Understanding Procedures

CMSC 491/691 – INTERACTIVE FICTION AND TEXT GENERATION

Slides adapted from Li “Harry” Zhang and Qing “Veronica” Lyu

Learning Objectives

See how procedures are used in modern NLP research

- Including datasets, evaluation

Determine how procedures can be used in interactive fiction

What are procedures?

- A procedure is “a series of **actions** conducted in a certain **order** or manner,” as defined by Oxford
- A more refined definition: “a series of **steps** happening to achieve some **goal**[1]”
  - Why?
- Examples of procedures: instructions (recipes, manuals, navigation info, how-to guide), algorithm, scientific processes, etc.
  - We focus on **instructions**, which is human-centered and task-oriented
- Examples of non-procedures: news articles, novels, descriptions, etc.
  - Those are often narrative: events do not have a specific goal
  - The umbrella term is **script**[2]

[1] (Momouchi, 1980) [2] (Schank, 1977)

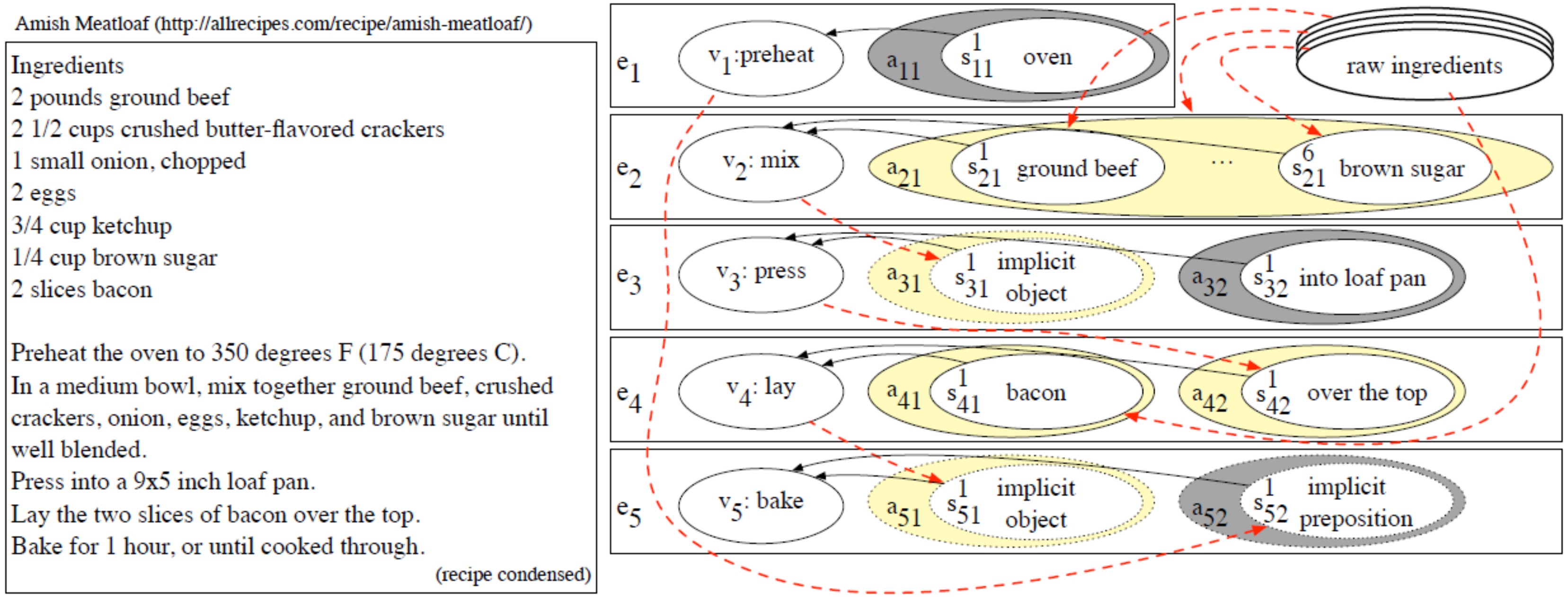
Why study procedures?

- Procedures are useful
  - Core to task-oriented systems (e.g., dialog agent)
  - Can be used for scaffolding in interactive fiction generation
  - Can be used to understand and analyze narrative texts (how?)
- Procedural events are a suitable scope to study **machine reasoning**
  - Procedural texts are more structured and less complex
  - Good starting point to study planning and reasoning

How to study procedures

- **Extract** procedural knowledge from instructions[1]
  - Usually a structured representation
- A subset of these works specifically focus on **recipes**[2]

Why would research on procedures focus on recipes?



[1] (Paris et al., 2002; Zhang et al., 2012; Maeta et al., 2015)
[2] (Kiddon et al., 2015; Bisk et al., 2019)

How to study procedures

- **Reason** about procedural events and serve **downstream** tasks
  - Early text classification (Takechi et al., 2003)
  - Answering how-to questions (Delpech and Saint-Dizier, 2008; our work)
  - Tracking entity states (Dalvi et al., 2018; Gupta and Durett, 2019; Tandon et al., 2020; our work)
  - Event relation reasoning (our work)
  - Intent classification (our work)
  - Application in task-oriented dialogs (our work)

