Human-in-the-Loop Approaches For Task Guidance In Manufacturing Settings

Ramesh Manuvinakurike Intel Labs
Santiago Miret Intel Labs
Richard Beckwith Intel Labs
Saurav Sahay Intel Labs

Giuseppe Raffa Intel Labs

Abstract

We introduce a task guidance framework for manufacturing settings aiming to improve the well-being and productivity of manufacturing workers completing a given task. The assistive technology proposed in this work centers on a dialogue system built upon semantic frame extraction of process specifications detailing a given manufacturing process. The dialogue system interacts with the technician performing the task by capturing their actions and assisting them in performing relevant steps. Specifically, we develop components to parse expert-authored natural language documents called specs and utilize the parse for task guidance and continual learning. While still in the early stages, we believe that an interactive, assistive AI framework similar to the one we are exploring will become an important component of high-volume manufacturing in the future.

1 Introduction

Manufacturing processes are a critical part of bringing automated materials design to a large scale. This is especially important for automated chemical synthesis and automated material characterization, both of which are necessary to enhance the quality and effectiveness in high-volume manufacturing. While we believe that current and future AI technologies have great potential to automate many of today's manual tasks, we also think that humans are going to continue to be a part of complex manufacturing operations. In this work, we introduce a human-in-the-loop framework to engage with technicians with the aim of enhancing the experience and productivity of human workers involved in manufacturing processes. We specifically study two tasks that are common in manufacturing processes, i.e., i) Understand process specification documentations (also called specs) & ii) Task guidance for technicians.

In today's factory settings, many workers are guided by a class of documents known as process specifications (specs), which consist of procedures and steps to be executed to achieve a consistent manufacturing/maintenance end goal. These documents consist of human interpretable steps (spec-items), often accompanied by images indicating the desired end goal or visual demonstrations. These documents (specs) contain information about the actions, objects to work on, tools to use, the desired end goal, directional & temporal attributes, and the purpose of the action, among others. We thus focus on **Spec document understanding** for the first part of our work. We first develop a data annotation scheme to represent the information in the spec documents via 'semantic frames' and baseline models to extract this information from the spec documents in manufacturing/maintenance use cases. The semantic frames representation Fillmore et al. [2006] used in this work assumes a natural language sentence consisting of an action and corresponding entities that help accomplish

Accepted at the NeurIPS 2022 AI for Accelerated Materials Design Workshop.

the action or indicate temporal aspects of the same (e.g., [Action: Turn] [Receiver: the nut] using [Tool: a screw driver] [Extent: until snug]). These semantic frames are often observable visually while the technicians execute a process. We develop a BERT-based semantic frame extraction (SFE) model, the goal of which is to extract these semantic frames and corresponding information that can be utilized for task guidance and developing other vision-based models. These semantic frames can then be used for task guidance (e.g., question-answering, helping with tools, informing the extent of action, etc.) by including these as a part of dialogue system architecture.

Task guidance is a framework by which an assistive technology (e.g., dialogue system) provides specific instructions to a user in performing a given task. In our case, we focus on workers in a manufacturing setting who would handle different tools and materials to produce or maintain certain products. Developing such an automated task guidance system is challenging since the execution of a process can have variability. Such systems also need continual learning capabilities since manufacturing processes are dynamic and evolve rapidly. Hence, a task guidance system also needs to have continual learning capabilities. As a first step towards developing such a system, we focus on developing a human remote-controlled (Wizard-of-Oz) agent that helps collect process execution data. This data can be used to train the underlying models that form integral parts of a task guidance system. We collect the video data and natural language narrations using the paradigm, which we refer to as 'guided think-aloud' (Please see A.1 as to why guided think-aloud is used in this work) where the technician narrates the process as they execute while the agent asks questions to extract the information for incorrect, uncertain, or missing model predictions. Such a mode of data collection additionally could help overcome the problem of time & cost-intensive data annotation steps. It also helps capture variations in execution of a spec and any change in the process over time thus making it comparatively scalable.

