Proposed Projects: SWP 2023/24

Prof. Dr. Anette Frank
Department of Computational Linguistics
Heidelberg University
frank@cl.uni-Heidelberg.de
October 18, 2023

Proposed Projects

Project 1: Reliable Large-scale Linguistic Annotation using LLMs

Dialogue Generation & Annotation
– with Quality Filtering Metrics

Proposed Projects

Project 1: Reliable Large-scale Linguistic Annotation using LLMs

Dialogue Generation & Annotation
– with Quality Filtering Metrics

Project 2: Analyzing Ambiguity and Biases in LLMs with (Interpretable) SBERT

Project 1: Generating Datasets with reliable Linguistic Annotations – enhanced by Quality Filtering Metrics

Motivation

- Conversational LLMs like ChatGPT have been shown to be good *NL generators* and *linguistic data annotators*
- But for NLP tasks, their performance often lags [Kocon etal; Bang etal. 2023]
  - ChatGPT: 25% avg. loss in quality compared to SOTA solutions
  - Critical: pragmatics, reasoning, hallucinations, biases from RLHF
- Still, they could be used to ***generate & label <u>training data</u> for NLP tasks***
  - + : reduce annotation costs for special tasks / low resource scenarios
  - – : limitations: hallucinations; not error-free; unstable; low diversity?

➡ **combat weaknesses: create reliable data annotators** for NLP

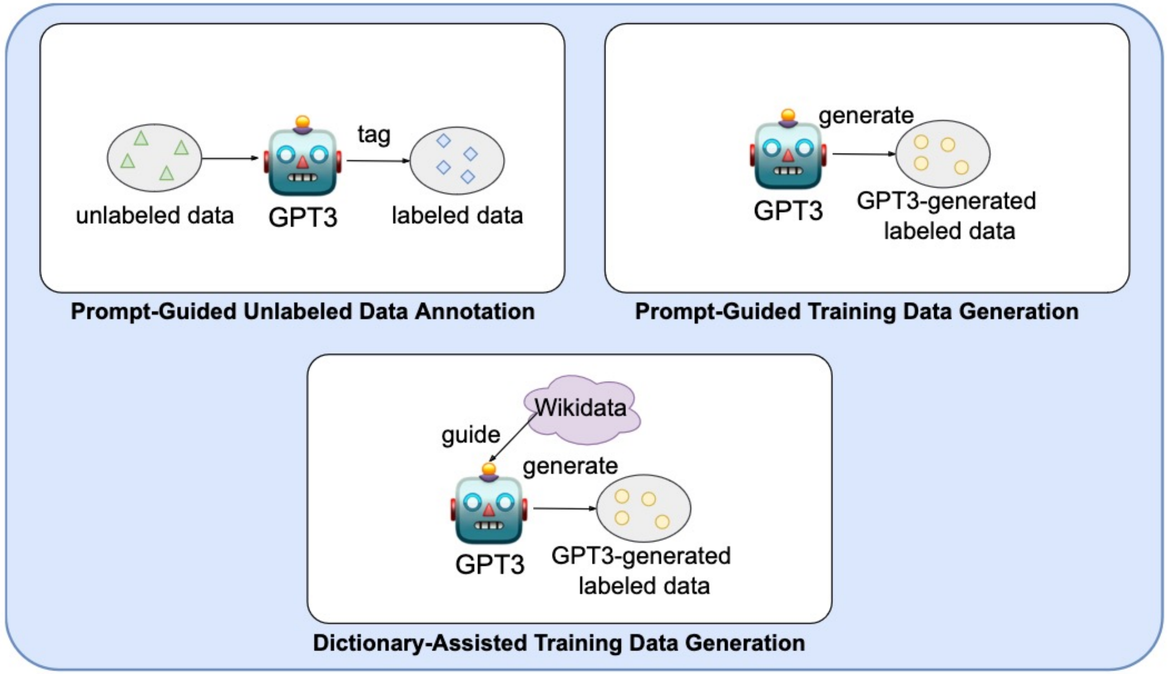
Project 1: Generating Datasets with reliable Linguistic Annotations – enhanced by Quality Filtering Metrics

➥ Aim

- generate datasets w/ linguistic annotations using (Chat)GPT(3)
- apply reliable & maximally general filtering methods
- focusing on a cost-intensive and challenging task: **dialogue!**

SOTA: Generating and annotating data with GPT3

Prompting (Chat)GPT(3) w/ labeled input pairs [Ding et al. 23, ACL]



y_i = GPT3 (I_{IOP} ,x_i)
x_i : input sample; y_i : annotated sample
I_{IOP} : demonstrations

PGDA: label ***unlabeled*** training data
PGDG: ***self-generate labeled*** data
DADG: self-generate labeled data ***guided*** by lexicon, ontology
GPI: ***0-shot*** annotation of testdata

➡ studied Tasks: **SST2, NER, FewRel** – no complex ones: **DepP, Coref, SRL, ...**
➡ GPT3 (vs. ChatGPT: equal but cheaper)

SOTA: Promising but challenging case: Dialogue

AI-Generated Goal-Oriented Dialogues & Annotations [Labruna et al 2023]

- 3 types of dialogues (***task-oriented, collaborative, explanatory***)
- interactive and one-shot (modeling user and system interaction)
- English and Italian

prompt to ***generate***

Dialogue Generation

- Develop *instruction prompts*: using 5 reference dialogues / type
- ChatGPT *generates 2 variants / reference* → 10 new + 5 original/ type

prompt to ***annotate***

Dialogue Annotation

Prompt for *annotation of new + original dialogues*

- high-level: what is expected to do
- content & format of different annotation types
- input: dialogue to be annotated