Reasoning about Goal-Step Relations

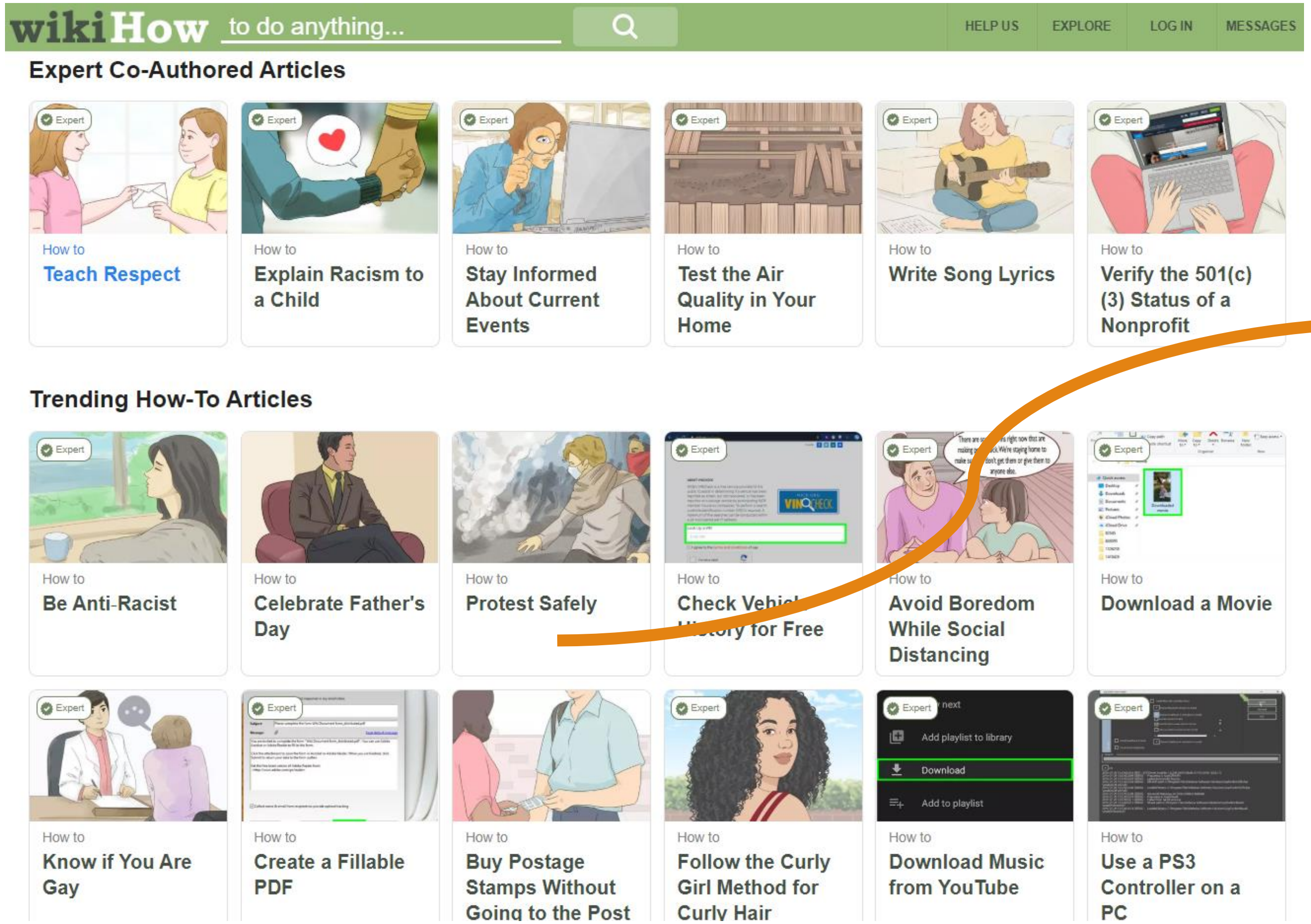
(ZHANG ET AL., 2020): REASONING ABOUT GOALS, STEPS, AND TEMPORAL ORDERING WITH WIKIHOW

Reasoning about Goal-Step Relations

- Procedure is “a series of **steps** happening to achieve some **goal**”
- First and foremost, a model should understand what steps and goals are
- Examples:
  - (Infer goal) If I preheat the oven to 350 F, what am I trying to do?
  - (Infer step) If I want to get a good GPA, should I study or play PS5?
  - (Infer order) To drive a car, do I first shift to drive or foot down the gas?
- Where can we get lots of procedural data?

(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

wikiHow.com



(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

Multiple-choice format

- Task #1 Goal Inference: Given a goal, choose the most likely step out of 4 candidates.
  - **Input**: “How to prevent coronavirus”
  - **Choices**: Wash your hands? Wash your cat? Clap your hands? Eat your protein?
- Task #2 Step Inference: Given a step, choose the most likely goal out of 4 candidates.
  - **Input**: “Blink repeatedly.”
  - **Choices**: Handle Pepper Spray in Your Eyes? Relax Your Eyes? Draw Eyes? Diagnose Pink Eye?
- Task #3 Step Ordering: Given a goal and two unordered steps, determine which comes first.
  - **Input**:

Goal: How to Act After Getting Arrested.

Step (a): Get a lawyer.   Step (b): Request bond from the judge

(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

Negative Sampling

- Finding the correct answer is easy; finding some wrong answers is hard
- Semantic similarity-based approach
  - BERT sentence embedding similarity based on KNN
  - Consider full sentence, only verbs, only nouns, etc.
- Example
  - In an article “Protest safely”, a step is “Wear sensible clothes.”
  - Find the sentence embedding of “Wear sensible clothes”, and all other steps
  - Find 3 other steps whose sentence embeddings have the highest cosine similarity with the sentence embedding of “Wear sensible clothes”
  - Result: “Wear sensible clothes,” “buy some clothes”, “put on costumes”, “try old clothing.”

(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

Statistical Cues

- Modern models like BERT can cheat!
- Sometimes models can tell the answer without looking at the question
- Zhang et al. resolve this by randomly reassigning the correct candidate
  - Once they get all 4 candidates, any of them could be the correct one
  - Any model is guaranteed to perform no better than chance given only the candidates

(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

Crowdsourcing validation

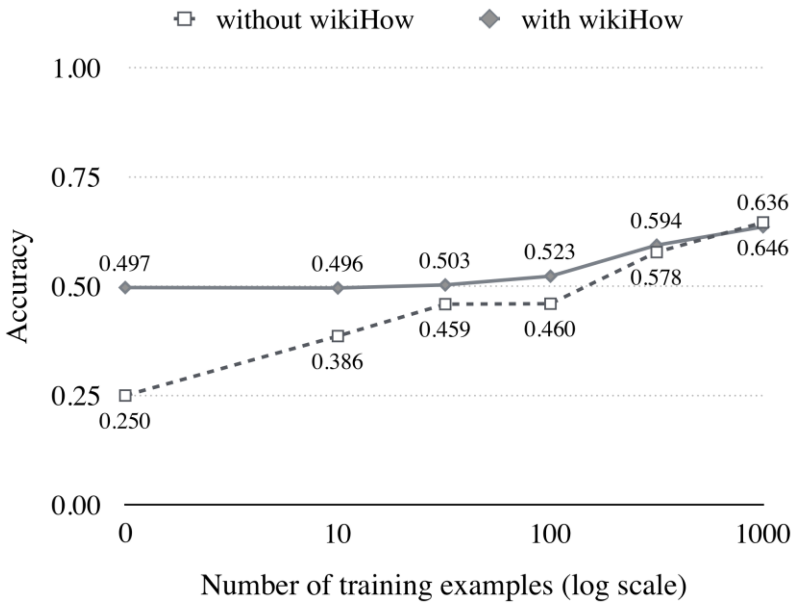
- We now have a large number of automatically generated examples
- These examples might be noisy, but we still want a clean test set
- For each task, randomly sample thousands of examples to pose to crowd workers
- If all of 3 crowd workers answer correctly based on the gold label, keep the example in the test set
- Use the rest as training and validation sets

