Learning Nearest Neighbour Informed Latent Word Embeddings to Improve Zero-Shot Machine Translation

Nishant Kambhatla **Logan Born** **Anoop Sarkar**
School of Computing Science, Simon Fraser University
8888 University Drive, Burnaby BC, Canada
{nkambhat, loborn, anoop}@sfu.ca

Abstract

Multilingual neural translation models exploit cross-lingual transfer to perform zero-shot translation between unseen language pairs. Past efforts to improve cross-lingual transfer have focused on aligning contextual *sentence*-level representations. This paper introduces three novel contributions to allow exploiting nearest neighbours at the *token* level during training, including: (i) an efficient, gradient-friendly way to share representations between neighboring tokens; (ii) an attentional semantic layer which extracts latent features from shared embeddings; and (iii) an agreement loss to harmonize predictions across different sentence representations. Experiments on two multilingual datasets demonstrate consistent gains in zero shot translation over strong baselines.

1 Introduction

Many-to-many multilingual neural translation models (Firat et al., 2016; Johnson et al., 2017; Khandelwal et al., 2020; Fan et al., 2022) share a single representation space across multiple language pairs, which enables them to perform *zero-shot* translations between unseen pairs (Ha et al., 2017; Chen et al., 2022; Wu et al., 2022). Prior work on zero-shot translation has focused on aligning *contextual*, *sentence*-level representations from multiple languages (Ji et al., 2020; Pan et al., 2021a), to make these more 'universal' or language-agnostic (Gu et al., 2018; Gu and Feng, 2022). *Non*-contextual, *token*-level representations offer another space in which this kind of alignment could be pursued, but this space has not been thoroughly explored in prior work. Even lexicon-based methods (Conneau et al., 2020; Reid and Artetxe, 2022), which exploit token-level anchors from multilingual dictionaries (Duan et al., 2020), still use these to align representations at the sentence level.

In this work, we explore a novel technique for sharing information across languages at the *token* level, which exploits nearest neighbours (NNs) to aggregate information from subwords across multiple languages. When analysing embedding spaces, many authors speak in terms of "neighborhoods" or "subspaces" which group together tokens from a particular semantic field or other natural class. These neighborhoods form implicitly as a model learns similarities between embedded words or subwords. We propose to make this neighborhood structure *explicit* by forcing a model to consider a token's neighbors when learning its embedding. Specifically, we dynamically perturb a translation model's token embeddings at training time by averaging them with their nearest neighbors; thus a token like *soccer* may end up mixed together with related tokens such as *football*, *fußball*, or *futbol* from potentially distinct languages (Figure 1). This encourages the model to organize its subword embeddings in such a way that nearby tokens convey similar information to one another. We hypothesize that this process will produce a more structured embedding space which will in turn enable more fluent outputs. This process only uses the model's embedding layer, and does not require any offline dictionaries or additional data.

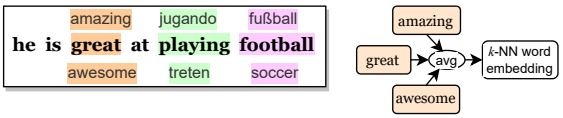

Figure 1: NN-informed embeddings average representations from nearby subwords in the embedding space.

Our experiments and ablations show that this simple technique significantly increases the effectiveness of translation models on the IWSLT17 and TED59 massively multilingual datasets. Concretely, our contributions include: (i) an efficient, gradient-friendly, soft representation-mixing technique which exploits token-level neighbors without changing the Transformer architecture; (ii) an attentional semantic layer which extracts features from

