PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 115 OCTOBER 2020 143–162

Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features?

David Mareček,[a] Hande Celikkanat,[b] Miikka Silfverberg,[b]
Vinit Ravishankar,[c] Jörg Tiedemann[b]

[a] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
[b] Department of Digital Humanities, University of Helsinki
[c] Department of Informatics, University of Oslo

Abstract

We investigate the effect of training NMT models on multiple target languages. We hypothesize that the integration of multiple languages and the increase of linguistic diversity will lead to a stronger representation of syntactic and semantic features captured by the model. We test our hypothesis on two different NMT architectures: The widely-used Transformer architecture and the Attention Bridge architecture. We train models on Europarl data and quantify the level of syntactic and semantic information discovered by the models using three different methods: SentEval linguistic probing tasks, an analysis of the attention structures regarding the inherent phrase and dependency information and a structural probe on contextualized word representations. Our results show evidence that with growing number of target languages the Attention Bridge model increasingly picks up certain linguistic properties including some syntactic and semantic aspects of the sentence whereas Transformer models are largely unaffected. The latter also applies to phrase structure and syntactic dependencies that do not seem to be developing in sentence representations when increasing the linguistic diversity in training to translate. This is rather surprising and may hint on the relatively little influence of grammatical structure on language understanding.

1. Introduction

There have been indications that explicitly modeling linguistic information can help performance of neural machine translation (NMT) models (Aharoni and Goldberg, 2017; Nadejde et al., 2017). Conversely, there is evidence that encoder-decoder

Corresponding author: marecek@ufal.mff.cuni.cz

Cite as: David Mareček, Hande Celikkanat, Miikka Silfverberg, Vinit Ravishankar, Jörg Tiedemann. Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features?. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 143–162. doi: 10.14712/00326585.009.

NMT models also discover linguistic properties without overt supervision while learning to translate (Conneau et al., 2018a; Mareček and Rosa, 2019). This paper provides a new perspective on the topic of linguistic information that is captured by NMT models. Specifically, we investigate the effect of training NMT models on multiple target languages using the assumption that the integration of multiple languages and the increase of linguistic diversity will lead to a stronger representation of syntactic and semantic features captured by the model. Indeed, our experiments show evidence that increasing the number of target languages forces the NMT model to generate more semantically rich representations for input sentences. However, our results do not provide strong support for the integration of additional syntactic properties in latent representations learned by multilingual translation models.

In a bilingual translation setting, especially when the source and target language are related, an NMT model can focus on shallow transformations between the input and output sentences. We hypothesize that this strategy is not sufficient anymore when the number and diversity of the target languages grow. Encoder representations for input sentences in a multilingual setup need to support a mapping to various target language realizations displaying a range of different linguistic properties. In other words, when faced with substantial linguistic diversity, the model will need to create additional abstractions reflecting syntactic and semantic structure that is essential for proper understanding and meaningful translation. In our research, we are interested in finding out what kind of structure is needed in such a setup and what kind of linguistic properties are picked up by current models of attentive neural machine translation.

In order to model a challenging level of linguistic coverage, we, therefore, apply a diverse set of target languages: Czech, Finnish, German, Greek and Italian. Each of these languages exhibit significantly different properties ranging from the complexity of their morphological system and rigidity of word order and syntactic structure up to differences in tense, aspect and lexical meaning. The source language is always English. Based on our experimental setup we now attempt to quantify and compare the semantic and syntactic information discovered by models with increasing amount of target language diversity and we test our hypothesis on two different NMT architectures: The widely-used Transformer architecture (Vaswani et al., 2017), a multi-headed attention based model, and the Attention Bridge architecture (Cífka and Bojar, 2018; Lu et al., 2018), an RNN-based model, which produces fixed-sized cross-lingual sentence representations.

In order to measure linguistic properties discovered by the models, we apply the following three methods: (1) the SentEval linguistic probing tasks on sentence representations, (Conneau et al., 2018a), (2) an analysis of the attention structures regarding the inherent phrase and dependency information (Mareček and Rosa, 2019), and (3) the structural probe on contextualized word representations proposed by (Hewitt and Manning, 2019).

2. Related Work

We learn sentence representations in a multilingual setting. In their seminal paper on multi-lingual neural machine translation, Johnson et al. (2017) show evidence that sentence representations learned for different source languages tend to cluster according to the semantics of the source sentence rather than its language. Schwenk and Douze (2017) train encoder-decoder systems on multiple source and target languages and investigate source sentence representations w.r.t. cross-lingual representation similarity.

Conneau et al. (2018b) train multilingual sentence representations for cross-lingual natural language inference by aligning source and target language representations instead of directly training the system to translate. Artetxe and Schwenk (2019) learn massively multilingual sentence representation on a training set encompassing 93 languages and show good performance on a number of downstream tasks.

Interpretation and evaluation of sentence representations has recently become a very active research area. Conneau et al. (2018a) investigate several ways to learn sentence representations for English and present a benchmark of probing tasks for syntax and semantics.

The structural probe presented by Hewitt and Manning (2019) investigates the relation between the syntax tree of a sentence and its contextualized word embeddings derived from a model. They show that monolingual English ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) embeddings encode syntactic structure whereas baselines do not. This approach is attractive because it directly investigates syntactic information captured by representations in contrast to probing, where an additional classifier is trained. We apply the structural probe as one of our evaluation methods.

