Harvesting Linguistically Related Languages for Low-Resource NMT

Angelos Nalmpantis, John Gkountouras, Vasiliki Vasileiou,
Konstantinos Papakostas, Irene Papadopoulou

University of Amsterdam, Netherlands

{ angelos.nalmpantis, john.gkountouras, vasiliki.vasileiou,
konstantinos.papakostas, irene.phone } @student.uva.nl

Abstract

Large Language Models have revolutionized the field of Natural Language Processing, with Transformers quickly becoming the prevalent choice for practitioners. Consequently, great advances have been made in Neural Machine Translation, with a notable increase in terms of BLEU score compared to the pre-Transformer era. Despite this success, the low-resource setting continues to pose significant challenges. Back-translation and knowledge distillation attempt to solve this issue, yet the model selection, to either generate the data or distill the knowledge, still remains an open question. In this paper, we investigate whether using multiple bi-lingual models can capture different but related linguistic information and, as a result, improve the models' performance.

1 Introduction

Progress in the field of Neural Machine Translation (NMT) has been rapid since the introduction of the Transformer architecture (Vaswani et al., 2017). High-resource language pairs, such as English to French, for which there are massive parallel corpora to train these million-parameter neural models, have seen significant increases in performance when moving from a recurrent neural network to a Transformer-based architecture.[1]

Despite the success of Transformers in many tasks, low-resource language pairs still pose a series of challenges (Koehn and Knowles, 2017), as there is a limited amount of data to train a model on. This is exacerbated by the fact that Transformers typically need more data to converge (Baevski et al., 2019) due to the absence of inductive biases that other architectures impose. As a result, there is a pressing need for efficient training regimes to maximize the returns on the data at our disposal.

[1]For reference, in English to French, LSTM networks achieve a BLEU score of 37.5 (Luong et al., 2015), while the current SotA model has a score of 46.4 (Liu et al., 2020).

Common approaches to this problem attempt to tackle this issue by integrating pre-trained Language Models in the decoding process (Gulcehre et al., 2015), employing self-learning to generate synthetic data in the source (Sennrich et al., 2016) or the target (Zhang and Zong, 2016) language, utilizing dual-learning (He et al., 2016) or unsupervised learning for non-parallel monolingual corpora (Lample et al., 2018), and using knowledge distillation techniques (Tan et al., 2018).

In this paper, we evaluate whether using multiple bi-lingual models can benefit existing approaches, such as *back-translation* (BT) (Sennrich et al., 2016) and *knowledge distillation* (KD) (Saleh et al., 2020), for low-resource NMT. We use T5 (Raffel et al., 2020) as our base model, and analyze the impact of using different language pairs for either BT or KD, by measuring the overlap in n-grams between the corpora. Finally, we investigate the failure modes of our models by examining the impact of the input length on the translation quality.

In summary, we find that utilizing the linguistic knowledge from multiple bi-lingual models does not offer substantial improvements, as we originally hypothesized. Back-translation was highly dependent on the models' performance, in order to avoid distributional shifts. As a result, the single bi-lingual model generated more in-distribution data and thus outperformed the multiple related models. However, in our knowledge distillation experiments, where the T5-base model was employed as a teacher, the difference between the two approaches was marginal.

2 Related Work

Machine translation (MT) has stimulated the interest of the research community for years. Traditional methods, deriving parameters from the statistical analysis of bilingual text corpora, paved the way for neural architectures, learning million to billion parameters in an end-to-end fashion.

Neural Machine Translation (NMT) NMT relies on the sequence-to-sequence encoder-decoder paradigm. The components of this architecture have gradually evolved, starting with the incorporation of data-efficient RNNs (Sutskever et al., 2014; Bahdanau et al., 2015). Subsequently, CNNs (Gehring et al., 2017) enabled parallelization over a sequence's elements. Ultimately, the Transformer architecture (Vaswani et al., 2017) relied exclusively on multi-head attention to convey spatial information. Followup work (So et al., 2019) has combined both convolutions and attention, in order to leverage the strengths of both elements.