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 291–301

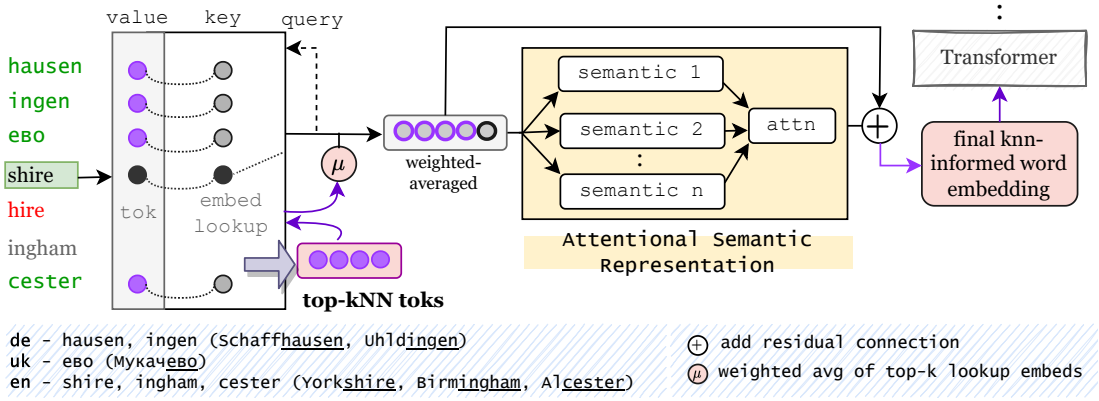

Figure 2: A NN-informed embedding for an arbitrary subword shire is produced by averaging across nearby subwords from various languages, and combining with a semantic representation extracted from this average.

mixed representations to give neighbour-informed latent word embeddings, and which is a drop-in replacement for a conventional embedding layer; and (iii) an agreement loss which harmonizes predictions with and without neighbor-informed embeddings.

2 Translation with Nearest Neighbour Augmented Embeddings

We describe our model for *nearest-neighbour informed token level embeddings* (Figure 2) of subwords from multiple source languages.

Nearest Neighbor Retrieval Let $\mathcal{L}_{emb}$ be a word embedding layer that performs a lookup $\text{EMB}(\cdot)$ using weights $\mathcal{W}_{emb} \in \mathbb{R}^{|\mathcal{V}| \times D}$, where $\mathcal{V}$ is a joint subword vocabulary over all languages and D is a fixed embedding dimension. Given the embedding $q = \text{EMB}(w) \in \mathbb{R}^{1 \times D}$ of a subword w, we wish to find q's nearest neighbour n (or neighbors $n_1, ..., n_k$) using maximum inner product search (MIPS) over the weight matrix $\mathcal{W}_{emb}$:

$$\begin{aligned} n &= \arg\min_{x \subset \mathcal{W}_{emb}} ||q - x||_2^2 \\ &= \arg\min_{x \subset \mathcal{W}_{emb}} (||x||_2^2 - 2q^T x) \end{aligned} \quad (1)$$

Approximate solutions to (1) can be efficiently computed on-the-fly using anisotropic vector quantization (Guo et al., 2020).[1]

Given the approximate nearest neighbors (ANNs) $n_1, ..., n_k$ of subword w, we compute a weighted average over these tokens' embeddings with a weighting term λ:

$$\text{EMB}_\mu(w) = \lambda \frac{1}{k} \sum_{i=1}^{k} (\text{EMB}(n_i)) + (1 - \lambda)\text{EMB}(w) \quad (2)$$

$\text{EMB}_\mu(\cdot)$ is computed directly from $\mathcal{W}_{emb}$, which ensures that our technique remains gradient-friendly[2] and does not need a separate warm-up step. Previous NN-based proposals for translation (Khandelwal et al., 2020) and language modeling (Khandelwal et al., 2019) have only explored NNs of contextualized representations, strictly for generation, and using neighbors from an offline *frozen datastore* of pretrained candidates. Their method proved effective for MT domain adaptation, rather than zero-shot translation which is the focus of this work. The ability to propagate gradients to a subword's neighbors during training is novel and unique compared to previous NN-based techniques.

Attentional Semantic Representation To extract contextually-salient information from $\text{EMB}_\mu(w)$, which combines information from many subwords in potentially disparate languages, we use a shared semantic embedding inspired by Gu et al. 2018; Wang et al. 2018 that shows a similar effect as topical modelling.

We introduce $\mathcal{W}_{sem} \in \mathbb{R}^{\mathcal{N} \times D}$, where each of the $\mathcal{N}$ rows is taken to be a language-agnostic semantic representation. $\mathcal{W}_{sem}$ is shared across all languages. We use attention (Luong et al., 2015) to compute a latent embedding $\text{EMB}_{latent}(w)$ using

[1] Exact and approximate solutions yield similar results, but approximation gives significant gains in training speed.