Chrupała and Alishahi (2019) use representational similarity analysis to compare the metrics induced by sentence representations and syntactic dependency trees. This approach is more flexible than the structural probe because it can compare metrics in unrelated spaces (for example continuous sentence representations and symbolic representations like syntax trees).

Another approach to investigate the syntactic information captured by transformer models is to relate self attentions to syntactic phrase or dependency structures. This approach was pioneered by Raganato and Tiedemann (2018), who analyze self attentions in terms of the dependency tree structures and Mareček and Rosa (2019), who train parsers based on self attentions of transformer models in monolingual and multilingual settings.

Whereas there is a large body of related work on interpretation of sentence representations learned by NMT models, few studies directly investigate the effect of multilinguality on sentence representations. Closely related to our work is the work by Ravishankar et al. (2019) which extends the probing tasks presented by Conneau et al. (2018a) into the multilingual domain. They train multilingual sentence representations for NLI by training an English NLI system and mapping sentences from other languages into the English representation space following Conneau et al. (2018b). They then conduct probing experiments on a multilingual dataset. Ravishankar et al. (2019) notice that, quite surprisingly, transferred representation can deliver better performance on some probing task than the original English representations.

Kudugunta et al. (2019) investigate massively multilingual NMT on a combination of 103 languages. In contrast to this paper, they investigate language representations using Singular Value Canonical Correlation Analysis. They show that encoder representations of different languages cluster according to language family and that the target language affects source language representations in a multilingual setting. In contrast to Kudugunta et al. (2019), our work investigates sentence representations instead of language representations and we investigate the impact of multilinguality on learning syntax and semantics.

To the best of our knowledge, this paper presents the first systematic study of the effect of target language diversity on syntactic and semantic performance for sentence representations learned by multilingual NMT models.

3. Data and Systems

In all our experiments, we use a multi-parallel[1] subset of the Europarl corpus (Koehn, 2005) spanning 391,306 aligned sentences in six languages: English, Czech, Finnish, German, Greek, and Italian. We choose these languages in order to include one representative from each of the major language families in the Europarl dataset allowing maximal diversity among target languages. The multi-parallel corpus is randomly divided into training (389,306 examples), development (1000 examples) and test (1000 examples) sets.

We always use English as the source language, while we vary the number of target languages. Specifically, we set up a systematic study starting with a single target language out of our set, and combining one additional target language at a time, until we reach the exhaustive combination of all the five target languages. Table 1 depicts all our settings. Note that we balance the number of occurrences of each language over training configurations in order to avoid biasing combinations toward particular languages.[2]

We use a multi-parallel corpus in order to avoid injecting additional source language information when increasing the number of target languages. Even when the number of target languages grows, the English source language data remains the same. The only difference is that each source sentence in the training data is paired with multiple translations in each of the target languages. This ensures that any addi-

[1] We took the intersection over the five parallel corpora.

[2] This means that each language occurs twice in 2-combinations, three times in 3-combinations and four times in 4-combinations of languages.

Source	**Target**	
{En}	**1 tgt**	{Cs}, {De}, {El}, {Fi}, {It}
	2 tgts	{Cs, De}, {De, El}, {El, Fi}, {Fi, It}, {It, Cs}
	3 tgts	{Cs, De, El}, {De, El, Fi}, {El, Fi, It}, {Fi, It, Cs}, {It, Cs, De}
	4 tgts	{Cs, De, El, Fi}, {De, El, Fi, It}, {El, Fi, It, Cs}, {Fi, It, Cs, De}, {It, Cs, De, El}
	5 tgts	{Cs, De, El, Fi, It}

Table 1. The configurations of the 21 different training scenarios. English is the source language in all configurations, while the combination of the target languages differs between scenarios.

tional syntax awareness in models trained on higher combinations of target languages cannot be due to additional English language data.

To preprocess our data, we first run a truecaser (Lita et al., 2003) before splitting into subword units using BPE (Lita et al., 2003). For the latter we train a model with 100k merge operations on the concatenation of all source and target language data.

3.1. Transformer

The first model architecture in our experimental setup is the widely used Transformer model by Vaswani et al. (2017). The Transformer is a multi-headed attention-based, feed-forward architecture. Each head can freely attend to any position, resulting in greater flexibility then competing sequential RNNs. Typically, several layers are stacked on top of each other, and each layer incorporates its own dedicated attention heads. Furthermore, the output from this attention mechanism is averaged with the original input vector via residual connections.

For the Transformer architecture we use a single encoder and decoder even in a multilingual setting using target language labels for informing the translation system about the language to be generated. Following (Artetxe and Schwenk, 2019), we add those labels to the beginning of target sentences rather than source sentences, which effectively hides target language information from the encoder guaranteeing a unified source sentence representation. During test time, we force-decode the initial target language label before continuing the standard decoding process that generates the translation in the desired language.