We envision a dialogue system that helps users perform tasks successfully by providing assistance via question-answering, process adherence nudges, and anomaly detection. The development of such a dialogue system is complex, requiring multimodal understanding, dialogue management, and generation capabilities. The goal for the dialogue system in this work is to develop an initial AI-enabled conversational assistant to gather information from the technicians to improve the underlying models, which ultimately serves the goal of enhancing technician welfare and productivity when executing a given task. The main contributions of this work are: i) Development of semantic frame extraction models for spec document understanding. ii) Human-AI collaborative interface for conversational data collection to help development of conversation task guidance.

2 Method

The flow for the agent is shown in Figure 1. The scenario involves a conversation between a wizard-controlled dialogue system (agent) and a technician executing the process in real time. For the data collection (for continual learning), the goal for the agent is to elicit narrations for the currently executed step. In order to facilitate the narrations, the wizard selects the spec item that is being executed (refer to Figure 3 see the wizard interface). The spec document understanding module parses the spec item selected by the wizard and loads the semantic frame. The wizard utilizes this information to elicit narrations from the technician or provide task guidance (e.g., inform tool). In this section, we first describe the semantic frames extraction (SFE) from the spec item. We then show how we leverage these semantic frames for the 'guided think-aloud' to help the technicians elicit better narrations for the processes. These narrations from humans are also used for model pre-training and show the improvements in SFE.

Spec document understanding A spec document for a manufacturing process consists of items (spec items) that typically describe the execution of a single step in a complex process. Each spec item in this work is assumed to consist of a single executable action [1] along with additional semantic information. A semantic frame is a dictionary consisting of slot-value as key-value pairs (e.g., Action, Receiver (object receiving the action), Tool (an object used to execute the action on the receiver), location (location of the action), etc. Please refer to Table 2 for all the slots and the number of tokens for each slot in the dataset). The values except the 'Action' are optional. These semantic annotations are performed at the token (words) level and then converted to B-I-O scheme (Please see Figure 2 for

[1]This is not always the case. A single sentence can consist of complex actions consisting of multiple observable actions. e.g., 'Clean' may mean wiping & rinsing.

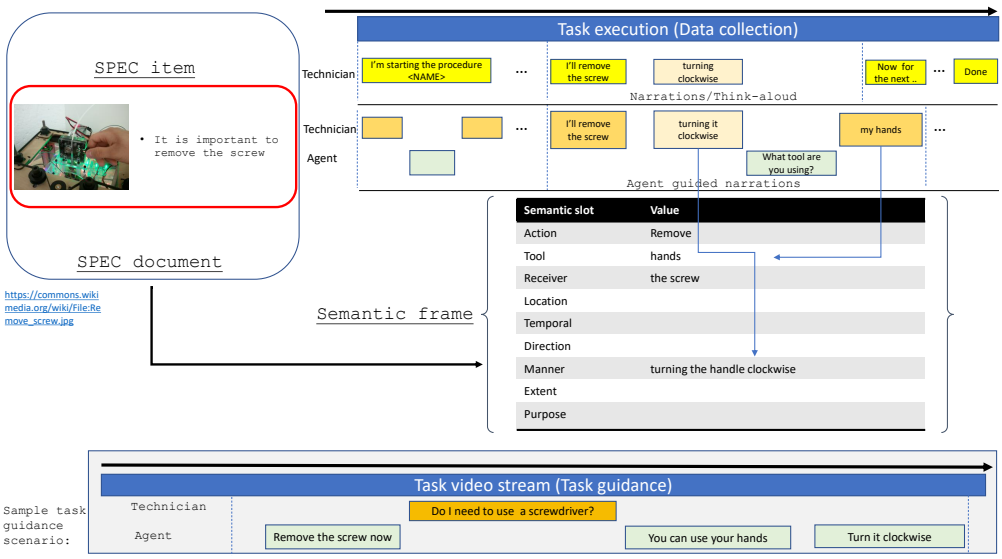

Figure 1: Figure shows hypothetical conversation about a step in the process.

an example) which allows us model the problem as using token classification proble to extract the semantic frames from an input spec item.