[2] In Section 3.3 we introduce a caching heuristic which is not gradient-friendly; however, this is simply an implementation detail to speed up training, and the gradient-friendly presentation in this section achieves equivalent performance.

	De - It		De - Nl		De - Ro		It - Nl		It - Ro		Nl - Ro		zero	sup.
	←	→	←	→	←	→	←	→	←	→	←	→		
Base M2M	15.64	15.28	18.46	18.14	14.42	14.98	18.16	18.79	17.91	20.14	15.81	16.41	17.01	**30.62**
SRA (2019)	16.44	16.45	18.44	19.15	15.07	15.83	**19.30**	19.10	18.52	21.52	16.83	17.66	17.85	30.41
SF (2019)	16.34	15.77	18.37	18.16	14.74	15.25	18.6	19.18	18.54	21.64	16.09	16.94	17.46	30.50
LV (2021)	16.82	15.81	18.74	18.64	15.12	**16.32**	18.92	19.29	18.70	22.13	16.21	18.22	17.91	30.51
CL (2021b)	17.31	16.21	19.70	**19.57**	15.32	16.25	18.90	20.09	19.07	**22.44**	17.14	17.99	18.33	30.29
DP (2021)	16.62	15.64	19.64	18.78	15.07	15.96	19.01	**20.15**	18.67	21.56	16.46	18.18	17.97	30.49
Ours	**17.41**	**16.89**	**19.71**	**19.21**	**15.60**	16.22	**19.30**	20.10	**19.60**	21.88	**17.25**	**18.40**	**18.47**	**30.62**

Table 1: BLEU on IWSLT17 test set (mean of 3 runs). `Zero` and `sup.` are average zero-shot and supervised results.

the averaged embedding $\text{EMB}_\mu(w)$ as query:

$$\text{EMB}_{latent}(w) = \text{Softmax}(\text{EMB}_\mu(w).\mathcal{W}_{sem}^T)\mathcal{W}_{sem} \quad (3)$$

A residual connection from $\text{EMB}_\mu(w)$ gives the final NN-informed word embedding:

$$\text{EMB}_{knn}(w) = \text{EMB}_{latent}(w) + \text{EMB}_\mu(w) \quad (4)$$

$\text{EMB}_{knn}(w)$ is a drop-in replacement for a conventional word embedding $\text{EMB}(w)$.

Modelling Prediction Consistency Given a source sentence represented using conventional word embeddings and using NN-informed embeddings, following Kambhatla et al. (2022b) we model the loss with respect to target sentence y_i as:

$$\begin{aligned}\mathcal{L}^i = & \underbrace{\alpha_1 \mathcal{L}^i_{NLL}(p_\Theta(y_i|x_i))}_{\text{source x-entropy}} \\ + & \underbrace{\alpha_2 \mathcal{L}^i_{NLL}(p_\Theta(y_i|kNN(x_i)))}_{k\text{-NN embeds. source x-entropy}} \\ + & \underbrace{\beta \mathcal{L}^i_{dist}(p_\Theta(y_i|x_i), p_\Theta(y_i|kNN(x_i)))}_{\text{agreement loss}}\end{aligned} \quad (5)$$

where $k\text{NN}(x_i)$ denotes the set of k-nearest neighbors to token x_i. This loss combines three terms: the first two are conventional negative log-likelihoods, while the third is an *agreement loss* measuring pairwise symmetric KL divergence between the output distributions for x_i and $kNN(x_i)$. This agreement-loss term performs *co-regularization* by allowing explicit interactions between source sentences with and without NN-informed embeddings.

3 Experiments

3.1 Datasets

We conduct experiments on 2 multilingual datasets, each with BPE (Sennrich et al., 2016) vocabulary size of 32k subwords:

IWSLT17 (Cettolo et al., 2012) is an English-centric dataset[3] totalling 1.8M parallel sentences. It has 8 supervised directions to and from German, Italian, Dutch and Romanian, each with about 220,000 parallel sentences, and 12 zero-shot directions. We use the official validation and test sets.

Ted59 (Qi et al., 2018) is a massively multilingual English-centric dataset[4] with 116 translation directions totalling 10.8M parallel sentences. The imbalanced data—from 0.25M to just 2000 parallel samples for some language pairs—makes it ideal to study the effects of our method. Following (Aharoni et al., 2019; Raganato et al., 2021) we evaluate on 16 supervised pairs and 4 zero-shot (Arabic ↔ French, Ukranian ↔ Russian).