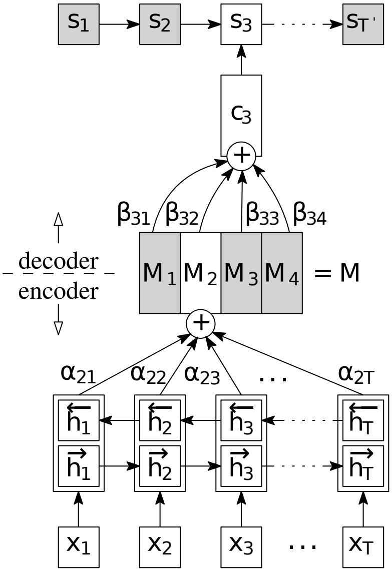
s_1
s_2
s_3
$s_{T'}$
c_3
β_{31}
β_{32}
β_{33}
β_{34}
decoder
encoder
M_1 M_2 M_3 M_4 = M
α_{21}
α_{22}
α_{23}
...
α_{2T}
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_T$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_T$
x_1 x_2 x_3 ... x_T

Figure 1. NMT architecture with the attention bridge (Cífka and Bojar, 2018)

3.2. Attention Bridge

Almost all recent NMT architectures (Bahdanau et al., 2015; Vaswani et al., 2017) utilize some kind of cross language attention that directly connects encoder with decoder representations. Cífka and Bojar (2018) introduced the idea of an attention bridge as it is depicted in Figure 1. Here, the whole sentence is encoded into one fixed-size matrix M that serves as an intermediate abstraction layer between attentive encoders and decoders. Sharing this layers across languages enables the effective combination of language-specific encoder and decoder modules to build an extensible multilingual translation architecture. A similar idea was proposed by Lu et al. (2018) but with a slightly different recurrent architecture in the intermediate layer.

In our experiments, we use a variant of the Attention Bridge re-implemented by Raganato et al. (2019) in the OpenNMT-py framework.[3] In this setup we have exactly one encoder for English and one to five separate decoders for our target languages. We run experiments for four different numbers of attention bridge heads: 10, 20, 40, and 80.

[3] Network parameters: 2 bidirectional GRU encoder layers of size 512, MLP attention bridge, 2 GRU decoder layers.

4. Evaluation of Syntax and Semantics

4.1. SentEval Probing Tasks

Our first measure for the degree of semantic and syntactic information captured by sentence representations is a set of ten linguistic classification tasks, so called probing tasks, presented by Conneau et al. (2018a) that look at different syntactic and semantic aspects of a sentence. We conduct experiments using the SentEval toolkit (Conneau and Kiela, 2018) which trains and evaluates models for each of them. Training, development and test data are provided by the SentEval toolkit and we extract the necessary representations for all sentences in those data sets from our Transformer and Attention Bridge models.

Three of the ten SentEval tasks probe for structural properties of the sentence and its syntax tree: **Depth** (depth of the syntax tree), **Length** (binned length of the input sentence) and **TopConsitutents** (the top-most non-root constituents in the syntax tree, for example **NP VP**). Three tasks probe for semantic properties of its main syntactic components: **SubjectNumber** (grammatical number of the subject), **ObjectNumber** (grammatical number of the object) and **Tense** (tense of the main verb). Three of the tasks perturb parts of the original sentences and ask the classifier to identify which of the sentences have been scrambled: **BigramShift** (recognize whether two tokens in the sentence have been transposed), **CoordinationInversion** (recognize whether two coordinated clauses have been transposed) and **SemanticOddManOut** (recognize whether a token in the sentence has been replaced by a random vocabulary item). Finally, **WordContent** is the task of predicting which of around 1,000 mid-frequency words occurs in the input sentence.

WordContent and Length represent surface properties of the sentence; BigramShift, Depth and TopConstituents are purely syntactic tasks; and SubjectNumber, ObjectNumber and Tense are semantic tasks which are related to the syntactic structure of the sentence. Finally, SemanticOddManOut and CoordinationInversion are purely semantic tasks.

We process the training, development and test data for probing tasks identically to the data used for NMT models: we use the same truecasing and BPE models for preprocessing. Subsequently, we extract sentence representations for the sentences to train the SentEval multi-layer perceptron classifier for each task and setting hyperparameters using grid search. Finally, the toolkit provides the classification accuracy on the test set.

4.2. Evaluating Transformer's Self-Attentions

Another way of measuring the amount of syntax captured by the translation encoder is to analyze its self-attention mechanisms and compare them to linguistically motivated syntactic trees (Raganato and Tiedemann, 2018; Mareček and Rosa, 2019).

For this, we partially adapted the approach used by Mareček and Rosa (2019). During the translation of the test data, we extract the weights of the self-attentions of all the attention heads from all six encoder layers, and compare them to syntactic structures of the source sentences automatically created by the Standford Parser (Klein and Manning, 2003) (for phrase-structure trees) and by UDPipe (Straka and Straková, 2017) (for syntactic dependency trees).

An example of typical distributions of weights in one encoder attention head is shown in Figure 2. For our parameter setting,[4] the attentions are very sharp and very often focused on just one token in the previous layer and we observe a kind of continuous phrase attending the same token from the previous layer. Such phrases may then be compared to the syntactic phrases we obtain by a syntactic parser.

