A contrast based heuristic for choosing transfer languages for multi-source neural machine translation with a low resource language

Job Gräber, Wenhua Hu, Alex Labro, Oline Ranum, Julius Wagenbach

Abstract

We propose and test a heuristic for choosing transfer languages for joint learning with a low-resource language (LRL) in neural machine translation (NMT). NMT from a LRL benefits substantially from training jointly with one or more other languages, but efficiently choosing such languages for more than one transfer language is an open problem. LangRank uses learning-to-rank to efficiently choose a single language. We propose an efficient method of choosing pairs of languages by taking into account to which degree the languages get their relevance score from LangRank for similar reasons. We observed preliminary results to the conclusion that high-contrast pairs can benefit learning more than low-contrast pairs.

1 Introduction

Recently, neural machine translation (NMT) has made advancements (Lakew et al., 2018; Vaswani et al., 2017) driven by large sequence-to-sequence models utilizing the attention mechanism (Bahdanau et al., 2015). However, these models require extensive parallel corpora which are unavailable for low-resource languages (LRLs). To address limited corpus sizes, methods such as data augmentation (Nishimura et al., 2018) and multilingual joint learning (Zoph et al., 2016) emerged. They respectively increase the data set synthetically and combine linguistically related corpora during training. For the latter, a significant challenge is how to efficiently select related corpora (Lin et al., 2019) and how to extend methods to include multiple transfer languages.

In this report, we investigate a heuristic for choosing transfer language pairs for joint learning. The intuition is that there could be diminished marginal returns arising from the specific ways in which the transfer operates (e.g. similar vocabulary or syntax), and as such it could be beneficial to select pairs in which the languages contribute to learning for differing reasons.

2 Background

Several studies have focused on exploring ways to enhance translation performance for LRLs. Edunov et al. (2018) improved translation performance by increasing the training set using back-translation, a process that translates a target language's monolingual data set and uses it as training data. Currey et al. (2017) demonstrated that LRL translation performance can be improved with synthetic data where the source is simply a copy of monolingual target data.

Language transfer is another effective approach that is widely performed for MT tasks regarding LRLs. The multilingual NMT introduced in (Johnson et al., 2016) improved the translation quality and performed well on unseen pairs of languages. Neubig and Hu (2018) trained models for LRLs with additional data from the most linguistically related high-resource language using joint learning, which showed better performance. Instead of sampling from a single language, Wang and Neubig (2019) constructed a data selection strategy to sample data from multiple high-resource languages based on distribution and brought significant gains by 2 BLEU on several multilingual NMT tasks. Lin et al. (2019) introduced a learning-to-rank model (LangRank) to efficiently select transfer language. LangRank was trained using a single transfer language, and as such it is unclear how LangRank can generalize when using multiple transfer languages.

3 Methodology

We assess the selection of transfer languages using an attribution technique with LangRank by analyzing the contrastive attributions of languages. In particular, we analyze Shapley values and use the results to inform multilingual models learned with a joint training methodology using XNMT.

3.1 Joint training

We utilize a unified neural machine translation model to facilitate translations across various languages, adopting the method of (Johnson et al., 2016). Specifically, we prepend an artificial token to the start of the input sentence to denote the desired target language. As an illustration, consider how we adjust the corpus pair:

"*gracias → thank you*", to
"*__en__ gracias → thank you*".

3.2 Contrast in LangRank

LangRank provides relevance scores to candidate transfer languages for joint learning with a given source language (Lin et al. (2019)). It uses the learning-to-rank approach with dataset-dependent and dataset-independent ranking features.[1] LangRank uses LGBMRanker from the LightGBM framework which we also use to compute the Shapley values (Ke et al., 2017). We use Shapley values to disaggregate the contributions of the ranking features to the relevance score (Shapley, 1953). These values can be used to compute the contrast between the explanations for different predictions (Merrick and Taly, 2020). Such a contrast is the degree to which differences in the predicted relevance scores for two languages can be attributed to different kinds of features. More specifically, we quantify this metric for contrast as the (Euclidean) distance between the vectors of Shapley values for two candidate languages.

3.3 Low and high contrastive pairs

Using the metric mentioned above, we can find pairs of candidates with low and high contrast in the explanation for their relevance score. They are found based on analysis using a series of quantitative tools. Performing T-SNE on the Shapley values of transfer candidates for a source language gives a first indication of their distribution, as shown in Figure 1. Heatmaps of the contrast metric for the narrowed-down candidates were used to show and verify the actual contrast values. Bar plots allow inspection of the different distributions of Shapley values over the features. To isolate the effect of the contrast we only consider candidates for the pairings that have similar overall relevance scores.[2]

[1]The specific ranking features are explained in Table 9 of Appendix A.2.

[2]All bar and T-SNE plots corresponding to the languages we ended up using are in Appendix A.1 (See Figures 2-5 and 6-9). The exact relevance scores are reported in Table 3 in Appendix A.1.

After conducting this analysis, we have chosen the language pairs outlined in Table 1 as the transfer languages to be utilized in our multilingual models.

