Follow the Knowledge: Structural Biases and Artefacts in Knowledge Grounded Dialog Datasets

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, Walter Daelemans
CLiPS Research Center
University of Antwerp, Belgium
firstname.lastname@uantwerpen.be

Abstract

Crowd-sourcing has been one of the primary ways to curate conversational data, specially for certain scenarios like grounding in knowledge. In this setting, using online platforms like AMT, non-expert participants are hired to converse with each other, following instructions which try to guide the outcome towards the desired format. The resulting data then is used for different parts of dialog modelling like knowledge selection and response selection/generation.

In this work, we take a closer look into two of the most popular knowledge grounded dialog (KGD) datasets. Investigating potential biases and artefacts in knowledge selection labels, we observe that in many cases the 'knowledge selection flow' simply follows the order of presented knowledge pieces. In Wizard of Wikipedia (the most popular KGD dataset) we use simple content-agnostic models based on this bias to get significant knowledge selection performance. In Topical-Chat we see a similar correlation between the knowledge selection sequence and the order of entities and their segments, as provided to crowd-source workers. We believe that the observed results, question the significance and origin of the presumed dialog-level attributes like 'knowledge flow' in these crowd-sourced datasets.

1 Introduction

Since the introduction of data hungry methods into dialog modeling, sizeable datasets have become an essential asset for researchers of the field. While generic conversational data can be harvested in large quantities from already existing resources like movie subtitles (Lison and Tiedemann, 2016) or website forums (Lowe et al., 2015), more specific datasets usually need to be curated under supervision. Knowledge grounded dialog is one of the fields that has remarkably benefited from crowd-sourced datasets like Wizard of Wikipedia or WoW (Dinan et al., 2018), Holl-E (Moghe et al., 2018) and Topical-Chat (Gopalakrishnan et al., 2019), which offer grounded utterances generated by non-expert annotators or Turkers. During the curation, participants are commonly asked to first choose a knowledge piece (or no knowledge) from a provided pool and then use the selected piece to ground their next utterance on.

Method	w/o response	w/ response
Random	2.7	2.7
GRU	20.0	66.0
Transformer	22.5	70.4
BERT	23.4	78.2
Human	17.1	83.7

Table 1: The prior-posterior gap in knowledge selection for the WoW seen-test dataset (from Kim et al. (2020)). Columns show the performance (accuracy) without and with access to the grounded response.

One common attribute in these datasets is the sizeable difference between the knowledge selection performance with and without access to the uttered response. The phenomenon –referred to as the prior-posterior gap– is demonstrated in Table 1 (from Kim et al. (2020)) for the WoW dataset. Looking for ways to improve the prior performance, studies have tried to design methods to capture higher-order patterns beyond the limiting (and seemingly insufficient) turn-level scope. A natural candidate for this, is modeling the 'knowledge flow'; i.e. how the history of knowledge selection affects the next selection.

In this work we investigate the potential spurious origins of 'knowledge flow' in crowd-sourced KGD datasets. Focusing on the most popular resources in the field, i.e. WoW, we show that competitive results can be obtained in the knowledge selection task, using very simple structural heuristics. We also show that these rudimentary patterns are not an isolated case and can be found in other knowledge grounded dialog datasets like Topical-

Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 109–121
July 13, 2023 ©2023 Association for Computational Linguistics

Dataset	dialogs	utterances	Kn access	Kn gold label	Kn pool	citations
Wizard of Wikipedia	22,311	201,999	A	sentence	dynamic, multi-topic	620
CMU_DoG	4,112	130,000	S/A	-	static, single-topic	170
Holl-E	9,071	90,810	S	sentence	static, single-topic	131
Topical-Chat	9,058	198,306	S/A	section	static, multi-topic	219

Table 2: Four most popular (English) datasets for knowledge grounded conversation (Kn:knowledge, S:symmetric, A:asymmetric). Citations are from Google scholar as of April 2023.

Chat, which –in our opinion– connects knowledge selection patterns to dataset curation choices and design.

While dataset artifacts and their relation to the curation process have been widely studied in NLU tasks and especially NLI (Nangia et al., 2021; Gururangan et al., 2018), it is an under-studied topic in dialog modeling. We hope our work draws attention to the issue and contributes to having better dialog datasets, which we believe is necessary for properly modeling higher-order dialog attributes.

