Cross-lingual Transfer Learning with Persian

Sepideh Mollanorozy
University of Malta
University of Groningen
sepid.mnorozy@gmail.com

Marc Tanti
University of Malta
marc.tanti@um.edu.mt

Malvina Nissim
University of Groningen
m.nissim@rug.nl

Abstract

The success of cross-lingual transfer learning for POS tagging has been shown to be strongly dependent, among other factors, on the (typological and/or genetic) similarity of the low-resource language used for testing and the language(s) used in pre-training or to fine-tune the model. We further unpack this finding in two directions by zooming in on a single language, namely Persian. First, still focusing on POS tagging we run an in-depth analysis of the behaviour of Persian with respect to closely related languages and languages that appear to benefit from cross-lingual transfer with Persian. To do so, we also use the World Atlas of Language Structures to determine which properties are shared between Persian and other languages included in the experiments. Based on our results, Persian seems to be a reasonable potential language for Kurmanji and Tagalog low-resource languages for other tasks as well. Second, we test whether previous findings also hold on a task other than POS tagging to pull apart the benefit of language similarity and the specific task for which such benefit has been shown to hold. We gather sentiment analysis datasets for 31 target languages and through a series of cross-lingual experiments analyse which languages most benefit from Persian as the source. The set of languages that benefit from Persian had very little overlap across the two tasks, suggesting a strong task-dependent component in the usefulness of language similarity in cross-lingual transfer.

1 Introduction and Background

Cross-lingual transfer learning consists in using a (usually high resource) language for fine-tuning a pre-trained model for a given task, but then using such model to obtain predictions for a different (usually low-resourced) language. This is advantageous if the lesser-resourced language lacks enough resources for training. While in early work on transfer learning English has often been used as source language, due to its high availability, more recent research has shown that this might not be the optimal choice. For example, de Vries et al. (2021) show that for POS tagging language similarity has a great impact on the success of transfer learning, and even with a small amount of data, one can achieve high accuracy.de Vries et al. (2022) expands this study by doing cross-lingual transfer learning between over 100 languages, in search of good combinations of source and target languages. They find that there is no single language that is a good source language for cross-lingual transfer learning with all other languages. Besides, the target language being included in the model pre-training is the most effective factor on performance of the model which does not play a role in low-resource settings. The next best predictor found for finding a good performing source-target language pair is the LDND distance (Wichmann et al., 2010) between them, considered as the language similarity measure. This measure is based on the Levenshtein distance between a set of selected words in two languages.

As a contribution to a better understanding of the properties of source and target languages towards successful transfer learning, and towards better processing for low-resource languages, we investigate cross-lingual transfer learning with a focus on Persian. We analyze the results of de Vries et al. (2022) experiments that include Persian as either the source or the target language to find the languages that are a good match with Persian for POS tagging. To explain the potential reasons for the results, we use the linguistic features from World Atlas of Language Structures (WALS).

We also examine the performance of ParsBERT, the pre-trained monolingual Persian model, in comparison to XLM-RoBERTa, a pre-trained multilingual model, for the POS tagging task.

Finally, we investigate whether the language pairs with Persian in the POS tagging are generalizable to other NLP tasks or not. We perform cross-

lingual transfer learning for sentiment analysis as there is Persian dataset available for this task and this task is a high-level NLP task compared to POS tagging as a low-level NLP task. This combination of tasks has been of interest for cross-lingual transfer learning in other studies as well (Dat, 2021).

We gather sentiment analysis datasets from various resources and carry out experiments using the pre-trained multilingual XLM-RoBERTa language model. We fine-tune this model using Persian data and then test it with other languages. In the end, we compare the best target languages with Persian as source in sentiment analysis and POS tagging.

Persian language is the official language of Iran, Afghanistan and Tajikestan. The variety of Persian in these countries is Iranian Persian (main and official variety of Persian), Dari, and Tajik. The writing system of Iranian Persian and Dari are the same, using Persian alphabet, whereas, the Tajik variety has a different writing system. Figure 1 shows the geographical location of people whose mother tongue is Persian.

