DMOps: Data Management Operations and Recipes

Eujeong Choi [1 *] **Chanjun Park** [1 *]

Abstract

Data-centric AI has shed light on the significance of data within the machine learning (ML) pipeline. Recognizing its significance, academia, industry, and government departments have suggested various NLP data research initiatives. While the ability to utilize existing data is essential, the ability to build a dataset has become more critical than ever, especially in the industry. In consideration of this trend, we propose a "Data Management Operations and Recipes" to guide the industry in optimizing the building of datasets for NLP products. This paper presents the concept of DMOps which is derived from real-world experiences with NLP data management and aims to streamline data operations by offering a baseline.

1. Introduction

With the emergence of Data-centric AI (Polyzotis & Zaharia, 2021; Mazumder et al., 2022), various in-depth natural language processing (NLP) data research has been introduced in academia alongside the wide range of policies from industry and government departments (Pencheva et al., 2020).

In the case of academia, there are studies boosting model performance through large-scale datasets (Liu et al., 2021; Costa-jussà et al., 2022) along with the production of benchmark datasets for objective performance comparison between models (Wang et al., 2018; Ruder, 2021). Furthermore, there are also benchmark datasets that specialize in specific tasks (Rajpurkar et al., 2016; Alt et al., 2020). The government contributes to the field by implementing public data open policies and releasing datasets from the National Statistics department (Panagos et al., 2012).

However, the industry frequently dives into an untapped and specialized domain, where a ready-to-go dataset is rarely there. Especially for B2B companies, there is usually an urgent demand for data that fits their customers or business items (Pustejovsky & Stubbs, 2012). Since the open source and benchmark datasets are normally insufficient to meet these specific demands, additional data production is always a necessary step to initiate a particular task. As a result, the majority of AI businesses started to build their own task-specific datasets, alongside the emergence of companies that specialize in operating crowd workers to meet these demands. Additionally, research on efficient data production on human-in-the-loop started to make appearance (Doan, 2018; Wu et al., 2022).

Despite its necessity, there has been a paucity of studies in the field of NLP data production from an industry perspective. To the best of our knowledge, there has not yet been research that proposes the entire process starting from analyzing the business standpoint to data annotation and evaluation. Therefore, we propose a "Data Management Operations and Recipes (DMOps)" that will assist in building an NLP dataset efficiently and economically. Specially, we propose a DMOps that can produce high-quality NLP data needed in manufacturing deep learning models.

2. DMOps

Data management operations involve the integration of human input and decision-making into a data management pipeline or system. This involves a series of tasks such as data annotation, data quality assurance, and other activities that require a human touch (Doan, 2018; Goosen, 2019; Solis et al., 2019; Eo et al., 2021). One way to implement data management operations is through the use of recipes. Recipes are step-by-step instructions for performing a specific task or set of tasks and can be used to guide human workers through the NLP data management process.

Our Data Recipes consists of 12 steps. Through these steps, we go over the entire process of data operations : from establishing the goal of the project to delivering the final data to the modeling team. The name and explanation of each step is as follows.

1. Establish the Project Goal : Analyzing the purpose and requirements of data production is the first step of the recipe. This step requires collaboration with a team of NLP engineers and business operators. Through communication,

[*]Equal contribution [1]Upstage, Gyeonggi-do, Republic of Korea. Correspondence to: Chanjun Park <chanjun.park@upstage.ai>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

we can decide the input and output format of data that is suitable to the model of choice, and also set data milestones that fit the timeline of the business operation team.

Unlike academia where research starts from related works or enhancing existing benchmark datasets, the industry starts with its users and customer needs (Tarafdar et al., 2019; Kerzel, 2021; Alenezi et al., 2022). This gap between the two areas must be considered when setting the goal in the first place. To build a good dataset in the industry, we need to start from the end-user's needs and requirements (Laato et al., 2022).

2. Secure Raw Data : Researching and collecting raw data is the second step of the recipes. Five possible cases of collecting raw data are 1) the client providing the raw data (Ahmed, 2004), 2) using open-sourced public data (Lhoest et al., 2021), and 3) purchasing the raw data from its source platform, 4) crawl from website (Pant et al., 2004) 5) crowdsourcing (Estellés-Arolas & González-Ladrón-de Guevara, 2012; Hossain & Kauranen, 2015). The key issue here is the copyright of each data source. License information must be checked thoroughly, and getting a legal review is recommended before its usage (Khayyat & Bannister, 2015).

