# Guiding Zero-Shot Paraphrase Generation with Fine-Grained Control Tokens

**Teemu Vahtola** and **Mathias Creutz** and **Jörg Tiedemann**
Department of Digital Humanities
Faculty of Arts
University of Helsinki
Finland

## Abstract

Sequence-to-sequence paraphrase generation models often struggle with the generation of diverse paraphrases. This deficiency constrains the viability of leveraging paraphrase generation in different Natural Language Processing tasks. We propose a translation-based guided paraphrase generation model that learns useful features for promoting surface form variation in generated paraphrases from cross-lingual parallel data. Our proposed method leverages multilingual neural machine translation pretraining to learn zero-shot paraphrasing. Furthermore, we incorporate dedicated prefix tokens into the training of the machine translation models to promote variation. The prefix tokens are designed to affect various linguistic features related to surface form realizations, and can be applied during inference to guide the decoding process towards a desired solution. We assess the proposed guided model on paraphrase generation in three languages, English, Finnish, and Swedish, and provide analysis on the feasibility of the prefix tokens to guided paraphrasing. Our analysis suggests that the attributes represented by the prefix tokens are useful in promoting variation, by pushing the paraphrases generated by the guided model to diverge from the input sentence while preserving semantics conveyed by the sentence well.

## 1 Introduction

Paraphrasing is a way of conveying some given meaning using different wording. Automatic paraphrase generation aims to produce sequences that carry similar semantics to some arbitrary input sentence but are realized in different surface forms. Table 1 presents examples of paraphrases. Approaches for natural language generation incorporating diverse paraphrasing can be highly influential for many natural language processing (NLP) tasks where it is important to recognize sequences that share contextual meaning regardless of their surface form realizations. Such tasks include, but are not limited to, question answering (Dong et al., 2017), machine translation (Callison-Burch et al., 2006; Mehdizadeh Seraj et al., 2015), summarization (Nema et al., 2017), and simplification (Nisioi et al., 2017). Models that reliably represent similar meanings regardless of their surface forms can also be highly useful for instance in style transfer (Krishna et al., 2020), conversational applications (Dopierre et al., 2021), and tracking how information changes across multiple domains (Wright et al., 2022). However, for generated paraphrases to be useful in various NLP tasks, their realizations must deviate enough from the original sequences while preserving the semantics of the original sequence well. Sequence-to-sequence-based paraphrasing is prone to generating sequences whose surface forms highly resemble the original sentence by producing trivial rewrites of the input sentence (Kumar et al., 2019). This impediment constrains their practical viability to the aforementioned tasks.

To increase variation, we propose the training of a guided multilingual neural machine translation (NMT) system that can be applied to diverse zero-shot paraphrase generation by leveraging dedicated prefix tokens designed to enhance variation. We train our multilingual translation system in English, Finnish, and Swedish, and apply it to guided zero-shot paraphrasing in the three languages. The model does not see parallel monolingual sentence pairs during training, but we guide it to produce monolingual paraphrases during inference.

During training, our proposed model learns the semantics of a set of dedicated prefix tokens that are designed to capture certain attributes of language, and can be used for promoting diversity in generated text during inference. The attributes we consider are length, lexical variation, word order, and negation. When generating paraphrases, we can thus guide the model to produce sentences that vary in the given attributes by assigning corresponding values to the prefix tokens. Apart from a



*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 323–337

| Original | Paraphrase |
|---|---|
| They are excellent dancers. | They dance extremely well. |
| The dinner will be served in the dining area. | The dining area is where the dinner will be served. |
| He enjoys playing the guitar. | Playing the guitar brings him joy. |

Table 1: Examples illustrating paraphrasing.

few language-specific rules for recognizing explicit negation, our control tokens are language-agnostic.

By evaluating the applicability of multilingual NMT pretraining with prefix tokens to paraphrasing, we analyze whether the dedicated prefix tokens increase variation in sequence-to-sequence-based paraphrasing. We asses the generated sequences with respect to the references using BLEU (Papineni et al., 2002), and analyze the ranking of generated correct references using Mean Reciprocal Rank. Additionally, we analyze how faithful the model is to the given instructions during decoding by comparing the accuracy of the guided model outputs to the prefix tokens. We also apply the models to a novel test suite (Vahtola et al., 2022), designed for analyzing how language models represent negation. Our analysis suggests that the paraphrases generated by the proposed model are more diverse compared to the baseline model, especially when selecting hypotheses from n-best lists with smaller n-sizes, and preserve semantics of the original sentence well.

The main advantage of our approach is that we train our system on parallel cross-lingual translation pairs rather than monolingual paraphrase data. Translation examples are available for a far larger number of languages and in larger quantities than monolingual paraphrases. As a result, our approach can be extended to a considerably larger number of languages than models that depend on existing paraphrase data. Furthermore, our model is not tied to diversity in the monolingual paraphrase examples in obtaining variation in the generated sequences. As we use cross-lingual training examples, the model can learn characteristics that might not be prominent in the existing paraphrase data sets. For instance, large language models do not reliably represent negation (Ettinger, 2020), which can be a result of not having a sufficient number of such examples in the training data. We show that the proposed model can learn the semantics of a set of dedicated guiding tokens, for instance a token for negation from sentence pairs where an explicit negation occurs, and that these tokens can then be used to guide the decoder to produce sentences with desired characteristics.