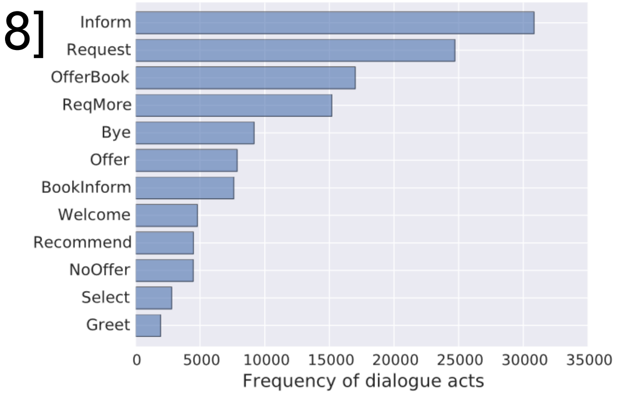
evaluate

Evaluation by crowd workers

- Using established quality criteria and annotation guidelines
- Rate quality of the ***dialogue*** itself and of the ***generated annotations***

Dialogue types and datasets

Task-oriented dialogue: **e.g. MultiWOZ** [Budzianowski et al., 2018]

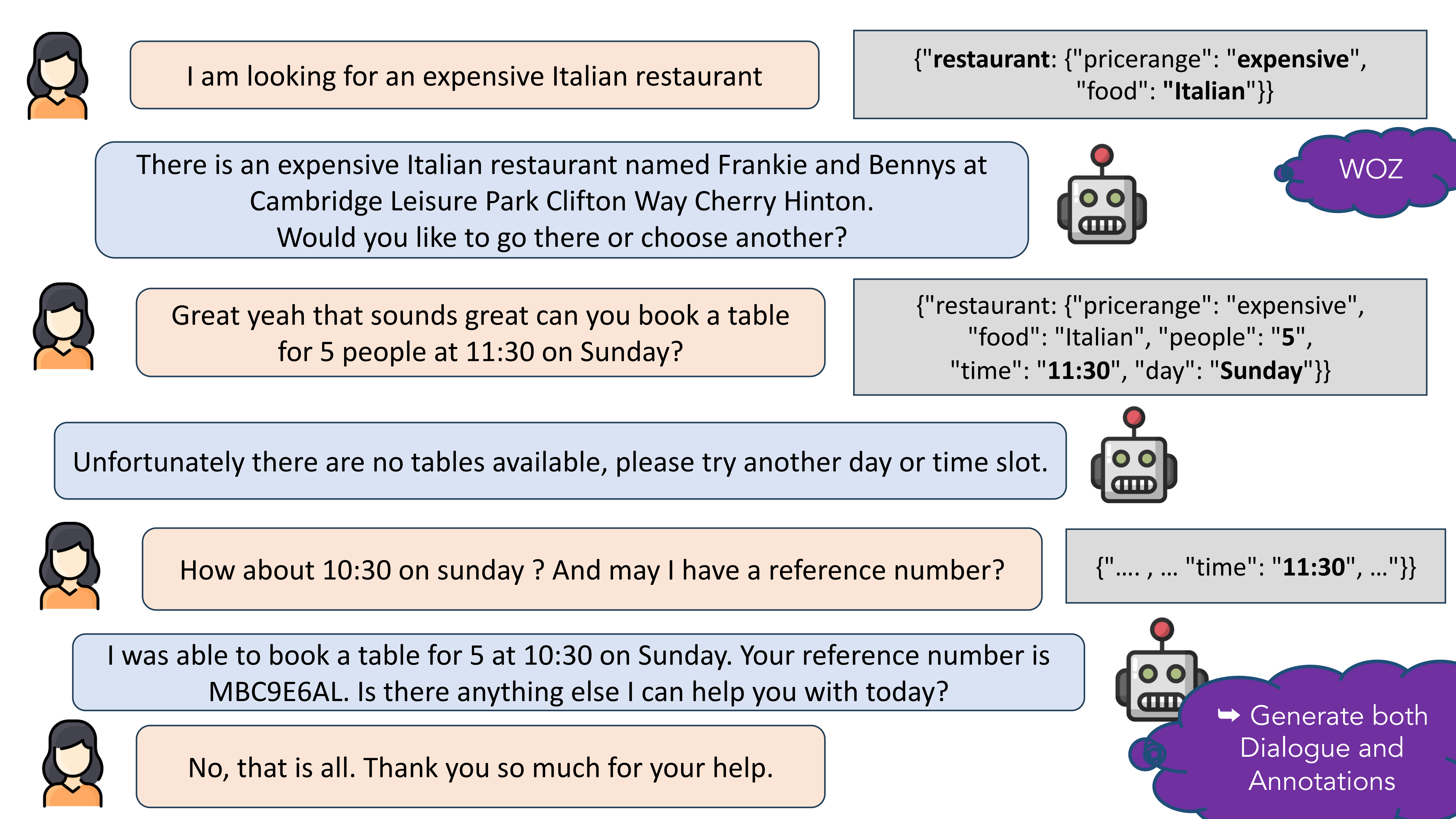


Main tasks:

- dialogue understanding (Louvan and Magnini, 2020)
- dialogue state tracking (Balaraman et al., 2021)

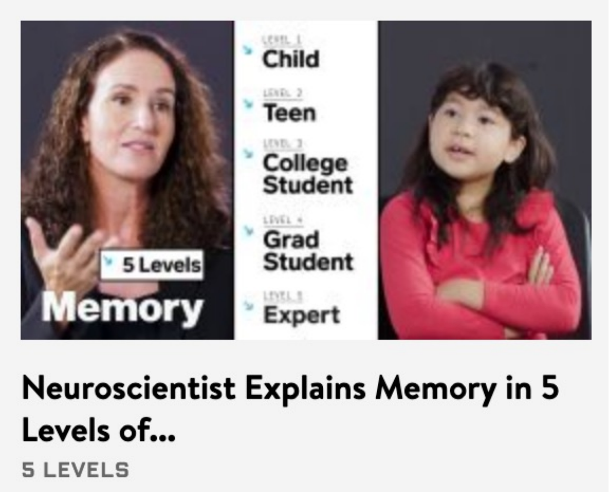
Annotations:

- **Dialogue acts** (e.g., welcome, inform, request, select, bye, …)
- **Dialogue states**: triples encoding facts about
  - domain (e.g. RESTAURANT), domain-relevant slots (FOOD), slot-values (ITALIAN)
  - using ontology of the conversational domain
- Annotations are constructed *incrementally for the evolving dialogue*
  - new slot-values are added to previous ones
  - slots represent the system's belief state of the user requirements at each step


I am looking for an expensive Italian restaurant
{"restaurant: {"pricerange": "expensive", "food": "Italian"}}
There is an expensive Italian restaurant named Frankie and Bennys at Cambridge Leisure Park Clifton Way Cherry Hinton. Would you like to go there or choose another?
WOZ
Great yeah that sounds great can you book a table for 5 people at 11:30 on Sunday?
{"restaurant: {"pricerange": "expensive", "food": "Italian", "people": "5", "time": "11:30", "day": "Sunday"}}
Unfortunately there are no tables available, please try another day or time slot.
How about 10:30 on sunday ? And may I have a reference number?
{".... , ... "time": "11:30", ..."}}
I was able to book a table for 5 at 10:30 on Sunday. Your reference number is MBC9E6AL. Is there anything else I can help you with today?
No, that is all. Thank you so much for your help.
Generate both Dialogue and Annotations

Explanatory dialogues

Serve to explain a concept in a collaborative way.
Explainer and explainee work together
to construct an understanding of a particular topic.



WIRED 5 Levels Corpus
[Wachsmuth & Alshomary 2022]

- Transcriptions from the WIRED video series 5 Levels (English)
- University teacher explains 13 topics (music harmony, …, machine learning) to 5 explainees of varying levels (child, teenager, undergrad, postgrad, colleague).
- 65 dialogues manually labelled for *topic, dialogue act, type of explanation*
- Labruna et al. use the 5-level dialogues for topic “machine learning”.

WIRED 5 Levels Corpus

Explaining dialogue on the main topic "blockchain"

01 Do you know what we're gonna talk about today? It's called blockchain.

02 What's blockchain?

03 That's a really good question. It's actually a way that we can trade. Do you know what trade is?

04 Mmm-hmm, it's when you take turns doing something. It's when you give up most of what you want, right?

05 When you give up most of what you want? Well, sometimes that definitely happens for sure. What if I told you that this is the kind of technology that I work on that means you could trade with any kid all over the world?

06 Really?

07 Yeah.

08 If I could trade with any kid, I would trade, well, I would trade something I don't like so much.

09 That's probably a good idea, maybe somebody else likes it more than you do. So normally, when people trade, they have to go to the store, or they have to know the person so they can get what they asked for. With blockchain, you can make that exact same trade, but you don't need the store, and you don't even necessarily need to know the other person.

10 Really?

11 Really.

Explainer (expert)

(child) **Explainee**

Results and Findings [Labruna et al 2023]

Quality of generated dialogues

- + **high or comparable to humans** (except for Italian datasets)
- – **reliability**: errors regarding hallucinations and instruction-following

Quality of annotations

- – weaknesses in slot accuracy and goal accuracy
- – long dialogues could not be annotated → ask model to generate shorter ones?
- + MultiWOZ: comparable to SOTA systems (auto vs. gold)

Found limitations:

- Instable annotation quality when same prompt is used multiple times

➥ apply ***error metrics to detect hallucinations***: domain correctness; etc.

➥ use ***different LMs as labelers/evaluators***

Project 1 proposal(s)

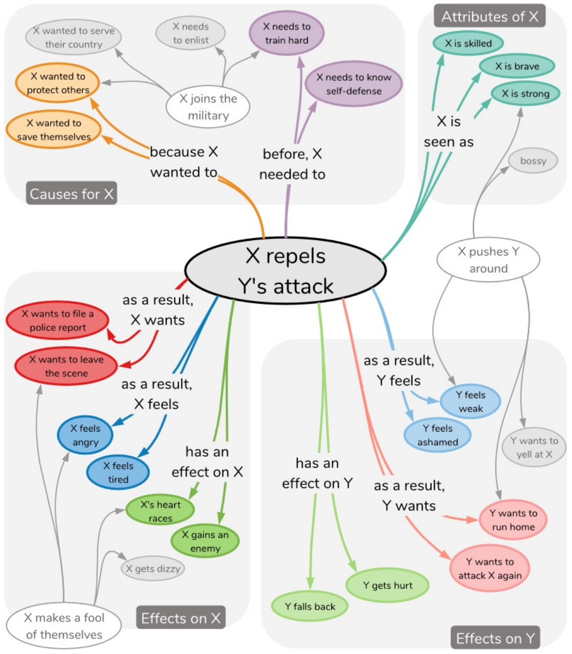
Use ChatGPT to *generate training data* for Goal-Oriented Dialogues *with Annotations*

(similar to Labruna et al. 2023)

- on ***different*** ***dialogue datasets (of similar types),*** or reproduce & ***enhance*** experiments on their datasets (smaller group)
- trying to ***improve*** by integrating
  - ***better guides (control) of generation*** and/or
  - post-hoc ***error detection methods or metrics*** to identify hallucinations

Option 1

esp. for interactive, also TOD dialogue (w/ domain ontology)