(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

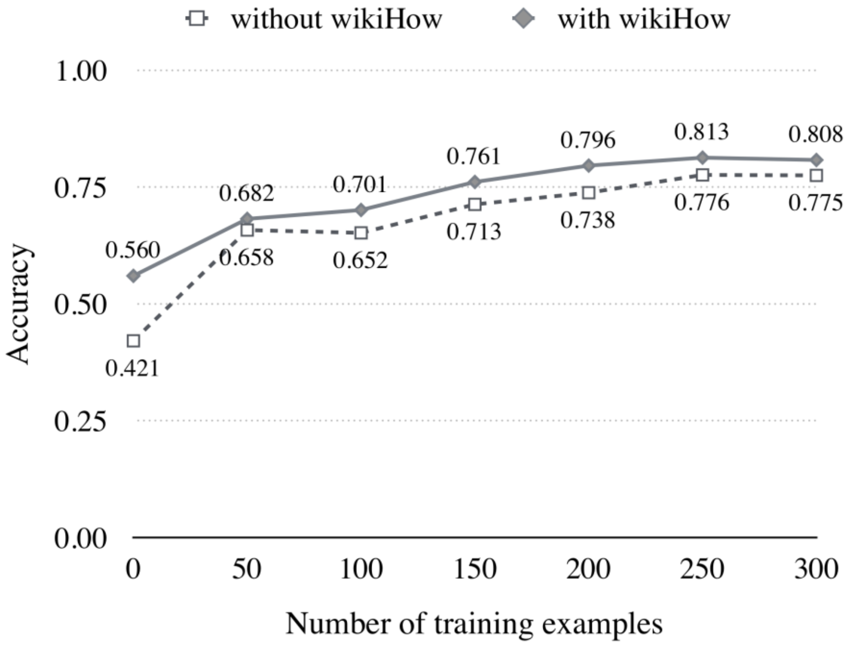
How can it be used?

- Large dataset for training or finetuning & testing
- A model trained on their data did well on other datasets (transfer learning)

SWAG



ROCStories



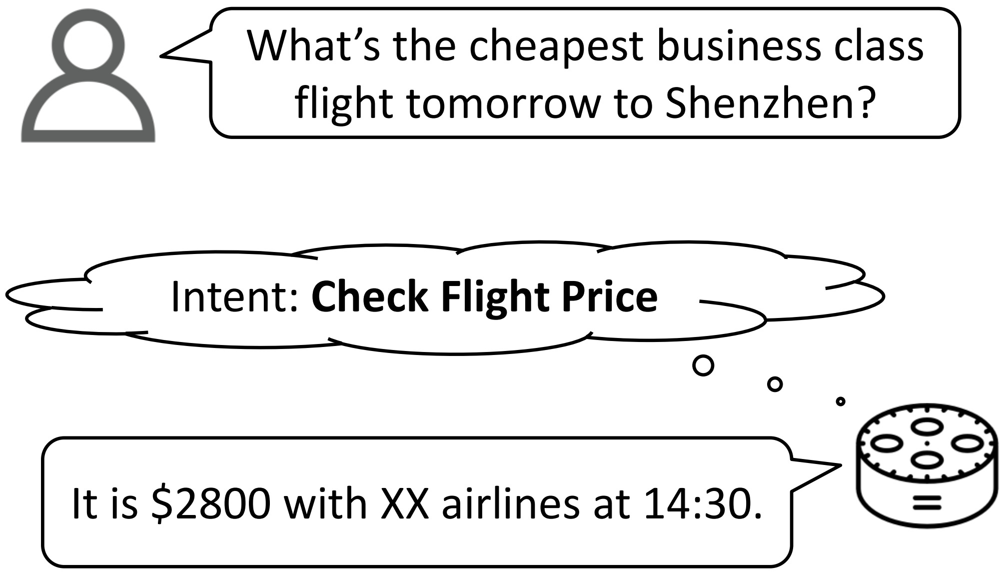
(Zhang et al., 2020): Reasoning about Goals, Steps, and Temporal Ordering with WikiHoW

Intent Classification with Goal-Step Relations

(ZHANG ET AL., 2020): INTENT DETECTION WITH WIKIHOW

Intent Detection

- Task-oriented dialog systems needs to match an **utterance** to an **intent**, before making informed responses
- Sentence classification task
  - Given an utterance, and some candidate intents
  - Choose the correct intent
  - Evaluated by accuracy



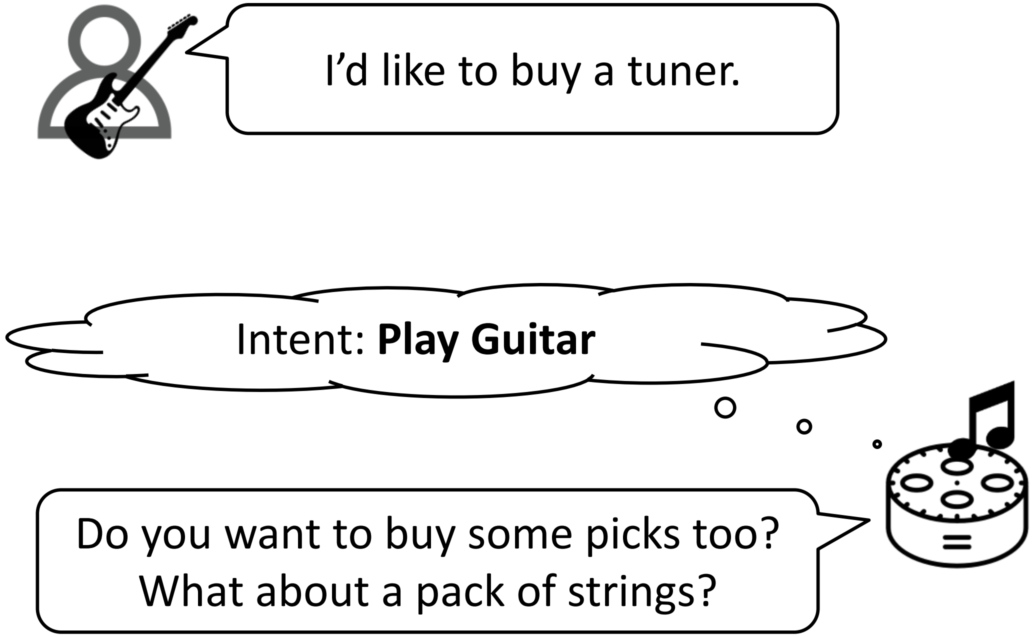
Example from Snips (Coucke et al., 2018)
Utterance: "Find the schedule at Star Theatres."
Candidate intents: Add to Playlist, Rate Book, Book Restaurant, Get Weather, Play Music, Search Creative Work, **Search Screening Event**

