pair?

To address these questions, we exploit the data collected in the IWSLT 2017 MT evaluation campaign (Cettolo et al., 2017) and made publicly available by the organizers. The campaign was the first featuring a multilingual shared MT task, spanning five languages (English, Dutch, German, Italian, and Romanian) and all their twenty possible translation directions. In addition to the official external single reference of the test sets, we can also rely on professional post-edits of the outputs of nine Romanian→Italian and of nine Dutch→German participants' systems. Hence, we exploit the availability of multiple Italian and German references to perform a thorough analysis for identifying, comparing and understanding the errors made by different neural system/architectures we are interested in; in particular, we consider pairs of both related languages (Romanian→Italian, Dutch→German) and unrelated languages (Romanian→German and Dutch→Italian). Furthermore, to explore the impact of using data from other related languages, French and Spanish are considered for training purposes as well, in particular for analyzing the behavior of zero-shot x→Italian systems, x representing any source language distant from Italian.

In the following sections, we begin with a brief review of related work on quantitative analysis of MT tasks (§2). Then, we give an overview of NMT (§3) with a contrast between the Recurrent (§3.1) and Transformer (§3.2) approaches, and a summary on multilingual and zero-shot translation (§3.3). Section (§4), describes the dataset and preprocessing pipeline (§4.1), qualitative evaluation data (§4.2), experimental setting (§4.3), models (§4.4) and the evaluation methods (§4.5). In Section (§5), we analyze the overall translation quality for related and unrelated language directions. Before the summary and conclusion, we will focus on lexical, morphological and word-order error types for the fine-grained analysis (§6).

2 Related Work

Recent trends in NMT evaluation show that post-editing helps to identify and address the weakness of systems (Bentivogli et al., 2018). Furthermore, the use of multiple post-edits in addition to the manual reference is gaining more and more ground (Bentivogli et al., 2016; Koehn and Knowles, 2017; Toral and Sánchez-Cartagena, 2017; Bentivogli et al., 2018). For our investigation, we follow the error analysis approach defined in Bentivogli et al. (2018), where multiple post-edits are exploited in order to quantify morphological, lexical, and word order errors, a simplified error classification with respect to that proposed in Vilar et al. (2006), which settles two additional classes, namely missing and extra words.

The first work that compares bilingual, multilingual, and zero-shot systems comes from the IWSLT 2017 evaluation campaign (Cettolo et al., 2017). The authors analyze the outputs of several systems through two human evaluation methods: direct assessment which focuses on the generic assessment of overall translation quality, and post-editing which directly measures the utility of a given MT output to translators. Post-edits are also exploited to run a fine-grained analysis of errors made by the systems. The main findings are that (i) a single multilingual system is an effective alternative to a bunch of bilingual systems, and that (ii) zero-shot translation is a viable solution even in low-resource settings. Motivated by

those outcomes, in this work we explore in more detail the practical feasibility of multilingual and zero-shot approaches. In particular, we explore the benefit of adding training data involving related languages in a zero-shot setting and, in that framework, we compare the behavior of state-of-the-art Transformer and Recurrent NMT models.

3 Neural Machine Translation

A standard state-of-the-art NMT system comprises of an encoder, a decoder and an attention mechanism, which are all trained with maximum likelihood in an end-to-end fashion (Bahdanau et al., 2014). Although there are different variants of the encoder-attention-decoder based approach, this work focuses on the Recurrent LSTM-based variant (Sutskever et al., 2014) and the Transformer model (Vaswani et al., 2017).

In both the Recurrent and Transformer approaches, the encoder is purposed to cipher a source sentence into hidden state vectors, whereas the decoder uses the last representation of the encoder to predict symbols in the target language. In a broad sense, the attention mechanism improves the prediction process by deciding which portion of the source sentence to emphasize at a time (Luong et al., 2015b). In the following two subsections, we briefly summarize the two architecture types.

