Itihāsa: A large-scale corpus for Sanskrit to English translation

Rahul Aralikatte[1] **Miryam de Lhoneux**[1] **Anoop Kunchukuttan**[2] **Anders Søgaard**[1]

[1]University of Copenhagen   [2]Microsoft AI and Research

[1]{rahul,ml,soegaard}@di.ku.dk
[2]ankunchu@microsoft.com

Abstract

This work introduces Itihāsa, a large-scale translation dataset containing 93,000 pairs of Sanskrit *shloka*s and their English translations. The *shloka*s are extracted from two Indian epics viz., The Rāmāyana and The Mahābhārata. We first describe the motivation behind the curation of such a dataset and follow up with empirical analysis to bring out its nuances. We then benchmark the performance of standard translation models on this corpus and show that even state-of-the-art transformer architectures perform poorly, emphasizing the complexity of the dataset.[1]

1 Introduction

Sanskrit is one of the oldest languages in the world and most Indo-European languages are influenced by it (Beekes, 1995). There are about 30 million pieces of Sanskrit literature available to us today (Goyal et al., 2012), most of which have not been digitized. Among those that have been, few have been translated. The main reason for this is the lack of expertise and funding. An automatic translation system would not only aid and accelerate this process, but it would also help in democratizing the knowledge, history, and culture present in this literature. In this work, we present Itihāsa, a large-scale Sanskrit-English translation corpus consisting of more than 93,000 *shlokas* and their translations.

Itihāsa, literally meaning 'it happened this way' is a collection of historical records of important events in Indian history. These bodies of work are mostly composed in the form of verses or *shloka*s, a poetic form which usually consists of four parts containing eight syllables each (Fig. 1). The most important among these works are The Rāmāyana

मा निषाद प्रतिष्ठां त्वमगमश्शाश्वतीस्समाः।
यत्क्रौञ्चमिथुनादेकमवधीः काममोहितम्॥

O fowler, since you have slain one of a pair of Krauñcas, you shall never attain prosperity (respect)!

Figure 1: An introductory *shloka* from The Rāmāyana. The four parts with eight syllables each are highlighted with different shades of gray.

and The Mahābhārata. The Rāmāyana, which describes the events in the life of Lord Rāma, consists of 24,000 verses. The Mahābhārata details the war between cousins of the Kuru dynasty, in 100,000 verses. The Mahābhārata is the longest poem ever written with about 1.8 million words in total and is roughly ten times the length of the Iliad and the Odyssey combined.

Only two authors have attempted to translate the unabridged versions of both The Rāmāyana and The Mahābhārata to English: Manmatha Nāth Dutt in the 1890s and Bibek Debroy in the 2010s. M. N. Dutt was a prolific translator whose works are now in the public domain. These works are published in a *shloka*-wise format as shown in Fig. 1 which makes it easy to automatically align *shloka*s with their translations. Though many of M. N. Dutt's works are freely available, we choose to extract data from The Rāmāyana (Vālmiki and Dutt, 1891), and The Mahābhārata (Dwaipāyana and Dutt, 1895), mainly due to its size and popularity. As per our knowledge, this is the biggest Sanskrit-English translation dataset to be released in the public domain.

We also train and evaluate standard translation systems on this dataset. In both translation directions, we use Moses as an SMT baseline, and Transformer-based seq2seq models as NMT baselines (see §4). We find that models which are generally on-par with human performance on other

