Abstract

Development of task guidance systems for aiding humans in a situated task remains a challenging problem. The role of search (information retrieval) and conversational systems for task guidance has immense potential to help the task performers achieve various goals. However, there are several technical challenges that need to be addressed to deliver such conversational systems, where common supervised approaches fail to deliver the expected results in terms of overall performance, user experience and adaptation to realistic conditions. In this preliminary work we first highlight some of the challenges involved during the development of such systems. We then provide an overview of existing datasets available and highlight their limitations. We finally develop a model-in-the-loop wizard-of-oz based data collection tool and perform a pilot experiment.

1 Introduction

In this work, we present the development of a conversational task guidance system. The goal of this work is to discuss the limitations of the current datasets and methods to fill the gap by developing a new dataset and method - based on human-in-the-loop approach - that helps address some of the issues. Such a system can find application in manufacturing, daily and do-it-yourself tasks (cooking, changing a flat tire, jump start a car, etc.), system maintenance, etc. Such systems can inform users about the next action to perform, tools to use, answer questions, provide insights about the current state, debug an issue, preemptively interject when an error occurs. Conversely, such a system can continually learn from the users to improve its performance and adapt to changing conditions (continual learning). Development of such systems is complex, including models for, but not limited to, knowledge retrieval, action recognition and segmentation, multi-modal dialogue understanding, dialogue management, language generation, common-sense reasoning, co-reference resolution, and grounding. Data availability remains a challenge for the development of such systems, although numerous datasets exist to study these questions in isolation. While several datasets have been developed focusing on problems mentioned, a unified dataset for conversational task guidance is not yet available. Such a dataset should include, i) Visual demonstrations: for visual understanding, ii) Conversations relevant for task guidance: for training dialogue modules, iii) Ground-truth (document) for reference/retrieval, iv) Ontology of task structure: for reasoning and personalization and, v) Annotations for developing supervised models. We will now define the components of such a conversational task guidance system.

Task: The conversations in a task guidance scenario is a dialogue between the task “performer” and the “guide”. The system acts as a task ‘guide’ that helps the task performer realize the goal of a complex situated task. The performer acts as the “human-in-the-loop” of the system (Amershi et al., 2014), who helps the dialogue system components to be trained, fine-tuned and adapted based on the dialogue.

Spec, Spec Item & actions: To complete the task, the performer is expected to follow a ground-truth document called the ‘spec’ (e.g., recipe) which provides details about the task. The spec consists ordered steps easily interpretable by human called ‘spec items’. Each step or a ‘spec item’ typically consists of a set of ‘atomic actions’ (referred as actions henceforth) that is needed to successfully complete the spec item. Note, these atomic actions are sometimes interchangeable or even optional (Figure 1). The guide (in this case the dialogue system) needs to recognize the actions performed by the performer in an online manner and simultaneously identify the spec item corresponding to the action. Thus the system is required to identify
actions being performed by the performer at two levels of abstraction. This is an important aspect as the guide can then recommend the next spec item or action to the performer. The process of identification of actions corresponding to a spec item bears resemblance to commonsense reasoning (Forbes et al., 2019; Losing et al., 2021).

Semantic frames: For representing a step in a process (spec item or actions), we use semantic frames which forms the understanding module. Table 1 shows the slot and values in the semantic frame and the definitions. The semantic frames bear resemblance to the slot-value representation for dialogue state (Williams et al., 2014). However, in this task each spec item consists of multiple actions and are represented as semantic frames, each step (spec item) in the execution can be thought of a semantic frame consisting of multiple sub semantic frame (action). Such a hierarchy bears resemblance to the task of frame (El Asri et al., 2017) representation of a state tracking task. However, the distinguishing factor is the hierarchical representation of a process spec item as a series of optional action items whose sequence is probabilistic and that the semantic frames are grounded in visual demonstrations (videos). Such a representation of the task enables development of a multi-modal dialogue system for task guidance.

The main contribution of this work is, i) Introducing a novel representation for a multimodal dialogue system development for task guidance, ii) data consisting of multiple sections for training the models discussed above, iii) A novel model-in-the-loop Wizard-of-oz framework for task guidance data collection.