3.2 Baselines and Related Work

We compare against methods for encoder manifold alignment. These include strong baselines such as sentence representation alignment (SRA; Arivazhagan et al. 2019), softmax forcing (SF; Pham et al. 2019), the contrastive multilingual model (CL; Pan et al. 2021b), multilingual Transformer with disentagled positional embedding (DP; Liu et al. 2021), and latent variable based denoising (LV; Wang et al. 2021), along with the vanilla many-to-many zero-shot model (M2M). On TED59, we compare against CL and 3 explicit multilingual alignment techniques proposed by Raganato et al. (2021): word-alignment, language tag alignment, and the union of the two. We also implement and compare against Raganato et al.'s (2021) sparse 1.5`entmax` cross-attention variant.

3.3 Model and Implementation Details

All models use the configuration in Vaswani et al. 2017 using the `fairseq` toolkit (Ott et al., 2019). See reproducibility details in Appendix A.

[3] https://wit3.fbk.eu/2017-01

[4] github.com/neulab/word-embeddings-for-nmt

	Θ	En→X	X→En	Zero-Shot	Acc_0
Aharoni et al. – 106 langs	473M	20.11	29.97	9.17	-
Aharoni et al. – 59 langs	93M	19.54	28.03	-	-
Transformer M2M reimp.	93M	18.98	27.22	7.12	74.10
Constrastive (2021b)	93M	19.09	27.29	8.16	73.90
Ours	77M	19.01	27.11	**10.03**	**95.81**
Raganato et al. (2021)					
ZS + 1.5entmax (ibid.)	93M	18.90	27.21	10.02	87.81
↳ Word Align (ibid.)	93M	18.99	27.58	8.38	73.12
↳ LangID Align (ibid.)	93M	18.98	27.48	6.35	65.01
↳ Word + LangID Align	93M	19.06	27.37	11.94	97.25
Ours + 1.5entmax	77M	18.94	27.42	**12.11**	**98.90**

Table 2: Average BLEU scores on the TED59 dataset. Our model produces zero-shot translations in the correct output language with high accuracy (Acc_0).

We use ScANN (Guo et al., 2020) for efficient ANN search [5] with $k = 3$. To increase training speeds, we cache each subword's ANNs for 400 iterations before recomputing them. We only (peridocally) cache subword IDs: the embedding $\text{EMB}_\mu(\cdot)$ is always computed directly from $\mathcal{W}_{emb}$. We set $\lambda = 0.5, \alpha_1, \alpha_2 = 1,$ and $\beta = 5$. The *attentional latent semantic representation* layer has 512 dim (same as the embedding layer) and a size $\mathcal{N}$ of 1000 for IWSLT17 (smaller dataset) and 5000 for TED59 (larger dataset). We did not tune this hyperparameter and chose the values based on the size of the datasets. For evaluation, we report `sacreBLEU` (Post, 2018).

3.4 Results

Main Results. Tables 1 and 2 show our main results. On IWSLT17, our latent k-NN embedding model outperforms several strong baselines, including sentence-representation alignment and contrastive learning, by an average of 0.62 and 0.11 BLEU respectively across the 12 zero-shot pairs. Compared to the baseline many-to-many model, our method yields a 1.5 BLEU gain on average. Our method is able to improve zero-shot performance without deteriorating supervised performance.

On the TED59 dataset, we follow Raganato et al. (2021) in comparing against two multilingual model variants: the standard Transformer, and the Transformer with sparse `entmax` instead of standard softmax cross-attention. Our approach gains ∼3 BLEU points against the baseline, and 2 BLEU against the stronger contrastive model. Further, our model consistently outperforms strong, explicitly alignment-based methods.

Target-language Accuracy. To supplement the evaluation, we provide the accuracy score for target language identification[6] in zero-shot scenarios, called Acc_0. While the classical many-to-many NMT models (Johnson et al., 2017; Aharoni et al., 2019) enable zero-shot translations, several studies have shown that these models fail to reliably generalize to unseen language pairs, ending up with an *off-target* translation issue (Zhang et al., 2020). The model ignores the language label and the wrong target language is produced as a result. We observe significant improvements in target language accuracy, up to nearly 99% (absolute).

