Explorations in Transfer Learning for OCR Post-Correction

Lindia Tjuatja
University of Texas, Austin
lindia.tjuatja@utexas.edu

Shruti Rijhwani, Graham Neubig
Carnegie Mellon University
srijhwan,gneubig@cs.cmu.edu

Abstract

In this abstract, we explore transfer learning to improve post-correction for optical character recognition (OCR), specifically for documents that contain endangered language texts. We extend an existing OCR post-correction model (Rijhwani et al., 2020) by introducing an additional pretraining step on related data, such as text in a related language or available target endangered language datasets that may differ in orthography. Although cross-lingual transfer is often successful in high-resource settings, our preliminary results show that transferring from data in another language decreases performance for this task. On the other hand, we observe small improvements in performance when transferring from additional target language data.

1 Introduction

Current advances in NLP tend to focus on languages with widely available machine-readable data (Joshi et al., 2020). However, for many of the world's languages, especially endangered languages, such data is scarce. Even so, resources for many of these languages can be found in printed materials such as cultural and educational texts, as well as linguistic documents. Optical character recognition (OCR) systems are a way to digitize these documents. However, training such systems often requires vast amounts of data. As a result, non-standard variants of languages and low-resource languages remain a common impediment to off-the-shelf OCR systems (Smith and Cordell, 2018).

Instead of training OCR systems from scratch, we can focus on *post-correcting* the output. Previous work, mainly in high-resource settings, has shown the efficacy of various post-correction approaches in reducing error rates of OCR system outputs, such as n-gram-based models (Tong and Evans, 1996), weighted finite-state transducers (Llobet et al., 2010), and the utilization of spell-checking (Bassil and Alwani, 2012), among others. More recent work by Rijhwani et al. (2020) has shown that post-correcting the output from OCR systems can significantly improve digitization performance on endangered language texts. Nevertheless, creating and fine-tuning these post-correction models still depends on the amount of manually annotated data available.

Since manual annotation can be time consuming and costly, we instead look to improving performance using resources that are more readily available. One such resource is data in a high-resource language that is related to the target endangered language genetically or geographically. In some cases, there may also be digitized data for the target language that is written with different spelling conventions or orthography and/or in a different domain. We explore transfer learning from these resources and evaluate its effect on OCR post-correction performance.

2 Method and Datasets

2.1 Method

Our proposed method builds upon an OCR post-correction model by Rijhwani et al. (2020) which is based on a sequence-to-sequence framework. The model uses an LSTM encoder-decoder with attention, along with added structural biases to improve low-resource learning. Notably, the model is pretrained on first-pass OCR data from the target endangered language before fine-tuning on the manually corrected

text. On top of this framework, we propose two extensions. First, we perform an additional pretraining step over either data in a *related* language or from another dataset in the *target* language (possibly with different orthographic conventions).

Second, we apply *(de)noising rules* on the related resource data. These rules specify probabilities for replacement, insertion, or deletion of a character, e.g., $P(\text{replace '?' with '?'}) = 0.7$, which were calculated from a small portion of the manually corrected endangered language target data. We apply the rules on the related resource data in one of two ways: to "noise" the data (introduce errors we would like the model learn to to fix), or to "denoise" it (which may be helpful in introducing uncommon diacritics or characters that are present in the target).

2.2 Datasets

We use the OCR post-correction dataset from Rijhwani et al. (2020) which contains transcribed data in three endangered languages: Ainu, Griko, and Yakkha. For each target language, we looked for related high-resource language data and, if possible, additional machine-readable endangered language data.

Ainu used related data from the Glossed Audio Corpus of Ainu Folklore (Nakagawa et al., 2021). The corpus contains thirty-eight folk stories recorded by a native Ainu speaker which were transcribed in Latin script and then translated into Japanese. We have *7038 lines of transcribed Ainu gloss data* and *7586 lines of Japanese data* (which we romanized) from the gloss corpus. The transcribed gloss data differs in spelling convention from the Ainu text in our post-correction dataset, particularly in the marking of glottal stops and personal suffixes. Although Ainu is considered a language isolate, we attemp to use Japanese as a transfer language, because they have similar word order and phonotactic structures.

Griko transferred from Greek, as they are partially mutually intelligible. We took data from approximately 100k Greek Wikipedia articles and romanized the text. There are approximately *143k lines of Greek data*. We found additional Griko data from a regional newpaper, *i Spitta*,[1] which contains different spelling conventions compared to the original Griko dataset, as well as spelling variation between articles. There are *8218 lines of Griko news data*.

Yakkha does not have any high-resource related languages that use the same script (Devanagari) in its language family. As an alternative, we use Nepali for cross-lingual transfer, which is a language that uses the Devanagari script and is also spoken in the same region. There are approximately *66k lines of Nepali data from Wikipedia articles*.

