Multilingual Agreement for Multilingual Neural Machine Translation

Jian Yang[1] [*], **Yuwei Yin**[1] [*], **Shuming Ma**[2], **Haoyang Huang**[2],
Dongdong Zhang[2], **Zhoujun Li**[1] [†], **Furu Wei**[2]
[1]State Key Lab of Software Development Environment, Beihang University
[2]Microsoft Research Asia
{jiaya, yuweiyin, lizj}@buaa.edu.cn; {shumma, haohua, dozhang, fuwei}@microsoft.com

Abstract

Although multilingual neural machine translation (MNMT) enables multiple language translations, the training process is based on independent multilingual objectives. Most multilingual models can not explicitly exploit different language pairs to assist each other, ignoring the relationships among them. In this work, we propose a novel agreement-based method to encourage multilingual agreement among different translation directions, which minimizes the differences among them. We combine the multilingual training objectives with the agreement term by randomly substituting some fragments of the source language with their counterpart translations of auxiliary languages. To examine the effectiveness of our method, we conduct experiments on the multilingual translation task of 10 language pairs. Experimental results show that our method achieves significant improvements over the previous multilingual baselines.

1 Introduction

Multilingual neural machine translation (MNMT) has experienced rapid growth in recent years (Johnson et al., 2017; Zhang et al., 2020; Aharoni et al., 2019; Wang et al., 2019). It is not only capable of translating among multiple language pairs by encouraging the crosslingual knowledge transfer to improve low-resource translation performance (Firat et al., 2016b; Zoph et al., 2016; Sen et al., 2019; Qin et al., 2020; Hedderich et al., 2020; Raffel et al., 2020), but also can handle multiple language pairs in a single model, reducing model parameters and training costs (Firat et al., 2016a; Blackwood et al., 2018; Wang et al., 2020; Sun et al., 2020).

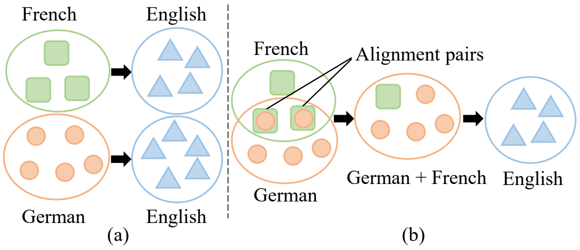

Figure 1: Comparison between (a) the multilingual translation and (b) our agreement-based method.

Previous works in MNMT simply optimize independent translation objectives and do not use arbitrary auxiliary languages to encourage the agreement across different translation directions. As shown in Figure 1, the multilingual baseline is separately trained on French-English and German-English directions and cannot explicitly promote each other. The German-English translation only implicitly helps the French-English translation since both translation directions share the same encoder. There still exists a gap between German-English and French-English translation directions. As a result, minimizing the difference across different translation directions by an explicit paradigm requires further exploration.

In this paper, we propose a novel agreement-based method, which explicitly models the shared semantic space for multiple languages and encourages the agreement across them. Our training procedure extends the multilingual translation with the agreement term, which encourages the model to produce the source sentence with multiple languages into the target sentence. As Figure 1 shows, we randomly substitute some source phrases with their counterparts of other languages to create code-switched sentences using word alignment. Our model is jointly trained with the multilingual translation and agreement objectives, where the code-switched sentences are translated into the target sentences. The key idea is to encourage the agreement among different translation directions simul-

[*]Contribution during internship at Microsoft Research Asia.

[†]Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 233–239
August 1–6, 2021. ©2021 Association for Computational Linguistics