Finally, we show that, especially when selecting hypotheses from smaller n-best lists, the guided paraphrase generation model goes beyond variation that can be achieved by filtering beam search (Kumar et al., 2019), as the prefix tokens provide more control for variation.

## 2 Previous Research

Previous research has studied paraphrase generation inspired by NMT systems. Prakash et al. (2016) use a deep LSTM network for paraphrase generation using monolingual parallel training data. Sjöblom et al. (2020) train encoder-decoder-based paraphrase generation systems for six languages, likewise using paraphrastic sentence pairs.

As an alternative to paraphrase data, cross-lingual parallel data has been used for finding paraphrases (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Ganitkevitch et al., 2013; *inter alia*). Mallinson et al. (2017) generate paraphrases via bilingual pivoting using a NMT system. Similarly, models based on NMT have been used in generating synthetic paraphrase pairs for learning paraphrastic sentence embeddings (Wieting et al., 2017; Wieting and Gimpel, 2018).

Additionally, multilingual NMT systems have been applied for paraphrase generation leveraging both parallel and monolingual data (Tiedemann and Scherrer, 2019), and assessing generalization to zero-shot paraphrasing while also promoting variation in the generated sequences by penalizing matching tokens in the source and output sentences (Thompson and Post, 2020). Zero-shot paraphrasing using large multilingual language models has also been explored (Guo et al., 2019).

Exploiting various linguistic features to control the decoding process of sequence-to-sequence models has been studied in different NLP tasks and granularities. Auxiliary control tokens have been used for controlling the language of the output in multilingual NMT (Johnson et al., 2017). Schioppa et al. (2021) use various features for controlling

output translations from a NMT system. In addition to prefix-based control tokens, they use vector-based interventions that guide the decoding process to certain directions. The complexity of the generated translations have been controlled by utilizing reading level tags, and by partitioning data based on reading or grade levels (Marchisio et al., 2019; Agrawal and Carpuat, 2019). Takeno et al. (2017) and Lakew et al. (2019) control length of the translated sequences with control tokens. Additionally, control tokens have been used with NMT systems for instance in domain adaptation (Kobus et al., 2017; Takeno et al., 2017), formality transfer (Sennrich et al., 2016; Niu et al., 2018), and voice control (Yamagishi et al., 2016).

Outside of machine translation, control tokens have been used successfully for instance in sentence simplification (Martin et al., 2020). Additionally, control tokens have been applied to sentences mined from the internet to obtain synthetic simplification data (Martin et al., 2022). In paraphrase generation, additional linguistic information obtained from the training data has been used for example in syntactic guiding (Iyyer et al., 2018; Huang and Chang, 2021; Sun et al., 2021).

Our approach to promoting variation is inspired by Schioppa et al. (2021) and Martin et al. (2022). We leverage existing translation corpora to learn controlled zero-shot paraphrasing using dedicated prefix tokens whose semantics the model learns directly from the training data. Our control tokens are designed to affect various properties of natural language. However, unlike Schioppa et al. (2021), we do not assess our model on machine translation, but take one step further, and evaluate it in zero-shot paraphrasing. We do not only attempt at increasing variation in lexical choices or divergence in syntactic realizations, for instance, but aim to affect both concurrently.

## 3 Guiding Attributes

To guide the decoding process, we need a method for signaling which decisions the decoder should take. Here, we use a prefix token-based approach, where we extract certain features from the source-target pairs in the training data, and concatenate the extracted information to the source side in the form of prefix tokens. We let the model learn to represent the semantics of each prefix token from the information incorporated in the translation pairs. Consequently, we can guide the decoding process of the proposed paraphrase model by applying these prefix tokens in monolingual transformation triggered by a target language token.

To promote variation in the generated paraphrases, we use the following attributes to control for various properties of natural language: length, lexical variation, word order, and negation.

### 3.1 Length

Inspired by automatic text simplification, we include a length-controlling token into our experiments. We represent the length attribute as a ratio between the lengths of source and target sentences after SentencePiece tokenization (Kudo and Richardson, 2018). We use pretrained SentencePiece models with a vocabulary size of 32 000 from the Opus-MT project (Tiedemann and Thottingal, 2020). If the sentences in a translation pair have exactly the same length after segmentation, the length ratio between the sentences is 100% (indicating that the target sequence should consist of 100% of the segments of the source sequence). Similarly, if the number of tokens in the target sentence is half of the number of tokens in the source, the length ratio is 50%. We round the length values to the nearest 10 to limit the number of features the model has to learn for controlling length.

### 3.2 Lexical Variation

Lexical variation could easily be measured in the monolingual case. However, we base our paraphrase generation model on multilingual machine translation and, therefore, need to apply a different mechanism to promote variation in lexical choices. We choose to base this prefix token on tf-idf. In previous research, tf-idf values have been used to measure lexical complexity of a sentence (Huang et al., 2021), but in our approach we apply them to promote lexical variation.