Furthermore, providing access and editing rights to the raw data must be dealt with caution. To prohibit unauthorized modifications from multiple parties, a sturdy data storage structure with limited access rights are necessary.

3. Data Pre-processing : The third step is improving the quality of the raw data through pre-processing. Basically the pre-processing consists of two main tasks: first, adjusting the format of data regarding its requirements, second, filtering non-ethical, privacy invading, and noisy data (Wiegand et al., 2018; Park et al., 2020). This step is all about practicing quality over quantity.

These preprocessing steps can be broadly divided into two tasks. The first task is to improve the quality of the data based on the inherent characteristics of the data (Rahm et al., 2000; Ridzuan & Zainon, 2019). A representative example of this is parallel corpus filtering in machine translation (Koehn et al., 2020), which is a data-centric method that improves the performance of the model without changing the structure of the model by only controlling the quality of the data (Park et al., 2022).

The second task is to address ethical issues with the data (Van Wel & Royakkers, 2004). This includes attaching data license information in advance or masking personal information if it exists. If these tasks are not clearly carried out in advance, there may be a risk of not being able to use the data even after data annotation and validation (Martin, 2020). This is why it is considered a highly important task.

4. Design a Data Schema : The fourth step involves designing an efficient annotation schema that captures all required information. This step is crucial, as it requires capturing the desired information fully while also ensuring efficiency to prevent cognitive overload for annotators. Also, figuring out parts that can be somewhat automated (pseudo-labeling) and parts that need human intervention (annotating) is essential in making the process efficient and moreover, accurate. With few pilot annotation iterations, the data scheme is expected to reach its optimal design (Gregor et al., 2020; He et al., 2022).

This step is one of the most important steps in designing data, such as "What kind of meaningful information to extract in Information extraction (IE)? (Hobbs & Riloff, 2010)", "What entity to tag in name entity recognition (NER) (Mansouri et al., 2008)", "In document summarization How much information should be compressed? (Yao et al., 2017)", "When constructing data for machine translation (MT), should paraphrase, literal translation, or transcendental translation be used? (Stahlberg, 2020)".

In academia, these elements are already pre-determined and research is conducted in that state, but in industry, this information must also be re-designed according to customer needs (Tseng et al., 2021).

5. Prepare a Guideline : The fifth step is the documentation of the data scheme. Its purpose is to deliver the designed annotation system to the expected annotators. The difficulty of the guideline should be monitored since the clarity and detailed explanation may be in a trade-off relationship.

6. Recruit Annotators : The sixth step is recruiting the annotators. The key is to select workers that are fit for the task for an efficient and accurate outcome. The best case would be selecting those who scored high on a test similar to the actual annotation task. Additionally, ethical considerations are also necessary. Good data is one that is created with fair compensation for the workers and without any unnecessary costs. It is important to take these factors into account as well (Foley et al., 2014).

7. Instruct Annotators : The seventh step is instructing the annotators with the guideline made above. In this stage, two-way communication that draws out questions and debates is the key whereas one-sided communication is discouraged.

8. Data Annotation : This step involves annotating the actual data, where annotators transfer their linguistic, cognitive, and visual intuition into the data. To ensure consistency among annotators, it is important to establish a unified approach and provide a safe environment for questions. Keeping a question log is also suggested to prevent inconsistencies within the corresponding answers.

In addition, in the field of NLP, the development of annotation tools is of paramount importance. This is due to several key factors, including the need for quality control, efficiency, and scalability (Grosman et al., 2020). In terms of quality control, an annotation tool allows for data annotation to be performed in a consistent and accurate manner. This is crucial for ensuring the quality of the data used for training NLP models. Additionally, an annotation tool can make the annotation process more efficient, which is especially important when building large datasets or for data that needs to be annotated quickly (Pei et al., 2022).

Scalability is also a crucial factor to consider when developing an annotation tool. As the amount of NLP data continues to grow, a tool that can handle the volume of data is essential. Furthermore, an annotation tool can be designed to assist human annotators, providing suggestions and feedback, thus increasing annotation speed and reducing human error (Perry, 2021).

To sum up, the development of a well-designed annotation tool is essential for NLP data management. It is a crucial step in training and evaluating NLP models and can greatly aid in preserving and analyzing linguistic data.