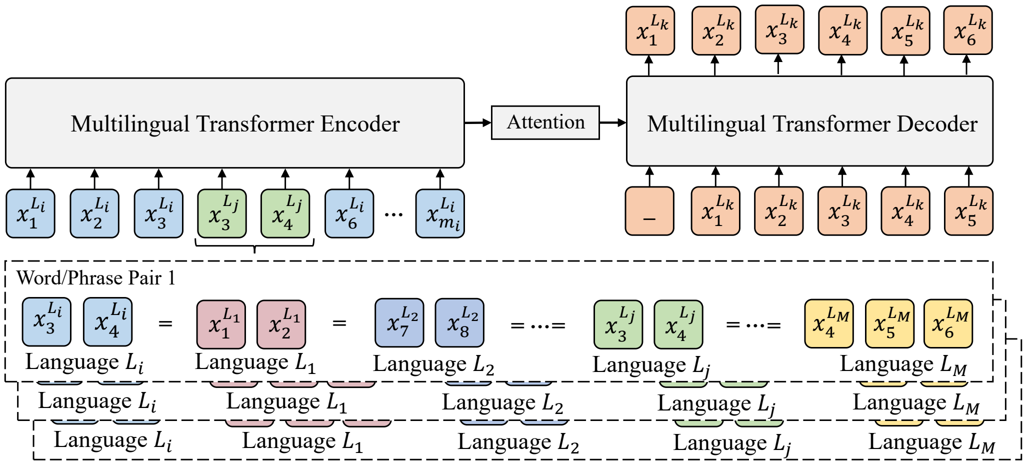

Figure 2: Overview of our method. $x_{m_i}^{L_i}$ denotes the m_i-th token in the sentence of language L_i. We randomly substitute source phrases of language $L_{src} = L_i$ with the translations of other languages $L_{aux} \in L_{all}$ to create code-switched sentences. Different words/phrases with the same meanings may contain different numbers of tokens. Then the code-switched source sentences are translated to the target language $L_{tgt} = L_k$ by the multilingual model. This process greatly encourages multilingual agreement across different translation directions.

taneously by leveraging alignment information of the bilingual source sentence pairs.

Experimental results on the multilingual translation task of WMT demonstrate that our method outperforms the multilingual baseline by a large margin. To better explain the BLEU improvements, we visualize the sentence-level crosslingual representations and the attention weights across different languages, which shows that our method effectively encourages the agreement between languages.

2 Our Approach

2.1 Multilingual Machine Translation

Our multilingual model is based on the single Transformer model (Vaswani et al., 2017) and shares all embedding matrices by a common vocabulary of all languages. Given M languages $L_{all} = \{L_1, \dots, L_M\}$, the multilingual model appends special symbols to the source text to indicate the translation direction from the source language L_{src} to the target language L_{tgt}.

2.2 Agreement-based Training

Multilingual models can translate multiple source-side languages into target-side languages. Given N bilingual corpora $D_B = \{D_{B_1}, \dots, D_{B_N}\}$, the multilingual model with parameters θ is jointly trained over N language directions to optimize the combined objective as below:

$$\mathcal{L}_{MT} = \sum_{n=1}^{N} \mathbb{E}_{x,y \in D_{B_n}} [-\log P_\theta(y|x)] \quad (1)$$

where x, y denote the sentence pair in the bilingual corpus D_{B_n}. $\mathcal{L}_{MT}$ is the combined translation objective of the multilingual model.

The agreement objective over the code-switched corpora D_C is calculated by:

$$\mathcal{L}_{AT} = \mathbb{E}_{x^{L_{src}/L_{aux}}, y \in D_C} [-\log P_\theta(y|x^{L_{src}/L_{aux}})] \quad (2)$$

where $x^{L_{src}/L_{aux}}$ is the code-switched sentence in which some phrases are substituted by their counterpart phrases in other languages and y is the target sentence. L_{aux} is the auxiliary language.

We combine the bilingual corpora D_B and code-switched corpora D_C to train our agreement-based model, which minimizes the gaps among different translation directions using word alignment:

$$\mathcal{L}_{ALL} = \mathcal{L}_{MT} + \mathcal{L}_{AT} \quad (3)$$

where $\mathcal{L}_{ALL}$ is the combined objective.

2.3 Constructing Training Samples

We use L_{src} as the source language, L_{tgt} as target language, and L_{aux} as auxiliary languages to construct training samples. As shown in Figure 2, $x^{L_{src}} = (x_1^{L_{src}}, \dots, x_m^{L_{src}})$ is the source sentence with m tokens and $x^{L_{aux}} = (x_1^{L_{aux}}, \dots, x_n^{L_{aux}})$ is the auxiliary sentence with n tokens. $x_{u:v}^{L_{src}}$ denotes the sentence fragment of $x^{L_{src}}$ from the u-th to v-th token and $x_{s:t}^{L_{aux}}$ denotes the fragment of $x^{L_{aux}}$ from the s-th to t-th token, where $x_{s:t}^{L_{aux}}$ of language L_{aux} is the translation of the $x_{u:v}^{L_{src}}$ of language L_{src}. Formally, the code-switched sequence

En → X	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Bilingual NMT	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
One-to-Many	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
One-to-Many + Pseudo	35.5	21.7	42.0	16.4	19.3	22.0	26.6	16.2	17.9	17.8	23.5
One-to-Many + AT (our method)	35.7	22.0	42.1	16.6	20.1	22.2	26.9	16.6	18.2	17.9	**23.9**

Table 1: En→X test results for bilingual and multilingual models of 10 language pairs on the WMT benchmark.