2 Knowledge Grounded Conversation

2.1 Problem Formulation

In general, the question of knowledge grounded dialog (KGD) modelling is defined over dialog and knowledge datasets $\mathcal{D}_d = \{(C_i, r_i)\}_{i=1}^{N}$ and $\mathcal{D}_k = \{(k_j)\}_{j=1}^{M}$ where $\forall i \in \{1, ..., N\}$, C_i and r_i represent context and response for a specific dialog turn, and $\forall j \in \{1, ..., M\}$, k_j is a knowledge piece (e.g. a sentence or paragraph). In most recent datasets, $\mathcal{D}_d$ and $\mathcal{D}_k$ are provided as parallel, which allows for a simpler formalization over $\mathcal{D} = \{(C_i, K_i, r_i)\}_{i=1}^{N}$, where K_i (or knowledge pool) is a subset of $\mathcal{D}_k$, and often includes one or more 'gold truth' (K_i^G), i.e. the knowledge piece(s) picked by the annotator during data curation.

The problem of knowledge selection (KS) in this context means designing a model f_{ks} to identify the relevant knowledge piece(s) in K_i: $f_{ks}(K_i) = K_i^G$. Ideally f_{ks} provides a ranking over K_i which can be used to retrieve top-k results for response generation.

2.2 Popular Datasets

The problem of modeling open-domain knowledge grounded conversation attracted increasing attention since the introduction and release of large scale crowd-sourced knowledge grounded dialog datasets with parallel dialog and knowledge corpora. Table 2 shows selected details of the four most popular KGD datasets: **Wizard of Wikipedia** (Dinan et al., 2018) includes conversations between a wizard (with access to knowledge) and an apprentice (no knowledge access) grounded on Wikipedia articles. **CMU_DoG** (Zhou et al., 2018) and **Holl-E** (Moghe et al., 2018) contain dialogs about movies grounded on Wikipedia information plus descriptions for 3 key scenes (CMU_DoG) or a selection of movie's plot, reviews, comments and facts (Holl-E). Finally **Topical-Chat** (Gopalakrishnan et al., 2019) includes conversations on various 'entities' and grounded in a combination of Wikipedia article, fun facts and news articles, in both symmetric and asymmetric knowledge access scenarios.

Among these, WoW is by far The most cited dataset in the field which can be attributed to qualities like proper size, gold knowledge labels and multi-topic knowledge pool. It is also the only dataset with dynamic pool, meaning that the knowledge choices are updated at each turn. Topical-Chat is another popular resource which creates distinction with pre-defined scenarios for knowledge access between collocutors. However it only provides section-level (and not sentence-level) labels for knowledge selection, which makes it less convenient for supervised knowledge selection modeling.

2.3 Knowledge Selection Methods

The popular approach of breaking the KGD problem into the knowledge selection (KS) and response generation (RG) tasks, became mainstream with WoW. Along with the dataset, the release paper (Dinan et al., 2018) also proposed a baseline model (Transformer MemNet) which addressed KGD in these two steps, acquiring 22.5% and 12.2% accuracy for knowledge selection on the seen and unseen test sets accordingly[1].

One of the first approaches to improve on this,

[1] In the seen set -unlike the unseen- dialog 'topics' are shared with the training set.

Model	Method	Seen	Unseen
Random	-	2.7	2.3
Baseline (Dinan et al., 2018)	memory network	22.5	12.2
PostKS (Lian et al., 2019)	posterior signal	22.5	15.8
SKT(BERT) (Kim et al., 2020)	sequential latent kn selection	26.8	18.3
DiffKS(BERT) (Zheng et al., 2020)	difference aware	25.6	20.1
DukeNet (Meng et al., 2020)	kn tracking & shifting	26.4	19.6
SKT+ (Chen et al., 2020)	SKT + posterior signal + distillation	27.7	19.4
MIKe (Meng et al., 2021)	initiative aware	28.4	21.5
SKT-KG (Zhan et al., 2021b)	kn transition with CRF	26	-
KMine* (Lotfi et al., 2021)	posterior signal via generation	27.9	27.0
CoLV (Zhan et al., 2021a)	collaborative latent spaces	30.1	18.9
DIALKI (Wu et al., 2021)	dial-doc contextualization	32.9	35.5
DSG (Li et al., 2022)	document semantic graph	29.4	30.8
TAKE (Yang et al., 2022)	modeling topic shift	28.8	25.8
RoBERTa-base	sequence classification (dialog+kn)	28.6	26.6