9. Data Internal Factor Verification : This ninth step is inspecting the annotated data. During this step, inspectors, who are usually selected from the annotator pool, must identify commonly occurring human errors and sort out the edge cases through discussions. Considering the nature of the Human-in-the-loop process, this step is essential to ensure the fidelity of the dataset.

In this stage, we recommend consensus labeling (Tang & Lease, 2011). Inter-annotator agreement (IAA) should be used to check the consistency of data labels (Ragheb & Dickinson, 2013; Boháč et al., 2017) due to the potential for human errors and misunderstandings of annotation guidelines. As human annotators are prone to inconsistencies and errors, it is necessary to verify the consistency of labels through IAA.

Certainly, data annotation can also be carried out in two steps. In the first round, rough annotation is performed on about 10 data, and IAA is checked in advance to correct the workers' misunderstandings. In the second round, using the corrected knowledge of the workers, the task of annotating all data is undertaken in a more thorough manner. In other words, by effectively combining the data annotation stage and the data inspection stage, the efficiency of the workers can be improved.

To sum up, data internal factor verification step is a process of validation for the inherent elements of the data. However, it does not verify external factors or the relationship between the data and the model (i.e. if the data truly helps improve the performance of the model). Therefore, additional steps such as ”data external factor verification” and ”data evaluation through model verification” should be considered to fully validate the data.

10. Data External Factor Verification : The tenth step is verifying the data external Factor. When inspecting data, it is necessary to first determine whether the work has been completed by observing the given guideline. Also, 1) data sufficiency, 2) data diversity, 3) data trustworthiness, 4) data privacy and security 5) data ethics suitability should be reviewed (Roh et al., 2019; Koo et al., 2022).

In other words, going beyond the internal information of the data, it is a step to thoroughly examine the sufficiency, diversity, reliability, security, and ethics of the data from various perspectives. The best way to conduct this verification is through Institutional Review Boards or external advisory committees (Blee & Currier, 2011).

11. Data Evaluation via Model Verification : The eleventh step is verifying the quality of data through actual modeling. In order to quantitatively verify whether the data is made as planned, various experiments are conducted such as checking data efficiency by increasing the amount of data or sectioning the data to check the consistency of its quality (Moon et al., 2021; Park et al., 2021). It is natural to find artifacts within one's data; after identifying the repeated errors, revisiting the recipes from step 5 is frequently required to enhance the quality of data. If there are parts that do not match our purpose while proceeding the steps, we should return to stage 5 and revise the guideline for another iteration.

To complement the 'human-in-the-loop' cycle, it's essential to detect errors through the model and clean them through human intervention. This cycle ensures error-free data and alignment with the model's outcomes, resulting in cohesive results. The goal is to create data that is both error-free and coherent with the model.

12. Data Deliverables : Final step of the recipes is delivering the final data outcome. In other words, it is the process of delivering annotated data to engineers or customers. When delivering, the versioning must be adapted to the protocol, and it is important to reveal the features of the data with its sample. Furthermore, after going through the exploratory data analysis (EDA) process, it is recommended to deliver the data analysis and the quality evaluation document together.

The quality of data in industry can be assessed based on several factors, including the informativeness of metadata, the legitimacy of data sources (e.g., compensated workers without unnecessary cost), well-established versioning systems, and intuitive and organized data storage structures. While these factors may seem obvious, they are crucial to creating high-quality data. In academia, these factors may not be given as much weight, but in industry, they are critical and together can elevate good data to the level of great data.

3. Discussion of DMOps

Why DMOps? Our proposed "DMOps" provides a universal and fixed process for data construction that can be applied to any NLP task or domain. This approach can serve as a baseline for data production, as it ensures a consistent and reliable process for generating high-quality data. The visualization of DMOps is shown in Figure 1.

Future of DMOps As DMOps evolve, automation will play a larger role in data production. This requires improving the efficiency and automation of current human-performed tasks. Additionally, synthetic data will become increasingly important, and methods for efficient inspection of this data through self-annotation by humans need to be developed. Managing data generated by large-scale language models like ChatGPT [1] and GPT-3 (Brown et al., 2020) efficiently is crucial in maintaining high data quality.

4. Conclusion and Future Works

In this paper, we presented DMOps, a task-agnostic methodology for efficiently producing high-quality NLP data with human annotation, which can serve as a baseline for any NLP data production. To increase the reliability of the proposed process, we plan to conduct quantitative verification at each stage of the process in the future. Additionally, we aim to conduct a study to investigate the difference in data quality when using the proposed data recipes compared to not using them.

