Transliteration for Cross-Lingual Morphological Inflection

Abstract

Cross-lingual transfer between typologically related languages has been proven successful for the task of morphological inflection. However, if the languages do not share the same script, current methods yield more modest improvements. We explore the use of transliteration between related languages as a data preprocessing method in order to alleviate this issue. We experimented with several diverse language pairs, finding that in most cases transliterating the transfer language data into the target one leads to accuracy improvements, even up to 9 percentage points.

1 Introduction

The majority of the world's languages are synthetic, meaning they have rich morphology. As a result, modeling morphological inflection computationally can have a significant impact on downstream quality, not only in analysis tasks such as named entity recognition and morphological analysis (Zhu et al., 2019), but also for language generation systems for morphologically-rich languages.

In recent years, morphological inflection has been extensively studied in monolingual high resource settings, especially through the recent SIGMORPHON challenges (Cotterell et al., 2016, 2017, 2018). The latest SIGMOPRHON 2019 challenge (McCarthy et al., 2019) focused on low-resource settings and encouraged cross-lingual training, an approach that has been successfully applied in other low-resource tasks such as Machine Translation (MT) or parsing. Cross-lingual learning is a particularly promising direction, due to its potential to utilize similarities across languages (often languages from the same linguistic family, which we will refer to as "related") in order to overcome the lack of training data. In fact, leveraging data from *several* related languages was crucial for the current state-of-the-art system over the SIGMORPHON 2019 dataset (Anastasopoulos and Neubig, 2019).

Transfer lang.		Test lang.	Acc.
(L$_1$)	*(script)*	(L$_2$)	
Urdu	*(Arabic)*	Bengali *(Bengali)*	42
Sanskrit	*(Devanagari)*		44
Hindi	*(Devanagari)*		49
Greek	*(Greek)*		42
Arabic	*(Arabic)*	Maltese *(Roman)*	18
Hebrew	*(Hebrew)*		22
Italian	*(Roman)*		34

Table 1: The languages' script can affect the effectiveness of cross-lingual transfer (using L$_1$ data to train a L$_2$ inflection system). Bengali results display low variance, as all transfer languages differ in script. Maltese is typologically closer to Arabic and Hebrew than Italian, but accuracy is higher when transferring from a same-script language.

However, as Anastasopoulos and Neubig (2019) point out, cross-lingual learning even between closely related languages can be impeded if the languages do not use the same script. We present a few examples taken from Anastasopoulos and Neubig (2019) in Table 1. The first example presents cross-lingual transfer for Bengali, with the transfer languages varying from very related (Hindi, Sanskrit, Urdu) to only distantly related (Greek). Nevertheless, there is notably little variance in the performance of the systems. We believe that the culprit is the difference in writing systems between all the transfer and test languages, which does not allow the system to easily leverage cross-lingual information: the Bengali data uses the Bengali script, Urdu data uses the Nastaliq script (a derivative of the Arabic alphabet), the Hindi and Sanskrit data uses Devanagari, and the Greek data uses the Greek alphabet. In the second example, with transfer from Arabic, Hebrew, and Italian for morphological inflection in Maltese, we note that although Maltese is much closer typologically to Arabic and Hebrew (they are all Semitic languages), the test accuracy is higher when transferring from Italian, which despite only sharing a few typological elements with Maltese happens to also share the same script.

The aim of this work is to investigate this potential issue further. We attempt to remedy this problem by bringing the representations of the transfer and the test languages in the same, shared space *before* training the morphological inflection system. We achieve this through transliteration of the transfer language. The aim of this work is to investigate this potential issue further. We first quantify the effect of script differences on the accuracy of morphological inflection systems through a series of controlled experiments (§2). Them, we attempt to remedy this problem by bringing the representations of the transfer and the test languages in the same, shared space *before* training the morphological inflection system. We achieve this through investigating transliteration of the transfer language as a preprocessing step and its effects on morphological inflection in low-resource settings (§3).

Our approach bears similarities to pseudo-corpus approaches that have been used in machine translation (MT), where low-resource language data are augmented with data generated from a related high-resource language. Among many, for instance, De Gispert and Marino (2006) built a Catalan-English MT by bridging through Spanish, while Xia et al. (2019) show that word-level substitutions can convert a high-resource (related) language corpus into a pseudo low-resource one leading to large improvements in MT quality. Such approaches typically operate at the word level, hence they do not need to handle script differences explicitly. NLP models that handle script differences do exist, but focus mostly on analysis tasks such as named entity recognition (Bharadwaj et al., 2016; Chaudhary et al., 2018; Rahimi et al., 2019) or entity linking (Rijhwani et al., 2019), whereas we focus in a *generation* task. Character-level transliteration was typically incorporated in phrase-based statistical MT systems (Durrani et al., 2014), but was only used to handle named entity translation. Notably, there exist NLP approaches such as the document classification approach of Zhang et al. (2018) showing that indeed shared character-level information can facilitate cross-lingual transfer, but limit their analysis to same-script languages only.

