An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation

Raj Dabre
Kyoto University,
Kyoto, Japan
prajdabre@gmail.com

Tetsuji Nakagawa
Google Japan,
Tokyo, Japan
tnaka@google.com

Hideto Kazawa
Google Japan,
Tokyo, Japan
kazawa@google.com

Abstract

Neural Machine Translation (NMT) is known to outperform Phrase Based Statistical Machine Translation (PBSMT) for resource rich language pairs but not for resource poor ones. Transfer Learning (Zoph et al., 2016) is a simple approach in which we can simply initialize an NMT model (child model) for a resource poor language pair using a previously trained model (parent model) for a resource rich language pair where the target languages are the same. This paper explores how different choices of parent models affect the performance of child models. We empirically show that using a parent model with the source language falling in the same or linguistically similar language family as the source language of the child model is the best.

1 Introduction

One of the most attractive features of Neural Machine Translation (NMT) (Bahdanau et al., 2015; Cho et al., 2014; Sutskever et al., 2014) is that it is possible to train an end to end system without the need to deal with word alignments, phrase tables and complicated decoding algorithms which are a characteristic of Phrase Based Statistical Machine Translation (PBSMT) systems (Koehn et al., 2003). It is reported that NMT works better than PBSMT only when there is an abundance of parallel corpora. In the case of low resource languages like Hausa, vanilla NMT is either worse than or comparable to PBSMT (Zoph et al., 2016). However, it is possible to use a previously trained X-Y model (parent model; X-Y being the resource rich language pair where X and Y represent the source and target languages respectively) to initialize the parameters of a Z-Y model (child model; Z-Y being the resource poor language pair) leading to significant improvements (Zoph et al., 2016) for the latter. This paper is about an empirical study of transfer learning for NMT for low resource languages. Our main focus is on translation to English for the following low resource languages: Hausa, Uzbek, Marathi, Malayalam, Punjabi, Malayalam, Kazakh, Luxembourgish, Javanese and Sundanese. Our main contribution is that we empirically (and exhaustively; within reason) show that using a resource rich language pair in which the source language is linguistically closer to the source language of the resource poor pair is much better than other choices of language pairs.

2 Related Work

Transfer learning for NMT (Zoph et al., 2016) is an approach where previously trained NMT models for French and German to English (resource rich pairs) were used to initialize models for Hausa, Uzbek, Spanish to English (resource poor pairs). They showed that French-English as a parent model was better than German-English when trying to improve the Spanish-English translation quality (since Spanish is linguistically closer to French than German) but they did not conduct an exhaustive investigation for multiple language pairs. In this paper we extend this work to explore how language relatedness impacts transfer learning.

3 Overview of Transfer Learning

Refer to Figure 1 for an overview of the method. It is essentially the same as described in (Zoph et al., 2016) where we learn a model (parent model) for a resource rich language pair (Hindi-English) and use it to initialize the model (child model) for the resource poor pair (Marathi-English). Henceforth the source languages of the parent model and

31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31), pages 282–286
Cebu City, Philippines, November 16-18, 2017

child models will be known as parent and child languages respectively and the corresponding language pairs will be known as the parent and child language pairs respectively. The target language vocabulary (English) should be the same for both the parent and the child models. Following the originally proposed method we focused on freezing[1] (by setting gradients to zero) the decoder embeddings and softmax layers when learning child models since they represent the majority of the decoder parameter space. This method can easily be applied in cases where we wish to use the X-Y pair to help the Z-Y pair where Y is usually English.

4 Experimental Settings

All of our experiments were performed using an encoder-decoder NMT system with attention for the various baselines and transfer learning experiments. We used an in house NMT system developed using the Tensorflow (Abadi et al., 2015) framework so as to exploit multiple GPUs to speed up training. To ensure replicability we use the same NMT model design as in the original work (Zoph et al., 2016). In order to enable infinite vocabulary we use the word piece model (WPM) (Schuster and Nakajima, 2012) as a segmentation model which is closely related to the Byte Pair Encoding (BPE) based segmentation approach (Sennrich et al., 2016). We evaluate our models using the standard BLEU (Papineni et al., 2002) metric[2] on the detokenized translations of the test set. However we report the only the difference between the BLEU scores of the transferred and the baseline models since our focus is not on the BLEU scores themselves but rather the improvement by using transfer learning and on observing the language relatedness phenomenon. Baseline models are simply ones trained from scratch by initializing the model parameters with random values.