(Zhang et al., 2020): Intent Detection with WikiHow

Intent Detection Gets Difficult

- Some intents are hard to infer
  - Require some world knowledge
  - Utterances and intents might span many domains, sometimes niche ones
  - Multilingual settings, especially low-resource

Image: https://safebooru.org/index.php?page=post&s=view&id=546415



(Zhang et al., 2020): Intent Detection with WikiHow

wikiHow can help

- A **goal** (title stripped of “how to”) can approximate an **intent**
- A **step** (step paragraph header) can approximate an **utterance**
- Not a perfect correspondence
  - Some goals are too specific to be an intent
  - Some steps are unlikely utterances
- Still, potentially very strong pre-training data

1 Learn the driving rules for your location.
2 Get your permit.
3 Practice driving.
…

Data: https://github.com/zharry29/wikihow-goal-step

(Zhang et al., 2020): Intent Detection with WikiHow

Results

English data

Siamese NN, prev SoTA

Data augmented with back-translation to & from Chinese

	Snips	SGD	FB-en
	Single-turn English	Schema-guided Dialogue (multi-turn, max 4)	Single-turn Multilingual
(Ren and Xue, 2020)	.993	N/A	.993
(Ma et al., 2019)	N/A	.948	N/A
+in-domain (+ID)	.990	.942	.993
(ours) +WH+ID	**.994**	**.951**†	**.995**†
(ours) +WH 0-shot	.713	.787	.445
Chance	.143	.250	.083

Table 2: The accuracy of intent detection on English datasets using RoBERTa. State-of-the-art performances are in bold; † indicates statistically significant improvement from the previous state-of-the-art.

Multilingual data

Multi-task learning with intent detection & slot filling

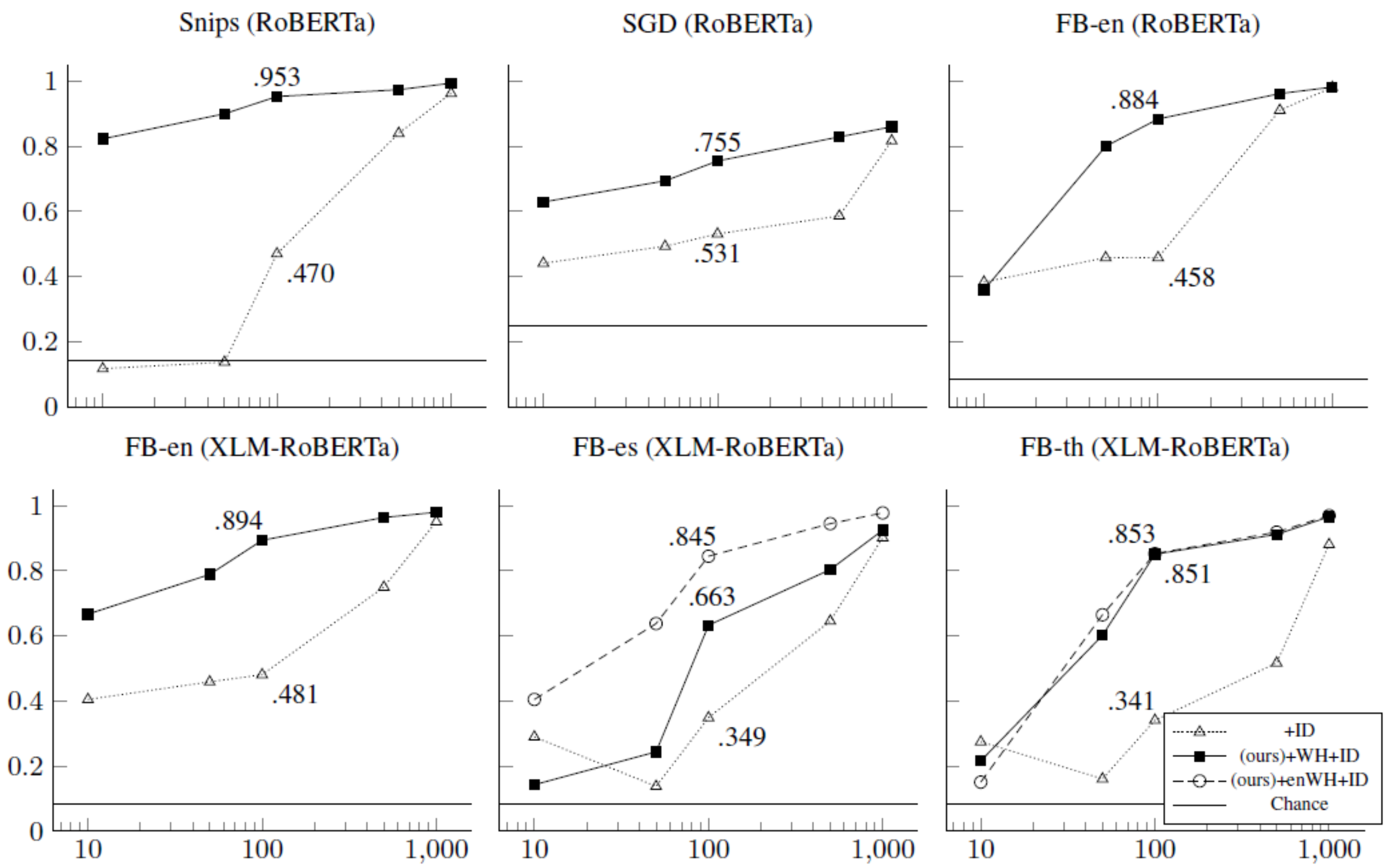
	FB-en	FB-es	FB-th
(Ren and Xue, 2020)	.993	N/A	N/A
(Zhang et al., 2019)	N/A	.978	.967
+in-domain (+ID)	.993	.986	.962
(ours) +WH+ID	**.995**	.988	.971
(ours) +enWH+ID	**.995**	**.990**†	**.976**†
(ours) +WH 0-shot	.416	.129	.119
(ours) +enWH 0-shot	.416	.288	.124
Chance	.083	.083	.083

Table 3: The accuracy of intent detection on multilingual datasets using XLM-RoBERTa.