3.1 Recurrent NMT

In this case, the source words are first mapped to vectors with which the encoder recurrent network is fed. When the <eos> (i.e. end of sentence) symbol is seen, the final time step initializes the decoder recurrent network. At each time step of the decoding, the attention mechanism is applied over the encoder hidden states and combined with the current hidden state of the decoder to predict the next target word. Then, the prediction is fed back to the decoder (i.e. input feeding), to predict the next word, until the <eos> symbol is generated (Sutskever et al., 2014; Cho et al., 2014).

3.2 Transformer NMT

The Transformer architecture works by relying on a self-attention (*intra-attention*) mechanism, removing all the recurrent operations that are found in the previous approach.In other words, the attention mechanism is repurposed to compute the latent space representation of both the encoder and the decoder sides. However, with the absence of recurrence, *positional-encoding* is added to the input and output embeddings. Similarly, as the time-step in a recurrent network, the positional information provides the Transformer network with the order of input and output sequences. In our work, we use the absolute positional encoding, but very recently the use of the relative positional information has been shown to improve performance (Shaw et al., 2018). The model is organized as a stack of encoder-decoder networks that works in an auto-regressive way, using the previously generated symbol as input for the next prediction. Both the decoder and encoder can be composed of uniform layers, each built of two sub-layers, i.e., a multi-head self-attention layer and a position wise feed-forward network (FFN) layer. The multi-head sub-layer enables the use of multiple attention functions with a similar cost of utilizing attention, while the FFN sub-layer is a fully connected network used to process the attention sublayers; as such, FFN applies two linear transformations on each position and a ReLU (Vaswani et al., 2017).

3.3 Multilingual NMT

Recent efforts in multilingual NMT using a single encoder-decoder and an attention mechanism (Johnson et al., 2016; Ha et al., 2016) have shown to improve translation performance with minimal complexity. Multilingual NMT models can be trained with parallel corpora of several language pairs in *many-to-one, one-to-many*, or *many-to-many* translation directions. The main idea that distinguishes multilingual NMT training and inference from a single language pair NMT is that in preprocessing, a *language-flag* is appended to the source side of each segment pair. The flag specifies the target language the source is paired with at training time. Moreover, it enables a zero-shot inference by directing the translation to a target language never seen at training time paired with the source. In addition to reducing training and maintenance complexity of several single language pair systems, the two main advantages of multilingual

	encoder-decoder type	embedding size	hidden units	encoder depth	decoder depth	batch size
Recurrent	LSTM	512	1024	4	4	128 seg
Transformer	Self-Attention	512	512	6	6	2048 tok

Table 1: Hyper-parameters used to train Recurrent and Transformer models, unless differently specified.

NMT is the performance gain for low-resource languages, and the possibility to perform a zero-shot translation.

However, the translations generated by multilingual and zero-shot systems have not been investigated in detail yet. This includes analyzing how the model behaves solely relying on a "language-flag" as a way to redirect the inference. Recent works have shown that the target language-flag is weaker in a low-resource language setting (Lakew et al., 2017). Thus, in addition to analyzing the behavior of bilingual and multilingual models, mainly, the zero-shot task requires a careful investigation.

4 Data and Experiments

4.1 Datasets and preprocessing

The experimental setting comprises seven languages; for each language pair, we use the ≈200,000 parallel sentences made publicly available by the IWSLT 2017 evaluation campaign (Cettolo et al., 2017), partitioned in training, development, and test sets. In the preprocessing pipeline, the raw data is first tokenized and cleaned by removing empty lines. Then, a shared byte pair encoding (BPE) model (Sennrich et al., 2015) is trained using the union of the source and target sides of the training data. The number of BPE segmentation rules is set to $8,000$, following the suggestion of Denkowski and Neubig (2017) for experiments in small training data condition. For the case of Transformer training, the internal sub-word segmentation (Wu et al., 2016) provided by the Tensor2Tensor library[1] is used. Note that prepending the "language-flag" on the source side of the corpus is specific to the multilingual and zero-shot models.

4.2 Evaluation data

For our investigation, we exploit the nine post-edits available from the IWSLT 2017 evaluation campaign. Post-editing regarded the bilingual, multilingual, and zero-shot runs of three different participants to the two tasks Dutch (Nl)→German (De) and Romanian (Ro)→Italian (It). Human evaluation was performed on a subset (603 sentences) of the nine runs, involving professional translators. Details on data preparation and the post-editing task can be found in Cettolo et al. (2017).

