TOKEN-LEVEL FITTING ISSUES OF SEQ2SEQ MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence-to-sequence (seq2seq) models have been widely used for natural language processing, computer vision, and other deep learning tasks. We find that seq2seq models trained with early-stopping suffer from issues at the token level. In particular, while some tokens in the vocabulary demonstrate overfitting, others underfit when training is stopped. Experiments show that the phenomena are pervasive in different models, even in fine-tuned large pretrained-models. We identify three major factors that influence token-level fitting, which include token frequency, parts-of-speech, and prediction discrepancy. Further, we find that external factors such as language, model size, domain, data scale, and pretraining can also influence the fitting of tokens.

We release our code for model and analysis on https://github.com/xxxx.

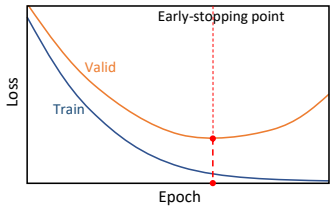
1 INTRODUCTION

Deep learning models tend to overfit because of their strong capacity and a massive number of parameters (Brownlee, 2018; Li et al., 2019; Rice et al., 2020; Bejani & Ghatee, 2021). Studies suggest regularization and early stopping to control the generalization error caused by overfitting (Hastie et al., 2009; Zhang et al., 2017; Chatterjee & Zielinski, 2022). Previous studies mainly analyze the generalization issue on image classification task (Arpit et al., 2017; Zhang et al., 2021), where the learning target is relatively simple. In contrast, NLP task such as machine translation (Zhang et al., 2015; Singh et al., 2017) is more complex with regard to the learning targets, which involve a sequence of tokens.

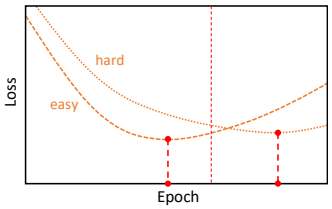
Natural languages exhibit a long-tailed distribution of tokens (Powers, 1998). The long-tail phenomena have been associated with performance degradation of NLP tasks (Gong et al., 2018; Raunak et al., 2020; Yu et al., 2022), where the rare (low frequency) tokens are ascribed as hard learning targets and popular (high frequency) tokens as easy learning targets. These criteria of easiness of learning targets are intuitive but coarse-grained, which are not associated with the training dynamics. In this paper, we study the easiness of tokens as learning targets from the perspective of overfitting and underfitting. Intuitively, the learning on hard tokens will be slower than that on easy tokens, which may result in underfitting on hard tokens and overfitting on easy tokens, as illustrated by Figure 1. We propose two measures to quantify fitting – *fitting-offset* and *potential-gain*. Fitting-offset measures the offset of the best fit from the early-stopping point, which reflects the degree of overfitting or underfitting. Potential-gain measures the accuracy gap between the early-stopping checkpoint and the best fit, which also estimates the accuracy decrease caused by overfitting or underfitting.

We use machine translation as our test bed, training models on English-German benchmark datasets, including News and Europarl domains. Our extensive experiments uncover multiple new and counter-intuitive findings: 1) Both overfitting and underfitting occur in a trained seq2seq model. 2) High-frequency tokens are expected to overfit, but some are found underfitted, and low-frequency tokens are expected to underfit, but some are found overfitted. 3) Large pretrained models reduce underfitting effectively during fine-tuning but are less effective on overfitting. Besides, we propose a direct indicator of easiness – prediction discrepancy, using the probability difference of predictions made by full context and local context as a criterion to group tokens.

In addition to tokens, sentences have also been considered as learning targets, where curriculum learning methods distinguish sentences as easy or hard (Kocmi & Bojar, 2017; Platanios et al., 2019; Xu et al., 2020). For example, the length of a sentence is used as a criterion to identify easy



(a) Idealized training and validation loss curves, where the model is selected by early stopping.



(b) Easy tokens overfit at the early-stopping point, while hard tokens underfit.

Figure 1: Seq2seq models trained with early stopping may suffer from overfitting or underfitting.

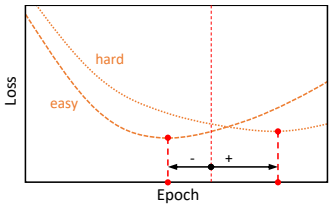


Figure 2: Fitting-offset measures the relative distance of the best fit from the early-stopping point, which is negative for the left side and positive for the right side.

sentences from hard ones, where the easy (short) sentences are first learned and then the hard (long) sentences. Using our metrics, we find that the length of a sentence is not a good indicator of easy or hard sentences from the perspective of overfitting and underfitting.

2 Related Work

Previous studies compare the fitting of real data and noise data, demonstrating that real data is easier to learn and has a faster convergence speed than noise data (Zhang et al., 2017; Arpit et al., 2017; Chatterjee & Zielinski, 2022). The different convergence speeds also provide an explanation of how early stopping prevents the memorization of noise data. However, these works do not compare the fitting among different learning targets inside the real data, neither between different samples nor between different parts of each sample. In this paper, we conduct experiments on a more complex seq2seq task instead of simple classification. We study the fitting of token-level learning targets, demonstrating that both overfitting and underfitting occur when training seq2seq models.

There are few works studying overfitting and underfitting in NLP. Sun et al. (2017) report that complex structure leads to overfitting in structured prediction. Wolfe & Caliskan (2021) demonstrate that low-frequency names exhibit bias and overfitting in the language model. Varis & Bojar (2021) illustrate that machine translation models generalize poorly on the test set with unseen sentence length. These works discuss overfitting issues on specific conditions, such as complex structure, frequent names, and unseen length. In comparison, we conduct a systematic analysis of the general phenomena of overfitting and underfitting in language. Specifically, we propose quantitative measures, identify major factors, and conduct statistical hypothesis testing on the phenomena.

