I hea- umm think that's what they say: A Dataset of Inferences from Natural Language Dialogues

Adam Ek[1], **Bill Noble**[1], **Stergios Chatzikyriakidis**[3], **Robin Cooper**[1], **Simon Dobnik**[1], **Eleni Gregoromichelaki**[1], **Christine Howes**[1], **Staffan Larsson**[2], **Vladislav Maraev**[1], **Gregory Mills**[4], **and Gijs Wijnholds**[5]

[1]University of Gothenburg first.last@gu.se; [2]first.last@ling.gu.se
[3]University of Crete stergios.chatzikyriakidis@uoc.gr
[4]Kingston University g.mills@kingston.ac.uk
[5]Leiden University g.j.wijnholds@liacs.leidenuniv.nl

Abstract

In this paper we describe a dataset for Natural Language Inference in the dialogue domain and present several baseline models that predict whether a given hypothesis can be inferred from the dialogue. We describe an approach for collecting hypotheses in the ENTAILMENT, CONTRADICTION and NEUTRAL categories, based on transcripts of natural spoken dialogue. We present the dataset and perform experiments using a flat-concatenating and a hierarchical neural network. We then compare these to baseline models that exploit lexical regularities at the utterance level. We also pre-train BERT with additional dialogue data and find that pre-training with additional data helps. Our experiments show that hierarchical models perform better when using a random split of the data, while flat-concatenation models perform better on Out-of-Domain data. Lastly, LLM prompting is performed on two models, Llama 2 and Zephyr, the former barely exceeding the baseline, while the latter showing an incremental increase in performance as context length increases.

1 Introduction

Natural Language Inference (NLI, or Textual Entailment, TE) is one of the core tasks for Natural Language Understanding (NLU) and central to NLU benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). The centrality and importance of NLI has been acknowledged early on by Cooper et al. (1996), arguing that NLI is the crux of Computational Semantics. Similarly, Bowman et al. (2015a) argue that understanding inference about entailment and contradiction, in effect the task of NLI, is an important aspect for constructing semantic representations, while on a more practical note, Nie et al. (2020) note that NLI is arguably the most canonical task in NLU.

Dialogue in particular and interactive reasoning more generally are an integral part of human language use. We, among others (Bender and Koller, 2020; Dobnik et al., 2021), believe that if we want to understand meaning in language we need to adapt language systems which attempt to understand language to language's central domain, namely spoken dialogue. However, there have only been a few attempts to combine dialogue and natural language inference. In this paper we outline our efforts to combine NLI and dialogue by reporting how we constructed a procedure for eliciting inference examples from dialogue data. Essentially, we take transcribed speech from naturally occurring dialogues and ask annotators to write hypotheses sentences with different inferential statuses based on the beliefs of the speakers. The final dataset contains examples like the following:[1]

A	so do you
B	I mean yeah but it wasn't that how many years ago was that ? eight years ?
A	oh when we graduated just six years ago wasn't it ?
B	yeah
A	two thousand and eight
HYPOTHESIS	they graduated last year
LABEL	**Contradiction**

Typically, information conveyed by a speaker is not limited to one turn only, but is spread out over several turns with other speakers asking clarification questions, expressing agreement and so on. Then the meaning of a turn can be considered as a joint construal of the interlocutors (Clark, 1996). A consequence of this for NLP systems is that meaning cannot be assigned to utterances independently of the dialogue history.

Thus, modelling dialogue involves both forming a representation of what has been said in the dialogue, and incorporating new utterances into this representation. By only looking at individual turns or utterances in a dialogue we are excluding the information conveyed by the *interaction* between the participants.

[1]The dataset is available at github.com/GU-CLASP/DNLI.

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

Another aspect of meaning in dialogue is that different speakers can have different interpretations given the same dialogue. Thus, when modeling dialogue and in particular multi-party dialogues, it becomes important to consider *whose* perspective we are modeling. The task of performing inference on dialogue examples presents models with a interesting set of challenges, not only does the model have to predict a label given a hypothesis, but also construct, or identify, a set of utterances that supports the hypothesis, both based on the semantic content expressed and pragmatic actions (speech acts etc.).

In sum, we present a dataset that contains natural language inference examples in the dialogue domain, named DNLI. The examples in our dataset differ crucially in at least three respects compared to existing NLI datasets that contain dialogue data: (1) a piece of dialogue can contain more than two participants (up to four), (2) a speaker may produce many utterances in one *turn* or core information may be spread out over several turns, and (3) the turns and utterances themselves might contain disfluencies like hesitations and also commonly found dialogue phenomena like repairs, split utterances and so on (Schegloff et al., 1977; Lerner, 1991; Purver et al., 2018).

2 Related work

The common ground-annotated dataset of Markowska et al. is the most similar work to ours to date. Dialogues from the CallHome dataset are annotated on the utterance level for (1) propositions that are introduced by the utterance and (2) the status of those propositions with respect to the common ground of the two speakers. By taking propositions that are considered common ground by both speakers at a given point in the dialogue, one could produce dialogue contexts and entailments along the lines of what is proposed in this paper. However, their dataset is much smaller (561 utterances), making it less suitable for machine learning. Moreover, our dataset also includes hypotheses labeled as *contradiction* and *neutral* with respect to the context, which is important for robustly training and assessing an inference model.

