Synthetic Data for Multilingual NLP: a Survey

Meet Doshi and Pushpak Bhattacharyya

CFILT, Indian Institute of Technology Bombay, India

{meetdoshi,pb}@cse.iitb.ac.in

Abstract

In this survey paper, we provide a comprehensive overview of approaches to effectively utilize synthetic data for various Natural Language Processing (NLP) applications. Many NLP tasks are solved by the paradigm of pre-training followed by fine-tuning, where the task solver is first initialized using large-scale unlabeled examples and then fine-tuned with strong supervision using examples that mimic the target task. Both stages can benefit from additional data generated synthetically, either through heuristics or end-to-end training. This synthetic data can be utilized in multiple ways, such as generating weak labels for downstream applications or creating entirely new samples from a generative model. Along with the benefits of synthetic data, this paper also highlights the different challenges associated with its generation and curation, with a key challenge being the identification of imposters among the acceptable samples. Some of these challenges are also seen in multilingual applications, where data scarcity is a major issue. Our aim with this survey is to provide readers with an understanding of current approaches to leveraging synthetic data in the context of multilingual applications.

1 Introduction

With the advent of new machine learning models every day, they have been able to show strong performance in many NLP applications. Most of their performance can be credited to the sheer size and amount of data that they have been trained on. Many efforts have been able to put together vast amounts of resources to make data available on the internet open to the public. A majority of this data is already being utilized in some of the recent LLMs (Touvron et al., 2023), which raises the question of whether we will ever run out of data. A study done by Villalobos et al. (2024) shows that we will run out of fresh text data by the year 2050. Even if we can leverage clean data until then, many problems exist like noisy samples, scarcity, privacy concerns, domain bias, and the amount of effort and time that can go into collecting or annotating datasets.

Synthetic data refers to any type of data that can be used to mimic a task at hand and follows the pattern of data from the real world. It can be generated by heuristics, algorithms, contextualized generation, or even imitation. The limitation of the synthetic data relies highly on the method of generation and the task at hand. Despite its cons, synthetic data can be very useful in scenarios where real-world data is infeasible to collect or make due to its ability to generate at scale. We will highlight some of the common practices and issues faced while utilizing synthetic data.

Large language models (LLMs) (Brown et al., 2020a; Workshop et al., 2022; Almazrouei et al., 2023; Lin et al., 2022) have been able to perform very well on downstream tasks like MMLU (Hendrycks et al., 2021), Big-Bench (Srivastava et al., 2022), etc, and have even started to reach human potential in many of these tasks. But this performance has very largely been credited to their scale and the vast amount of data that they have been fed. Due to an increase in language understanding and generation abilities, LLMs have become a prime candidate for generating synthetic data. Most of these language models (LMs) perform well in languages like English where abundant data is available (Kudugunta et al., 2023), but a vast majority of languages don't have comparable data as compared to English. As a consequence, many LLMs, both monolingual and multilingual, involving these languages still show poor performance for various downstream tasks. For example, the largest open source multilingual model BLOOM (Workshop et al., 2022) covers 46 natural languages spanning 9 language families, but the top 5 languages comprise 74.14% of the data. Despite the benefits of multilingualism (Dabre et al., 2020),

this skew in data still means that the low-resource languages will not perform well.

Fortunately, synthetic data has shown to be useful for many multilingual NLP applications like using back translations (Sennrich et al., 2016a), (Edunov et al., 2018) for improving Machine translation performance (Marie et al., 2020),(Bogoychev and Sennrich, 2019),(Ni et al., 2022) or for classification tasks like native language identification (Goldin et al., 2018), etc. But there has been very little work on utilizing synthetic data for pretraining LMs, this is mostly because synthetic data generated from open-ended generations poses many problems like hallucination (Maynez et al., 2020), ungrounded non-factual text (Thorne et al., 2018), etc. Instead, machine translation alleviates this problem by translating a source text in one language into another. However, translation errors exist and can reduce the quality of synthetic text. A naive approach would be to use round-trip-translation (RTT) BLEU scores to evaluate the quality of synthetic text but this requires twice the compute and RTT errors make these scores unreliable. Other machine translation evaluation approaches like BARTScore (Yuan et al., 2021), T5Score (Qin et al., 2023), MQM & COMET (Rei et al., 2020) either rely on large scale models to evaluate the quality of synthetic text or use large scale human annotated scores which makes it hard to scale for evaluating large amounts of synthetic text. Approaches like KenLM (Heafield, 2011) have been used to filter monolingual corpora based on perplexity.