(Zhang et al., 2020): Intent Detection with WikiHow

Few-Shot Transfer

- Learning curves of models in low-resource settings
- Vanilla transformers (+ID) struggle with low-resource setting



Vertical axis: accuracy of intent detection
Horizontal axis: # in-domain training examples of each task, distorted to log-scale

(Zhang et al., 2020): Intent Detection with WikiHow

Procedural Construction

(LYU ET AL., 2021): GOAL-ORIENTED SCRIPT CONSTRUCTION

Motivation

- Models can infer goals, steps, and ordering in **existing procedures**. Can we go one step further?
- **Creating new procedures**:

  If you know

  how to make an apple pie and how to make a banana cake

  can you infer

  how to make a banana pie ?

- This is commonsense knowledge to humans; do language models have it?

(Lyu et al., 2021): Goal-Oriented Script Construction

The Script Construction Task

- **Script:** a standardized sequence of events
- **Script Construction:** Given the goal, narrate the full script
  - **Example**: If someone **goes to a restaurant**, what will typically happen?

1. Greeted by a waiter
2. Sit down at a table
3. Look at menu
4. Order food
5. Eat
6. Ask for the check
7. Pay
8. ....

(Lyu et al., 2021): Goal-Oriented Script Construction

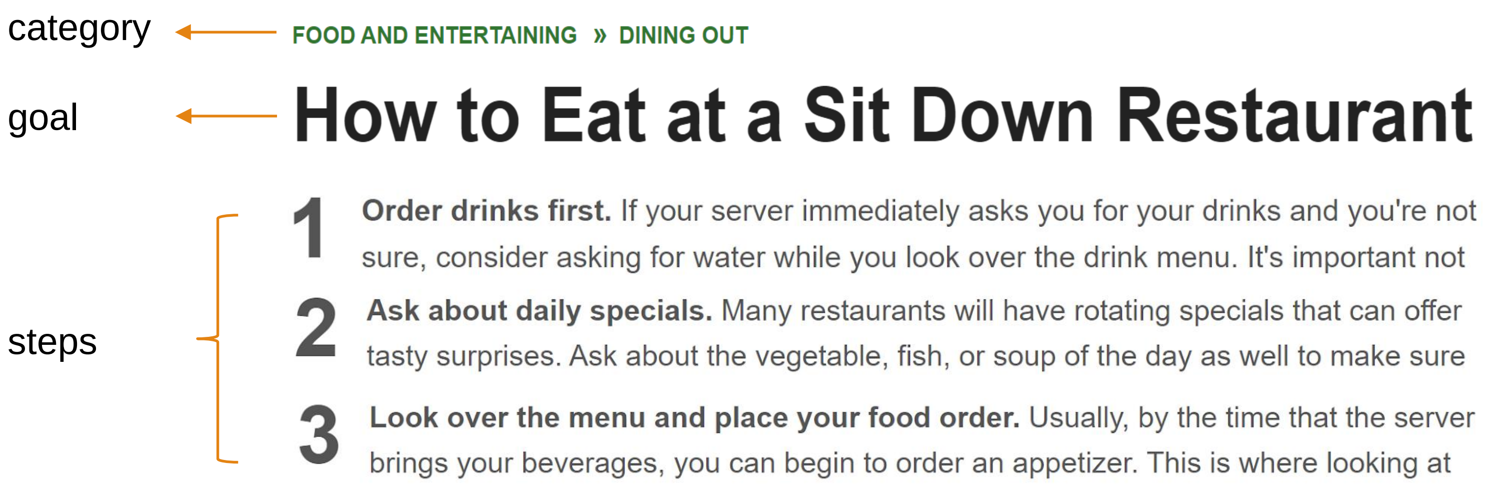
How do we set up the task?

- **Generation** setting
  - Input: a goal
  - Output: a list of steps
  - Evaluation: Perplexity; Human
  - Concerns: might be too difficult; output might be "unfocused"
- **Retrieval** setting
  - Input: a goal, a set of candidate steps
  - Output: a list of steps
  - Evaluation: Accuracy; Recall; Normalized Discounted Cumulative Gain + Kendall's Tau; Human
  - Concerns: where to get a suitable set of candidate steps?

(Lyu et al., 2021): Goal-Oriented Script Construction

Building a Dataset from wikiHow

- They built a script construction dataset of **18 languages** using **wikiHow articles**



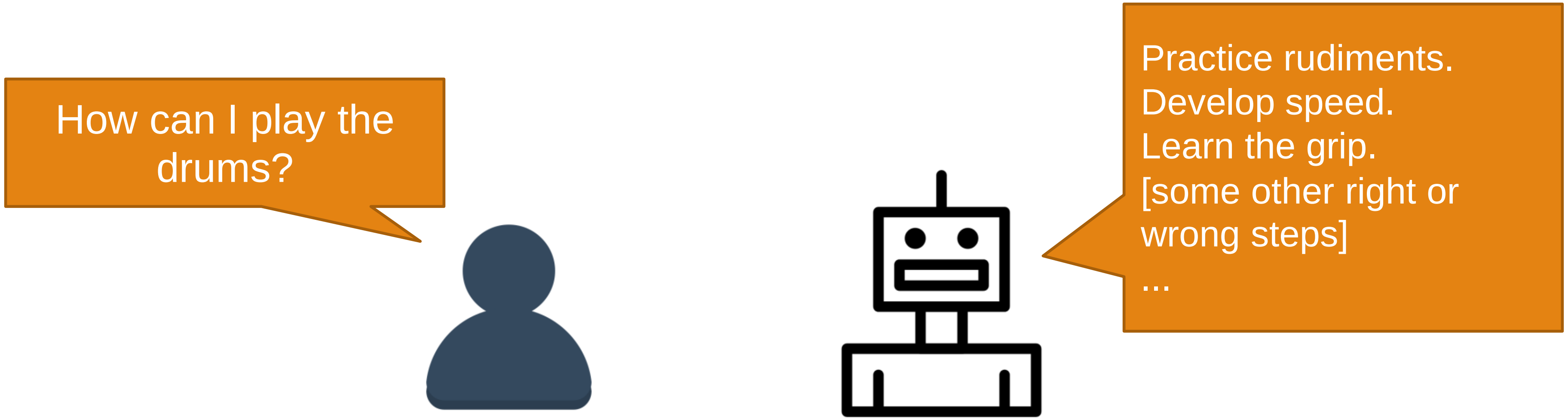
- For the Retrieval Setting, the set of candidate steps are from all articles of the same category

(Lyu et al., 2021): Goal-Oriented Script Construction

How to construct scripts?