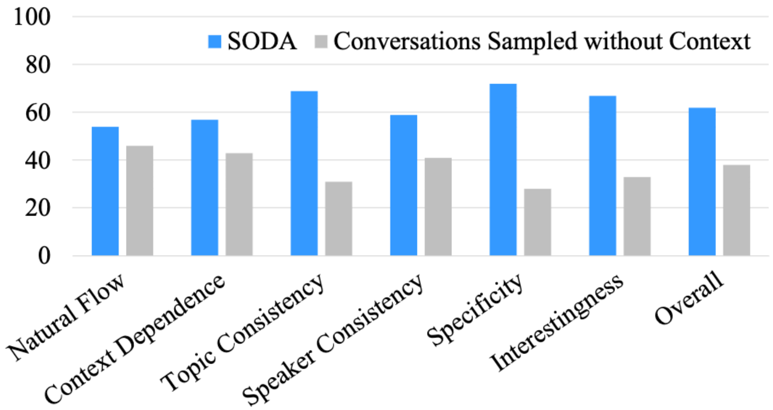


Using commonsense knowledge to *guide and control generation*

(cf. Kim et al. 2023, Jiang et al. 2021)

- trigger knowledge-guided questions:
  - What has happened?
  - Why did it happen?
  - What would you want to do now?
  - Who is capable of doing this?

Common keywords for each relation (excluding the above)	
xAttr (18%)	kindness, anger, intelligent, responsibility, friend, trust, conversation, food, generosity, smart
xEffect (17%)	gratitude, anger, upset, hard work, happy, money, friend, boss, party, kindness
xIntent (23%)	independence, hard work, determination, money, relaxation, anger, kindness, store, understanding
xNeed (7%)	job, money, confidence, comfort, advice, interest, conversation, listening, store, park
xReact (25%)	frustration, anger, confidence, happy, pride, relief, disappointment, relaxation, anxiety, satisfaction
xWant (11%)	conversation, store, determination, apology, learning, doctor, job, friend, improvement, marriage

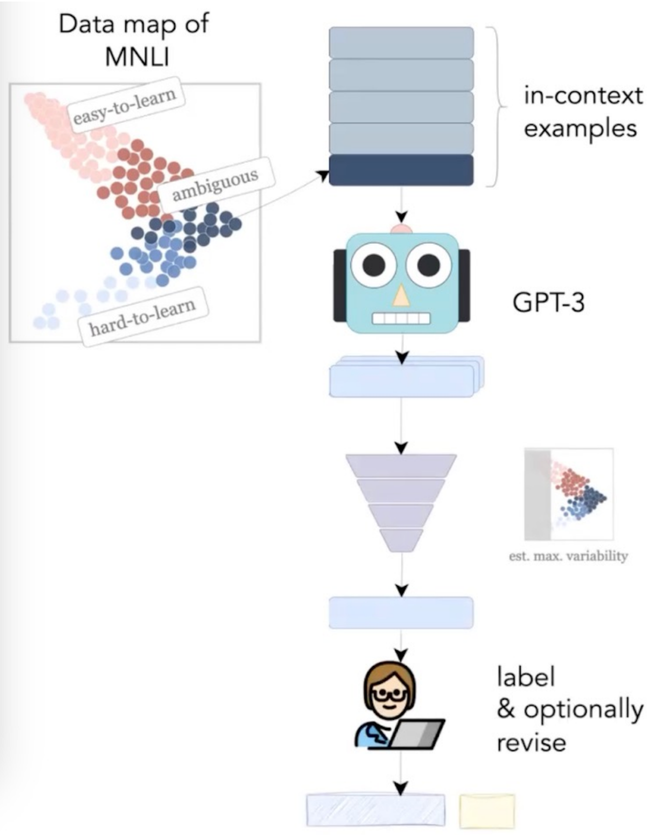


Option 1
esp. for interactive dialogue

How to judge the outcomes? – using data maps (Swayamdipta et al 2020)

Deploy metrics for quality estimation to annotate ➥ filter, improve

Metrics to check for:
relevant keywords,
coherence with rules,
factuality, consistency,
coherence, diversity ..



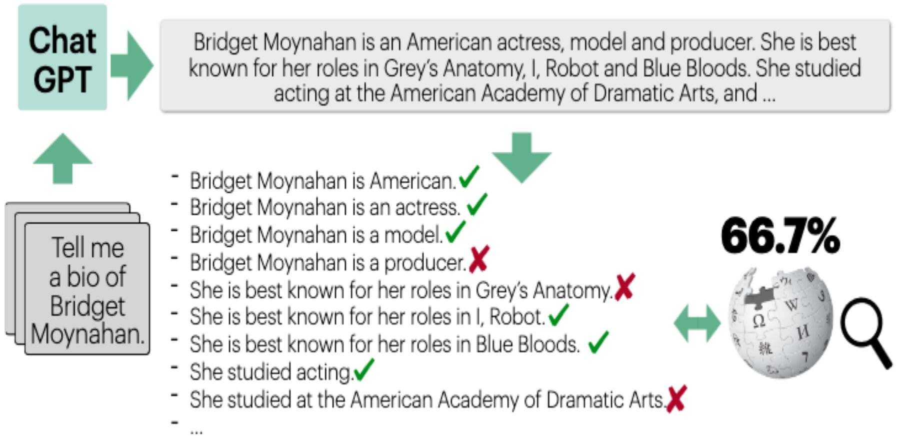
1. ***Exemplar collection***: *collect groups of tricky examples found using data maps*

2. ***Overgeneration***: *prompt GPT-3 to create more similarly tricky examples!*

3. ***Filtering***: *develop metric based on data maps for filtering*

4. ***Human annotation***: *humans do what humans are good at, evaluating & improving examples!*

Option 2: Factuality metrics
esp. for TOD



Implement FactScore
[Min et al. 2023]
(existing metric, no public code)

- LLM_{Subj} decomposes model-generated text into atomic statements
- LLM_{Eval} judges each statement: supported (or not) in given domain?

$$f(y) = \frac{1}{|\mathcal{A}_y|}\sum_{a\in\mathcal{A}_y}[a \text{ is supported by } \mathcal{C}],$$

$$\textsc{FactScore}(\mathcal{M}) = \mathbb{E}_{x\in\mathcal{X}}[f(\mathcal{M}_x)|\mathcal{M}_x \text{ responds}]$$

Limitations:
- Poor measure of coverage
- Requires undebated factuality of atomic facts
- Weighting individual facts
- Overlapping or inconsistencies in context

Mainly TOD!

