Investigating Multilingual NMT Representations at Scale

Sneha Kudugunta **Ankur Bapna** **Isaac Caswell** **Orhan Firat**
Google AI
{snehark,ankurbpn,icaswell,orhanf}@google.com

Abstract

Multilingual Neural Machine Translation (NMT) models have yielded large empirical success in transfer learning settings. However, these black-box representations are poorly understood, and their mode of transfer remains elusive. In this work, we attempt to understand massively multilingual NMT representations (with 103 languages) using Singular Value Canonical Correlation Analysis (SVCCA), a representation similarity framework that allows us to compare representations across different languages, layers and models. Our analysis validates several empirical results and long-standing intuitions, and unveils new observations regarding how representations evolve in a multilingual translation model. We draw three major conclusions from our analysis, with implications on cross-lingual transfer learning: (i) Encoder representations of different languages cluster based on linguistic similarity, (ii) Representations of a source language learned by the encoder are dependent on the target language, and vice-versa, and (iii) Representations of high resource and/or linguistically similar languages are more robust when fine-tuning on an arbitrary language pair, which is critical to determining how much cross-lingual transfer can be expected in a zero or few-shot setting. We further connect our findings with existing empirical observations in multilingual NMT and transfer learning.

1 Introduction

Multilingual Neural Machine Translation (NMT) models have demonstrated great improvements for cross-lingual transfer, on tasks including low-resource language translation (Zoph et al., 2016; Nguyen and Chiang, 2017; Neubig and Hu, 2018) and zero or few-shot transfer learning for downstream tasks (Eriguchi et al., 2018; Lample and Conneau, 2019; Wu and Dredze, 2019). A possible explanation is the ability of multilingual models to encode text from different languages in a shared representation space, resulting in similar sentences being aligned together (Firat et al., 2016; Johnson et al., 2017; Aharoni et al., 2019; Arivazhagan et al., 2019b). This is justified by the success of multilingual representations on tasks like sentence alignment across languages (Artetxe and Schwenk, 2018), zero-shot cross-lingual classification (Eriguchi et al., 2018) and XNLI (Lample and Conneau, 2019). Although there exist empirical results that suggest that transfer is stronger across similar languages (Zoph et al., 2016; Neubig and Hu, 2018; Wu and Dredze, 2019), the mechanism of generalization in multilingual representations is poorly understood.

While interpretability is still a nascent field, there has been some work on investigating the learning dynamics and nature of representations learnt by neural models (Olah et al., 2018). Singular Value Canonical Correlation Analysis (SVCCA) is one such method that allows us to analyze the similarity between noisy, high-dimensional representations of the same datapoints learnt across different models, layers and tasks (Raghu et al., 2017). SVCCA has been used to understand the learning dynamics and representational similarity in a variety of computer vision (Morcos et al., 2018), language models (Saphra and Lopez, 2018) and NMT (Bau et al., 2018).

In this work, we attempt to peek into the black-box of massively multilingual NMT models, trained on 103 languages, with the lens of SVCCA. We attempt to answer:

- Which factors determine the extent of overlap in the learned representations?
- Is the extent of representational overlap similar throughout the model?

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1565–1575,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

- How robust are multilingual NMT representations to fine-tuning on an arbitrary other language?

Answers to the above questions might have large implications on how we approach multilingual models and cross lingual transfer learning. Our work is the first that attempts to understand the nature of multilingual representations and cross-lingual transfer in deep neural networks, based on analyzing a model trained on 103 languages simultaneously.

We structure the study into these sections: In Section 2, we describe the experimental setup and tools used to train and analyze our multilingual NMT model. Following that, we attempt to answer each of the above questions in Sections 3 and 4. Finally in Section 5 we summarize our findings with a discussion of the potential implications.[1]

2 Experimental Setup

2.1 Data and Model

We study multilingual NMT on a massive scale, using an in-house training corpus (Arivazhagan et al., 2019b) generated by crawling and extracting parallel sentences from the web (Uszkoreit et al., 2010). Our dataset contains more than 25 billion sentence pairs for 102 languages to and from English, adding up to 204 direct language pairs.

Having being crawled from the web, our dataset has some important characteristics worth mentioning.

