Structural Biases for Improving Transformers on Translation into Morphologically Rich Languages

Paul Soulos[*1], **Sudha Rao**[2], **Caitlin Smith**[*1], **Eric Rosen**[*1], **Asli Celikyilmaz**[2], **R. Thomas McCoy**[*1], **Yichen Jiang**[*3], **Coleman Haley**[*1], **Roland Fernandez**[2], **Hamid Palangi**[2], **Jianfeng Gao**[2], **Paul Smolensky**[1,2]
[1]Johns Hopkins University [2]Microsoft Research, Redmond [3]UNC Chapel Hill
{psoulos1, csmit372, erosen27, tom.mccoy, chaley7}@jhu.edu
{sudha.rao, aslicel, rfernand, hpalangi, jfgao, psmo}@microsoft.com
{yichenj}@cs.unc.edu

Abstract

Machine translation has seen rapid progress with the advent of Transformer-based models. These models have no explicit linguistic structure built into them, yet they may still implicitly learn structured relationships by attending to relevant tokens. We hypothesize that this structural learning could be made more robust by explicitly endowing Transformers with a structural bias, and we investigate two methods for building in such a bias. One method, the TP-Transformer, augments the traditional Transformer architecture to include an additional component to represent structure. The second method imbues structure at the data level by segmenting the data with morphological tokenization. We test these methods on translating from English into morphologically rich languages, Turkish and Inuktitut, and consider both automatic metrics and human evaluations. We find that each of these two approaches allows the network to achieve better performance, but this improvement is dependent on the size of the dataset. In sum, structural encoding methods make Transformers more sample-efficient, enabling them to perform better from smaller amounts of data.

1 Introduction

The task of machine translation has seen major progress in recent times with the advent of large-scale Transformer-based models (e.g., Vaswani et al., 2017; Dehghani et al., 2019; Liu et al., 2020a). However, there has been less progress on language pairs that specifically involve morphologically rich languages. Moreover, although there has been previous work that builds linguistic structure into translation models to deal with morphological complexity (Sennrich and Haddow, 2016; Dalvi et al., 2017; Matthews et al., 2018), to the best to our knowledge there has not been work that applies such strategies to large-scale Transformer-based models. We hypothesize that providing Transformers access to structured linguistic representations can significantly boost their performance on translation into languages with complex morphology that encodes linguistic structure.

In this work, we investigate two methods for introducing such structural bias into Transformer-based models. In the first method, we use the TP-Transformer (TPT) (Schlag et al., 2019), in which a traditional Transformer is augmented with Tensor Product Representations (TPRs) (Smolensky, 1990) (§ 2). At a high level, TPRs use a composition of *roles*

[*]Work partially done while at Microsoft Research.

English:	I want people to raise their hands who are in favour of the motion to report progress. (17)
Turkish:	Ilerleme raporunun talep edilmesinden yana olanların el kaldırmalarını istiyorum. (9)
Inuktitut:	isaaquvaksi taikkua nangmaksaqtut pigiaqtitausimajumut nuqqarumaliqtu. (5)

Table 1: Parallel sentence in English, Turkish, and Inuktitut. The number of words in each translation (marked in parentheses) is indicative of their information density and, hence, their morphological complexity.

and *fillers* where roles encode structural information (e.g., the part-of-speech of a word) and fillers encode the content (e.g., the meaning of a word). This enables learned internal structured representations. In the second method, we encode structure external to the model by segmenting training data using morphological tokenization (§3): morphological segmentation is done by existing parsers prior to training the Transformer. Since all neural models that operate over sequences tokenize the training data, through this method, we aim to answer the question of whether linguistically-informed tokenization that respects morphological structure can be helpful in processing morphologically-rich languages. Through the use of TPT, we aim to examine whether enabling a Transformer to learn its own structured *internal* representations will help it learn linguistic structure including structure which is encoded morphologically in morphologically-rich languages. Unlike the morphological tokenizer, the TPT architecture is language-agnostic and can be used on arbitrary datasets without feature engineering. We further investigate how the biases of these two approaches work together. We experiment on the task of translating from English into two morphologically rich languages: Turkish and Inukitut (Inuit; Eastern Canada). For Turkish, we train on several different dataset sizes from Open Subtitles (1.4M, 5M and 36M), a spoken-language domain, and also fine-tune on SETimes (200K), a news-wire domain. For Inuktitut, we train on the Nunavut Hansard Corpus (1.3M). We test models' performance using both an automatic metric and human evaluation (§5).

In the English to Turkish translation task, we find that the TP-Transformer beats the Transformer when evaluated for nuances such as morphology, word-order and subject/object-verb agreement. TPT provides a significant improvement on small datasets segmented with language agnostic BPE ($\sim$ 1 BLEU for Open Subtitles 1.4m and $\sim$ 2.5 BLEU for Hansard) and a more modest improvement on larger datasets (0.16 BLEU for Open Subtitles 5m and 0.36 BLEU for Open Subtitles 36m). Using morphologically segmented data helps substantially with models that are trained on small datasets. This is true for both pre-training (Open Subtitles 1.4m and Inuktitut Hansard), as well as models that are trained on large datasets and later finetuned using a smaller dataset (SETimes). This suggests that the method of encoding structure directly in the training data helps substantially with sample efficiency and transfer learning. When our two techniques are used together, we achieve an **8 BLEU** improvement over the state of the art on translation into Inuktitut (Joanis et al., 2020).