Low-resource MT In language pairs with few parallel corpora available, it is not always possible to get multiple pairs of words occurring in different contexts, which poses challenges for the training process. Data augmentation can be used to efficiently increase the number of pairs (Fadaee and Monz, 2018) and consequently overcome the aforementioned problem. Zoph et al. (2016) first trained a high-resource language pair and used the resulting network to initialize and train a model on the low-resource language pair. Sequence-level distillation (Kim and Rush, 2016), which generalizes knowledge distillation to sequence generation, can be considered an instance of transfer learning. In this line of research, Saleh et al. (2020) proposes an adaptive approach to dynamically adjust the contribution of the teacher models during distillation.

3 Method

3.1 Data Augmentation

We draw inspiration from Sennrich et al. (2016), who address the problem of data scarcity in low-resource NMT through *back-translation*. In summary, after training an NMT model on the available parallel data, they use it to translate a monolingual corpus, either in the source or the target language, thus creating new synthetic examples. Even though this results in a noisy dataset, as the model is bound to generate imperfect translations, the method has been proven to substantially improve performance. Hence, we explore two approaches:

- Use a single NMT model, *directly* fine-tuned on the low-resource language pair, to generate new samples from a large monolingual corpus. We call this **direct synthesis** (DS).
- Use a mix of NMT models, pre-trained on *linguistically related* high-resource language pairs and fine-tuned to the low-resource pair, to generate multiple synthetic corpora. We hypothesize that such a procedure could improve performance, as each pre-trained model may be better at capturing different grammatical and lexical phenomena from the source language. We call this **related synthesis** (RS).

3.2 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) is the process of transferring the knowledge of a large model to a smaller one. Large models are typically *over-parameterized*, hinting that they are not fully utilizing their capacity (Michel et al., 2019).

We assume a setup where we initially train a student model S on a parallel corpus $\mathcal{D}_{\text{LR}}$ of a low-resource language pair. Additionally, we have access to a collection $\mathcal{T}$ of teacher models that are pre-trained on high-resource language pairs, with the same source (or target) language as the low-resource pair. We further fine-tune these teacher models on $\mathcal{D}_{\text{LR}}$ to perform knowledge distillation with samples in the low-resource language pair.

We opted for *response-based* KD (Gou et al., 2021), where the teachers generate a distribution of soft labels corresponding to the SentencePiece subwords (Kudo and Richardson, 2018). Given a training sample $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{LR}}$, similar to Saleh et al. (2020), we define the distillation loss ℓ_{KD} as the sum of the *Kullback–Leibler divergence* between each teacher and the student:

$$\sum_{\tau=1}^{|\mathbf{y}|} \sum_{T_i \in \mathcal{T}} \sum_{v \in V} q_{\boldsymbol{\theta}_{T_i}}(v|\mathbf{y}_{<\tau}, \mathbf{x}) \log p_{\boldsymbol{\theta}_S}(v|\mathbf{y}_{<\tau}, \mathbf{x})$$

where $\boldsymbol{\theta}_{T_i}$ and $\boldsymbol{\theta}_S$ are the parameters of the teacher and student models respectively, $p_{\boldsymbol{\theta}_S}$ is the conditional probability of the student model, $q_{\boldsymbol{\theta}_{T_i}}$ is the output distribution of the teacher model, and V is the vocabulary. We sum over the losses in the training set to obtain the distillation objective $\mathcal{L}_{\text{KD}}$.

Following the literature (Cui et al., 2017), our final training objective becomes a linear combination of the distillation loss $\mathcal{L}_{\text{KD}}$ and a supervision signal, i.e. the cross entropy loss $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L} \triangleq \lambda \mathcal{L}_{\text{CE}} + (1 - \lambda)\mathcal{L}_{\text{KD}}, \quad \lambda \in [0, 1]$$

4 Experiments

In this section, we describe our experimental setup regarding our overall training pipeline, the corpora and models used, as well as other modeling choices.