1. **Heavy imbalance between language pairs:** The number of parallel sentences per language pair ranges between 10^4 to 10^9. Figure 1 illustrates the data distribution for all the language pairs we study. Although this skew introduces optimization challenges (see Appendix A.1), it also creates a plausible setup for maximizing the positive language transfer from high-resource to low-resource language pairs, making it possible to study low-resource languages, that would otherwise have been very low quality.[2]

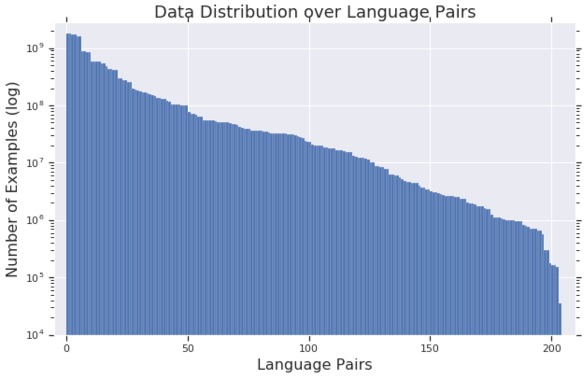


Figure 1: Per language pair data distribution of the dataset used to train our multilingual model. The y-axis depicts the number of training examples available per language pair on a logarithmic scale. Dataset sizes range from 10^4 for the lowest resource language pairs to 10^9 for the largest.

2. **Diversity:** Our corpus has languages belonging to a wide variety of scripts and linguistic families. These characteristics of our dataset make the problem that we study as realistic as possible. Models trained on this massive open-domain dataset are expected to yield rich, complex representations which we attempt to study in this paper.

To minimize confounding factors and control the evaluation set size and domain, we created a multi-way aligned evaluation set containing nearly 3k sentence pairs for all languages.[3] This also allows us to analyze the representational difference and similarity while controlling for semantics.

We use the Transformer-Big (Vaswani et al., 2017) architecture described in Artetxe and Schwenk (2018) for our experiments and share all parameters across language pairs including softmax layer and input/output word embeddings. For vocabulary, we use a Sentence Piece Model (Kudo and Richardson, 2018) with 64k tokens shared on both the encoder and decoder side.

Each set of parallel sentences has a `<2xx>` token prepended to the source sentence to indicate the target language, as in (Johnson et al., 2017). [4]

2.2 SVCCA

In this study we use Singular Value Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017)

[1] Tools for online visualization and representation similarity used in our work will be open-sourced to facilitate further analysis.

[2] We provide baseline BLEU scores for all languages in Appendix A.2, notice the improvements for low-resource languages in our multilingual setup.

[3] Each sentence in our evaluation set is semantically identical across all other languages.

[4] Further details of the model and training routines are given in Appendix A.1.

as our investigation tool. SVCCA is a technique to compare vector representations in a way that is both invariant to affine transformations and fast to compute. We express a layer's representations as its activations on a set of n examples, $X = \{x_1, \ldots, x_n\}$. Let $l_1 \in \mathbb{R}^{n \times d_1}$ and $l_2 \in \mathbb{R}^{n \times d_2}$ be the representations of two neural network layers, with d_1 and d_2 being the dimensions of the layers corresponding to l_1 and l_2 respectively. Given layer activations over the set X, SVCCA does the following:

1. Computes SVD decompositions of l_1 and l_2 to get subspaces $l'_1 \in \mathbb{R}^{n \times d'_1}$ and $l'_2 \in \mathbb{R}^{n \times d'_2}$ that capture most of the variance in the subspace.[5]

2. Uses Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) to linearly transform l'_1 and l'_2 to be as aligned as possible, i.e., CCA computes $\tilde{l_1} = W_1 l'_1$ and $\tilde{l_2} = W_2 l'_2$ to maximize the correlations $\bar{\rho} = \{\rho_1, ..., \rho_{min(d'_1, d'_2)}\}$ between the new subspaces.

As done in (Raghu et al., 2017), we use the mean of the correlations, $\bar{\rho}$, as an approximate measure of the relatedness of representations.

SVCCA for Sequences

Recent work on interpretability for NLU tasks uses methods such as diagnostic tasks (Belinkov et al., 2017; Tenney et al., 2019; Belinkov et al., 2018), attention based methods (Raganato and Tiedemann, 2018) or task analysis (Zhang and Bowman, 2018) and is primarily focused on understanding the linguistic features encoded by a trained model. Some recent work has applied SVCCA (or CCA) to language modeling (Morcos et al., 2018; Saphra and Lopez, 2018; Dalvi et al., 2019) and NMT (Bau et al., 2018; Dalvi et al., 2019). However, while Raghu et al. (2017) compare the learning dynamics of different classes in an image classifier, to the best of our knowledge, we are the first to apply SVCCA to a multilingual NMT or multitask setting. SVCCA was originally proposed for feed-forward neural networks, but our domain requires comparing unaligned sequences in different languages.