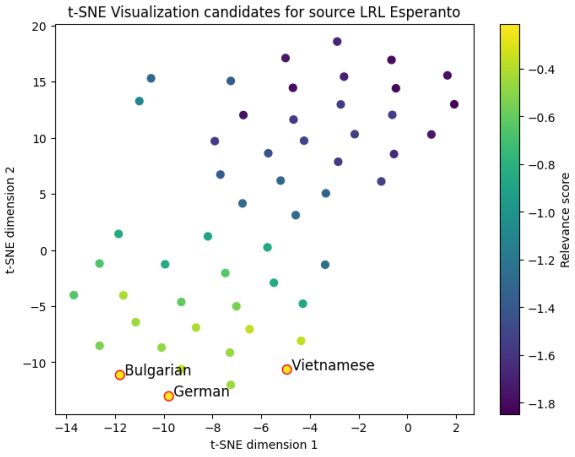

Figure 1: Example: T-SNE based on the Shapley values of all candidates for $EO \rightarrow EN$. The vertical direction can roughly be interpreted as variance in the overall score and the orthogonal horizontal direction as different ways of getting to the same score.

LRL	Contrast	Transfer Pair
AZ	High (1.0251)	Persian (FA)-Turkish (TU)
	Low (0.2673)	Persian (FA)-Hungarian (HU)
BE	High (0.8809)	Russian (RU)-Hungarian (HU)
	Low (0.2871)	Russian (RU)-Ukrainian (UK)
EO	High (0.2474)	German (DE)-Vietnamese (VI)
	Low (0.1210)	German (DE)-Bulgarian (BE)
ES	High (1.2036)	Portuguese (PT)-Kazakh (KA)
	Low (0.5095)	Portuguese (PT)-Galician (GA)

Table 1: Selected transfer language pairs for multilingual training. The source LRLs are respectively Azerbaijani, Belarusian, Esperanto and Spanish (synthetic). The numbers in the 'Contrast' column are the scores of our contrast metric rounded to the fourth decimal.

3.4 XNMT

We utilize the XNMT framework (Neubig et al., 2018) to train our models. More specifically, we employ a standard attention-based sequence-to-sequence model. Before training, the sentences are tokenized using the UnicodeTokenizer (Mielke and Eisner, 2019) provided by XNMT. After tokenization, the sentences are lowercased and filtered to contain a maximum sentence length of 60

words. Finally, during inference the output is detokenized using byte-pair encoding (Sennrich et al., 2016). We use the same hyperparameters as (Qi et al., 2018) in order to align our baseline with established literature.[3] We use the BLEU metric to assess the models (Papineni et al., 2002).[4]

4 Experiments

4.1 Dataset

We train our models using a parallel corpus from TED talk transcripts.[5] The languages and dimensions of the sampled subsets are presented in Table 7 in Appendix A.2. For Spanish and Esperanto, we synthesize a low-resource dataset by sampling a limited number of parallel sentences, either sequentially or randomly.

4.2 Bilingual baseline models

Initially, we train four bilingual baseline models targeting Belarusian, Azerbaijani, Esperanto, and a synthesized low-resource sample set of Spanish. To verify whether the data quality is uniform across the corpus, i.e. if sampling sequentially does not introduce additional bias, we train three baseline Spanish models where sentences are randomly sampled from the same corpus. The results presented are the mean and standard deviation obtained from three independently trained models. Lastly, we train a bilingual Spanish model using an equivalent number of sentence pairs to that used in training the multilingual models. This allows us to provide an additional assessment of the performance of models trained with transfer languages.

4.3 Multilingual models

We proceed by training multilingual models for each language group outlined in Table 1. We keep a relatively small data set size due to resource constraints that limit our computational capacity (see Table 7 in Appendix A.2 for further details). We jointly train the models on either one or two transfer languages to evaluate the benefit of learning from multiple languages simultaneously.

5 Results

The results from training the language models are presented in Table 2. The key observations from the results can be summarized as follows:

Model	SRC	Sampling	**BLEU**	**Relative Increase**
M^B_{BE}	BE	Sequential	1.24	
M^B_{AZ}	AZ	Sequential	1.29	
M^B_{EO}	EO	Sequential	2.50	
M^B_{ES}	ES	Sequential	1.39	
$M^B_{ES,F}$	ES	Sequential	15.85	
$M^B_{ES,R}$	ES	Random	1.37	($\pm$0.02)
M^M_{BE}	BE, RU	Sequential	**6.35**	4.12
$M^M_{BE,H}$	BE, RU, HU	Sequential	2.34	0.89
$M^M_{BE,L}$	BE, RU, UK	Sequential	3.96	2.19
M^M_{AZ}	AZ, FA	Seq.	1.46	0.13
$M^M_{AZ,H}$	AZ, FA, TU	Sequential	1.94	0.50
$M^M_{AZ,L}$	AZ, FA, HU	Sequential	**2.80**	1.17
M^M_{ES}	ES, PT	Sequential	16.6	
$M^M_{ES,H}$	ES, PT, GL	Sequential	10.11	6.27
$M^M_{ES,L}$	ES, PT, KA	Sequential	**13.35**	8.60
M^M_{EO}	EO, DE	Seq.	**10.61**	7.84
$M^M_{EO,H}$	EO, DE, VI	Random	7.65	2.06
$M^M_{EO,L}$	EO, DE, BG	Random	6.44	4.37

