Multilingual KERMIT: It's Not Easy Being Generative

Harris Chan[*]
University of Toronto
Vector Institute
hchan@cs.toronto.edu

Jamie Kiros, William Chan
Google Research, Brain Team
{kiros,williamchan}@google.com

Abstract

We present multilingual KERMIT, a generative model over multiple languages. Multilingual KERMIT models the joint distribution over multiple languages, and all its decompositions using a single neural network. KERMIT can be trained by feeding it N way parallel-data, bilingual data, or monolingual data. At inference, KERMIT can generate translations for a particular target language, or up to $N-1$ languages in parallel. It can also unconditionally generate sentences in multiple languages. Our experiments on the Multi30K dataset containing English, French, Czech, and German languages suggest that the multitask training with the joint objective leads to improvements in bilingual translations. We provide a quantitative analysis of the quality-diversity trade-offs for different variants of KERMIT for conditional generation, and a measurement of self-consistency during unconditional generation. We provide qualitative examples for parallel greedy decoding across languages and sampling from the joint distribution of the 4 languages.

1 Introduction

Traditional autoregressive approach [14, 3] models the conditional probability $p(y \mid x)$ of an output sequence y conditioned on the input sequence x with a left-to-right factorization. The model decomposes $p(y \mid x)$ as predicting one output token at time, conditioning on the previously generated output tokens $y_{<t}$ and the input sequence x:

$$p(y \mid x) = \prod_t p(y_t \mid x, y_{<t}). \tag{1}$$

Recent encoder-decoder models with attention such as Transformer [15] have been successful in various domains, including machine translation.

Instead of assuming a fixed left-to-right decomposition, recent insertion-based conditional modeling frameworks [13, 16, 7] consider arbitrary factorization of the output sequence by using insertion operation, which predicts both (1) content token $c \in \mathcal{C}$ from the vocabulary, and (2) location l insert, relative to the current partial output $\hat{y}_t$:

$$p(c, l | x, \hat{y}_t) = \text{InsertionTransformer}(x, \hat{y}_t) \tag{2}$$

Subsequent work, KERMIT (Kontextuell Encoder Representations Made by Insertion Transformations) [2], simplified the Insertion transformer model to be purely decoder based with no causal masking, by concatenating the original input and output sequence as a single sequence.

In this work, we investigate applying KERMIT to *model the joint distribution over multiple sequences*. Specifically, we train a multilingual KERMIT on the Multi30K [6] machine translation

[*] Work done during an internship at Google Research, Brain Team

Workshop on Perception as Generative Reasoning, NeurIPS 2019, Vancouver, Canada.

task, consisting of four languages: English (EN), French (FR), Czech (CS), and German (DE). One advantage of multilingual KERMIT is during inference, we can generate translation for a single target language, or generate translations for $N-1$ languages in parallel in logarithmic time in the token length per language. Our preliminary experiments on the Multi30K dataset that the multitask training with the joint objective leads to improvements in bilingual translations. We illustrate qualitative examples for parallel greedy decoding across languages and sampling from the joint distribution of the 4 languages.

To summarize, our contributions in this work are:

- Extending KERMIT beyond a pair of sequences, specifically to multilingual sequences.
- Analyzing of the quality-diversity trade off for *conditional* generation using variants of KERMITs, evaluated with traditional BLEU for quality and Self-BLEU for diversity.
- Demonstrating *unconditional* multilingual generation using the joint KERMIT model, investigating its self-consistency quantitatively via pseudo-targets BLEU, with qualitative sample comparison to bilingual model baseline sampled in a chain.

2 Background

We briefly reintroduce the insertion transformer framework for training and inference. This section is taken abridgedly from the KERMIT paper, section 3 [2]. We provide it for the reader's convenience. Unlike with the left-to-right autoregressive approach, the exact computation of the log-likelihood (Equation 3) of a sequence x is not possible with insertion models due to the intractable marginalization over the generation order z, where S_n denotes the set of all possible permutations on n elements. Instead, KERMIT optimizes a lower bound the log-likelihood via Jensen's inequality:

$$\log p(x) = \log \sum_{z \in S_n} p(z) p(x \mid z) \tag{3}$$

$$\geq \sum_{z \in S_n} p(z) \log p(x \mid z) \quad =: \mathcal{L}(x) \tag{4}$$

The $p(x \mid z)$ term can be expanded as a product of probability of insertions (c_i^z, l_i^z) conditioned on the partial output $x_{1:i-1}^{z,i-1}$ at time i according to the permutation order z. The loss term can be simplified by changing the summation and decomposing the permutation, leading to:

$$\begin{aligned}\mathcal{L}(x) &= \sum_{z \in S_n} p(z) \log \prod_{i=1}^{n} p((c_i^z, l_i^z) \mid x_{1:i-1}^{z,i-1}) \\ &= \sum_{i=1}^{n} \sum_{z_{1:i-1}} p(z_{1:i-1}) \sum_{z_i} p(z_i \mid z_{1:i-1}) \log p((c_i^z, l_i^z) \mid x_{1:i-1}^{z,i-1})\end{aligned}$$