Sahbi (2018) proposes an alignment agnostic CCA approach to comparing unaligned data. However, the likelihood matrix D specifying alignment of datapoints across different datasets (say, language A and B) is application specific and infeasible to obtain in a multilingual setting. A simple heuristic is to summarize a set of activations by applying a pooling operation over the set. This is equivalent to assuming that a given token in a sentence from language A is equally likely to be aligned to each token in an equivalent sentence in language B. We perform SVCCA on the hidden representations of the model, averaged over sequence time-steps, for each sentence in our evaluation set. We compare this approach with a naive token level SVCCA strategy in A.3.

SVCCA Across Languages

In all known work using SVCCA, representations of the same data are used for analysis. However, in order to compare representations across languages, we leverage our multi-way parallel evaluation set to compare representations across different languages, since each data point is semantically equivalent across languages.

3 Multilingual NMT Learns Language Similarity

In this section, we use SVCCA to examine the relationship between representations of different languages learned by our massively multilingual NMT model. We compute SVCCA scores of layer-wise activations of a fully trained model between 103 languages pairs in both the Any-to-English and English-to-Any directions.[6]

Visualizing the Representations We first visualize the relationship between languages in their representation space for each layer using Spectral Embeddings (Belkin and Niyogi, 2003) [7]. In our case, we use mean SVCCA scores described in Section 2.2 as a similarity measure. Due to the differing nature of translating multiple languages to English and vice versa, the representation space of these two sets of languages, All-to-English and English-to-Any, behave differently and separate quite clearly (Figure 11 in the Appendix). We

[5] We retain 99% of the variance in our studies.

[6] Our multilingual NMT model is trained on the available training data which is English centric, hence an All-to-All multilingual model internally decomposes into All-to-English (X-En) and English-to-All (En-X) translation bundles, excluding zero-shot directions.

[7] We use the Laplacian Eigenmaps implementation of Spectral Embeddings in scikit-learn as of August 2019.

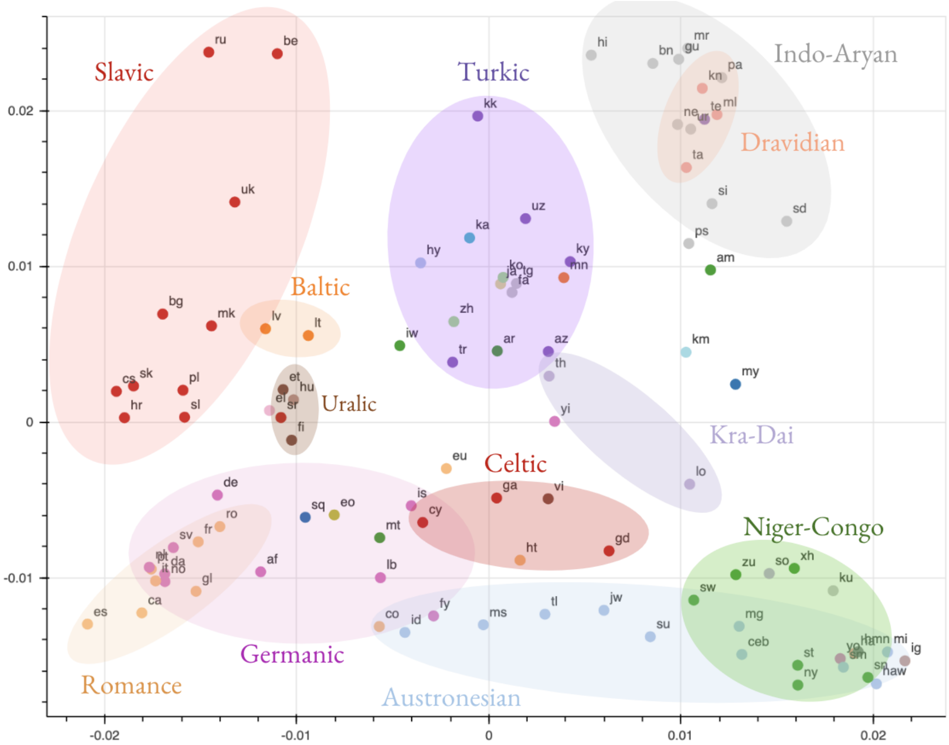


Figure 2: Visualizing clustering of the encoder representations of all languages, based on their SVCCA similarity. Languages are color-coded by their linguistic family. Best viewed in color.

first visualize the encoder representation of all languages in the All-to-English language pair set in Figure 2. For the sake of analysis, we then visualize subsets of the aforementioned 103 languages in Figures 3 and 4. We include visualizations of representations extracted from the embeddings and top layers of the encoder and decoder in the Appendix.