The MNLI dataset (Williams et al., 2018), which is a multi-genre NLI dataset, includes some examples that can be classified as dialogue—a little over a fifth of the examples are drawn from transcripts of telephone calls from the Switchboard corpus (Godfrey et al., 1992). However, none of the important characteristics of dialogue, which may influence (e.g. disfluencies, split utterances, repairs, interactivity, incrementality, or turn-taking) the beliefs of the speakers, are taken into account. Rather, the examples seem to function in exactly the same way as typical NLI cases: a single sentence from the source data is paired with a hypothesis generated by annotators. This is to be expected as the aim of MNLI was to include a multiplicity of text genres, rather than dealing with the intricacies of reasoning in dialogue settings specifically.

Welleck et al. (2019) presents an NLI dataset based on Persona-Chat (Zhang et al.). It is also referred to as Dialogue NLI, but has substantially different aims from the present work. The dataset consists of premise-hypothesis pairs, where the hypothesis is drawn from a set of *persona sentences* (facts about the speaker) and the premise is either a persona sentence or an utterance from the dialogue. The dataset seeks to improve the ability of chit-chat dialogue models to generate utterances consistent with the "persona" of the agent. In contrast, our dataset is interested in the ability to keep track of what is entailed by the dialogue itself, which requires reasoning over a dialogue context that includes multiple utterances. Moreover, our dataset is based on naturally-occurring transcribed face-to-face interaction, whereas Persona-Chat consists of text chat conversations between crowd workers play-acting as an assigned persona.

Khanuja et al. (2020) introduce a dataset for Natural Language Inference (NLI) from code-mixed Hindi-English conversations of Bollywood movies. It is comprised of 400 premises and 2240 hypotheses, annotated by Hindi-English bilinguals.The paper evaluates the dataset using an mBERT-based pipeline, revealing that existing multilingual models are not yet competent in handling code-mixed NLI tasks. Again, this is a different setup than ours, involving single premise-hypothesis examples, and does not require reasoning over a dialogue context that includes multiple utterances.

There exist a number of other dialogue and discourse datasets which might be helpful for natural language inference tasks. Many such dataset are summarised within the ParlAI (Miller et al., 2017) tool. In particular question-answering datasets may be relevant for the dialogue inference task, although they typically do not involve answering questions about the dialogue itself (e.g., Antol et al., (VQA)). While it is not dialogue, The bAbI (Weston et al., 2015) is another relevant question-answering dataset. Inputs consist of a sequence of statements representing an emergent context, followed by a question about the context. Paperno et al. (2016) put together the LAMBADA dataset, where context is comprised of a list of passages (including dialogical exchanges) and the task is to predict the last word of a target sentence which follows from the context.

3 Dialogue NLI

Typically, a Natural Language Inference example contains a *premise* statement and a *hypothesis* statement. Then, the task is to determine whether the hypothesis is *entailed* or not from the premise. That is, does the premise contain sufficient evidence to determine whether the hypothesis is true or not.

In our dataset we cast the premise as a continuous sequence of *utterances* from a dialogue. The hypothesis is a statement that one of the dialogue participants

would take to be true, false or neither true nor false. In particular we take the perspective of the speaker who most recently produced an utterance when evaluating the hypothesis. This is because the common-ground of different dialogue interlocutors may diverge without being acknowledged by the participants, but an outside observer could potentially observe this.[2]

To elicit statements about a particular speaker's point of view, we ask annotators to produce a statement which one of the speakers make a judgement about, based on what has been said in the dialogue. We follow Bowman et al. (2015a); Williams et al. (2018) and consider three types of inference judgements: TRUE (ENTAILMENT), FALSE (CONTRADICTION) or NEITHER (NEUTRAL), presented as follows to our annotators:

ENTAILMENT: A statement that the last speaker would take to be true at this point in the dialogue.

NEUTRAL: A statement for which there is no evidence that the last speaker would take it to be true or false at this point in the dialogue.

CONTRADICTION: A statement that the last speaker would take to not be true at this point in the dialogue.

Thus, each hypothesis is based on a particular point in time. As such, we can't know whether a participant would make the same judgement again if the dialogue continues, as new information can be expressed in the dialogue which may change what the participants believes. As an example, let us consider a dialogue whose hypothesis involves what the topic is being discussed. The hypotheses "they are talking about wine" will only be an entailment as long as they continue talking about wine. However, if the topic changes from "wine" to "saunas" as in the dialogue below, the previous hypothesis will no longer be an ENTAILMENT but a CONTRADICTION (since they are *not* talking about wine anymore).

D	so what was the conclusion with the wine thing should you pour it? is it
	...
HYPOTHESIS	they are talking about wine
LABEL	**Entailment**
A	I mean it does alter the taste
	...
C	I'd much prefer sitting in a sauna nice and dry and hot

To summarize, we consider a sequence of utterances $u_0^{S_i} ... u_n^{S_j}$ to be the premise, and a hypothesis h. The label of the hypothesis is based on the beliefs of speaker S_j (the one who uttered u_n) when u_n was uttered.

[2] A project that explores this is Ghosal et al. (2021), who considers dyadic dialogues and what common-sense inferences that one can draw from those.