We annotated the semantic frames from a manufacturing spec (referred to as technical domain) consisting of 841 tokens (285 unique tokens) spanning 6 processes and 56 spec-items. Due to the limited availability of real technical specs used in manufacturing, we only utilize this data as the test set. Since limited data availability is an issue and recent BERT-based (Devlin et al. [2018]) models rely on large amounts of data either for pre-training or fine-tuning, we identify 'related domains' that bear a resemblance to (referred to as related domains) the manufacturing process. Such data is available more freely (e.g., cooking & daily tasks) and in abundance compared to the manufacturing specs. This data section (related domains) consists of 12963 tokens (1803 unique) spanning 7 domains. Table 2 shows a more detailed view of the data. We collected this data and annotated them with the same annotation scheme to explore if the models learned using this data can be utilized for frame extraction in the technical domain. We also augment the data in the related domains using data augmentation techniques, i) Synonym replacement, ii) Paraphrasing and iii) Back-translation approaches. During the augmentation process, we ensure that the number of tokens in the augmented input sentence and the Part-of-speech is the same as that in the original sentence. This allows us to train the models using the augmented sentences as the input and original annotations as the labels.

We train the models using BERT using only the annotated samples and with augmentation. We use the 'bert-base-uncased' model from the Huggingface repository (Wolf et al. [2019]). The original spec document also consists of additional textual instructions not related to the manufacturing process (guidelines & supplementary explanation). We utilize these sentences for pre-training the models using TSDAE (Wang et al. [2021]) and then finetune the models using only the real and with augmentation. We run the model with a learning rate of 3e-5 for 3 epochs with the rest of the parameters in the default configurations. We use MSE loss and Adam optimizer for training.

Dialogue system (Wizard-of-Oz) Our goal for the spoken dialogue task guidance agent is to demonstrate the applicability of the dialogue in the manufacturing setting. The agent assists in collecting the user narrations/descriptions of the processes for learning the models required to completely automate it. We believe that this approach of collecting the narrations is more scalable and time-efficient compared to time intensive annotation efforts. (Such narrations are referred to as weak labels in the literature in recent years, have shown appreciable results in learning robust embeddings (Miech et al. [2019])) Nudging the technicians to narrate during process execution is not straightforward since the technicians are not used to narrating during process execution and can

Model	Train data	Test data	P / R / F1 / Acc
DistilBERT	Real	Technical	0.27 / 0.30 / 0.29 / 0.44
BERT	Real	Technical	0.38 / 0.43 / 0.40 / 0.47
BERT + TSDAE (Spec)	Real	Technical	0.43 / 0.50 / 0.46 / 0.50
BERT + TSDAE (Transc)	Real	Technical	0.43 / 0.51 / 0.46 / 0.52
BERT + TSDAE (Transc+Spec)	Real	Technical	**0.51 / 0.59 / 0.55 / 0.54**
DistilBERT	Aug+Real	Technical	0.51 / 0.61 / 0.55 / 0.64
BERT	Aug+Real	Technical	0.49 / 0.57 / 0.52 / 0.64
BERT + TSDAE (Spec)	Aug+Real	Technical	**0.52** / 0.63 / **0.57** / **0.65**
BERT + TSDAE (Transc)	Aug+Real	Technical	**0.52** / 0.63 / **0.57** / 0.62
BERT + TSDAE (Transc+Spec)	Aug+Real	Technical	0.51 / **0.64** / **0.57** / 0.62

Table 1: Shows the results of semantic frame extractor models on technical spec documents evaluated using token level Precision, Recall, F1 and Accuracy. We can observe that the data augmentation accompanied by unsupervised domain adaptation increases the P,R,F1 and Accuracy.

be interruptive. The agent used in this work is a mixed-initiative turn taking agent that nudges the technicians to narrate sparingly to avoid being interruptive.