In order to better understand our models, we conduct detailed analysis, including error analysis, on sample outputs from different model variations (Appendix G). We also separate results out into different bins as defined by the morphological density of the target outputs to understand how results vary with morphological complexity §6. We find that morphological tokenization is strongly correlated with improved performance on complex sentences.

2 Using the TP-Transformer

The TP-Transformer (TPT) was introduced by Schlag et al. (2019) to improve performance on mathematical problem solving, a highly symbolic task. The model introduces an additional component to the attention mechanism which represents relational structure. In addition to the standard key K, query Q, and value V vectors used in attention, they introduce the role vector R.

Let the input for token $i \in 1, .., N$ at layer l be represented as X_i^l. For head h, the vectors are:

$$Q_i^{lh} = X_i^l W_q^{lh} + b_q^{lh} \quad K_i^{lh} = X_i^l W_k^{lh} + b_k^{lh} \quad V_i^{lh} = X_i^l W_v^{lh} + b_v^{lh} \quad R_i^{lh} = X_i^l W_r^{lh} + b_r^{lh}$$

The output of soft attention $\bar{V}_i^{lh}$ is: $\bar{V}_i^{lh} = \sum_{t=1}^{N} \text{softmax}(\frac{Q_i^{lh} K_t^{lh}}{\sqrt{d_k}})^\top V_t^{lh}$

In a Tensor Product Representation, role vectors are bound to their corresponding filler vectors by the tensor product $\otimes$ or some compression of it: in the TPT, we use the compression of discarding the off-diagonal elements, resulting in the elementwise or Hadamard product $\odot$. The query Q_i^{lh} is interpreted as probing for a filler for the role R_i^{lh}, so the output of attention $\bar{V}_i^{lh}$ is taken to be the filler of that role; thus for the original TPT, this yielded: $Z_i^{lh} = \bar{V}_i^{lh} \odot R_i^{lh}$.

The role vector R is intended to act as a structural encoding independent of that structure's content (which is encoded in $\bar{V}$). We hypothesize that, by disentangling structure and content in this way, we can improve the model's ability to place familiar linguistic units in novel structures (e.g., using a suffix with a word stem that never had that suffix during training). Such structural flexibility is crucial for morphologically-rich languages.

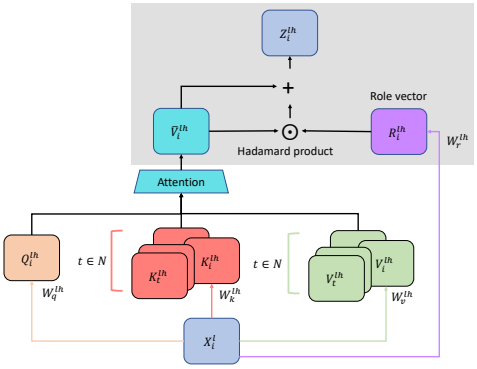


Figure 1: Architectural diagram of TPT attention mechanism. Highlighted section shows the additional components added to standard Transformer attention.

We make two modifications to the TPT used in Schlag et al. (2019)[1]. First, we use relative position embeddings (Shaw et al., 2018). We also use a residual connection to produce gradients that are not zero; $\bar{V} \odot R$ is a multiplicative interaction, so values of R near 0 will produce activation values and gradients of 0. This is similar to the model detail in Perez et al. (2017) Section 7.2. A schematic of our attention is shown in Figure 1. The rest of the architecture follows the standard residual connections and encoder-decoder architecture defined in Vaswani et al. (2017)

3 Using morphological segmentation

Our target languages, Turkish and Inuktitut, both exhibit a high degree of morphological complexity. Words in both languages consist of a root followed by potentially many suffixes, each of which may have multiple surface forms.

We used two methods of subword tokenization: one utilizing a type of character-level byte-pair encoding (Gage, 1994), and one incorporating morphological parsing plus byte-pair encoding. The first method (which we label 'BPE') used SentencePiece (Kudo and Richardson, 2018), a tokenizer that builds subword tokens using a combination of byte-pair encoding and unigram language modeling. BPE relies only on character frequencies and incorporates no morphological information.

Language	Segmented Word
Turkish	anla-t-ma-yacak
Gloss	understand-CAUS-NEG-FUT
English	will not tell
BPE	anlat-mayacak
Inuktitut	miv-vi-liar-uma-lauq-tur-uuq
Gloss	land-place-go-want-PAST-3S-say.3S
English	He said he wanted to go to the landing strip.
BPE	mivvi-lia-ruma-lau-qturuuq

Table 2: Morpheme breakdown, gloss, English, and BPE tokenization of Turkish and Inuktitut morphologically complex words

[1]Code available at `https://github.com/psoulos/tpt`

The second method (which we call 'morphological tokenization') incorporated morphological information by parsing all words (i.e. breaking them up into their composite morphemes) in our morphologically complex target languages before tokenizing them. For Turkish, we used the morphological parser from Zemberek (Akın and Akın, 2007), an open-source Turkish NLP toolkit. Zemberek uses sentence-level disambiguation to produce the most likely parse of each word given its sentential context. For Inuktitut, we used the morphological parsing method adopted by Joanis et al. (2020), incorporating a symbolic parser with a neural parser backoff. See Appendix D for implementational details on morphological segmentation.