To aid in the reproducibility of our work, we have published our code on GitHub.[2]

Datasets CCMatrix (Schwenk et al., 2021) is a large collection of parallel corpora between 90 languages, extracted from 32 different snapshots of the CommonCrawl dataset. As there are dozens of millions of sentences for high-resource pairs, we *randomly* sample 1M of the English-Romanian and English-French sets to use with our computational resources, and 15K of the English-Italian set to simulate a scenario with low-resource data.

Model T5 (Raffel et al., 2020) is a Transformer-based encoder-decoder architecture that is pre-trained on a mixture of NLP tasks, such as summarization, question answering, and translation. For our experiments, we use the HuggingFace implementations of T5-small[3] and T5-base[4].

Language Selection The motivation for the selection of French and Romanian as our high-resource languages is twofold. Primarily, the public checkpoint of T5 used has been trained to translate between English, Romanian, French, and German passages. This means that its tokenizer is also fit to process sentences in these languages with a shared vocabulary. Moreover, both French and Romanian belong to the Romance language family, which Italian is also a member of. This also motivates our use of the latter as an artificially low-resource language, to overcome the limitation of training an NMT model with vastly different vocabulary terms from scratch. We *simulate* the low-resource setting, as has been done previously in the literature (Marton et al., 2009; Duong et al., 2015).

Training Setup For the low-resource language pair, we used 1.5K, 1.5K, and 12K samples for the validation, testing, and training set correspondingly. Similarly, for the high-resource pairs, we used 100K, 100K, and 800K samples. When fine-tuning on the low-resource pair, we set the number of epochs to 40, the learning rate to $2 \cdot 10^{-5}$, and the weight decay for `AdamW` (Loshchilov and Hutter, 2017) to 0.01. We used a batch size of 64 for T5-small and 32 for T5-base. For the knowledge distillation setup, we set λ to 0.5, the batch size to 32, the number of epochs to 10, and the learning rate to $2 \cdot 10^{-5}$ using the `Adam` (Kingma and Ba, 2014) optimizer with no weight decay.

[2] https://github.com/j0hngou/LRNMT
[3] https://huggingface.co/t5-small
[4] https://huggingface.co/t5-base

Back-translation data In all of our augmentation experiments, we generate $2\times$ the amount of the low-resource training samples, which we found to be a balance between excessive noise and insufficient data. Additionally, for the *direct synthesis* scenario, we explore both directions of back-translation. Namely, we fine-tune T5 either from English to Italian (denoted with $\rightarrow$), or from Italian to English (denoted with $\leftarrow$), to generate the synthetic data.

Distillation architectures We initialize the student with a T5-small model that is fine-tuned on the low-resource pair. For the teachers, we experimented with using either a *single* teacher, which is a T5-base model directly fine-tuned on the low-resource pair, or *dual* teachers, which are T5-base models that are pre-trained on each high-resource pair and then fine-tuned on the low-resource one.

5 Results and Analysis

In this section, we examine the similarity of the corpora used to train our models, as well as present our experimental results. Further evaluation, such as the effect of the input length on the model's performance, can be found in Appendix A.

5.1 Data Analysis

In order to observe the influence that linguistically related languages can have in back-translation and knowledge distillation, we analyze the similarity of the monolingual corpora in terms of n-gram overlap. High overlap between two datasets in different languages sets a strong foundation for knowledge transfer when fine-tuning an NMT model that is trained on one of the two to the other.

Our assumption holds for the original monolingual corpora, as there is a notable intersection between the Italian, Romanian, and French n-grams.

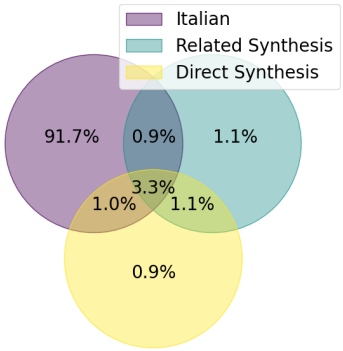


Figure 1: **Trigram** overlap between the Italian corpus and the synthetic datasets generated with RS and DS.