Table 2: BLEU scores for baseline bilingual models (B) and multilingual models (M). The subscripts indicate if the transfer languages were high in contrast H, low in contrast L and the LRL. F indicates that the model was trained using the same dataset size as in the multilingual models. For the Spanish baseline model $M^B_{ES,R}$ trained with a random corpus selection we report the mean and variance across 3 trained models.

i. In the case of EO, we observe that joint training with the high-contrast language pair yields better performance than training with the low-contrast pair. For all other baseline languages, we observe that training with the low-contrast pair outperforms training with the high-contrast pairs.

ii. All multilingual models achieve better performance than their respective baselines.

iii. Utilizing a singular transfer language over multiple typically yields higher performance, with the exception of Azerbaijani.

iv. We observed no consistent trend when training with languages from different writing systems. For Spanish (Latin script), training with Kazakh (Cyrillic script) showed notable improvement. However, Belarusian performed best when trained with only Cyrillic-based languages. See Table 8 in Appendix A.2.

[3] For the values see Table 5 in Appendix A.2

[4] Additional architecture and implementation details can be found in Table 6 in the Appendix A.2

[5] https://www.ted.com/participate/translate

v. We observed no consistent trends in the linguistic families of transfer languages. While Belarusian translation gained from using exclusively East Slavic languages, Azerbaijani and Spanish translation excelled when trained with languages of distinct linguistic families.

Minor observations include that our baseline BLEU scores align with results reported in literature (Qi et al., 2018). Qu et al. estimated baseline performances of 1.3 BLEU for (AZ → EN) and 1.6 for (BE → EN). Additionally, we make the observation that the variations across using random and sequential sampling methods are minor.

6 Discussion

We observe that our multilingual models outperform the bilingual models, similar to observations made in literature (Johnson et al., 2016). This implies that the models were capable of learning from the selected transfer language groups. In particular, the Spanish model trained alongside Portuguese and Kazakh achieved a BLEU score only two points lower than a model trained using only Spanish data (13.35 vs. 15.85). This suggests that multilingual transfer learning has the potential to greatly support the training of LRLs.

In our research, we aimed to develop a contrast-based heuristic for choosing transfer languages for multi-source NMT. However, in our experiments, we observed that training with both high-contrast and low-contrast language pairs could prove beneficial depending on the language group.

The greater performance of the high-contrast pair in the case of the EO models provides some evidence supporting the heuristic that high-contrast pairs are preferable, all other factors being equal. Most probably, the discrepancy with other language groups where training with low-contrast pairs yielded greater performance may be explained by a ranking bias. In the case of EO, the high-contrast pair includes VI which boasts a marginally higher overall relevance score compared to BG from the low-contrast pair. This introduces a minor relevance-score-based bias in favor of the high-contrast pair. However, for other source languages, the bias goes in the opposite direction and is up to almost two orders of magnitude greater.[6] This pronounced bias can explain why the low-contrast pair performs better. However, the bias favoring the high contrast pair in the case of EO is too small to solely account for the observed performance increase. This suggests that while high contrast can be advantageous, it does not strongly outweigh the significance of a language's overall relevance score.

Furthermore, we observed that using a single transfer language often yielded a higher BLEU score than using two languages. However, learning from two languages may still be beneficial when transferring from other low-resource languages.

A big limitation is that we tested only on a small number of data points due to the computational cost of the joint training. This also means that we did not obtain enough results to do a systematic analysis of the significance of feature-level patterns between the selected languages. Inspection of Figures 2-5 in Appendix A.1, for example, suggests the prevalence of contrast in sub-word level overlap but more results are needed for systemic analysis. Furthermore, we have looked for cases of languages with roughly the top relevance but for different contrastive attributions and still ended up with mostly significantly biased pairs. This suggests that the existence of such unbiased pairings is relatively rare, which limits the scope of applications of the heuristic.

7 Conclusion

In this report, we assessed the potential of a contrast-based heuristic for selecting transfer languages in multi-lingual neural machine translation targeting low-resource languages. We aimed to formulate this heuristic by leveraging Shapely values and language rankings from the LangRank framework, and then evaluate it using jointly trained multilingual models.

Our analysis revealed challenges with distinguishing the contributions from linguistic features in the learning process from a bias associated with the ranking scores. To mitigate this bias in our assessment, we selected language pairs for Esperanto that had closely matched ranking scores from both high-contrast and low-contrast pairs. We inferred that given the magnitude and direction of the bias, the bias could not be the sole explanation for why the high-contrast pair achieved a BLEU score greater than that of the low-contrast pair. However, in order to determine the potential of leveraging distinct linguistic traits in training multilingual models, more extensive experiments with diverse language groups and larger data sets are required.

[6] For the specific values related to these biases see Table 3 in Appendix A.1.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. 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*.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 148–156. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3146–3154.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 641–652. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3125–3135. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori S. Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 8–14. Association for Computational Linguistics.

Luke Merrick and Ankur Taly. 2020. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference on Machine Learning and Knowledge Extraction*.

S. J. Mielke and Jason Eisner. 2019. Spell once, summon anywhere: A two-level open-vocabulary language model. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6843–6850. AAAI Press.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 875–880. Association for Computational Linguistics.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: The extensible neural machine translation toolkit. In *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*, Boston.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with data augmentation. *arXiv preprint arXiv:1810.06826*.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics.

Lloyd S Shapley. 1953. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5823–5828. Association for Computational Linguistics.