When calculating the tf-idf values, we treat each target sentence as a document, and calculate tf-idf over all the sentences in a given language pair. We consider the highest value in the resulting vector as a rough proxy of the lexical complexity of the sentence.

We automatically assign the obtained values into quartiles. Intuitively, sentences assigned into the first quartile should consist of simpler and more frequent tokens, whereas sentences in the subsequent quartiles should include less frequent, and increasingly difficult tokens. We hypothesize that controlling for tf-idf quartiles will promote divergence in

terms of lexical variation in sentence-to-sentence paraphrasing. Additionally, it could provide a simplifying effect if applied to simplification tasks.

### 3.3 Word Order

As an attribute of word order, we use the monotonicity of word alignments as proposed by Schioppa et al. (2021). Here, monotonicity refers to the degree of preservation of word order in the source compared to the target sentence. First, we apply `fast_align` (Dyer et al., 2013) to encode sentence pair alignment in the "Pharaoh" format, where the *i*th token of the input sentence is paired with the *j*th token of the output sentence, and the alignments are indicated by the corresponding word indices (e.g., 0-0 1-1 2-2 for a bijective alignment of two sentences with three tokens, or 0-2 1-1 2-0 for reversed word order). Next, we apply the following calculation from Schioppa et al. (2021):

$$\delta(s) = \frac{1}{\#\{(i,j)\}} \sum_{(i,j)} \left| \frac{i}{n} - \frac{j}{m} \right| + 0.1 \quad (1)$$

where $\#\{(i,j)\}$ stands for the cardinality of the alignments.

We assign the obtained monotonicity values $\delta(s)$ for each sentence pair automatically into quartiles, similarly as with the lexical variation tokens. We hypothesize that during inference, keeping other prefix token features constant, controlling for monotonicity promotes variation in word order in relation to the input sentence by guiding the model for either more monotone or more varied choices of word order.

### 3.4 Negation

Previous research has suggested that language models do not reliably represent negation (Ettinger, 2020; Hartmann et al., 2021). Therefore, we include a prefix token for controlling polarity of a generated sentence. By applying polarity change, we focus on one specific case of paraphrase formulation, namely, antonym substitution (Bhagat and Hovy, 2013). In this paradigm, some word in a sentence is substituted to a word that carries the opposite meaning to the original word, that is, its antonym. Concurrently, to maintain the original meaning, a negation is either inserted to or deleted from a corresponding position in the sequence. As an example, the sentence *My brother is asleep* could be paraphrased as *My brother is not awake*, by using antonym substitution as defined in Bhagat and Hovy (2013).

As a control token for polarity change, we use Boolean values to indicate whether an explicit negation occurs in the target sentence. We use handwritten rules to automatically recognize negation in each target sentence. These rules are designed to only grasp explicit negation (e.g., *not*) as opposed to alternative ways of conveying opposite meanings, such as negative prefixes (e.g., *un-*, *im-*, *dis-*, *il-*, *ir-*, and *in-* in English). Our hypothesis is that by explicitly expressing the presence of a negation token in a target sentence, the prefix token can be used for controlling polarity of a paraphrase.

## 4 Experiments

We train two multilingual NMT models for English, Finnish and Swedish from scratch: a baseline model without prefix tokens apart from the target language token, and our proposed model with prefix tokens. Both models are based on the Transformer architecture (Vaswani et al., 2017), and trained using OpenNMT (Klein et al., 2017) with standard hyperparameters for training a Transformer. We gather training data from OpenSubtitles (Lison and Tiedemann, 2016) using OpusTools (Aulamo et al., 2020), by filtering for one-to-one aligned sentences with a time stamp overlap threshold of 0.85.[1] The obtained training set consists of approximately 17 million sentence pairs in three language pairs (en-fi, en-sv, fi-sv). We extract 10 000 sentence pairs from each direction to serve as validation data for tuning the translation models. We train the models for all cross-lingual directions on two GPUs for one million steps or until early stopping criteria is met.

We evaluate the models on true paraphrase pairs extracted from the Opusparcus test sets (Creutz, 2018). The sizes of the filtered English, Finnish and Swedish test sets are 723, 669, and 732 sentence pairs, respectively. Opusparcus is a sentential paraphrase corpus that consists of paraphrastic bi-texts in six languages, English, Finnish, and Swedish included. The data is collected from the OpenSubtitles corpus, and therefore matches the domain of the training data. Consequently, the translation models may have seen some of the sentences included in the test sets during training, either on the encoder or the decoder side, but not as parallel

[1] Time-overlap ratio based on the time information given for each pair of aligned subtitle lines in the corpus.