- Which models to use? InstructGPT (paid) for break-down – ChatGPT or LLAMA-7B – FLAN for eval
- Retrieve relevant facts from ontology and/or background text
- Baselines to explore: NLI via, e.g., RoBERTa or T5 (cf. Steen et al. 2023)

Possible metrics

Diversity

- distinct-n
- Benchmark: Evaluating the Evaluation of Diversity in Natural Language Generation

Similarity

- SBERT, S3BERT (Opitz & Frank 2022), BERTScore

Dialogue Coherence

- DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations
- GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems
- ACCENT: An Automatic Event Commonsense Evaluation Metric for Open-Domain Dialogue Systems

Dialogue State Tracking

- Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking
- Survey: "Do you follow me?": A Survey of Recent Approaches in Dialogue State Tracking
- Towards Fair Evaluation of Dialogue State Tracking by Flexible Incorporation of Turn-level Performances

Useful resources and tools

- Hallucination Detection Benchmark:
  - Liu et al. 2022: A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation, ACL.
- For inspiration
  - Knowledge-Aware Audio-Grounded Generative Slot Filling for Limited Annotated Data (how to generate slots using knowledge)
- Hugging Face Inference Endpoints
  - only use for inference with chosen model
    - to perform support/not-support queries with prompts for LLM_{Eval} in FactScore
    - to perform controlled generation (alternative to open ChatGPT if needed)
  - video: Deploy models with Hugging Face Inference Endpoints

Dataset options

Name	Description	Domain
DailyDialog	A dataset consisting of daily dialogues, annotated with conversation intention and emotion information	Open-domain Dialogue
PersonaChat	A chit-chat dataset where paired Turkers are given assigned personas and chat to try to get to know each other.	Open-domain Dialogue
Switchboard Dialog Act	A collection of 1,155 five-minute telephone conversations between two participants, annotated with speech act tags.	Open-domain Dialogue
MuTual	A dialogue reasoning dataset containing English listening comprehension exams	Dialogue Reasoning
MultiWOZ	A fully-labeled collection of human-human written conversations spanning over multiple domains and topics.	Task Oriented Dialogue
Curiosity [could use Wikipedia]	An open-domain dataset annotated with preexisting user knowledge and dialogue acts.	Knowledge-Grounded System
EmoryNLP	Collected from Friends' TV series, annotated with emotion labels	Empathetic Response

	#Dialog	Avg. #Turns	Avg. Utt. Length	Lexical Diversity
DailyDialog	13K	7.9	14.6	63.0
PersonaChat	11K	14.8	14.2	43.6
WizardOfWikipedia	22K	9.1	16.4	60.3
EmpatheticDialogue	25K	4.3	13.7	64.2
BlendedSkillTalk	7K	11.2	13.6	64.2
ProsocialDialog	58K	5.7	20.0	60.2
SODA	1.5M	7.6	16.1	68.0

References for Project 1

SOTA

Bang et al. 2023: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, arXiv.

Ding et al. 2023: Is GPT-3 a Good Data Annotator?, ACL.

Kocon et al, 2023: ChatGPT: Jack of all trades, master of none, arXiv.

Laskar et al. 2023: A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets, ACL.

Labruna et al. 2023: Unraveling ChatGPT: A Critical Analysis of AI-Generated Goal-Oriented Dialogues and Annotations, arXiv.

ChatGPT CheatSheets: The Great ChatGPT CheatSheet

Methods: Knowledge grounding

Kim et al. 2023: SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization, arXiv.

Jiang et al. 2021: "I'm Not Mad": Commonsense Implications of Negation and Contradiction, NAACL.

References for Project 1

Metrics

Min et al. 2023: FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, arXiv.

Steen et al. 2023: With a Little Push, NLI Models can Robustly and Efficiently Predict Faithfulness, ACL.

Related: (see Project 2):

Opitz & Frank 2021: Towards a Decomposable Metric for Explainable Evaluation of Text Generation from AMR, EACL.

Opitz & Frank 2022: SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features, TACL.

Misc.

Swayamdipta et al 2020: **Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics, EMNLP**

Liu et al. 2022: A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation, ACL.

Sun et al. 2023: Knowledge-Aware Audio-Grounded Generative Slot Filling for Limited Annotated Data

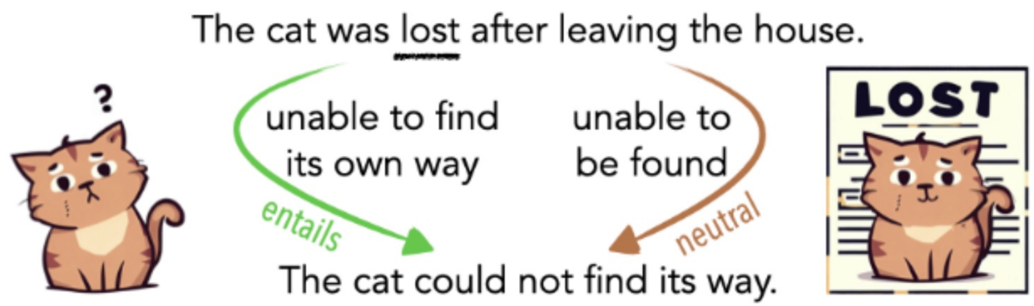
Proposed Projects

Project 2: Analyzing Ambiguity and Biases in LLMs with (Interpretable) SBERT

Project 2: Analyzing Ambiguity & Biases in LLMs

Motivation:

Can LLMs detect ambiguity?