The differences in how these tokenizers divide multi-morphemic Turkish and Inuktitut words into subwords are illustrated in Table 2. The boundaries determined by BPE do not reflect the internal morphological structure of these words.

4 Dataset description

4.1 English-Turkish data

For pretraining of the English-Turkish translation model, we used the Open Subtitles corpus (Lison and Tiedemann, 2016). This corpus consists of a large number of aligned pairs of subtitles from film and television. In order to test the effect of dataset size on model performance, we utilized three splits of this corpus: the full-size corpus, a sample of five million sentence pairs, and a sample of approximately one million sentence pairs. For fine-tuning of the English-Turkish model, we used the South-East European Parallel (SETimes) Corpus. SETimes is a collection of short written news stories in ten languages. For this task, we used the subset of this corpus that was used for the WMT 2018 English-Turkish shared translation task (Bojar et al., 2018).

4.2 English-Inuktitut data

Corpus	Training	Validation	Test
Open Subtitles 36m	28,694,211	3,586,776	3,586,777
Open Subtitles 5m	4,000,000	500,000	500,000
Open Subtitles 1.4m	1,300,000	65,000	65,000
SETimes	207,678	3,007	3,000
Nunavut Hansard	1,312,791	5,494	6,181

Table 3: Number of training, validation, and test samples in the different datasets.

Like Turkish, Inuktitut is a morphologically complex language. Words may consist of a root, a prefix, and potentially many suffixes. Table 2 contains an example of a multi-morphemic Inuktitut word. For training of the English-Inuktitut translation model, we used the Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020), the only sizable publicly available bilingual corpus. The dataset consists of over one million aligned sentence pairs from government proceedings. The size of the dataset splits are reported in Table 3.

5 Experimental Results

We aim to answer the following research questions (RQ) through our experimentation:

1. Do either or both of our structural methods improve translation?
2. If so, how does that advantage interact with:
   (a) Training data quantity?
   (b) Transfer learning?
   (c) Morphological richness of language?

As a baseline, we trained the standard Transformer model (Vaswani et al., 2017) with the addition of relative position embeddings (Shaw et al., 2018). Model training details and computing resources can be found in Section 1 and 2 of the supplementary materials. For each model, we used either byte pair encoding (BPE) (Sennrich et al., 2016) or morphological tokenization as described in §3. In order to see how our changes relate to sample efficiency, we

vary the size of the subset of the Open Subtitles dataset used for training. We used the SETimes dataset to finetune these models to test whether either structural bias improves transfer learning. We also trained models on the Inuktitut dataset to compare the results from languages with differing morphological richness.

5.1 Automatic Metric Results

	Transformer	TP-Transformer
1.4m	7.5 ±.43	8.44 ±.25
1.4m morph	16.63 ±.19	16.89 ±.07
5m	18.70	18.86
5m morph	18.84	19.19
36m	20.95	21.31
36m morph	21.05	21.32

Table 4: BLEU scores on the test set of Open Subtitles separated by training set size and tokenization method. For the 1.4m runs, we show the mean and standard deviation of three randomly initialized models. The larger datasets only have one run each due to computational resource reasons.

Table 4 shows the test set BLEU[2] scores for the different size splits of the Open Subtitles dataset (Research Question RQ2a). For the smallest data split of 1.4m samples, TPT provides almost 1 BLEU improvement over a standard Transformer. Using a morphological tokenization provides an 8 BLEU improvement on the small split. Using TPT with morphologically tokenized data does not provide any additional benefit on the 1.4m split. For the two larger splits, TPT (across columns) and morphological parsing (across rows) provides minor improvements (0.1–0.36 BLEU), and this improvement becomes more modest when both are combined (top left cell to bottom right cell) (0.49 BLEU on the 5m split and 0.37 on the full 36m split). Next, in order to analyze whether either structural bias helps with transfer learning (RQ2b), we take the best performing models shown in Table 4 and finetune them on the SETimes dataset.

	Transformer	TP-Transformer
5m	14.19	14.25
5m morph	15.16	15.39
36m	16.77	17.01
36m morph	18.35	18.82

Table 5: BLEU scores on the test set of SETimes from models pretrained on OpenSubtitles (5m) and finetuned on SETimes (200K) divided by training set size and tokenization.

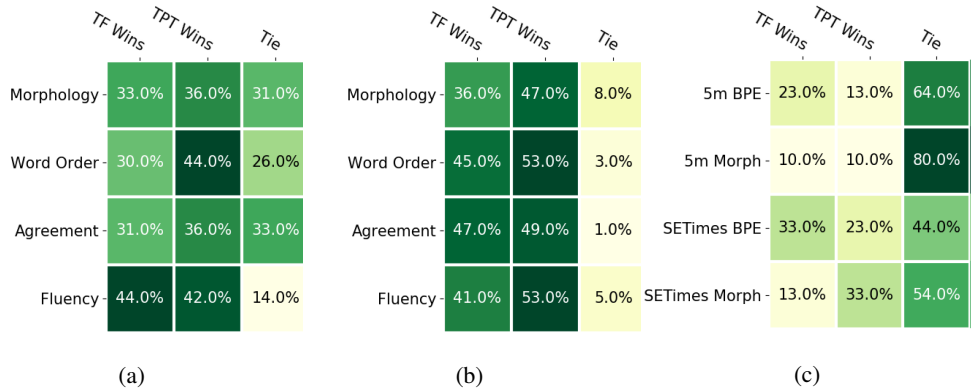
The BLEU scores for these finetuned models can be seen in Table 5. There is a large increase in BLEU score across rows between models that use either BPE encoding or morphological tokenization. This provides further evidence for the findings from the 1.4m split in Table 4 that morphological tokenization provides a large improvement in low data regimes. While morphological tokenization does not provide much of an improvement during large-scale pretraining, it is beneficial for transfer learning on a smaller domain.