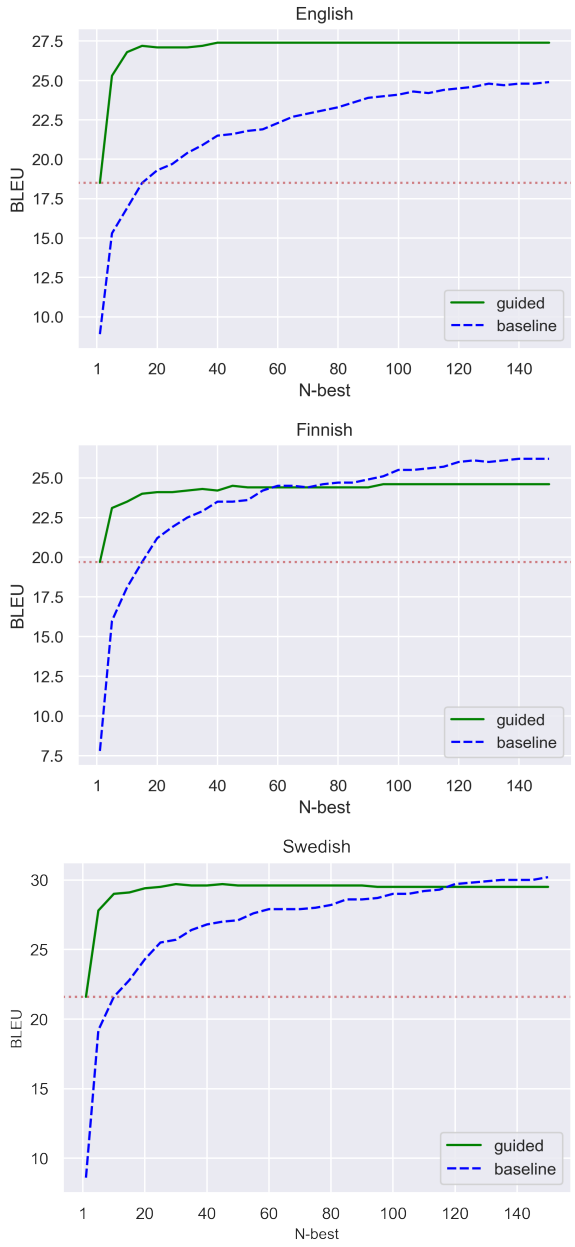


Figure 1: Obtained BLEU scores calculated on the Opusparcus test sets for the guided and the baseline models for sentences selected from different n-best lists. The x-axis denotes the size of the n-best list where the best hypothesis is selected from. The y-axis denotes the obtained BLEU scores. The horizontal line indicates the obtained accuracy of the 1-best translation from the guided model.

monolingual pairs.

## 5 Results

### 5.1 Automatic Evaluation

We assess the alignment of the generated paraphrases to their reference sentence based on BLEU, and further analyze the quality of the systems in terms of Mean Reciprocal Rank.

#### 5.1.1 BLEU

We evaluate our model by testing how well it can generate a paraphrase of a source sentence that closely aligns to the desired target sentence. To quantify this, we use BLEU, which is an established metric in machine translation, for comparing a produced translation to a given reference. As the test examples are designed to exhibit surface form variation (Creutz, 2018), increase in BLEU implies increased variation in sentences generated by the models.

During inference, the guided model requires prefix tokens to perform guided paraphrase generation. We calculate the true guiding values for the prefix tokens from the test set examples, and input them together with the source sentence to the guided model. For calculating the test set prefix tokens, we first train `fast_align` parameters for each language using the first 500 000 paraphrase pairs from the corresponding Opusparcus training sets, and use these alignment parameters for calculating the word order features. For the lexical variation attribute, we use the tf-idf weights learnt from the training data to assign the lexical variation values of each target sentence. Consequently, the guided model can leverage this information about the ground truth reference during decoding. The baseline model, however, has no information about the reference sentence during decoding. As such, this evaluation setup would result in an unfair comparison of the models. Therefore, we use beam search with a beam size of 250 to generate n-best hypotheses from both models. From the n-best lists, we choose the hypothesis that most accurately matches the desired prefix tokens. Now, also the baseline model has a fair chance of producing a sentence the matches the desired guiding values, if such a hypothesis is available in the n-best list.[2]

Figure 1 presents BLEU scores of the models for n-best lists ranging in size from 1 to 150. The results indicate that our proposed guided paraphrase

[2]When determining which hypothesis is the best match for the desired guiding values, we treat the prefix token values as vectors. The negation tokens are mapped from Boolean values into their binary feature representation $\{0, 1\}$ and the other prefix tokens are normalized in the range $[0, 1]$ using min-max normalization. We calculate the cosine similarity of the ground truth prefix token values and all the hypotheses' prefix token values, and choose the one that maximizes cosine similarity. If multiple hypotheses maximize the similarity (e.g., multiple hypotheses have cosine similarity of 1.0), we choose the hypothesis with the highest translation score.

generation model greatly benefits from the information provided by the prefix tokens. Considering only the 1-best hypotheses for each language, the guided paraphrase generation model obtains significantly higher BLEU scores than the baseline model (18.6 vs. 8.9, 19.7 vs. 7.8, and 21.6 vs. 8.6 for English, Finnish and Swedish, respectively). Increasing the pool of hypotheses to 5-best increases BLEU scores of both models.

