# An Empirical Study on Many-to-Many Simultaneous Machine Translation

**Erenay Dayanik***
IMS, University of Stuttgart
erenay.dayanik@ims.uni-stuttgart.de

**Ran Xue**
Alexa AI
ranxue@amazon.com

**Ching-Yun Chang**
Alexa AI
cychang@amazon.com

## Abstract

Simultaneous machine translation (SimulMT) is a challenging task which aims to translate a source sequence to the target language with low latency. Despite significant progress in SimulMT, there has not been much work in the area of multilingual SimulMT where a single model is capable of translating between multiple language pairs. This paper studies SimulMT from a multilingual perspective. Through our experiments, we first compare several language tag strategies, and show that language tag strategies can effectively adapt a unidirectional SimulMT model to translate multiple language arcs. Second, we find that SimulMT models trained on a language family perform better than a global model. Finally, we demonstrate that it is possible to improve the performances of multilingual SimulMT models by transferring embeddings from a pre-trained language model such as multilingual BERT.

## 1 Introduction

Simultaneous machine translation (SimulMT) (Ma et al., 2019, 2020; Arivazhagan et al., 2020; Arthur et al., 2021) is one of the research areas in Natural Language Processing (NLP) that has attracted much attention lately. In contrast to conventional machine translation (MT) systems, SimulMT models can begin translating a source sentence before it is finished, and hence significantly reduce translation latency. However, this makes the translation process more challenging compared to the conventional MT because in order to balance the translation quality and latency, SimulMT needs to have a policy agent to decide when to read the input and when to continue translating.

Previous work on SimulMT proposed various policies which can be grouped into two main classes according to whether the policy rules are predetermined and fixed (fixed policies) (Cho and Esipova, 2016; Dalvi et al., 2018; Ma et al., 2019), or whether they are learned from the data (flexible policies) (Gu et al., 2017; Raffel et al., 2017; Arivazhagan et al., 2019; Ma et al., 2020). While fixed policies follow some simple and easy-to-implement rules to determine the next action, they can be too aggressive or too conservative at times. Flexible policies, on the other hand, allow the translation latency to be adjusted during the decoding time, but they require more sophisticated algorithms and longer training time.

To the best of the authors' knowledge, all of the previous studies in SimulMT have focused on one-to-one simultaneous machine translation, except Arthur et al. (2021) which investigated a many-to-one model. One aim of the paper is to fill the gap in the current literature by exploring the SimulMT task from a many-to-many perspective. We utilized the hard monotonic multihead attention (MMA-Hard) model proposed by Ma et al. (2020), a state-of-the-art transformer-based SimulMT system with flexible policy based on monotonic attention. In this paper, we empirically answered the following research questions:

- RQ1: Can we build an effective multi-source multi-target SimulMT model using the language tag (LT) strategy (Johnson et al., 2017) which simply prefixes an artificial token to a sentence to provide the language signal, and what is the best way to configure them? — Yes, using target LT alone at the decoder yields the best results, with an average of +3.85 BLEU points and +1.3 Average Lagging (AL) compared with the one-to-many SimulMT models from Arthur et al. (2021).
- RQ2: Does a many-to-many SimulMT model trained using languages in the same language

*This work has been done during the author's internship at Alexa AI, Amazon.

family perform better than a big multilingual model mixing different language families? — Yes. Our experiments on Germanic and Romance language families show that language family specific models achieve better translation quality, albeit with a slightly longer latency.

- RQ3: Can we improve translation quality and latency of a SimulMT model by utilizing embeddings from a pre-trained language model? — To some extent, yes. We observe that transferring embeddings from a pre-trained multilingual BERT (Devlin et al., 2019) improves both of the translation quality (up to +0.7 BLEU) and latency (up to -0.4 AL reduction) over a model trained from scratch.