2 Background

We limit ourselves to only text-based methods and do not extend to multi-modal synthetic data generation techniques. With this, we classify synthetic data generation methods in NLP into the following types:

1. **Weak labelers/augmentation**: A lot of labeled NLP tasks like classification-based (NER Tagging, POS tagging, sentiment classification, etc.) and retrieval-based (passage retrieval, LaBSE filtering, NLI) tasks require large-scale corpora for attaining strong performance as compared to humans. But most of these tasks do not have data available in the real world and occur as latent variables (E.g. POS Tags). Creating a sufficient amount of data for these tasks using human annotators is a big challenge. Instead, weak supervision can act as a proxy for generating large-scale noise data for these tasks. Methods like iterative fine-tuning, label mining, and similarity-based methods have served as a strong proxy for mining these labels and showing competitive performance. However, these methods are not applicable to many generative tasks because they rely on heuristics and provide less control.

2. **Referenceless Generation**: The concept of "large" has evolved rapidly over the past decades. In the early days of NLP, models with 10,000 to 100,000 parameters were considered large. Today, state-of-the-art models often have 100 billion parameters. This scaling of language models has led to the creation of some of the largest NLP models to date, such as GPT-3 (Brown et al., 2020b), with 175 billion parameters, and PaLM (Chowdhery et al., 2022), with approximately 540 billion parameters. This increase in size has significantly enhanced the performance of these models across a wide range of NLP tasks, including language modeling, question answering, and language translation. Referenceless generation methods rely on minimal or zero supervision to generate purely new content that has not been seen in the model's training data before. These LLMs can imitate human-like text very easily and can serve as a proxy for generative new text for a particular domain or task. Even with such strong text generation capabilities, these methods face issues from hallucinated, non-factual text and biased content. These are very open research problems for current LLMs. Recent study on TinyLMs that are trained on purely synthetically generated text from models like GPT3.5 and being as small as 10M parameters have been shown to produce fluent and consistent stories with almost perfect grammar (Eldan and Li, 2023) which means LMs even at a small scale have language understanding. Challenges like BabyLM (Warstadt et al., 2023) focus on improving LMs with a fixed data budget which enables exhaustive study of LM development methodologies, which can then be applied to larger LMs.

3. **Reference based filling**: Although reference-

less generation methods suffer due to lack of contextualized representations, reference-based filling mechanisms rely on the model's contextualized training abilities E.g. BERT (Devlin et al., 2018) to allow the model to generate reliable labels. These can also be extended to Seq-2-Seq models or Decoder-only models by using prompting. NLP prompt-based learning techniques aim to learn an LM that models the probability $P(x;\theta)$ of text x, enabling y prediction without large supervised datasets. For tasks that involve text generation or standard auto-regressive language models, prefix prompts are typically more effective as they align well with the model's left-to-right nature. Cloze prompts, on the other hand, are better suited for tasks that utilize masked language models as they closely match the form of the pre-training task. Full-text reconstruction models are more versatile and can use either cloze or prefix prompts. In tasks that require multiple inputs, such as text pair classification, prompt templates must accommodate space for two inputs. The most natural approach to creating prompts is to manually create intuitive templates based on human introspection. For instance, the LAMA dataset (Petroni et al., 2019) provides manually crafted cloze templates for probing knowledge in language models. GPT-3 (Brown et al., 2020b) uses manually crafted prefix prompts to handle a variety of tasks such as question answering, translation, and probing for common sense reasoning. Reference-based filling methods also include machine-generated text where a reference sentence X in the source language is used to generate a target sentence in Y language, these synthetically generated translations have shown improvements over many Machine Translation tasks e.g. using back translations (Sennrich et al., 2016a), (Edunov et al., 2018) to enhance Machine translation (Marie et al., 2020), (Bogoychev and Sennrich, 2019), (Ni et al., 2022), or for tasks such as native language identification (Goldin et al., 2018). However, there's limited exploration of using synthetic data for pretraining LMs due to issues like hallucination (Maynez et al., 2020), and ungrounded non-factual text (Thorne et al., 2018).