	Transformer	TP-Transformer
BPE	18.56 ±1.92	21.12 ±.70
Morphological	26.05 ±.90	28.3 ±.50

Table 6: BLEU scores on the test set of Inuktitut divided by tokenization. We show the mean and standard deviation of three randomly initialized models.

In addition to Turkish, we trained models on the Inuktitut dataset described in §4.2 to understand the variance of model performance by the morphological richness of languages (RQ2c). We trained models using both data tokenized by BPE encoding as well as by an Inuktitut morphological parser. The results are shown in Table 6. As we saw on both the 1.4m Open Subtitles split and SETimes,

[2]We calculated BLEU using SacreBLEU (Post, 2018) and the signature is "BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.0". All models were also tested with CHRF (Popović, 2015) and the results can be found in Appendix E.



(a) (b) (c)

Figure 2: Human judgment results: (a) Comparison between Transformer (TF) and TPT on different criteria when trained on Open Subtitles (5m) using BPE encoding. (b) Comparison between Transformer and TPT when trained on Open Subtitles (5m) using morphologically segmented data. (c) Comparison between Transformer (TF) and TPT on meaning preservation when trained on different datasets.

morphological tokenization provides a huge improvement in BLEU. TPT provides a large average improvement regardless of the tokenization scheme, although the BPE Tranformer in particular has a high variance and is sensitive to random initialization. Inuktitut is more morphologically complex than Turkish across several measures of morphological complexity[3] and it is possible that TPT models perform better with more complex morphology. For example, compare the improvements from using TPT over a standard transformer for BPE on the Open Subtitles 1.4m split and Inuktitut. TPT provides $\sim$1 BLEU improvement on Turkish, and this improvement increases to $\sim$2.5 on Inuktitut. The previous state-of-the-art on the Hansard dataset is 20.3 BLEU on the test set (Joanis et al., 2020). Both methods proposed in this paper improve on that, and together they improve the state-of-the-art by 8 BLEU.

5.2 Human-based Evaluation Results

The BLEU scores in the previous section give us a single number summarizing the quality of our translations. We now evaluate some of the finer-grained characteristics of the outputs. We focus on four aspects of the output that are likely to benefit from more robust encodings of structure: morphology, word-order, subject-verb agreement and fluency.

We use Amazon Mechanical Turk to get human judgements. We perform a comparative study where we show annotators two Turkish translations from the transformer and the TPT models trained on the 5m Open Subtitles split. We do not show the English source sentence since the four criteria of evaluation in this study does not require looking at the source sentence. We collect three annotations per comparison and use only those instances where at least 2 out of the 3 annotators agree on the same answer. We collect annotations on 180 instances for each of the two comparative studies. See Appendix F for the questions asked to annotators.

Figure 2a shows the result of this comparison when we use BPE encoding to tokenize the data whereas Figure 2b shows the result of this comparison when we use morphological segmenter to tokenize the data. Under BPE encoding, we find that TPT has slightly less morphological and agreement errors and has significantly less word-order issues. This suggests that the structural bias introduced by the TPT helps in forming sentences that are overall morphologically better

[3]Using the parallel test sets from Mielke et al. (2019), we measured a type-token ratio of 0.42 for Inuktitut and 0.19 for Turkish, as well as a relative entropy of word structure of 1.75 for Inuktitut and 1.21 for Turkish

formed. On the other hand, annotators find translations from the Transformer to be slightly more fluent than those from the TPT. Under morphologically segmented data, annotators find translations from TPT are significantly better than the Transformers in morphological form and word-order and slightly better in subject-verb agreement, providing further evidence that the structural bias introduced by the TPT is helpful. Moreover, annotators also find translations from TPT in this case to be more fluent than those from the Transformer.

We perform an additional study to understand which of the two model translations best preserves the meaning of the English source sentence. We ask an expert, a linguistically-trained native Turkish speaker, to annotate 30 instances each from eight model outputs (5m Open Subtitles BPE & morphologically tokenized, SETimes BPE & morphologically tokenized for both Transformer and TPT). We show them the English sentence and two Turkish translations. We ask them "*Grammatical issues aside, which of the two translations better preserves the meaning of the English sentence?*" and let them choose from A, B or Both preserve equally. Figure 2c shows the results of this study. In the Open Subtitles dataset, we find the difference between Transformer and TPT performance is too small under both BPE encoding and morphological segmentation. In the SETimes dataset, we find the same trend under BPE encoding. Only under morphologically segmented data in SETimes, TPT significantly wins over Transformer. These results show that when we include the English source sentence, it is inconclusive if TPT or Transformer is better. This suggests that although TPT improves the ability to compose Turkish text (as found by the first study), it does not affect the ability to determine which Turkish output should go with a given English input.