- **Generation-based approach:**

  End-to-end finetuning of a generative language model, Google's Multilingual-T5 (Xue et al., 2021).
    - During training, input is goal + steps
    - During inference, input is goal; output is steps

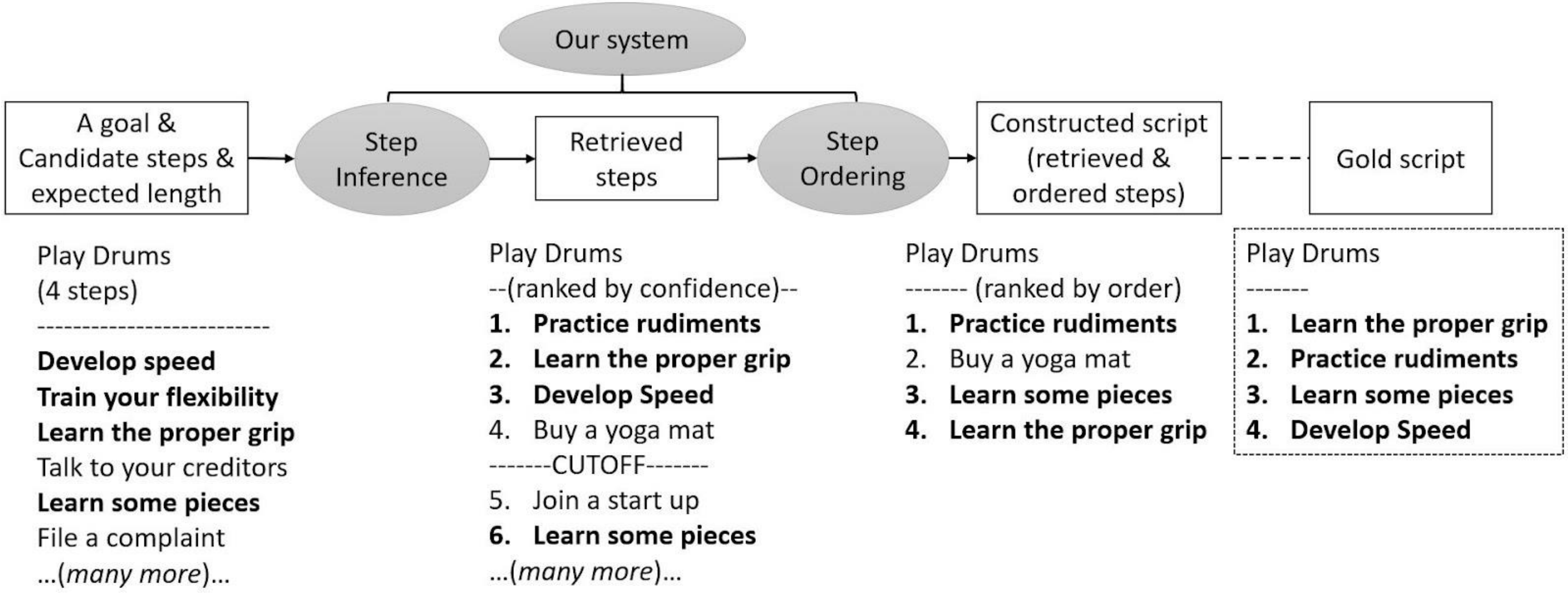


(Lyu et al., 2021): Goal-Oriented Script Construction

How to construct scripts?

[Reasoning about Goals, Steps, and Temporal Ordering with WikiHow]

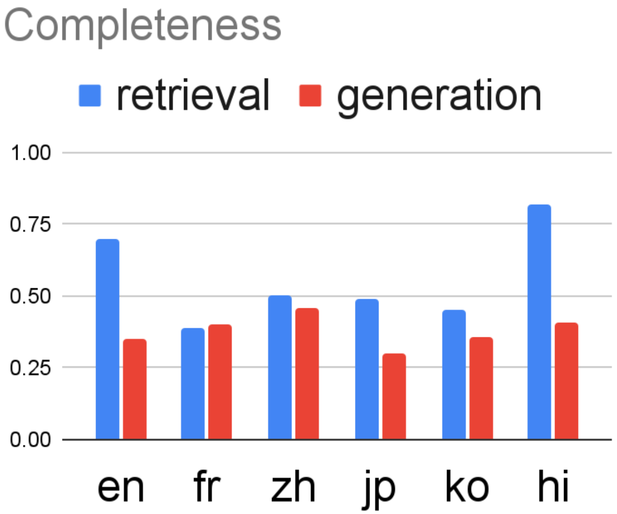
- **Retrieval-based approach:**
  A pipeline using the Step Inference model & Step Ordering models from their previous EMNLP 2020 work

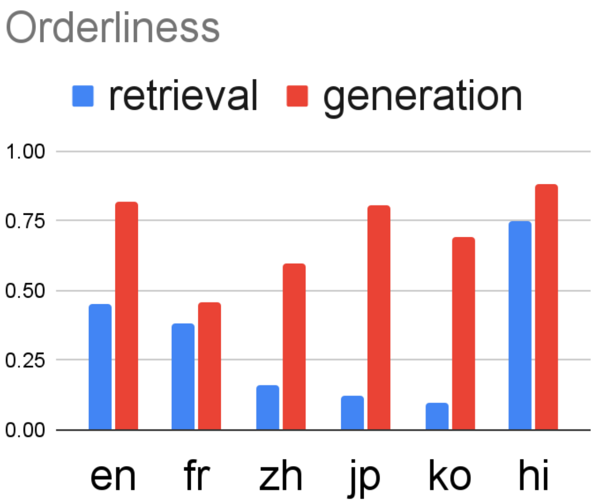


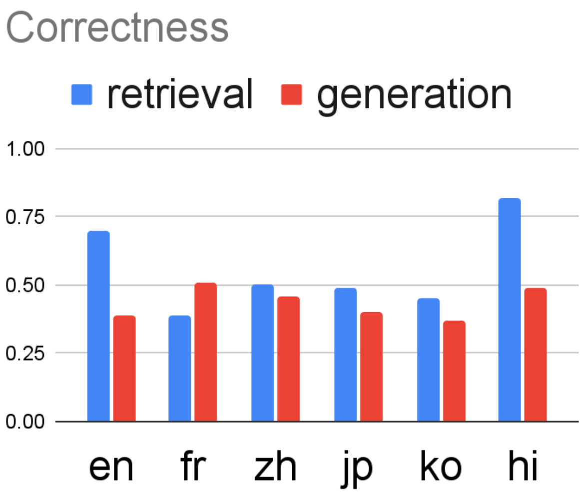
(Lyu et al., 2021): Goal-Oriented Script Construction

How well can models construct scripts?