3.1 What is Language Similarity?

In the following sections we draw comparisons between the representational similarity of languages learned by our models, and the linguistic similarity between those languages. While there are entire sub-fields in linguistics devoted to studying similarity (e.g. Comparative Linguistics and Linguistic Typology (Ivić, 2011)), in this paper we define language similarity in terms of membership in the same language family (e.g. Turkic languages), or branch within that family (e.g. Oghuz Turkic languages). Families are groups of languages believed to share a common ancestor, and therefore tend to have similar vocabulary and grammatical constructs. An example phylogenetic language tree is given in Figure 3, on the right.

We also discuss writing systems, including scripts like Cyrillic, Roman, and Ge'ez. While similar languages frequently share the same script, that is not always true. Note that all of these categories[8] are muddled by various factors that are difficult to tease apart, and might be affected by the web-crawled data that we train on. For instance, languages sharing a script may also be part of the same political bloc, influencing what text is on the web. This and other confounding factors make a rigorous comparison exceedingly difficult. For brevity, we label languages in images with their BCP-47 language codes (Phillips and Davis, 2009), which are enumerated in the Appendix, Table 3.

3.2 Representations cluster by language similarity

We first visualize a clustering for all languages together in Figure 2. While there are a few outliers, we can observe some overlapping clusters, including the Slavic cluster on the top-left, the Germanic and Romance clusters on the bottom-left, the Indo-Aryan and Dravidian clusters on the top-right, etc. To analyze language clustering in more detail we

[8] When we refer to languages from a certain category, we only list those that are in our dataset. For example, when listing Turkic languages we exclude Bashkir, because we do not have English-Bashkir parallel data.

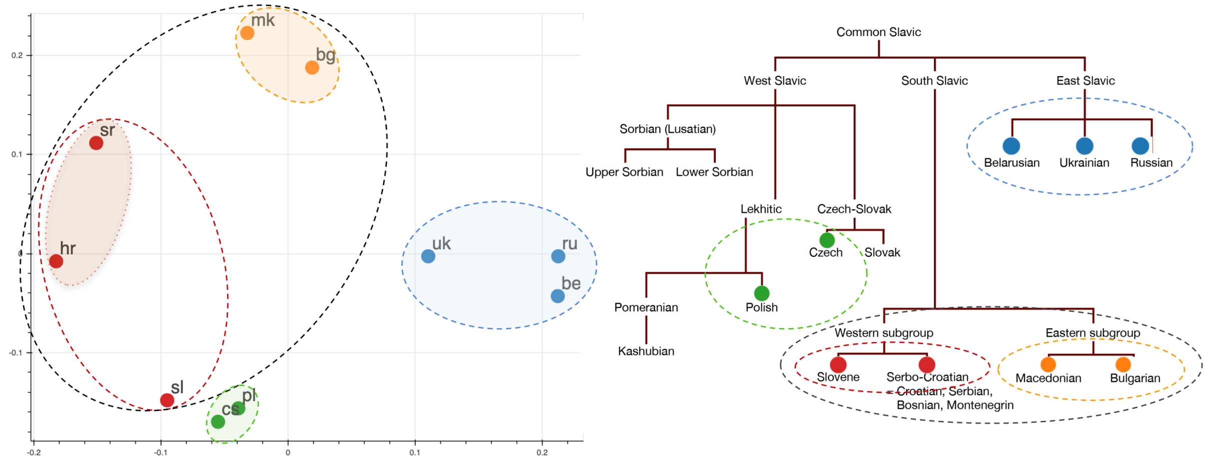


Figure 3: Comparing clusterings in the subword embeddings of the Slavic languages in our dataset with their family tree, which is the result of centuries of scholarship by linguists ((Browne and Ivanov)). Best seen in color.

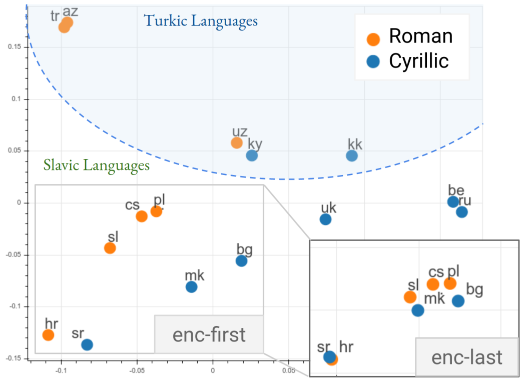


Figure 4: Representations of the Turkic and Slavic languages at the subword embeddings, compared with the top of the encoder, and colored by script. The inset shows a portion of the same image at the top of the encoder, focusing on the South-Western Slavic languages. Both images are at the same scale, see Appendix-Fig. 14 for more details. Best seen in color.