The translation directions we consider in this work are Nl/Ro→De and Nl/Ro→It. The choice of *German* and *Italian* as the target languages is motivated by (i) the availability of multiple post-edits for the fine-grained analysis and (ii) the possibility of varying the linguistic distance between the source and the target languages, allowing experimental configurations suitable to answer the research questions raised in Section §1.

As said, for Nl→De and Ro→It the human evaluation sets consist of 603 segments. Since post-editing involved only those two language pairs, for the other two directions considered in this work, namely Nl→It and Ro→De, we tried to exploit at best the available post-edits by looking for all and only the segment pairs of the Nl→It and Ro→De tst2017 sets for which the target side exactly matches (at least) one of the segment pairs of the Ro→It and Nl→De human evaluation sets. This way, we were able to find 478 matches on the Italian sides and 444 on the German sides, which therefore become the sizes of the human evaluation sets of Ro→De and Nl→It, respectively, for which we can re-use the available post-edits.

It is worth to note that in general, the post-edits from the evaluation campaign are not actual post-edits of MT outputs generated in our experiments, with some exceptions discussed later, therefore they should rather be considered as multiple external references.

[1] https://github.com/tensorflow/tensor2tensor

Model	#Directions	System description
NMT	1	*Four pure bilingual models for the Nl→De/It and Ro→De/It directions*
M-NMT	20	*Multilingual, trained on all directions in the set {En,De,Nl,It,Ro}*
ZST	16	*Zero-shot, trained as multilingual but removing also Nl↔De and It↔Ro data*
ZST_A	12	*Zero-shot, trained as ZST but removing also De↔Ro and Nl↔It data*
ZST_B	16	*Zero-shot, trained as ZST_A but adding En↔Fr/Es data*

Table 2: The training setting of 4*bilingual, 1*multilingual, and 3*zero-shot systems.

4.3 Training setting

Each of the three system types, namely bilingual, multilingual and zero-shot, is trained using both Recurrent and Transformer architectures, with the proper training data provided in the IWSLT 2017 evaluation campaign. Meta training parameters were set in a preliminary stage with the aim of maximizing the quality of each approach. Recurrent NMT experiments are carried out using the open source OpenNMT-py[2] (Klein et al., 2017), whereas the Transformer models are trained using the Tensor2Tensor toolkit. Hence, we took the precaution of selecting the optimal training and inference parameters for both approaches and toolkits. For instance, for our low-resource setting characterized by a high data sparsity, the dropout (Srivastava et al., 2014) is set to 0.3 (Gal and Ghahramani, 2016) in Recurrent models and to 0.2 in Transformer models to prevent over-fitting. Similarly, Adam (Kingma and Ba, 2014) optimizer with an initial learning rate of either 0.001 (RNN) or 0.2 (Transformer) is used. If the perplexity does not decrease on the validation set or the number of epochs is above 7, a learning rate decay of 0.7 is applied in the Recurrent case. For the Transformer case, the learning rate is increased linearly in the early stages (*warmup_training_steps*=16000); after that, it is decreased with an inverse square root of training step (Vaswani et al., 2017). Table 1 summarizes the list of hyper-parameters.

4.4 Models

To address the research questions listed in Section §1, we train five types of models using either the Recurrent or the Transformer approaches. All models are trained up to convergence, eventually the best performing checkpoint on the dev set is selected. Table 2 summarizes the systems tested in our experiments. As references, we consider four bilingual systems (in short NMT) trained on the following directions: Nl→De/It and Ro→De/It. The first term of comparison is a many-to-many multilingual system (in short M-NMT) trained in all directions in the set {En,De,Nl,It,Ro}. Then, we test zero-shot translation (ZST) between related languages, namely Nl→De and Ro→It, by training a multilingual NMT without any data for these language pairs. We also test zero-shot translation between unrelated languages (ZST_A), namely Ro→De and Nl→It, by excluding parallel data between these languages. Finally, for the same unrelated zero-shot directions we also train multi-lingual systems (ZST_B) that include data related to Romanian and Italian, namely En↔Fr/Es.