3 Experimental Settings

3.1 Datasets

We experiment on two machine translation benchmark datasets. We use the News corpus as a major dataset for our experiments and analysis, and we use the Europarl corpus for the comparison of different domains and data scales.

News We use News Commentary v11 for training, newstest2015 for validation, and newstest2016 for testing. The English-German machine translation dataset contains $236,287$ sentence pairs for training, $2,169$ pairs for validation, and $2,999$ pairs for testing.

Europarl We use English-German Europarl v7, following Bao et al. (2021) to split the train, validation, and test sets. The dataset contains contains $1,666,904$ sentence pairs for training, $3,587$ pairs for validation, and $5,134$ pairs for testing.

We tokenize the sentences using MOSES (Koehn et al., 2007). We use truecase and a BPE (Sennrich et al., 2015) with $30,000$ merging operations. We use separate embedding tables for source and target languages in the model.

3.2 MODEL CONFIGURATIONS

We study the overfitting and underfitting issues on three model configurations.

Base Model We use the standard Transformer base model (Vaswani et al., 2017), which has 6 layers, 8 heads, 512 output dimensions, and 2048 hidden dimensions. We train the model with a learning rate of 5×10^{-4}, a dropout of 0.3, a label smoothing of 0.1, and an Adam optimizer (Kingma & Ba, 2014).

Big Model Following the standard Transformer big model (Vaswani et al., 2017), we use 6 layers, 16 heads, 1024 output dimensions, and 4096 hidden dimensions. We train the model with a learning rate of 3×10^{-4}, a dropout of 0.3, a label smoothing of 0.1, and an Adam optimizer.

Pretrained Large Model We use mBART25 (Liu et al., 2020), which has the similar setting as BART large model (Lewis et al., 2020), using 12 layers, 16 heads, 1024 output dimensions, and 4096 hidden dimensions. We fine-tune the model with a learning rate of 3×10^{-5}, a dropout of 0.3, an attention-dropout of 0.1, a label smoothing of 0.2, and an Adam optimizer.

For each experiment, we train 40 models using random seeds from 1 to 40, obtaining 40 samples for the significance test. During the training of the base or big model, we keep the last 20 checkpoints for analysis, where the checkpoint of early-stopping is the 10-th of the 20 checkpoints. For mBART25, we keep the last 10 checkpoints, and the early-stopping checkpoint is at the 5-th of the checkpoints.

3.3 EVALUATION METRIC

Measures We propose two measures: fitting-offset and potential-gain.

Fitting-offset represents how far (i.e., number of epochs) the best fit of a group of tokens diverges from the point of early stopping. In this paper, we use epoch as its unit because we evaluate the model using the validation set at the end of each training epoch. As Figure 2 shows, for the easy tokens, the fitting-offset is negative, denoting overfitting, where the best fit is before the early-stopping epoch. For the hard tokens, the fitting-offset is positive, denoting an underfitting, where the best fit is after the early-stopping epoch. Using fitting-offset, we can quantify the degree of overfitting and underfitting.

Potential-gain represents the potential accuracy increase if we move the best fit to the early-stopping epoch. We calculate the measure by subtracting the accuracy of the early-stopping checkpoint from the accuracy of the best fit. Using this measure, we can quantitatively estimate the potential benefits by fixing the overfitting or underfitting issue.

Significance Test Since the distribution of fitting-offset is unknown, we use a non-parametric sign-test (Dixon & Mood, 1946; Hodges, 1955) to test our hypothesis. We train the model N times to obtain N observations on the fitting-offset. The hypothesis about the overfitting and underfitting can be expressed by

$$\begin{cases} H_0 : \text{fitting-offset} = 0, & \text{there is not overfitting or underfitting;} \\ H_1 : \text{fitting-offset} \neq 0, & \text{there is overfitting or underfitting.} \end{cases} \tag{1}$$

If H_0 is true, the N observations are expected to be half positive and half negative. The total number of positive observations N_+ follows a binomial distribution, through which we decide the rejection region according to a significance level α.

Grouping Ideally, we can calculate the two measures on each token to tell which tokens are over-fitted and which tokens are underfitted at the early-stopping epoch. However, direct observation of each token is noisy and does not show obvious patterns. We group the tokens and average the valid losses to reduce the noise, through which the pattern emerges, and we obtain stable measures.

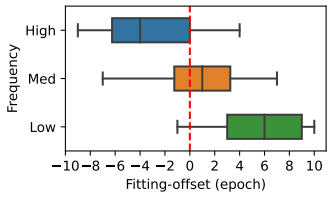


Figure 3: Fitting-offset of tokens grouped by token *frequency*.

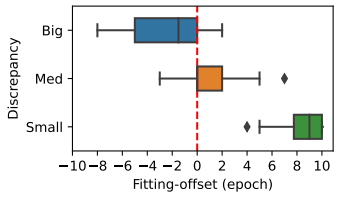


Figure 4: Fitting-offset of tokens grouped by prediction *discrepancy*.

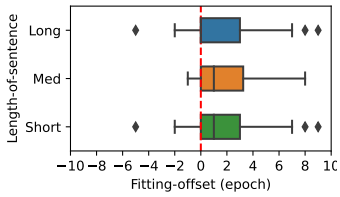


Figure 5: Fitting-offset of tokens grouped by *length-of-sentence*.