2 Related work

Datasets: Several datasets have been developed in recent years focused on the task of task guidance. We discuss a small subset of the relevant datasets. Activitynet (Fabian Caba Heilbron and Niebles, 2015) focused on developing a dataset to identify activities and the related hierarchies in general web videos. Youcook2 dataset (Zhou et al., 2018) consists of cooking videos from youtube consisting of annotated instructions. Annotation of the dataset is resource intensive. Various benchmarks have been developed on datasets with focus on unaligned videos and annotations (typically narrations or weak labels). Howto100m (Miech et al., 2019) dataset consists of more than 134k hours of videos and corresponding narrations collected from youtube intended to learn a joint embedding representation of video and text. COIN (Tang et al., 2019) dataset consists of >475 hours of instructional contents from youtube.

Grounding the videos in a spec document has also garnered interest. Videos for video classification (MedvidCL) and question answering (MedvidQA) was introduced recently (Gupta et al., 2022). MedvidCL task involves classifying the video into medical instructional, non-instructional and others. MedvidQA involves answering questions by highlighting relevant section of the videos that answers the question. The authors in this work leverage wikihow (spec) during the process of data curation. Similarly, Colas et al. (2020) have developed a question-answering dataset to highlight the video segment from tutorial videos that answer the question.

One of the major challenges with the videos typically crawled from the web is that they’re edited with the purpose of conveying the message as

| Action: Action term |
| Tool: Object used for action |
| Location: where action is performed |
| Temporal: temporal aspects of the action |
| Direction: direction of the action |
| Manner: how the action is performed |
| Extent: until when the action is performed |
| Purpose: Purpose of the action |
| Note: Additional non-actionable notes |

Table 1: Semantic frame used for understanding modules.
Figure 2: Figure shows a sample data from our dataset. The spec document consists of instructional images and natural language instruction for carrying out the process to completion. Each process has accompanying videos with narrations.

quickly, rather than to demonstrate the real-time execution of a particular task. This is important in the task guidance scenario, since the temporal aspects of the actions are useful for the temporal aspects of actions which in-turn is useful for the planning problems. Dataset that capture the real-time egocentric have also been developed. EPIC-Kitchens (Damen et al., 2018, 2022) consists of videos and narrations/annotations where users with a head-mounted gopro camera perform various tasks (e.g., cooking, cleaning) in a kitchen. EGO4D (Grauman et al., 2021) is a recent dataset consisting of users performing large number of daily activities. Recent assembly 101 (Sener et al.) consists of videos performing assembling tasks with hierarchy of tasks also represented.

**Semantic frames:** Conversational task guidance systems require fine-grained multi-modal understanding abilities such that they can provide assistance relevant to the user. For understanding we use semantic frames representation (Traum, 2003; DeVault et al., 2011). Our definition of semantic frames for task guidance scenarios is motivated by the semantic role labeling from propbank (Kingsbury and Palmer, 2002) annotation schemes. The semantic roles of a given sentence are verb centric, where each verb in a given sentence is labeled with arguments such as patient, agent, benefactive etc. The task of identifying such semantic roles from the visual context (image or video) has been studied (Gupta and Malik, 2015; Yang et al., 2016; Sadhu et al., 2021). In our work, we define a semantic frame where the arguments are centered around actions of interest, rather than verb/auxiliary as present in the propbank annotation scheme (Bonial et al., 2012).

**Retrieval:** The process of retrieval in a task guidance domain, specifically cooking domain has been studied as a problem involving retrieving images from recipes/ingredients or as a problem of generating recipes from the images (Salvador et al., 2017, 2019). In our work, one of the focus is to highlight the spec items incrementally as the demonstrations occur in real-time. The spec is authored by experts, and this can be considered a ground truth document that needs to be aligned to the demonstrations. Closest to our work is that of Multimodal-aligned-recipes (Lin et al., 2020) where the authors construct a dataset of youtube videos where the segments from the videos are automatically aligned to recipes and recipe items. This is however not incremental, which is a requisite for our work. For building the models for spec retrieval, we explore recent contrastive learning models which have given promising results (Chen et al., 2020; He et al., 2020; Radford et al., 2021;
Xu et al., 2021; Jia et al., 2021).

Conversational task guidance systems: Task guidance system development has been of interest for sometime. Such systems have been developed to help navigate through operation of daily household equipment (Rich and Sidner, 2007), medical domains helping conduct surgery (Escobar-Castillejos et al., 2020; Tillou et al., 2016), cooking for differently-abled (Ramil Brick et al., 2021) etc. Given the complex nature of the recent works have focused on a tractable set of problems such as question-answering (Anantha et al., 2020; Ramil Brick et al., 2021; Feng et al., 2021). It is important to note the distinction between task guidance systems discussed in this work with systems leveraging human guidance (Zhang et al., 2021) for learning. We believe such methods to learn from human demonstrations as an important characteristic of the task guidance systems.