Table 3: Knowledge selection performance (accuracy) on the WoW seen and unseen test sets for various models. Numbers are for the highest performing variance (when multiples were present). All models except for Baseline and PostKS benefit from pretrained transformers. ∗: KMine is unsupervised (no gold knowledge labels).

was addressing and exploiting the prior-posterior gap, which uses the posterior knowledge distribution to provide additional learning signals for the KS module, usually via a KL-divergence loss (Lian et al., 2019; Chen et al., 2020; Zhan et al., 2021a; Lotfi et al., 2021).

But probably the most popular approach is trying to address the problem on the dialog level (rather than turn level), and model higher-order 'flows' or sequential patterns that could guide the knowledge selection process. Li et al. (2019) used an incremental transformer to incorporate the knowledge selection history. Jiang et al. (2020) enhanced the posterior signal by modeling the 'topic drift'. Kim et al. (2020) introduced sequential latent knowledge selection to incorporate the selection history. Zheng et al. (2020) took a more specific approach by providing a positive bias for new or different knowledge choices. Meng et al. (2020) explicitly modeled 'knowledge tracking' and 'knowledge shifting' during a conversation while Meng et al. (2021) tried to incorporate speakers' initiative. Zhan et al. (2021b) used conditional random fields to model knowledge transition and Wu et al. (2021) leveraged the document structure to provide dialog-contextualized passage encodings while adding an auxiliary loss to capture the history of dialog-document connections. Li et al. (2022) used document semantic graphs to guide the knowledge selection, and Yang et al. (2022) proposed a topic-shift aware knowledge selector.

More recently models like RAG (Lewis et al., 2020) and FID (Kim et al., 2020) improved the question answering performance by shifting the final knowledge selection to the decoding process. Extending this to dialog (which was implemented differently by Lin et al. (2020)), studies have incorporated the fine-grained decoding-stage selection for better knowledge grounding (Shuster et al., 2021) or combined it with the posterior signal (Paranjape et al., 2021).

Table 3 summarizes a selection of these approaches, which mostly try to incorporate dynamic knowledge patterns by modeling attributes like topic shift, knowledge transition, knowledge tracking/shifting, knowledge difference etc.

3 Knowledge Selection Biases and Artefacts

Our main objective in this work is to explore the structural biases and artefacts in the knowledge selection labels of popular KGD datasets. For this, we use different methods depending on the way knowledge pools are constructed and presented to crowd-source workers, but in both cases, we essentially investigate the same hypothesis:

> *Crowd-source workers often base their knowledge selection on the structure and order of the knowledge pool, as presented to them.*

In other words, when selecting the knowledge piece for the next utterance, they tend to just follow the knowledge document and pick the 'next' item, instead of coming up with a more sophisticated 'flow'. In the following sections, we explore this hypothesis separately for Wizard of Wikipedia and Topical-Chat.

3.1 Wizard of Wikipedia (WoW)

As mentioned before, in WoW, dialogs happen between a 'wizard' and 'apprentice', with the former having access to unstructured knowledge. For each dialog either the wizard or apprentice is picked to choose the topic and speak first (the other player receives the topic information). The conversation begins and at each turn the wizard (system) can select from a knowledge pool which has been curated from a collection of Wikipedia articles via basic retrieval methods. Then the (potentially) selected knowledge piece is used by the wizard to generate the next utterance (system response). Out of 83247 wizard turns in the training set, 77523 (93%) are 'knowledge-grounded'; i.e. turns where the annotator has chosen a knowledge piece to ground their next utterance on.