During inference, the model can perform: (1) autoregressive greedy decoding with 1 insertion at a time: $(\hat{c}, \hat{l}) = \operatorname{argmax}_{c,l} p(c, l|\hat{x}_t)$, or (2) partially autoregressive parallel decoding by inserting at all non-finished slots simultaneously: $\hat{c}_l = \operatorname{argmax}_c p(c \mid l, \hat{x}_t)$. For the latter, the Insertion Transformer [13] has shown that using a binary tree prior for $p(z)$ led to $\approx \log_2 n$ iterations for n token generation.

3 Multilingual KERMIT

We extended the work of [2] to investigate applying the KERMIT objective on tasks with more than 2 sequences, in order to learn the joint distribution $p(L_1, \ldots, L_N)$ over N sequences. These sequences can be for example sentences in different languages, i.e. $p(EN, FR, CS, DE)$.

4 Experiments

We experiment on a multilingual dataset to demonstrate that we can learn a multilingual KERMIT, which in several cases outperforms a bilingual only KERMIT. We provide some qualitative examples of sampling from the joint distribution of languages and parallel decoding across several languages.

4.1 Settings

We experiment on the Multi30k [6, 5, 1], a multilingual dataset which consists of 29000 parallel training sentences in English (EN), French (FR), Czech (CS), and German (DE) describing an image. We implement our model as a base Transformer decoder, without any causal masking (i.e. dense attention), with 6 hidden layers and 1024 dimensional hidden representation. We concatenate all 4 language raw text training examples and use SentencePiece [11] to learn a subword unigram [10] tokenizer with a shared 32K vocabulary size. We follow a similar training set up to BERT [4], using Adam [9] optimizer with learning rate of 1e-4, warmup over the first 10% of the total training iterations varying between 10k to 50k iterations. We train 3 different variants of KERMIT by altering the training data (i.e. identical architecture size):

1. *Bilingual* (e.g. EN → FR), a uni-directional model for a specific language pair
2. *Multi-target* (Any 1 → Rest), where given a single source language sentence in any one of the languages, the model is tasked to predict the translation for the remaining languages
3. *Joint*, which is trained to predict slot tokens for all languages, given partial (or none) sentences in each of the languages.

Figure 5 in Appendix A.2 illustrates the subset of data being used to train each of the five models, and the possible decoding inference modes: a single target language (top right), or multiple target languages in parallel (bottom right).

4.2 Conditional Bilingual Generation: Quality-Diversity Trade-off

We first evaluated the models on *conditional* generation task by sampling bilingual translations (1 source, 1 target language) for each of the 12 language pair directions. We used the Gumbel-Max trick [8] for sampling the token and location $(c, l) \sim p(c, l|x, \hat{y})$ from the partial canvas at each iteration, sampling 100 hypothesis translations per source sentence, at softmax temperature $\tau = 0.1, 0.5, 1.0$. At each temperature and model, we computed the *quality* of the generated samples by computing the BLEU [12] score between the reference translation and the samples, and the *diversity* by computing the pairwise BLEU between the 100 samples per source, also known as Self-BLEU [17]. Lower Self-BLEU indicates the higher the diversity as there is less overlap between the samples.