Data collection: The development of such systems is challenging and requires large amounts of annotated data. Following Damen et al. (2018); Grauman et al. (2021) we follow the notion of developing the data with weakly aligned human narrations as labels for visual data. For building understanding models for semantic frames extraction, given the promising aspect of controlled crowd-sourcing, we’re exploring the development of a corpus leveraging crowd-sourcing paradigm (He et al., 2015; Roit et al., 2020). we believe Question-answering SRL paradigm of leveraging crowd-workers to answer the questions given a context to fill semantic frames as a promising direction (Roit et al., 2020).

3 Data

The task guidance we are interested in this work involves humans performing a task where they’re expected to follow a spec document. It is important to note that the spec document is authored by a task expert for the human task performer and such documents are typically available during task performance. Such document assumes prerequisites (e.g., in recipes to understand ‘mince garlic’ the human performer needs to know how to do this) that might not all be interpretable by novice humans or machines. Hence, we develop a task breakdown structure where items in the spec document is broken down into atomic actions which a novice can interpret (it is important to note that it is possible that a task performer cannot still understand these). These atomic actions can also be represented as semantic frames.

Think-aloud/narrations have gained popularity (for example Damen et al. (2018); Miech et al. (2019); Tang et al. (2019)) since they can be leveraged as noisy weak labels overcoming the annotation phase which can be expensive. Such noisy labels have yielded promising results in various vision based tasks (Radford et al., 2021; Jia et al., 2021; Xu et al., 2021; Wang et al., 2021) outperforming models trained explicitly on limited annotated samples. The main advantage of such an approach is that the narrations can be provided by novice annotators who don’t require additional training. However, noisy labels due to inconsistent vocabulary usage and misalignment between the visual frames and narrations can be a drawback. These drawbacks can be overcome potentially by carefully designing self-supervision pre-text tasks to learn a joint representation (Xu et al., 2021; Wang et al., 2021). One of the ideas we are currently exploring that has yielding promising preliminary results is the ‘conversational system guided think-aloud/narrations’. We collect these conversational system guided narrations data collection either ‘post-hoc’ or ‘during’ task demonstration. We leverage wizard-of-oz approach to collect the narrations data for think-aloud/narrations data post-hoc and during task demonstrations.

In the ‘post-hoc’ system, we assume that there exists large amounts of videos without any narrations. To collect the weak labels, we make the users (typically task experts) watch the pre-recorded videos post-hoc and narrate them. When the users are not sure what to narrate or narrate incompletely (for e.g., don’t narrate the action or objects) the conversational system assists them by generating questions specific to slots in the semantic frames. This is challenging since additional time associated with question asking could further exacerbate the problem of alignment. To help mitigate this issue, we need capability for the dialogue system to have video manipulation abilities (play, pause, rewind, forward, loop, zoom) and collect narrations from the users specific to the video frames. Although this approach doesn’t solve the problem of alignment, it could potentially mitigate some of the issues. We are also interested in developing the dialogue system for collecting the narrations ‘during’ the task performance. This phase is important as it can accelerate the system development by decreasing the
time involved in collecting the narrations post-hoc. However, there are additional challenges involved in deploying such systems. The models need to operate online and incrementally. The goal for the system to help collect the weak labels during task demonstration is to, i) prod the performers to provide narration, ii) if information is missing in the narrations (slots in semantic frames) ask questions to fill the incomplete slots, iii) highlight the relevant spec item for the task demonstrator to collect feedback about the errors in retrieval model predictions.

4 Wizard system

The wizarding system was developed to collect spoken conversational data between the human narrator (post-hoc)/task performer (during) and the agent. The wizard interface and the user interface for both the ‘post-hoc’ and ‘during’ task data collection is similar in design (Figure 4). The only difference being that the video section is either recorded or live streamed. The wizard interface contains pre-filled utterances that can be played back to the user and is designed to aid the goal of collecting more relevant narrations. The wizard also has access to the predictions from semantic frame extractor that can assist in asking relevant questions to aid better narrations from the user. The semantic frame extractor is a BERT based (Devlin et al., 2018) model. The questions are also pre-generated using BART based sequence to sequence model (Lewis et al., 2020). The wizard can either choose to use these questions as is, edit them or choose to enter a new question. The atomic-actions and an image representing each step are provided manually for every item in the spec document by the expert.