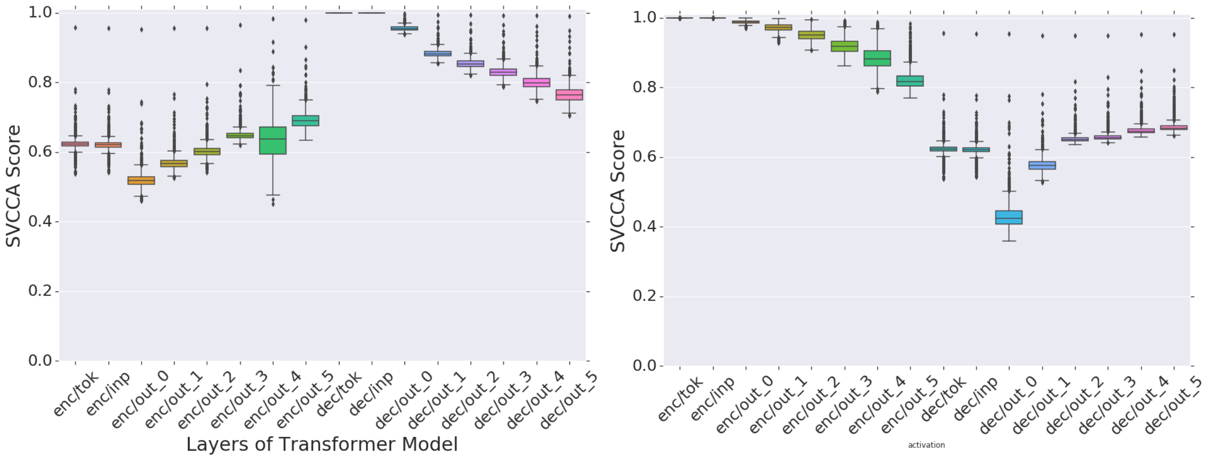
visualize sub-sets of the above languages.

In Figures 3 and 4, we visualize the Slavic and Turkic languages in our dataset. These languages come from two distinct families with very different linguistic properties, and within each family there are languages that are written in Cyrillic and Roman alphabets. This makes them ideal for understanding the interaction between superficial similarity (having the same alphabet and thus sharing many subwords) and linguistic similarity (sharing similar grammatical properties).

The first remarkable phenomenon we observe is that the clusters resulting from our model are grouped not only by family (Slavic), but branches within it (e.g. South Slavic), branches within those branches (e.g. Western Subgroup), and dialects within those (e.g. Serbo-Croatian). Figure 3 provides a more detailed look into the Slavic languages, and how this compositionality maps to the established family tree for Slavic languages. As can be seen in Figure 4, this phenomenon can also be observed for Turkic languages, with the Oghuz languages (Turkish and Azeri) forming one cluster, and the two Eastern branches Kipchak and Karluk (Uzbek, Kyrgyz, Kazakh) forming another.

A point worth special notice is the closeness between Serbian (sr) and Croatian (hr). Although these two are widely considered registers of the same language (Sussex and Cubberley), Serbian is written in Cyrillic, whereas Croatian is written in the Roman script. However, we see in both Figure 3 (middle left of plot) and Figure 4 (bottom left of plot) that they are each others' closest neighbors. Since they have no overlap in subword vocabulary, we conclude that they cluster purely based on distributional similarity – even at the level of sub-word embeddings.

Although we see strong clustering by linguistic family, we also notice the importance of script and lexical overlap, especially (and unsurprisingly) in the embeddings. In Figure 4 we visualize the Turkic and Slavic languages, and color by script. Although the linguistic cluster is probably stronger, there is also a distinct grouping by script, with the Roman-scripted languages on the left and the Cyrillic-scripted languages on the right. However, as we move up the encoder, the script associations become weaker and the language family as-



(a) X-En Language Pairs

(b) En-X Language Pairs

Figure 5: The change in distribution of pairwise SVCCA scores between language pairs across layers of a multilingual NMT model, with SVCCA scores between English-to-Any and Any-to-English language pairs visualized separately. We see that while the encoder in (a) and decoder in (b) have dissimilar representations across languages, the English representations of the decoder in (a) and the encoder in (b) diverge depending on the language X.

sociations become stronger. The inset in Figure 4 shows the seven South-Western Slavic languages at the top of the encoder, where they have clustered closer together. Again, Serbian and Croatian are an excellent example: *by the top of the encoder, they have become superimposed, diminishing the effect of the difference in script*.