- Lyu et al. ask humans to ***edit*** the predicted script by either deleting or moving a step
- They then calculate:
  - "*Correctness*": len(edited script) / len(predicted script)
  - "*Completeness*": len(edited script) / len(gold script)
  - "*Orderliness*": Kendall's Tau of steps in the edited script







(Lyu et al., 2021): Goal-Oriented Script Construction

How can it be used?

- Construct scripts in new domains:
  - Models trained on wikiHow (**daily activities**) can transfer to the **news** (**political & military**) domain
  - Example: A script of "**Roadside Improvised Explosive Device Attack**"

1. Movement.Transportation
   (Example: "Taliban transported explosives in car to centre place.")
2. Conflict.Attack.DetonateExplode
   (Example: "Someone detonated ied explosive device at ghazni place.")
3. Conflict.Attack
   (Example: "Group attacked convoy using explosive.")
4. Life.Injure
   (Example: "861 was injured by contractor using explosive.")
5. ArtifactExistence.DamageDestroyDisableDismantle.Damage
   (Example: "Someone damaged vehicle.")
6. ArtifactExistence.ManufactureAssemble
   (Example: "Someone manufactured or assembled or produced weapons.")
7. Life.Die
   (Example: "Members died, killed by efps killer.")
8. Justice.TrialHearing
   (Example: "Someone tried someone before dunford court or judge.")

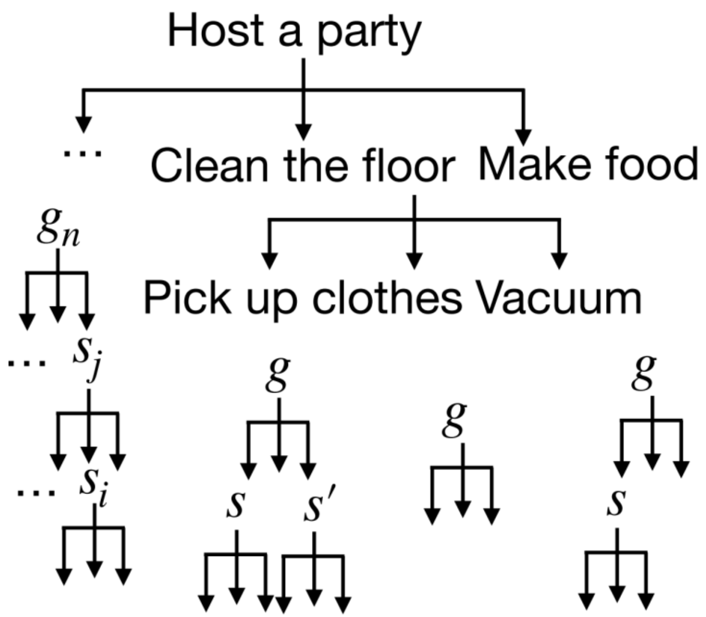
(Lyu et al., 2021): Goal-Oriented Script Construction

Hierarchical Procedures

(ZHOU ET AL., 2022): SHOW ME MORE DETAILS: DISCOVERING HIERARCHIES OF PROCEDURES FROM SEMI-STRUCTURED WEB DATA

Procedures are Hierarchical

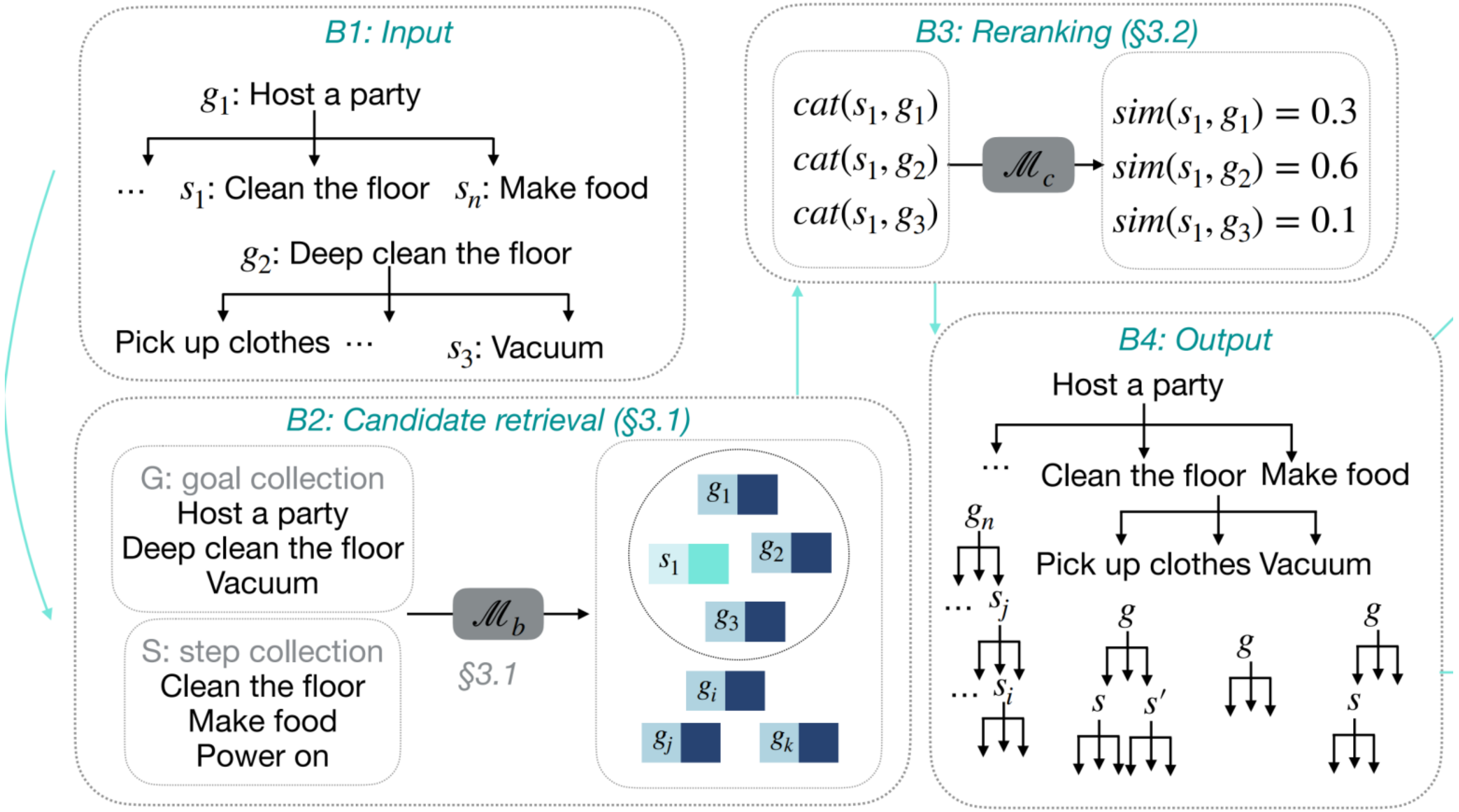
- An event can simultaneously be a **goal** of one procedure, and a **step** in another
- A procedural hierarchy… So what?
  - Can "explain in more details" by expansion
  - Can shed light on event **granularity** (why?)
- How do you build such hierarchy?
  - To "host a party", I need to "clean the floor"; to "clean the floor", I need to do what?