## 2 Related Work

The previous work on SimulMT can be grouped under two categories based on the type of the policy that is used for deciding when to read another source token or when to make a generation. The models in the first group use fixed policies. Due to its general applicability and high effectiveness, the "wait-k" policy (Dalvi et al., 2018; Ma et al., 2019) has become one of the most widely used fixed policies in SimulMT. Wait-k is a strategy that forces models to always wait for exactly k tokens before beginning to translate, and then alternates between reading and writing until the source sentence ends. Dalvi et al. (2018) train a conventional full-sentence translation model and use the wait-k policy only during testing. Ma et al. (2019), on the other hand, integrate it directly into the model training to address the problem of mismatching between training and testing in Dalvi et al. (2018).

The second group of work uses adaptive policies. Among these, Cho and Esipova (2016) is the first study experimenting with adaptive policy. They use greedy heuristic decoding to modify the decoder of a conventional MT model such that at each iteration the decoder decides between writing and reading tokens. Later, Gu et al. (2017) utilize an adaptive policy based on reinforcement learning (RL) with the aim of learning a READ/WRITE policy, and Alinejad et al. (2018) extend this approach by adding the PREDICT action which predicts the next source word. One drawback of RL based methods is that they have stability and robustness issues due to the sparse reward signals. To overcome this, Arthur et al. (2021) propose using limitation learning with READ/WRITE sequences generated from parallel text. In addition to heuristic and RL based methods, there is also a line of work that focuses on joint learning of MT models and adaptive policy. Arivazhagan et al. (2019) present Monotonic Infinite Lookback (MILk) attention which allowed the building of the first simultaneous MT system to learn an adaptive schedule jointly with an RNN-based NMT model that attends over all source tokens read thus far. Following on from Arivazhagan et al. (2019), Ma et al. (2020) proposed an extension of the MILk, named Multi-head Monotonic Attention (MMA), which can be integrated into transformer-based models.

| | EN | DE | IT | NL | RO |
|---|---|---|---|---|---|
| EN | - | 206112 | 231619 | 237240 | 220538 |
| DE | 206112 | - | 205465 | 213628 | 201455 |
| IT | 231619 | 205465 | - | 233415 | 217551 |
| NL | 237240 | 213628 | 233415 | - | 206920 |
| RO | 220538 | 201455 | 217551 | 206920 | - |

Table 1: Number of sentences in the training set.

| | EN | DE | IT | NL | RO |
|---|---|---|---|---|---|
| EN | - | 1138 | 1147 | 1181 | 1129 |
| DE | 1138 | - | 1133 | 1174 | 1121 |
| IT | 1147 | 1133 | - | 1183 | 1127 |
| NL | 1181 | 1174 | 1183 | - | 1123 |
| RO | 1129 | 1121 | 1127 | 1123 | - |

Table 2: Number of sentences in the test set.

## 3 Experimental Setup

In this section we firstly describe the datasets and evaluation metrics used in our experiments, followed by a brief introduction of the MMA-Hard SimulMT model (Ma et al., 2020) and baselines used for benchmarking model performances.

**Datasets.** We use IWSLT 2017 datasets (Cettolo et al., 2012) to perform many-to-many SimulMT experiments. The datasets consist of transcriptions of TED talks in five languages, English (EN), German (DE), Italian (IT), Dutch (NL) and Romanian (RO), and are split into predefined training, validation and test sets. We use official[1] training, development, and test splits provided in all these datasets, and the dataset statistics are shown in Table 1 and

[1] https://sites.google.com/site/iwsltevaluation2017/TED-tasks

Table 2.

For all the data, we firstly apply tokenization and punctuation normalization using the Moses tokenizer (Koehn et al., 2007). We then learn byte pair encoding (BPE) (Sennrich et al., 2016) jointly on the source and target text, and construct a shared vocabulary of 16K BPE tokens.

**Evaluation Metrics.** SimulMT models are evaluated by the translation quality and latency, and we use the SimulEval toolkit[2] to compute the results. We use BLEU (Papineni et al., 2002) to measure the translation quality on detokenized text, the higher the better. For latency, we report Average Lagging (AL) (Ma et al., 2019) that measures the average rate by which the MT system lags behind an perfectly simultaneous translator, the lower the better.