Figure 1 shows how the knowledge pool is created for the wizard: At each turn, the last two utterances are used as queries by a TF-IDF retrieval module to get 14 (7 for each) relevant articles from a Wikipedia collection (title + first paragraphs). The dialog-topic article (title + first 10 sentences) is added to this set to create the final pool, which on average contains 63 knowledge *sentences* from up to 15 *passages* or articles[2]. Figure 5 (Appendix A) shows how this pool is presented to annotators.

To investigate our hypothesis, we consider 3 heuristic content-agnostic models for knowledge selection:

- **Topic-First (T0)**: Picks the first sentence of the dialog-topic article in all turns.
- **Topic-Next (T+)**: Starts from the dialog-topic article's first sentence, but proceeds to the next sentence at each successive turn.
- **Last-Next (L+)**: Picks the next sentence in the (gold) passage that was selected in the previous turn[3].

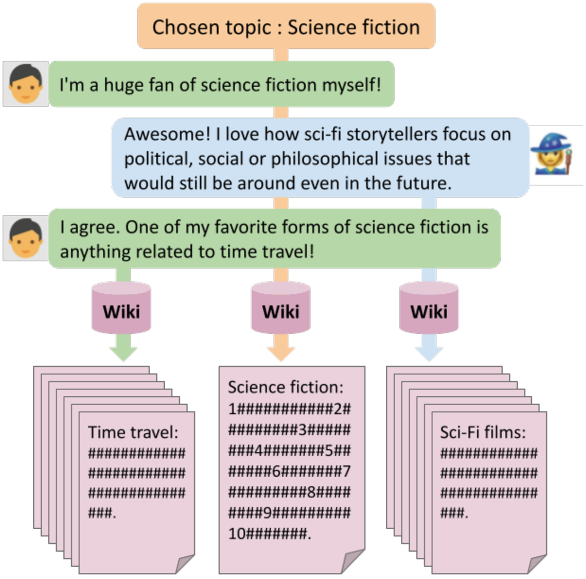

Chosen topic : Science fiction
I'm a huge fan of science fiction myself!
Awesome! I love how sci-fi storytellers focus on political, social or philosophical issues that would still be around even in the future.
I agree. One of my favorite forms of science fiction is anything related to time travel!
Wiki
Wiki
Wiki
Time travel:
#############
#############
#############
###.
Science fiction:
1###########2#
########3#####
###4#######5##
#####6#######7
#########8####
####9#########
10#######.
Sci-Fi films:
#############
#############
#############
###.

Figure 1: Curating the knowledge pool for wizard (right) in WoW dataset. At each turn a Wikipedia article collection is consulted using the last two utterances as queries, and the first paragraph of the 7 most relevant articles for each query plus the article for the dialog's chosen topic (first 10 sentences) are returned to create the knowledge pool to be used for the next wizard turn.

T0 is a static ('flow-less') model. T+ precisely and strictly follows the topic-article's narrative for dialog grounding, and L+ exploits the knowledge selection history for the next move.

Table 4 shows the performance of these 3 models in selecting the gold passage and sentence from the knowledge pool. T0 does not score very high but it offers a strong baseline for knowledge selection in WoW. In particular the T0 performance on the unseen test set already beats a handful of the models in Table 3 including the original baseline (18.9 vs. 12.2). Adding the basic 'flow' (T+) significantly improves the KS performance (an additional ~6% accuracy), and following the L+ selection policy adds another 5% boost to accuracy, making the content-agnostic L+ model highly competitive among the KS models. These performances show a strong bias towards picking the dialog-topic article among the passages, as well as picking the 'next' sentence within the current passage.

3.2 Topical-Chat

Unlike WoW, in Topical-Chat (Gopalakrishnan et al., 2019) the partners do not have explicitly defined roles. Instead, the authors leveraged infor-

[2] Since each passage corresponds to an article with a unique topic, passage, article and topic can be used interchangeably in the WoW context.

[3] If not available, the model returns the next (unused) sentence in the dialog-topic article (e.g. in the first turn, the dialog-topic article's first sentence will be selected.)

