Development of Urdu-English Religious Domain Parallel Corpus

Noor e Hira noorehira94@gmail.com
Department of Computer Science, Fatima Jinnah Women University, Pakistan
Sadaf Abdul Rauf sadaf.abdulrauf@gmail.com
Department of Computer Science, Fatima Jinnah Women University, Pakistan

Abstract

Despite the abundance of monolingual corpora accessible online, there remains a scarcity of domain specific parallel corpora. This scarcity poses a challenge in the development of robust translation systems tailored for such specialized domains. Addressing this gap, we have developed a parallel religious domain corpus for Urdu-English. This corpus consists of 18,426 parallel sentences from Sunan Dawood, carefully curated to capture the unique linguistic and contextual aspects of religious texts. The developed corpus is then used to train Urdu-English religious domain Neural Machine Translation (NMT) systems, the best system scored 27.9 BLEU points.

1 Introduction

Neural Machine Translation Bahdanau et al. (2014) has been a field of intense attention for researchers since its advent. It has shown explosive increase in research, introducing new paradigms, revealing new approaches, achieving new milestones and ultimately gaining far better accuracy levels than the previous statistical machine translation (SMT) approaches. NMT Research is not only focused on improving the translation quality of high-resource language pairs, but it also investigates techniques to train machines under different scenarios including monolingual Gibadullin et al. (2019), low-resource Ranathunga et al. (2023), multilingual Dabre et al. (2020), document level NMT Maruf et al. (2021), and much more.

These investigations open new hopes for NMT, but the availability of parallel corpus for training NMT systems is the bottle neck factor to improve translation quality. The more this factor is important the more it is difficult to obtain Munteanu and Marcu (2005); Abdul-Rauf and Schwenk (2009). After years of research on MT till today, only a few languages have huge parallel corpora available, some others have moderate parallel corpus whereas many languages still lack the availability of any parallel corpus for their training.

Training standard NMT systems is a real challenge in low resource settings. Scarcity of available parallel corpus for low resource languages affect translation quality of NMT systems. Same is the case for training domain specific NMT systems, which is subject to the availability of domain specific parallel corpus. Hence, there is a need to investigate and analyse different NMT techniques for low resource settings including domain specific NMT training and adopt the possible ways to improve Urdu-English machine translation which falls under the category of low resource language.

Development of parallel corpus for languages is a time-consuming and tedious task, which

Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation, pages 14–21
September 5, 2023, Macau SAR, China.

sometimes requires the input of native speakers as well Callison-Burch et al. (2011). The Urdu-English parallel corpora as investigated by Abdul Rauf et al. (2020) are not available in abundance. David M. et al. (2021) provided statistics about Urdu highlighting the need of parallel corpora.

The availability of massive monolingual religious translations in English and Urdu motivated our research to develop a religious domain Urdu-English parallel corpus. Despite the abundant availability of religious corpora in multiple languages, parallel corpora are still limited. To our knowledge, $UMC005$ is the only religious domain parallel Urdu-English corpus publicly available (Jawaid and Zeman, 2011; Abdul Rauf et al., 2020). The creation of such corpora for Urdu, a low-resource language holds immense significance, as it enables the adaptation of machine translation systems tailored to this specialized domain.

We have developed a bilingual Urdu-English religious corpus of 18,426 sentences [1]. Section 3 of this paper outlines the detailed steps and procedures taken for the development of this religious parallel corpus. We have also trained NMT models specialized for this domain where the best BLUE score is 27.9. Our NMT experiments are described in Section 4.

2 Related Work

We report the works related to publicly available religious domain parallel corpora specifically the hadith corpora. Altammami et al. (2020) publish the first publicly available bilingual parallel corpus of Islamic Hadith extracted from the six canonical Hadith books; using a domain-specific tool for Hadith segmentation, resulting in bilingual English-Arabic parallel corpus[2] of 39,038 annotated Hadiths. However, Sunan Dawood is automatically aligned in their work where they report an accuracy of 92%, whereas our corpus is aligned and checked manually. Abdul Rauf et al. (2020) provide details about all the publicly available corpora for Urdu-English language pair in biomedical, religious, technological, and general domain. We have used all the corpora mentioned in the study for our NMT experiments.