**Model.** The MMA-Hard SimulMT model (Ma et al., 2020) is a transformer-based model with a flexible policy agent making use of monotonic multihead attention. Monotonic attention is an alternative attention mechanism to softmax-based attention (Raffel et al., 2017), and it does not require a translation model to observe the entire input sequence before the model starts producing the output sequence. Monotonic attention is integrated into a transformer-based MT model by replacing each encoder-decoder attention head in the decoder with a monotonic attention head. Since there are multiple encoder-decoder attention heads, and each of them is replaced with an independent monotonic attention head, a MMA-Hard model can focus on multiple positions in an input sequence at each time step and learn complex alignments between the input and output sequences even in the absence of softmax attention. We direct readers to Ma et al. (2020) for more details regarding the MMA-Hard model.

**Baseline.** We use two baseline models in our experiments: (1) A publicly available version[3] of the M2M-100 model (Fan et al., 2021), which is a transformer-based multilingual encoder-decoder model for conventional machine translation. We use this model to estimate the upper bound of the translation quality. (2) The many-to-one SimulMT models from Arthur et al. (2021). We extract their best reported translation quality results for each translation direction.

**Training Details.** We use the fairseq[4] library. Each model has 96,296,960 trainable parameters consisting of 6 transformer layers in the encoder and decoder; each transformer layer has 4 self-attention heads and 4 encoder-decoder attention heads. The size of the hidden states and the feed-forward layer are 512 and 1024, respectively. We set dropout to 0.3 and the head divergence loss coefficient to 0.1; Adam optimizer and inverse square root scheduler are used with initial learning rate of 5e-4. We set warm-up phase to 4000 and the training batch size to a maximum of 3584, and update parameters every 8 batches. The maximum number of iterations is set to 50000, and early stopping strategy is applied when the perplexity on the validation set starts to increase. Each model is trained on 4 NVIDIA V100 GPUs for around 48 hours.

## 4 Results and Analyses

In this section, we present our experimental results and analyses for the three research questions listed in the introduction section. For all the result tables, rows and columns denote source and target languages, respectively; AL and BL are Average Lagging and BLEU score, respectively.

**RQ1: Can we build effective many-to-many SimulMT models using the LT strategy, and what is the best way to configure them?** LT is a simple but effective technique which allows us to train a single translation model for multiple translation directions without making changes to the model architecture, and it can be implemented in various ways. Johnson et al. (2017) prefix a target language tag (TLT) to each source sentence to indicate the target language the model should translate to, while Fan et al. (2021) add a source language tag (SLT) in the encoder indicating the source language, and a target language tag in the decoder indicating the target language. Apart from these two configurations, we also explore another option where a target language tag is forced at the beginning of the decoder, and rely on the model to learn the source language automatically.

Following Arthur et al. (2021), we group the parallel datasets into Germanic language family: Dutch (NL), English (EN) and German (DE), and Romance language family: Italian (IT) and Roma-

[2] https://github.com/facebookresearch/SimulEval

[3] https://huggingface.co/docs/transformers/model_doc/m2m_100

[4] https://github.com/pytorch/fairseq

| | EN | | DE | | NL | |
|---|---|---|---|---|---|---|
| | AL | BL. | AL | BL. | AL | BL. |
| EN | - | - | **6.8** | **21.5** | 7.6 | 26.9 |
| DE | 6.2 | 24.6 | - | - | 7.0 | 18.7 |
| NL | 5.7 | 29.8 | 8.2 | 18.4 | - | - |

(a) [TLT] src → tgt

| | EN | | DE | | NL | |
|---|---|---|---|---|---|---|
| | AL | BL. | AL | BL. | AL | BL. |
| EN | - | - | 4.5 | 21.2 | 5.0 | 26.6 |
| DE | 4.5 | 24.9 | - | - | **5.2** | **19.1** |
| NL | 4.1 | 29.6 | 4.4 | 18.1 | - | - |