3 Coverage of Papers

Now we cover papers related to synthetic data in the following three settings.

3.1 Translationese Pretraining

"Translationese" is a term used to describe peculiarities in the text translated into a specific language, differentiating it from content originally written in that language (Gellerstam, 1986). Translated texts into the target language (via humans or machine-generated) often show distinctive features that differentiate them from their original counterparts in the target language. These disparities arise from either the influence of the translation process itself on the final product or the inherent "fingerprints" of the source language subtly present in the target language rendition (Rabinovich and Wintner, 2015). This is a common phenomenon in translation models where the target language translations often show characteristics of the source language and add bias to the evaluation of downstream tasks (Toral et al., 2018), (Zhang and Toral, 2019), (Graham et al., 2019). So far a lot of work on synthetic translated data has been done for using back translations (Sennrich et al., 2016a), (Edunov et al., 2018) for improving Machine translation performance (Marie et al., 2020),(Bogoychev and Sennrich, 2019),(Ni et al., 2022) or for classification tasks like native language identification (Goldin et al., 2018), etc. Tranlationese data has been used for many tasks but here highlights the efficacy of using translationese data for pretraining language models.

Existing approaches like Conneau et al. (2018) focus on transfer learning where a similar baseline is used called translate-train. In translate-train, a multilingual PLM (e.g., multilingual BERT) is fine-tuned using the original source language and machine-translated target language and then evaluated on the target language. This approach utilizes task-specific data translated into target language for fine-tuning, whereas our work focuses on pretraining rather than fine-tuning these language models and the effects synthetic text can have for pretraining and diverse downstream NLU and NLG tasks. Oh et al. (2022) also focus on leveraging translate-train and translate-test together for better cross-lingual fine-tuning.

Now we describe our framework (Doshi et al., 2024) for leveraging synthetic data for LM training. This process consists of collecting monolingual

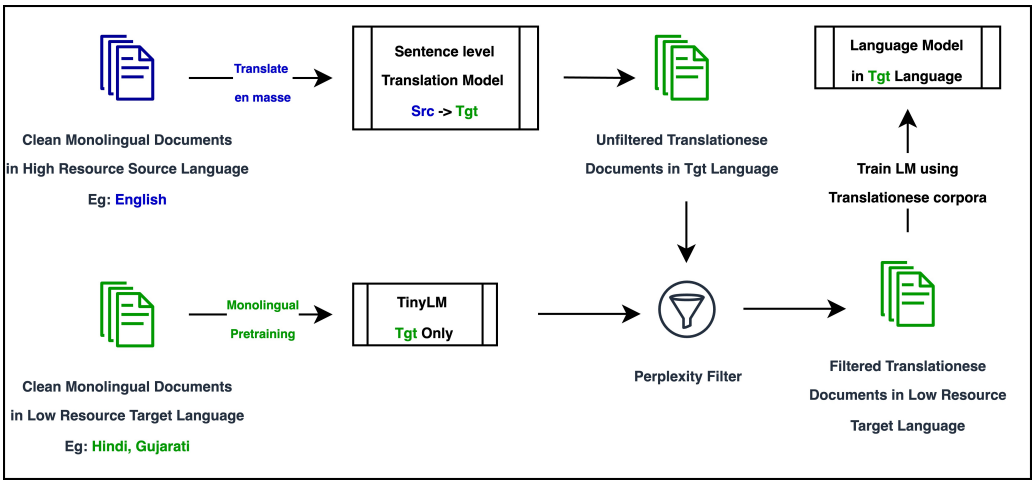

Figure 1: Overview of approach given by Doshi et al. (2024) to pre-train language models using translationese data.