4 Analysis

Ablation Study. Table 3 reports ablations on the IWSLT17 test set. We find that kNN embeddings alone yield improvements over the baseline many-to-many model. By contrast, absent the other parts of our model, the attentional semantic layer *deteriorates* model performance. Only in combination with the agreement loss do we observe a benefit from this component.

Embedding Analysis. Figure 3 visualizes subword representations from models trained on IWSLT17. Each subword is colored according to the language in which it is most frequent. The overall layout of the two spaces is similar, although the

[5] We use asymmetric hashing with 2-dimensional blocks and a quantization threshold of 0.2, and re-order the top 100 ANN candidates.

[6] We utilize FastText (Joulin et al., 2017) as a language identification tool to compare the translation language with the reference target language and keep count of the number of matches.

ID	Component	dev.2010	test.2010
1	many-to-many (zero-shot)	15.95	18.46
2	① + attn. semantic repr.	15.43	17.83
3	① + kNN embeds	17.11	19.69
4	② + kNN embeds	16.60	19.08
5	③ + agreement loss	17.99	20.91
6	④ + agreement loss	**18.31**	**21.01**

Table 3: Effect of different components of our model on the IWSLT17 datasets. We report sacreBLEU scores on the two official validation sets with beam size 1.

baseline model (left) exhibits a clear ring-shaped gap dividing the embeddings into two groups. With ANN embeddings (right), this gap is eliminated and the layout of the embeddings appears more homogeneous. Quantitatively, the average distance from a subword to its neighbors exhibits a smaller variance in the ANN model than in the baseline, which further supports the reading that ANN training creates a more homogeneous representation space in which subwords are more uniformly distributed.

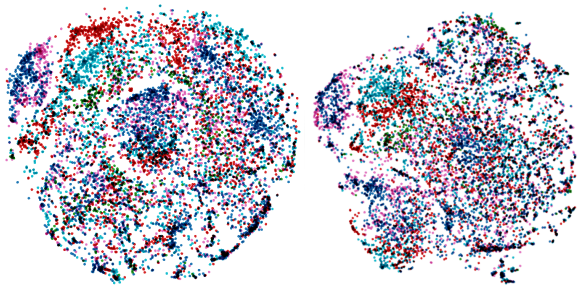

Figure 3: t-SNE visualization of subword embeddings from IWSLT17 models trained without (left) and with (right) ANN embeddings. Points are colored according to the language where the corresponding subword is most frequent. ANN embeddings decrease the separation between some monolingual subspaces, and remove others entirely.

Table 4 shows nearest neighbors for a random sample of subwords (additional examples in Table 5 in Appendix B). With ANN training, a subword's nearest neighbors are generally its synonyms (e.g. `_wonderful`, `_large` `_tremendous`, and `_big` as neighbors to `_great`) or derived forms (e.g. `_încep`, `_începem`, `_început`, `_începe` beside `_început`). In the baseline, it is more likely to find neighbors with no apparent relation, such as `_erzählen` 'tell' and `_stemmen` 'hoist' or 'accomplish' beside `_America`. This suggests that ANN embeddings help a model to better organize its subword embedding space into coherent, semantically-related subspaces.

We quantify this trend by labeling each subword according to the language in which it is most frequently attested. In the baseline model, we find that on average only 2.7 of a subword's 6 nearest neighbors come from the same language as that subword. This average rises to 3.6 in the ANN model, demonstrating that ANN training significantly increases the number of same-language neighbors on average.

In the ANN model, a few rare subwords (√, ž, ć) are disproportionately common among the nearest neighbors of many other subwords. We speculate that these tokens may act as pivots for information to flow between their many neighbours. Their high centrality means that these tokens provide avenues for information to flow between a large number of subwords, even those which never occur in sentences together. Because these tokens are rare, there is also very little penalty for the model to "corrupt" their representations with information from neighboring subwords.