We now describe the wizard system that is used to collect the narrations. The wizard selects the process that is being executed, which loads the spec items from the spec document. As the technician executes the process, if they're not narrating, the wizard chooses one of the open-ended action directives to make the users narrate the current step (e.g., Can you please tell me what you're doing right now? what is that? etc.). The wizard can also choose the spec item and generate questions specific to a slot. For example, For a spec item "wipe the residue", the wizard can click on the text which loads the questions relevant to each slot in the semantic frame (E.g., for Tool, the question will be "what is being used for wiping"). These questions can either be input by the developers or generated by an automated question generation model. We use the questions written by the developers in this work. This shows another avenue to integrate the question generation NLP models. To prevent the agent from interrupting by asking too many questions, we limited the number/type of questions the agent asks. We collect the user narrations transcriptions generated automatically by an off-the-shelf speech-to-text system.

3 Results & Future Work

Table 1 shows the results of the SFE models. We found that the models utilizing the TSDAE (Wang et al. [2021]) approach outperform the models that don't use the TSDAE pre-training regimen. We found that the augmentations of the dataset helped the models achieve better precision, recall, F1, and accuracy, highlighting the problem of sparse data availability. We collected 49 interactions from the technicians on the factory floor spanning 12.3 hours across 6 processes. We show that the think-aloud data collection paradigm shows promise in collecting narrations that can aid better model building (e.g., learning better models for SFE). This demonstrates that weak labels could not only help develop better models for vision training but also helps learn better NLP models via domain adaptation. While still in the early stages, we intend to explore utilizing this data for learning joint embeddings (Vision & language). This paradigm shows promise in learning better models for AI-based systems in manufacturing. We thus highlight three main contributions of this work, i) domain adaptation yields better SFE outputs, ii) Using data from completely different domains with our frame annotation scheme still yields promising results for SFE, iii) dialogue system for manufacturing applications (data collection & task guidance).

As a next step towards the development of an automated system in manufacturing, we're integrating visual action recognition systems. Such a vision-based system will be integral for a system as it can deliver information about the actions being performed by the technicians without relying completely on the narrations. The SFE outputs can also be used to further improve the visual action recognition models. While such a system is in the early stages of development, we see promising results. Towards automating the dialogue system without the wizard, we're developing dialogue system components (Intent/Dialogue act recognizers, Dialogue manager, Dialogue state trackers) for task guidance scenarios. The goal for the development of such a system is to help identify anomalies and deviations from process execution that can help improve overall productivity during the manufacturing process. Such a system could increase the quality and scalability of next-generation materials and materials-

related products, such as batteries or fuel cells, produced by automated materials synthesis techniques and analyzed by automated materials characterization techniques.

References

André da Silva Barbosa, Felipe Pinheiro Silva, Lucas Rafael dos Santos Crestani, and Rodrigo Bueno Otto. Virtual assistant to real time training on industrial environment. In *Transdisciplinary Engineering Methods for Social Innovation of Industry 4.0*, pages 33–42. IOS Press, 2018.

Abhijeet Sandeep Bhardwaj, Akash Deep, Dharmaraj Veeramani, and Shiyu Zhou. A custom word embedding model for clustering of maintenance records. *IEEE Transactions on Industrial Informatics*, 18(2):816–826, 2021.

Michael P Brundage, Thurston Sexton, Melinda Hodkiewicz, Alden Dima, and Sarah Lukens. Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, 27: 42–46, 2021.

Mario Casillo, Francesco Colace, Loretta Fabbri, Marco Lombardi, Alessandra Romano, and Domenico Santaniello. Chatbot in industry 4.0: An approach for training new employees. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 371–376. IEEE, 2020.

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

Alden Dima, Sarah Lukens, Melinda Hodkiewicz, Thurston Sexton, and Michael P Brundage. Adapting natural language processing for technical text. *Applied AI Letters*, 2(3):e33, 2021.

Charles J Fillmore et al. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400, 2006.