The evaluation procedure is the following: First, we "sharpen" the soft attention matrix by only keeping the maximal attention weight on each row of the attention matrix, setting the weights on all other positions to 0:

$$A_{o,i} = \begin{cases} A'_{o,i} & \text{if } A'_{o,i} = \max_{j \in [1,N]} A'_{o,j} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where A' is the original self-attention weight matrix, i and o is the input and output state index respectively, and N is the length of the sentence. Second, we compute the weights for each possible continuous phrase by averaging the individual weights:

$$w_{a,b} = \frac{\sum_{i \in [1,N]} \sum_{o \in [a,b]} A_{o,i}}{b - a + 1}, \tag{2}$$

where a and b is the beginning and the end of the phrase. Such weights are computed for each attention head and for each layer. Then, we can compute layer-wise precision and recall:

$$\text{PhrPrec}_L = \frac{\sum_{h \in H_L} \sum_{[a,b] \in P} w^h_{a,b}}{\sum_{h \in H_L} \sum_{[a,b]} w^h_{a,b}} \tag{3}$$

$$\text{PhrRec}_L = \frac{\sum_{h \in H_L} \sum_{[a,b] \in P} w^h_{a,b}}{|P| \cdot |H|} \tag{4}$$

Where w^h are the phrase weights from attention head h which is chosen from the heads H_L on layer L. P are the phrases present in the constituency tree created by the Stanford Parser.

We can also evaluate the attention matrices with respect to a dependency trees. We simply take the pixels of the attention matrix corresponding to the dependency edges of the dependency tree obtained by UDPipe parser. Since it is not clear whether

[4] layers: 6, heads: 16, ff-size: 4096, normalization: tokens

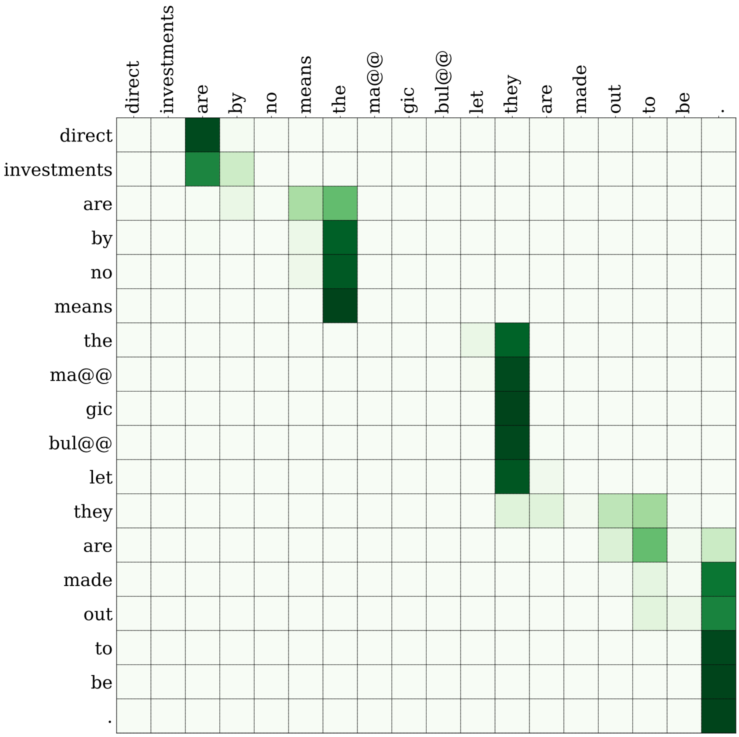

Figure 2. Example of a self-attention head (this one is head 4 on the 3th layer) in transformer encoder. Such continuous phrases attending to the same token are typical for many of the attention heads through all layers.

the dependents should attend to their governors or vice versa, we count both the possibilities. The precision is computed as sum of all "dependency" attention weights divided by the sum of all attention weights.

$$\mathrm{DepPrec}_L = \frac{\sum_{[i,j]\in D} \sum_{h\in H_L} A^h_{i,j} + A^h_{j,i}}{\sum_{h\in H_L} \sum_{i\in[1,N]} \sum_{j\in[1,N]} A^h i,j} \tag{5}$$

The recall is computed as an average weight of "dependency" attention.

$$\mathrm{DepRec}_L = \frac{\sum_{[i,j]\in D} \sum_{h\in H_L} A^h_{i,j} + A^h_{j,i}}{|D| \cdot |H|} \tag{6}$$

4.3. Evaluating Attention Bridge Cross-Attentions

In the attention-bridge architecture, there is one fixed-size vector representation of the input sentence M divided into n vectors composed by the individual attention bridge heads (see Figure 1). Each of them can possibly attend to all sentence tokens but, in practice, they tend to focus on continuous parts of the sentence. An example is included in Figure 3.

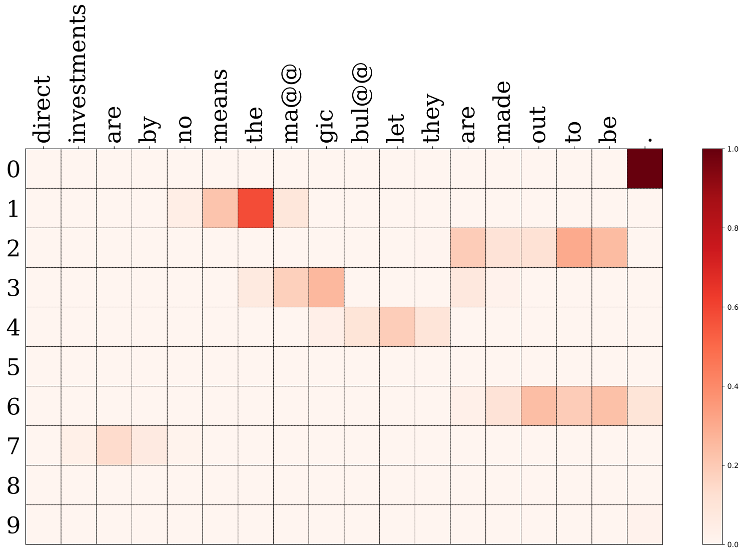

Figure 3. Example of distribution of weights in a 10-headed attention bridge.