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. The Association for Computational Linguistics.

A Appendix

A.1 Extra material on the results

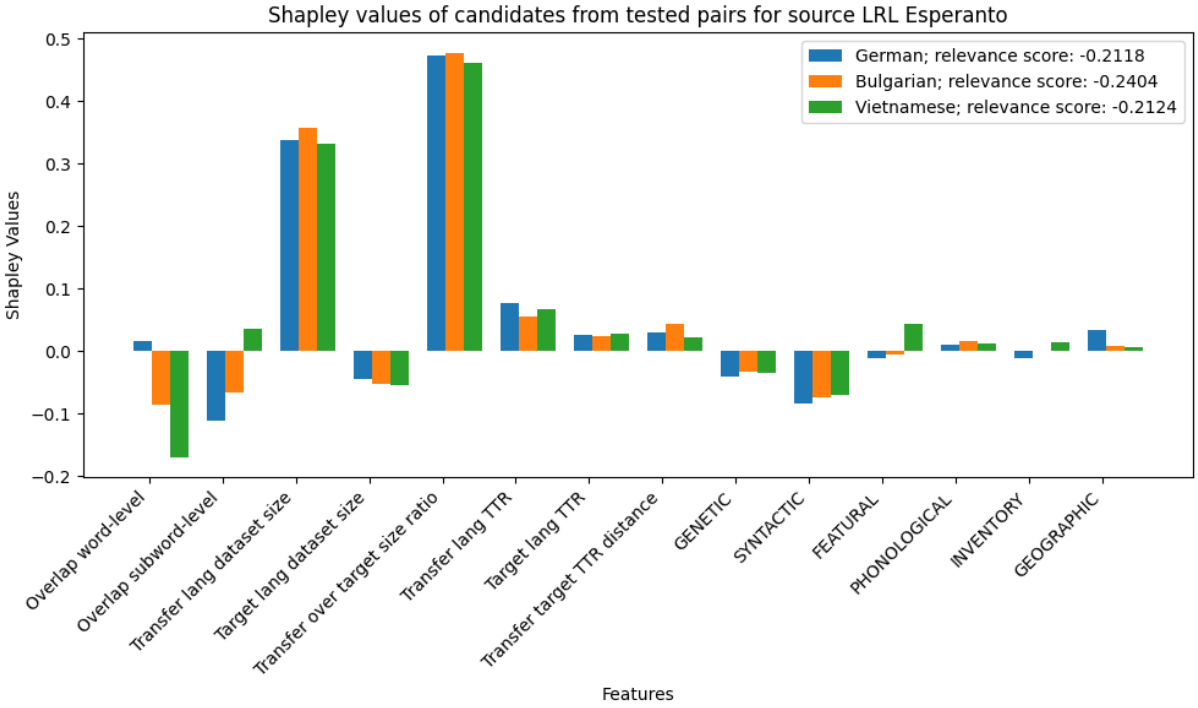

Figure 2: Shapley values for German, Bulgarian and Vietnamese with source language Esperanto and target language English

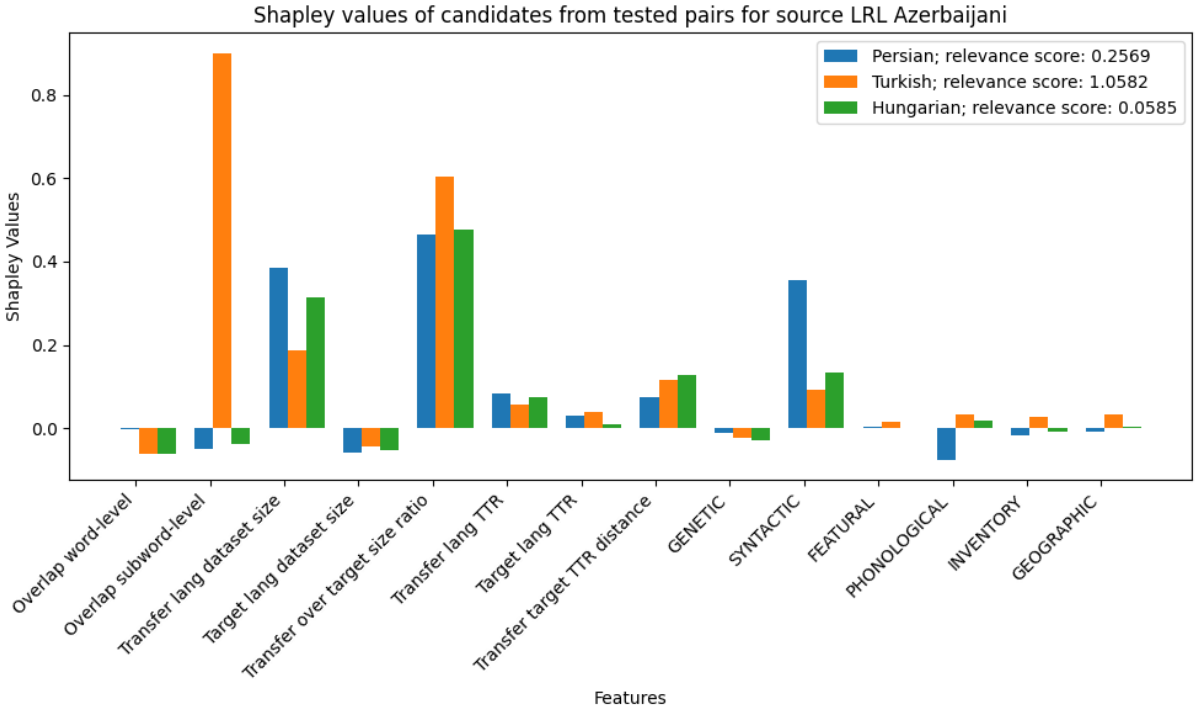

Figure 3: Shapley values for Persian, Turkish and Hungarian with source language Azerbaijani and target language English