Persian is an Indo-European language, with a subject-object-verb word order, and it has words borrowed from French and English. Additionally, its grammar is similar to many Indo-European languages. But also it has many words in common with Arabic, as Iran has Iraq as one of its neighbour countries and the official religious book for both countries is in Arabic.

Figure 1: Regions that the majority of people's mother tongue is Persian (Commons, 2021b)

Considering Iran's population of 85 million people, the number of Persian speakers is considerably large. According to Figure 2, Persian speakers are widely spread around the world. These observations show the importance of research with Persian language as it is used by a lot of people around the world, and it can result in applications benefiting a large group of people.

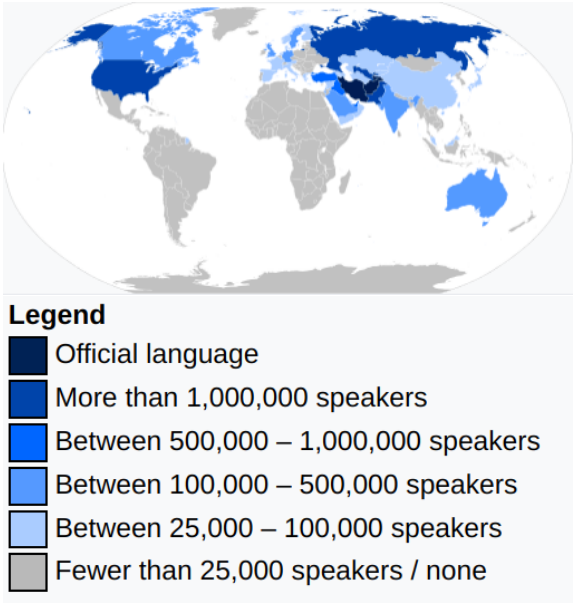

Figure 2: Persian speakers spread around the world (Commons, 2021a)

2 Experimental Setup

2.1 Datasets

For the POS tagging analysis, we a subset of the Universal Dependencies (UD) dataset as de Vries et al. (2022)[1] that has a tag set of 17 tags. There are 105 languages in this dataset in total, all having at least 10 samples as test data that we consider as target language in our analysis. Among them, 65 languages also have at least 25 samples as train data which we consider as source languages. We also obtained the accuracy scores of these 6825 different source-target language pairs using the XLM-RoBERTa model from de Vries et al. (2022).

We use the LDND distance measure between 90 different languages from the ASJP database[2] (Wichmann et al., 2022) as a measure of language similarity. In addition, we use the WALS dataset (wal, 2013)[3] including 192 different phonological, grammatical, and lexical properties of 2 676 unique languages. The number of common linguistics features is the second language similarity measure that we use in our analysis.

A multilingual sentiment analysis dataset containing all the languages that exist in UD dataset for

[1] https://huggingface.co/datasets/wietsedv/udpos28

[2] https://asjp.clld.org/

[3] https://www.kaggle.com/datasets/rtatman/world-atlas-of-language-structures

POS tagging does not exist. The largest one that we found contains negative and positive tagged data including 23 languages[4] as follows: Algerian, Arabic, Basque, Bulgarian, Cantonese, Chinese, Croatian, English, Finnish, German, Greek, Hebrew, Indonesian, Japanese, Korean, Maltese, Norwegian, Russian, Slovak, Spanish, Thai, Turkish, and Vietnamese. In addition, we gather data for 8 languages namely Persian, Urdu, Hindi, Welsh, Polish, Romanian, Bambara, and Uyghur from multiple resources, resulting in 31 languages in total. Details about the datasets is provided in appendix A. We converted all of them to the same structure and only kept the positive and negative data entries [5].