A more detailed study on the corpus overlap in the original languages can be found in Appendix C.

Furthermore, we investigate the similarity between the Italian corpus and the synthetic datasets generated through related synthesis (RS) and direct synthesis (DS). In Figure 1 we display the trigram overlap of the aforementioned datasets. We see that the RS dataset has a slightly smaller intersection with the original corpus compared to DS, which is an indication that the latter might contain more in-distribution samples and thus less noisy data.

5.2 Experimental Evaluation

Table 1 displays the BLEU (Papineni et al., 2002) scores[5] achieved on the test set of the English-to-Italian corpus. We observe that the model quickly adapts to the low-resource pair, possibly due to its similarity to T5's pre-training languages. We see significant improvements when using either back-translation or knowledge distillation, with their combination leading to the best result for T5-small.

We divide the results into three categories. First, we examine the behavior of the models when fine-tuning directly on the low-resource pair. Overall, the performance increase is more significant for T5-base, which suggests that larger models benefit more from few-shot learning (Brown et al., 2020). Surprisingly, pre-training on a high-resource pair doesn't improve, in the case of T5-base, or even harms performance, in the case of T5-small. This indicates that T5's original multi-lingual representations are a better baseline for translation in Italian.

Next, we consider the contribution of data augmentation. We observe that the quality of the model generating the synthetic data is of paramount importance. This is evident by looking at the degradation in performance when using RS in T5-small, compared to DS, as the model that was solely fine-tuned on Italian outperforms the ones pre-trained on the high-resource pairs. When it comes to direct synthesis, we see that generating data from a monolingual corpus in the *target* language leads to a greater performance increase as opposed to the reverse direction. This is in line with previous research (Fadaee and Monz, 2018) which hints that the encoder architecture is less affected by noise in the input, rather than having noisy labels for the supervision signal in the decoder architecture.

Finally, we analyze the distillation results, where we condense the knowledge from either a single or a dual teacher setup to a smaller-sized student. Interestingly, both approaches lead to comparable improvements in performance, which are consistently higher than simply fine-tuning on the low-resource pair. The application of knowledge distillation on an augmented dataset through back-translation results in the best performance in terms of BLEU score for the T5-small model. Although the difference between single and dual teacher distillation is not vast, we speculate that in cases where the teachers capture more orthogonal information compared to each other, this method may prove useful.

[5]We use the official SacreBLEU implementation.

	T5-small	**T5-base**
No fine-tuning	2.7	2.5
Fine-tuning on Italian		
Direct fine-tuning	9.3*	20.7*
+ pre-train on Fr.	8.7†	20.6†
+ pre-train on Ro.	8.8‡	20.7‡
Fine-tuning w/ data augmentation		
Related Synthesis ($\rightarrow$)	9.0†‡	21.4†‡
Direct Synthesis ($\rightarrow$)	9.5*	22.2*
Direct Synthesis ($\leftarrow$)	16.6*	<u>27.5</u>*
Knowledge Distillation*		
Single-teacher	15.1*	-
+ back-translation ($\leftarrow$)	18.5*	-
Dual-teacher	15.1†‡	-
+ back-translation ($\leftarrow$)	<u>18.6</u>*	-

Table 1: **BLEU** scores for English-to-Italian translation with different pre-training & fine-tuning regimes. Symbols $\{*, \dagger, \ddagger, \star\}$ indicate the models used in each case. Bold symbols $\{*, \dagger, \ddagger, \star\}$ indicate the T5-base variant.

6 Conclusion

In this study, we followed several approaches to increase the performance of an NMT model in the low-resource scenario. We noticed T5's ability to quickly adapt to data from a parallel corpus of English and Italian sentences, with larger models proving superior in the few-shot setting. Techniques such as back-translation and knowledge distillation led to a drastic improvement in BLEU score, with their combination resulting in a T5-small model with similar translation capabilities to T5-base. We experimented with data augmentation using a monolingual corpus on either the source or the target language, with the latter proving to be the most efficient. Although we hypothesized that using NMT models trained on linguistically similar languages could be beneficial, our experiments did not conclude in a significant performance increase.