(*clean*) data from the web for low-resource languages, training TinyLMs with it, translating *clean* data from a high resource language such as English into low-resource languages, using the aforementioned TinyLMs to filter *synthetic* data, and then using this filtered data to train LMs for downstream tasks. Our framework is illustrated in Figure 1. If the generated data is too noisy or lacks diversity, the potential for performance improvement may be limited (Epaliyana et al., 2021).

3.2 Synthetic Data in Multi-Source Machine Translation

Multi-source machine translation revolves around leveraging a source and a relatively high-resource pivot language jointly in a multi-source ensembling setup to improve translation into a low-resource target language.

Back-translation augmentation is a widely used data augmentation method in multilingual language models. This technique creates synthetic parallel training data from monolingual sources (Xu et al., 2022; Bi et al., 2021; Caswell et al., 2019; Liao et al., 2021; Marie et al., 2020; Pham et al., 2021; Sennrich et al., 2016b). For example, Sennrich et al. (2016b) back-translated monolingual target data into source language data, generating additional parallel training samples that significantly enhanced translation tasks. However, generating synthetic data through back-translation has limitations. The performance of the back-translation method significantly affects the quality and diversity of the synthetic data.

During the statistical MT era, pivot language MT was implemented using either a cascading or phrase table triangulation approach (Utiyama and Isahara, 2007). With the advent of NMT, various pivoting methods emerged that leveraged transfer learning(Zoph et al., 2016; Kim et al., 2019; Li et al., 2022). Furthermore, the rise of multilingual NMT enabled pivoting to be performed explicitly through cascading or implicitly via zero-shot translation (Dabre et al., 2020).

Zoph and Knight (2016) introduced the multi-source technique in NMT, which exploits multiple source languages to improve translation accuracy in the target language. They employed an encoder-decoder framework with multiple encoders, each dedicated to one source language, and combined the representations from these encoders for the decoder to generate the target sentence. (Firat et al., 2016) introduced an approach called late averaging with multiple encoder-decoder pairs each mapped to a source and the probabilities produced by all the decoders are averaged to produce the final probabilities for generating the next token. They also develop an approach called early averaging that centers on the concept of merging two distinct translation paths when calculating time-dependent context vectors within the decoder. For each time step in the decoder, dedicated context vectors are computed for each source language. These two context vectors are averaged and used as the final context vector. Additionally, Garmash and Monz (2016) proposed a multi-source ensembling method using a mixture of experts. They trained multiple models with different initializations and used them as experts to perform translation.

Nishimura et al. (2018b) leverage an incomplete multilingual corpus to enhance translation quality

using multi-source NMT. They use this incomplete multiway parallel data and incorporate a <NULL> token for missing source sentences during training, their approach effectively utilizes partially available multiway parallel data, resulting in significant improvements over conventional one-to-one NMT systems. Nishimura et al. (2018a) highlight the challenge of training multi-source Neural Machine Translation (NMT) systems due to the scarcity of n-way parallel corpora. To address this issue, they propose a data augmentation technique wherein they train two multi-source NMT systems and use them to create synthetic data and train systems again using this synthetic data. This iterative process continues until both systems' performance converges.

Libovický and Helcl (2017) investigate various attention mechanisms in the context of multi-source NMT. They present three strategies: serial, parallel, and hierarchical. In the serial strategy, encoder-decoder attention is computed sequentially for each input encoder. The query set for each cross-attention is derived from the preceding self-attention's context vectors. In the parallel combination strategy, each encoder is attended to independently, with the resulting context vectors summed up. All encoders are attended using the same set of queries from the self-attention sub-layer. The hierarchical combination involves computing attention independently for each input, treating the resulting contexts as states for another input, and then computing attention again over these states.