2.2 Methods

For POS tagging analysis, we analyze the results of experiments that de Vries et al. (2022) did with Persian and other languages. In each experiment, de Vries et al. (2022) fine-tuned the XLM-RoBERTa model with a source language and then tested it with a target language. We focus on the result of experiments that have Persian as the source or target language and attempt to find languages that result in a high score with Persian in each scenario. We find the target languages that have Persian as one of their top 10 source languages based on accuracy score. Then, we consider Persian as the target language, and find the source languages that have Persian as one of their top 10 target languages.

We also explore the linguistic features of the languages that are a good pair with Persian using the WALS data. We get all the features of the languages and measure their Hamming distance to the Persian features.

In our last experiment for POS tagging, we fine-tune the ParsBERT language model for 3 epochs with Persian data. At this stage, we achieved a high performance with an accuracy score of 95.99% on the validation set. As this score is higher than the XLM-RoBERTa Persian monolingual score, we kept this model and did not continue the training procedure. Then, we test this model with Persian and other languages that are a good match with it for POS tagging.

For the sentiment analysis experiments, we use the XLM-RoBERTa[6] pre-trained model, the same model that is used by de Vries et al. (2022) for the POS tagging experiments. We fine-tune the model with Persian data for 10 epochs with the best score occurring at the 5th epoch, yielding an accuracy of 87.21%. Model fine-tuning details are provided in appendix A. We take the model checkpoint at epoch 5 and test it with Persian and other target languages to predict the sentiment of the input text as positive or negative.

3 Results and discussion

3.1 POS-tagging

Using the XLM-RoBERTa model, the monolingual Persian experiment [7] has the highest accuracy of 91.43%. Considering Persian as the target language, Persian itself is the best source language, as the accuracy score drops under 81% in other experiments. Only two languages: Gothic and Arabic have Persian as one of their top 10 target languages but with low accuracies of 53.12% and 76.08%. Therefore, for POS tagging, Persian as target does not benefit from other languages as the source language. Details of source languages and scores is provided in appendix3

Nevertheless, considering Persian as the source language yields interesting results. The list of languages that have Persian as one of their top 10 source languages is as follows: Akkadian (low resource), Assyrian (low resource), Bambara (low resource), Bhojpuri (low resource), Hindi, Kurmanji (low resource), Persian, Tagalog (low resource), Urdu, Uyghur, and Welsh. Among these 11 languages, 6 languages are low resource languages which draw our interest. For Tagalog (78.96%) and Kurmanji (78.90%) we observe a score of roughly 79%, which is higher than the other low-resource languages. In addition, among the languages resulting in a high accuracy for Kurmanji listed in appendix 4, Persian is the most similar language to it regarding the LDND distance measure. Also from another perspective to assess languages similarity, we use the linguistic features from WALS dataset. We observe that for Kurmanji there are only 12 features in WALS and 10 of them are shared with Persian. Therefore, we propose that Persian is a good source languages for Kurmanji. Besides, Persian and Kurmanji are spoken in close geographical locations (Iran, Turkey, Iraq, Syria).

[4] https://github.com/jerbarnes/typology_of_crosslingual

[5] The whole dataset is accessible from https://huggingface.co/sepidmnorozy

[6] xlm-roberta-base

[7] The source language and the target language are the same

For languages that have Persian as one of their top 10 source languages, we provide the number of features available for each language in WALS and the number of common ones with Persian in appendix 6. According to this table, first Hindi and second Tagalog have the most common features with Persian. Although Tagalog is a low-resource language, it has 145 features listed in WALS. Besides, Persian has 147 features and has 54 features in common with Tagalog. In addition, among the list of top 10 source languages for Tagalog shown in appendix 5, Persian has the lowest LDND distance. Therefore we propose Persian as a potential source language for Tagalog in other cross-lingual tasks.

Using the Pars-BERT model, fine-tuning it with Persian as source, and test it with Persian and others as target, Persian as target has a score of 95.99% which is higher than the monolingual Persian experiment with XLM-RoBERTa. However, only with Persian Pars-BERT outperforms XLM-RoBERTa. Therefore, the monolingual Persian model is not enough for transfer learning and other languages' existence in the pre-training of the model has a significant effect on both high-resource and low-resource languages.