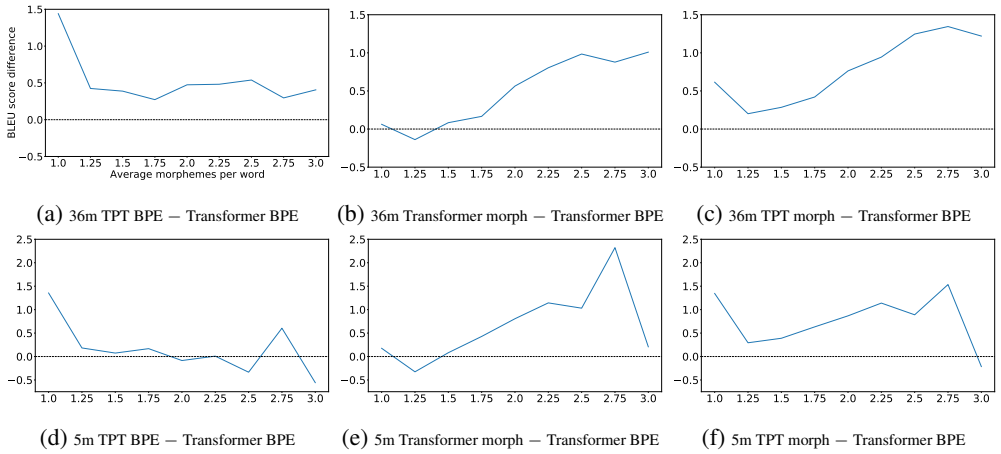
6 Morphological density analysis

Given the rich morphology of the target languages, we are interested in whether either structural bias or morphological segmentation improves performance on more morphologically complex sentences. To answer this question, we used our Turkish morphological segmenter on sequences from the test set and binned sentences based on the average morphemes per word in a sentence. For example, a long sentence with simple words that are all a single morpheme would have an average morpheme per word of 1, whereas a sentence that is made of complex words would have a larger average morpheme per word. We then calculated the BLEU score for each of these buckets so that we could see if our models performed better on sentences that are morphologically complex.

The results are shown in Figure 3. On the 36m training set (top row), both of our methods provide an improvement at almost every morpheme density. Comparing TPT to a standard Transformer, Figure 3a shows a relatively consistent improvement of around 0.4 BLEU with a large increase for simple sentences. Comparing standard transformers with morphological parsing against BPE, Figure 3b shows that as the morphological complexity of sequences increases, the model using morphological tokenization improves over BPE tokenization. The same trend is visible when comparing TPT with morphological tokenization with a standard transformer using BPE tokenization (Figure 3c), except the magnitude of the increase is greater.

The morphological analysis on the 5m training set (bottom row) is less conclusive. TPT does not appear to have any impact as the morphological density increases (Figure 3d). Morphological tokenization shows a similar upward trend as on the 36m dataset, but this improvement disappears suddenly at 3.0 morphemes per word (Figure 3e). As the morphological density increases, the number of samples for each bucket on the test set decreases, so it is possible that the sudden drop is the result of too few samples.

Our results also show some correspondence with the overall morphological complexity of the dataset. We computed a modified version of the C_D measure (the "relative entropy of word structure") from Bentz et al. (2016), as we found it to be the most robust to the meaning



(a) 36m TPT BPE − Transformer BPE (b) 36m Transformer morph − Transformer BPE (c) 36m TPT morph − Transformer BPE

(d) 5m TPT BPE − Transformer BPE (e) 5m Transformer morph − Transformer BPE (f) 5m TPT morph − Transformer BPE

Figure 3: BLEU score differences between models on the Turkish Open Subtitles 36m (top row) and 5m (bottom row) training sets bucketed by morphological density (average number of morphemes per word in a sentence).

variations between corpora (Supplementary materials section 4). Higher values of this measure correspond to more regular structure/information in words, and thus, greater morphological complexity. We computed the measure over the first 100,000 characters of the test set of each dataset. We computed C_D as 1.89 for the Hansard dataset, while the Turkish datasets ranged from C_D 1.45-1.49. This corresponds to the relatively large increase in BLEU seen for Inuktitut.

7 Related Work

Translating into Morphologically-rich languages Previous work has leveraged morphology for translating into morphologically-rich languages. Turhan (1997) uses a recursive symbolic system to translate from English into Turkish including a morphological generator. Ataman et al. (2020) use hierarchical latent variable models to model both character and morpheme level statistics for translating into morphologically rich languages (Arabic, Czech, Turkish) with GRUs. Passban et al. (2018a) introduce a character-level neural machine translation model for translating into morphologically rich languages which incorporates a morphology lookup table into the decoder whereas Passban et al. (2018b) propose a subword-level model that uses separate embedding for stem and affix. Joanis et al. (2020) introduced the dataset that we use for Inuktitut and also explored using morphological segmentation for alignment as well as neural and statistical machine translation. This work was followed up by Knowles et al. (2020) who introduce additional methods techniques on the Inuktitut dataset. Roest et al. (2020) and Scherrer et al. (2020) also investigated morphological segmentation in Inuktitut in addition to data augmentation and pretraining.