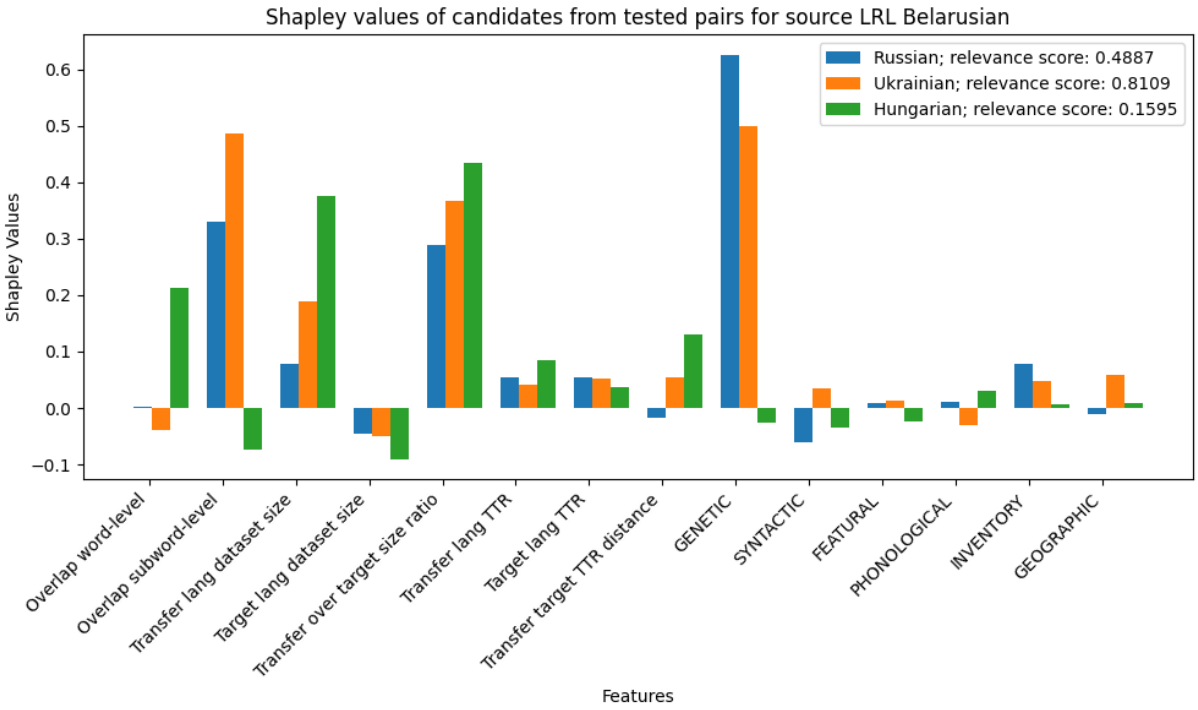

Figure 4: Shapley values for Russian, Ukrainian and Hungarian with source language Belarusian and target language English