---

[1]The processed and split dataset can be found at https://github.com/rahular/itihasa and a human-readable version can be found at http://rahular.com/itihasa.
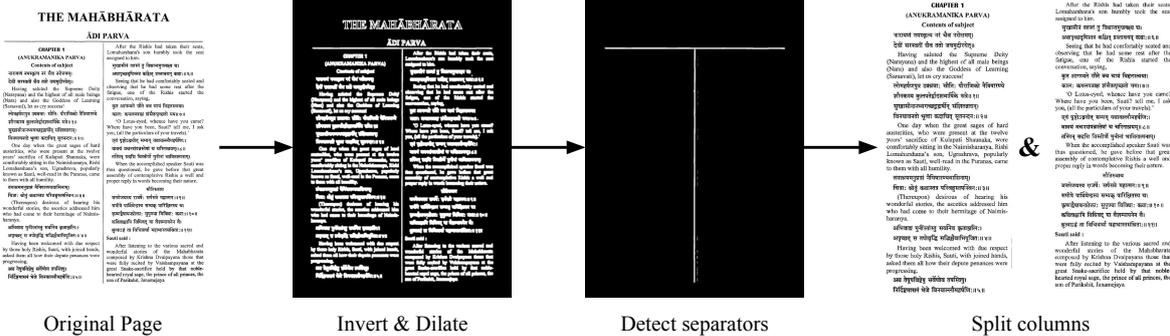
Figure 2: Pre-processing Pipeline: The three steps shown here are: (i) invert the colour scheme of the PDF and dilate every detectable edge, (ii) find the indices of the longest vertical and horizontal lines in the page, and (iii) split the original PDF along the found separator lines.

translation tasks, perform poorly on Itihāsa, with the best models scoring between 7-8 BLEU points. This indicates the complex nature of the dataset (see §3 for a detailed analysis of the dataset and its vocabulary).

Motivation The main motivation behind this work is to provide an impetus for the Indic NLP community to build better translation systems for Sanskrit. Additionally, since The Rāmāyana and The Mahābhārata are so pervasive in Indian culture, and have been translated to all major Indian languages, there is a possibility of creating an *n*-way parallel corpus with Sanskrit as the pivot language, similar to Europarl (Koehn, 2005) and PMIndia (Haddow and Kirefu, 2020) datasets.

The existence of Sanskrit-English parallel data has other advantages as well. Due to Sanskrit being a morphologically rich, agglutinative, and highly inflexive, complex concepts can be expressed in compact forms by combining individual words through *Sandhi* and *Samasa*.[2] This also enables a speaker to potentially create an infinite number of unique words in Sanskrit. Having a parallel corpus can help us induce word translations through bilingual dictionary induction (Søgaard et al., 2018). It also allows us to use English as a surrogate language for tasks like knowledge base population. Constituency or dependency parsing, NER, and word sense disambiguation can be improved using indirect supervision (Täckström, 2013). Essentially, a parallel corpus allows us to apply a plethora of transfer learning techniques to improve

NLP tools for Sanskrit.

2 Data Preparation

The translated works of The Rāmāyana and The Mahābhārata were published in four and nine volumes respectively.[3] All volumes have a standard two-column format as shown in Fig. 2. Each page has a header with the chapter name and page number separated from the main text by a horizontal line. The two columns of text are separated by a vertical line. The process of data preparation can be divided into (i) automatic OCR extraction, and (ii) manual inspection for alignment errors.

Automatic Extraction The OCR systems we experimented with performed poorly on digitized documents due to their two-column format. They often fail to recognize line breaks which result in the concatenation of text present in different columns. To mitigate this issue, we use an edge detector[4] to find the largest horizontal and vertical lines, and using the indices of the detected lines, split the original page horizontally and vertically to remove the header and separate the columns (see Fig. 2). We then input the single-column documents to Google Cloud's OCR API[5] to extract text from them. To verify the accuracy of the extracted text, one chapter from each volume (13 chapters in total) is manually checked for mistakes. We find that the extracted text is more than 99% and 97% accurate in Sanskrit and English respectively. The surprising accuracy of Devanagari OCR can be attributed to

---

[2]*Sandhi* refers to the concatenation of words, where the edge characters combine to form a new one. *Samasa* can be thought of as being similar to elliptic constructions in English where certain phrases are elided since their meaning is obvious from the context.