4 Data collection

In this section we describe the corpora used to create the Dialogue Natural Language Inference dataset and the way annotations were elicited.

4.1 Dialogue Corpora

Our corpus of annotated dialogues draws from the BNC2014 (Love et al., 2017)[3] and the CHILDES (MacWhinney, 2000) corpora. The corpus contains 13,856 annotations distributed over 938 dialogues from the BNC data and 287 annotations on 17 dialogues from the CHILDES corpus.[4]

CHILDES is a collection of corpora of transcribed spontaneous conversations, mostly between children and their adult caregivers. We draw from dialogues in the Warren-Leubecker and Iii (1984) portion of the corpus, which is comprised of conversations between English speaking two- and five-year olds from suburban Atlanta and their parents. As CHILDES is a widely used resource in both the child language acquisition and computational modeling communities, a number of other annotation resources are available, including morphological and syntactic annotations (Buttery and Korhonen, 2005; Sagae et al., 2010; Villavicencio et al., 2012) and utterance-level semantic annotations (Bergey et al., 2021). The Warren-Leubecker and Iii (1984) portion of the corpus also includes intonation annotations.

BNC (Love et al., 2017) is a follow-up to the 1994 version of the BNC, comprised of conversations between adult native speakers of British English. A key component of the BNC dataset which makes it particularly interesting for NLI is that it is naturally occurring speech that has been annotated faithfully, such that repairs, disfluencies, and so on are included in the data. The dialogues in the dataset have 2-4 participants making it interesting for NLI as a model potentially has to learn 2-4 different belief representations if their beliefs diverge. The dialogues on the BNC dataset are also long and touch on many different topics (even within one dialogue). Thus, to successfully model these dialogues, a system must be able to handle that participants changing topics, or talk about different topics in the dialogues. Moreover, because the dialogues are naturally-occurring, a model must also learn to accurately model an open-ended range of topics (Chen and Gao, 2017; Shalyminov et al., 2020). This is in contrast to many other dialogue datasets such as MultiWoZ (Budzianowski et al., 2018) in which topics are constrained to a pre-determined set of task-oriented scenarios.

One major argument for including both adult-adult and adult-child dialogues is that in real-life dialogues, participants can not always assume symmetry in the

[3] Henceforth, BNC.

[4] An *annotation* refers to a hypothesis of one of the three labels (ENTAILMENT, CONTRADICTION, NEUTRAL) elicited from an annotator. See §4.3 for details.

linguistic resources of their interlocutors. Thus, there will be cases when a model is forced to interpret a speakers intention, even when it is not expressed in the most conventional or explicit way, just as a human would have to when speaking to a child.[5]

We believe the linguistic diversity of speakers is a key aspect of dialogue often neglected in dialogue research. The use-cases for dialogue systems often involve a *random person* talking to someone. This becomes very problematic when we consider the case of a child interacting with some QA system. If the system is developed with data collected from adult speakers only, it won't be able to take into account linguistic variations associated with children.

4.2 Dialogue formatting

The data from both the BNC and CHILDES corpus contain longer dialogues, with up to 15000 utterances in a dialogue (but about 900 on average). This presents a problem as we would like to give the annotators a dialogue that can easily be read and subsequently annotated. To make the annotation process feasible, we split each dialogue into n sub-dialogues, where each sub-dialogue contains around 50 utterances each. For each sub-dialogue we select $1-5$ utterances at random and insert an annotation flag, as a constraint to this we do not allow for consecutive annotation flags they must be separated by more than two utterances. Then for each annotation flag we elicit an annotation.

4.3 Annotations

In the annotation process we utilize both Amazon Mechanical Turk workers and Master students in the Language Technology program at the University of Gothenburg We noted that the task was difficult for AMT workers, which prompted us to manually go through all the AMT annotations and select the reasonable ones. The masters students were hired on an hourly basis and got paid 12 USD per hour. The AMT workers were paid around 3 USD per dialogue.

We created an online annotation tool that the workers used to do their annotationk. On the web page, the dialogue is presented incrementally, such that the annotators have the same information as the participants in the dialogue. So neither the annotators nor the dialogue participants are able to see future utterances.

We ask the annotators to take the perspective of the *last speaker* and freely write a hypothesis statement conforming to one of the three labels: *true*, *false* or *unknown*. Our approach to NLI is similar to other large scale datasets, such as (Bowman et al., 2015b; Williams et al., 2018; Khot et al., 2018), where the logical constraints have been relaxed.[6]

[5]Two speakers interacting NEVER share the 'same' language, so the model has to be able to deal with asymmetries. Child-adult conversations or dialogues between native and non-native speakers are just obvious examples of this (Clark, 1998).

[6]It is important to note that the notion of entailment is not uniform across all NLI datasets. An overview can be found in (Chatzikyriakidis et al., 2017; Bernardy and Chatzikyriakidis, 2019; Poliak, 2020).

B	see you in a year
A	so what do we do like what do I do if with the birthday card? can I send it to you? like will you have an address?
D	what birthday card?
B	yours
A	well you'll be away for your birthday
B	yeah
D	no don't bother
HYPOTHESIS	Speaker D don't want birthday cards
LABEL	**Entailment**

One core feature of this dataset is that for a model to accurately predict the label of a statement, the model must compose the information given over several turns and take into account a speaker's perspective which will be different depending on who the speaker is. This will be affected by the usage of pronouns and spatial perspective but also other facts about the speaker and their role in the conversation. In the example above, the model must infer that when speaker D says *"no don't bother"* they are referring to the giving of a birthday card which was proposed by A five turns earlier.