3 Methodology

Despite the fact that religious books and documents are available over the internet in massive amounts, along with their translations in many languages including English and Urdu, the creation of a religious domain Urdu-English parallel corpus is not easy as both languages have far different sentence segmentation and arrangement. The text of the available translations is coherent, as per the needs of language proficiency and flow. The difference in sentence structure of both the languages and the coherency of the text makes automatic sentence segmentation almost impossible. Our corpus development cycle includes four different stages, collection of available translations, manual filtering of collected data, extraction of parallel translations, and sentence-level segmentation of parallel texts.

3.1 Source Data Collection

The first step of our corpus development cycle included the search and collection of available translations of religious texts. Although, abundant religious text is available over the internet for English Urdu language pair, but the format of the documents is not suitable for MT corpus development research. The foremost hurdle faced during corpus collection was to search for books or documents in Unicode format. We were able to find and download *Sunan Abu Dawood*, a hadith book among the six major hadith books collected by Abu Dawud al-Sijistani from

[1] `https://github.com/sabdul111/SunanDaud-Urdu-English-Parallel-Corpus`

[2] The corpus is named as *Leeds and King Saud University (LK) Hadith corpus*

Figure 1: Sample of Sunan Abu Dawood files after extraction from PDF.

IslamicUrduBooks [3]. The website provides access to many hadith books in unicode format, but only *Sunan Abu Dawood* was available with English and Urdu translations. Arabic text of each hadith is followed by its Urdu and English translation respectively. Few hadiths had some extra information embedded in between Urdu and English translations of Arabic text. Figure 1 shows the format of Sunan Abu Dawood file.

[3]https://islamicurdubooks.com/books/word-files/

Source	Files	Words		lines
		English	Urdu	
SD1	3,194	83,093	99,770	8,160
SD2	2,952	80,775	96,784	8,495
SD3	1,878	19,639	25,594	1,771
Total	8,024	183,507	222,148	18,426

Table 1: Urdu-English Religious Domain Corpus, SD1 represents Sunan Abu Dawood volume 1, SD2 volume 2, and SD3 volume 3

3.2 Data Filtering

We manually inspected the files and applied different filtering steps to convert them to parallel bi-texts. Document filtration included removal of content tables, figures, and objects. The text file was then manually inspected to identify the extra information embedded in between the translations. Such information had specific keywords such as Takhreej Darul Da'wah, Wazahat etc. Scripts were used where appropriate to remove extra content using specified keywords. Additionally, hadith numbers and blank lines were eliminated.

3.3 Parallel Translation Extraction

In this step, the filtered files were further examined to ensure that each hadith contained translations in both English and Urdu languages. The line numbers of the English text were observed, as each hadith consisted of Arabic text on the first line, Urdu text on the second line, and English text on the third line. Any discrepancies in the line numbers were manually corrected by backtracking through the file to identify and remove the problematic content. In cases where translations were missing in one language, a placeholder text such as "translation not available" was added in the respective language to maintain line number consistency without compromising the contents for the other two languages. Scripts were utilized to separate the text of each hadith into three distinct files: one for Arabic, one for Urdu, and one for English. A manual inspection of the main file was conducted to verify the accurate extraction of each language's text. If successful, the process moved forward; otherwise, steps were retraced and adjustments were made to address any errors.

3.4 Parallel Sentence Splitting

This step involved splitting the extracted text into parallel smaller phrases or sentences, focusing on the English and Urdu files. Manual splitting was chosen over automatic methods to ensure corpus content accuracy. Volunteers with proficiency in these languages were chosen from graduate students. To assess the volunteers' understanding, an initial submission of a few hadiths was evaluated. Only a small percentage demonstrated complete comprehension, prompting adjustments, and the provision of a demo video. Subsequent submissions showed significant improvement, reinforcing the chosen approach. Each student's work was reviewed to ensure correct alignment, and files with errors were reassigned to students with greater accuracy.