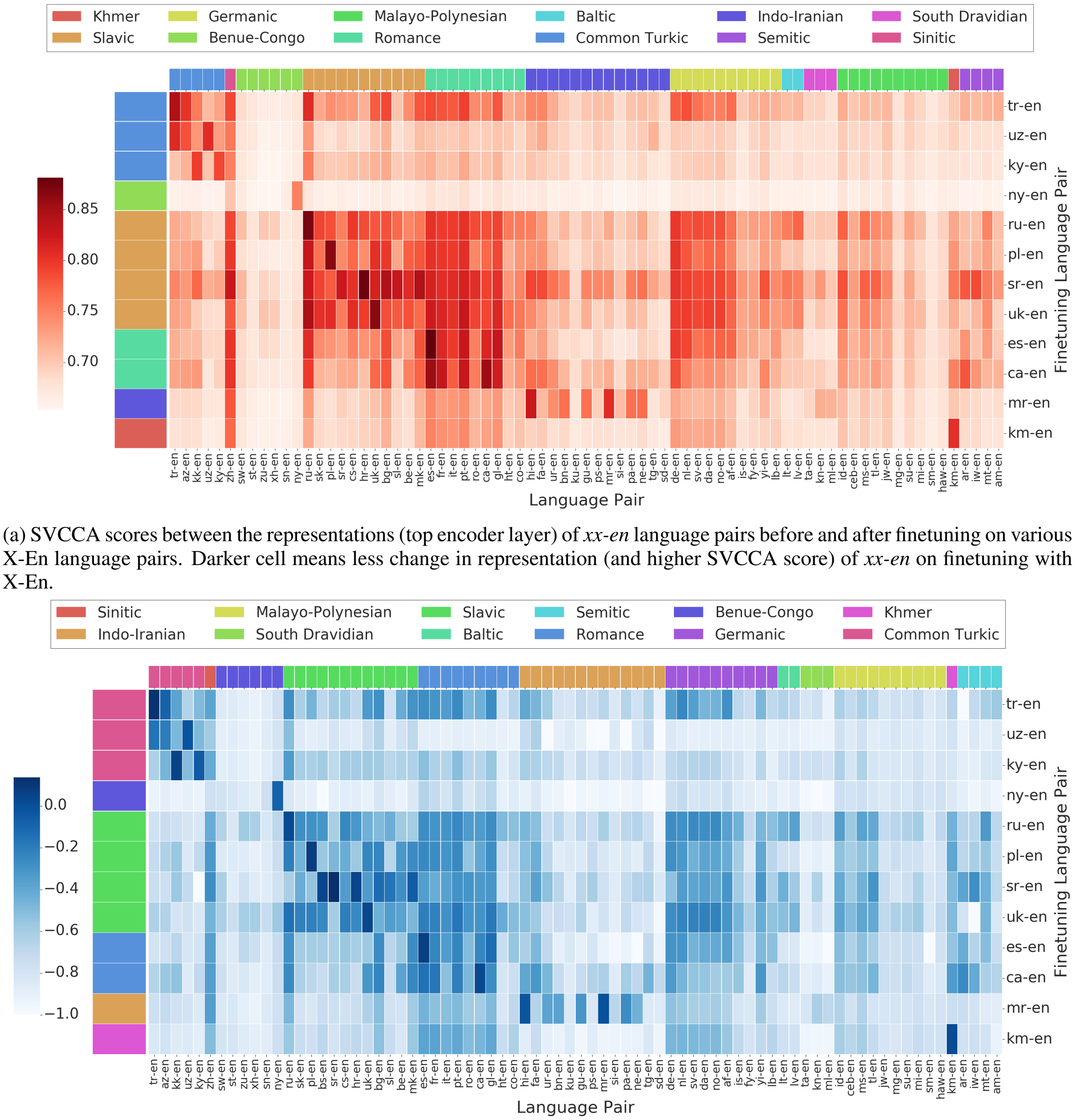
We find that the trends discussed above are generally true for other language groupings too. The Appendix shows an example with the Dravidian, Indo-Aryan, and Iranian language families, demonstrating the same phenomena discussed above (Appendix Figure 12). Section A.5 further analyzes how the nearest neighbors of languages in SVCCA space become more linguistically coherent as one moves up the encoder.