4.1 Languages

The set of parent languages (and abbreviations) we considered is: Hindi (Hi), Indonesian (Id), Turkish (Tr), Russian (Ru), German (De) and French (Fr). The set of child languages (and abbreviations) consists of: Luxembourgish (Lb), Hausa (Ha), Somali (So), Malayalam (Ml), Punjabi (Pa), Marathi (Mr), Uzbek (Uz), Javanese (Jw), Kazakah (Kk) and Sundanese (Su). Table 1 groups the languages into language families. For each child model we try around 3 to 4 parent models out of which one is mostly learned from a linguistically close parent language pair. The source languages vary but the target language is always English. Since there are no standard training sets for many of these language pairs, we use parallel data automatically mined from the web using an in-house crawler. For evaluation, we use a set of 9K English sentences collected from the web and translated by humans into each of the source languages mentioned above. Each sentence has one reference translation. We use 5K sentences for evaluation and the rest form the development set.

Group	Languages
European	French, German, Luxembourgish
Slavic	Russian
Afro-Asiatic	Hausa, Somali
Turkic	Turkish, Uzbek, Kazakh
Austronesian	Indonesian, Javanese, Sundanese
Indian	Hindi, Marathi, Punjabi, Malayalam

Table 1: Language Groups in Experiments

To give a rough idea of the corpora sizes consider the WMT14 dataset for German-English which contains around 5M lines of parallel corpora for training. The child language pair corpora sizes vary from being one decimal order of magnitude smaller to one decimal order of magnitude larger than the WMT14 German-English corpus. However the parent language pair corpora are two to three decimal orders of magnitude larger than the aforementioned dataset. From left to right, the languages above are ordered according to the size of their corpora with the leftmost being the one with the smallest dataset. Since these datasets are mined from the open web they represent a realistic scenario and hence it should be evident that the child language pairs are truly resource poor.

Our choice of languages was influenced by two factors:

- a. We wanted to replicate the basic transfer learning results (Zoph et al., 2016) and hence chose French, German for Hausa and Uzbek.
- b. We wanted to compare the effects of using parent languages belonging to the same lan-

[1] We also tried experiments where we froze the decoder LSTM layers as well but we omit the results for brevity.

[2] This is computed by the multi-bleu.pl script, which can be downloaded from the public implementation of Moses (Koehn et al., 2007).

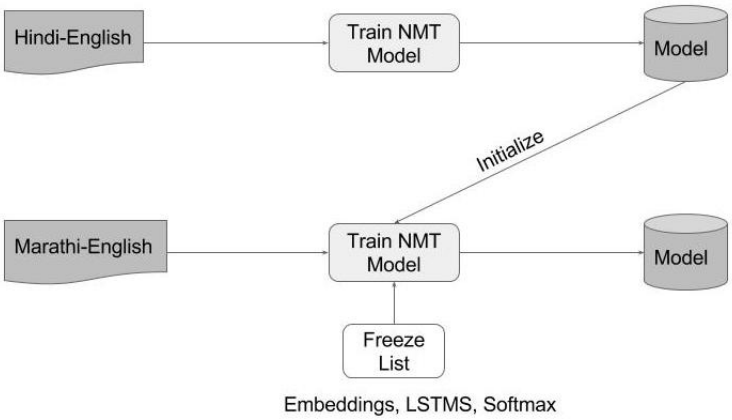

Figure 1: Transfer Learning for Low Resource Languages

guage family as the child languages (Hindi for Marathi) as opposed to unrelated parent languages (German for Marathi).

4.2 Settings

Following the aforementioned factors influencing our language choices we conducted our experiments in two stages as below:

- Exhaustive experimentation on 6 child languages (Hausa, Uzbek, Marathi, Malayalam, Punjabi and Somali) by using 4 parent languages (French, German, Russian and Hindi). This was done in order to verify whether there is any language relatedness phenomenon worth exploring or not. Based on these experiments we proposed a hypothesis that a parent language from the same or a closely related language family should be a lot more helpful than any other parent language.
- Opportunistic experimentation on 4 child languages (Kazakh, Javanese, Sundanese and Luxembourgish) by using 3 parent languages out of which one is from the same language family and the other two are from another language family. Turkish being the related language for Kazakh, German for Luxembourgish and Indonesian for Javanese and Sundanese.

The model and training details are the same as that in the original work (Zoph et al., 2016) but following are some specific settings:

- Model parts frozen (only when doing transfer learning): softmax and decoder embeddings layers (Decoder LSTMs were retrained)
- Embeddings: 512 nodes
- LSTM: 4 layers, 512 nodes output
- Attention: 512 nodes hidden layer

Child	Parent			
	Fr	De	Hi	Ru
Ha	+2.85	+2.17	+2.03	**+2.99**
Uz	+0.12	+0.22	**+0.46**	+0.34
Mr	-1.62	-0.38	**+0.57***	-0.55
Ml	+1.31	+1.89	**+2.80***	+1.45
Pa	+0.80	+0.67	**+2.41***	+0.69
So	**+3.17**	+2.69	+2.26	+2.89

Table 2: BLEU deltas for Exhaustive experimentation

- WPM vocabulary size: 16k (separate models for source and target)
- Batch size: 128
- Training steps: 5M
- Optimization algorithms: Adam for 60k iterations followed by SGD
- Annealing: Starts at 2M iterations followed by halving learning rate every 200k iterations
- Choosing the best model: Evaluate saved checkpoints on the development set and select checkpoint with best BLEU.