(b) [SLT] src → [TLT] tgt

| | EN | | DE | | NL | |
|---|---|---|---|---|---|---|
| | AL | BL. | AL | BL. | AL | BL. |
| EN | - | - | 4.3 | 21.2 | **4.8** | **27.4** |
| DE | **4.4** | **25.0** | - | - | 5.0 | 19.0 |
| NL | **4.0** | **29.8** | **4.1** | **18.9** | - | - |

(c) src → [TLT] tgt

Table 3: LT strategy results for Germanic language family.

| | EN | | IT | | RO | |
|---|---|---|---|---|---|---|
| | AL | BL. | AL | BL. | AL | BL. |
| EN | - | - | 6.1 | 29.8 | 6.2 | 23.0 |
| IT | **5.6** | **34.1** | - | - | 6.4 | 20.1 |
| RO | **5.6** | **28.8** | 6.0 | 21.8 | - | - |

(a) [TLT] src → tgt

| | EN | | IT | | RO | |
|---|---|---|---|---|---|---|
| | AL | BL. | AL | BL. | AL | BL. |
| EN | - | - | 4.5 | 30.2 | 4.3 | 22.9 |
| IT | 4.4 | 33.8 | - | - | 4.3 | 20.2 |
| RO | 4.5 | 28.5 | 4.6 | 21.9 | - | - |

(b) [SLT] src → [TLT] tgt

| | EN | | IT | | RO | |
|---|---|---|---|---|---|---|
| | AL | BL. | AL | BL. | AL | BL. |
| EN | - | - | **3.8** | **31.2** | **3.6** | **24.2** |
| IT | 3.7 | 33.8 | - | - | **3.6** | **21.4** |
| RO | 3.7 | 28.4 | **3.8** | **22.2** | - | - |

(c) src → [TLT] tgt

Table 4: LT strategy results for Romance language family.

| Source language: DE | Target language: EN |
|---|---|
| ...wurden die Forderungen der Regierung an Risen fallen gelassen ... | ...letterlijk the government's *Forderungen* dropped gek klinkt to risk ... |

Table 5: An example where the predicted hypothesis includes words in multiple languages. Underlined and Italic words are Dutch and German, respectively.

| | Arthur et al. (2021) | | | | M2M-100 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $EN_{GLOB}$ | | $EN_{LANG}$ | | EN | DE | IT | NL | RO |
| | AL | BLEU | AL | BLEU | BLEU | BLEU | BLEU | BLEU | BLEU |
| EN | - | - | - | - | - | 20.3 | 28.5 | 24.0 | 21.7 |
| DE | 2.6 | 23.0 | 2.6 | 23.9 | 24.4 | - | 17.6 | 19.0 | 15.8 |
| IT | 2.6 | 25.1 | 2.8 | 25.2 | 33.6 | 17.1 | - | 19.3 | 19.4 |
| NL | 2.4 | 27.4 | 2.5 | 28.4 | 28.9 | 19.7 | 19.9 | - | 17.0 |
| RO | 2.5 | 24.6 | 2.7 | 24.1 | 28.3 | 16.6 | 22.0 | 17.8 | - |

Table 6: The baseline results from Arthur et al. (2021) and the M2M-100 model. $EN_{GLOB}$ and $EN_{LANG}$ are the global and language family specific models, respectively.

nian (RO) plus English (EN)[5], and train MMA-Hard SimulMT models with different LT strategies. Table 3 and Table 4 show the results for Germanic and Romance models respectively. From the results we can see that 8 out of the 12 translation directions achieve the best AL and BLEU performances when a TLT is added to the decoder. In addition, we observe that when prefixing TLT to a source sentence, the model may forget the target language it should translate to, as shown in Table 5. This may be because of the characteristic of monotonic attention whereby an attention head is not allowed to attend to previous states.

[5]Although English does not belong to the Romance language family, English has been highly influenced by Romance languages. We add the English parallel datasets to the Romance language family to increase the translation directions.

We compare the best performing many-to-many SimulMT models that use only TLT at the decoder, with the best results from the language family specific models reported by Arthur et al. (2021) listed in Table 6. Note that, unlike ours, the models proposed in Arthur et al. (2021) are multi-source single-target SimulMT models. We observe that our many-to-many models outperform the many-to-one models in all the translation directions, with an average of +3.85 BLEU points including a significant +8.6 BLEU points for IT → EN, and an average of +1.3 AL.