X → En	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Bilingual NMT	36.2	28.5	40.2	19.2	17.5	19.7	29.8	14.1	15.1	9.3	23.0
Many-to-One	34.8	29.0	40.1	21.2	20.4	26.2	34.8	22.8	23.8	19.2	27.2
Many-to-One + Pseudo	35.4	30.1	42.1	22.0	21.2	29.0	35.8	27.3	26.0	22.6	29.1
Many-to-One + AT (our method)	35.7	30.2	42.6	22.3	21.8	29.5	36.4	27.6	26.7	22.8	**29.6**

Table 2: X→En test results for bilingual and multilingual models of 10 language pairs on the WMT benchmark.

$x^{L_{src}/L_{aux}}$ is described as:

$$x^{L_{src}/L_{aux}} = (x_1^{L_{src}}, \dots, x_{s:t}^{L_{aux}}, \dots, x_m^{L_{src}}) \quad (4)$$

where most words in the code-switched sentence $x^{L_{src}/L_{aux}}$ are derived from $x^{L_{src}}$, while some source phrases $x_{u:v}^{L_{src}}$ are substituted by their counterpart phrases $x_{s:t}^{L_{aux}}$.

Given the parallel sentences among M different languages, we can construct code-switched source sentence $x^{L_{src}/L_{aux}}$ with different auxiliary languages. Therefore, the code-switched corpora D_C can be constructed in a similar way for other languages to encourage the agreement across different translation directions to help each other.

3 Experiment Setup

3.1 Multilingual Data

We use the same training, valid, and test sets as the previous work (Wang et al., 2020) to evaluate multilingual models by parallel data from multiple WMT datasets with various languages, including English (En), French (Fr), Czech (Cs), German (De), Finnish (Fi), Latvian (Lv), Estonian (Et), Romanian (Ro), Hindi (Hi), Turkish (Tr), and Gujarati (Gu). For each language, we concatenate the WMT data of the latest available year and get at most 10M sentences by randomly sampling. Detailed statistics of datasets are listed in Table 3. All sentences in our experiments are tokenized by SentencePiece[1] (Kudo and Richardson, 2018).

	Train Size	Valid	Test
En-Fr	10.00M	newstest13	newstest15
En-Cs	10.00M	newstest16	newstest18
En-De	4.60M	newstest16	newstest18
En-Fi	4.80M	newstest16	newstest18
En-Lv	1.40M	newsdev17	newstest17
En-Et	0.70M	newsdev18	newstest18
En-Ro	0.50M	newsdev16	newstest16
En-Hi	0.26M	newsdev14	newstest14
En-Tr	0.18M	newstest16	newstest18
En-Gu	0.08M	newsdev19	newstest19

Table 3: The statistics of the training, valid, and test sets on WMT datasets of 10 language pairs.

3.2 Baselines and Evaluation

We compare our method against the following baselines. **Bilingual baseline** is trained on each language pair separately. **One-to-Many** and **Many-to-One** are trained on the En→X and X→En directions respectively. We collect all English sentences (33M) of the bilingual corpora described above and translate them into other languages sentences. We extract alignment pairs (Dyer et al., 2013) across different languages for our method. **One-to-Many + Pseudo** and **Many-to-One + Pseudo** are trained on multilingual data combined with the pseudo data. We average the last 5 checkpoints and employ the beam search strategy with a beam size of 5 for evaluation. The evaluation metric is case-sensitive detokenized sacreBLEU[2] (Post, 2018).

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

[2] BLEU+case.mixed+lang.{src}-{tgt}+numrefs.1+smooth.exp+tok.13a+version.1.4.14

3.3 Training Details

We adopt the `Transformer_big` architecture as the backbone model for all our experiments, which has 6 layers with an embedding size of 1024, a dropout of 0.1, the feed-forward network size of 4096, and 16 attention heads. We train multilingual models with Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$). The learning rate is set as 5e-4 with a warm-up step of 4,000. The models are trained with the label smoothing cross-entropy with a smoothing ratio of 0.1. The batch size is 5,120 tokens and the parameters are updated every 16 iterations to simulate a 128-GPU environment.

4 Results

The results of our model are separately listed in Table 1 and Table 2. Table 1 shows that **One-to-Many** outperforms **bilingual NMT** by +1.8 BLEU points on average. Our method further improves over both **One-to-Many** and **One-to-Many + Pseudo** consistently. Using pseudo and code-switched data brings more improvements to the low-resource languages (Et, Ro, Hi, Tr, and Gu) than high-resource languages (Fr, Cs, De, Fi, and Lv). These results suggest that our model encourages the agreement between different translation directions.