3.2 Sentiment analysis

The evaluation metrics for top 10 languages based on the accuracy score are shown in figure 3. Surprisingly the accuracy of the monolingual Persian experiment is only 87.69%, and Persian is not on the top of the list. However, Slovak has the highest accuracy of 93.38% occupying the first rank.

In this binary sentiment analysis task, most of the languages shown in Figure 3 have higher precision than recall. High precision values show that the model is not labeling negative samples as positive. The opposite case happens for Polish and sharply for German. For these two languages, the model has a higher recall, better at predicting the positive case and performs poorly on negative samples.

Considering Persian as the source language, the target languages that have a high score for POS tagging (listed in appendix7) and for sentiment analysis (listed in figure 3) only have two languages in common: "Polish" and "Bulgarian" Therefore, based on our results, cross-lingual transfer learning with Persian is task-dependent, and not the same group of languages appeared for both tasks.

4 Conclusion

All in all, we analyse the result of previous experiments for POS tagging and investigate whether having Persian as source or target language in cross-lingual transfer learning would be beneficial for Persian and other languages. We observe that Persian is the best source for itself as target and achieves a score of 91.43% for POS tagging. Besides, it can serve as a good source for 6 low-resource languages. We use LDND distance measure and linguistic features from WALS to reason that Persian can be a potential good source for Kurmanji and Tagalog for other tasks than POS tagging as well. Lastly for POS tagging, we observe that ParsBERT outperforms XLM-RoBETa only for monolingual Persian experiment and achieves a score of 96%. Then, we gather data and perform sentiment analysis to investigate whether the same target languages found for POS tagging would also benefit from Persian as the source language for sentiment analysis. We observe different target languages from the POS tagging results and only two languages: Polish and Bulgarian appear for both tasks. In addition, monolingual Persian experiment does not achieve the highest accuracy and Slovak is the best performing target. Therefore, we conclude that cross-lingual transfer learning with Persian is task dependent.

5 Limitations

The main challenge of this work was to find sentiment analysis dataset for various languages, especially the low-resource ones.

References

2013. Wals online.

2021. *Evaluating morphological typology in zero-shot cross-lingual transfer*. Association for Computational Linguistics, Online.

Wikimedia Commons. 2021a. File:map of persian speakers.svg — wikimedia commons, the free media repository. [Online; accessed 13-February-2022].

Wikimedia Commons. 2021b. File:persian language location map.svg — wikimedia commons, the free media repository. [Online; accessed 13-February-2022].