[3]The digitized (scanned) PDF versions of these books are available at https://hinduscriptures.in

[4]We invert the color scheme and apply a small dilation for better edge detection using OpenCV (Bradski, 2000).

[5]More information can be found at https://cloud.google.com/vision/docs/pdf

इति मत्वा प्रियं पुत्रं भीष्मादाय ततो ह्वहम्।
पूर्वस्नेहानुरागित्वात् सदारः सौमकिं गतः॥६०॥

O Bhishma, thus resolved and remembering my former friendship for him (Drupada) I regarded myself very much blessed. I went joyfully to the Saumaka, taking my beloved son and wife me.

(a) Print error.

वृत्तानि रथयुद्धानि विचित्राणि पदे पदे॥२९॥
इदमेकं गदायुद्धं भवत्वद्याद्भुतं महत्।

Many wonderful single combats have takes place on cars. Let his one great and wonderful encounter with the mace take place today.

(b) Input error.

उच्छृष्टैनं तथा वाक्यं मानुषाणां त्वमीश्वरः॥६८॥
असिना धर्मगर्भेण पालयस्व प्रजा इति।

At the time of giving it to Manu, they said— You are the lord of all men. Protect all creatures with this sword having religion within its womb.

(c) Subjective error.

Figure 3: Different types of errors found in the original text while performing manual inspection.

the distinctness of its alphabet. For English, this number decreases as the OCR system often misclassifies similar-looking characters (viz., *e* and *c*, *i* and *l*, etc.).

Manual Inspection An important limitation of the OCR system is its misclassification of alignment spaces and line breaks. It sometimes wrongly treats large gaps between words as line breaks and the rest of the text on the line is moved to the end of the paragraph which results in translations being misaligned with its *shloka*s. Therefore, the output of all 13 volumes was manually inspected and such misalignments were corrected.[6]

Upon manual inspection, other kinds of errors were discovered and corrected where possible.[7] These errors can be categorized as follows: (i) *print errors*: this type of error is caused by occluded or faded text, smudged ink, etc. An example can be seen in Fig. 3a, (ii) *input errors*: these are human errors during typesetting the volumes which include typos (Fig, 3b), exclusion of words, inclusion of spurious words, etc., (iii) *subjective errors*: these are contextual errors in the translation itself. For example, in Fig. 3c, the word *dharma* is incorrectly translated as 'religion' instead of 'righteousness', and (iv) *OCR errors*: these errors arise from the underlying OCR system. An example of such errors is the improper handling of split words across lines in the Devanagari script. If the OCR system encounters a hyphen as the last character of a line, the entire line is ignored. In general, print errors are corrected as much as possible, subjective errors are retained for originality, and other types of errors are corrected when encountered.

---

[6]This was a time-consuming process and the first author inspected the output manually over the course of one year.

[7]It was not feasible for the authors to correct every error, especially the lexical ones. The most common error that exists in the corpus is the swapping of *e* and *c*. For example, 'thcir' instead of 'their'. Though these errors can easily be corrected using automated tools like the one proposed in (Boyd, 2018), it is out-of-scope of this paper and is left for future work.

	Train	Dev.	Test	Total
Rāmayana				
Chapters	514	42	86	642
Shlokas	15,834	1,115	2,422	19,371
Mahābhārata				
Chapters	1,688	139	283	2,110
Shlokas	59,327	5,033	9,299	73,659
Overall				
Chapters	2202	181	369	2,752
Shlokas	75,161	6,148	11,721	93,030

Table 1: Size of training, development, and test sets.

3 Analysis

In total, we extract 19,371 translation pairs from 642 chapters of The Rāmāyana and 73,659 translation pairs from 2,110 chapters of The Mahābhārata. It should be noted that these numbers do not correspond to the number of *shloka*s because, in the original volumes, *shloka*s are sometimes split and often combined to make the English translations flow better. We reserve 80% of the data from each text for training MT systems and use the rest for evaluation. From the evaluation set, 33% is used for development and 67% for testing. The absolute sizes of the split data are shown in Tab. 1.