Yiyang Gao, Caitlin Woods, Wei Liu, Tim French, and Melinda Hodkiewicz. Pipeline for machine reading of unstructured maintenance work order records. In *Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference (ESREL)*, 2020.

Yunyi Jia, Lanbo She, Yu Cheng, Jiatong Bao, Joyce Y Chai, and Ning Xi. Program robots manufacturing tasks by natural language instructions. In *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 633–638. IEEE, 2016.

Aman Kumar and Binil Starly. "fabner": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, pages 1–15, 2021.

Aman Kumar, Akshay G Bharadwaj, Binil Starly, and Collin Lynch. Fabkg: A knowledge graph of manufacturing science domain utilizing structured and unconventional unstructured knowledge source. *arXiv preprint arXiv:2206.10318*, 2022.

Chen Li, Andreas Kornmaaler Hansen, Dimitrios Chrysostomou, Simon Bøgh, and Ole Madsen. Bringing a natural language-enabled virtual assistant to industrial mobile robots for learning, training and assistance of manufacturing tasks. In *2022 IEEE/SICE International Symposium on System Integration (SII)*, pages 238–243. IEEE, 2022.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

Syed Meesam Raza Naqvi, Christophe Varnier, Jean-Marc Nicod, Noureddine Zerhouni, and Mohammad Ghufran. Leveraging free-form text in maintenance logs through bert transfer learning. In *International Conference on Deep Learning, Artificial Intelligence and Robotics*, pages 63–75. Springer, 2022.

Thurston Sexton, Melinda Hodkiewicz, Michael P Brundage, and Thomas Smoker. Benchmarking for keyword extraction methodologies in maintenance work orders. In *Proceedings of the Annual Conference of the PHM Society*, volume 10, 2018.

Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.59. URL https://aclanthology.org/2021.findings-emnlp.59.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A Appendix

A.1 Experimentation with think-aloud

We conducted numerous experiments with different variations to collect the think-aloud data before settling on 'guided think-aloud' which contains an agent in the loop. We found that the users (technicians) do not interact with the system even if there is an intention to help the developers of the system. This could be because the users after a few minutes tend to forget to narrate since the task being conducted is not cognitively easy. We also found that absence of a listener and lack of feedback from the listener (agent) was not sufficient. It remains to be seen if a completely automated agent continues to engage the users to provide relevant narrations.

Figure 2: Figure shows the applicability of the semantic frame representations for technical language processing in different domains.

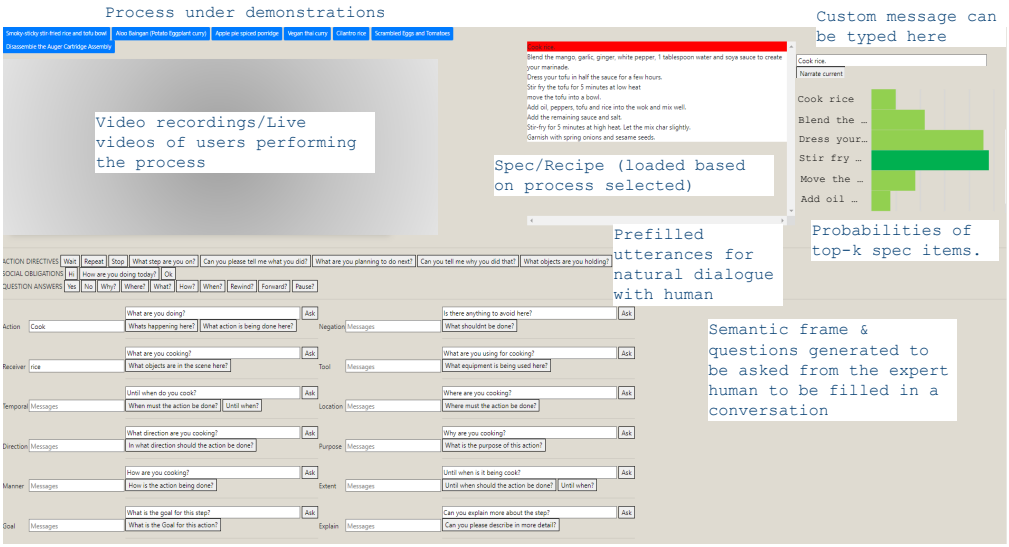

Figure 3: Wizard interface used for collecting the data.