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

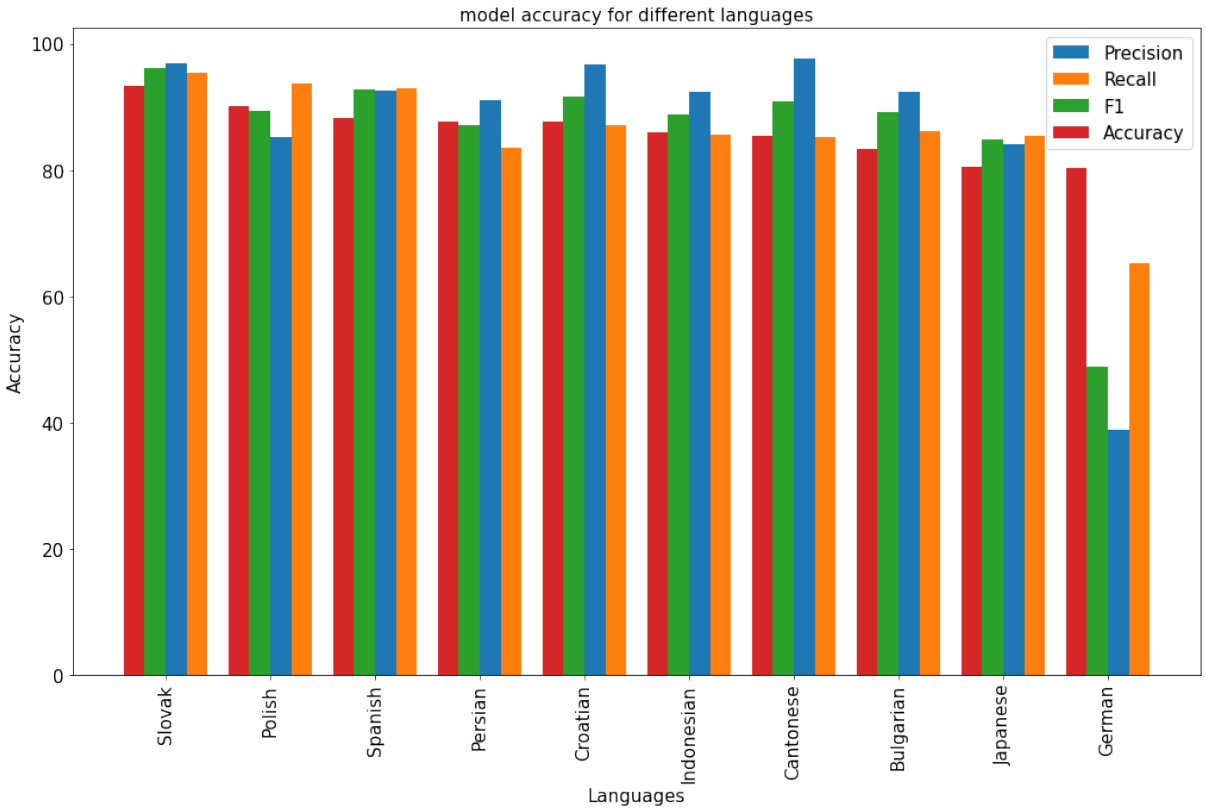

Figure 3: Evaluation metrics for sentiment analysis testing

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Mountaga Diallo, Chayma Fourati, and Hatem Haddad. 2021. Bambara language dataset for sentiment analysis.

Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasic, and Dawn Knight. 2021. English–welsh cross-lingual embeddings. *Applied Sciences*, 11:6541.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding.

Muhammad Yaseen Khan and Muhammad Suffian Nizami. 2020. Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled datasetfor urdu sentiment analysis. In *2020 IEEE 2nd International Conference On Information Science Communication Technology (ICISCT)*. IEEE.

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. pages 980–991.

Siyu Li, Kui Zhao, Jin Yang, Xinyun Jiang, Zhengji Li, and Zicheng Ma. 2022. Senti-exlm: Uyghur enhanced sentiment analysis model based on xlm. *Electronics Letters*, 58.

Søren Wichmann, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389(17):3632–3639.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2022. The ASJP Database.

A Sentiment Analysis Details

Table 1 shows the details of different datasets we gathered for sentiment analysis. Table 2 shows the evaluation metrics while fine-tuning the XLM-RoBERTa model for sentiment analysis.

B POS Tagging Details

Table 3, table 4, and table 5 show the top 10 source languages for target languages Persian, Kurmanji, and Tagalog respectively. Table 6 shows the number of features from WALS dataset for languages that have Persian as one of their top 10 source languages. Table 7 shows the languages achieving the highest accuracies when Persian is the source.

Lang	#pos	#neg	content	source	#train	#val	#test
Persian	35k	35k	food reviews	(Farahani et al., 2020)	56.7k	6.3k	7k
Urdu	500	480	political tweets	Khan and Nizami (2020)	685	-	294
Hindi			movie reviews	Kaggle	513	115	-
Welsh	25k	25k	movie reviews	Espinosa-Anke et al. (2021)	25k	-	25k
Polish	1762	2455	school, products, medicine, hotels reviews	Kocoń et al. (2019)	3737	-	480
Romanian	17271	11675	products and movie reviews	Huggingface	17941	-	11005
Bambara	1663	579	sports, politics, music, etc	Diallo et al. (2021)	1569	-	673
Uyghur	2450	353	Common-crawl	Li et al. (2022)	1962	-	841