2 Quantifying the Issue

In Table 1 we offered a few examples from the literature to indicate that differences in script between the transfer and test language in a cross-lingual learning setting can be a potential issue. In this section, we provide additional evidence that this is indeed the case.

The intuition behind our analysis is that a model trained cross-lingually can only claim to indeed learn cross-lingually if it ends up sharing the representations of the different inputs, at least to some extent. This observation of a learned shared space has also been noted in massively multilingual models like the multilingual BERT (Pires et al., 2019), or for cross-lingual learning of word-level representations (Wang et al., 2020). For a character-level model, such as the ones typically used for neural morphological inflection, this implies a learned mapping between the characters of the two inputs. Our hypothesis is that such a learned character mapping, and in particular between related languages, should resemble a transliteration mapping, assuming that both languages use a phonographic writing system (such as the Latin or the Cyrillic alphabet and their variations), to use the notation of Faber (1992).[1]

To verify whether this intuition holds, we trained models on Armenian–Kabardian and Bashkir–Tatar (see details in Section §3). In the first setting, the transfer language (Armenian) uses the Armenian alphabet, while the test language (Kabardian) uses the Cyrillic one. In the second, we are transferring from Bashkir, which currently uses the Cyrillic alphabet, to Tatar, which is written with the Latin alphabet. We obtain the character representations from the final trained models, and we perform a simple search over the embedding space, returning for each of the transfer language characters the nearest neighbor from the test language alphabet. Our findings are that this type of mapping does not resemble a transliteration one, at all.

For example, one would expect that the Bashkir characters е, ә, or э would map to the Tatar e character, or at least to another vowel. Bashkir е indeed maps to Tatar e, but ә maps to Tatar i (which might

[1] In contrast, one should not expect this to hold if one of the scripts is logographic, like the Chinese one, or if the two languages are coded differently, e.g. one script is syllabic and segmentally coded, like the Japanese kana, but the other is segmentally linear using a complete alphabet like the Latin script. If both scripts use the same level of coding, then the intuition holds (i.e. between Hebrew and Arabic).

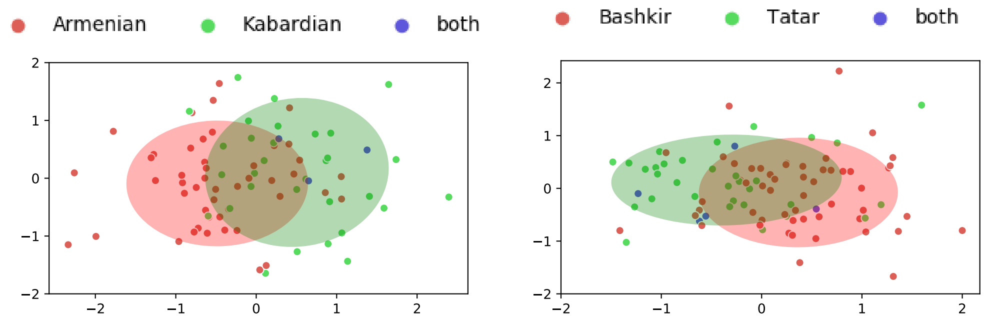

Figure 1: 2-D projection of the character embeddings learned after cross-lingual learning in two settings (Armenian–Kabardian and Bashkir–Tatar). The shaded area denotes the mean ± three standard deviations.

be somewhat fine since they are both vowels), while Bashkir э maps to Tatar r. After a manual annotation of the mappings in both language pairs, we find that the absolute accuracy is less than 5% in both settings (2 of 54 are correct in Bashkir–Tatar, and 1 of 47 in Armenian–Kabardian). We also present a visualization (obtained through PCA (Wold et al., 1987)) of the character embeddings in Figure 1 for these two settings, which shows that the two languages are still, to an extent, separable.

In an attempt to also take into account potential slight differences in pronunciation, which are common across related languages, we also count mappings that agree in coarse phonetic categories as correct. We obtain rough grapheme-to-phoneme mappings from Omniglot[2] (Ager, 2008) which allows us to classify each character as mapping to a vowel, or a consonant category (we devise categories across both manner and place). For instance, the Bashkir characters с,ҫ,һ,ҙ,ш map to sibilant fricatives, so we count any mapping to Tatar characters that also map to sibilant fricatives (ç, z, s, ş) as correct. Overall, however, even this more flexible evaluation only leads to an accuracy of less than 30% (16 out of 54 characters for Bashkir–Tatar, 12 of 47 in Armenian–Kabardian).