Acknowledgements

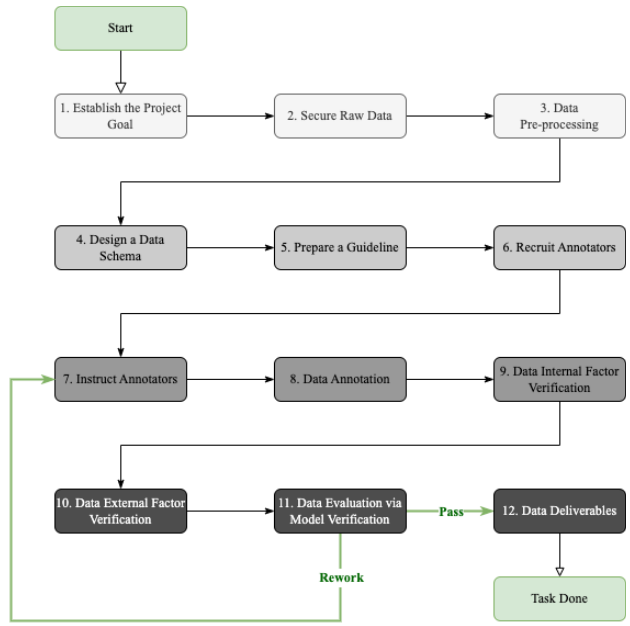

Figure 1. Image of the Data Management Operation Recipes

References

Ahmed, S. R. Applications of data mining in retail business. In *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, volume 2, pp. 455–459. IEEE, 2004.

Alenezi, M., Zarour, M., and Akour, M. Can artificial intelligence transform devops? *arXiv preprint arXiv:2206.00225*, 2022.

Alt, C., Gabryszak, A., and Hennig, L. Tacred revisited: A thorough evaluation of the tacred relation extraction task. *arXiv preprint arXiv:2004.14855*, 2020.

Blee, K. M. and Currier, A. Ethics beyond the irb: An introductory essay. *Qualitative Sociology*, 34:401–413, 2011.

Boháč, M., Rott, M., and Kovář, V. Text punctuation: an inter-annotator agreement study. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pp. 120–128. Springer, 2017.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Doan, A. Human-in-the-loop data analysis: a personal perspective. In *Proceedings of the workshop on human-in-the-loop data analytics*, pp. 1–6, 2018.

Eo, S., Park, C., Moon, H., Seo, J., and Lim, H. Comparative analysis of current approaches to quality estimation for neural machine translation. *Applied Sciences*, 11(14): 6584, 2021.

Estellés-Arolas, E. and González-Ladrón-de Guevara, F. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[1] https://chat.openai.com/

Foley, M., Ruser, J., Shor, G., Shuford, H., and Sygnatur, E. Contingent workers: Workers' compensation data analysis strategies and limitations. *American journal of industrial medicine*, 57(7):764–775, 2014.

Goosen, A. E. *A system to quantify industrial data quality*. PhD thesis, North-West University (South Africa), 2019.

Gregor, S., Chandra Kruse, L., Seidel, S., et al. Research perspectives: the anatomy of a design principle. Association for Information Systems, 2020.

Grosman, J. S., Furtado, P. H., Rodrigues, A. M., Schardong, G. G., Barbosa, S. D., and Lopes, H. C. Eras: Improving the quality control in the annotation process for natural language processing tasks. *Information Systems*, 93: 101553, 2020.

He, X., Nassar, I., Kiros, J., Haffari, G., and Norouzi, M. Generate, annotate, and learn: Nlp with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842, 2022.

Hobbs, J. R. and Riloff, E. Information extraction. *Handbook of natural language processing*, 15:16, 2010.

Hossain, M. and Kauranen, I. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):2–22, 2015.

Kerzel, U. Enterprise ai canvas integrating artificial intelligence into business. *Applied Artificial Intelligence*, 35 (1):1–12, 2021.

Khayyat, M. and Bannister, F. Open data licensing: More than meets the eye. *Information Polity*, 20(4):231–252, 2015.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 726–742, 2020.

Koo, S., Park, C., Seo, J., Lee, S., Moon, H., Lee, J., and Lim, H. K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria. *IEEE Access*, 10:118167–118175, 2022.