The steep increase of BLEU scores between 1-best and 5-best, which is obtained by the guided model, may seem surprising at first. Why does the 1-best translation not match the guidance values the best? We hypothesize that this is caused by the model balancing between what it considers the best translation and the decisions it is supposed to be making based on the prefix tokens. In practice, the model might find a solution that it considers a better translation, even if it means partly ignoring the guiding tokens. Consequently, increasing n-best size results in the model selecting a sentence that better matches the guiding tokens, which in turn increases the obtained BLEU score. However, on average, the guided model does not benefit from n-best sizes larger than 15. At this point, the model has found a solution that maximizes the similarity to the ground truth prefix tokens for each input sentence.

The baseline models benefit greatly from filtering from a larger collection of hypotheses. Albeit beginning from a very low BLEU score in all languages, the results for Finnish and Swedish surpass the ones obtained by the guided model when selecting from a sufficiently large set of hypotheses (approximately 60-best hypotheses for Finnish, and 110-best hypotheses for Swedish). In terms of the guided model, the prefix tokens constrain the options where the model can choose from during decoding, since it also needs to consider the given instructions. As a result, the output sequences are close to the desired outputs in terms of the prefix tokens to begin with. In case of multiple hypotheses that maximize the similarity to the reference prefix tokens, we choose the first such occurrence in the n-best list. This is not always the one that maximizes alignment to the reference translation.

### 5.1.2 Mean Reciprocal Rank

Mean Reciprocal Rank (MRR) calculates the average of the reciprocals of the first generated sequence that exactly matches the reference:

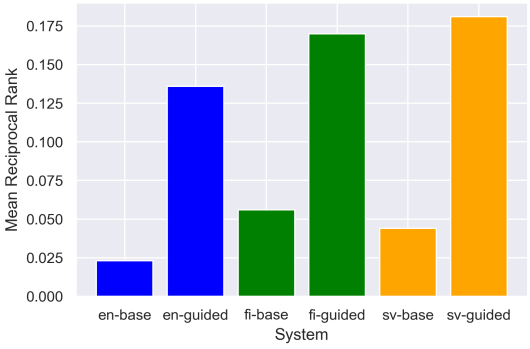


Figure 2: Obtained Mean Reciprocal Rank of the baseline and guided models calculated from the 250-best lists. The x-axis denotes the system, and the y-axis indicates the obtained MRR score.

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}, \quad (2)$$

where $N$ is the number of paraphrased sentences, and $rank_i$ refers to the position in the n-best list of the first sentence that matches the reference.[3]

The score indicates how consistently the models retrieve and rank the correct references high in the generated n-best lists. Figure 2 presents the MRR scores of the models. The guided models consistently rank the generated sentence that matches the reference higher compared to the baseline models, whereas the baseline model struggles in ranking matching sentences high in the n-best list. We believe that the low ranking performance of the baseline model is mainly caused by the decoding algorithm. Beam search is known to produce bland outputs (Holtzman et al., 2020), and when applied in paraphrasing, this realizes in copies or trivial rewrites of the input sentence. In fact, the 1-best outputs of the baseline model exactly match the source sentence in 71% of the cases in English, whereas the guided model ranks a copy of the source as the best paraphrase only in 12% of the cases (46% vs. 10%, and 60% vs. 10% in Finnish and Swedish, respectively). If the reference sentence is produced by the baseline model, it is ranked lower in the n-best list, since the model prefers repetitions of the input. When decoding is restricted with the guiding tokens, the decoder works with a notion of assumed diversity, resulting in outputs that may be closer to the

[3]The rank is defined as 0 if none of the proposed target sentences in the n-best list match the desired reference.

| Input | I haven't been contacted by anybody. |
|---|---|
| Baseline | Guided |
| I haven't been contacted by anyone. | Nobody has contacted me yet. |
| I haven't been contacted by anybody. | I have not been contacted. |
| I have not been contacted by anyone. | No one has contacted me. |
| I haven't been approached by anyone. | I was contacted by nobody. |
| I've never been contacted by anyone. | *Nobody's contacted me.* |

Table 2: Top-5 generated sequences from the baseline and the guided model for the input sentence: I haven't been contacted by anybody. The gold reference is highlighted in cursive.

| Language | Negation | Length | Lexical Variation | Word Order |
|---|---|---|---|---|
| English | 99.72 | 99.31 | 87.00 | 55.46 |
| Finnish | 100.0 | 98.51 | 79.07 | 63.86 |
| Swedish | 100.0 | 98.09 | 88.93 | 58.20 |

Table 3: Prefix token accuracy [%] calculated from the observed realizations of the guided models' 1-best hypotheses with respect to the ground truth reference prefix tokens.

reference to begin with. An example illustrating this phenomenon is provided in Table 2. Finally, even when the baseline model generates sentences that better match the reference, as indicated by increase in BLEU with large n-best sizes, it does not generate exact matches of the references, or fails to rank them high in the n-best list.