Table 2 reports the results on the X→En test sets. **Many-to-One** outperforms the **bilingual** NMT by +4.2 BLEU points on average. We combine the parallel data with the pseudo data, leading to an improvement of +1.9 BLEU points over **Many-to-One**. Our method further outperforms **Many-to-One + Pseudo** by a large gain of +0.5 BLEU points on average, showing the effectiveness of our agreement-based method and the significance of multilingual agreement.

5 Analysis

Attention Visualization The representations of attention in Figures 3 and 4 are averaged over all 16 heads of the last layer. Figure 3 shows the self-attention weights of a code-switched English sentence, where the source phrase "coordination between law enforcement" is substituted by the German phrase "Koordinierung zwischen Strafverfolgung sbehörden". Similar to the common attention pattern, our model can learn better crosslingual representations in this code-switching case. Figure 4 shows that the cross-attention weights between the input code-switched English sentence and the output German sentence. The words with similar meanings are aligned together between the code-switched input and target output.

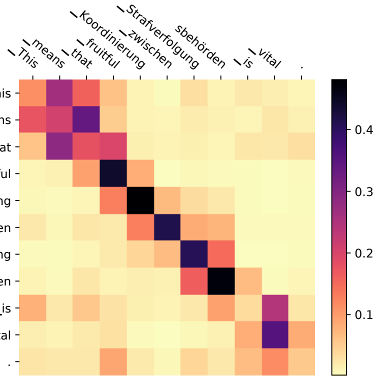

Figure 3: Visualization for the self-attention weights.

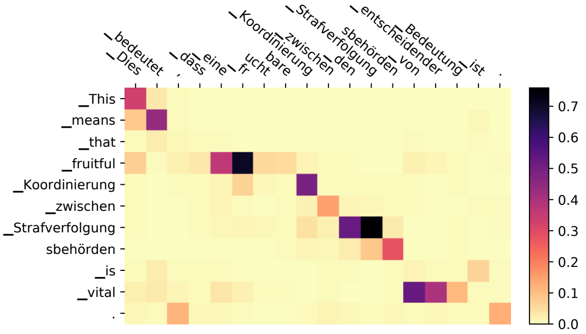

Figure 4: Visualization for the cross-attention weights between the code-switched input and target sentence.

Crosslingual Representation We select 500 parallel sentences across different languages and visualize their sentence vectors of multilingual baseline and our method in Figure 5. The vector of the special language symbol of the source sentence is used as the sentence representation for visualization. Compared to Figure 5(a), different languages become closer and overlap with each other in Figure 5(b), which shows our method aligns representations and minimizes the differences among different languages.

Substitution Strategy We employ both word-level and phrase-level substitution strategies for code-switching. The word-level and phrase-level methods replace some words or spans of the source sentence with other languages. In Table 4, phrase-level substitution works better. Furthermore, we investigate the effect of the substitution ratio of the source words. From Figure 6, the best substitution ratio is 10%. When increasing the ratio to 30%, the performance gets worse, which indicates substitut-

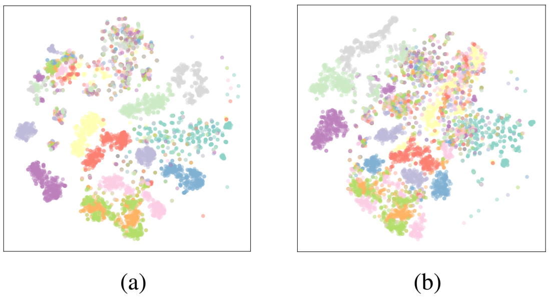

Figure 5: t-SNE (Maaten and Hinton, 2008) visualization of parallel sentences vector space of all languages from the multilingual baseline (a) and our method (b). Each color denotes one language.

X → En	De	Lv	Ro	Tr	Avg
Word-level	42.5	21.5	35.9	26.2	31.5
Phrase-level	42.6	21.8	36.4	26.7	31.9

Table 4: Comparison of BLEU points between the word-level and the phrase-level substitution strategies on X→En directions.

ing too many words may degrade the performance.

As Equation 3 formulates, our method uses both the original corpora and code-switched corpora simultaneously to reduce the effect of the word alignment errors. Besides, `fast_align` (Dyer et al., 2013) is a simple, fast, and effective tool with a lower alignment error rate. Therefore, our method can avoid the disturbance introduced by the word alignment errors as much as possible.

Time Cost of Word Alignment In this work, we try a large pseudo parallel corpus (33M) to train the multilingual corpora. In most scenarios, the size of the parallel corpus is less than 33M and thus consumes less time to generate the alignment pairs. All the alignment pairs are offline generated only once before the training phase. Therefore, the time cost of the word alignment is much smaller than that of the model training.