Huang et al. (2020) introduced a multi-stage training approach for multi-source NMT. In the first stage, the model is trained on monolingual corpora to learn the sequence generation task. Subsequently, the model is trained using parallel data for one-to-one translation tasks. Finally, fine-tuning is conducted for multi-sourcing.

More recently, Macháček et al. (2023) have used a multisourcing approach for the Automatic Speech Translation (AST) task. AST is susceptible to errors in speech recognition. However, since speech recognition systems in different languages may make different errors, they proposed that these diverse sources could complement each other in terms of the information they provide. Consequently, they developed a multi-source AST system, much more robust compared to the individual AST systems.

3.3 Synthetic Data in OGD Systems

This section covers research related to three topics: 1) Open Government Data (OGD), the current de facto system for data transparency in government; 2) Similar technology implementations from non-government domains; and 3) Utilizing synthetic data for OGD systems.

Attard et al. (2015) and Tang and Jiang (2021) provide a comprehensive overview of OGD terminology. Wibowo et al. (2023) focus on how OGD can be used as an effective tool for citizen engagement across various cities and use cases. Many OGD initiatives fail due to usability issues, as discussed by Hossain et al. (2021). Chakravarty (2018) analyze why OGD initiatives in India face challenges.

Mamalis et al. (2024) demonstrate how large language models (LLMs) like ChatGPT can retrieve relevant statistics from OGD systems in a zero-shot manner. Peña et al. (2023) explore the use of LLMs for topic classification in public affairs documents.

Doddapaneni et al. (2023) and Kakwani et al. (2020) have released training and evaluation sets, as well as pre-trained BERT (Devlin et al., 2019) models for Indian languages. Dabre et al. (2022) show that script unification for Indian languages can improve performance in low-resource languages. Haq et al. (2023) illustrate how machine-translated data can enhance retriever performance.

4 Challenges

Creating synthetic data for multilingual natural language processing (NLP) poses several challenges due to the complexities involved in accurately representing diverse languages. One significant issue is maintaining linguistic diversity across different languages. Each language has unique grammatical structures, idiomatic expressions, and cultural contexts, which can be difficult to replicate accurately. The creation of high-quality synthetic data necessitates an intricate understanding of these linguistic characteristics, but achieving this level of accuracy is challenging.

Additionally, balancing data quantity and quality presents a complex problem. While having a large dataset is beneficial, over-reliance on synthetic data may introduce biases or errors that are not present in naturally occurring text. This can compromise the integrity of the data and negatively impact the performance of NLP models. Ensuring the usefulness of synthetic data is crucial, as synthetic data

may lack certain elements imitating real-world data. This lack of authenticity can limit the generalizability and robustness of NLP models, making them less effective in real-world applications.

Moreover, the process of generating synthetic data at scale demands substantial computational resources and sophisticated algorithms. This process is resource-intensive and costly, posing a significant barrier to the widespread adoption of synthetic data in multilingual NLP, especially for pretraining. The need for advanced algorithms and significant computational power adds another layer of complexity to the already challenging task of creating synthetic data. As the field progresses, addressing these challenges will be crucial to harnessing the full potential of synthetic data in enhancing multilingual NLP.

5 Summary

In this survey, we explored how synthetic data enhances Natural Language Processing (NLP), focusing on its role in improving model training and addressing data challenges like scarcity, noise, privacy issues, and bias. Synthetic data helps create weak labels, generate new samples, and enhance machine learning models, especially for languages with limited resources. We analyzed various methods of synthetic data generation and filtering, including weak labelers and methods for creating new samples. These methods benefit tasks such as machine translation and language model pre-training. While synthetic data shows promise, it also brings challenges like maintaining linguistic diversity, balancing data quality and quantity, and managing computational demands, especially in multilingual applications. In conclusion, synthetic data offers significant potential for advancing NLP, particularly in multilingual settings. Future research should focus on refining algorithms to generate high-quality synthetic data efficiently, addressing these challenges to improve NLP model performance across languages and tasks.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418.