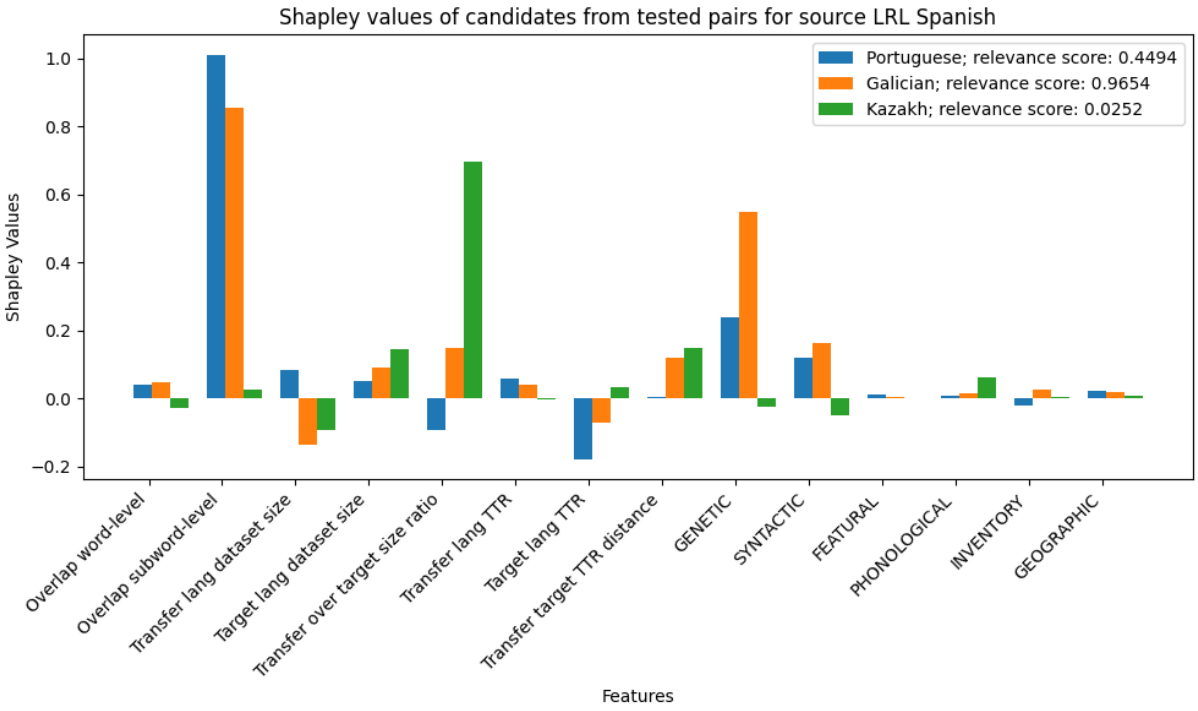

Figure 5: Shapley values for Portuguese, Galician and Kazakh with source language Spanish and target language English

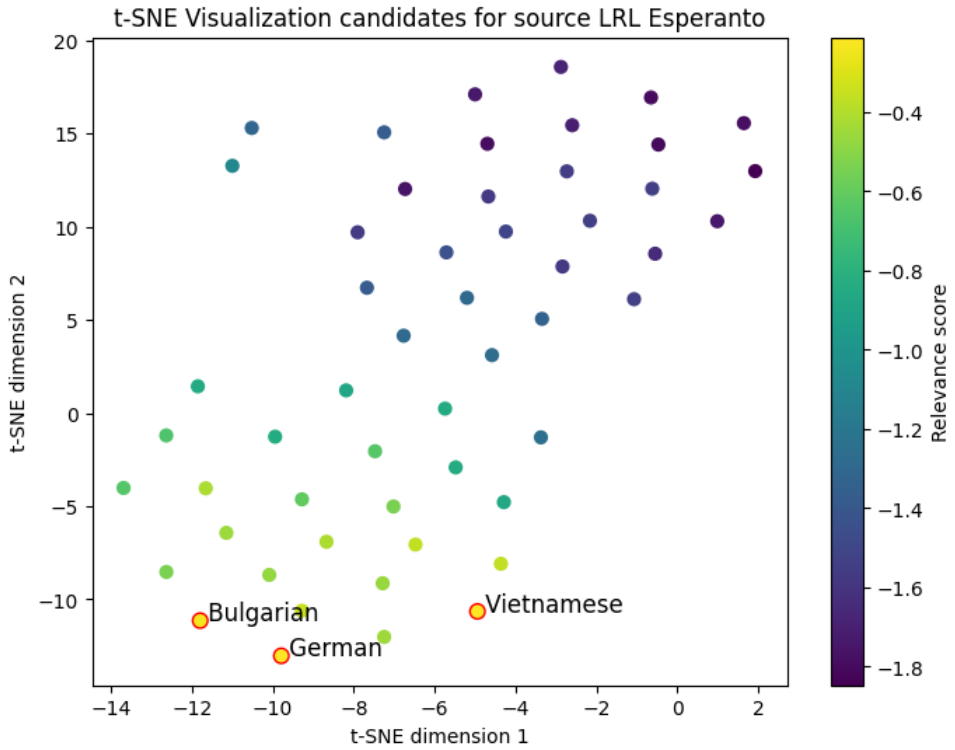

Figure 6: T-SNE based on the Shapley values of all candidates with source language Esperanto and target language English. Relevance score is according to LangRank.