Table 1: Groups aggregating parts of speech.

Group	Parts-of-speech (POS)
Noun	NOUN, PRON, PROPN
Verb	VERB, AUX
Adjv	ADJ, ADV
Num	NUM
Func	ADP, CONJ, CCONJ, DET, PART, SCONJ
Symb	PUNCT, SYM

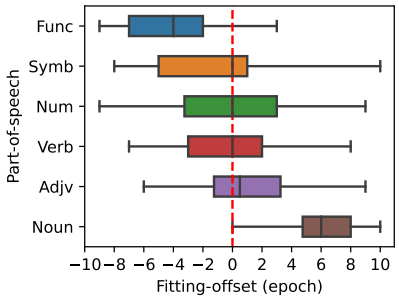


Figure 6: Fitting-offset of tokens grouped by *parts-of-speech*.

4 Token-level Results

4.1 Fitting of Rare Tokens in Seq2seq Model Training

Previous studies suggest that long-tail token distribution affects the performance of NLP tasks (Gong et al., 2018; Raunak et al., 2020; Yu et al., 2022). We hypothesize that the low-frequency tokens underfit during training and conduct verification experiments as follows.

Settings We experiment on the News English-German translation dataset, using a Transformer base model (Vaswani et al., 2017). We categorize the target tokens into high/medium/low-frequency according to their distribution in the training set, with balanced probability mass on the three buckets.

Hypothesis Testing For each group of high/medium/low-frequency tokens, we measure the fitting-offset using the checkpoints of each model, obtaining 40 samples of fitting-offset for each group. We test our hypothesis on each group using a sign-test, as described in Eq. 1. As a result, we obtain a p-value of 1.9×10^{-5} for high-frequency and 7.5×10^{-11} for low-frequency, which strongly supports the hypothesis that the high/low-frequency tokens either overfit or underfit.

Further as Figure 3 shows, the average fitting-offset for high-frequency tokens is -3.7 with a standard deviation of 3.4. The negative value of the fitting-offset indicates that the high-frequency tokens overfit, where the best fit happens at an average of 3.7 epochs, before the early-stopping point. The average fitting-offset for low-frequency tokens is 5.8 with a standard deviation of 3.3. The positive value of the fitting-offset indicates underfitting, where the best fit happens at 5.8 epochs, after the early-stopping point on average.

Based on this evidence, we conclude that

Both overfitting and underfitting occur at the token level when training seq2seq models.

Analysis The significant divergence of fitting-offset between the high/low-frequency tokens suggests that the frequency of tokens has a significant influence on their fitting. We quantify the influence using the potential-gain. In particular, take the low-frequency tokens as an example. The potential-gain is 0.73, which means that the average accuracy is expected to be increased from 45.61 to 46.34 if we move the best fit to the early-stopping epoch. The potential-gain of the high-frequency

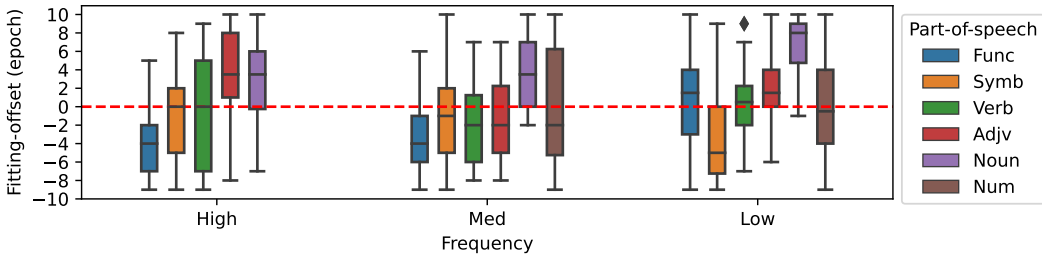


Figure 7: Fitting-offset of tokens grouped by *frequency* and *parts-of-speech*.

Table 2: Potential-gain for each category grouped by *frequency* and *parts-of-speech*. The column is in a format of "averaged-accuracy potential-gain", where the "+" in the potential-gain indicates an increase in the accuracy and the "-" indicates a decrease in the accuracy. We mark potential-gains bigger than 0.5 with the bold font to indicate their significance.

Frequency	Function	Symbol	Number	Verb	Adj/Adv	Noun
High	56.82 +0.15	84.77 **+0.59**	nan +nan	59.13 **+1.09**	71.02 **+1.67**	60.87 **+0.89**
Med	49.59 +0.16	69.63 **+3.41**	73.47 **+1.50**	47.64 +0.12	44.56 +0.24	59.21 **+0.76**
Low	43.76 **+1.38**	72.24 **+3.55**	74.84 **+1.20**	36.25 **+0.51**	41.18 +0.43	48.03 **+0.88**

tokens is 0.05, and that of the medium-frequency tokens is 0.23, which is relatively smaller than that of the low-frequency tokens, suggesting underfitting of the low-frequency tokens is the major issue.

4.2 Linguistic Factors to Token-level Fitting

In section 4.1, we find that the high-frequency tokens tend to overfit and the low-frequency tokens tend to underfit in the seq2seq model as a group. In order to further understand a fine-grained correlation between the frequency and the fitting of a token, we further split the high/low-frequency tokens into smaller groups and conduct experiments on the specific categories. Linguistic factors are considered in the detailed experiments.