Using Transformer-based models for translation In recent times, there have been several work that use variations of Transformer (Vaswani et al., 2017) model for the task of machine translation. Chen et al. (2018) combine the power of recurrent neural network and transformer. Dehghani et al. (2019) introduce universal transformers as a generalization of transformers whereas Deng et al. (2018) combine transformer architecture with several other techniques such as BPE, back translation, data selection, model ensembling and reranking. Bugliarello and Okazaki (2020) incorporate syntactic knowledge into transformer model to show improvements on English to German, Turkish and Japanese translation tasks. Currey and Heafield (2019) introduce two methods to incorporate English syntax when translating from English into other

languages with Transformers. Liu et al. (2020b) introduce mBART, an auto-encoder pretrained on large-scale monolingual corpora and show gains on several languages.

Using TPRs TPRs have gained traction recently with the interest in neurosymbolic computation to achieve out-of-domain generalization. They have been used in a variety of domains, including mathematical problem solving (Schlag et al., 2019), reasoning (Schlag and Schmidhuber, 2018), image captioning (Huang et al., 2018), question-answering (Palangi et al., 2018), and program synthesis (Chen et al., 2020). A separate line of work uses TPRs as an interpretation tool to understand representations in networks that do not explicitly use TPRs (McCoy et al., 2019; Soulos et al., 2020).

8 Conclusion

We investigated two methods for improving translation into morphologically rich languages with Transformers. The TP-Transformer adds an additional component to Transformer attention to represent relational structure. This model had the largest improvement on smaller datasets and modest improvement on larger datasets. We also investigated morphological tokenization which had substantial improvements on small datasets and transfer learning. When used together, our methods improve on the state of the art for translation from English into Inuktitut by 8 BLEU. The models were analyzed by human evaluators to tease apart different dimensions along which our models excel; TP-Transformer had fewer morphological, word-order, and agreement issues. We analyzed the performance of our networks under varying morphological complexity and found that morphological tokenization provides a large benefit for more complex sentences.

References

Akın, A. A. and Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure*, 10.

Ataman, D., Aziz, W., and Birch, A. (2020). A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. In *ICLR 2020*.

Bentz, C., Ruzsics, T., Koplenig, A., and Samardžić, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Bugliarello, E. and Okazaki, N. (2020). Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627.

Chen, K., Huang, Q., Palangi, H., Smolensky, P., Forbus, K., and Gao, J. (2020). Mapping natural-language problems to formal-language solutions using structured neural representations. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1566–1575. PMLR.

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Schuster, M., Shazeer, N., Parmar, N., et al. (2018). The best of both worlds: Combining recent advances in neural machine

translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

Currey, A. and Heafield, K. (2019). Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Florence, Italy. Association for Computational Linguistics.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., and Vogel, S. (2017). Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, L. (2019). Universal transformers. In *International Conference on Learning Representations*.

Deng, Y., Cheng, S., Lu, J., Song, K., Wang, J., Wu, S., Yao, L., Zhang, G., Zhang, H., Zhang, P., et al. (2018). Alibaba's neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376.

Gage, P. (1994). A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(50):1469–1484.

Huang, Q., Smolensky, P., He, X., Deng, L., and Wu, D. (2018). Tensor product generation networks for deep NLP modeling. 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 1263–1273, New Orleans, Louisiana. Association for Computational Linguistics.

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D., and Micher, J. (2020). The Nunavut hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Knowles, R., Stewart, D., Larkin, S., and Littell, P. (2020). NRC systems for the 2020 Inuktitut-English news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 156–170, Online. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (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.

Lison, P. and Tiedemann, J. (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).

Liu, X., Duh, K., Liu, L., and Gao, J. (2020a). Very deep transformers for neural machine translation. In *arXiv:2008.07772 [cs]*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020b). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Matthews, A., Neubig, G., and Dyer, C. (2018). Using morphological knowledge in open-vocabulary neural language models. 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 1435–1445, New Orleans, Louisiana. Association for Computational Linguistics.

McCoy, R. T., Linzen, T., Dunbar, E., and Smolensky, P. (2019). RNNs implicitly implement tensor-product representations. In *International Conference on Learning Representations*.

Mielke, S. J., Cotterell, R., Gorman, K., Roark, B., and Eisner, J. (2019). What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Palangi, H., Smolensky, P., He, X., and Deng, L. (2018). Question-answering with grammatically-interpretable representations. In *AAAI*.

Passban, P., Liu, Q., and Way, A. (2018a). Improving character-based decoding using target-side morphological information for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 58–68.

Passban, P., Way, A., and Liu, Q. (2018b). Tailoring neural architectures for translating from morphologically rich languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3134–3145.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. (2017). Film: Visual reasoning with a general conditioning layer. *CoRR*, abs/1709.07871.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Roest, C., Edman, L., Minnema, G., Kelly, K., Spenader, J., and Toral, A. (2020). Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Scherrer, Y., Grönroos, S.-A., and Virpioja, S. (2020). The University of Helsinki and aalto university submissions to the WMT 2020 news and low-resource translation tasks. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1129–1138, Online. Association for Computational Linguistics.