The wizard in the ‘post-hoc’ phase controls the video playback and asks the narrator to describe the video by selecting relevant buttons on their screen. The wizard’s selection is played back to the user using TTS (Text to speech). The user’s narration is relayed to the wizard screen live during the interaction using webRTC. The conversations are recorded on the server with timestamps. We’ll now briefly discuss the models developed for semantic frame extraction, action recognition, spec item retrieval and question generation.

5 Models

The models have access to video frames from the beginning of the execution until current n-th frame. $F = \{F_0, F_1, \ldots, F_n\}$. The spoken narrations by the users are available as transcription chunks $T = \{T_0, T_1, \ldots, T_p\}$. These transcription chunks are extracted from the ASR (Automatic speech recognizer). Each task consists of m spec items $S = \{S_0, S_1, \ldots, S_m\}$. Each spec item correspond to a system state. It is ensured that each spec item corresponds to a single semantic frame. These semantic frames may be incomplete.

5.1 Action Segmentation and Recognition:

The objective of the action segmentation and recognition model is to recognize actions being performed at any frame $F_t$ at the $t^{th}$ time in the online manner. Each action has long horizontal temporal dependency. For example, an action ‘wash rice’ is temporally dependent on ‘get rice’ action as shown in Fig. 3. Thus, we use MSTCN (Farha and Gall, 2019; Li et al., 2020) that recognize and segment actions based on sequence of frame features. A frame feature $F_t$ at time $t$, in our model, is captured using 3DCNN such as Slowfast (Feichtenhofer et al., 2019) over a sliding window of 64 frames. A standard MSTCN requires the entire video to perform action segmentation and recognition, which is a key restriction in our online case. Thus, we modify the individual temporal convolutions of MSTCN to causal temporal convolutions (Van Den Oord et al., 2016) that only intakes all the frame features up to time $t$ to predict the action. Due to the use of the causal convolutions, the model is suitably modified
for the online nature of the action recognition.

5.2 Spec item matching

The problem of spec item matching involves utilizing the current video frames, $F$ and/or the narrations, $T$ (if available) and estimate the current spec item $S_i$ that is ‘currently’ being performed at a cadence as selected by the user. This is important to the task performer as they can use this information to plan the next set of actions. For baseline we use the models in a zero shot manner, i.e without fine-tuning the models.

**Using transcriptions:** We use sentence transformers as text encoders to get the text embeddings from the transcriptions $T$ and spec items $S$. We use sentence BERT$^1$ (Reimers and Gurevych, 2019) to get the sentence embedding. The semantic similarity between the narrations and the spec items are calculated using cosine similarity metric.

**Using video:** We use CLIP$^2$ (Radford et al., 2021) for the baseline model. CLIP model during inference takes as input an image ($F_j$) and text candidates ($S$) and ranks the candidates based on cosine similarity between the text embeddings and vision embeddings generated by the text and vision encoders of the CLIP model. We input each frame $F_j$ and get the similarity scores for each $S_i$. The $S_i$ with the highest aggregate similarity score for last ‘n’ frames is selected as the best matching candidate. We find that this approach yields promising first results. There are several limitations with this approach, i) We find that image encoders from CLIP yield better results than using video encoders. However, use of video encoders self-supervised training could potentially yield better results. We leave this to the next steps. ii) Since, we aggregate the results (irrespective of image or video encoders) there will be a lag in displaying the current spec item. A segmentation model could potentially help resolve this issue. We leave this to the future work and point to the importance of such segmentation models.

5.3 Semantic frame extraction

The semantic frame extractor (SFE) module is a token classifier module which extracts the semantic frame from the input text ($S$ or $T$). The SFE module is adapted from the SRL (Semantic Role Labeling) BERT model (Shi and Lin, 2019) trained on Ontonotes dataset (Pradhan et al., 2013). We use this SRL BERT and stack a linear layer mapping the original labels into the semantic tags definitions shown in Table 1 in this work. An expert provides a map from the original propbank

---

$^1$https://sentence-transformers.all-MiniLM-L6-v2
$^2$https://huggingface.co/openai/clip-vit-base-patch16
labels into the new semantic frames (Example: $B - Verb \rightarrow B - Action, I - ARG1 \rightarrow I - Tool$ etc.). These mapping can then be used to initialize the weights of the newly stacked linear layer with 1 if the mapping is true, 0 otherwise. One of the earlier works applying this approach of rule integration was performed by mapping the input rules using the bag of words model (Hatzilygeroudis and Prentzas, 2004). We apply similar approach at the newly stacked linear layers. While this new layer is initialized with the human provided mapping, they still can be further finetuned with new data.