To conclude, we observe two opposite factors working in favor of the models: On the one hand, the use of explicit prefix tokens in the guided models produces high BLEU values instantly, even for very small-sized n-best lists. This makes it possible to use smaller beam sizes, which leads to faster inference. On the other hand, the absence of explicit guiding tokens in the baseline models seems to constrain the decoding process less, which may eventually result in translations that match the references better, if we can afford large n-best lists. However, that requires larger beam sizes and heavier computation. Additionally, the favorable trend for the baseline model is observed only for Finnish and Swedish.

### 5.2 Faithfulness to the Control Tokens

Automatic evaluation suggests that the guided paraphrase generation model obtains more variation and increases the quality of the generated paraphrases compared to the baseline model, especially when paraphrase hypotheses are selected from smaller n-best lists. To analyze how faithful the model is to the given prefix tokens, we calculate the accuracy of each prefix token of the generated sequences with respect to the ground truth prefix tokens. Table 3 presents the results.

The model seems to learn the semantics of two tokens, negation and length, especially well, but somewhat struggles with the features designed for promoting variation in lexical choices and in word order. The word order attribute seems particularly difficult to the model. This weakness can be a consequence of two aspects. First, when assigning the feature values for the word order feature, we binned the sentences into four (nearly) equally sized buckets automatically. Hence, sentences appointed in adjacent buckets may only have minor differences. This, in turn, makes recognizing differences between the adjacent quantiles unnecessarily difficult for the model, and the model can not generalize to this information. Secondly, the sentences in the Opusparcus test sets are rather short, which restricts the possibilities for finding solutions that incorporate variation in word order.

In addition to analyzing how accurately the model learns to follow the given prefix tokens, we assess whether the prefix tokens affect the output as expected by focusing on each prefix token separately. We generate hypotheses from the guided model using a beam size of 5 and only consider the top-1 hypothesis. Now, we do not rely on the prefix token values calculated from the reference sentences. Instead, we manually tune the values of the prefix tokens to obtain diverse paraphrases with the desired surface form variation. Controlling for different attributes demonstrates how changing the prefix token values affect the generated sequences. We present examples of English paraphrases with the given prefix token values in Tables 4–7. Examples for Finnish and Swedish paraphrasing are provided in the Appendix A.

| Negation | Length | Lexical Variation | Word Order | Input | Output |
|---|---|---|---|---|---|
| True | 100 | 1 | 4 | Time's short. | Not much time left. |
| True | 100 | 1 | 2 | He must remain here. | He cannot leave here. |
| True | 100 | 1 | 2 | Has this ever happened to you? | This has never happened to you? |
| False | 100 | 3 | 4 | Don't be silly. | Stop fooling around here. |
| False | 100 | 1 | 4 | I didn't have much choice. | I had little choice, though. |
| False | 100 | 3 | 4 | I'm not feeling very well. | I'm feeling a little poorly. |

Table 4: Generated sentences from the guided model using different prefix token values for controlling negation in the output. The prefix token for negation indicates whether there should be an explicit negation in the output sequence or not.

| Input | | | | Can I ask a simple question? |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 50 | 1 | 3 | A question? |
| False | 80 | 1 | 3 | Can I ask you? |
| False | 100 | 1 | 3 | Can I ask you a question? |
| False | 120 | 1 | 3 | Can I ask you a very easy question? |
| False | 150 | 1 | 3 | Do you mind if I ask you a simple question? |

Table 5: Generated sentences from the guided model using different prefix token values for guiding for the length of the output. The prefix token value denotes the ratio between the tokenized input and output sequences.

**Negation** Table 4 provides examples of how the negation token affects the generated outputs. To further analyze the prefix token that controls negation, we use a recent test suite for analyzing vector-based representations of antonymy and negation (Vahtola et al., 2022). The data consists of approximately 3000 test examples where an input sentence, for instance *I'm guilty*, is paired with three hypothetical paraphrases: *I'm innocent, I'm not guilty*, and *I'm not innocent*. The first two hypotheses semantically oppose the input sentence, whereas the last hypothesis carries the closest meaning to the input sentence. Using this test suite, we analyze how our proposed model learns the semantics of the negation token.

In practice, we use the translation probabilities to find which of the three hypotheses each model would translate the input sentence to, and calculate the accuracy of the model over the test set based on the preferred output. The baseline model obtains an accuracy of 30%, which is lower than acquired by random choice (33%). The guided model obtains a higher accuracy, 41%, suggesting that explicit information about negation assists the model in generating better representations of negation. However, the model does not seem to reliably learn the interplay of negation and antonymy in sentence semantics. Regardless, examples given in Table 4 show that the guided model learns, at least to some extent, to reformulate sentences with polarity change while maintaining meaning close to the original.

**Length** Table 5 provides examples of how the length guiding feature effects the generated output. Keeping other prefix tokens constant, but guiding for five different values for length (50, 80, 100, 120, and 150), the model does follow the given instructions faithfully, further validating the results obtained with accuracy on the different guiding tokens.