Due to Sanskrit's agglutinative nature, the dataset is asymmetric in the sense that, the number of words required to convey the same information, is less in Sanskrit when compared with English. The Rāmāyana's English translations, on average, have 2.54 words for every word in its *shloka*. This value is even larger in The Mahābhārata with 2.82 translated words per *shloka* word.

This effect is clearly seen when we consider the vocabulary sizes and the percentage of common tokens between the texts. For this, we tokenize the data with two different tokenization schemes: word-level and byte-pair encoding (Sennrich et al., 2016, BPE). For word-level tokenization, the translations of The Rāmāyana (The Mahābhārata) have 16,820 (31,055) unique word tokens, and the *shloka*s have 66,072 (184,407) tokens. The English vocabularies
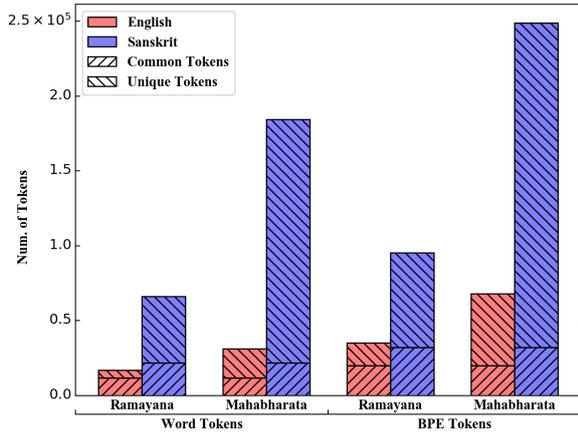
Figure 4: Comparison of vocabulary sizes. Sanskrit's morphological and agglutinative nature accounts for the large number of unique tokens in the vocabularies.

Model	chrF	Tok. Acc.	BLEU	TER ($\downarrow$)
English to Sanskrit				
Moses	25.9	5.48	0.21	1.53
B2B-`Tiny`	15.1	8.07	2.85	1.04
B2B-`Mini`	21.6	8.52	5.66	1.03
B2B-`Small`	23.4	8.75	6.93	1.03
B2B-`Medium`	23.7	8.67	6.94	**1.02**
B2B-`Base`	**24.3**	**8.89**	**7.59**	1.04
Sanskrit to English				
Moses	29.3	8.03	5.67	**0.91**
B2B-`Tiny`	24.5	**8.61**	5.64	0.98
B2B-`Mini`	29.3	8.58	7.28	0.95
B2B-`Small`	30.1	8.55	**7.49**	0.95
B2B-`Medium`	30.4	8.49	7.48	0.94
B2B-`Base`	**30.5**	8.38	7.09	0.93

Table 2: Character F1, Token accuracy, BLEU, and TER scores for Moses and Transformer models. Scores marked with ($\downarrow$) are better if they are lower.

have 11,579 common tokens which is 68.8% of The Rāmāyana's and 37.3% of The Mahābhārata's. But the overlap percentages drop significantly for the Sanskrit vocabularies. In this case, we find 21,635 common tokens which amount to an overlap of 32.7% and 11.7% respectively. As shown in Fig. 4, this trend holds for BPE tokenization as well.

4 Experiments

We train one SMT and five NMT systems in both directions and report the (i) character n-gram F-score, (ii) token accuracy, (iii) BLEU (Papineni et al., 2002), and (iv) Translation Edit Ratio (Snover et al., 2006, TER) scores in Tab. 2. For SMT, we use Moses (Koehn et al., 2007) and for NMT, we use sequence-to-sequence (seq2seq) Transformers (Vaswani et al., 2017). We train the seq2seq models from scratch by initializing the encoders and decoders with standard BERT (B2B) architectures. These `Tiny`, `Mini`, `Small`, `Medium`, and `Base` models have 2/128, 4/256, 4/512, 8/512, and 12/768 layers/dimensions respectively. See Turc et al. (2019) for more details. In our early experiments, we also tried initializing the encoders and decoders with weights from pre-trained Indic language models like MuRIL (Khanuja et al., 2021), but they showed poor performance and thus are not reported here.