	Train				Test-seen				Test-unseen			
	Pssg. Acc.		Sent. Acc.		Pssg. Acc.		Sent. Acc.		Pssg. Acc.		Sent. Acc.	
Model	all	kng	all	kng	all	kng	all	kng	all	kng	all	kng
T0	67.4	72.4	18.3	19.7	67.9	72.4	18.1	19.3	70.9	75.2	18.9	20.1
T+	67.4	72.4	23.3	25.0	67.9	72.4	24.0	25.6	70.9	75.2	24.6	26.2
L+	68.5	73.5	27.9	29.9	68.9	73.5	28.3	30.1	72.3	76.8	30.2	32.1

Table 4: Knowledge selection accuracy for the heuristic content-agnostic models T0, T+ and L+ on WoW subsets. 'kng' refers to the knowledge-grounded subset. Underlined values can be compared with Table 3.

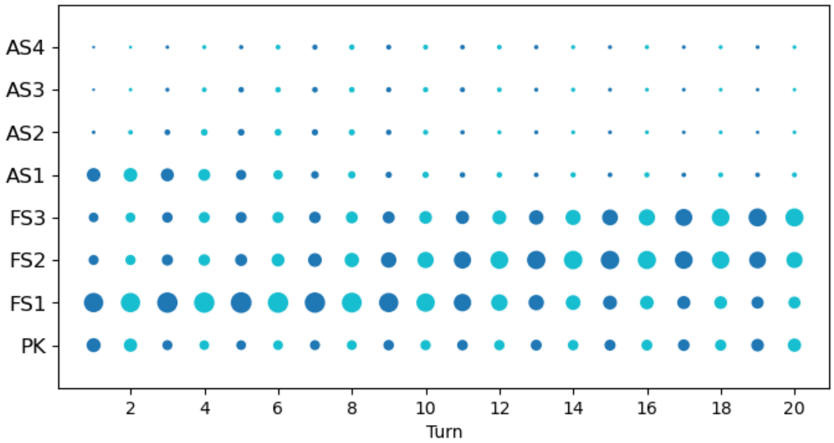


Figure 2: Grounding density for different knowledge sections at each turn in TopicalChat train set. Colors represent speakers. (PK: Personal Knowledge; FS: Factual Section; AS: Article Section.)

mation asymmetry to implicitly cause both partners to serve dual roles of a teacher and a participant which more accurately reflects real-world conversations.

For each dialog, 3 entities (from a pool of 300) were picked, plus their Wikipedia lead section, 8-10 fun facts, and a news article referencing all 3. These resources then were divided or modified according to one of 4 configurations (Figure 6 in Appendix B) to provide 2 identical (conf. A and B) or different (conf. C and D) knowledge pools. Finally Mechanical Turk workers were partnered up and assigned to these reading sets, and asked to chat about them for at least 20 turns. To present the reading set, information about an entity E (i.e. Wikipedia sections and fun facts) were displayed as a group titled Factual Section (FS), and the news article about the entities was chunked into 4 similar-sized sections (AS1-4). Turkers were asked to specify the knowledge source (FS1-3, AS1-4 and/or Personal Knowledge (PK)) used to generate their message at each turn. Selecting Personal Knowledge as the source means that the utterance is not grounded in external knowledge.

Figure 2 shows how the grounding evolves as conversations proceed in TopicalChat. Knowledge sections are arranged along the Y-axis, and point sizes represent normalized (per turn) frequency, or density. We can see that:

1. Grounding is mainly done on Factual Sections (FS), rather than Article Sections (AS).
2. The first part of the Article (AS1) is used significantly more than the rest for grounding, mainly in the beginning of the conversation[4].
3. As the conversation proceeds, the grounding density peak moves from FS1 to FS2 and FS3.
4. Personal Knowledge has a higher density in the beginning and ending turns which agrees with greeting patterns.

In other words, the 'average' TopicalChat conversation is likely to follow the PK-AS1-FS1-FS2-FS3-PK pattern for grounding. Since the numbering of entities and corresponding Factual Sections (i.e. FS1-3) in each conversation is independent

[4] This usually corresponds to opening utterances like *Do you know/Have you heard about X?*