Table 1: Details of sentiment analysis data

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.3645	0.4315	0.8603	0.8466	0.9386	0.7711
2	0.374	0.4015	0.8713	0.8648	0.9105	0.8235
3	0.3363	0.4772	0.8705	0.8615	0.9256	0.8057
4	0.3131	0.4579	0.8702	0.8650	0.9007	0.8321
5	0.3097	0.4160	0.8721	0.8663	0.9069	0.8292
6	0.2921	0.4638	0.8673	0.8630	0.8917	0.8362
7	0.272	0.5183	0.8654	0.8602	0.8947	0.8283
8	0.2481	0.5846	0.8649	0.8624	0.8787	0.8467
9	0.192	0.6481	0.8610	0.8596	0.8680	0.8514
10	0.1945	0.7030	0.8603	0.8585	0.8699	0.8473

Table 2: XLM-RoBERTa fine-tuning results for sentiment analysis

Idx	Source	Target	Score	dist
1	Persian	Persian	91.43	nan
2	Urdu	Persian	80.63	78.87
3	Czech	Persian	80.09	94.62
4	Irish	Persian	79.73	98.25
5	Croatian	Persian	79.39	93.12
6	Armenian	Persian	79.23	98.0
7	Romanian	Persian	79.05	92.91
8	Galician	Persian	78.88	92.96
9	Welsh	Persian	78.7	97.71
10	Russian	Persian	78.7	93.02

Table 3: Top 10 best source languages for Persian as target

Idx	Source	Target	Score	Dist
1	Romanian	Kurmanji	79.52	89.76
2	Galician	Kurmanji	79.38	93.39
3	Czech	Kurmanji	79.28	95.59
4	Persian	Kurmanji	78.9	79.4
5	French	Kurmanji	78.88	90.9
6	Icelandic	Kurmanji	78.56	95.49
7	Croatian	Kurmanji	78.51	93.89
8	Bulgarian	Kurmanji	78.47	93.55
9	Dutch	Kurmanji	78.32	90.39
10	Italian	Kurmanji	78.24	89.86

Table 4: Top 10 best source languages for Kurmanji as target

Idx	Source	Target	Score	Dist
1	Bulgarian	Tagalog	81.56	102.73
2	Russian	Tagalog	80.91	101.1
3	Polish	Tagalog	80.17	98.98
4	Icelandic	Tagalog	79.98	100.87
5	Hebrew	Tagalog	79.24	101.8
6	Persian	Tagalog	78.96	96.05
7	Urdu	Tagalog	78.49	99.32
8	Serbian	Tagalog	77.47	97.51
9	Faroese	Tagalog	76.07	102.85
10	Spanish	Tagalog	74.39	96.76

Table 5: Top 10 best source languages for Tagalog as target

Idx	Lang	#features	#Common
0	Persian	147	147
1	Hindi	144	71
2	Tagalog	145	54
3	Bambara	90	33
4	Welsh	69	28
5	Urdu	42	20
6	Bhojpuri	36	17
7	Uyghur	35	11
8	Kurmanji	12	10
9	Arabic	30	10
10	Assyrian	3	2

Table 6: WALS features for languages related to Persian

idx	Lang	Score	Mono score	Dist
1	Hebrew	89.58	93.75	99.16
2	Marathi	84.05	88.96	91.65
3	Estonian	83.52	96.80	100.19
4	Bulgarian	83.452	99.30	97.11
5	Polish	82.692	98.22	91.71
6	Serbian	82.472	99.06	93.93
7	Icelandic	82.32	95.64	98.67
8	Telugu	82.11	94.87	98.47
9	Tamil	82.00	85.64	97.17
10	Arabic	81.70	75.93	97.46

Table 7: Top 10 target languages for Persian as Source language based on POS tagging score