Note that the target language (English) vocabulary is same for all settings and the WPM is learned on the English side of the French-English corpus since it is the largest one amongst all our pairs. We deliberately chose this since we wished to maintain the same target side vocabulary for all our experiments (both baseline and transfer) for fair comparison. The parent source vocabulary (and hence embeddings) is randomly mapped to child source vocabulary since it was shown that NMT is less sensitive to it (Zoph et al., 2016).

Child	Parent			
	De	Hi	Tr	Id
Kk	+0.21	+0.40	**+0.48**	-
Jw	+1.10	+0.44	-	**+2.47***
Su	-0.13	+0.41	-	**+1.10***
Lb	**+8.58***	+6.44	+6.01	-

Table 3: BLEU deltas for Opportunistic experimentation

5 Results

Refer to Table 2 for the results of the exhaustive experimentation round and Table 3 for those of the opportunistic experimentation round. As mentioned before we only report the difference between the BLEU scores of the transferred and the baseline model. Entries in bold indicate the parent-child pair that performed the best amongst others. Furthermore, entries that have an "*" mark represent the parent-child pair with a BLEU difference that is statistically significant compared to the BLEU difference of other parent-child pairs.

5.1 Observations

One thing that stood out during the exhaustive experimentation phase (Table 2) is that Hindi as a parent language led to better gains (from +0.57 to +2.8) for all Indian languages as opposed to gains (-1.62 to +1.89) due to other parents. In the case of Marathi all other parent languages led to degradation in performance and Punjabi gained the most (+2.41) from Hindi as a parent where as the gains due to the others were at most +0.8. It makes sense that Punjabi being the closest language (linguistically speaking) to Hindi would gain the most followed by Marathi. It is also important to note that amongst all parent languages Hindi had the least amount of data and French had the most. This led us to believe that beyond a certain amount the size of the training data is not the real factor behind the gains observed due to transfer learning. Amongst the child languages Uzbek and Marathi were the most resource abundant ones and hence the gains to the transfer learning (less than 1 BLEU point) are notable only in cases where the baseline systems are not that strong.

Following this we decided to verify our hypothesis that: "A parent language from the same (or linguistically similar) language family as the child language will have a larger impact on transfer learning." From Table 3 it can be seen that this hypothesis is mostly true. The gain (+8.58) in the case of German as a parent for Luxembourgish is quite striking since the latter is known to be closely related to the former. Moreover using German gives an additional improvement of around 2 BLEU points over other parents. Indonesian, Javanese and Sundanese are close to each other in the same way that Punjabi is similar to Hindi. Thus Indonesian as a parent gives around 1 to 2 BLEU improvement for these language pairs over when other parents are chosen. Indonesian, Javanese and Sundanese use the same script but Hindi and Punjabi do not. In spite of this Hindi still acts as a better parent as compared to the others which means that the NMT system does learn certain grammatical features which provide the child models with a good prior when transferring the parameters. Finally, Kazakh received maximum benefit when using Turkish as a parent but the baseline model for Kazakh was too strong and thus it is difficult to draw any proper conclusion in this case since Hindi as a parent helped almost as much. We did try a scenario where Turkish was used as a parent for Uzbek (not in the tables) but failed to see any particular improvement over when other parents are used but it should be noted that, linguistically speaking, Turkish is a lot closer to Kazakh than it is to Uzbek.

Although we do not give details here due to lack of space transfer learning helps cut down the training time by more than half in most cases since more than half the model is already pre-trained.

6 Conclusions and Future Work

We presented our work on an empirical study of language relatedness for transfer learning in Neural Machine Translation. We showed that in general, transfer learning done on a X-Y language pair to Z-Y language pair has maximum impact when Z-Y is resource scarce and when X and Z fall in the same or linguistically similar language family. We did exhaustive experimentation to validate our hypothesis and it stands to be true in most cases. In the future we would like to experiment with transfer learning where we use Spanish as a parent for Italian with a slight modification where we force the Spanish vocabulary to resemble Italian by applying a segmentation mechanism (like BPE or WPM) trained on Italian to Spanish. This should help exploit cognates between closely related languages.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, USA. International Conference on Learning Representations.

Kyunghyun Cho, Bart van Merriënboer, ÇaÄ§lar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1568–1575.