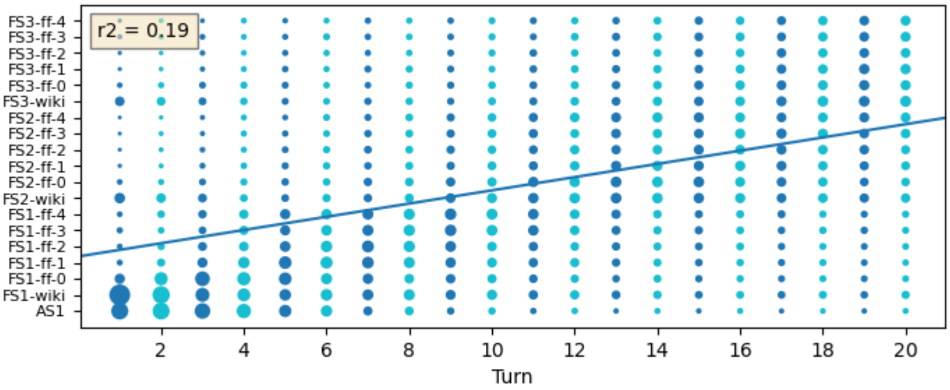


Figure 3: Fine-grained grounding density for different knowledge sections at each turn in TopicalChat train set. Colors indicate speakers.

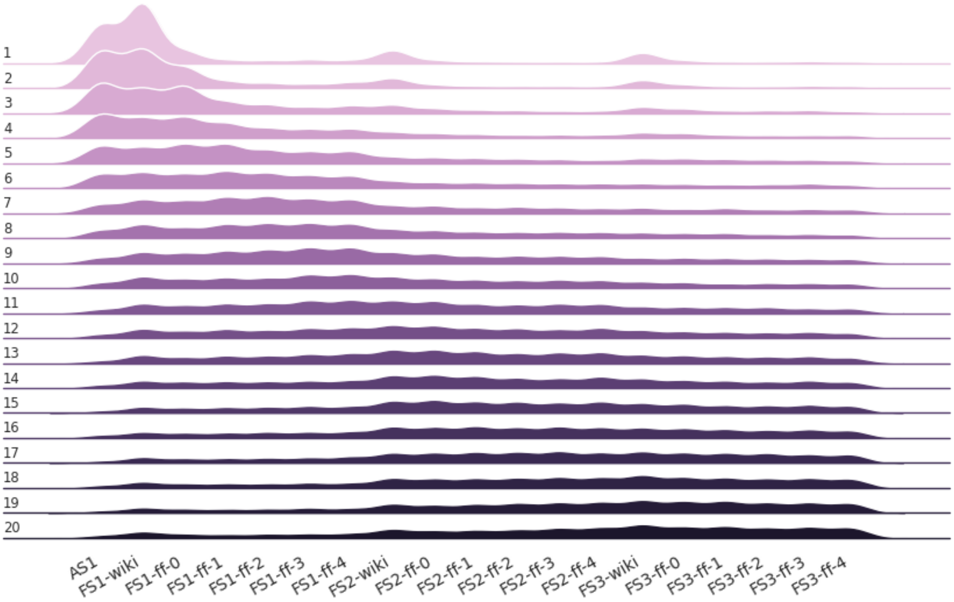


Figure 4: Grounding distribution over different knowledge parts (x-axis) for different turns (y-axis) in TopicalChat train set.

of the Article content[5], the observed pattern suggests that the grounding order is biased towards a pre-determined arbitrary parameter.

To have a more fine-grained view of this pattern, we exploit the response-knowledge overlap and employ a pre-trained sentence embedding model to estimate the gold knowledge sentence within the gold section (details in Appendix B). We then use these labels to expand each FS section into FS-wiki (the Wikipedia part) and FS-ff-{0-4} (the 5 fun facts). Figure 3 shows the resulting chart (limited to AS1, FS1, FS2 and FS3 sections), which demonstrates the same overall tendency of grounding on later sentences/sections as the conversation proceeds. The straight line is the linear regression fit (assuming sections' order as their value; i.e. 1, 2, ..., 19) with the slope and $r2$ value of 0.41 and 0.19 respectively (The slope of the diagonal line is 0.95).

The proceeding pattern is better illustrated in Ridgeline plots. For this, we switch the axes and invert the Y-axis direction so that the conversation starts at the top of the Y-axis. The result (Figure 4) displays a dispersing distribution with a clear tendency to 'move' forward; i.e. towards later sentences/sections. As Figure 7 (Appendix B) shows, this is shared in different knowledge configurations (i.e. A, B, C, D in Figure 6) with slight fluctuations.