Parts-of-speech (POS) We speculate that parts-of-speech, as an important linguistic feature, may provide a different perspective to study the overfitting and underfitting issues. We group tokens according to their parts-of-speech as listed in Table 1. Specifically, we first obtain POS tagging on each word using spaCy [1]. Then we map the POS of words to tokens by labeling all the tokens of a word with the same POS. Last, we group these tokens according to their POS. Take the group Noun as an example. We group tokens with the POS of NOUN, PRON, and PROPN into one category, naming Noun. We aggregate the major parts of speech into six groups according to their functional similarity, as shown in the Table.

As Figure 6 shows, parts-of-speech has a significant influence on the fitting of tokens. The function words are most likely to overfit, which is likely because they are close-set and easier to learn from the linguistic perspective. On the contrary, nouns are most likely to underfit, which can be due to the openness of the set and the challenging context dependencies.

The potential-gain of nouns is 0.69, increasing the accuracy from 52.38 to 53.07. The potential-gains of numbers, symbols, verbs, and adj/adv words are 1.09, 0.58, 0.26, and 0.22, respectively. Surprisingly, the potential-gain of function words is negligible, even though they obviously overfit. We attribute it to the overall high frequency of function words because sufficient training samples reduce the negative impact of overfitting. It is confirmed by the detailed potential-gains shown in Table 2, where the function words with low frequency have a much higher potential-gain of 1.38.

Frequency and Parts-of-speech We combine frequency and POS to make a detailed analysis of the high/low-frequency tokens. As Figure 7 shows, frequency and POS work independently. Among the high-frequency tokens, the function words tend to overfit, while adjvs and nouns tend to underfit. Among the low-frequency tokens, the symbols tend to overfit, while the adjvs and nouns tend to

[1] https://spacy.io/

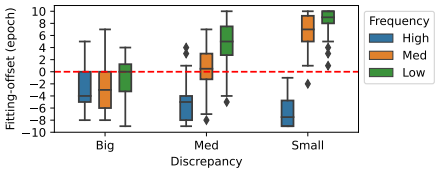


Figure 8: Fitting-offset of tokens grouped by *discrepancy* and *frequency*.

Table 3: Potential-gain for each category grouped by *frequency* and *discrepancy*.

Discrepancy	Frequency		
	High	**Med**	**Low**
Big	63.91 +0.09	26.88 -0.22	10.57 +0.11
Med	66.96 +0.09	47.77 +0.31	20.22 **+0.86**
Small	72.49 +0.09	80.95 **+0.62**	73.19 **+0.85**

underfit. Based on this evidence, we arrive at a counter-intuitive conclusion that

In a seq2seq model, the high-frequency tokens (popular tokens) mostly overfit but can also underfit and the low-frequency tokens (rare tokens) mostly underfit but can also overfit.

When we look into the potential-gains, as shown in Table 2, we see higher potential-gains than in the previous section. The potential-gain of low-frequency function words, symbols, and numbers are 1.38, 3.55, and 1.20, respectively. The potential-gains on med-frequency symbols and numbers are 3.41 and 1.50, respectively. Overall the high-frequency tokens have low potential-gains, and the verbs and adjvs have potential-gains of 1.09 and 1.67, respectively. These results demonstrate that combining the frequency and linguistic factors reveals stronger overfitting and underfitting, forecasting higher potential-gains in specific categories.

4.3 Using Prediction Discrepancy as A Measure for Token-level Fitting

Given that neither frequency nor parts of speech are decisive factors, we consider one additional factor, which relies on the context. We use prediction discrepancy to measure the degree of dependence on long context, which is calculated as

$$D_j = |P(Y_j|Y_{<j}, X) - P(Y_j|Y_{j-1}, X)|, \tag{2}$$

where X is the source sequence, and Y is the target. For each token Y_j, we predict it using its full context $Y_{<j}$ and its local context Y_j. We use the discrepancy between these two predictions to indicate its dependence on long context. We train an altered Transformer model to do the predictions using two decoders, where one decoder uses the full context of a target while another decoder uses the local context. The two decoders share the same token embedding table and encoder. According to the value of discrepancy, we categorize the tokens into three groups, with big, medium, and small discrepancy, respectively.

Results As Figure 4 shows, the big-discrepancy tokens have an average fitting-offset of -2.7 with a standard deviation of 3.0. The medium-discrepancy tokens have an average fitting-offset of 1.0 with a standard deviation of 2.1. The small-discrepancy tokens have an average fitting-offset of 8.2 with a standard deviation of 1.8, showing a trend to exceed the boundary of 10. In comparison with frequency, the bigger range of the average fitting-offsets and the smaller standard deviations suggest that discrepancy is a better indicator than frequency. This indicates that the discrepancy is a good indicator of overfitting and underfitting.

The potential-gain of the small-discrepancy tokens is 0.63, increasing the average accuracy of the tokens from 75.85 to 76.48. In comparison with the potential-gain of 0.75 for the low-frequency tokens, which increases the average accuracy from 45.61 to 46.34, the baseline accuracy of small-discrepancy is much higher, suggesting the effectiveness of discrepancy in discovering fitting issues among high accuracy predictions.

Discrepancy and Frequency Intuitively, discrepancy and frequency are two independent factors, given that discrepancy relies on context and frequency relies on the token itself. As Figure 8 shows, the most significant difference between high-frequency and med/low-frequency tokens is that med/small discrepancy tokens with high frequencies tend to overfit, while the med/small discrepancy tokens with med/low frequencies tend to underfit.

In addition, as shown in Table 3, when frequency and discrepancy are combined to predict the overfitting and underfitting, the biggest potential-gain of low-frequency tokens increases from 0.73 to 0.86, suggesting that frequency and discrepancy are two independent factors.