Spec statistics	All combined	Technical	Related domains
# spec items	940	56	884
# tokens	13804	841	12963
# augmented items	36453	-	36453
# augmented tokens	465005	-	465005
Semantic slot - Annotated	**# tokens**	**%**	**%**
Actions	1228	9.86	7.97
Tool	483	6.08	3.29
Receiver	2614	22.48	18.56
Location	1286	9.75	9.24
Temporal	157	0.12	1.19
Direction	30	0.46	0.19
Manner	488	0.0	3.75
Extent	1432	8.03	10.48
Purpose	421	3.8	2.98
Other slots			
Negation	16	0.11	0.10
Explain	1594	3.67	12.03
# Think-aloud stats (in progress)			
# Sessions	49	49	-
# Recording duration	12.3 hours	12.3	-
# Total transcription tokens	25519	255519	-
# Unique think-aloud tokens	1867	1867	-

Table 2: Shows the statistics of the data and the semantic frames.

B Related work

Understanding the language from the technical specs or maintenance logs is also referred to as Technical language processing Brundage et al. [2021], Dima et al. [2021] where some of the challenges specific to the domain are highlighted. For instance, referring terms of the objects (receivers or tools) are idiosyncratic and the tokenizers are unable to tokenize these words. The handwritten notes for instance contains typographical errors, abbreviations, incomplete entries, inconsistent referring expressions and even erroneous making it difficult to develop models or leverage out of the box models. Recent works have developed custom embedding models to generate embeddings for technical language representation and show promise Bhardwaj et al. [2021]. Recent works have also leveraged Named entity recognitions (NER) Kumar and Starly [2021], Kumar et al. [2022] models showing adoption of deep learning approaches for entity extraction for technical language understanding. Another approach of understanding such technical language is by classifying the maintenance work orders into problems categories Naqvi et al. [2022] which can then be interpreted by human technicians.

Representation of such technical language has been explored by annotation of data as Item-Activity-State (e.g., [Activity: Repair] [State: cracked] [Item: hand rail] near right hand [Item: mirror]) Gao et al. [2020] and also as item/problem-action/solution-action ([Solution-action: Replace] [Problem-action:missing] [Item: tip]) Sexton et al. [2018]. In this work, we extend the representations for manufacturing processes or technical language by leveraging semantic frames. It's important to keep in mind that these representations are for maintenance work orders (MWO) where the language is more terse compared to the spec documents.

Data annotation for manufacturing processes remains a challenge. In this work, we develop an approach for a 'learning' dialogue system which interacts with users to learn about the process being executed. The process is human-in-the-loop approach where expert human technicians not only provide further insights on the process but also clarify the mistakes committed by the system. Leveraging human experts in manufacturing has garnered interest recently Jia et al. [2016], Dima et al. [2021]. Human expertise has been shown to be valuable in labeling the data for entities extraction in technical language processingDima et al. [2021]. In Jia et al., the authors utilize human feedback to instruct the robots to perform tasks in manufacturing setting. The instructions are pre-programmed since they're dependent on the capabilities of the robot performing the task. In our approach, we

allow experts to converse freely which could be used to help guide the development of capabilities in the robot and/or to utilize this information to train new technicians via dialogue systems. Dialogue systems have also been developed in manufacturing setting Barbosa et al. [2018], Casillo et al. [2020], Li et al. [2022]. Such virtual assistants have been utilized to assist humans by teaching them about the task. These systems leverage an expert provided knowledge graphs to assist the human technicians by answering the questions asked. In our work, we're interested in the developing a learning dialogue system that helps the system learn a knowledge base consisting of task model.