4.5 Evaluation methods

Systems are compared in terms of BLEU (Papineni et al., 2002) (as implemented in *multi-bleu.perl* [3]) and TER (Snover et al., 2006) scores, on the single references of the official IWSLT test sets.
In addition, two TER-based scores are reported, namely the multiple-reference TER (mTER) and a lemma-based TER (lmmTER), which are instead computed on the nine post-edits of the IWSLT 2017 human evaluation set. In mTER, TER is computed by counting, for each segment of the MT output, the minimum number of edits across all the references and dividing by the average length of references. lmmTER is computed similarly to mTER but looking for matches at the lemma level instead of surface forms. Significance tests for all scores are reported using Multeval (Clark et al., 2011) tool.
Systems are also compared in terms of three well known and widely used error categories, that is lexical, morphological, and word order errors, exploiting TER and post-edits as follows. First, the MT outputs

[2]https://github.com/OpenNMT/OpenNMT-py
[3]A script from the Moses SMT toolkithttp://www.statmt.org/moses

Direction	System	Recurrent				Transformer			
		BLEU	TER	mTER	lmmTER	BLEU	TER	mTER	lmmTER
Nl→De	NMT	18.05	64.61	23.70	20.60	**18.37**	**63.74**	27.95	23.86
	M-NMT	17.79	66.18	21.75	18.28	↑**19.95**	**61.90**	23.62	20.05
	ZST	17.06	65.73	26.35	22.29	↑**19.13**	**62.69**	**25.19**	**21.53**
Ro→It	NMT	22.16	59.35	22.99	20.39	**22.48**	**57.34**	26.60	23.36
	M-NMT	21.69	59.50	21.12	18.46	↑**22.12**	**57.51**	25.05	21.57
	ZST	18.72	62.08	29.66	26.15	↑**21.29**	**59.08**	**26.93**	**23.33**

Table 3: Automatic scores on tasks involving related languages. BLEU and TER are computed on test2017, while mTER and lmmTER are reported for human evaluation sets. Best scores of the Transformer model against the Recurrent are highlighted in bold, whereas arrow ↑ indicates statistically significant differences ($p < 0.05$).

and the corresponding post-edits are lemmatized and POS-tagged; for that, we used ParZu (Sennrich et al., 2013) for German and TreeTagger (Schmid, 1994) for Italian. Then, the lemmatized outputs are evaluated against the corresponding post-edits via a variant of the *tercom* implementation[4] of TER: in addition to computing TER, the tool provides complete information about matching lemmas, as well as shift (matches after displacements), insertion, deletion, and substitution operations. Since for each lemma the tool keeps track of the corresponding original word form and POS tag, we are able to measure the number of errors falling in the three error categories, following the scheme described in detail in Bentivogli et al. (2018).

5 Translation Analysis

5.1 Related languages

First, we compare the bilingual (NMT), multilingual (M-NMT), and zero-shot (ZST) systems on the two tasks *Nl→De* and *Ro→It*, implemented as either Recurrent or Transformer networks, in terms of automatic metrics. As stated above, BLEU and TER exploit the official external reference of the whole test sets, while mTER and lmmTER are utilize the multiple post-edits of the (smaller) IWSLT human evaluation test set. Scores are given in Table 3.

Looking at the BLEU/TER scores, it is evident that Transformer performs better in all the three model variants. In particular, for the multilingual and the zero-shot models, the gain is statistically significant. On the contrary, the mTER and lmmTER scores are better for the Recurrent architecture; in this case, the outcome is misleading since the nine post-edits include those generated by correcting the outputs of the three Recurrent systems. As such, the translations of the Recurrent systems are rewarded over the translations produced by the Transformer systems, thus making the comparison not fair.

As far as the models are compared, the bilingual one is the best in three out of four cases, the exception being the Transformer/Nl→De. Nonetheless, it is worth to note the good performance of the multilingual model in terms of mTER and lmmTER. This result holds true in both Recurrent and Transformer approaches, regardless of the BLEU score. We hypothesize that the main reason behind this is the higher number of linguistic phenomena observed in training, thanks to the use of data from multiple languages, which makes the multilingual models more *robust* than the bilingual models.