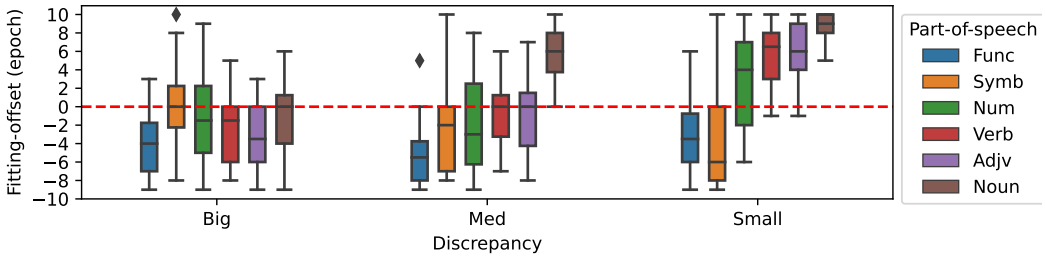


Figure 9: Fitting-offset of tokens grouped by *discrepancy* and *parts-of-speech*.

Table 4: Potential-gain for each category grouped by *discrepancy* and *parts-of-speech*.

Discrepancy	Function	Symbol	Number	Verb	Adj/Adv	Noun
Big	45.43 -0.07	83.38 **+0.51**	31.58 **+2.72**	20.31 +0.12	13.38 -0.10	20.58 -0.06
Med	58.33 +0.08	78.79 **+0.96**	36.49 **+2.02**	39.69 +0.25	29.51 +0.14	33.30 **+0.93**
Small	77.06 +0.44	77.98 **+1.63**	89.20 **+1.15**	69.90 **+0.84**	71.93 **+0.75**	77.08 **+0.85**

Discrepancy and Parts-of-speech (POS) As Figure 7 shows, discrepancy and POS also work orthogonally. Overall, tokens with a smaller discrepancy have a larger fitting-offset, which consistently appears on numbers, verbs, adjvs, and nouns. Function words and symbols show a different pattern that the med-discrepancy tokens tend to have smaller fitting-offset than high-discrepancy.

Looking into Table 4, we can see that small-discrepancy tokens have a potential-gain of 1.63 and 1.15 on symbols and numbers, respectively. The potential-gain of numbers on big/med-discrepancy tokens are 2.72 and 2.02, respectively, suggesting the effectiveness of combining the two factors.

Summary We have identified three independent factors that affect token-level fitting in seq2seq model training, including frequency, parts-of-speech, and discrepancy. While the former two are internal to the token, the third is external and context-dependent. These indicate that the fitting of tokens results from interestingly complex factors.

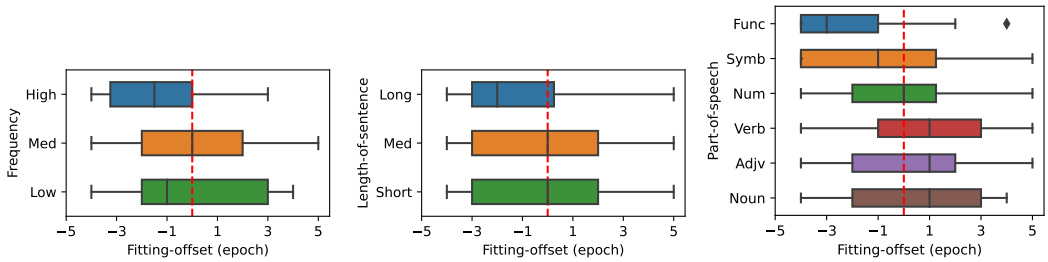
5 Easy Sentences vs Hard Sentences

Curriculum learning starts with easy sentences and then with hard sentences (Kocmi & Bojar, 2017; Zhang et al., 2018; Platanios et al., 2019; Xu et al., 2020; Zhou et al., 2020), whereas different criteria are used to measure the difficulty of sentences. Among these criteria, the length-of-sentence is the simplest and most popular one, which hypothesizes that short sentences will be easy to learn and long sentences will be hard to learn. We test the hypothesis by evaluating whether short sentences overfit and long sentences underfit in a trained seq2seq model.

Length-of-sentence We categorize sentences into short/medium/long sentences according to the sentence length that each bucket is allocated with almost the same number of sentences. On the News dataset, the length of short sentences is between 1 and 18 tokens, the length of medium sentences between 19 and 31, and the length of long sentences between 32 and 792.

Hypothesis Testing. We test our hypothesis using sign-test on News English-German dataset, obtaining a p-value of 3.6×10^{-5} for short-sentence, 2.1×10^{-5} for medium-sentence, and 2.6×10^{-3} for long-sentence, which indicates overfitting or underfitting. The fitting-offset has an average of 1.95, 2.0, and 1.38 for short/medium/long-sentences, respectively. The positive fitting-offsets suggest that they overfit in the trained models. However, as Figure 5 shows, the degree of overfitting and underfitting is less than that of frequency (Figure 3) and discrepancy (Figure 4).

Summary The above experiments suggest that although the length-of-sentence can differentiate easy sentences from hard sentences, its effectiveness may not be as significant as other factors such as frequency, discrepancy, and parts-of-speech. More surprisingly, short sentences are more likely to underfit than long sentences, which is also confirmed by experiments on pretraining settings in section 6, suggesting that we could not simply judge the short-sentences as easy and long-sentences as hard.



(a) Fitting-offset of tokens grouped by *frequency*.

(b) Fitting-offset of tokens grouped by *length-of-sentence*.

(c) Fitting-offset of tokens grouped by *parts-of-speech*.