Once in a while, we can find more then one phrase per head. However, we treat such cases as one long phrase. For each head we simply take the beginning of the phrase as the leftmost token with weight higher than a threshold t and the end of the phrase as the rightmost token with weight higher than t. We set the threshold t to 0.1. We also tested other thresholds controlling the phrase lengths, but the final results were all very similar and, therefore, we keep the original setting in the results presented hereafter.

Having the set of phrases extracted from the attention bridge, we can now compare it to the phrases of constituency trees obtained by Stanford parser measuring precision in the usual way.

4.4. Structural Probe

We also attempt to evaluate the syntax our representations store by extending Hewitt and Manning's (2019) probe to a multilingual domain. The probe they describe is capable of learning to reliably extract some form of dependency structure, via a combination of two independent distance and depth components. For a detailed mathematical description of either component, we refer the reader to the original paper. Whilst the original probe returns undirected edge weights and depths separately, we (trivially) combine these by forcing edges to point from shallower to deeper nodes. We employ Chu-Liu/Edmonds' algorithm (Chu and Liu, 1965; McDonald et al., 2005) to extract the minimum spanning arborescence of this graph, which is equivalent to a conventional dependency tree.

5. Results

SentEval Probing Tasks: The results of SentEval evaluations are illustrated in Figure 4. For the Attention Bridge, accuracy on all probing tasks except WordContent and SemanticOddManOut generally improves when the number of target languages goes up. The same trend can be seen with all sizes of the attention bridge.

For the Transformer, the effect of adding more target languages does not result in a clear change in probing task accuracy. For Length and Tense, we can discern a small improvement but for the other tasks, performance seems largely independent of the number of target languages. Interesting is that the performance of higher layers is better than for lower layers in almost all cases. SemanticOddManOut is a clear exception. Furthermore, we can also see that the Attention bridge model performs better on most of the probing tasks when adding multiple target languages and increasing the size of the attention bridge. This especially true with the semantic tasks in SentEval.

Syntactic Evaluation of Attentions: Next, we try to assess the attention vectors from the two models in terms of the syntactic information they include. Figure 5 shows the precision and recall results for the phrase trees and the dependency relations. We observe almost no changes or even a slight decreases for the Attention Bridge model when adding more languages to the model. For the Transformer models, we see a slight increase of Phrase precision and recall on the last two layers (4 and 5), whereas the measures on the lower layers are slightly decreasing with the number of target languages.

Structural Probe: Finally, we perform an analysis of the contextualized word representations of the Transformer.[5] Figure 6 describes the variation in UAS with sentence

[5] Note that the Attention Bridge does not produce a per-token representation, and, therefore, this part of the analysis is not applicable for that model.