5.2 Unrelated languages

In unrelated language directions, our experimental setting is aimed at evaluating the impact of source *language-relatedness* with the target. Particularly, we focus on the zero-shot setup given its intrinsic difficulty, by taking the bilingual systems as references. Table 4 provides BLEU and TER based scores for the Ro→De and Nl→It directions.

Concerning the ZST_A training condition, in one case (Recurrent Ro→De) it outstandingly allows to outperform the pure bilingual system, while in the other cases there is no significant difference between

[4]Available at `wit3.fbk.eu/show.php?release=2016-02&page=subjeval`

Direction	System	Recurrent				Transformer			
		BLEU	TER	mTER	lmmTER	BLEU	TER	mTER	lmmTER
Ro→De	NMT	13.99	72.70	61.82	54.61	$^{\uparrow}$**16.52**	**66.71**	**55.68**	**48.44**
	ZST_A	14.93	69.38	58.26	51.08	$^{\uparrow}$**16.46**	**66.88**	**54.72**	**48.25**
	ZST_B	14.75	69.29	58.26	51.37	$^{\uparrow}$**16.55**	**67.18**	**55.29**	**48.03**
Nl→It	NMT	18.88	63.79	58.79	52.16	$^{\uparrow}$**20.22**	**60.88**	**55.52**	**48.56**
	ZST_A	18.77	62.97	58.80	51.32	$^{\uparrow}$**19.80**	**60.24**	**54.06**	**47.16**
	ZST_B	18.87	62.40	57.34	50.17	$^{\uparrow}$**20.61**	**59.41**	**53.04**	**46.17**

Table 4: Evaluation results for the unrelated language directions. BLEU and TER scores are computed with single references, while mTER and lmmTER are computed with nine post-edits. Best scores of the Transformer over the corresponding Recurrent architectures are highlighted in bold, whereas arrow $^{\uparrow}$ indicates statistically significant differences ($p < 0.05$).

ZST_A and NMT, proving once again that zero-shot translation built on the "language-flag" of M-NMT is really effective (Johnson et al., 2016): in fact, at most a slight performance degradation is recorded as the number of pairs used in training decreases (Lakew et al., 2017). Although gains are rather limited, adding training data involving Romance target languages (French and Spanish, ZST_B) close to Italian impacts as hoped: ZST_B scores are in general better than both NMT and ZST_A in Nl→It, while they do not degrade with respect to ZST_A in Ro→De.
Similarly to what is observed for related pairs (Table 3), the Transformer architecture shows definitely higher quality than the RNN one, confirming the capability of the approach to infer unseen directions.
The overall outcomes from Tables 3 and 4 are: (i) multilingual systems have the potential to effectively model the translation either in zero-shot or non zero-shot conditions; (ii) zero-shot translation is a viable option to enable translation without training samples; (iii) the Transformer is the best performing approach, particularly in the zero-shot directions.
The next section is devoted to a fine-grained analysis of errors made by the various systems at hand, with the aim of assessing the outcomes based on automatic metrics.

6 Fine-grained Analysis

Following the error classification defined in Section 4.5, now we focus on lexical, morphological, and reordering error distributions to characterize the behavior of the three types of models and the two sequence-to-sequence learning approaches considered in this work.
As anticipated in the previous section, it is expected that scores computed with reference to post-edits penalize Transformer over Recurrent systems because the outputs of the latter were post-edited, but not those of the former. We try to mitigate this bias by relying on the availability of multiple post-edits which likely allows to better match the Transformer runs than having a single reference would do. For the fine-grained analysis, we use instead the expedient of computing error distributions that are normalized with respect to the error counts observed in a bilingual reference system. In the next two sections, the fine-grained analysis is reported for related and unrelated languages pairs, consecutively.