[5] As opposed to –for example– numbered by importance or coverage order in the supporting Article.

3.3 WoW: A Markov View

As the final experiment, we revisit WoW dataset with a more complicated knowledge selection strategy. Inspired by the noticeable performance of content-agnostic models, we now consider a Markov model for KS with two stochastic parameters; Passage (P), and Line (L), with the following domains:

$$\begin{aligned} P &\in \{\texttt{None, DTopic, Other}\} \\ L &\in \{\texttt{None, Next, Other}\} \end{aligned} \quad (1)$$

P can be None (no grounding), DTopic (grounding on the dialog-topic passage) or Other (grounding on any other passage). Similarly, L can be None (no grounding), Next (picking the next sentence in the selected passage)[6], or Other (picking any other sentence. These choices follow the observed biases towards picking the dialog-topic article among the passages, and the 'next' sentence within the current passage, as discussed in 3.1.

Using this model, we can calculate initial state and transition probabilities from the WoW training set. Table 5 shows $P0$ and $L0$ probabilities for the initial state (S0); i.e first turn. 0/0 and 0/1 refer to the first turns in which wizard (0/0) or apprentice (0/1) start the conversation. As one can see, there is a strong bias towards picking the dialog-topic passage (DTopic) which is expected, especially for 0/0 where DTopic is the only grounding choice. More interesting is the tendency to start from the first sentence, especially when DTopic is chosen as passage (random probability: ~0.1).

P0	Turn = 0	0/0	0/1
None	0.048	0.061	0.034
DTopic	0.909	0.939	0.880
Other	0.043	0.0	0.085

L0	Turn = 0	0/0	0/1
Next (= first)	0.647	0.712	0.584
Other	0.353	0.288	0.416

Table 5: Initial state (S0) probabilities for the Passage (P) and Line (L) variables in WoW (here Next is equivalent to picking the first line in the passage).

Table 6 shows the transition probabilities for Passage (P) and Line (L) between successive states (full grounding memory), which demonstrates a strong tendency to 'stay' in DTopic (~0.75) and an overall preference for picking the Next sentence ((random probability: ~0.19).

P	None	DTopic	Other
None	0.208	0.415	0.377
DTopic	0.055	0.754	0.191
Other	0.102	0.192	0.704

L	Next	Other
	0.348	0.652

Table 6: Transition probabilities for the Passage (P) and Line (L) variables with full grounding memory.

Equation 1 along with the $P0$, $L0$, P and L values provides a fine-grained content-agnostic distribution (CAG) over the knowledge choices at each turn, which can be used in combination with any content-aware (CAW) KS model. Here we examine three ways to do so (all CAW models are based on RoBERTa-base):

- **Ensemble**: We simply use the CAG predictions in a mean-value ensemble.
- **TokenCues**: Instead of directly incorporating the CAG values, we provide corresponding bias cues as special tokens in the input sequences. In particular we add `<topic>`, `<next>` and `<prev_next>` to respectively mark the topic-article sentences, the successive sentence in each passage (w.r.t the last visited one in that passage) and the sentence after the one selected in the previous turn.
- **Both**: We use the token-cues model in combination with CAG in a mean-value ensemble.

Model	S	U
Baseline	28.6	26.6
Ensemble (CAG + Baseline)	31.9	33.8
TokenCues	32.8	33.8
Ensemble (CAG + TokenCues)	32.9	34.6

Table 7: Knowledge selection accuracy on WoW test subsets (S: seen, U:unseen) for various incorporations of the content-agnostic knowledge.

Table 7 shows the KS performance of these variations compared with the conventional sequence classification approach (Baseline). As one can see,

[6] Here Next is meant with respect to the grounding history; i.e. picking up from the last time the passage was visited. In the case of no grounding memory (first-order Markov), L starts from 0 every time the grounding topic changes.

incorporating the content-agnostic knowledge (directly or indirectly), results in a significant performance improvement. Moreover it seems that the transformer model is capable of learning the KS biases once proper cues are provided in the training data: the TokenCues model matches the Baseline Ensemble while gaining only marginal improvement from the explicit CAG values.