**Lexical Variation** Table 6 provides examples of the effect of changing the lexical variation value while keeping other prefix tokens constant. Increasing the value for lexical variation does not only promote for varied lexical choices, but can also push for potentially less frequent word types (e.g., *'bout* and *wanna*) for sequences guided with larger values (3 and 4).

**Word Order** Learning the semantics related to the attribute guiding for variation in word order is difficult for the model, as indicated by the obtained accuracies on the prefix token (Table 3). Similarly, the examples in Table 7 demonstrate that the prefix token does not work exactly as expected, as sentences with word order values 1 and 2 are identical. However, when pushing for more variation in word order with larger values, the model generates sequences with syntactic alteration. The results suggest that as such the prefix token may not be

| Input | | | | Would you like a drink? |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 120 | 1 | 4 | Can I get you a drink? |
| False | 120 | 2 | 4 | May I offer you a drink? |
| False | 120 | 3 | 4 | How 'bout a drink? |
| False | 120 | 4 | 4 | Wanna have a drink '? |

Table 6: Generated sentences from the guided model using different prefix token values for promoting lexical variation in the output sequences. Sentences in bucket 1 should only include frequent tokens, and subsequent buckets should contain sentences where also less frequent and potentially difficult tokens are present.

| Input | | | | There's really nothing you can do. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| True | 80 | 1 | 1 | There is nothing you can do. |
| True | 80 | 1 | 2 | There is nothing you can do. |
| True | 80 | 1 | 3 | There really is nothing to do. |
| True | 80 | 1 | 4 | You really can't do anything. |

Table 7: Generated sentences from the guided model using different prefix token values for guiding output sequence's word order in relation to the input sentence. The sentences with lower values should preserve the word order of the input well, whereas sentences with larger values should deviate more from the input sentence in terms of word order.

optimized perfectly, but with careful redesigning of the attribute, it could provide a method of promoting variation in word order.

## 6 Conclusions

We propose a paraphrase generation model that is based on multilingual NMT, leveraging cross-lingual parallel examples as diverse paraphrase data. We apply dedicated diversity-promoting prefix tokens to the training of the model in order to obtain a paraphrase model designed for guided zero-shot paraphrasing, and compare the model to a baseline paraphrase generation model based on multilingual NMT without prefix guiding. Compared to the baseline model, the results suggest that the proposed guided paraphrase generation model benefits significantly from the guiding information, and produces paraphrases that deviate more from the original sentence but maintain the meaning of the original sentence well, especially with lower n-sizes of n-best decoding. The analysis also suggests that there is still room for improvement, and especially the prefix tokens promoting lexical and word order variation are not perfectly optimized.

In future work, we would like to further improve the aforementioned prefix tokens by either optimizing the bucketing based on the observed values better, or by modeling the variation promoting attributes directly within a paraphrase generation model. We would also like to evaluate the applicability of dedicated guiding attributes with different data sets or transfer tasks, such as simplification. The method could also be expanded to a larger number of languages by fine-tuning existing multilingual NMT models for guided paraphrasing. Finally, we plan to explore modular architectures for diverse paraphrasing.

## Acknowledgments

This work was supported by the Behind the Words project, funded by the Academy of Finland. We wish to acknowledge *CSC - The Finnish IT Center for Science* for the generous computational resources they have provided. We thank the anonymous reviewers for their insightful comments.

## References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and parallel corpus diagnostics. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3782–3789, Marseille, France. European Language Resources Association.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. PROTAUGMENT: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. 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 2454–2466, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Yinpeng Guo, Yi Liao, Xin Jiang, Qing Zhang, Yibo Zhang, and Qun Liu. 2019. Zero-shot paraphrase generation with multilingual language models. *CoRR*, abs/1911.03597.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Bo Huang, Yang Bai, and Xiaobing Zhou. 2021. hub at SemEval-2021 task 1: Fusion of sentence and word frequency to predict lexical complexity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 598–602, Online. Association for Computational Linguistics.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1022–1033, Online. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. 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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In

*Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

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

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited 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 881–893, Valencia, Spain. Association for Computational Linguistics.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1379–1390, Lisbon, Portugal. Association for Computational Linguistics.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multitask neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. 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.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, Osaka, Japan. The COLING 2016 Organizing Committee.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Eetu Sjöblom, Mathias Creutz, and Yves Scherrer. 2020. Paraphrase generation and evaluation on colloquial-style sentences. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1814–1822, Marseille, France. European Language Resources Association.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. 2017. Controlling target features in neural machine translation via prefix constraints. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 55–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2019. Measuring semantic abstraction of multilingual NMT with paraphrase recognition and generation tasks. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 35–42, Minneapolis, USA. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, Abu Dhabi, United Arab Emirates (Hybrid). 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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

## A Appendix. Finnish and Swedish Examples

Tables 8–11 present examples of paraphrasing in Finnish, and tables 12–15 in Swedish. Similarly as for English, we use the guided model with a beam size of 5 and only select the top-1 hypothesis.