Implementation Details All models are trained using HuggingFace Transformers (Wolf et al., 2020). Both source and target sequences are truncated at 128 tokens. We train WordPiece tokenizers on our dataset and use them for all models. Adam optimizer (Kingma and Ba, 2014) with weight-

decay of 0.01, and learning rate of 5×10^{-5} is used. All models are trained for 100 epochs. The learning rate is warmed up over 8,000 steps and decayed later with a linear scheduler. We use a batch size of 128, and use standard cross-entropy loss with no label smoothing. We run into memory errors on bigger models (`medium` and `base`), but maintain the effective batch-size and optimization steps by introducing gradient accumulation and increasing the number of epochs, respectively. Also, to reduce the total training time of bigger models, we stop training if the BLEU score does not improve over 10 epochs. During generation, we use a beam size of 5 and compute all metrics against truncated references.

Discussion We see that all models perform poorly, with low token accuracy and high TER. While the English to Sanskrit (E2S) models get better with size, this pattern is not clearly seen in Sanskrit to English (S2E) models. Surprisingly for S2E models, the token accuracy progressively decreases as their size increases. Also, Moses has the best TER among S2E models which suggests that the seq2seq models have not been able to learn even simple co-occurrences between source and target tokens. This leads us to hypothesize that the Sanskrit encoders produce sub-optimal representations. One way to improve them would be to add a *Sandhi-splitting* step to the tokenization pipeline, thereby decreasing the Sanskrit vocabulary size. Another natural extension to improve the quality of representations would be to initialize the encoders with a pre-trained language model.

EN	**Hearing the words of Viśvāmitra, Rāghava, together with Laksmana, was struck with amazement, and spoke to Viśvāmitra, saying,**
SN	विश्वामित्रवचः श्रुत्वा राघवः सहलक्ष्मणः। विस्मयं परमं गत्वा विश्वामित्रमथाब्रवीत्॥
Pred.	विश्वामित्रवचः श्रुत्वा लक्ष्मणः सहलक्ष्मणः। विश्वामित्रवचः श्रुत्वा विश्वामित्रोऽब्रवीदिदम्॥

Figure 5: A gold sentence and *shloka* from the test set, and its corresponding `small` model prediction.

Though it is clear that there is a large scope for improvement, the models are able to learn some interesting features of the dataset. Fig. 5 shows a random gold translation pair and the `small` model's prediction. Though we see repetitions of phrases and semantic errors, the prediction follows the meter in which the original *shloka*s are written, i.e. it also consists of 4 parts containing 8 syllables each.

5   Related Work

Early translation efforts from Sanskrit to English were limited to the construction of dictionaries by Western Indologists (Müller, 1866; Monier-Williams, 1899). Over the years, though notable translation works like Ganguli (1883) have been published, the lack of digitization has been a bottleneck hindering any meaningful progress towards automatic translation systems. This has changed recently, at least for monolingual data, with the curation of digital libraries like GRETIL[8] and DCS[9]. Currently, the largest freely available repository of translations are for The Bhagavadgita (Prabhakar et al., 2000) and The Rāmāyana (Geervani et al., 1989).

However, labeled datasets for other tasks, like the ones proposed in (Kulkarni, 2013; Bhardwaj et al., 2018; Krishnan et al., 2020) have resulted in parsers (Krishna et al., 2020, 2021) and sandhi splitters (Aralikatte et al., 2018; Krishnan and Kulkarni, 2020) which are pre-cursors to modular translation systems. Though there have been attempts at building Sanskrit translation tools (Bharati and Kulkarni, 2009), they are mostly rule-based and rely on manual intervention. We hope that the availability of the Itihāsa corpus pushes the domain towards end-to-end systems.