➥ Liu et al. 23:
[We're Afraid Language Models Aren't Modeling Ambiguity]

Human-detected (hidden) ambiguity

Disambiguating hypothesis (w/ label prediction)

Political claim (premise)	Generated paraphrase (hypothesis)	Rating	Prediction	Explanation of ambiguity (ours)
When President Obama was elected, the market crashed...	The stock market reacted immediately to President Obama's election in 2008, ...	Barely-true	{ENTAIL, NEUTRAL}	The claim implies a causal relationship

Multi-label NLI rating reflecting ambiguity

Project 2: Analyzing Ambiguity & Biases in LLMs

Motivation:

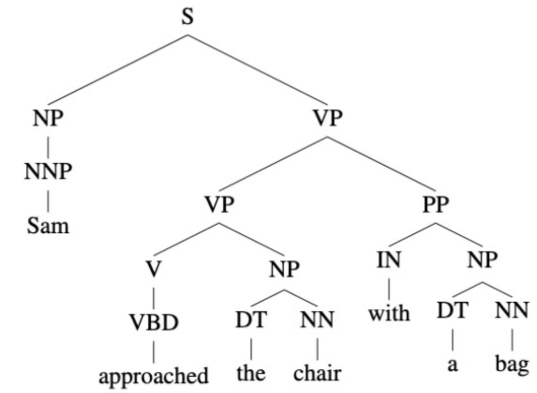
Can LLMs detect ambiguity?

➥ Berzak et al. 2015: [Do You See What I Mean? Visual Resolution of Linguistic Ambiguities]

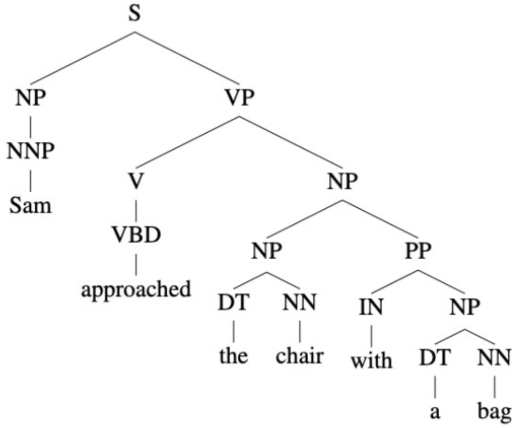
Ambiguities can be resolved
by contextualization,
in text or in visual situations.

In vision & language, the relevant
reading is often directly ‚visible'.

Sam approached the chair with a bag.



(a) First interpretation



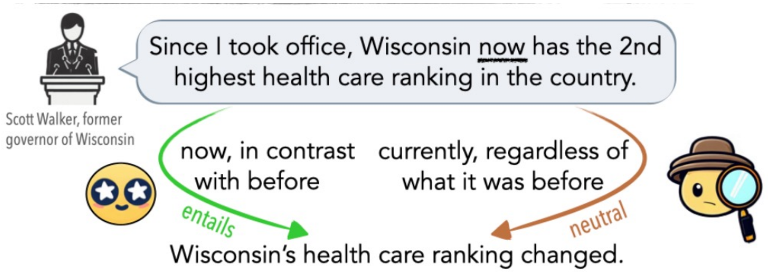
(b) Second interpretation

(c) Visual context

Ambient: Ambiguity in Entailment [Liu et al. 2023]

Many open research questions

- Can language models 'perceive' ambiguities (as humans – sometimes – do)?
- To what extent are they (we) guided by context?
- How much of the model's contextualization results from a (pre)training bias? Does this differ from humans?



New Ambiguity Benchmark AMBIENT

- 1,645 sentences with lexical, syntactic and pragmatic ambiguities (convey multiple readings)
- ***Ambiguity*** is represented via ***natural language inference (NLI)*** in premise and/or hypothesis, by the ***effect it takes on entailment relations***.
- AMBIENT instances:
  - premise and hypothesis pairs with each a ***set of assigned labels*** (E, N, C)
  - a ***disambiguating rewrite*** of P or H for each assigned label (i.e., reading)

Ambient: Ambiguity in Entailment [Liu et al. 2023]