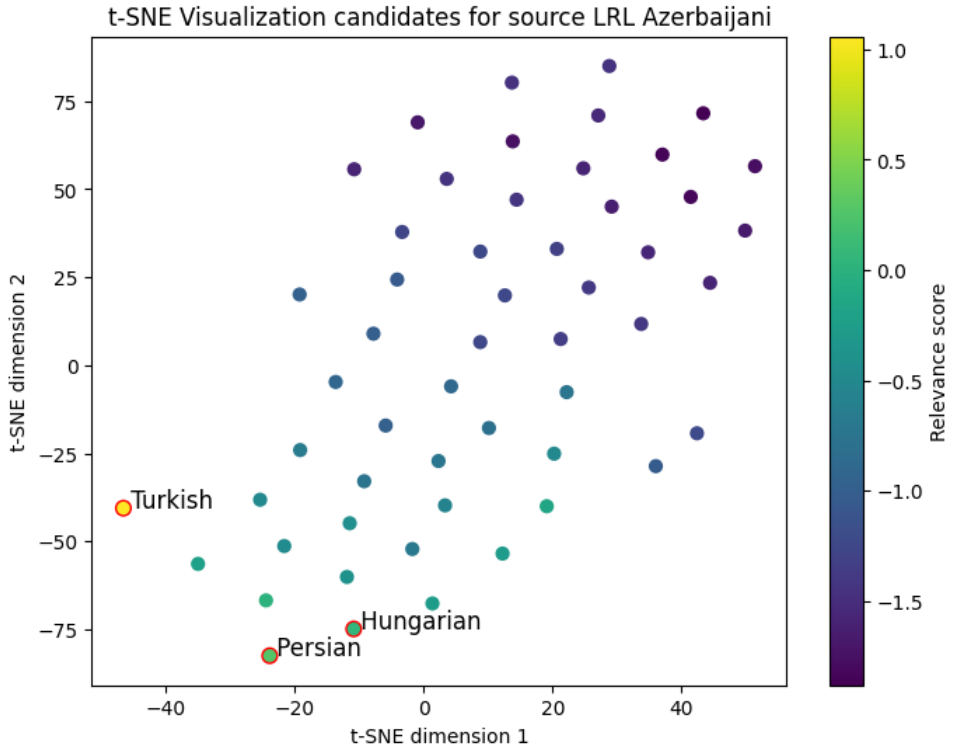

Figure 7: T-SNE based on the Shapley values of all candidates with source language Azerbaijani and target language English. Relevance score is according to LangRank.

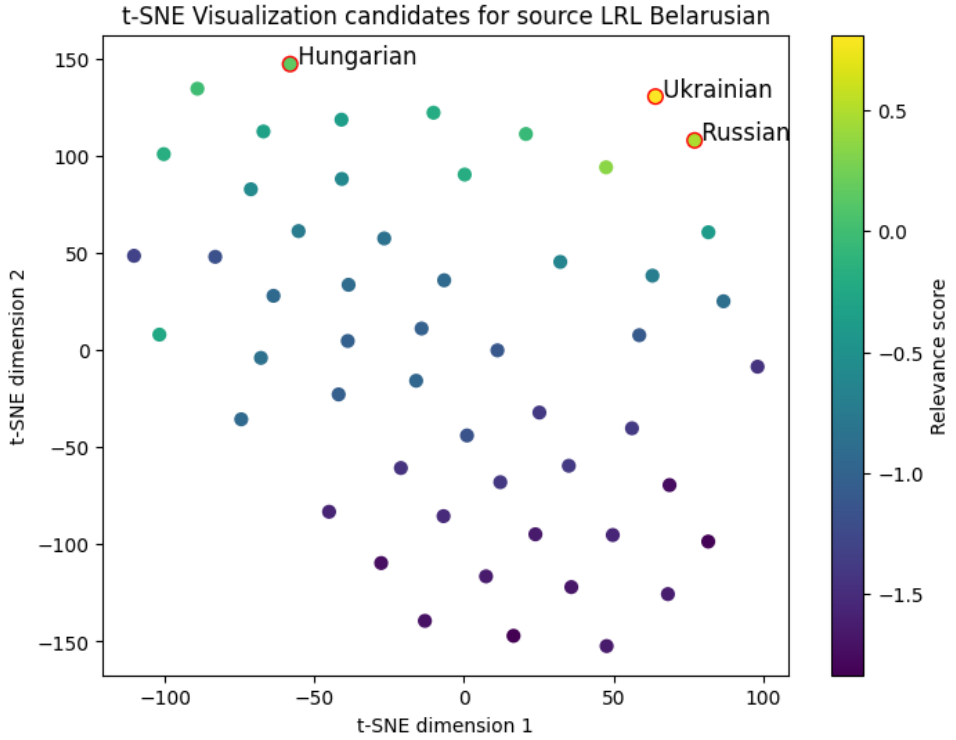

Figure 8: T-SNE based on the Shapley values of all candidates with source language Belarusian and target language English. Relevance score is according to LangRank.

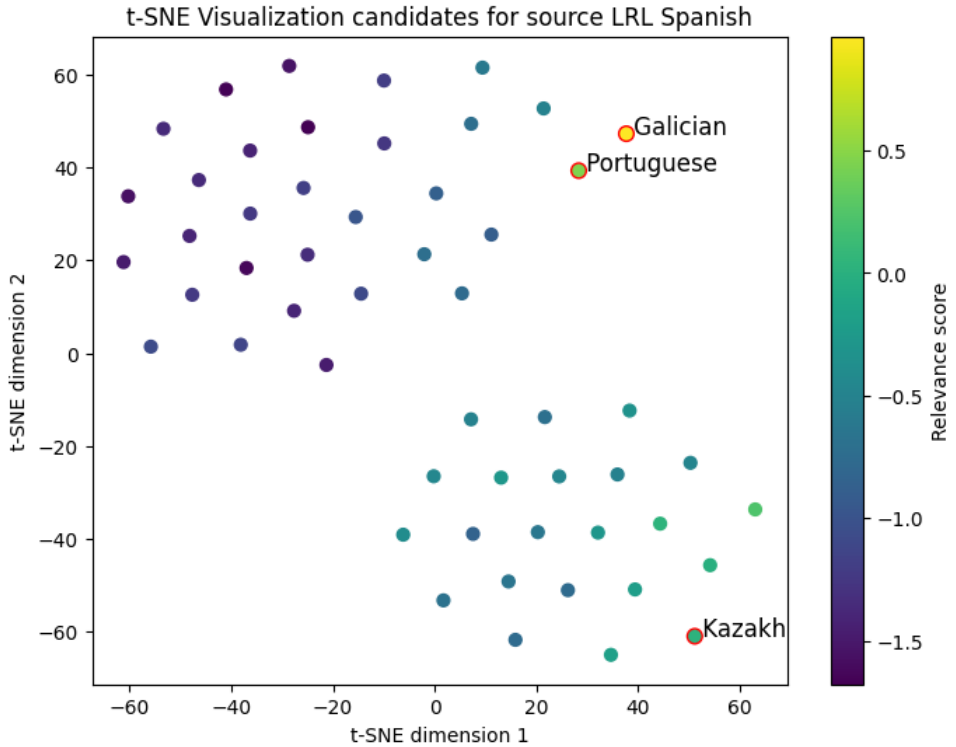

Figure 9: T-SNE based on the Shapley values of all candidates with source language Spanish and target language English. Relevance score is according to LangRank.