Currently, the SFE is stateless, meaning it is not cognizant of the previously extracted incomplete frame which might be a useful to distinguish the semantic entities (e.g., Tool vs Receiver when the user has uttered only a referent). We believe this is another interesting area for future work.

5.4 Question generation

The question generator is a trained sequence to sequence model that is trained using the BART model architecture (Lewis et al., 2020). The model converts the input semantic frame entities and a prompt (slot for which question needs to be generated) into a question. We leverage QA-SRL dataset for (Klein et al., 2020) for training our initial question generation model for filling a semantic slot. We preprocess the data to fit our needs in this work.

Table 2 shows the model performance of spec item retrieval using the CLIP and Sentence BERT models every 1 second of the video. When utilizing the frame features $F$, we experiment with including a window size of ‘n’, where ‘n’ from the from the past is used to visual features extraction. We observe that the CLIP model outperforms the sentence BERT model. This can be attributed to the misalignment between the narrations and the video demonstrations. The spec item was annotated in the video using the visual demonstrations and not the narrations (e.g., person in the video demonstrations says ‘I’ll be using jack’ with the video focused on the person’s face). The person then proceeds to get the jack without narrating anything. The annotations are corresponding the visual demonstrations and not the narrations. The narrations from the users are spoken (semantic similarity model is trained on written text) and thus contains various disfluencies. We also see that increasing the window size (i.e increasing amount of visual context) helps until certain point (n=6s). We further investigated individual tasks (performing CPR, changing tire, jump starting a car, making coffee using moka pot, repotting a plant) from Instructional videos dataset (Alayrac et al., 2016) and found that the performance varies across tasks. This is primarily because each task is unique and the duration of each step (corresponding to a spec item) varies depending on the nature of the task (e.g., chopping onion corresponds to different lengths depending the amount of onions being used in a recipe). It
Table 2: Shows the spec retrieval performance (ROUGE scores and Accuracy) on videos 50 videos Instructional videos dataset (Alayrac et al., 2016). The videos were annotated by spec documents for the videos obtained from wikihow. The videos belonged to 5 types of tasks (performing CPR, changing tire, jump starting a car, making coffee using moka pot, repotting a plant).

|          | R1   | R2    | RL    | Acc  |
|----------|------|-------|-------|------|
| SBERT    | 39.37| 16.324| 26.89 | 19.14|
| CLIP (n = 1s) | 43.51 | 26.28 | 33.14 | 31.6 |
| CLIP (n = 2s) | 53.94 | 34.79 | 42.26 | 41.68|
| CLIP (n = 4s) | 53.84 | 36.51 | 44.16 | 44.35|
| CLIP (n = 6s) | 55.05 | 33.38 | 41.45 | 42.25|
| CLIP (n = 8s) | 53.05 | 33.38 | 41.45 | 42.25|

is thus important to include process information when modeling the spec retrieval for a given task. As a next step we plan to learn a joint vision and text encoder. When the narrations/transcriptions are absent it is not possible to run the time window analysis using the text.

**Semantic frame extraction** We find that on a small test set created from instruction videos spec data (50 sentences) we find the average precision=0.39 and recall=0.29. It is important to note that, although the tags for the semantic frames were motivated from the propbank labeling scheme there are numerous differences (for instance, SRL frames are created for every predicate (verb & AUX) while the semantic frames in our corresponds to an action performed).

We finally collected 10 interactions (5 for post-hoc think-aloud/narrations, 2 during think-aloud narrations, 3 task guidance) interaction between the wizard and the user for the task of ‘cooking’. The total audio collected spans 3 hours. The users in all the scenarios successfully completed the task. We’re currently in the process of collecting and analyzing the dataset.