First two volumes of Sunan Abu Dawood, along with selected hadiths from the third volume, were successfully processed. Table 1 presents the statistics for the developed religious corpus, with SD1 representing Sunan Abu Dawood Volume 1, SD2 denoting Volume 2, and SD3 referring to Volume 3.

4 Urdu-English Neural Machine Translation

This section describes the results of NMT systems trained using our developed corpus and other publicly available Urdu-English corpora.

4.1 Corpora

The study of Abdul Rauf et al. (2020) provides details about all the publicly available corpora for Urdu-English language pair. We have used all the corpora mentioned in the study and some additional corpora as explained below and listed in Table 2.

- The Emille[4] (Baker et al., 2002) is a 97 million word corpus developed under a joint project of Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. It is a collection of monolingual, parallel and annotated corpora for fourteen South Asian Languages including Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu and Urdu. The corpus comprises of data in both textual and spoken formats and is freely distributed by ELRA (European Language Resource Association) for research purposes.
- Indic[5] is a corpus comprising texts for six indian languages including Bengali, Hindi, Malayalam, Tamil, Telugu and Urdu. The corpus was developed from top 100 most visited documents of Wikipedia. Corpus was constructed using Amazon's Mechanical Turk (MTurk) for crowd sourcing. (Post et al., 2012).
- OPUS[6] (Tiedemann, 2012) is a resource which provides access to freely available annotated parallel corpora, collected from web resources and processed automatically. OPUS contains fourteen different corpora for Urdu-English language pair, including CCAligned, CCMatrix, GlobalVoices, GNOME, Mozilla, OpenSubtitles, QED, Tanzil, Tatoeba, TED, Tico, Ubuntu, Wikimedia and XLEnt. We used all these corpora for our experiments.
- Jawaid and Zeman (2011) collected translations of Quran and Bible from web, which is different from Tanzil corpus provided by OPUS. Their collection, UMC005[7], contains two other corpora for Urdu-English language pair, but Only Quran and Bible are available for free.
- Urdu translations of Wall Street Journal (WSJ), a subset of *Penn Treebank* Marcus et al. (1993) have been released by CLE[8]. We collected the Urdu translations of this corpus form official website of CLE and their corresponding English translations were awarded from LDC as data scholarship we applied for.
- QBJ (Quran+Bible+Joshua) corpus is another collection of freely available Urdu-English corpus. It has 1.02M English words and 1.13M Urdu words.
- PMindia is a parallel corpus of Indian languages extracted from the website of the Prime Minister of India (www.pmindia.gov.in). The corpus provides parallel sentences for thirteen major languages of India.
- SD is the Urdu-English religious domain corpus having parallel ahadith from 3 volumes of Sunan Abu Dawood that we developed during this work.

[4]The Emille/CIIL Corpus:ID:ELRA-W0037
[5]http://joshua-decoder.org/indian-parallel-corpora/
[6]http://opus.nlpl.eu/
[7]https://ufal.mff.cuni.cz/umc/005-en-ur/
[8]http://www.cle.org.pk/

Category	Corpus	tokens		Sentences
		English	Urdu	
Out-domain	CCAligned	18M	23M	1,371,930
	CCMatrix	67M	80M	6,094,149
	Emily	89K	0.1M	5,877
	Global Voices	72K	82K	4,103
	Gnome	42K	50,k	11,535
	Indic	0.5M	0.6M	35,139
	Open-Subtitles	0.17M	2.0M	29,074
	PMindia	0.2M	0.26M	11,167
	QED	0.25M	0.29M	19,053
	Tatoeba	10K	12k	1,667
	TED	0.26	0.32	15,755
	Tico	70K	91K	3,071
	Treebank	0.13M	0.18	5,693
	Ubuntu	10K	12K	3,025
	Wikimedia	2.0M	3M	43,168
	XLEnt	2.0M	2.1M	746,804
	Total	**91.5M**	**111M**	**8.4M**
In-domain	Tanzil	19M	23M	748320
	OBJ	1.0M	1.1M	49510
	Bible	0.21M	0.20M	7957
	Quran	0.25M	0.24M	6414
	SunanDawood	0.19M	0.23M	20678
	Total	**20.1M**	**24.8M**	**832879**