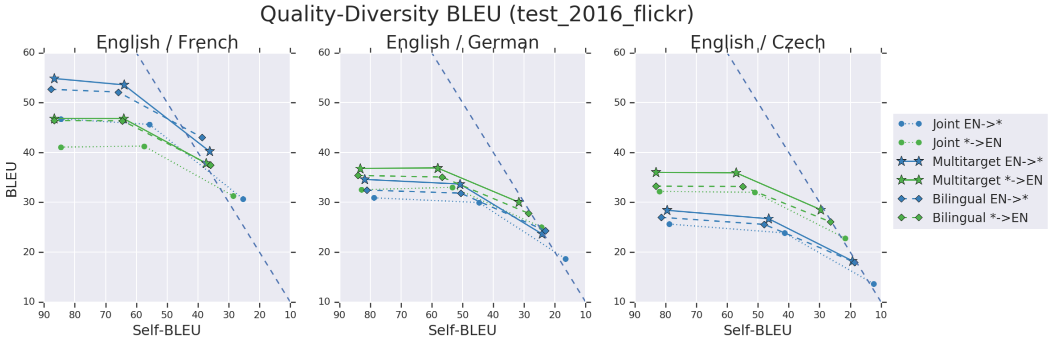

Figure 1: Quality-Diversity BLEU curve for several KERMIT models (bilingual, multitarget, joint) on the Multi30k `text_2016_flickr` test set. Dotted diagonal line signifies BLEU equals Self-BLEU. Points indicate different temperatures, from 0.1 (low diversity, left in graph) to 1.0 (high diversity, right in graph)

Figure 1 illustrates the Quality-Diversity trade-off for the three models for different translation pairs involving English as one of the language. The top right portion of the graph is the ideal area. We observed that the Multitarget model outperformed the Bilingual model at lower temperature (both higher quality and diversity), and at higher temperature slightly above or below in quality but still higher diversity. Note that only one single Multitarget model was used for all language pair at inference time, while each bilingual model was different for each language pair curve. Therefore, a single Multitarget KERMIT model could outperform specialized bilingual KERMIT models.

Model	Language	Generated Sentences
Joint	English	A young man in a blue jacket walking up a mountain.
	French	Un jeune homme en veste bleue descendant une paroi rocheuse en horu.
	German	Ein junger Mann in einer blauen Jacke klettert eine Felswand hoch.
	Czech	Mladý muž v modré bundě stoupá po horách.
		≈"*Young men in blue jackets ascend and climb mountains.*" ✓
Biling.	English	Two small white dogs are holding the duck in a fenced yard.
	French	Deux petits chiens blancs tenant un canard dans une cour clôturée.
	German	Zwei kleine weiße Hunde halten eine gelbe Ente in einem eingezäunten Hof.
	Czech	Dva malí chlapci drží žlutou panou venku u žlutého oploceném nádvoří.
		≈"*Two little boys holding a yellow gentleman outside by a yellow fenced courtyard.*" ✗

Table 1: Example unconditional text generation samples from the Joint (top) and chain of Bilingual model (bottom). Note that the Joint model generates one long sequence and we split them into the resulting four sentences in each language here, while Bilingual generate a complete sentence in each language conditioned on previous sentence.

4.3 Unconditional Multilingual Generation

We then evaluated the models on *unconditional* multilingual generation task, to generate a sentence each in all 4 languages such that they correspond to each other. For the Joint model, we sampled one (token, location) at each iteration starting from an empty canvas, allowing the model to insert a token in any language, until all slots were marked as completed. For the Bilingual model, we trained a separate English language model, then first sampled a complete English sentence. We then sampled a French sentence conditioned on the English sentence, followed by German conditioned on the generated French, and finally Czech conditioned on German (i.e., EN → FR → DE → CS). For each pair of language direction, we computed a *pseudo* target by using a separately trained (on Multi30k) vanilla Transformer [15] and performed beam search (size 5) to translate the source language sample. Figure 2 illustrates the pseudo target BLEU score for different source-target language pairs. We observed that the Bilingual model with a phone booth game-style sampling exhibits worse performance than the Joint model, especially towards the upper right and lower left corner where there was a longer hop (i.e. EN → CS) for the Bilingual model. Table 1 illustrates some example samples from our Joint versus Bilingual model. See 3 in Appendix A.4 for the Joint model's sampling process for that particular example.

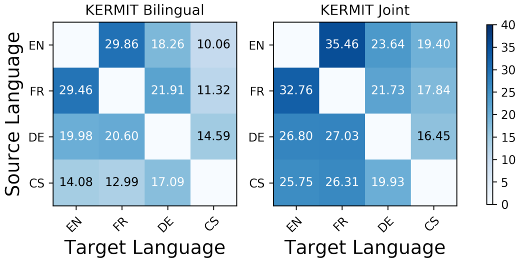

Figure 2: Unconditional multilingual generation Pseudo-Target BLEU for self-consistency when generating sentences in multiple languages.

4.4 Parallel Greedy Decoding: Parallel in Target Languages

Similarly to KERMIT, the Multilingual KERMIT can also perform parallel greedy decoding that is also *parallel in number of target languages*. We illustrate this process in Table 2 in Appendix A.3. By starting with K initial slots for K target languages, Multilingual KERMIT can decode K target languages that has n tokens per language in $\mathcal{O}(\log n)$, i.e. constant in number of target languages.