**Task guidance dialogue system** A task guidance dialogue system interacts with the task performer and assists them in performing a given task. A task guidance dialogue system is comprised of followed features, i) Multi-modality: The system monitors the actions and objects present in the environment. This contains action recognition and object recognition tasks. ii) Mixed initiative: The system can interject and provide corrections to the task performer if it predicts that the task is not being carried out satisfactorily. The system can also ask questions if it is uncertain about the task execution. iii) Document grounded question-answering: The system needs to be able to answer questions about the objects, actions, point the users to the next steps to be performed etc. The answers provided by the agent must be grounded in the documents/specs which provide ‘ideal’ execution protocol of a given process. iv) Execution flexibility: While, the spec document provides the ideal execution of a given process, it is by no means the only way in which a task can be completed successfully. For instance, in a cooking recipes dataset if the spec document mentions chop the onions and then boil pasta, there is no impact on the outcome if the task performer chooses to put pasta for boiling before chopping the onion. However, the order of ingredients used for sauteing could impact the outcome. The system needs to account for the flexibility in task execution.

References

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In Computer Vision and Pattern Recognition (CVPR).

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. Ai Magazine, 35(4):105–120.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Sriniivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. arXiv preprint arXiv:2010.04898.

Claire Bonial, Jena Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. 2012. English propbank annotation guidelines. Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder, 48.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Dooho Kim. 2020. Tutorialqa: Question answering dataset for tutorial videos. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 5450–5455.
Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The epic-kitchens dataset. In European Conference on Computer Vision (ECCV).

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision, 130(1):33–55.

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialogue & Discourse, 2(1):143–170.

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.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A video corpus for adding memory to goal-oriented dialogue systems. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue.

David Escobar-Castillejos, Julieta Noguez, Fernando Bello, Luis Neri, Alejandra J Magana, and Bedrich Benes. 2020. A review of training and guidance systems in medical surgery. Applied Sciences, 10(17):5752.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 961–970.

Yazan Abu Farha and Juergen Gall. 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In 2019 ieee. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3570–3579.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6202–6211.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6162–6176.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? arXiv preprint arXiv:1908.02899.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. arXiv preprint arXiv:2110.07058, 3.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2022. A dataset for medical instructional video classification and question answering. arXiv preprint arXiv:2201.12888.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. arXiv preprint arXiv:1505.04474.

Ioannis Hatzilygeroudis and Jim Prentzas. 2004. Integrating (rules, neural networks) and cases for knowledge representation and reasoning in expert systems. Expert Systems with Applications, 27(1):63–75.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 643–653.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916. PMLR.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In LREC, pages 1989–1993. Citeseer.

Ayal Klein, Jonathan Mamou, Valentina Pyatin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. Qanom: Question-answer driven srl for nominalizations. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3069–3083.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880.

Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. 2020. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepa Dey, and William B Dolan. 2020. A recipe for creating multimodal aligned datasets for sequential tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4871–4884.

Viktor Losing, Lydia Fischer, and Jörg Deimöller. 2021. Extraction of common-sense relations from procedural task instructions using bert. In Proceedings of the 11th Global Wordnet Conference, pages 81–90.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. 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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 143–152.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Elisa Ramil Brick, Vanesa Caballero Alonso, Conor O’Brien, Sheron Tong, Emilie Tavernier, Amit Parekh, Angus Addlessee, and Oliver Lemon. 2021. Am i allergic to this? assisting sight impaired people in the kitchen. In Proceedings of the 2021 International Conference on Multimodal Interaction, pages 92–102.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Charles Rich and Candace I. Sidner. 2007. Diamond-help: A generic collaborative task guidance system. AI Magazine, 28(2):33–33.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality qa-srl annotation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7008–7013.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5589–5600.

Amaia Salvador, Michal Drozdzal, Xavier Giró-i Nieto, and Adriana Romero. 2019. Inverse cooking: Recipe generation from food images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10453–10462.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3020–3028.

F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. CVPR 2022.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. arXiv preprint arXiv:1904.05255.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1207–1216.

Xavier Tillou, Sylvie Collon, Sandrine Martin-Francois, and Arnaud Doerfler. 2016. Robotic surgery simulator: elements to build a training program. Journal of surgical education, 73(5):870–878.

David Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In Proceedings of the International Workshop on Computational Semantics, pages 380–394.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. volume 125, page 2.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472.

Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. AI Magazine, 35(4):121–124.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Chai. 2016. Grounded semantic role labeling. In Proceedings of...
Ruohan Zhang, Faraz Torabi, Garrett Warnell, and Peter Stone. 2021. Recent advances in leveraging human guidance for sequential decision-making tasks. *Autonomous Agents and Multi-Agent Systems*, 35(2):1–39.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.