Laato, S., Tiainen, M., Islam, A. N., and Mäntymäki, M. How to explain ai systems to end users: a systematic literature review and research agenda. *Internet Research*, 32(7):1–31, 2022.

Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

Liu, J., Fang, Y., Zhu, D., Ma, N., Pan, J., and Meng, M. Q.-H. A large-scale dataset for benchmarking elevator button segmentation and character recognition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14018–14024. IEEE, 2021.

Mansouri, A., Affendey, L. S., and Mamat, A. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344, 2008.

Martin, K. E. Ethical issues in the big data industry. In *Strategic Information Management*, pp. 450–471. Routledge, 2020.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Rojas, W. G., Diamos, S., Diamos, G., He, L., Kiela, D., Jurado, D., et al. Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062*, 2022.

Moon, H., Park, C., Eo, S., Park, J., and Lim, H. Filter-mbart based neural machine translation using parallel corpus filtering. *Journal of the Korea Convergence Society*, 12 (5):1–7, 2021.

Panagos, P., Van Liedekerke, M., Jones, A., and Montanarella, L. European soil data centre: Response to european policy support and public data requirements. *Land use policy*, 29(2):329–338, 2012.

Pant, G., Srinivasan, P., and Menczer, F. Crawling the web. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, 153, 2004.

Park, C., Lee, Y., Lee, C., and Lim, H. Quality, not quantity?: Effect of parallel corpus quantity and quality on neural machine translation. In *Annual Conference on Human and Language Technology*, pp. 363–368. Human and Language Technology, 2020.

Park, C., Lee, S., Moon, H., Eo, S., Seo, J., and Lim, H. How should human translation coexist with nmt? efficient tool for building high quality parallel corpus. *arXiv preprint arXiv:2111.00191*, 2021.

Park, C., Seo, J., Lee, S., Lee, C., and Lim, H. The asr post-processor performance challenges of backtranscription (bts): Data-centric and model-centric approaches. *Mathematics*, 10(19):3618, 2022.

Pei, J., Ananthasubramaniam, A., Wang, X., Zhou, N., Sargent, J., Dedeloudis, A., and Jurgens, D. Potato: The portable text annotation tool. *arXiv preprint arXiv:2212.08620*, 2022.

Pencheva, I., Esteve, M., and Mikhaylov, S. J. Big data and ai–a transformational shift for government: So, what next for research? *Public Policy and Administration*, 35(1): 24–44, 2020.

Perry, T. Lighttag: Text annotation platform. *arXiv preprint arXiv:2109.02320*, 2021.

Polyzotis, N. and Zaharia, M. What can data-centric ai learn from data and ml engineering? *arXiv preprint arXiv:2112.06439*, 2021.

Pustejovsky, J. and Stubbs, A. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.", 2012.

Ragheb, M. and Dickinson, M. Inter-annotator agreement for dependency annotation of learner language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 169–179, 2013.

Rahm, E., Do, H. H., et al. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Ridzuan, F. and Zainon, W. M. N. W. A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738, 2019.

Roh, Y., Heo, G., and Whang, S. E. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019.

Ruder, S. Challenges and opportunities in nlp benchmarking, 2021.

Solis, R., Pakbin, A., Akbari, A., Mortazavi, B. J., and Jafari, R. A human-centered wearable sensing platform with intelligent automated data annotation capabilities. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, pp. 255–260, 2019.

Stahlberg, F. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

Tang, W. and Lease, M. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*, pp. 1–6, 2011.

Tarafdar, M., Beath, C. M., and Ross, J. W. Using ai to enhance business operations. *MIT Sloan Management Review*, 60(4), 2019.

Tseng, M.-L., Tran, T. P. T., Ha, H. M., Bui, T.-D., and Lim, M. K. Sustainable industrial and operation engineering trends and challenges toward industry 4.0: A data driven analysis. *Journal of Industrial and Production Engineering*, 38(8):581–598, 2021.

Van Wel, L. and Royakkers, L. Ethical issues in web data mining. *Ethics and Information Technology*, 6(2):129–140, 2004.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Wiegand, M., Siegel, M., and Ruppenhofer, J. Overview of the germeval 2018 shared task on the identification of offensive language. 2018.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 2022.

Yao, J.-g., Wan, X., and Xiao, J. Recent advances in document summarization. *Knowledge and Information Systems*, 53:297–336, 2017.