Acknowledgments

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. We would also like to thank our TA Kata Naszádi for her continuous supervision and her suggestions for research directions during the development of this project.

References

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. 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 5360–5369, Hong Kong, China. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, Tom Sercu, Kartik Audhkhasi, Abhinav Sethy, Markus Nußbaum-Thom, and Andrew Rosenberg. 2017. Knowledge distillation across ensembles of multilingual models for low-resource languages. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4825–4829.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A neural network model for low-resource Universal Dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348, Lisbon, Portugal. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vision*, 129(6):1789–1819.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. 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 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

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

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 820–828, Red Hook, NY, USA. Curran Associates Inc.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

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

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

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Fahimeh Saleh, Wray Buntine, and Gholamreza Haffari. 2020. Collective wisdom: Improving low-resource neural machine translation using adaptive knowledge distillation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3413–3421, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5877–5886. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2018. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Effect of input length on performance

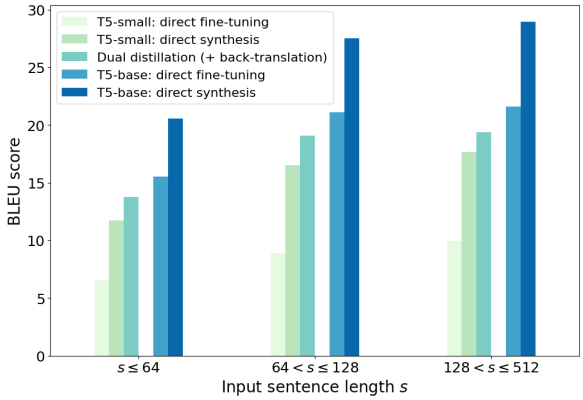


Figure 2: **BLEU** scores for discrete input sequence length buckets with different training regimes.

Figure 2 provides an insightful visualization of our results, where we experiment with different values for the (tokenized) input sequence length. We display the BLEU scores of the T5-small and T5-base models with the best performance on each segment of Table 1. To be more precise, we experimented with a sentence length of less than 64, between 65 and 128, and between 129 and 512 tokens. The selected values were determined from the distribution of our data. In all three cases, we observe the same pattern, which also comes to an agreement with the results reported in Table 1. Furthermore, by increasing the tokenized sequence length we achieve a better BLEU score, regardless of the training regime. This hints that the models may be utilizing information from longer sentences that leads to better translation performance.

B Model Parameters

Fine-tuning		
	low-resource	**high-resource**
# train set samples	12K	800K
# test set samples	1.5K	100K
# dev set samples	1.5K	100K
# of epochs	40	12
Batch Size	T5-small: 64 – T5-base: 32	
Optimizer	`AdamW` (weight decay: 0.01)	
Learning Rate	$2 \cdot 10^{-5}$	

Table 2: Hyper-parameters chosen for our fine-tuning experiments. Any option not explicitly mentioned uses the default value in the HF trainer options.

In Table 2 we report the size of our datasets and other model-specific metadata of our fine-tuning experiments. Note that we opted for 12K samples in our low-resource training set, which –in contemporary literature (Gu et al., 2018)– is considered an *extremely* low resource setting.