6.1 Related languages

Table 5 provides the distribution over the error types by the bilingual (NMT), multi-lingual (M-NMT), and zero-shot (ZST) models, implemented either with Recurrent or Transformer architectures, for the Nl→De translation direction. We also report, for each error type and M-NMT and ZST system, the observed relative difference of errors with respect to the bilingual reference model (NMT).
Considering each error category, we observe the same general trend for all systems: the lexical errors represent by far the most frequent category (76-77%), followed by morphology (15-16%) and reordering (3-6%) errors; cases of words whose morphology and positioning are both wrong, represent about 1-2% of the total errors. Beyond the similar error distributions, it is worth to note the variation of errors made by M-NMT and ZST models with respect to those of the NMT model: for the Recurrent

Nl→De	Recurrent					Transformer				
	NMT	M-NMT	Δ_{NMT}	ZST	Δ_{NMT}	NMT	M-NMT	Δ_{NMT}	ZST	Δ_{NMT}
Lexical	77.29	69.65	-7.64	83.73	+6.44	76.47	64.83	-11.64	69.53	-6.94
Morph	15.41	16.51	+1.10	19.1	+3.69	15.70	13.96	-1.74	14.13	-1.57
Reordering	5.53	3.14	-2.39	5.41	-0.12	6.20	4.97	-1.23	5.41	-0.79
Morph. & Reo.	1.76	1.02	-0.74	1.61	-0.15	1.63	1.36	-0.27	1.53	-0.10
Total	100	90.31	-9.69	109.84	+9.84	100	85.12	**-14.88**	90.6	**-9.40**

Table 5: Distribution of lexical, morphological, and reordering error types from the two MT approaches. Reported values are normalized with respect to the total error count of the respective bilingual reference model (NMT). Δ_{NMT} are variations with respect to the bilingual reference models (NMT).

Ro→It	Recurrent					Transformer				
	NMT	M-NMT	Δ_{NMT}	ZST	Δ_{NMT}	NMT	M-NMT	Δ_{NMT}	ZST	Δ_{NMT}
Lexical	80.63	73.81	-6.82	102.79	+22.16	81.97	76.01	-5.96	84.12	+2.15
Morph	12.33	12.86	+0.53	16.00	+3.67	11.49	11.79	+0.30	12.44	+0.95
Reordering	5.74	3.71	-2.03	6.09	+0.35	5.35	4.64	-0.71	4.81	-0.54
Morph. & Reo.	1.30	1.15	-0.15	2.18	+0.88	1.19	1.09	-0.10	1.09	-0.10
Total	100	91.54	**-8.46**	127.07	+27.07	100	93.52	-6.48	102.45	**+2.45**

Table 6: Distribution of the error types in the Ro→It direction for the Recurrent and Transformer approaches. From the variation of errors that compare M-NMT and ZST models with the bilingual reference (NMT), a larger margin of error is observed in case of Transformer ZST model.

Ro→It	Recurrent					Transformer				
	NMT	ZST_A	Δ_{NMT}	ZST_B	Δ_{NMT}	NMT	ZST_A	Δ_{NMT}	ZST_B	Δ_{NMT}
Lexical	80.63	108.27	+27.64	100.31	+19.68	81.97	82.11	+0.14	76.76	-5.21
Morph	12.33	17.11	+4.78	17.23	+4.90	11.49	13.09	+1.60	11.59	+0.10
Reordering	5.74	6.20	+0.46	6.16	+0.42	5.35	5.18	-0.17	5.59	+0.24
Morph. & Reo.	1.30	2.22	+0.92	2.30	+1.00	1.19	1.16	-0.03	1.02	-0.17
Total	100	133.81	+33.81	126	+26.00	100	101.53	**+1.53**	94.96	**-5.04**

Table 7: Error distribution of ZST_A and ZST_B models for the Recurrent and Transformer variants. Transformer achieves the highest error reduction, particularly in the ZST_B model setting.

architecture, there is a decrease of 9.69 and an increase of 9.84 points, respectively. On the contrary, the Transformer architecture yields improvements for both models: total errors reduce by 14.88 and 9.40 points, respectively. The result for the Transformer ZST system is particularly valuable since the average error reduction comes from remarkable improvements across all error categories.