Table 2: Indomain and out domain Urdu-English training corpora

ID	Train Set (No of sentences)	Size	scores
M1	Out_D	8,401,210	14.5
M2	In_D	832,879	**27.9**
M3	$Out_D \xrightarrow{adapt} In_D$	832,879	21.4

Table 3: BLEU scores

4.2 Preprocessing

Corpus preprocessing is an essential part of building machine learning systems. Three of the corpora, Emillie, NLT and Penn Tree-bank were partially aligned. We used LF sentence aligner[9] to align these corpora but due to the topological differences between the two languages results obtained from LF aligner were not accurate and, thus manual alignment was done to ensure correctness. Tokenization, using mosses tokenizer[10], truecasing and BPE (Sennrich et al., 2016), were applied to all the corpora during pre-processing.

[9]https://sourceforge.net/projects/aligner/

[10]`https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl`

4.3 NMT Experiments

We trained three NMT models using the transformer (Vaswani et al., 2017) architecture. The models were evaluated using religious domain test set as our objective was to build and improve the accuracy of religious domain translation models. The religious domain Urdu-English corpora was split in a ratio of 8:1:1 for train, validation and test-set respectively.

The $M1$ model was trained using out-domain corpus, i.e. all the Urdu-English corpus other than the religious domain and it scored 14.5 BLEU points.

$M2$, the model trained on in-domain data outscored $M1$ by 13.4 BLEU points. This result is inline with existing research highlighting the importance of domain for the training corpora. A system built on the same domain as the test set will give better translations.

Lastly we experimented with domain adaptation $M3$, i.e. improve domain-specific machine translation using indomain data to adapt the out domain model towards the religious domain. For $M3$, though performance improved as compared to $M1$ giving 21.4 BLEU scores on the test-set but still it did not outperform $M2$.

Our results show the importance of in-domain corpus. System trained on only small amount of religious domain corpus is better than system trained on large general domain data and fine tuned on in domain corpora.

5 Conclusion

In this paper, we have successfully tackled the challenges of developing a parallel Urdu-English corpus in the religious domain. The meticulous process of acquiring, processing, and aligning the data resulted in a corpus comprising 18,426 lines. The developed corpus underwent a thorough analysis to ensure the accuracy and integrity of data. It is then used to train Urdu-English religious domain NMT systems, the best systems scored 27.9 BLEU points. These findings underscore the effectiveness of the corpus in enabling accurate and meaningful translations within the religious context.

Acknowledgments

This study is funded by the National Research Program for Universities (NRPU) by Higher Education Commission of Pakistan (5469/Punjab/NRPU/R&D/HEC/2016).

References

Abdul Rauf, S., Abida, S., Hira, N.-e., Zahra, S., Parvez, D., Bashir, J., and Majid, Q.-u.-a. (2020). On the exploration of English to Urdu machine translation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 285–293, Marseille, France. European Language Resources association.

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece. Association for Computational Linguistics.

Altammami, S., Atwell, E., and Alsalka, A. (2020). The arabic-english parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 8(2).

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Baker, P., Hardie, A., McEnery, T., Cunningham, H., and Gaizauskas, R. J. (2002). Emille, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation. In *LREC*.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. F. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

David M., E., Simons, G. F., and Fennig, C. D. (2021). Ethnologue: Languages of the world. twenty-fourth edition. dallas, texas: Sil international.

Gibadullin, I., Valeev, A., Khusainova, A., and Khan, A. (2019). A survey of methods to leverage monolingual data in low-resource neural machine translation. *arXiv preprint arXiv:1910.00373*.

Jawaid, B. and Zeman, D. (2011). Word-order issues in english-to-urdu statistical machine translation. *Prague Bull. Math. Linguistics*, 95:87–106.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Maruf, S., Saleh, F., and Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. *Wmt-2012*, pages 401–409.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Sennrich, R., Haddow, B., and Birch, A. (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.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Ugur Dogan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.