From the data points, we are confident that using the LT strategy can effectively adapt a one-to-one SimulMT model to perform many-to-many translations. In addition, the experiments on MMA-Hard

| | EN | | DE | | IT | | NL | | RO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AL | BLEU | AL | BLEU | AL | BLEU | AL | BLEU | AL. | BLEU |
| EN | - | - | 3.4 | 20.5 | 3.6 | 29.1 | 4.0. | 26.1 | 3.4 | 22.3 |
| DE | 3.8 | 23.6 | - | - | 3.9 | 16.2 | 4.1 | 18.0 | 3.7 | 14.8 |
| IT | 3.3 | 33.1 | 3.4 | 16.0 | - | - | 3.8 | 19.4 | 3.3 | 19.4 |
| NL | 3.5 | 28.8 | 3.5 | 18.2 | 3.7 | 19.0 | - | - | 3.5 | 16.2 |
| RO | 3.4 | 27.9 | 3.5 | 15.6 | 3.6 | 21.2 | 3.8 | 17.8 | - | - |

Table 7: A many-to-many SimulMT model trained on the 20 translation directions available in IWSLT 2017.

| Source sentence | Romance lang family model | Global model |
|---|---|---|
| È | – | – |
| È difficile | – | It's difficult |
| È difficile che | – | It's difficult that |
| È difficile che la | – | It's difficult that |
| È difficile che la gente | It's hard for people | It's difficult that |
| È difficile che la gente lasci | It's hard for people to leave | It's difficult that people are leaving |
| È difficile che la gente lasci cattive | It's hard for people to leave | It's difficult that people are leaving |
| È difficile che la gente lasci cattive recensioni. | It's hard for people to leave bad reviews. | It's difficult that people are leaving bad reviews. |

Table 8: An example of generating simultaneous translations using the Romance language family model (Column-2) and the global model (Column-3).

SimulMT models show that placing a target language prefix token at the decoder achieves on-par or better performances than conventional MT as well as SimulMT models.

**RQ2: Are language family specific many-to-many models better than a single global model?** To address this question, we train a global many-to-many MMA-Hard SimulMT model using all the IWSLT 2017 parallel data with TLT at the decoder. The model performances are presented in Table 7. Comparing the results of the Germanic language family model in Table 3(c) against the global model, we see that the BLEU scores of the language family specific model outperform the global model for all the translation directions, with an average of +1.0 BLEU points. There is, however, a trade-off in AL of +0.7. Similar comparison outcomes are also observed for the Roman language family model, with an average of +1.4 BLEU points improvement over the global model, and +0.3 AL.

To better observe the trade-off, we sample some data to qualitatively compare the language family specific models with the global model. Table 8 presents an Italian source sentence "È difficile che la gente lasci cattive recensioni." (the English reference is "It's hard for people to leave bad reviews.") alongside its translations output by the Romance language family model and the global model. Each row in the table is a time step where one additional source word is incrementally available to a SimulMT model. We can see that while the Romance language family model starts outputting a translation slightly later than the global model (and hence has a higher latency), it yields a better translation.

We also compare the language family specific many-to-many models with the M2M-100 model, a multilingual conventional MT model, the results of which are shown in Table 6.[6] Our models outperform the M2M-100 model in all of the translation directions except NL $\rightarrow$ DE, by an average of

[6]As the M2M-100 is a regular MT model, no AL is given.

| | Random initialization | | | | | | Pre-trained embeddings | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN | | DE | | NL | | EN | | DE | | NL | |
| | AL | BLUE | AL | BLUE | AL | BLUE | AL | BLUE | AL | BLUE | AL | BLUE |
| EN | - | - | 3.6 | 19.2 | **4.3** | **25.1** | - | - | **3.3** | **19.9** | 4.2 | 25.0 |
| DE | 3.9 | 23.1 | - | - | 4.7 | 17.2 | **3.5** | **23.6** | - | - | **4.7** | **17.7** |
| NL | 3.7 | 27.7 | 3.9 | 16.8 | - | - | **3.4** | **28.2** | **3.8** | **17.3** | - | - |