3 Experiments

We test three main variations of the additional pretraining step: encoder only, decoder only, and the entire seq2seq model. For the first two, we pretrain the encoder/decoder on $\boldsymbol{x}$, a sequence of characters from the related resource data, with a language model objective. The sequence $\boldsymbol{x}$ can be denoised or noised. For pretraining the entire seq2seq model, the sequence $\boldsymbol{x}$ serves as the input and the target sequence $\boldsymbol{y}$ can either be a copy of $\boldsymbol{x}$ or a denoised version of $\boldsymbol{x}$.

We perform 10-fold cross-validation on the original endangered language dataset and evaluate experiments on two metrics, character error rate and word error rate (Berg-Kirkpatrick et al., 2013). Variations of our proposed extension are compared to the baseline method from Rijhwani et al. (2020).

Table 1 shows that transferring from high-resource language data hurt performance across all target languages. For the Greek and Nepali pretraining experiments, common errors compared to the baseline were largely a result of differences in character and diacritic usage compared to the target Griko and Yakkha text. Examples of such errors can be seen in Figure 1, where the model mistakes accented characters.

ćiupànu	ćiupanu
ivò."	ivo."
tôvale	tèvale
iss'eména	iss'emèna

Figure 1: Examples of common errors from the Griko-Greek model due to differences in diacritic usage. Left shows the gold-standard transcription, right is the incorrect output.

In contrast, transferring from other sources of endangered language text marginally improved performance

[1] https://www.rizegrike.com/spitta.php

Transferring from high-resource language data

	Ainu (+Japanese)		Griko (+Greek)		Yakkha (+Nepali)	
Model	CER	WER	CER	WER	CER	WER
BASELINE	**0.745**	**5.08**	**1.48**	**7.37**	**8.41**	**21.6**
ENC_ONLY	1.06	6.29	2.90	10.2	20.9	41.4
DEC_ONLY	0.837	5.23	2.57	9.51	30.4	58.1
DEC_ONLY [den]	—	—	2.09	8.83	—	—
ENC_DEC	1.17	6.45	—	—	22.5	46.0
ENC_DEC [den]	—	—	3.54	11.0	—	—
ENC_DEC [noi]	1.64	7.17	—	—	15.2	41.9

Table 1: Comparison between baseline and variations of pretraining using high-resource language data. **Bold** indicates per-language best accuracies.

Transferring from additional endangered language data

	Ainu (+Gloss)		Griko (+Newspaper)	
Model	CER	WER	CER	WER
BASELINE	0.745	5.08	1.48	7.37
ENC_ONLY	1.08	6.03	**1.48**	**7.31**
DEC_ONLY	**0.724**	**5.05**	1.70	7.72
DEC_ONLY [den]	—	—	1.51	7.02
ENC_DEC	1.28	6.74	1.96	8.06
ENC_DEC [den]	—	—	1.78	7.83
ENC_DEC [noi]	1.48	7.18	—	—

Table 2: Comparison between baseline and variations of pretraining using additional endangered language data.

for Ainu and Griko, despite differences in orthography between the pretraining and target data. Compared to the baseline, the Ainu gloss pretraining model corrected more punctuation errors, and the Griko newspaper pretraining model made fewer errors with accent markings.

4 Conclusion and Future Work

In conclusion, in the context of OCR post-correction for endangered language text, our results show that effective usage of data from other languages is not straight-forward, but usage of different sources of endangered language text is promising. As our high-resource language data differed in domain from the target endangered language data, further work could be done to determine the impact that domain shift has on transfer learning in this setting, if any. In addition, it would be interesting to see if transferring *between* endangered languages within the same language family would be effective.

Acknowledgements

This work was supported by the National Science Foundation under award number 1761548.

References

Youssef Bassil and Mohammad Alwani. 2012. OCR post-processing error correction algorithm using google online spelling suggestion. *CoRR*, abs/1204.0191.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria, August. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

Rafael Llobet, Jose-Ramon Cerdan-Navarro, Juan-Carlos Perez-Cortes, and Joaquim Arlandis. 2010. Ocr post-processing using weighted finite-state transducers. In *2010 20th International Conference on Pattern Recognition*, pages 2021–2024.

Hiroshi Nakagawa, Anna Bugaeva, Miki Kobayashi, and Yoshikawa Yoshimi. 2021. A Glossed Audio Corpus of Ainu Folklore. `https://ainucorpus.ninjal.ac.jp`.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November.

David A Smith and Ryan Cordell. 2018. A research agenda for historical and multilingual optical character recognition.

Xiang Tong and David A. Evans. 1996. A statistical approach to automatic OCR error correction in context. In *Fourth Workshop on Very Large Corpora*, Herstmonceux Castle, Sussex, UK, June. Association for Computational Linguistics.