3 Methodology

The previous section showcases that different scripts can inhibit the model's ability to represent both languages in a shared space, which can be damaging for downstream performance in cross-lingual learning scenarios. In order to bring the transfer and test languages into a shared space we take a straightforward approach: we first transliterate the transfer language data into the script of the test language, and then use the data to train an inflection model. As our baseline or control experiment, we use the exact same data, model, and process, only removing the transliteration preprocessing step. The following sections provide details on transliteration, the inflection model, and the data that we use for training and evaluation.

Transliteration In the absence of some sort of a universal transliteration approach, we rely on various libraries for our experiments. For transliterating between the Indic scripts (Devanagari, Bengali, Kannada, and Telugu in our experiments) we rely on the IndicNLP library.[3] We also use the URoman[4] library (Hermjakob et al., 2018) to transliterate into the Roman alphabet for the Arabic, Hebrew, Armenian, and Cyrillic scripts.

The lack of resources and transliteration tools for some directions severely limited the extent of the experiments that we could conduct. Notably, even though romanization is fairly well-studied and are easily attainable through tools like URoman, the opposite direction is fairly understudied. Most of the related work has focused on either to-English transliteration specifically (Lin et al., 2016; Durrani et al., 2014) or on *named entity* transliteration (Kundu et al., 2018; Grundkiewicz and Heafield, 2018). Even then, the state-of-the-art results on the recent NEWS named entity transliteration task (Chen et al., 2018) ranged from 10% to 80% in terms of accuracy across several scripts. The high variance in expected quality depending on the transliteration direction showcases the need for further work towards tackling hard transliteration problems.

Inflection Model We use the morphological inflection model of Anastasopoulos and Neubig (2019) which achieved the highest rank in terms of average accuracy in the SIGMORPHON 2019

[2] https://omniglot.com/

[3] https://github.com/anoopkunchukuttan/indic_nlp_library

[4] https://github.com/isi-nlp/uroman

Transfer	Test	Baseline	with Transliteration		Baseline	with Transliteration
L$_1$	L$_2$	L$_1$+L$_2$	L$_1$ Conversion	Tr(L$_1$)+L$_2$	L$_1$+L$_2$+$\mathcal{H}$	Tr(L$_1$)+L$_2$+$\mathcal{H}$
Hindi		33	Devanagari	**42**	47	**56**
Sanskrit	Bengali	27	→	32	66	65
both		39	Bengali	41	58	63
Arabic		18	Arabic→Roman	**27**	29	27
Hebrew	Maltese	22	Hebrew→Roman	27	29	33
both		18		21	25	28
Kannada	Telugu	66	Kannada→Telugu	66	70	62
Bashkir	Crimean Tatar	73	Cyrillic→Roman	69	70	73
Bashkir	Tatar	**74**		59	73	74
Russian	Portuguese	34	Cyrillic→Roman	**43**	61.5	**63.5**
(*) Adyghe		90	no conversion	–	96	–
(*)Armenian	Kabardian	80	Armenian→Roman	78	86	86

Table 2: Transliteration of the transfer language (L$_1$) into the test language (L$_2$) improves accuracy in some cases (top), with and without hallucinated data ($\mathcal{H}$). In some language pairs (bottom) it can be harmful. We report exact match accuracy on the test set. We **highlight** statistically significant improvements ($p < 0.05$) over the baseline. "both" denotes that both L$_1$ languages are used for transfer. * marks an additional control experiment.

shared task, using the publicly available code.[5] The authors also use a data hallucination technique similar to the one of Silfverberg et al. (2017), which we also use in ablation experiments.[6]

Data and Evaluation We use the data from the SIGMORPHON 2019 Shared Task on Morphological Inflection(McCarthy et al., 2019). We stick to the transfer learning cases that were studied in the shared task, but limit ourselves to the language pairs where (1) the two languages use different writing scripts, and (2) we have access to a transliteration model from the transfer to the test language. As a result, we evaluate our approach on the following language pairs: {Hindi,Sanskrit}–Bengali, Kannada–Telugu, {Arabic,Hebrew}–Maltese, Bashkir–Tatar, Bashkir–Crimean Tatar, Armenian–Kabardian, and Russian–Portuguese. We compare our systems' performance with the baselines using exact match accuracy over the test set. We also perform statistical significance testing using bootstrap resampling (Koehn, 2004).[7]

4 Experiments and Results

We perform experiments both with single-language transfer as well as transfer from multiple related languages, if available. We also perform ablations in two settings, with and without hallucinated data. Table 2 presents the exact match accuracy obtained on the test set for a total of 12 language settings. In 7 of them, we observe improvements due to our transliteration preprocessing step, some of them statistically significant.