During the annotation process only simple instructions on how to refer to participants were given, resulting in a variety of strategies. We extracted these using a simple regular expression to get an idea how the annotators did this, as shown in Table 1. Primarily, we observe two ways of referring to a specific speaker, "Speaker X" or "Person X", additionally pronouns such as "they" were used often. We also note that "he" and "she" occur but much less frequently. It is often not clear from the dialogue alone which gender a speaker is (exemplified in the above dialogue). There are slightly more occurrences of male over female pronouns.

Referring expr.	Count
Speaker X	2963
Person X	2193
They	2152
He	240
She	229
Her	20
His	26

Table 1: Expressions used by annotators to refer to speakers in the dialogue.

5 Data analysis

In this section we describe some descriptive statistics of the dataset. The distribution of labels shown in Table 2, we see that the labels are distributed roughly evenly across the labels, slightly favoring Entailment.

Label	Count	Proportion
Entailment	4799	0.338
Contradiction	4677	0.329
Neutral	4723	0.333

Table 2: Distribution of labels in the dataset.

In total, we collected 14 179 hypotheses from the dialogues in our corpora and show the number of dialogues from each corpora in Table 3.

Source	Dialogues	Annotations
BNC	938	13 892
CHILDES	17	287

Table 3: Number of dialogues and annotations from BNC and CHILDES.

We collect data from 955 dialogues in total, where 938 of the dialogues are from BNC are 17 are from the CHILDES. In the BNC portion of the dataset there are 13 892 hypotheses annotated and in the CHILDES portion 287 hypotheses annotated.

One feature of our dataset is that some dialogues have more utterances than others. For example, the shortest dialogue contains 125 utterances, while the longest one contains 15 054 utterances. As shown in Appendix A, Figure 1, the number of utterances in the dialogues follow roughly a Zipfian distribution.

During the annotation process we randomly select sub-dialogues (see Section 4.2). A consequence of this is that longer dialogues tend to receive more annotations. We show the number of annotations available for each dialogue in Appendix A, Figure 2.

Additionally, we look at the number of tokens in both the premises and hypotheses, shown in Table 4. We see in Table 4 that the number of tokens per utterance can vary a lot with a standard deviation of 7.31 tokens and a mean of 6.05 tokens.

	Mean	STD
Tokens/Utterance	6.05	7.31
Tokens/Hypothesis	8.13	3.14

Table 4: Distribution of tokens in the dataset.

This is caused by for example utterances only containing back-channels and disfluencies. In general, this poses an interesting problem for models that also occur in real-life: namely to select the utterances that provide useful information to some belief of a speaker. For the number of tokens per hypothesis this contain less variation, but they tend to be longer than the premises.

5.1 Data splits

We provide a split of the data into a standard split, with the following data distribution: 80% training, 10% validation and 10% testing, ensuring that the label distribution is roughly uniform between the different data splits. Additionally, because the BNC corpora is larger and we elicited more annotations from this dataset we ensure that the validation and test splits contain more than two dialogues from the CHILDES corpora.

Another split we consider is an Out-of-Domain split (Zheng et al., 2020; Haddow and Koehn, 2012), where training and development data is randomly sampled from the BNC and the test data taken from CHILDES. This type of splitting allows us to estimate how much we can learn about dialogues regardless of domain (chit-chat versus more task-oriented dialogues). Another feature of this type of split is that we can evaluate how dialogues between adults transfer to dialogues between caregivers and children. As we have mentioned earlier, dialogues occur between different types of people and systems of dialogue need to handle this.

For reproducibility we perform all experiments in this paper with standard and Out-of-Domain split, but encourage future work to explore other data splits (see Gorman and Bedrick, 2019; Søgaard et al., 2020).

6 Experiments & Results

We perform experiments both on the standard split and Out-of-Domain split, and investigate the performance of two model architectures: flat-concatenation (Smith et al., 2020; Zhang et al., 2020; Li et al., 2021) and hierarchical (Serban et al., 2016; Tran et al., 2017).

In the flat-concatenation architecture the utterances preceding an annotation are concatenated together and fed to the model as one sequence. We apply max pooling over the sequence to get a dialogue representation. In the hierarchical architecture we consider two levels of representation: a token level representation where the tokens in each utterance are encoded (and as in the case of flat-concatenation, we use max pooling), and an utterances level representation where the representation from the token level are modeled. To get a dialogue representation D we use additive attention

$$D = \mathsf{softmax}(w^T \tanh(W_a^T k + W_b^T u))\, u$$

where u is the utterance representation and k the max-pooled hypothesis representations. We experimented with other ways of compiling this information (dot-product attention, self-attention, last hidden state, max/mean pooling) but found that additive attention yielded the best performance. The main idea is that tokens and utterances are distinct units of information, as such it could be beneficial to model these in a hierarchical fashion. An overview of the hierarchical architecture we employ is given in Figure 1.