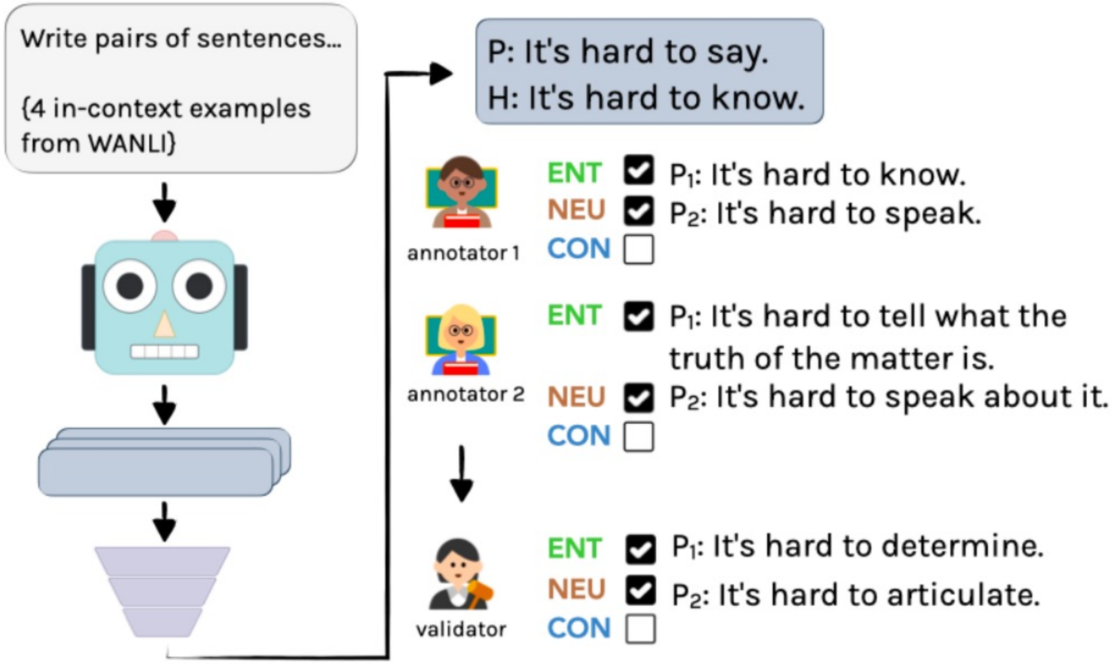
Ambiguity benchmark

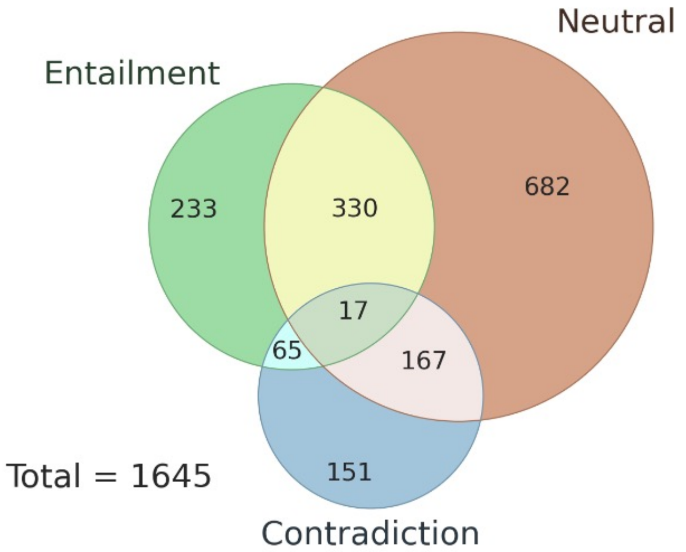
- 1,645 sentences with lexical, syntactic and pragmatic ambiguities (convey multiple readings/messages)
- Ambiguity is represented via natural language inference (NLI) in premise and/or hypothesis, by the effect it takes on entailment relations.
- AMBIENT instances:

Example	Disambiguation 1	Disambiguation 2	Type
P: I'm afraid the cat was hit by a car. H: The cat was not hit by a car. {NEUTRAL, CONTRADICT} : [7 N, 2 C]	P: I'm worried... NEUTRAL : [9 N]	P: I'm sorry to share that... CONTRADICT : [9 C]	*Pragmatic* (44.8%)
P: John and Anna are married. H: John and Anna are not a couple. {NEUTRAL, CONTRADICT} : [5 N, 4 C]	P: ... are both married. NEUTRAL : [7 N, 2 E]	P: ... are married to each other. CONTRADICT : [9 C]	*Lexical* (20.0%)

Ambient Dataset

- Premise – Hypothesis pairs with sets of conflicting NLI labels





Label distribution in Ambient

Category
Lexical
Syntactic
Figurative
Pragmatic
Scopal
Coreferential
Other

Start from 142 samples (handwritten, from NLI datasets & linguistics textbooks)

Automatically generate unlabeled, ambiguous NLI samples, in an overgeneration – filtering process, using examples from WANLI.

Prompt InstructGPT:
" Write pairs of sentences that are *related to each other in the same way* " ➥ get 5 continuations

Ambient Dataset

➥ generating disambiguations via prompting:

Ask model ***to restate ambiguous sentences with additional context that directly affirms or negates the hypothesis***.

P: *He always ignores his mother's advice to follow his own dreams.*
H: *He follows his dreams.*

ChatGPT disambiguates P:
[P] "***and therefore does follow his dreams***"
versus
[P] "***and therefore does not follow his dreams***"

Instruction	In each example, you will be given some **context** and a claim, where the correctness of the claim is affected by some ambiguity in the **context**. Enumerate two or three interpretations of the **context** that lead to different judgments about the claim.
Example	**Context**: `{premise}` **Claim**: `{hypothesis}` Given the context alone, is this claim **true**, **false**, or **inconclusive**? We don't know, because the **context** can be interpreted in many different ways: 1. `{disambiguation 1}` Then the claim is **true**. 2. `{disambiguation 2}` Then the claim is **false**. 3. `{disambiguation 3}` Then the claim is **inconclusive**.

Ambiguity in political claims

Political claim (premise)	Generated paraphrase (hypothesis)	Rating	Prediction	Explanation of ambiguity (ours)
When President Obama was elected, the market crashed...	The stock market reacted immediately to President Obama's election in 2008, ...	Barely-true	{ENTAIL, NEUTRAL}	The claim implies a causal relationship
Rhode Island is "almost dead last"... in the length of time first-degree murderers must spend in prison before they're eligible for parole.	Rhode Island is one of the states... where murderers must spend the longest time in prison before being eligible for parole.	True	{ENTAIL, NEUTRAL, CONTRADICT}	"*dead last*" may mean shortest or longest, depending on stance
Donald Trump even said, on his very first day in office, he would require every school in America to let people carry guns into our classrooms.	Donald Trump said on his first day in office that every school in America would have to allow people to carry guns in classrooms.	True	{ENTAIL, NEUTRAL}	"*on his first day*" may describe either the *saying* or the *requiring*