For the Ro→It direction, results are given in Table 6. Although to a different extent, we observe a picture similar to that of Nl→De discussed above: lexical errors is the type of error committed to a greater extent, multilingual models outperform their bilingual correspondents (more for the Recurrent than for the Transformer models), and ZST is competitive with bilingual NMT only if the Transformer architecture is adopted.

Training under the zero-shot conditions ZST_A and ZST_B assume less training data available and permit to measure the impact of introducing additional parallel data from related languages. We considered training conditions ZST_A and ZST_B here to perform Ro→It zero shot translation and report the outcomes in Table 7.

Results show error counts for each condition normalized with respect to the corresponding bilingual reference models (NMT). The most interesting aspect comes from the fact that global variations in the normalized error counts of the zero-shot translation can be here associated with the relatedness and variety of languages in the training data. As recently reported (Lakew et al., 2017), zero-shot performance

Ro→De	Recurrent					Transformer				
	NMT	ZST_A	Δ_{NMT}	ZST_B	Δ_{NMT}	NMT	ZST_A	Δ_{NMT}	ZST_B	Δ_{NMT}
Lexical	79.18	74.42	-4.76	74.09	-5.09	79.21	79.11	-0.10	78.52	-0.69
Morph	9.91	10.35	+0.44	10.07	0.16	9.92	10.05	+0.13	10.87	+0.95
Reordering	7.33	6.16	-1.17	6.16	-1.17	7.19	6.88	-0.31	7.22	+0.03
Morph. & Reo.	3.58	3.47	-0.11	3.47	-0.11	3.68	3.52	-0.16	3.60	-0.08
Total	100	94.4	**-5.60**	93.79	**-6.21**	100	99.55	-0.45	100.21	+0.21

Table 8: Error distribution of the bilingual (NMT), ZST_A and ZST_B model runs for the unrelated Ro→De direction. The Transformer moder shows the smallest sensitivity to the change in the number of training language pairs.

Nl→It	Recurrent					Transformer				
	NMT	ZST_A	Δ_{NMT}	ZST_B	Δ_{NMT}	NMT	ZST_A	Δ_{NMT}	ZST_B	Δ_{NMT}
Lexical	81.08	80.7	-0.38	78.79	-2.29	81.15	79.36	-1.79	77.48	-3.67
Morph	8.47	9.03	+0.56	8.38	-0.09	9.01	9.2	+0.19	9.03	+0.02
Reordering	7.78	6.63	-1.15	6.32	-1.46	7.51	6.74	-0.77	6.51	-1.00
Morph & Reo	2.67	2.54	-0.13	2.38	-0.29	2.34	2.41	+0.07	2.45	+0.11
Total	100	98.89	-1.11	95.86	-4.14	100	97.71	**-2.29**	95.47	**-4.53**

Table 9: Error distribution of the bilingual (NMT), ZST_A and ZST_B model runs. Δ_{NMT} shows the relative change in the error distribution of the zero-shot models with respect to the bilingual reference models.

of Recurrent models in a low resource setting seems highly associated with the number of languages provided in the training data. This is also confirmed by comparing performance of Recurrent models across the ZST (Table 6), ZST_A and ZST_B conditions. In particular, variations from the bilingual reference model, show significant degradation when some language directions are removed (from +27.07 to +33.81) and a significant improvement when two related languages are added (from +33.81 to +26.00). Remarkably, the Transformer zero-shot model seems less sensitive to the removal or addition of languages: actually a slight improvement is observed after removing Nl→It and De→Ro (ZST_A), i.e., from +2.45 to +1.53, followed by a large improvement when En→Fr/Es (ZST_B) are added, i.e. from +1.53 to -5.04. Notice that the latter results outperform the bilingual model. Overall, across all experiments, we see slight changes in the distribution of errors types. On the other hand, increases or drops of specific error types with respect to the bilingual reference model show sharper differences across the different conditions. For instance, the best performing Transformer model (ZST_B in Table 7) seems to gain over the reference bilingual systems only in terms of lexical errors (-5.21). The zero-shot Transformer model trained under the ZST condition (Table 6) although globally worse than the bilingual reference, seems instead slightly better than the reference concerning reordering error (-0.54), which account for 5.35% of the total number of errors.