| Negation | Length | Lexical Variation | Word Order | Input | Output |
|---|---|---|---|---|---|
| True | 190 | 4 | 4 | Huono idea. | Ei kuulosta hyvältä idealta. |
| True | 130 | 2 | 3 | Taidan viihtyä täällä. | Eiköhän tämä ole mukava paikka. |
| False | 120 | 1 | 4 | En ole turvassa täällä. | Tämä paikka on minulle vaarallinen. |
| False | 60 | 1 | 1 | Ei hän ole vainaa. | Hän on elossa. |

Table 8: Generated sentences from the guided model for Finnish paraphrasing using different prefix token values for controlling negation in the output. The prefix token for negation indicates whether there should be an explicit negation in the output sequence or not.

| Input | | | | Minusta se näyttää hienolta. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 50 | 3 | 4 | Upealta. |
| False | 80 | 3 | 4 | Näyttääpä hienolta. |
| False | 100 | 3 | 4 | Se näyttää minusta hienolta. |
| False | 120 | 3 | 4 | Näyttääpä se hienolta minusta. |
| False | 150 | 3 | 4 | Se näyttää hienolta, jos minulta kysytään. |

Table 9: Generated sentences from the guided model using different prefix token values for guiding for the length of the output. The prefix token value denotes the ratio between the tokenized input and output sequences.

| Input | | | | Taidan viihtyä täällä. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 130 | 1 | 4 | Minusta tuntuu, että pidän tästä. |
| False | 130 | 2 | 4 | Luulen, että viihdyn täällä. |
| False | 130 | 3 | 4 | Viihdyn täällä mainiosti. |
| False | 130 | 4 | 4 | Viihdyn täällä mainiosti. |

Table 10: Generated sentences from the guided model using different prefix token values for promoting lexical variation in the output sequences. Sentences in bucket 1 should only include frequent tokens, and subsequent buckets should contain sentences where also less frequent and potentially difficult tokens are present.

| Input | | | | Uskoakseni olet kuullut hänestä. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 100 | 1 | 1 | Uskon, että olet kuullut hänestä. |
| False | 100 | 1 | 2 | Uskon, että olet kuullut hänestä. |
| False | 100 | 1 | 3 | Uskon, että olet kuullut hänestä. |
| False | 100 | 1 | 4 | Olet tainnut kuulla hänestä jo. |

Table 11: Generated sentences from the guided model for Finnish paraphrasing using different prefix token values for guiding output sequence's word order in relation to the input sentence. The sentences with lower values should preserve the word order of the input well, whereas sentences with larger values should deviate more from the input sentence in terms of word order.

| Negation | Length | Lexical Variation | Word Order | Input | Output |
|---|---|---|---|---|---|
| False | 70 | 2 | 2 | Det är inte över än. | Det pågår fortfarande. |
| False | 80 | 2 | 3 | Faktiskt inte så bra. | Faktiskt ganska dåligt. |
| True | 120 | 3 | 4 | Det här är allt vi kan göra. | Vi kan inte göra nåt annat än det här. |
| True | 100 | 1 | 2 | Det är nåt helt annat. | Det är inte samma sak. |

Table 12: Generated sentences from the guided model for Swedish paraphrasing using different prefix token values for controlling negation in the output. The prefix token for negation indicates whether there should be an explicit negation in the output sequence or not.

| Input | | | | Det är min bröllopsdag. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 50 | 2 | 4 | Mitt bröllop |
| False | 80 | 2 | 4 | Jag gifter mig. |
| False | 100 | 2 | 4 | Det är mitt bröllop. |
| False | 120 | 2 | 4 | Det är mitt bröllop idag. |
| False | 150 | 2 | 4 | Det är mitt bröllop i dag. |

Table 13: Generated sentences from the guided model using different prefix token values for guiding for the length of the output. The prefix token value denotes the ratio between the tokenized input and output sequences.

| Input | | | | Det kommer att gå jättebra. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 80 | 1 | 1 | Det kommer gå bra. |
| False | 80 | 2 | 1 | Det kommer gå jättebra. |
| False | 80 | 3 | 1 | Det kommer gå smidigt. |
| False | 80 | 4 | 1 | Det blir skit bra. |

Table 14: Generated sentences from the guided model using different prefix token values for promoting lexical variation in the output sequences. Sentences in bucket 1 should only include frequent tokens, and subsequent buckets should contain sentences where also less frequent and potentially difficult tokens are present.

| Input | | | | Det har jag redan sagt. |
|---|---|---|---|---|
| Negation | Length | Lexical Variation | Word Order | Output |
| False | 110 | 2 | 1 | Det har jag redan talat om. |
| False | 110 | 2 | 2 | Det har jag redan talat om. |
| False | 110 | 2 | 3 | Det har jag ju redan berättat. |
| False | 110 | 2 | 4 | Jag har redan talat om det. |

Table 15: Generated sentences from the guided model for Swedish paraphrasing using different prefix token values for guiding output sequence's word order in relation to the input sentence. The sentences with lower values should preserve the word order of the input well, whereas sentences with larger values should deviate more from the input sentence in terms of word order.