Figure 10: The distribution of fitting-offset on *pretraining* setting.

6 Fine-tuning of Pretrained Language Models

Fine-tuning on a large pretrained model has become the dominant setting for NLP tasks (Kenton & Toutanova, 2019; Lewis et al., 2020; Brown et al., 2020; Liu et al., 2020). We investigate the overfitting and underfitting issues, particularly in the pretraining setting.

Hypothesis Testing We first test whether the overfitting and underfitting issues exist under the pretraining setting. We experiment by fine-tuning mBART25 on the News English-German dataset.

First, we evaluate frequency as an indicator of overfitting and underfitting. As Figure 10a shows, we obtain a p-value of 1.2×10^{-3} and an average fitting-offset of -1.48 on high-frequency tokens, suggesting the tendency of overfitting on high-frequency tokens. Results on medium/low-frequency tokens do not show significance, although the average fitting-offsets of 0.08 on medium-frequency tokens suggest slight underfitting.

Next, we consider length-of-sentence. As Figure 10b shows, the fine-tuning tends to overfit for long-sentence and we obtain a p-value of 1.1×10^{-2} and an average fitting-offset of -1.2 on long-sentences. It suggests obvious overfitting of long sentences, which is counter-intuitive because it is widely assumed that long-sentence is harder to learn than short-sentence.

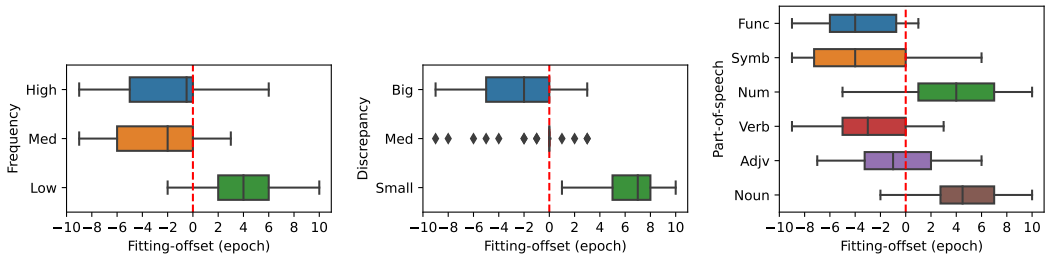
Last, we examine parts of speech. As Figure 10c shows, fine-tuning on function words tends to overfit and that on nouns underfit, which is consistent with the result on non-pretraining setting (Figure 6). A difference is that the fine-tuning of verbs shows underfitting, which is not significantly observed in the non-pretraining settings. We obtain a p-value of 5.8×10^{-2} and an average fitting-offset of 0.8 on verbs, suggesting an observable underfitting on verbs. The issue on function words is more significant than on verbs, on which we obtain a p-value of 4.3×10^{-6} and an average fitting-offset of -2.3, suggesting significant overfitting on function words.

The experiments above confirm that the overfitting and underfitting issues exist in the pretraining setting, although it is not as significant as that in the non-pretraining settings. In addition, it shows that overfitting is the major issue in comparison with underfitting. We attribute it to the effectiveness of large pretraining to prevent underfitting.

7 Additional Factors to the Fitting Issue

Most of the figures and tables for this section are placed in Appendix A.

The Language As a comparison of languages, we study the issues on the News dataset but with a reversed direction of languages, translating German to English instead of English to German. As Figure 11 shows, our observations on target English tokens are consistent with previous observations on target German tokens. First, the high-frequency tokens tend to overfit, while the low-frequency tokens tend to underfit. Second, the big-discrepancy tokens tend to overfit, while the small-discrepancy tokens tend to underfit. Third, the function words tend to overfit, while the nouns tend to underfit. Last, we also obtain bigger potential-gains by combining the factors and the most significant potential-gains happen on the consistent categories, such as low-frequency symbols,



(a) Fitting-offset of tokens grouped by *frequency*.

(b) Fitting-offset of tokens grouped by *discrepancy*.

(c) Fitting-offset of tokens grouped by *parts-of-speech*.

Figure 11: The Language: Fitting-offset of English tokens evaluated on News *German-English* dataset.

high-frequency adj/advs, and big/med-discrepancy numbers, as the bold numbers in Table 5, 6, and 7 show.

The Model Size To evaluate the influence of model size, we conduct experiments with a big model on the News English-German dataset. As Figure 15 shows, the distribution of fitting-offset on each category is very close to that of the base model but with a smaller range. We attribute it to the faster convergence of bigger model (Li et al., 2020). One significant difference between the big model and the base model is that the fitting-offset of symbols moves toward the negative region, suggesting overfitting for symbols in the big model. We attribute it to the stronger memorization ability of the bigger model.

The Domain Previous experiments are done on the News dataset. To justify that the phenomena are not domain specific, we conduct the same experiments on the Europarl English-German dataset, which is from the Europarl domain. We randomly sample $250,000$ sentence pairs from the Europarl training set for a fair comparison with the News dataset, which contains $236,287$ samples.

As Figure 16 shows, the observations described in "The Language" block hold but with slight differences in the distribution of fitting-offset of tokens grouped by parts-of-speech, in comparison with Figure 6 evaluated on News dataset. The major difference happens in verbs, adjvs, and nouns, reflecting a different distribution of topics of Europarl in comparison with News.

The Data Scale Intuitively, overfitting and underfitting should be more severe on small datasets. For comparison, we test on a bigger dataset from Europarl. We sample $500,000$ sentence pairs from the Europarl training set in comparison with the model trained using $250,000$ samples.