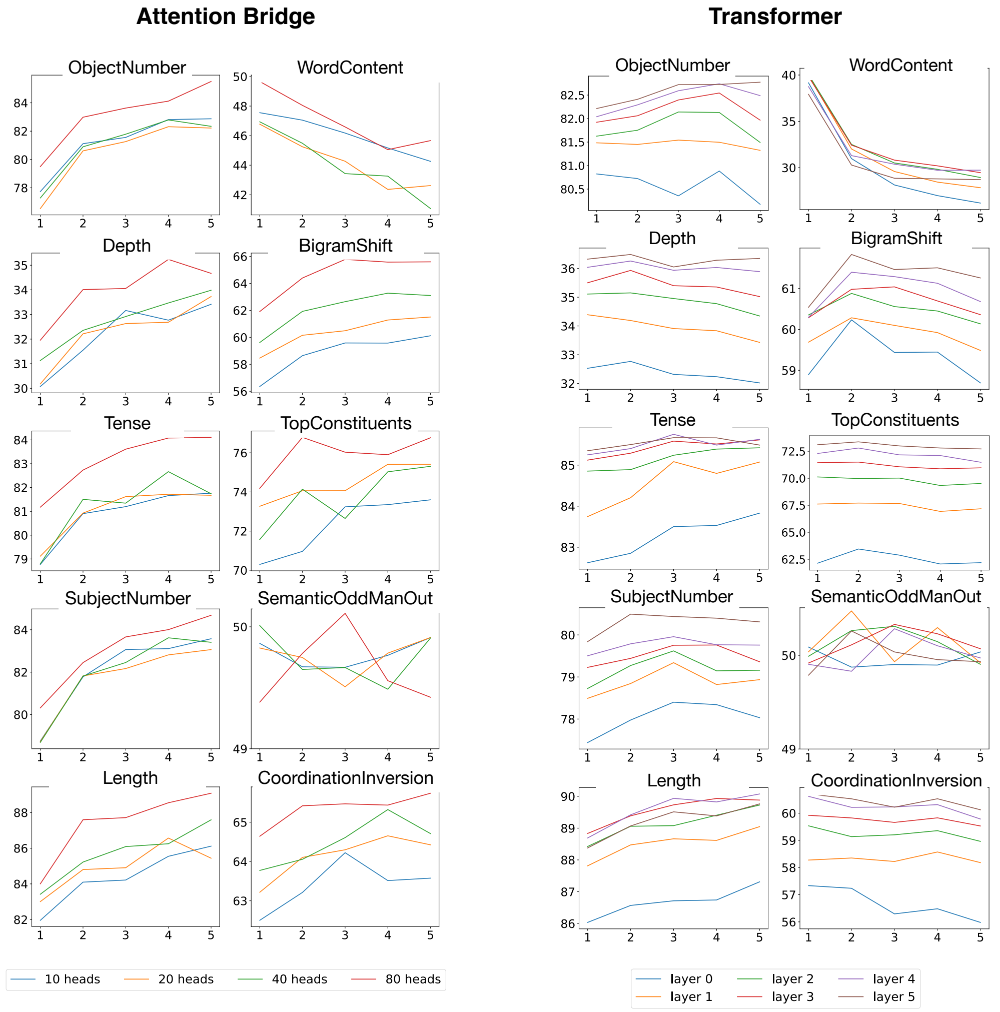

Figure 4. SentEval results for all probing tasks for both the Attention Bridge and Transformer models. The average classification accuracies on the corresponding SentEval task for increasing number of target languages in the models (x-axis) are depicted. For Attention Bridge models, different plot colors indicate different numbers of heads (10, 20, 40, or 80). For Transformer models, different plot colors indicate the layer number (from 0 to 5).

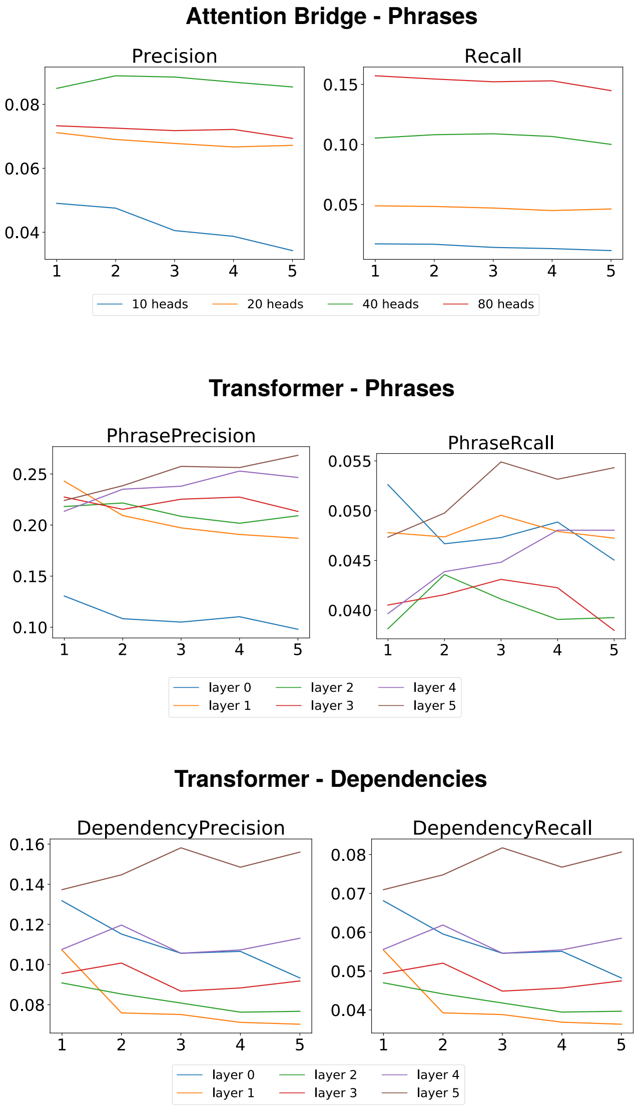

Figure 5. The precision and recall graphs for the continuous phrases extracted from the attention vectors of Attention Bridge and for the continuous phrases and dependency relations form the Transformer models. X-axis denotes the number of target languages.

length for increasing number of languages, and Figure 7 shows UAS variation per token, for three token 'categories' based on their POS.

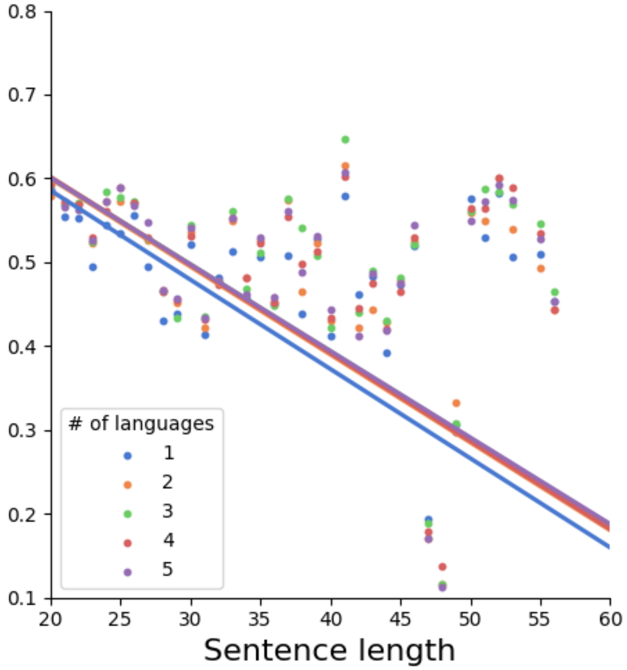

Figure 6. UAS plotted against sentence length. Lines represent trend lines.