Identifying types of ambiguity
attachment in semantic parse; cause relation; …

Project 2: Possible Project Aims

1. Understand whether LLMs are *aware* of linguistic ambiguity, and *what knowledge they need to resolve them*

Methods

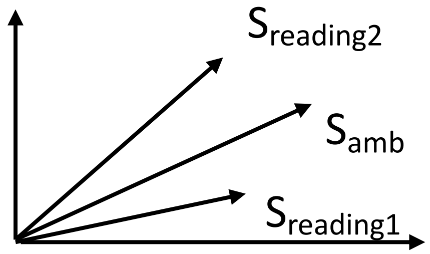
- prompt models to generate explanations for specific readings (+ evaluate against ground truth from dataset)
- **on failure**: try *in-context-learning* or *chain-of-thought prompting*
- **on failure**: *retrieve relevant knowledge from appropriate knowledge resources*
  - structured: ConceptNet / ATOMIC / GLUCOSE / DBpedia
  - textual: Wikipedia, textual CSK knowledge resources
  - Possible Datasets: WinoWhy → WinoGrande

Possible Project Aims

2. How do LLMs represent ambiguous readings?

Analyze model representations and biases using fine-grained metrics

Methods



- Discriminate readings
- Construct sentence embeddings for each reading
- Compute similarities: $sim(S_{amb}, S_{r1})$; $sim(S_{amb} - S_{r2})$; $sim(S_{r1} - S_{r2})$
- Evaluate model decisions:
  - Is the model biased? Does it suffer from insufficient knowledge?
  - To what extent can appropriate contexts resolve ambiguity in LLMs?
  - Ask models to generate explanations for their interpretation
- Datasets: WinoGender, WinoGrande, AMBIENT

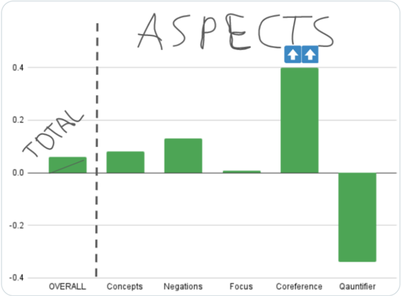
Methods

WinoGrande / WinoGender examples

Amb: The trophy would not fit in the suitcase because it was too [big/small].

➥ R1: The trophy would not fit in the suitcase because the trophy was too big.

➥ R2: The trophy would not fit in the suitcase because the suitcase was too big.



Sentence similarity

- Unstructured: SBERT, BERTScore
- Structured S3BERT: Opitz & Frank 2022: [SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features](), TACL.

Project 2 References

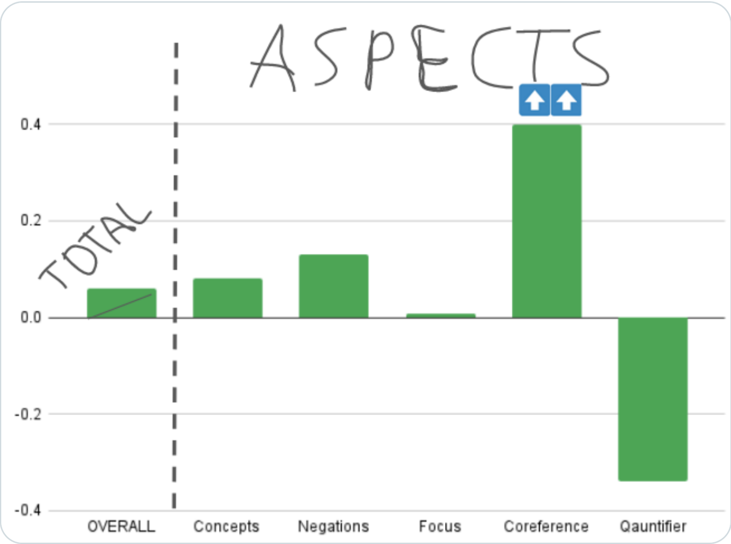
About Winograd Schemata

- Winograd Schema Challenge
- The Defeat of the Winograd Schema Challenge (big review)

Approaches

- Addressing the Winograd Schema Challenge as a Sequence Ranking Task, 2018
- A Simple Method for Commonsense Reasoning, 2018
- A Surprisingly Robust Trick for the Winograd Schema Challenge, 2019

Project 2 References: Methods



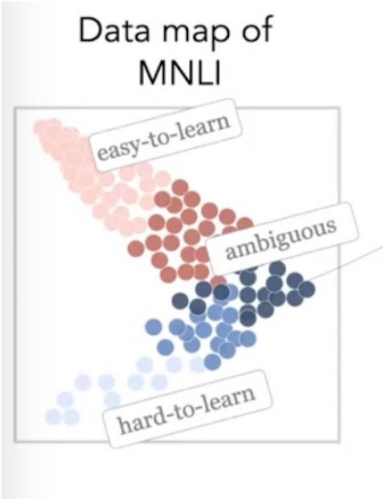
Metrics

Sentence similarity

- Unstructured: SBERT, BERTScore
- Structured S3BERT: Opitz & Frank 2022: SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features, TACL.

Evaluation

- Swayamdipta et al 2020: Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics, EMNLP



Project 2 References

Task Datasets

- WinoGrande: an adversarial winograd schema challenge at scale
- Gender Bias in Coreference Resolution
- A Balanced Corpus of Gendered Ambiguous Pronouns [Dataset]
- Multilingual: Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution
- Visual: Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality [data]
- Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality [data]

Datasets with explanations

- WinoLogic: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge
- Few-Shot Out-of-Domain Transfer Learning of Natural Language Explanations in a Label-Abundant Setup
- WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge [Video]