As Figure 17 shows, the observations described in "The Language" block still hold but the range of the distribution increases. Looking into the potential-gains, we see that they decrease by about $1/4$ compared to that of the experiments with $250,000$ samples. The results suggest that the fitting-offset is more challenging to measure, and the potential-gain decreases when the model is trained on a larger dataset, which is expected due to the larger dataset reducing overfitting and underfitting.

8 Conclusion

We study overfitting and underfitting issues of learning targets in the context of neural machine translation. Our experiments demonstrate that overall rare tokens tend to underfit and frequent tokens overfit. We explored detailed factors related to the overfitting and underfitting issues and identified three major influencing factors, which include frequency, parts-of-speech, and discrepancy. This shows that fitting is the result of a complex interaction between multiple factors. Further experiments demonstrate that the issues exist as a general problem for both non-pretraining and pretraining settings. Future work includes the investigation of strategies to alleviate the overfitting and underfitting issues.

REFERENCES

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 06–11 Aug 2017.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3442–3455, 2021.

Mohammad Mahdi Bejani and Mehdi Ghatee. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, 54(8):6391–6438, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jason Brownlee. *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine Learning Mastery, 2018.

Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning. *arXiv preprint arXiv:2203.10036*, 2022.

Wilfrid J Dixon and Alexander M Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Frage: Frequency-agnostic word representation. *Advances in neural information processing systems*, 31, 2018.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Joseph L Hodges. A bivariate sign test. *The Annals of Mathematical Statistics*, 26(3):523–527, 1955.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tom Kocmi and Ondřej Bojar. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 379–386, 2017.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pp. 177–180, 2007.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Haidong Li, Jiongcheng Li, Xiaoming Guan, Binghao Liang, Yuting Lai, and Xinglong Luo. Research on overfitting of deep learning. In *2019 15th International Conference on Computational Intelligence and Security (CIS)*, pp. 78–81. IEEE, 2019.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on Machine Learning*, pp. 5958–5968. PMLR, 2020.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of NAACL-HLT*, pp. 1162–1172, 2019.

David MW Powers. Applications and explanations of zipf's law. In *New methods in language processing and computational natural language learning*, 1998.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3088–3095, 2020.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anshika Rastogi, and Shikha Jain. Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)*, pp. 162–167. IEEE, 2017.

Xu Sun, Weiwei Sun, Shuming Ma, Xuancheng Ren, Yi Zhang, Wenjie Li, and Houfeng Wang. Complex structure leads to overfitting: A structure regularization decoding method for natural language processing. *arXiv preprint arXiv:1711.10331*, 2017.

Dusan Varis and Ondřej Bojar. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8246–8257, 2021.

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

Robert Wolfe and Aylin Caliskan. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 518–532, 2021.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3977–3989, 2020.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29–45, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Jiajun Zhang, Chengqing Zong, et al. Deep neural networks in machine translation: An overview. *IEEE Intell. Syst.*, 30(5):16–25, 2015.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*, 2018.

Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6934–6944, 2020.

A Appendix

We present experimental results of the additional factors here, prefixing the name of each factor on the caption.

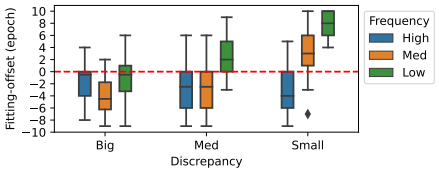


Figure 12: The Language: Fitting-offset of tokens grouped by *discrepancy* and *frequency*.

Table 5: The Language: Potential-gain for each category grouped by *frequency* and *discrepancy*.

Discrepancy	**Frequency**		
	High	**Med**	**Low**
Big	67.14 +0.21	27.14 -0.27	9.90 +0.20
Med	69.91 +0.07	47.08 -0.08	14.67 **+0.78**
Small	62.99 +0.33	81.85 +0.33	66.38 **+1.35**

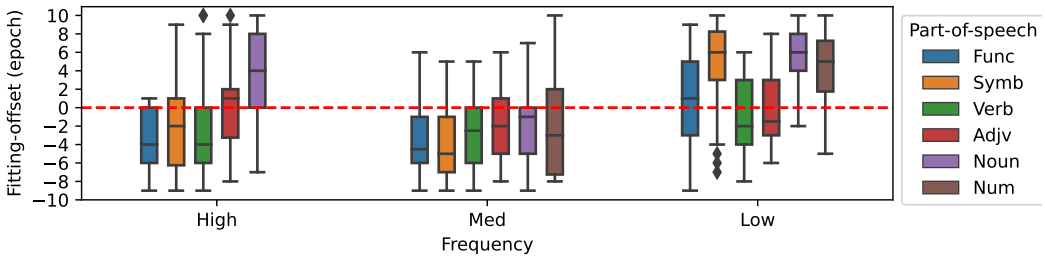


Figure 13: The Language: Fitting-offset of tokens grouped by *frequency* and *POS*.

Table 6: The Language: Potential-gain for tokens grouped by *frequency* and *POS*.