---

[8] http://gretil.sub.uni-goettingen.de/gretil.html
[9] http://www.sanskrit-linguistics.org/dcs/index.php

6   Conclusion

In this work, we introduce Itihāsa, a large-scale dataset containing more than 93,000 pairs of Sanskrit *shlokas* and their English translations from The Rāmāyana and The Mahābhārata. First, we detail the extraction process which includes an automated OCR phase and a manual alignment phase. Next, we analyze the dataset to give an intuition of its asymmetric nature and to showcase its complexities. Lastly, we train state-of-the-art translation models which perform poorly, proving the necessity for more work in this area.

Acknowledgements

References

Rahul Aralikatte, Neelamadhav Gantayat, Naveen Panwar, Anush Sankaran, and Senthil Mani. 2018. Sanskrit sandhi splitting using seq2(seq)2. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4909–4914, Brussels, Belgium. Association for Computational Linguistics.

Robert Stephen Paul Beekes. 1995. *Comparative Indo-European Linguistics*. Benjamins Amsterdam.

Akshar Bharati and Amba Kulkarni. 2009. Anusaaraka: an accessor cum machine translator. *Department of Sanskrit Studies, University of Hyderabad, Hyderabad*, pages 1–75.

Shubham Bhardwaj, Neelamadhav Gantayat, Nikhil Chaturvedi, Rahul Garg, and Sumeet Agarwal. 2018. SandhiKosh: A benchmark corpus for evaluating Sanskrit sandhi tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Krishna Dwaipāyana and Manmatha Nāth Dutt. 1895. *Mahabharata*. Elysium Press, Calcutta.

Kisari Mohan Ganguli. 1883. *The Mahabharata*.

P Geervani, K Kamala, and V. V. Subba Rao. 1989. Valmiki ramayana.

Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf, and Ralph Bunker. 2012. A distributed platform for Sanskrit processing. In *Proceedings of COLING 2012*, pages 1011–1028, Mumbai, India. The COLING 2012 Organizing Committee.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Amrith Krishna, Ashim Gupta, Deepak Garasangi, Pavankumar Satuluri, and Pawan Goyal. 2020. Keep it surprisingly simple: A simple first order graph based parsing model for joint morphosyntactic parsing in Sanskrit. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4791–4797, Online. Association for Computational Linguistics.

Amrith Krishna, Bishal Santra, Ashim Gupta, Pavankumar Satuluri, and Pawan Goyal. 2021. A Graph-Based Framework for Structured Prediction Tasks in Sanskrit. *Computational Linguistics*, 46(4):785–845.

Sriram Krishnan and Amba Kulkarni. 2020. Sanskrit segmentation revisited.

Sriram Krishnan, Amba Kulkarni, and Gérard Huet. 2020. Validation and normalization of dcs corpus using sanskrit heritage tools to build a tagged gold corpus.

Amba Kulkarni. 2013. A deterministic dependency parser with dynamic programming for Sanskrit. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 157–166, Prague, Czech Republic. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Monier Monier-Williams. 1899. *A Sanskrit-English dictionary : etymologically and philologically arranged with special reference to cognate Indo-European languages*. Clarendon Press.

F.M. Müller. 1866. *A Sanskrit Grammar for Beginners: In Devanâgarî and Roman Letters Throughout*. Handbooks for the study of Sanskrit. Longmans, Green, and Company.

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.

T. V Prabhakar, Ravi Mula, T Archana, Harish Karnick, Nitin Gautam, Murat Dhwaj Singh, Madhu Kumar Dhavala, Rajeev Bhatia, and Saurabh Kumar. 2000. Gita supersite.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Oscar Täckström. 2013. *Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision*. Ph.D. thesis, Uppsala University.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefine-dukasz Kaiser, and Illia Polosukhin. 2017. Attention

is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Vālmiki and Manmatha Nāth Dutt. 1891. *Ramayana*. Deva Press, Calcutta.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.