For all architectures we model the interaction between the premise and hypothesis representations by concatenating u, h, $|u - h|$ and $u \odot h$ (element-wise multiplication) (Conneau et al., 2017), where u is the premise and h is the hypothesis. An overview of this procedure is given in Figure 2.

All experiments were conducted on a RTX Titan 12GB card with a batch size of 4 over 20 epochs with

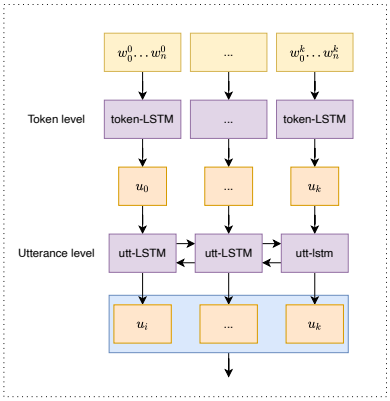


Figure 1: Overview of the dialogue encoder in the hierarchical architecture, where $w_0^k ... w_0^n$ are the tokens of utterance 0, and $u_0 ... u_k$ the utterance representation.

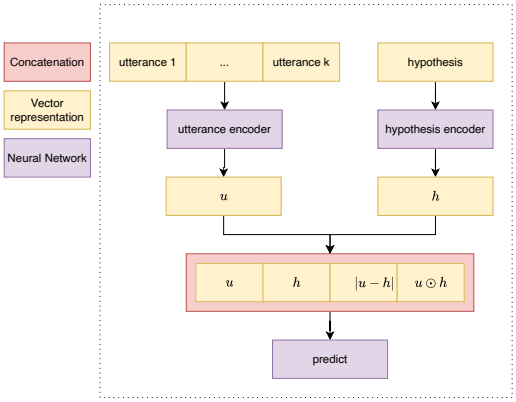


Figure 2: Base architecture of our dialogue NLI models. We encode the utterances and hypothesis separately, then before we predict a label we learn how the two representations interact by concatenating u, h, $|u - h|$ and $u \odot h$ (element-wise multiplication).

early stopping (two epochs of no improvement on the dev data). We use the Adam (Kingma and Ba, 2015) optimizer with default parameters, a Cosine Annealing learning-rate scheduler (Loshchilov and Hutter, 2017), with an initial learning rate of $1e{-}3$ and a minimal learning rate of $1e{-}7$, and weight decay of 0.01. For both architectures, we experiment using a transformer model (`bert-base-uncased`) [7] or a LSTM as the utterance and hypothesis encoders.

As annotators had limited access to context we explore how many utterances to use as the premise. Too much utterance context could introduce noise and too little could miss the utterances where relevant information is expressed. To investigate this question we perform our experiments using different numbers of utterances as context, namely contexts of size $1, 3, 5, 7, 9, 11, 13$ and 15. The performance of our models on the standard and Out-of-Domain split is given in Figure 3 and Figure 4 respectively.

[7] We also performed experiments with GPT2, but found no clear difference compared to BERT.

In the experiment on the standard data split we can observe than for context $1, 3$, and 5 the BERT based models (both flat-concatenation and hierarchical) perform slightly better than LSTM based approaches and in all other cases outperform the LSTM based approach. As we increase the context (and thus information available to the model), BERT-based approaches start to perform better than the LSTM approach, and the outlier here is for context 13, where the flat-concatenation BERT model performs roughly the same as hierarchical LSTMs. We see clearly that hierarchical BERT is most effective with 5 or more utterances as the premise, where we get a substantial increase in performance. The performance of the flat-concatenation model varies across contexts, only out performing the hierarchical BERT with contexts sizes of $1, 3, 9$ and 11. For the LSTM models there is a clear preference for the hierarchical architecture.

In the Out-of-Domain split we see a lower performance across setups. However, another pattern appears, namely that the flat-concatenation models perform better than the hierarchical models.

6.1 Baselines

We consider a number of baselines whose primary aim is to probe biases in the data and explore how far we can get without actually modelling real dialogues. We consider the majority class as one of these baselines, and the hypothesis-only baseline. In the hypothesis-only approach we simply try to predict the label based on the hypothesis and *not* the premise (i.e. the dialogue utterances) (Poliak et al., 2018). This baseline probes for biases in the hypothesis statements associated with different labels. For example, if the word “not" occurs in every contradiction, the model will likely learn to exploit that regularity in the hypothesis rather than modelling the relationship between premise and hypothesis. The performance of the baselines is shown in Table 5.

Model	Standard Split	Out-of-Domain
Majority Class	33.8	35.5
LSTM Hyp. only	51.3 ± 0.4	42.4 ± 0.2
BERT Hyp. only	58.9 ± 0.9	44.4 ± 0.4

Table 5: Baseline performance on the standard split and for training on BNC and testing on CHILDES (Out-of-Domain).

The majority class baseline reveals that the labels in both the standard split and Out-of-Domain data is more-or-less balanced, with the Out-of-Domain data showing a slightly higher bias to neutral hypotheses.