5 Other Related Work

A vast body of work addresses zero-shot translation. Most methods focus on producing language-agnostic encoder outputs (Pham et al., 2019). Wei et al. (2021) introduce multilingual contrastive learning, while Yang et al. (2021) adopt auxiliary target language prediction. To enable the input tokens to be positioned without constraints, Liu et al. (2021) eliminate the residual connections within a middle layer of the encoder. Yang et al. (2022); Gu and Feng (2022) employ optimal transport to improve contextual cross-alignments, in contrast to our method which performs *soft*, non-contextual alignment between subwords in the continuously-updating embedding space. Other methods extend the training data using monolingual data (Al-Shedivat and Parikh, 2019) to pretrain the decoder (Gu et al., 2019), and random-online backtranslation (Zhang et al., 2020). Lin et al. (2021); Reid and Artetxe (2022) use dictionary based alignments to produce pseudo-cross-lingual sentences. Other approaches that enhance token level representations include multiple subword segmentations (Wu et al., 2020; Kambhatla et al., 2022a), enciphered source text (Kambhatla et al., 2022b) and stroke sequence modelling (Wang et al., 2022). While all these techniques rely on multilingual training paradigm for machine translation, they either rely on external data and use explicit augmentations. We do not

Subword	Nearest Neighbors (Baseline)						Nearest Neighbors (Ours)					
_great	_gesproken	_schaffen	ppy	ită	_prosper	_senior	_wonderful	_large	_tremendous	_big	_great	√
_începeapă	_popolare	_condotto	_mişcă	_bekijken	_crească	_creeze	_gepubliceerd	_încep	_începem	_început	_începe	muovono
_America	tate	_erzählen	_stemmen	dine	_facultate	_chestiune	_USA	_Asia	_Africa	_American	_America	ć
_play	_lavori	eranno	_tenuto	_bekijken	-	möglichkeiten	play	_playing	_Play	_played	_play	√
_football	_pesci	bon	_surf	_betrachten	_Hintergrund	möglichkeiten	_weather	_baseball	ball	_montagna	_biodiversità	_football
ing	ificazione	izăm	amento	tung	erende	ende	ling	ting	ung	ž	ingen	ing
_fish	_petrec	schen	_Sachen	_feed	_chestii	möglichkeiten	fisch	_pesce	_pesca	_Fisch	_fish	√

Table 4: Approximate nearest neighbors for a sample of subwords, computed with (right) and without (left) ANN training.

use any external data or explicit alignments and our model can be trained end-to-end like a regular multilingual model.

6 Conclusion

We described a novel approach to harness nearest neighbors at the token level and learn *nearest-neighbour informed word embeddings* for every word in a source language for many-to-many multilingual translation. Our experiments show that this simple yet effective approach results in consistently better zero-shot translations across multiple multilingual datasets. Additionally, our model produces translations in the right target language with high accuracy. Our analysis shows that our model learns to organize subwords into semantically-related neighborhoods, and reduces the separation between monolingual subspaces in the embedding space.

Limitations

While our method is effective in zero-shot settings, we find that it has limited implications in supervised settings. This is because improving zero-shot translation presents a tug-of-war between language-agnostic and language-specific representations, each of which has a distinct effect on the model. Another major downside is reduced training speed relative to the baseline many-to-many model. We note that this is an artifact of the agreement loss (KLDiv.) which entails two forward-passes for each update. Finally, in the present work, we compute k-NNs for every source word in a sentence. Although this has yielded strong results, we would like to explore a more explainable setting where k-NNs can be applied to specific source words. We leave such explorations to future work.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. The research was partially supported by the Natural Sciences and Engineering Research Council of Canada grants NSERC RGPIN-2018-06437 and RGPAS-2018-522574 and a Department of National Defence (DND) and NSERC grant DGDND-2018-00025 to the third author, and by an NSERC award CGSD3-547773-2020 to the second author.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. 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. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2022. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. 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. Association for Computational Linguistics.

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.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. 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. Association for Computational Linguistics.

Shuhao Gu and Yang Feng. 2022. Improving zero-shot multilingual translation with universal representations and cross-mappings. In *Proceedings of the EMNLP 2022 Long Findings*.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pages 3887–3896. PMLR.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective strategies in zero-shot neural machine translation. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 105–112, Tokyo, Japan. International Workshop on Spoken Language Translation.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 115–122.