3.3 Representational Similarity evolves across Layers

To visualize the how the representational overlap across languages evolves in the model, we plot how the distribution of pairwise SVCCA scores change across layers. For each layer, we first compute the pair-wise similarity between all pairs of languages. These similarity scores are then aggregated into a distribution and represented in Figures 5a and 5b, separately for the Any-to-English (X-En) and English-to-Any (En-X) language pairs.

For the Any-to-English (X-En) language pairs (Figure 5a), we notice that similarity between the source languages (X) increase as we move up the encoder, from embeddings towards higher level encoder layers, suggesting that the encoder attempts to learn a common representation for all source languages. This might also be motivated by training on a shared target language (En). However, representations at the top of the encoder are far from perfectly aligned, possibly indicating that different languages are represented in only partially overlapping sub-spaces. On the other hand, as the decoder incorporates more information from the source language (X), representations of the target (En) diverge. This is also in line with some findings of studies on translationese (Koppel and Ordan, 2011), where the authors show that that the translated text is predictive of the source language. For English-to-Any (En-X) language pairs (Figure 5b) we observe a similar trend. Representations of the source language (En) diverge as we move up the encoder, indicating that the representations of English sentences separate conditioned on the target language.

While it is a natural assumption that the encoder in a seq2seq model encodes the source, and the decoder decodes it into the target (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015), our results indicate that this change from source to target might be more gradual, and the boundary between encoder and decoder, in terms of the localization of the representation is blurry.



(a) SVCCA scores between the representations (top encoder layer) of *xx-en* language pairs before and after finetuning on various X-En language pairs. Darker cell means less change in representation (and higher SVCCA score) of *xx-en* on finetuning with X-En.

(b) Relative change in BLEU scores of *xx-en* language pairs after finetuning with various language pairs of the form X-En. Darker cell means less change in BLEU score of *xx-en* on finetuning with X-En.

Figure 6: Visualization depicting the (a) change in representations (using SVCCA) and (b) relative change in performance (in terms of test BLEU) of *xx-en* language pairs (x-axis), after fine-tuning a large multilingual model on various X-En language pairs (y-axis). Language sub-families on the axes are color-coded. We notice that representations of high resource languages are relatively robust to fine-tuning on any language. Additionally, languages linguistically similar to the one being fine-tuned on observe less distortion from their original representations.

4 Analyzing representation robustness to fine-tuning

In this section, we try to analyze the robustness of encoder representations, when fine-tuning a multilingual model on a single language pair. Note that here we define robustness to mean representational rigidity to fine-tuning, aka robustness to catastrophic forgetting (McCloskey and Cohen, 1989). Understanding the factors that affect robustness during fine-tuning is critical to determining how much cross-lingual transfer can be expected for individual languages in a zero or few-shot setting.

Analysis Setup: We fine-tune a fully trained multilingual model separately on 12 Any-to-English language pairs for 40k steps using the optimizer as described in A.1. These selected languages form a mix of high and low resource language pairs, from 6 language sub-families.[9]

We first attempt to quantify the extent of distortion in language representations caused by the fine-tuning process. To this end, we calculate the SVCCA similarity between the encoder representations of a language, before and after fine-tuning. We do this for all languages, in order to understand which factors determine the extent of distortion. We visualize these changes in Figure 6a for the final encoder layer, for all X-En language pairs. To complement our analysis of representational similarity, we visualize the relative change in BLEU score after fine-tuning in Figure 6b.

Observations: The first observation from Figures 6a and 6b is that the variations in SVCCA scores correspond very well with changes in BLEU; degradation in translation quality is strongly correlated with the magnitude of change in representations during fine-tuning.

We find that representations of high resource languages are quite robust to fine-tuning on any language. In Figure 6a, we see that high resource languages such as Chinese, German, Russian and Italian do not change much, irrespective of the language the model is fine-tuned on.

In general, we find that language representations are relatively robust to fine-tuning on a language pair from the same linguistic family. For example, on fine-tuning with tr-en (Turkish) or ky-en (Kyrgyz), the Turkic language group does not experience much shift in representation. We see a similar pattern with models fine-tuned on es-en (Spanish), ca-en (Catalan) and the Romance languages, uk-en (Ukranian), sr-en (Serbian), ru-en (Russian) and the Slavic languages.

An exception to these general trends seems to be fine-tuning on ny-en (Nyanja: Benue-Congo sub-family); all language pairs degrade by roughly the same extent, irrespective of language similarity or resource size. It's worth noting that all of the languages from the Benue-Congo sub-family are low-resource in our corpus.