Table 9: Germanic language family models using random initialization and pre-trained embeddings.

| | Random initialization | | | | | | Pre-trained embeddings | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN | | IT | | RO | | EN | | IT | | RO | |
| | AL | BLUE | AL | BLUE | AL | BLUE | AL | BLUE | AL | BLUE | AL | BLUE |
| EN | - | - | 3.3 | 27.9 | 3.1 | 19.9 | - | - | 3.3 | 27.9 | **3.0** | **20.2** |
| IT | **3.2** | 30.7 | - | - | 3.1 | 17.2 | 3.3 | **31.1** | - | - | 3.2 | **17.4** |
| RO | 3.4 | 24.5 | 3.7 | 19.5 | - | - | 3.4 | **25.4** | **3.6** | **20.1** | - | - |

Table 10: Romance language family models using random initialization and pre-trained embeddings.

+1.1 BLEU points. This is particularly encouraging, considering that SimulMT is more challenging than the full sentence translation due to the lack of complete source-side information during decoding. In addition, we find that the performance of our global model achieves better translation quality with an average of +1.5 BLEU points compared with the global model of Arthur et al. (2021), and is on par with the M2M-100 model.

With these experimental results, we suggest that a SimulMT model trained with languages from the same language family has a better translation quality compared to a global model trained using all languages, albeit with a slightly longer latency.

**RQ3: Can we improve SimulMT performance by utilizing pre-trained token embeddings?** Lastly, we investigate whether it is possible to improve the performance of our language family specific multilingual SimulMT models further using pre-trained multilingual embeddings, which has shown to be effective for many NLP tasks (Lample et al., 2016; Qi et al., 2018). To do so, we replace the embedding layer of our MMA-Hard SimulMT model with that of the pre-trained multilingual BERT (mBERT) model (Devlin et al., 2019). Since mBERT uses WordPiece (Schuster and Nakajima, 2012), a different tokenization algorithm than our models, and also its embedding size is different than our models, we can't perform a fair comparison using the models evaluated above. Therefore, we train two new language family specific models (for German and Romance language families) using the same tokenizer and embedding size as the mBERT model, and use them for comparison in this part. Tables 9 and 10 show results of the models with random initialization and pre-trained mBERT embeddings for Germanic and Romance language families respectively. For Germanic language family model, we observe that for all of the translation arcs except EN→NL, using pre-trained embeddings has positive effect on both the translation quality and latency. We see a similar pattern for the Romance language family, with a larger difference between target-source language pairs: While pretrained embeddings almost always improve the translation quality, its affect on the latency is very limited.

## 5 Conclusions

In this paper, we present an empirical study discussing the task of simultaneous machine translation from a many-to-many perspective. Our results suggest that: (1) Using the language tag strategy, we can easily adapt a state-of-the-art one-to-one SimulMT model to a many-to-many SimulMT setting. In addition, placing the language tag only at the target side achieves the best translation performance. (2) Language-family focused many-to-many SimulMT models perform better than a single global model trained on all languages. (3) We can improve a SimulMT model's performance, to some extent, by utilizing pre-trained token embeddings.

For the future work, we plan to apply the tech-

nique to more language families as well as different domain datasets, and study the quality/latency trade-off for different text properties. Furthermore, we would like to experiment with other transfer learning strategies and auxiliary learning to advance the area of many-to-many SimulMT.

## References


Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. Prediction improves simultaneous neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3022–3027, Brussels, Belgium. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Philip Arthur, Dongwon Ryu, and Gholamreza Haffari. 2021. Multilingual simultaneous neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4758–4766, Online. Association for Computational Linguistics.

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, Trento, Italy. European Association for Machine Translation.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in 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 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. 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, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020. Monotonic multihead attention.

In *International Conference on Learning Representations*.

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.

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.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2837–2846. JMLR.org.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

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.