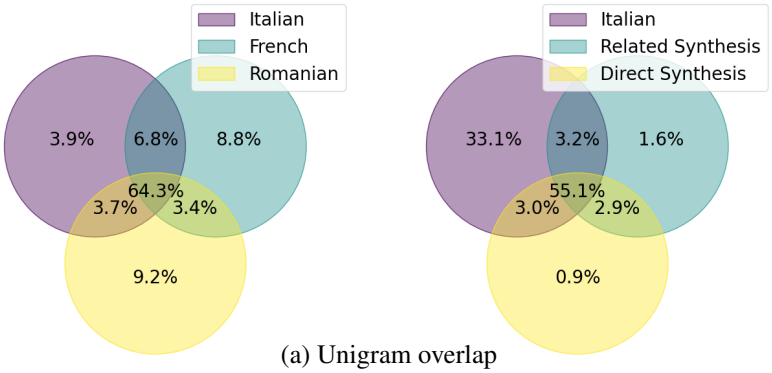
Knowledge Distillation	
Student Architecture	T5-small
Teacher Architecture	T5-base
λ	0.5
Batch Size	32
Optimizer	`Adam`
Learning Rate	$2 \cdot 10^{-5}$

Table 3: Hyper-parameters chosen for our knowledge distillation experiments. Any option not explicitly mentioned uses the default value in the PL trainer options.

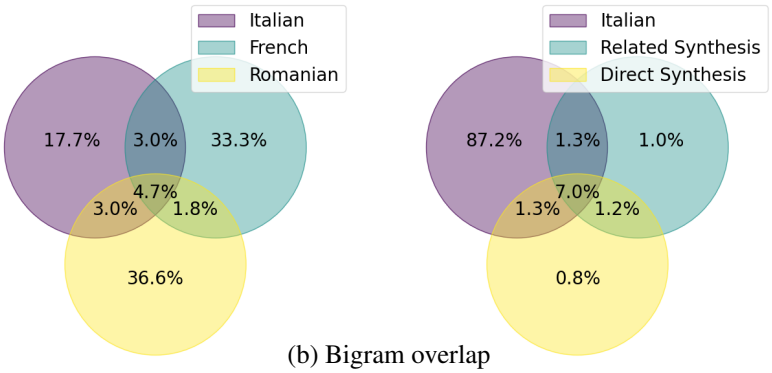
Table 3 summarizes the training setup for our distillation experiments. Notably, the pre-training of the teachers was exclusively performed in the dual-teacher scenario. In the case of single-teacher knowledge distillation, we directly fine-tuned it on the low-resource language pair.

C Corpus Similarity

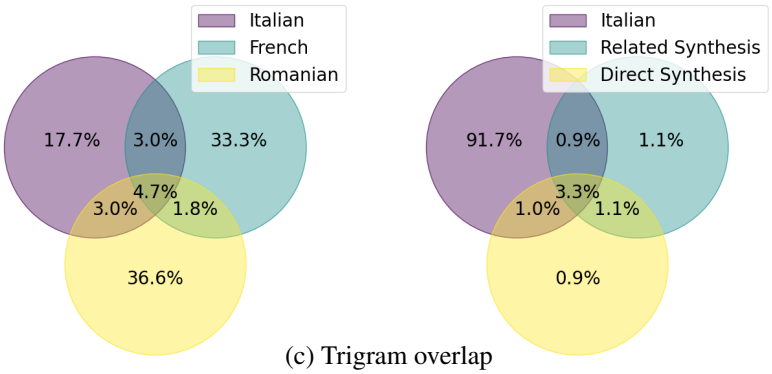
In the left side of Figure 3, we report the overlap between the Italian, French, and Romanian corpora, for the tokenized unigrams, bigrams, and trigrams. We notice a high similarity in unigrams, which is expected as all languages belong to the Romance family. However, as we progress to bigrams and trigrams, the overlap drops drastically. Similarly, on the right side, we observe the overlap for the Italian, Related Synthesis, and Direct Synthesis corpora. The overlap between the Italian and the synthetic data indicates the number of in-distribution data that were generated. We notice that the RS suffered from a higher distributional shift than DS, which corresponds to the inferior results we reported.



(a) Unigram overlap



(b) Bigram overlap



(c) Trigram overlap

Figure 3: Venn diagrams of the unigram (**top**), bigram (**middle**), and trigram (**bottom**) overlap between the Italian-Romanian-French (1M pairs each) corpora (**left**), and Italian-RS-DS (1M-24K-24K) (**right**).