The hypothesis-only model does perform better than the majority class baseline, suggesting that there is some bias in the hypothesis statements for the models to exploit. A BERT-based approach to this baseline yields a higher accuracy of 58.9% versus 51.3% for the LSTM. Interestingly, the performance of the hypothesis only baseline is lower relative to the ma-

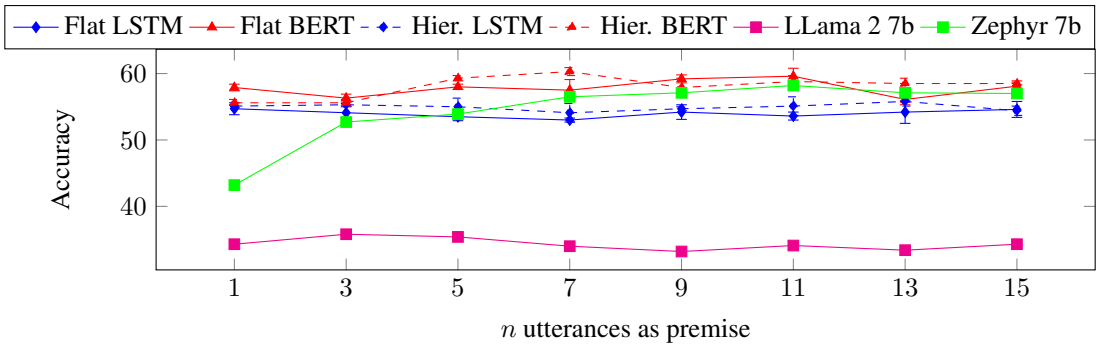


Figure 3: Mean accuracy and standard deviation over three runs on the standard split. We consider both a LSTM and a BERT-based approach. Additionally, we show the performance of the Llama 2 7b, which was prompted with three examples from the training set.

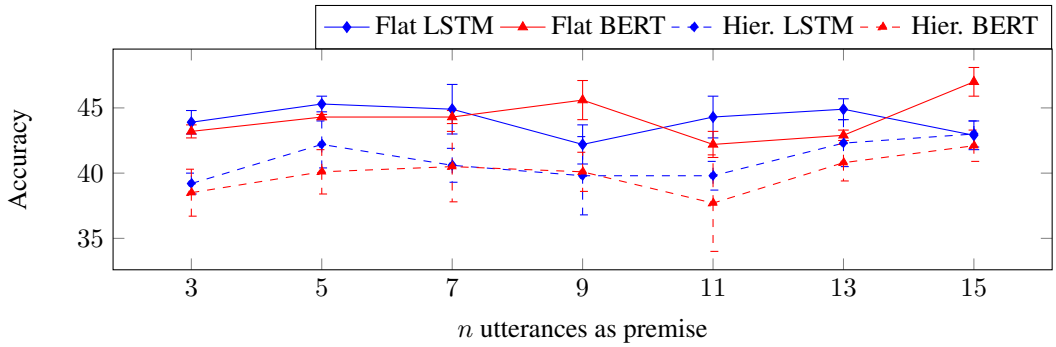


Figure 4: Mean accuracy and standard deviation over three runs on the Out-of-Domain test data. We consider both a LSTM and a BERT-based approach.

jority class for the Out-of-Domain data (8.6% for the LSTM vs 25.1% for the BERT hypothesis only). Thus, it seems that the hypotheses generated from CHILDES dialogues are less biased with respect to the label.

6.2 Dialogue pre-training for utterance encoders

While there is widespread evidence that large language models improve performance on Out-of-Domain tasks, models such as BERT, trained on text-only corpora, may have trouble representing features specific to spoken dialogue. Indeed, there is evidence that these models require fine-tuning to perform well on dialogue-specific tasks and that additional pre-training in the dialogue domain can be helpful (Noble and Maraev, 2021). For that reason, we experiment with a BERT utterance encoder that has been additionally pre-trained on in-domain data. In particular, we create a corpus from the dialogues in the spoken section of the BNC which were not included in the DNLI dataset. This amounts to 1,252 dialogues and 1,119,747 utterances (about 30% larger than the DNLI dataset). To assess the contribution of the original BERT pre-training, we train two BERT models: one with randomly-initialized parameters (BERT RandomInit), and one initialized with the standard pre-trained BERT-base parameters (i.e., the BERT model that is used in other experiments).

In each case, the model was trained with a masked language modelling objective (Devlin et al., 2018) over 100 epochs on the BNC pre-training corpus described above. Models were trained with a batch size of 64, though gradients were accumulated every 8 batches, making the effective batch size 512. We took the model from the epoch with lowest development loss.[8] These were 0.16 (epoch 98) and 0.53 (epoch 85) for pre-trained and from-scratch BERT models, respectively. These results suggest that BERT is able to leverage its text pre-training in the masked language modelling objective, but it remains to be seen whether the text pre-training is useful for the downstream inference task.

When applied to the DNLI dataset we see that the regular pre-training of BERT appears to be helpful, as downstream performance of BERT RandomInit consistently decreases, both for the standard and Out-of-Domain data splits. We also see that the pre-training of BERT using BNC improves the performance of the flat concatenation model while it decrease the performance of the hierarchical model, as shown in Table 6. We argue that the flat-concatenation model, operating on the token-level only, has a closer connection to the pretraining objective during finetuning and thus can more easily exploit the dialogue information provided by BNC, unlike the hierarchical model that operates on the utterance-level.

[8]The utterances from the validation split of the DNLI dataset were used for development.