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. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022a. Auxiliary subword segmentations as related languages for low resource multilingual translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 131–140, Ghent, Belgium. European Association for Machine Translation.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022b. CipherDAug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 201–218, Dublin, Ireland. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Yusen Lin, Jiayong Lin, Shuaicheng Zhang, and Haoying Dai. 2021. Bilingual dictionary-based language model pretraining for neural machine translation.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. 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 1259–1273, Online. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021a. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of ACL 2021*.

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

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.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alex Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful 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 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2021. An empirical investigation of word alignment supervision for zero-shot multilingual neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8449–8456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Machel Reid and Mikel Artetxe. 2022. PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810, Seattle, United States. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weizhi Wang, Zhirui Zhang, Yichao Du, Boxing Chen, Jun Xie, and Weihua Luo. 2021. Rethinking zero-shot neural machine translation: From a perspective of latent variables. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4321–4327, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2018. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Zhijun Wang, Xuebo Liu, and Min Zhang. 2022. Breaking the representation bottleneck of Chinese characters: Neural machine translation with stroke sequence modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6473–6484, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

Lijun Wu, Shufang Xie, Yingce Xia, Yang Fan, Jian-Huang Lai, Tao Qin, and Tieyan Liu. 2020. Sequence generation with mixed representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10388–10398. PMLR.

Shijie Wu, Benjamin Van Durme, and Mark Dredze. 2022. Zero-shot cross-lingual transfer is under-specified optimization. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 236–248, Dublin, Ireland. Association for Computational Linguistics.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation

and gradient regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7266–7279, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhe Yang, Qingkai Fang, and Yang Feng. 2022. Low-resource neural machine translation with cross-modal alignment. pages arXiv–2210.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Reproducibility Details

A.1 Data

IWSLT17 (Cettolo et al., 2012) is an English-centric dataset[7] totalling 1.8M parallel sentences. It has 8 supervised directions to and from German, Italian, Dutch and Romanian, each with about 220,000 parallel sentences, and 12 zero-shot directions. We use the official validation and test sets.

Ted59 (Qi et al., 2018) is a massively multilingual English-centric dataset[8] with 116 translation directions totalling 10.8M parallel sentences. The imbalanced data—from 0.25M to just 2000 parallel samples for some language pairs—makes it ideal to study the effects of our method. Following (Aharoni et al., 2019; Raganato et al., 2021) we evaluate on 16 supervised pairs (Azerbaijani, Belarusian, Galician, Slovak, Arabic, German, Hebrew, and Italian to and from English) and 4 zero-shot (Arabic ↔ French, Ukranian ↔ Russian). Note that of these languages, Azerbaijani, Belarusian, Galician, and Slovak are low resource with only 5.9k, 4.5k, 10k and 61.5k paralle samples to/from English.

All settings and baselines use `sentencepiece`[9] for subword tokenization using byte-pair encodings (BPEs; Sennrich et al. 2016) with 32000 merge operations.

A.2 Model and Hyperparameters

All models follow the basic configuration of Vaswani et al. (2017), using the `fairseq` toolkit (Ott et al., 2019) in PyTorch. This includes 6 layers of encoder and eecoder each with 512 dim and 2048 feed-forward dimension. The 512 dim word embedding layer has a vocabulary size of 32000. All word-embeddings in the model (encoder, decoder input/output) are shared, although the latent embedding layer alone is specific to encoder only. This implies that any updates to the actual embedding layer because of k-NN tokens also impacts the decoder.

The *attentional latent semantic representation* layer has 512 dim (same as the embedding layer) and a size $\mathcal{N}$ of 1000 for IWSLT17 (smaller dataset) and 5000 for TED59 (larger dataset). We did not tune this hyperparameter and chose the values based on the size of the datasets. This implies that this layer adds 0.5M trainable parameters to the IWSLT17 model and 2.5M parameters to the TED59 model. However, note that the total trainable parameters are still much lower than that of the baselines – this because our models have shared embedding layers.

We use the Adam optimizer with inverse square root learning scheduling and 6k warm steps, $lr = 0.0007$ and dropout of 0.3 (IWSLT17), or 10k warmup steps, $lr = 0.005$ and dropout of 0.2 (TED59). The batch size is 4096 tokens for each of four A100 GPUs.