6. Discussion

Our results support a connection between the number of target languages in an NMT model and the linguistic properties it picks up at least in the Attention Bridge model as evidenced by the SentEval probing tasks. In that model, all probing tasks except WordContent and SemanticOddManOut significantly increase when the number of target languages in the model grows.

At the same time, BLEU scores for translation performance actually degrade for smaller models (Attention Bridge with 10 and 20 heads) and remains constant for larger models (Attention Bridge with 40 and 80 heads, as well as Transformer), see Figure 8. Degradation of translation performance in itself is not unusual. For example, Kudugunta et al. (2019) notice that performance of high resource languages degrades

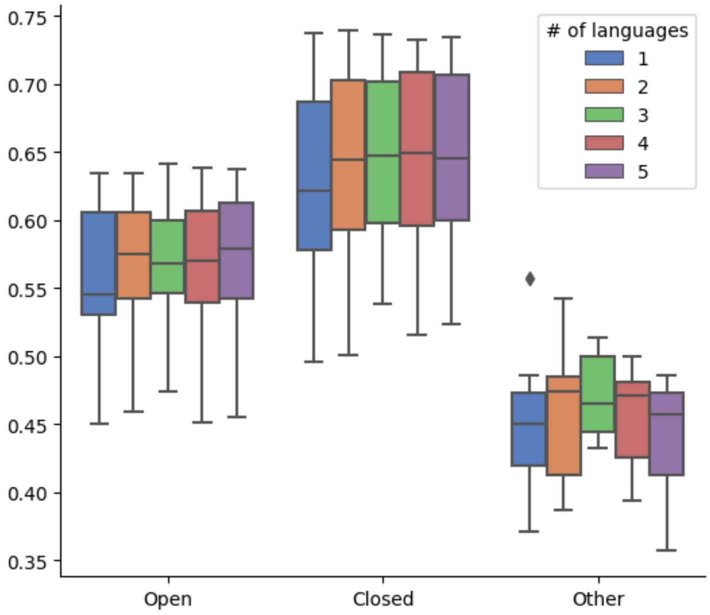

Figure 7. UAS for different groupings of (dependent) tokens by POS. Mappings are the same as in `universaldependencies.org/u/pos/`

in multilingual models. However, it is very interesting that this is accompanied by improved performance on linguistic probing tasks.

For the Transformer model, only the Tense and Length probing tasks seem to show consistent improvement when the number of target languages increases. In general, higher layers tend to deliver a better performance. The overall result for the Transformer model is lower on SentEval tasks than for the Attention Bridge model. This is consistent with some earlier observations, eg., (Tran et al., 2018) who show that RNN-based models tend to outperform the Transformer in subject-verb agreement.

The WordContent task shows a clearly degrading performance when the number of target languages increases. The SemanticOddManOut task in turn shows a very diffuse picture. Those trends are visible in both model architectures, However, these probing tasks differ from all the other ones in the sense that the output label is a word type rather than a category from a limited set or a small integer value as explained in Section 4.1. We believe that the confusion might be due to the BPE segmentation of the input data which generates sub-word level tokens and thus increases the difficulty of

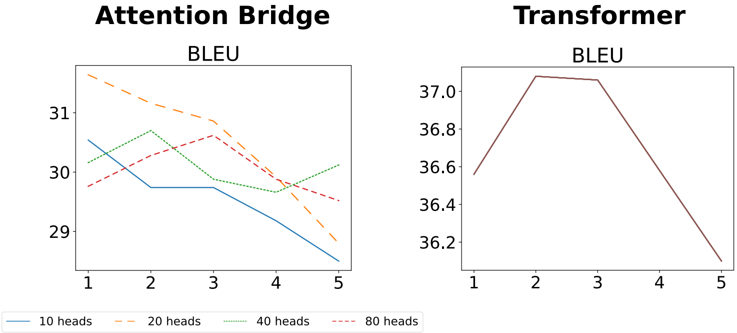

Figure 8. Averaged BLEU scores for Attention Bridge and Transformer. x-axis denotes the number of target languages. Evaluation was done on the test *part of our data.*

the classification task. Furthermore, we note that (Conneau et al., 2018a) also report fluctuating performance for WordContent, which reduces the trust in this particular probing task.

Applying a structural probe to our representations results in several interesting observations. Figure 7 seems to indicate that the jump in *median* syntactic performance is largest when as few as two languages are used as target languages; indicating that the marginal value of further target languages is, as far as syntax is concerned, minimal. Figure 6 also seems to indicate that this holds true across all sentence lengths although the gap widens slightly for longer sentences. We also observe that the increase in median performance is greater for open-class words than closed-class words; this intuitively makes sense, as open-class terms are likelier to have a broader range of semantic values, which are likelier to be better defined with multiple target tokens. Moreover, this observation further corroborates our other results, that exhibit more noticeable improvements in semantic-level tasks than syntactic ones: tokens that receive more reference translations are more likely to be able to better contextualise a broader range of semantic values, particularly from a perspective of lexical disambiguation.