5 Conclusion and Future Work

We have demonstrated that a multilingual KERMIT can learn a joint distribution over more than two sequences, as shown in our Multi30K experiments. In addition to improvements in bilingual translation via multi-task training in the multi-target models, multilingual KERMIT also allows for efficient inference of multiple target languages in parallel using a single model. For future work, we want to demonstrate KERMIT's generative modeling ability on larger multilingual datasets, and tasks requiring modeling of multiple sequences such as long-form question answering.

References

[1] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, 2018.

[2] William Chan, Nikita Kitaev, Kelvin Guu, Mitchell Stern, and Jakob Uszkoreit. KERMIT: Generative Insertion-Based Modeling for Sequences, 2019.

[3] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, 2014.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

[5] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[6] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

[7] Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based Decoding with Automatically Inferred Generation Order. In *arXiv*, 2019.

[8] Emil Julius Gumbel. Statistical theory of extreme values and some practical applications. *NBS Applied Mathematics Series*, 33, 1954.

[9] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[10] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018.

[11] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

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

[13] Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. Insertion Transformer: Flexible Sequence Generation via Insertion Operations. In *ICML*, 2019.

[14] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS*, 2017.

[16] Sean Welleck, Kiante Brantley, Hal Daume, and Kyunghyun Cho. Non-Monotonic Sequential Text Generation. In *ICML*, 2019.

[17] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100. ACM, 2018.

A Appendices

A.1 Additional Quality-Diversity Curves For Conditional Generation

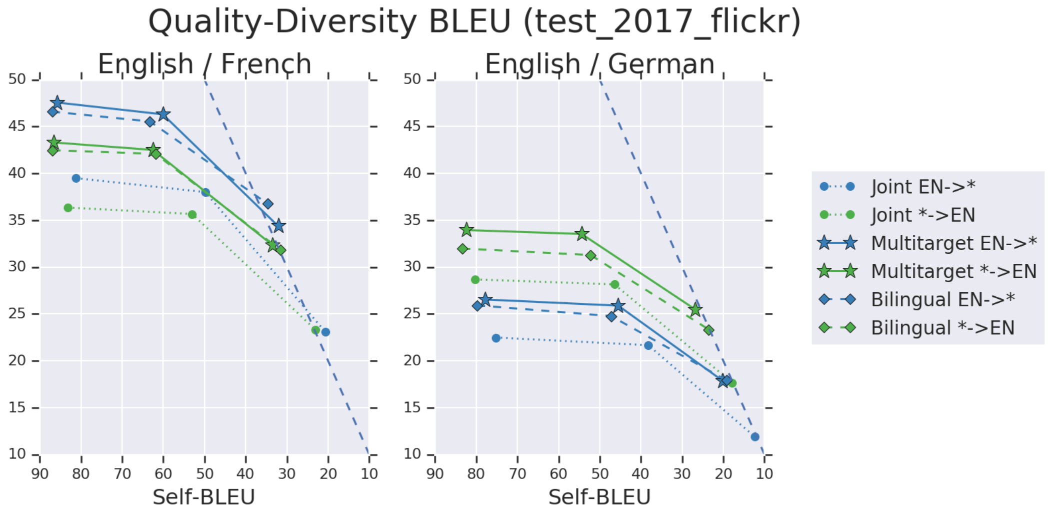

Figure 3: Quality-Diversity BLEU curve for several KERMIT models (bilingual, multitarget, joint) on the Multi30k text_2017_flickr test set. Dotted diagonal line signifies BLEU equals Self-BLEU. Points indicate different temperatures, from 0.1 (low diversity, left in graph) to 1.0 (high diversity, right in graph)