Source	Transfer	Relevance score
EO	DE	-0.2118
EO	BG	-0.2404
EO	VI	-0.2124
BE	RU	0.4887
BE	UK	0.8109
BE	HU	0.1595
AZ	FA	0.2569
AZ	TR	1.0582
AZ	HU	0.0585
ES	PT	0.4494
ES	GL	0.9654
ES	KK	0.0252

Table 3: Relevance score (rounded to the fourth decimal) according to LangRank for the languages from the low and high contrast pairs. The magnitudes of these scores are hard to interpret on their own and naive comparison between source languages can be misleading because their purpose is just to help ordering per source language. The variances of the scores for all candidates for a given source are however similar for all source languages, so the absolute difference between scores within a source does seem like an appropriate proxy for difference in relevance. In other words, these transfer candidates for EO seem to be significantly more similar in relevance to each other than the candidates shown here for the other sources.

Experiment	BLEU
$M_{ES,Random-sample-1}$	1.49
$M_{ES,Random-sample-2}$	1.43
$M_{ES,Random-sample-3}$	1.20

Table 4: List of results from individual runs in evaluating data quality across distinct, random samples.

A.2 Extra material on the data and models

Hyperparameter	Value
Batch size	32
Optimizer	Adam
Initial learning rate	0.0002
Learning rate decay factor	0.5
Epochs	100

Table 5: Hyperparameters used in training of translation models

Architecture	Value
Encoder	XNMT-based BiLSTM Sequence Transducer
Decoder	XNMT-based Autoregressive Decoder
Word embedding dimension	300
Hidden layer size	128
Desegmentation method	XNMT-based join-bpe

Table 6: Architecture used in training of translation models

Model	train	dev	test
Baseline Models			
AZ	5946	671	903
BE	4509	248	664
EO	4500	758	450
ES	5000	600	1000
Multilingual models			
AZ	4500	450	903
BE	4500	248	664
EO	4500	450	5571
ES	4500	450	5571
BG	10,000	1000	NA
DE	10,000	1000	NA
FA	10,000	1000	NA
GL	10,000	1000	NA
HU	10,000	1000	NA
KA	10,000	1000	NA
PT	10,000	1000	NA
RU	10,000	1000	NA
TU	10,000	1000	NA
UK	10,000	1000	NA
VI	10,000	1000	NA

Table 7: Number of sentences used in training and evaluation of each model

Language	Language code	Language family	Writing system
German	DE	Germanic	Latin
Farsi (Persian)	FA	Indo-Iranian	Perso-Arabic
Azerbaijani	AZ	Turkic	Cyrillic
Turkish	TR	Turkic	Latin
Kazakh	KZ	Turkic	Cyrillic
Hungarian	HU	Uralic	Latin
Spanish	ES	Romance	Latin
Portuguese	PT	Romance	Latin
Galician	GL	Romance	Latin
Belarussian	BE	East Slavic	Cyrillic
Bulgarian	BG	South Slavic	Cyrillic
Russian	RU	East Slavic	Cyrillic
Ukrainian	UA	East Slavic	Cyrillic
Esperanto	EO	Constructed Language	Latin (primarily)
Vietnamese	VI	Austroasiatic	Latin

Table 8: Language code, family, and writing system of all languages considered.

Type	Feature	Description
Data dependent	Overlap word-level	Similarity of target- and transfer-language corpora in vocabularies by word overlap.
	Overlap subword-level	Similarity of target- and transfer-language corpora in vocabularies by subword overlap.
	Transfer lang dataset size	Number of sentences in the corpus of a transfer language.
	Target lang dataset size	Number of sentences in the corpus of a target language.
	Transfer over target size ratio	Ratio of the number of sentences from corpora between transfer and target languages.
	Transfer lang TTR	Ratio of the number of unique words and tokens in the corpus of a transfer language.
	Target lang TTR	Ratio of the number of unique words and tokens in the corpus of a target language.
	Transfer target TTR distance	The distance (morphological similarity) of the TTRs of the transfer- and target language corpora.
Data independent	Genetic distance	Genealogical distance of languages derived from hypothesized tree of language descent.
	Syntactic distance	Cosine distance between feature vectors derived from syntactic structures of languages.
	Phonological distance	Cosine distance between the phonological feature vectors.
	Featural distance	Cosine distance between feature vectors combining all 5 single data-dependent features.
	Inventory distance	Cosine distance between the phonological feature vectors derived.
	Geographic distance	Orthodromic distance between the languages on earth's surface, divided by the antipodal distance.

Table 9: LangRank uses 8 data-dependent and 6 data-independent ranking features for learning-to-ranking. The data-independent features are based on Littell et al. (2017).