Frequency	**Function**	**Symbol**	**Number**	**Verb**	**Adj/Adv**	**Noun**
High	64.8 +0.15	80.08 **+0.86**	nan +nan	64.51 **+1.20**	38.24 **+1.93**	49.15 **+4.01**
Med	49.82 +0.13	66.23 **+1.67**	69.43 **+1.29**	47.03 +0.38	55.78 +0.34	61.54 +0.21
Low	11.12 **+1.75**	36.09 **+5.98**	72.28 **+1.29**	30.92 **+0.57**	37.76 **+0.78**	42.13 **+1.20**

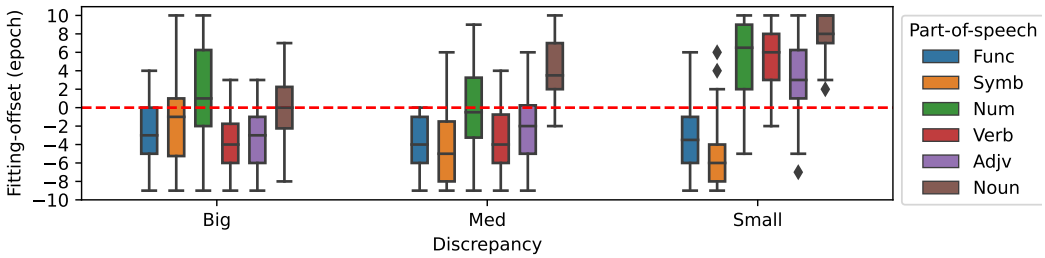
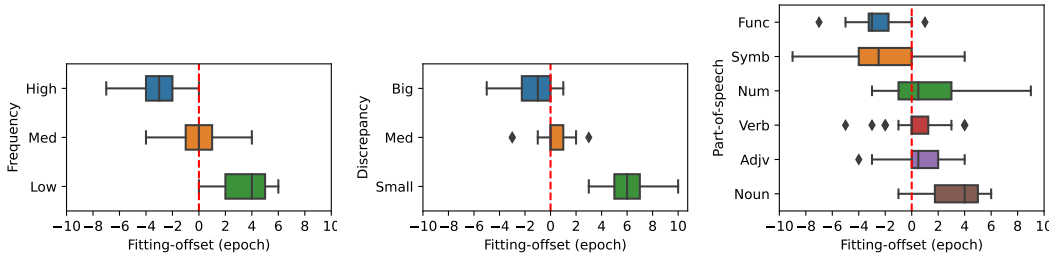


Figure 14: The Language: Fitting-offset of tokens grouped by *discrepancy* and *POS*.

Table 7: The Language: Potential-gain for tokens grouped by *discrepancy* and *POS*.

Discrepancy	Function	Symbol	Number	Verb	Adj/Adv	Noun
Big	55.19 +0.10	77.67 **+0.58**	28.34 **+1.78**	20.76 +0.04	15.50 -0.20	16.61 +0.19
Med	63.84 +0.03	75.72 **+1.05**	38.57 **+2.92**	32.44 -0.10	29.21 +0.35	26.49 **+0.92**
Small	71.99 +0.20	75.22 **+1.32**	86.89 **+1.08**	65.93 **+1.14**	72.88 **+0.81**	72.12 **+1.20**

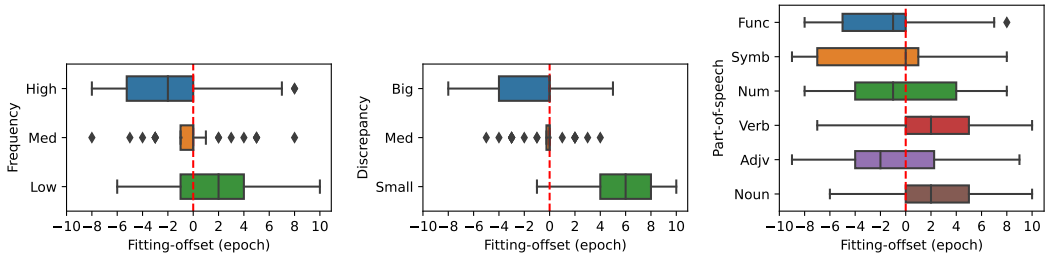


(a) Fitting-offset of tokens grouped by *frequency*.

(b) Fitting-offset of tokens grouped by discrepancy.

(c) Fitting-offset of tokens grouped by *parts-of-speech*.

Figure 15: The Model Size: Fitting-offset of German tokens evaluated on News English-German with *big* model.

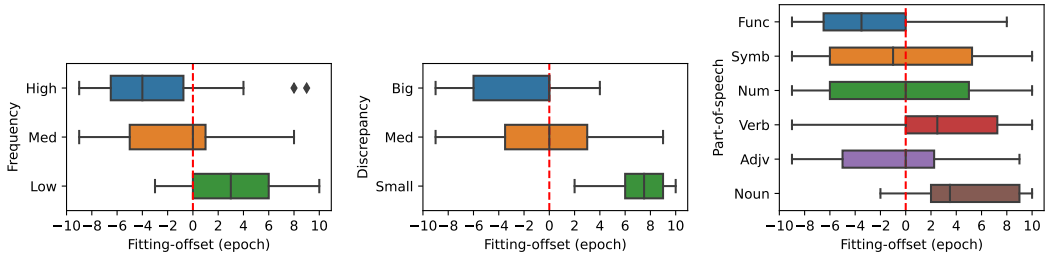


(a) Fitting-offset of tokens grouped by *frequency*.

(b) Fitting-offset of tokens grouped by *discrepancy*.

(c) Fitting-offset of tokens grouped by *parts-of-speech*.

Figure 16: The Domain: Fitting-offset of tokens evaluated on *Europarl* English-German ($250,000$ samples).



(a) Fitting-offset of tokens grouped by *frequency*.

(b) Fitting-offset of tokens grouped by *discrepancy*.

(c) Fitting-offset of tokens grouped by *parts-of-speech*.

Figure 17: The Data Scale: Fitting-offset of tokens evaluated on Europarl English-German ($500,000$ *samples*)..