Results for generating syntax trees seem to be largely negative. There is no discernible tendency for the precision or recall on phrase structure for the Attention Bridge model. For the transformer, we see a slight increasing trend in the precision and recall when the number of target languages grows both for phrase structure and dependency parsing for the final layer in the model. There is no clear tendency for the other layers.

An important question our results raise is why the Attention Bridge model shows a much more clear on probing tasks as compared to the Transformer. We hypothesize that this difference may be due to the much greater number of parameters that the

Transformer employs. As a result of having access to a much larger representational space, the Transformer may not have needed to abstract so drastically over several target languages, resorting instead to dedicate some specific part of the representational space to each language. In contrast, the Attention Bridge model with a much more restricted parameter space might have been under more pressure to abstract useful syntactic representations when confronted with a large number of different languages.

7. Conclusion

In this paper, we investigate the impact of additional target languages in multilingual NMT systems on syntactic and semantic information captured by its sentence representations. We analyze two models, the Attention Bridge and the Transformer, using three different evaluation methods. We show evidence that performance on linguistic probing tasks improve for the Attention Bridge when the number of target languages grows. We also show that a transition from a bilingual to a multilingual setting improves performance for the structural probe presented by (Hewitt and Manning, 2019). While we find evidence for improved performance on probing tasks, many of which are related to the semantics of the sentence, our results on syntax performance are inconclusive.

Several interesting unresolved questions remain. Although we tried to cover substantial linguistic variety by using languages from different families, the effect of an even larger typological diversity is still an open question. Additionally, we would also like to know how multiple source languages would affect the results and whether they depend on other latent variables and parameters in the model.

Acknowledgements

This work has been supported by the grant 18-02196S of the Czech Science Foundation and by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 771113). We have been using language resources and tools developed, stored and distributed by theLINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

Bibliography

Aharoni, Roee and Yoav Goldberg. Towards String-To-Tree Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–140, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2021. URL https://www.aclweb.org/anthology/P17-2021.

Artetxe, Mikel and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

Chrupała, Grzegorz and Afra Alishahi. Correlating neural and symbolic representations of language. *arXiv preprint arXiv:1905.06401*, 2019.

Chu, Y. J. and T. H. Liu. On the Shortest Arborescence of a Directed Graph. *Science Sinica*, 14: 1396–1400, 1965.

Cífka, Ondřej and Ondřej Bojar. Are BLEU and Meaning Representation in Opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1126. URL `https://www.aclweb.org/anthology/P18-1126`.

Conneau, Alexis and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations. *arXiv preprint arXiv:1803.05449*, 2018.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018a. doi: 10.18653/v1/P18-1198.

Conneau, Alexis, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018b. doi: 10.18653/v1/D18-1269.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Hewitt, John and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

Johnson, Melvin, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5(0), 2017. doi: 10.1162/tacl_a_00065.

Klein, Dan and Christopher D Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10, 2003.

Koehn, Philipp. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

Kudugunta, Sneha Reddy, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*, 2019. doi: 10.18653/v1/D19-1167.

Lita, Lucian Vlad, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. Truecasing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 152–159. Association for Computational Linguistics, 2003. doi: 10.3115/1075096.1075116.

Lu, Yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6309. URL `https://www.aclweb.org/anthology/W18-6309`.

Mareček, David and Rudolf Rosa. From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4827. URL `https://www.aclweb.org/anthology/W19-4827`.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *HLT-EMNLP*, pages 523–530, 2005. doi: 10.3115/1220575.1220641.

Nadejde, Maria, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. Predicting target language ccg supertags improves neural machine translation. *arXiv preprint arXiv:1702.01147*, 2017. doi: 10.18653/v1/W17-4707.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. doi: 10.18653/v1/N18-1202.

Raganato, Alessandro and Jörg Tiedemann. An Analysis of Encoder Representations in Transformer-Based Machine Translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5431. URL `https://www.aclweb.org/anthology/W18-5431`.

Raganato, Alessandro, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. An Evaluation of Language-Agnostic Inner-Attention-Based Representations in Machine Translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 27–32, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4304. URL `https://www.aclweb.org/anthology/W19-4304`.

Ravishankar, Vinit, Lilja Øvrelid, and Erik Velldal. Probing Multilingual Sentence Representations With X-Probe. *arXiv preprint arXiv:1906.05061*, 2019. doi: 10.18653/v1/W19-4318.

Schwenk, Holger and Matthijs Douze. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Rep4NLP@ACL*, 2017. doi: 10.18653/v1/W17-2619.

Straka, Milan and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw*

Text to Universal Dependencies, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3009. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Tran, Ke, Arianna Bisazza, and Christof Monz. The Importance of Being Recurrent for Modeling Hierarchical Structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1503. URL https://www.aclweb.org/anthology/D18-1503.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Address for correspondence:
David Mareček
marecek@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Malostranské náměstí 25, 118 00 Praha, Czechia