6 Related Work

Multilingual Machine Translation Previous works (Zoph et al., 2016; Firat et al., 2016b; Johnson et al., 2017) have explored different settings of the multilingual neural machine translation (MNMT). Recent studies show that MNMT (Blackwood et al., 2018; Platanios et al., 2018; Gu et al., 2018) helps improve the performance of the low-resource or zero-shot translation. Some researchers use the sentence pairs to enhance the bilingual neural machine translation (Conneau and Lample, 2019; Song et al., 2019; Yang et al., 2020b).

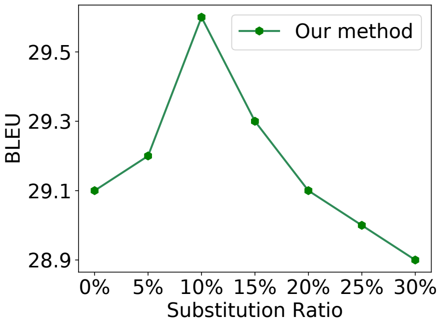

Figure 6: Average results of X→En directions on different substitution ratio settings. Large substitution ratio may degrade the model performance and is even worse than the multilingual baseline.

Agreement-based Learning Many works try to use the agreement-based method (Liang et al., 2007, 2006; Al-Shedivat and Parikh, 2019) to encourage agreement among different translation orders and directions (Liang et al., 2006; Castilho, 2020; Yang et al., 2020a; Cheng et al., 2016; Zhang et al., 2019). Besides, the agreement-based method is also used to minimize the difference between the representation of source and target sentence (Yang et al., 2019). Our method further explores the approach of the multilingual agreement.

7 Conclusion

We propose a novel agreement-based framework to encourage multilingual agreement across different translation directions by the agreement term. Experimental results on the multilingual translation task demonstrate that our method effectively minimizes the gaps among different translation directions and significantly outperforms the multilingual baselines. The analytic experiment about the crosslingual representation shows the effectiveness of our multilingual agreement in minimizing the differences among different languages.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos.U1636211, 61672081, 61370126), the 2020 Tencent WeChat Rhino-Bird Focused Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No.SKLSDE2019ZX-17).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL 2019*, pages 3874–3884.

Maruan Al-Shedivat and Ankur P. Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *NAACL 2019*, pages 1184–1197.

Graeme W. Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *COLING 2018*, pages 3112–3122.

Sheila Castilho. 2020. Document-level machine translation evaluation project: Methodology, effort and inter-annotator agreement. In *EAMT 2020*, pages 455–456.

Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Agreement-based joint training for bidirectional attention-based neural machine translation. In *IJCAI 2016*, pages 2761–2767.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NeurIPS 2019*, pages 7057–7067.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *NAACL 2013*, pages 644–648.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL 2016*, pages 866–875.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *EMNLP 2016*, pages 268–277.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *NAACL 2018*, pages 344–354.

Michael A. Hedderich, David Ifeoluwa Adelani, Dawei Zhu, Jesujoba O. Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. In *EMNLP 2020*, pages 2580–2591.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018*, pages 66–71.

Percy Liang, Dan Klein, and Michael I. Jordan. 2007. Agreement-based learning. In *NIPS 2007*, pages 913–920.

Percy Liang, Benjamin Taskar, and Dan Klein. 2006. Alignment by agreement. In *NAACL 2006*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom M. Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *EMNLP 2018*, pages 425–435.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT 2018*, pages 186–191.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *IJCAI 2020*, pages 3853–3860.

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. *JMLR*, 21:140:1–140:67.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *ACL 2019*, pages 3083–3089.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *NAACL 2019*, pages 449–459.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *ACL 2020*, pages 3525–3535.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS 2017*, pages 5998–6008.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *ICLR 2019*.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *EMNLP 2020*, pages 1022–1034.

Mingming Yang, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Min Zhang, and Tiejun Zhao. 2019. Sentence-level agreement for neural machine translation. In *ACL 2019*, pages 3076–3082.

Mingming Yang, Xing Wang, Min Zhang, and Tiejun Zhao. 2020a. Incorporating phrase-level agreement into neural machine translation. In *NLPCC 2020*, pages 416–428.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020b. CSP: code-switching pre-training for neural machine translation. In *EMNLP 2020*, pages 2624–2636.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *ACL 2020*, pages 1628–1639.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. 2019. Regularizing neural machine translation by target-bidirectional agreement. In *AAAI 2019*, pages 443–450.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *EMNLP 2016*, pages 1568–1575.