Schlag, I. and Schmidhuber, J. (2018). Learning to reason with third order tensor products. *Advances in neural information processing systems*, 31:9981–9993.

Schlag, I., Smolensky, P., Fernandez, R., Jojic, N., Schmidhuber, J., and Gao, J. (2019). Enhancing the transformer with explicit relational encoding for math problem solving. *arXiv preprint arXiv:1910.06611*.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604, Stockholmsmässan, Stockholm Sweden. PMLR.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216.

Soulos, P., McCoy, R. T., Linzen, T., and Smolensky, P. (2020). Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 238–254, Online. Association for Computational Linguistics.

Turhan, C. K. (1997). An english to turkish machine translation system using structural mapping. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, page 320–323, USA. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, abs/1706.03762.

Appendix

A Model Training Parameters

Both the standard Transformer and the TP-Transformer (TPT) use 6 layers and 8 heads per layer. TPT has key/value/query/role dimensions of 64, whereas the standard Transformer has key/value/query dimensions of 80. The reason for this increase is so that the resulting models match in terms of parameter count, and we add parameters are the most homologous area. The standard Transformer has 74,375,936 parameters, and the TP-Transformer has 74,385,152 parameters. Both networks use a token dimension of 512, a feedforward dimension of 2048, and 32 relative positioning buckets Shaw et al. (2018). The input vocabulary size is 50,000. We set a training batch size of 80 per GPU and used the Adafactor Shazeer and Stern (2018) optimizer with square root learning rate decay. Throughout the model, we used a commonly used dropout rate of .1.

B Computing Resources

The models were all trained with 8 Tesla V100 GPUs. The models trained on the small Hansard and Open Subtitles 1.4m datasets converged in about 8 hours. The larger Open Subtitles 5m models coverged in around 40 hours, and the Open Subtitles 32m models coverged in 15 days.

C Corpora Morphological Complexity

Studies have considered what corpus-based measures are correlated with linguistic measures of morphological complexity. Most notably, Bentz et al. (2016) found several corpus-based measures that correlate strongly with complex morphological typology. This measure computes the regularity of structure within words by taking the character-level entropy of the corpus and subtracting that from the entropy of a "masked" version of the corpus, where all non-whitespace characters have been replaced with random samples from the uniform distribution over the characters in the corpus. Rather than the approximation used in Bentz et al. (2016) for character-level entropy, we directly computed the character-level Shannon's entropy using a James-Stein shrinkage estimator as in Hausser and Strimmer (2009).

D Morphological parser process

For each target language, its parser was used to insert morpheme boundaries into all multi-morphemic words in the dataset. Due to the comparatively low level of morphological complexity of the English source data, no parsing of English words was conducted. From here, each SentencePiece tokenizer's vocabulary was built over a dataset's training data (both the source and target language) with a target size of 50,000 vocabulary items. SentencePiece allows the user to specify special characters that cannot be crossed when constructing subword tokens, both during training of the tokenizer and during tokenization of a sentence. The symbol used to represent morpheme boundaries was specified as such a special symbol. As a result, morpheme boundaries in Turkish and Inuktitut (as identified by their respective parsers) always served as subword token boundaries.

Each SentencePiece tokenizer's vocabulary was built over a dataset's training data (both source and target language) with a target size of 50,000 vocabulary items. This tokenization method (which we label simply 'BPE') relies only on character frequencies and incorporates no morphological information, so many multi-morphemic words may each be assigned to a single token, and there is no guarantee that a word's subword boundaries align with its morpheme boundaries.

E CHRF Results

The same models used to measure BLEU scores are also tested using CHRF (Popović, 2015). The results are shown in Tables 7, 8, and 9.

	Transformer	TP-Transformer
1.4m	.351 $\pm$.005	.365 $\pm$.004
1.4m morph	.438 $\pm$.001	.440 $\pm$.001
5m	.461	.463
5m morph	.467	.469
36m	.486	.488
36m morph	.490	.492

Table 7: CHRF scores on the test set of Open Subtitles separated by training set size and tokenization method. For the 1.4m runs, we show the mean and standard deviation of three randomly initialized models. The larger datasets only have one run each due to computational resource reasons.

	Transformer	TP-Transformer
5m	.502	.502
5m morph	.509	.514
36m	.532	.537
36m morph	.540	.543

Table 8: CHRF scores on the test set of SETimes from models pretrained on OpenSubtitles (5m) and finetuned on SETimes (200K) divided by training set size and tokenization.

F Annotator Questions

We ask annotators the following questions:

Morphology: "*Which of the two sentences has more morphological issues (i.e. incorrect suffixes)?*" and let annotators choose from A, B, Both or None.

Word-order: "*Which of the two sentences has word-order issues?*" and let annotators choose from A, B, Both or None.

Agreement: "*Which of the two sentences has more agreement errors between the subject/object and the verb (i.e. the suffixes for the verbs and/or the nouns do not agree with each other)?*" and let annotators choose from A, B, Both or None.

	Transformer	TP-Transformer
BPE	.498 $\pm$.011	.513 $\pm$.003
Morphological	.526 $\pm$.007	.539 $\pm$.006

Table 9: CHRF scores on the test set of Inuktitut divided by tokenization. We show the mean and standard deviation of three randomly initialized models.