(Zhou et al., 2022): Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data

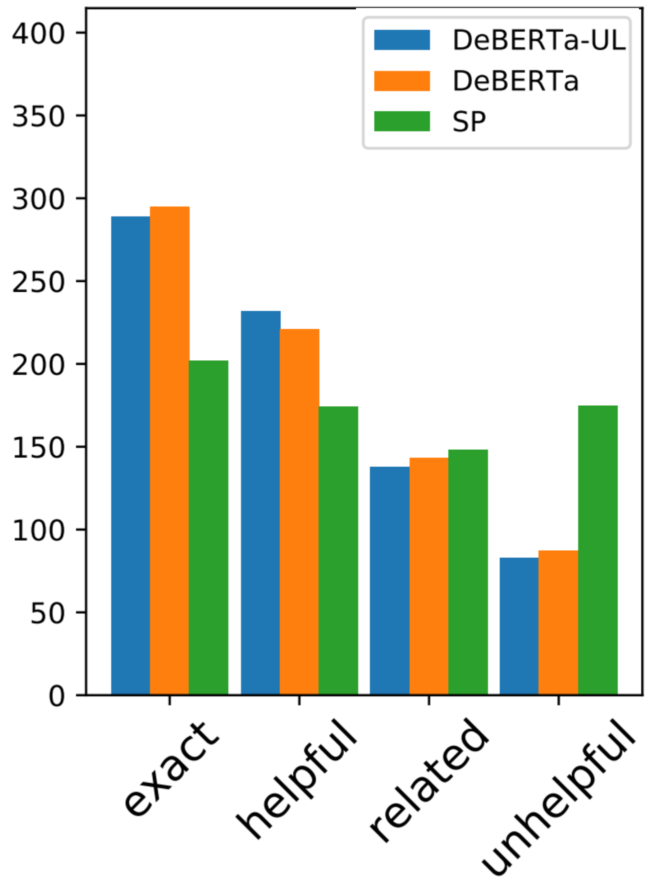
Linking Steps and Goals

- For each procedure, for each **step**, find another procedure whose **goal** has the same meaning
- Effectively a paraphrasing matching or semantic search problem (is it really?)
- Method: retrieve-then-rank



Procedural Hierarchy is Useful

- Crowdworkers rate linked **step-procedure** pair, and decide if the instructions of the **procedure** is helpful for doing the **step**
  - About 80% are helpful
- Can be directly applied to a task-oriented dialog system (e.g., Alexa Prize Taskbot)



(Zhou et al., 2022): Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data

Event Granularity is Not So Obvious

- Words lower in the hierarchy (supposedly more low level): push, continue, learn, decide, repeat, avoid, finish, move, wait, accord
- Words higher in the hierarchy (supposedly more low level): decorate, knead, season, beat, paint, simmer, dig, sew, build, melt
- They don't seem clustered by granularity

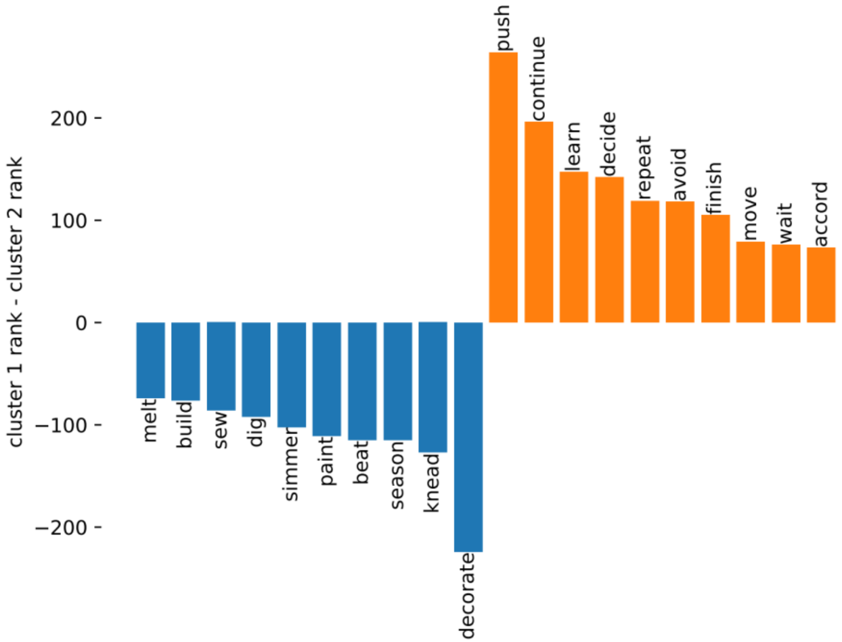


Figure 3: The verbs with largest rank difference in two clusters. The blue bars are words becoming less frequent in cluster 2 (decomposed steps) and the orange bars are words becoming more frequent.

(Zhou et al., 2022): Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data

Procedures in IF (Think-Pair-Share)

How might you use procedures in interactive fiction?

Potential Ways to Use Procedures in IF

- Infer likely steps given a goal (c.f. our goal-step relation work)
  - If a user decides to “conquer a dragon”, they need to “buy potions”, “level up”, “get equipment”, “swing the sword”, etc.; they won’t need to “get a PhD”, “play music”, etc.

Potential Ways to Use Procedures in IF

- Reason about entity states
  - If a user "uses a *key* to open a *door*", the *key* would remain intact, but the *door* changes from "locked" to "unlocked", both of which are implicit (c.f. Tandon et al., 2020)
  - If a user "throws away the *only* key they have", they would have 0 keys. (c.f. Li et al., 2021)

Potential Ways to Use Procedures in IF

- Use procedures as scaffolding of the story
  - Language models tend to hallucinate given too much freedom
  - Instead, use procedures to guide them