Wei Bi, Huayang Li, and Jiacheng Huang. 2021. Data augmentation for text generation without any augmented data.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation.

Rupak Chakravarty. 2018. *Open Government Data (OGD) Initiative in India: An Empirical Analysis*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*,

pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Do not worry if you do not have data: Building pretrained language models using translationese.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Koshiya Epaliyana, Surangika Ranathunga, and Sanath Jayasena. 2021. Improving back-translation with iterative filtering and data selection for sinhala-english nmt. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 438–443.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Brussels, Belgium. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Saiful Haq, Ashutosh Sharma, and Pushpak Bhattacharyya. 2023. Indicirsuite: Multilingual dataset and neural information models for indian languages.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mohammad Alamgir Hossain, Shams Rahman, Mohammed Quaddus, Elsie Hooi, and Abdus-Samad Temitope Olanrewaju. 2021. Factors affecting performance of open government data initiatives: A multi-method approach using sem and fsqca. *Journal of Organizational Computing and Electronic Commerce*, 31:300 – 319.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2022. ConsistTL: Modeling consistency in transfer learning for low-resource neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8383–8394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. 2021. Back-translation for large-scale multilingual machine translation.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dominik Macháček, Peter Polák, Ondřej Bojar, and Raj Dabre. 2023. Robustness of multi-source MT to transcription errors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3707–3723, Toronto, Canada. Association for Computational Linguistics.

Marios Evangelos Mamalis, Evangelos Kalampokis, Areti Karamanou, Petros Brimos, and Konstantinos Tarabanis. 2024. Can large language models revolutionalize open government data portals? a case of using chatgpt in statistics.gov.scot. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, PCI '23, page 53–59, New York, NY, USA. Association for Computing Machinery.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018a. Multi-source neural machine translation with data augmentation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 48–53, Brussels. International Conference on Spoken Language Translation.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018b. Multi-source neural machine translation with missing data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 92–99, Melbourne, Australia. Association for Computational Linguistics.

Jaehoon Oh, Jongwoo Ko, and Se-Young Yun. 2022. Synergy with translation artifacts for training and inference in multilingual tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6747–6754, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Íñigo Puente, Jorge Córdova, and Gonzalo Córdova. 2023. *Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs*, page 20–33. Springer Nature Switzerland.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. 2021. Meta back-translation.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2023. T5Score: Discriminative fine-tuning of generative evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15185–15202, Singapore. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Rong Tang and Jie Jiang. 2021. Characteristics of open government data (ogd) around the world: A country-based comparative meta-analysis. *Data and Information Management*, 5(1):11–26.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Will we run out of data? limits of llm scaling based on human-generated data.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Wahyu Setiawan Wibowo, Dana Indra Sensuse, Sofian Lusa, Prasetyo Adi Wibowo Putro, and Alivia Yulfitri. 2023. A systematic literature review on open government data: Challenges and mapped solutions. *Journal of Theoretical and Applied Information Technology*, 101(5):1806–1818. Funding Information: The authors would like to express their deepest gratitude to the Ministry of Communication and Information Technology (KOMINFO) for financial aid in carrying out this research. WSW would also offer sincere appreciation to KOMINFO for supporting him during his study at Universitas Indonesia. Publisher Copyright: © 2023 Little Lion Scientific.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra,

Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Jiahao Xu, Yubin Ruan, Wei Bi, Guoping Huang, Shuming Shi, Lihui Chen, and Lemao Liu. 2022. On synthetic data for back translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. *CoRR*, abs/1906.08069.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 30–34, San Diego, California. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.