Specifically, in the top two cases (for Bengali and Maltese as test languages) where the transfer and test languages are closely related, we see improvements across the board. In fact, for Hindi–Bengali and Arabic–Maltese the improvement is statistically significant with $p < 0.05$. Interestingly, the improvements are significant also when we use hallucinated data, which indicates that our transliteration preprocessing step is orthogonal to monolingual data augmentation through hallucination. For the case of Kannada–Telugu, despite the exact match accuracy being the same (66%) for the case without hallucinated data, we observed small improvements on the average Levenshtein distance between the produced and the gold forms (not reported in Table 2).

On the other hand, when transferring from Bashkir to Tatar and Crimean Tatar, even though

[5] https://github.com/antonisa/inflection

[6] We direct the reader to (Anastasopoulos and Neubig, 2019) for further details on the model.

[7] We use 10,000 bootstrap samples and a $\frac{1}{2}$ ratio of samples in each iteration.

all three languages belong to the same branch (Kipchak) of the Turkic language family, transliterating Bashkir into the Roman alphabet that Tatar and Crimean Tatar use leads to performance degradation. In the case of Bashkir–Tatar, the degradation is statistically significant. It is of note, though, that hallucination also does not offer any improvements in these language pairs.

In a surprising result, transliterating Russian into the Roman alphabet, and using it for cross-lingual transfer to Portuguese also leads to statistically significant improvements. Both languages are Indo-European ones, but belong to different branches (Slavic and Romance). Nevertheless, both with and without hallucinated data the performance improves with transliteration, a finding that surely warrants further study.

Last, we discuss the control experiment of Armenian–Kabardian. Kabardian (and Adyghe, displayed for comparison) belong to the Circassian branch of the Northwest Caucasian languages, and are considered closely related, also both using the Cyrillic alphabet; Armenian, in contrast, is an Indo-European language spoken in the same region. First, transferring from Adyghe leads to better performance compared to transfer from Armenian. Converting Armenian to the Roman script has no effect on downstream performance, as expected.

5 Conclusion

With this work we study whether using transliteration as a preprocessing step can improve the accuracy of morphological inflection models under cross-lingual learning regimes. With a few exceptions, most cases indeed show accuracy improvements, some of them statistically significant. We also note that the improvements are orthogonal to those obtained by data augmentation through hallucination, even in typologically distant languages.

While this work represents a first step in the direction of understanding the effect of script differences in morphological inflection, it is still limited in scope, as the experiments were restricted by the lack of reliable transliteration tools for most scripts. Additionally, some of the transliteration models do not account for phenomena that could have an impact in downstream performance, such as vowelization for Abjad scripts like Arabic. As we aim to expand the scale of this study, a future direction will involve training transliteration models between most scripts of the world. This will allow more extensive experimentation, both by incorporating more language pairs and by allowing more control experiments across various scripts. We will also explore the possibility of using the International Phonetic Alphabet as a universal shared space, creating grapheme-to-phoneme and phoneme-to-grapheme models (which are currently not available for most languages).

References

Simon Ager. 2008. Omniglot writing systems and languages of the world. Retrieved March 30, 2020.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proc. EMNLP*, Hong Kong.

Akash Bharadwaj, David R Mortensen, Chris Dyer, and Jaime G Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. 2018. Report of NEWS 2018 named entity transliteration shared task. In *Proceedings of the Seventh Named Entities Workshop*, pages 55–73, Melbourne, Australia. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proc. CoNLL–SIGMORPHON*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proc. CoNLL SIGMORPHON*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden.

2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proc. SIGMORPHON*.

Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68. Citeseer.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153, Gothenburg, Sweden. Association for Computational Linguistics.

Alice Faber. 1992. Phonemic segmentation as epiphenomenon. *The linguistics of literacy*, 21:111–134.

Roman Grundkiewicz and Kenneth Heafield. 2018. Neural machine translation techniques for named entity transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 89–94, Melbourne, Australia. Association for Computational Linguistics.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Soumyadeep Kundu, Sayantan Paul, and Santanu Pal. 2018. A deep learning based approach to transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 79–83, Melbourne, Australia. Association for Computational Linguistics.

Ying Lin, Xiaoman Pan, Aliya Deri, Heng Ji, and Kevin Knight. 2016. Leveraging entity linking and related language projection to improve name transliteration. In *Proceedings of the Sixth Named Entity Workshop*, pages 1–10, Berlin, Germany. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. *Proc. SIGMORPHON*.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber. 2018. Exploiting cross-lingual subword similarities in low-resource document classification. *arXiv preprint arXiv:1812.09617*.

Yi Zhu, Benjamin Heinzerling, Ivan Vulić, Michael Strube, Roi Reichart, and Anna Korhonen. 2019. On the importance of subword information for morphological tasks in truly low-resource languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.