	Standard Split		Out-of-Domain	
	Flat	Hierarchical	Flat	Hierarchical
BERT RandomInit	55.4(−2.1)	54.6(−5.7)	42.5(−1.8)	41.8(−3.1)
BERT+BNC	58.5(+1.0)	58.4(−1.9)	47.4(+3.1)	43.6(−1.3)

Table 6: Performance with BERT trained on only BNC (BERT RandomInit) or with BNC as additional pre-training (BERT+BNC), with 7 utterances as context on our DNLI dataset, the difference from standard BERT in terms of percentage points is shown within parenthesis.

6.3 LLM prompting

We also report the performance of two large language models (LLMs), Llama 2 7b (Touvron et al., 2023)[9] and Zephyr 7b (Tunstall et al., 2023)[10]. In both cases, the model was provided with a short prompt describing the task and three examples from the training set (see Appendix B for an example). The model's generation was constrained to one of the three target labels[11]. An example prompt in shown in the appendix. We observe a strong difference in results; where LLama 2 barely beats the majority class baseline, Zephyr displays a largely continual increase in performance as the context length increases. Overall it reaches performance slightly below the hierarchical BERT-based model.

7 Discussion

7.1 Annotations

During the annotation process we noted that several different strategies were used to refer to speakers, and sometimes pronouns or other referring expressions were used in the hypotheses. To know if a certain hypothesis is true or not requires anaphora resolution. Then, should this responsibility be placed upon the annotators or the models? We would like a model to disambiguate referring expression as it is a part of language use. But the hypotheses are written in a "meta-language", that describe beliefs of a speaker, and the question is: do we actually want the model to learn this meta-language or not? The goal of our dataset is to allow for dialogue understanding and how meaning is obtained, and then the task of disambiguating referring expressions is redundant. But we would also like systems to do this, so they can function in a real-world application, where meta-language does occur. In the dataset we put the burden of disambiguation on the models rather than the annotators.

7.2 Experiments

In our experiments, we observe that the hypothesis-only baseline is outperformed by a full model in LSTM-based approaches, but the converse occurs when using BERT.

Given that hypotheses are generally framed meta-linguistically, we are in fact already modeling two domains, the dialogue and the meta-language. So the question is, why doesn't dialogue context always help? As mentioned earlier, one issue may be that the model has to model two domains and also disambiguate referring expressions. Another issue is dialogue phenomena such as repairs, disfluencies and split-turns, can be interpreted as noise by BERT. To properly use these features of dialogue systems must recognize that they serve a *pragmatic function*, that a dialogue is a joint effort of two or more dialogue participants. If a model can not do this, utterances such as "umm" or repairs, that provide no clear semantic meaning will be modelled improperly. This may be one of the failings of the hierarchical model as each utterance, however small, gets a representation.

Among the large language models, we see that Zephyr greatly out-performs Llama 2 and that Zephyr is able to take advantage of longer context windows. This may be a result the fact that Zephyr was fine-tuned as a chat model using Direct Preference Optimization.

8 Conclusion

We have presented our data collection process for a dataset of natural language inference in dialogues, the first of its kind that uses both natural dialogues and hand-annotated hypotheses. We performed experiments using LSTM, BERT and prompted LLM baselines. The dataset is hard to model properly as shown by our experiments, where the best performance we obtained was about 2% better (Figure 3 with 7 utterances as context for hierarchical BERT) than a hypothesis-only baseline. It is unclear if the LSTM and BERT models are able to recognize pragmatic functions of dialogue phenomena such as disfluencies and back-channels, or make sense of split utterances or repairs. We believe this dataset fills a gap for both dialogue systems and natural language inference systems, presenting a challenging dataset in both research directions.

Future work includes collecting additional annotations, such as paraphrases or meaning-reversing modification of the current annotations. Additionally, we plan to continue annotating the BNC corpora to achieve full coverage. Another avenue in this direction is to explore how models deal with dialogue phenomena that serve a pragmatic function such as back-channels and disfluencies, and how to properly model these with neural networks.

[9] We use the AWQ-quantized version available from TheBloke/Llama-2-7B-AWQ on HuggingFace.

[10] Available from HuggingFaceH4/zephyr-7b-beta. We use the 4-bit bitsandbytes quantization configuration.

[11] We employ constrained decoding from the guidance library: https://github.com/guidance-ai/guidance

Acknowledgments

The research detailed in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Stergios Chatzikyriakidis gratefully acknowledges funding from the Special Account for Research Funding of the Technical University of Crete (grant number: 11218), as well as funding from the TALOS-AI4SSH ERA Chair in Artificial Intelligence for Humanities and Social Sciences grant (grant agreement: 101087269).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Claire Bergey, Zoe Marshall, Simon DeDeo, and Daniel Yurovsky. 2021. Learning communicative acts in children's conversations: A Hidden Topic Markov Model analysis of the CHILDES corpus.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *ICAART (2)*, pages 919–931.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Paula Buttery and Anna Korhonen. 2005. Large Scale Analysis of Verb Subcategorization differences between Child Directed Speech and Adult Speech. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, and Staffan Larsson. 2017. An overview of natural language inference data collection: The way forward? In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics, Workshop on Computing Natural Language Inference*, pages 1–6, Montpellier, France. Association for Computational Linguistics.