6.2 Unrelated languages

In our second scenario, we evaluate the relative changes in the error distribution for the unrelated language directions (Ro→De and Nl→It). This section complements the translation results reported in Table 4, analyzing the runs from the ZST_A and ZST_B models in a different manner.

In the Ro→De unrelated direction (Table 8), the Recurrent model shows a reduction in the error rate of 5.60 points (ZST_A) and 6.21 points (ZST_B) with respect to the bilingual (NMT) reference model, while for the Transformer no significant differences are observed. These results confirm what observed in the automatic evaluation on the reference translations (Table 4). The gain observed by the Recurrent model on the ZST_B condition is mainly on lexical (-4.76 points) and reordering errors (-1.17 points) is probably due to the poor performance of its bilingual counterpart.

As far as the the Nl→It unrelated direction (Table 9) is concerned, both Recurrent and Transformer

ZST models show to reduce the error counts over the bilingual reference model. Actually, a similar trend occurs in Ro→De (Table 8), but with a relatively higher error reduction in case of the Transformer model. In particular, the Transformer model shows reductions of -2.29 points for ZST_A and -4.53 for ZST_B, whereas for the Recurrent model the improvements are slightly lower, namely -1.11 (ZST_A) and -4.14 (ZST_B) points. Remarkably, both the Recurrent and Transformer models benefit from additional training data related to Italian (compare ZST_A and ZST_B).

In conclusion, we observe that error counts of the zero-shot models in unrelated directions can increase (Table 8) when compared to the bilingual model. However, in the related language direction the most interesting aspect is observed with the discount of error in the Nl→It direction (Table 9). In particular, the ZST_B zero-shot model showed $>2.0\%$ error reduction over the ZST_A model. This gain is directly related to the newly introduced training data (i.e., English↔French/Spanish) in case of ZST_B.

Summary and Conclusions

In this work, we showed how bilingual, multilingual, and zero-shot models perform in terms of overall translation quality, as well as the errors types produced by each system. Our analysis compared Recurrent models with the recently introduced Transformer architecture. Furthermore, we explored the impact of grouping related languages for a zero-shot translation task. In order to make the overall evaluation more sound, BLEU and TER scores were complemented with mTER and lmmTER, leveraging multiple professional post-edits. Our investigation on the translation quality and the results of the fine-grained analysis shows that:

- Multilingual models consistently outperform bilingual models with respect to all considered error types, i.e., lexical, morphological, and reordering.
- The Transformer approach delivers the best performing multilingual models, with a larger gain over corresponding bilingual models than observed with RNNs.
- Multilingual models between related languages achieve the best performance scores and relative gains over corresponding bilingual models.
- When comparing zero-shot and bilingual models, relatedness of the source and target languages does not play a crucial role.
- The Transformer model delivers the best quality in all considered zero-shot condition and translation directions.

Our fine-grained analysis looking at three types of errors (lexical, reordering, morphology) show significant differences in the error distributions across the different translation directions, even when switching the source language with another source language of the same family. No particular differences in the error distributions were observed across neural MT architectures (Recurrent vs. Transformer), while some marked differences were observed when comparing bilingual, multilingual, and zero-shot systems. A more in-depth analysis of these differences will be carried out in future work.

Acknowledgements

This work has been partially supported by the EC-funded projects ModernMT (H2020 grant agreement no. 645487) and QT21 (H2020 grant agreement no. 645452). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. Neural versus phrase-based mt quality: An in-depth analysis on english–german and english–french. *Computer Speech & Language*, 49:52–70.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. *arXiv preprint arXiv:1706.09733*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL (1)*, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Surafel M Lakew, Mattia A Di Gangi, and Marcello Federico. 2017. Multilingual neural machine translation for low resource languages. In *CLiC-it 2017 4th Italian Conference on Computational linguistics*.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015a. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015b. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

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

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of Recent Advances in Natural Language Processing*, number September, pages 601–609.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, US-MA, August.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

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

Antonio Toral and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

David Vilar, Jia Xu, Luis Fernando dHaro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.