These observations suggest that high resource languages might be responsible for partitioning the representation space, while low-resource languages become closely intertwined with linguistically similar high-resource languages. Low resource languages unrelated to any high resource languages have representations spread out across multiple partitions.

While these observations are based on representations from the top of the encoder, we further analyze sensitivity of representations to fine-tuning across different layers in the Appendix A.6. One key observation from that analysis is the robustness of embeddings to fine-tuning on any individual language; there is no significant change in the embedding representations (Correlation between representation of any language before and after finetuning $\bar{\rho} > 0.98$).

5 Discussion

Our work uncovers a few observations that might be of interest to practitioners working in multilingual NMT and cross-lingual transfer.

Our analysis reveals that language representations cluster based on language similarity. While language similarity has been exploited for adaptation previously (Neubig and Hu, 2018), our work is the first to concretely highlight which factors affect the overlap in representations across languages. This has potential implications for transfer learning: for example, it is possible to identify and exploit the nearest neighbors of a low resource language to maximize adaptation performance.

We also highlight how representation overlap evolves across layers, which is, again, of interest for cross-lingual transfer. For example, our analysis reveals that embeddings of different languages are less overlapping than the final encoder outputs. This hints that it might not be might not be effective to utilize input embeddings learned in multilingual NMT, since they don't overlap as much as the final encoder outputs. We also notice that encoder representation overlap across languages is not perfect, which explains why explicit language alignment or consistency losses might be needed to enable zero-shot NMT (Arivazhagan et al., 2019a; Al-Shedivat and Parikh, 2019).

We further analyze the robustness of language representations to fine-tuning, and notice that high-resource and linguistically similar languages are more robust to fine-tuning on an arbitrary language. This might help explain why linguistically distant languages typically result in poor zero-shot transfer (Wu and Dredze, 2019). Applying explicit losses, like elastic-weight consolidation (Kirkpatrick et al., 2017), to force language rep-

[9] More details on the relative resource sizes of different language pairs can be found in the Appendix A.6.

resentations of distant languages from getting distorted might help improve transfer performance.

6 Conclusion

To conclude, we analyzed factors that affect the overlap in representations learned by multilingual NMT models. We used SVCCA to show that multilingual neural networks share representations across languages strongly along the lines of linguistic similarity, and encoder representations diverge based on the target language. With this work we hope to inspire future work on understanding multitask and multilingual NLP models.

Acknowledgments

First, we would like to thank Naveen Arivazhagan for foundational contributions to the project. We would like to thank Aditya Siddhant, Dmitry (Dima) Lepikhin, Klaus Macherey and Macduff Hughes for their helpful feedback on the draft. We would also like to thank Maithra Raghu for multiple insightful discussions about our methodology.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Maruan Al-Shedivat and Ankur P. Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. *CoRR*, abs/1904.02338.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv preprint arXiv:1801.07772*.

Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Wayles Browne and Vyacheslav Vsevolodovich Ivanov. Slavic languages' family tree. Encyclopdia Britannica.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 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, Melbourne, Australia. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Sandagomi Coperahawa. 2007. Language Contact and Linguistic Area: the Sinhala - Tamil Contact Situation. *Journal of the Royal Asiatic Society of Sri Lanka*, 53:133 – 152.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, D Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 7.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. *Hindustani*.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Marvin Irving Herzog. West Germanic Languages.

Pavle Ivić. 2011. Encyclopaedia Britannica.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1318–1326. Association for Computational Linguistics.

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

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Etukoori Balaraama Moorti. 2011. Proto Dravidian. *Study of Dravidian Linguistics and Civilization*.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 5732–5741, USA. Curran Associates Inc.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proc. IJCNLP*, volume 2, pages 296–301.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill*, 3(3):e10.

A. Phillips and M Davis. 2009. Tags for Identifying Languages. RFC 5646, RFC Editor.

Alessandro Raganato and Jörg Tiedemann. 2018. 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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6076–6085.

Hichem Sahbi. 2018. Learning cca representations for misaligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*.

Mohammad Tahsin Siddiqi. 1994. *Hindustani-English code-mixing in modern literary texts*. University of Wisconsin.

Roland Sussex and Paul Cubberley. The Slavic Languages.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Jakob Uszkoreit, Jay M Ponte, Ashok C Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1101–1109. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *CoRR*, abs/1904.09077.

Kelly W Zhang and Samuel R Bowman. 2018. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv preprint arXiv:1809.10040*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.