We use ScANN (Guo et al., 2020) for efficient ANN search[10] with $k = 3$. To increase training speeds, we cache each subword's ANNs for 400 iterations before recomputing them. We only (peridocally) cache subword IDs: the embedding $\text{EMB}_\mu(\cdot)$ is always computed directly from $\mathcal{W}_{emb}$. The value of λ is set to 0.5 (Equation 1). We follow Kambhatla et al. (2022b) to set the values of α_1, α_2 to 1, and β to 5 (Equation 5).

Evaluation. For evaluation, all translations are generated with beam size 5. We report case-sensitive BLEU scores (Papineni et al., 2002) using `sacreBLEU`[11] (Post, 2018). We report detokenized BLEU for IWSLT17 and tokenized BLEU for TED59 for fair comparison with prior work (Aharoni et al., 2019; Raganato et al., 2021).

B Nearest Neighbor Examples

See Table 5.

[7] https://wit3.fbk.eu/2017-01

[8] github.com/neulab/word-embeddings-for-nmt

[9] https://github.com/google/sentencepiece

[10] https://github.com/google-research/google-research/tree/master/scann . We use asymmetric hashing with 2-dimensional blocks and a quantization threshold of 0.2, and re-order the top 100 ANN candidates.

[11] case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

Subword	Nearest Neighbors (Baseline)						Nearest Neighbors (Ours)					
_Fisch	_findet	œ	_chestii	_Netz	fisch	möglichkeiten	erei	fisch	_pesca	_fish	_Fisch	ž
schaft	hood	erung	ungen	gaat	_gehabt	schaft	würdig	lichkeit	ship	nisse	äglich	schaft
the	tje	ped	own	asta	by	_solamente	tech	ther	th	by	the	√
_the	isce	izăm	_erzählen	”&	_gehabt	oara	_your	_their	_our	_the	ć	ž
_Music	mat	_cartoon	hood	_connessione	zia	_şcoala	_musica	_music	ž	_Music	dine	ć
_picior	_sfârşit	_plaatje	_mesaj	_teren	_avion	_gehabt	_corpul	_braț	_pagină	_picior	ž	√
ern	eien	iere	eren	erung	_tenuto	_gehabt	uren	ungen	ert	eren	stern	ern
_începapă	_popolare	_condotto	_mişcă	_bekijken	_crească	_creeze	_gepubliceerd	_încep	_începem	_început	_începe	muovono
_democrația	analisi	_înțelege	_popolare	izăm	_şcoala	_deshalb	_terorism	muovono	_democratic	dine	_biodiversità	ć
_pure	rische	_giovane	_appena	_tare	\u0e22	_avesse	_semplicemente	_unique	_tragic	_complete	_sole	_pure
_genomic	_finanzia	Â	_popolare	_răspândi	_genomen	möglichkeiten	_electronic	_genome	_robotic	ž	_genetic	_genomic
_Abbiamo	_perciò	_gehabt	_spunem	_condotto	izăm	_avesse	abbiamo	mmo	iamo	_Abbiamo	_abbiamo	ć
izări	amento	isieren	ierung	izzazione	_răspândi	izare	izare	ități	aţie	izări	muovono	nelli
_negative	_altele	azioni	iere	_bune	_enormous	oase	_illegal	_alternative	_evil	_positive	_negativ	_negative
_take	_solamente	_gemacht	_spinge	_accompagna	_preso	_tenuto	_takes	_taken	_taking	_took	ć	_take
_muziek	_percorso	_besef	_onderwijs	_erzählen	_vreugde	oara	_music	muovono	_Musik	_musica	_muziek	ć
_Karte	_Bibliothek	_lavori	strategie	_chestii	_cifre	kaart	_Weise	_Sprache	_carta	_montagna	kjes	_Karte
_funcționa	_mişcă	_munci	matig	_realiza	_funcție	_funcționa	_funcţionează	_funcționează	_funziona	_funcţiona	_funcționa	ć
_național	_popolare	iere	_bază	_condotto	_esențial	_politic	juist	_rural	äglich	_National	_național	_national
_America	tate	_erzählen	_stemmen	dine	_facultate	_chestiune	_USA	_Asia	_Africa	_American	_America	ć

Table 5: Approximate nearest-neighbors for a sample of subwords, computed with (right) and without (left) ANN training.