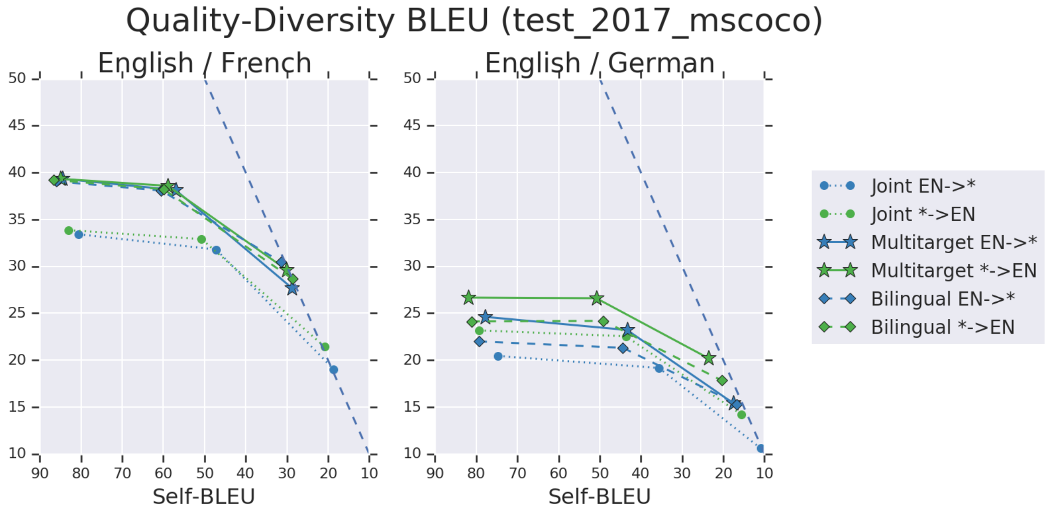

Figure 4: Quality-Diversity BLEU curve for several KERMIT models (bilingual, multitarget, joint) on the Multi30k text_2017_mscoco test set. Dotted diagonal line signifies BLEU equals Self-BLEU. Points indicate different temperatures, from 0.1 (low diversity, left in graph) to 1.0 (high diversity, right in graph)

A.2 Training Data Diagram

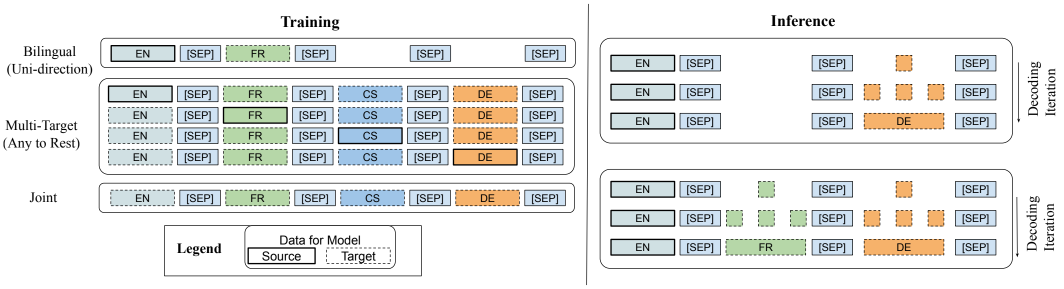

Figure 5: *(Left)* Training data used for various translation models. Within each rounded box contain example row(s) of data used to train a model. A solid outlined box indicates the source sentence (provided in full) while dashed outlined box depicts target languages to predict. *(Right)* During inference, our model can generate translation for a single target language (top), or for multiple languages in parallel (bottom), conditioning on source sentence and partial translations of multiple languages.

A.3 Parallel Greedy Decoding Example

We illustrate the parallel generation across multiple target languages below:

Input: A man sits on a bench holding his dog and looking at the water.
Parallel Decode:

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR:_Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS:_Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

Table 2: Example parallel greedy decode using the Multi-target (Any → Rest) KERMIT model, starting with an English sentence. Blue underlined tokens are the inserted tokens at each iteration, and the gray tokens are the final output that have not been generated yet.

A.4 Unconditional Sampling Generation

Table 3 illustrates the serial sampling (one token at a time) from the joint model, every 20 timesteps

Iterations	Language	Generated Sentence from Joint Model
1	English	
	French	
	Czech	Mladý
	German	
20	English	
	French	descendant
	Czech	Mladý muž v modré bundě stoupá po
	German	Mann klettert.
40	English	blue jacket walking up a mountain.
	French	veste descendant paroi rocheuse en
	Czech	Mladý muž v modré bundě stoupá po horách.
	German	Mann klettert.
60	English	A man blue jacket walking up a mountain.
	French	veste bleue descendant une paroi rocheuse en horu.
	Czech	Mladý muž v modré bundě stoupá po horách.
	German	Mann einer blauen klettert eine hoch.
80	English	A young man in blue jacket walking up a mountain.
	French	veste bleue descendant une paroi rocheuse en horu.
	Czech	Mladý muž v modré bundě stoupá po horách.
	German	Ein junger Mann in einer blauen Jacke klettert eine Felswand hoch.
96	English	A young man in a blue jacket walking up a mountain.
	French	Un jeune homme en veste bleue descendant une paroi rocheuse en horu.
	Czech	Mladý muž v modré bundě stoupá po horách.
	German	Ein junger Mann in einer blauen Jacke klettert eine Felswand hoch.

Table 3: Example of serial sampling unconditional text generation from the joint $p(EN, FR, CS, DE)$ model, over 96 insertion time steps. Note that the model generates one long sequence and we split them into the resulting four sentences in each language here.