Yun-Nung Chen and Jianfeng Gao. 2017. Open-domain neural dialogue systems. In *Proceedings of the IJCNLP 2017, Tutorial Abstracts*, pages 6–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press.

Herbert H Clark. 1998. Communal lexicons. In Kirsten Malmkjær and John Williams, editors, *Context in Language Learning and Language Understanding*, chapter 4, pages 63–87. Cambridge University Press, Cambridge.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680. Association for Computational Linguistics.

R. Cooper, D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman. 1996. Using the framework. Technical report LRE 62-051r, The FraCaS consortium. `ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz`.

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*.

Simon Dobnik, Robin Cooper, Adam Ek, Bill Noble, Staffan Larsson, Nikolai Ilinykh, Vladislav Maraev, and Vidya Somashekarappa. 2021. We went to look for meaning and all we got were these lousy representations: aspects of meaning representation for computational semantics. *Preprint*, arXiv:2109.04949.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Cider: Commonsense inference for dialogue explanation and reasoning. *arXiv preprint arXiv:2106.00510*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International*

Conference on, volume 1, pages 517–520. IEEE Computer Society.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2786–2791.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on smt systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gene H. Lerner. 1991. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. *Preprint*, arXiv:2106.02227.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed edition. Lawrence Erlbaum, Mahwah, NJ.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. Finding Common Ground: Annotating and Predicting Common Ground in Spoken Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Bill Noble and Vladislav Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics*, pages 166–172, Groningen, the Netherlands (online). Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Adam Poliak. 2020. A survey on recognizing textual entailment as an nlp evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451.

Kenji Sagae, Eric Davis, Alon Lavie, Brian Macwhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729.

E.A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Hybrid generative-retrieval transformers for dialogue domain adaptation. *arXiv preprint arXiv:2003.01680*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2020. We need to talk about random splits. *arXiv preprint arXiv:2005.00636*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arxiv:2302.13971.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, von Leandro Werra, Fourrier, Clémentine, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *Preprint*, arxiv:2310.16944.

Aline Villavicencio, Beracah Yankama, Rodrigo Wilkens, Marco Idiart, and Robert Berwick. 2012. An annotated English child language database. In *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*, pages 23–25, Avignon, France. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Amye Warren-Leubecker and John Neil Bohannon Iii. 1984. Intonation Patterns in Child-Directed Speech: Mother-Father Differences. *Child Development*, 55(4):1379.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

J. Weston, Antoine Bordes, S. Chopra, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks.

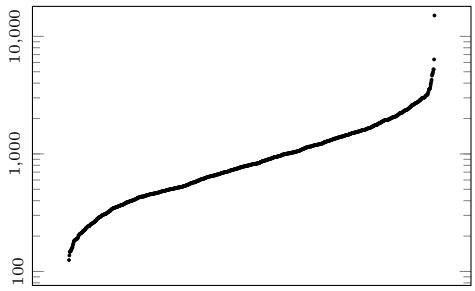
Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

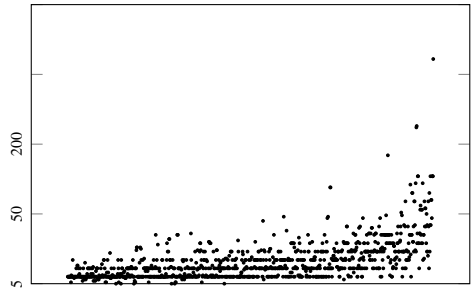
Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

A Additional descriptive statistics



Appendix Figure 1: Number of utterances in each dialogue sorted on a logarithmic scale. The y-axis indicate the number of utterances in each dialogue and the x-axis each dialogue.



Appendix Figure 2: Number of annotations in each dialogue are shown on the y-axis, and the dialogues sorted by number of utterances of the x-axis.

B LLM prompt

Given a dialogue excerpt and a Hypothesis, decide on the semantic relation between them, choosing between Entailment, Contradiction, and Neutral.

SPEAKER B: well his his his brother had cancer his brother died
SPEAKER C: did he?
SPEAKER A: and then his mum got ill
SPEAKER B: and then his mum got really ill he dropped he
SPEAKER C: mm
HYPOTHESIS: they are talking about fathers
RELATION: Contradiction
SPEAKER B: he was ex-
SPEAKER C: yeah
SPEAKER B: so it seemed to be a bit of a stopgap bit like when dad
SPEAKER A: yeah
SPEAKER C: yeah
HYPOTHESIS: they are not sure about dad
RELATION: Neutral
SPEAKER A: yeah
SPEAKER B: mm yeah yeah
SPEAKER C: but I I was paid for it and I got bonuses and everything and it was good money
SPEAKER A: yeah
SPEAKER C: in the early eighties early to mid-eighties
HYPOTHESIS: they are talking about eighties
RELATION: Entailment
SPEAKER C: yeah saucepan
SPEAKER D: yeah it should be a saucepan
SPEAKER C: small one heavy bottomed
SPEAKER A: and then like
SPEAKER D: low heat do it low heat
HYPOTHESIS: to make proper scrambled eggs, you must use a pot
RELATION: [FILL]

Figure 5: An example prompt for LLM prompting. We use three examples, one for each NLI label, [FILL] indicates the generation of the model which is constrained to the three possible labels.