Fluency: "*Which of the two sentences is more fluent i.e. reads more like it was written by a native Turkish speaker?*" and let annotators choose from A, B, Both are equally fluent.

G Output analysis

Here we present an error analysis of a few sample translations from Transformer and TPT models. We group errors according to the aspects used to perform human-based evaluation in §5.2. Table 10 shows the result of this analysis. Under fluency issues, Transformer introduces an unnecessary word 'zamaninda' making it less fluent compared to the TPT translation. Under meaning preservation, the translation by Transformer incorrectly suggests "in exchange for money" whereas TPT correctly preserves the meaning. Under agreement issues, TPT includes incorrect use of first person suffix whereas Transformer does not

	Fluency Issues
English	*"I want to carry on living," he said at the time of the CPJ award.*
Turkish Transformer	*CPJ **ödülünün zamanında** konuşan Jovanoviç, "Yaşamak istiyorum." dedi.*
Turkish TPT	*CPJ ödülünde konuşan bakan, "Yaşamayı sürdürmek istiyorum." dedi.*
Reason	unnecessary use of the word 'zamaninda'.
	Meaning Preservation
English	*Some say you chose Turkey for money.*
Turkish Transformer	*Bazıları Türkiye'yi **para karşılığında** seçtiğinizi söylüyor.*
Turkish TPT	*Bazıları, para için Türkiye'yi seçtiğinizi söylüyorlar.*
Reason	"para karşılığında" suggests 'in exchange for money'
	Subject to verb agreement Issues
English	*Maybe because I go to bed listening to the message you left, saying how much you liked missing me.*
Turkish Transformer	*Belki de yatağa gidip, beni özlemeyi ne kadar sevdiğini söyleyen mesajını dinlediğim için.*
Turkish TPT	*Belki de yatağa gidip bıraktığın mesajı dinleyip beni özlediğini **söy-le-di-ğ-im** için.*
Reason	incorrect use of first person ('-im') instead of second person ('-in')
	Morphology Issues
English	*So far we have not received any news nor found any clues.*
Turkish Transformer	*Şimdiye kadar hiçbir haber alamadık ve hiçbir ipucu bulamadık*
Turkish TPT	*Bugüne kadar ne haber aldık ne de ipucu **bul-a-ma-dı-k***
Reason	Highlighted word has a double negative instead of the correct form **bul-a-bil-di-k/bul-du-k**.

Table 10: Sample outputs showing issues relating to fluency, meaning preservation, agreement and morphology from Transformer and TPT models.

have any subject to verb agreement issues. Under morphology issues, TPT incorrectly includes a negation suffix making the sentence a double negative whereas Transformer correctly translates the English sentence.

Table 11 includes analysis of some additional sample outputs from Transformer and TPT models. Under morphology issues, Transformer includes an unnecessary plural suffix. The TPT translation is okay but would have been better with the addition of the '-mu' suffix. Under meaning preservation, Transformer incorrectly translates "Bank of England" as "Bank of England", thus losing out on the meaning. Whereas TPT correctly translates that named entity into Turkish. Under tense issues, Transformer uses an incorrect past tense suffix whereas TPT correctly preserves the tense of the English sentence. Under repetition issues, Transformer repeats a word which is not required in written-language but might be okay in spoken-language.

	Morphology Issues
English	*First we have to decide if those lost six minutes will be coming out of game time, bathroom time or the pizza break.*
Turkish Transformer	*İlk önce, bu altı dakika kaybet-me-ler-i-n oyun zamanından mı yoksa banyo zamanından mı olacağına karar vermeliyiz.*
Turkish TPT	*İlk olarak, o 6 dakikanın maçtan, banyo saatinden veya pizza molasından (-mı) çıkıp çıkmayacağına karar vermeliyiz.*
Reason	Unnecessary plural suffix (-ler)
	Meaning Preservation
English	*Bank of England to keep interest rates at 0.25%*
Turkish Transformer	*Bank of England faiz oranlarını %0,25 oranında tutacak.*
Turkish TPT	*İngiltere Merkez Bankası faiz oranlarını %0,25 oranında tutacak.*
Reason	Incorrect translation of named entity
	Tense Issues
English	*Barely out of bed and already on the phone.*
Turkish Transformer	*Yataktan zar zor çıktım ve telefonla konuştum bile.*
Turkish TPT	*Yataktan zar zor çıktım ve telefondayım.*
Reason	Incorrect use of past tense suffix (‘-tum’) instead of present tense suffix (‘yorum’)
	Repetition Issues
English	*Specific criteria, such as an asteroid’s size and collision angle, are the factors that would determine the depth of its crater and the damage that its impact would cause.*
Turkish Transformer	*Asteroidin büyüklüğü ve çarpışma açısı gibi belli kriterler, kraterin derinliğini belirleyecek ve etkisinin yaratacağı hasarı belirleyecek faktörler*
Turkish TPT	*Bir asteroidin büyüklüğü ve çarpışma açısı gibi belirli kriterler, kraterinin derinliğini ve etkisinin yol açacağı hasarı belirleyecek faktörler*
Reason	The word “belirleyecek” is repeated which is unnecessary in written-language but would be okay in spoken-language.

Table 11: Sample outputs from Transformer and TPT models showing issues relating to morphology, meaning preservation, tense and repetition.