4 Discussion and Conclusion

In this work we investigated the potential knowledge selection biases and artefacts in two popular KGD datasets. Our central governing hypothesis was that crowd-source workers tend to simply follow the structure and order of knowledge pieces, as presented to them. For the WoW dataset, we showed that using this hypothesis, content-agnostic models can achieve noticeable knowledge selection performance, and combined with simple sequence classification training are able to compete with sophisticated solutions. For Topical-Chat we observed a noisy alignment between the KS sequence and the order of entities and their segments, as provided to crowd-sources.

Although following the existing order of knowledge pieces is not strange or unexpected (at least within one document), we believe that the way knowledge options are curated and presented to crowd-source workers can be an exacerbating factor. All 4 datasets provide a large number of retrieved knowledge pieces at each turn (usually more than 60) which is statistically beneficial to the dataset, but it could also encourage an ‘easy solution’ regime in which annotators opt for the safe and convenient choice of following the already existing structure of knowledge articles, instead of trying to create and maintain a novel ‘flow’. In its extreme case, this leads to conversations similar to reciting an article line by line[7].

In terms of dialog modeling, these results can suggest that the origin and significance of higher-order attributes in the dataset can be questioned. In particular, the concept of ‘flow’ as governing the dialog-level pattern of knowledge selection seems to be rooted substantially in the structure of knowledge documents. This does not rule out the existence or learnability of genuine patterns/flows, but the very low human performance for this task (~17%; Table 1) imposes a serious higher-bound on its discerning power; i.e. in most cases, there seems to be not enough semantic cues in the conversational history to uniquely and clearly bound it to a single knowledge piece.

Although the ultimate goal in KGD modeling is generating proper responses (and not mastering the knowledge selection part), but in order to model higher-order and dialog-level conversational phenomena, we probably need better datasets. One important factor in producing such resources is considering the process from annotators’ point of view, and how design choices (e.g. annotation interface and instructions, size of knowledge pool, etc.) can persuade them towards or away from ‘easy solution’ regimes which are prone to artifacts. Another approach is providing explicit ‘scenarios’ for the way dialogs are supposed to unfold. This is how DuConv (Wu et al., 2019) and NaturalConv (Wang et al., 2021) datasets (both Chinese) have been curated, but whether this mitigates the problem or introduces new artifacts should be studied.

5 Limitations

The main limitation of our work is its focus on English datasets. While this was due to their popularity and extensive usage (and our limited language skills), it overlooks datasets like DuConv (Wu et al., 2019) and NaturalConv (Wang et al., 2021) (both Chinese) which employ more explicit annotation instructions regarding dialog ‘path’ and topic transitions. Studying the way these restrictions affect conversational attributes, is necessary for a more comprehensive understanding of the problem.

Another limitation is the lack of an empirical investigation on how/if these artefacts and biases affect the final objective of KGD modeling, i.e. response generation. This of course is not easy in the absence of a less biased dataset, but synthetic datasets –which have become much better in quality and flexibility thanks to large language models– can probably provide reliable estimations, which we plan to explore in future studies.

Acknowledgements

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

[7] There are also case-specific factors. for example in WoW the utterance-based knowledge pieces are subject to change at each turn, and therefore there is no guarantee that the passage used for grounding in the current turn will be present in the provided pool for the next turn. This makes grounding on the dialog-topic article a safe choice, since it is always in the pool.

References

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Bin Jiang, Jingxu Yang, Chao Yang, Wanyue Zhou, Liang Pang, and Xiaokang Zhou. 2020. Knowledge augmented dialogue generation with divergent facts selection. *Knowledge-Based Systems*, 210:106479.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Seattle, United States. Association for Computational Linguistics.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2021. Teach me what to say and I will learn what to pick: Unsupervised knowledge selection through response generation with pretrained generative models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 254–262, Online. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 522–532, New York, NY, USA. Association for Computing Machinery.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1151–1160, New York, NY, USA. Association for Computing Machinery.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D. Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation.

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

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *AAAI Conference on Artificial Intelligence*.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. TAKE: Topic-shift aware knowledge sElection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 253–265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021a. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021b. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630, Online. Association for Computational Linguistics.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

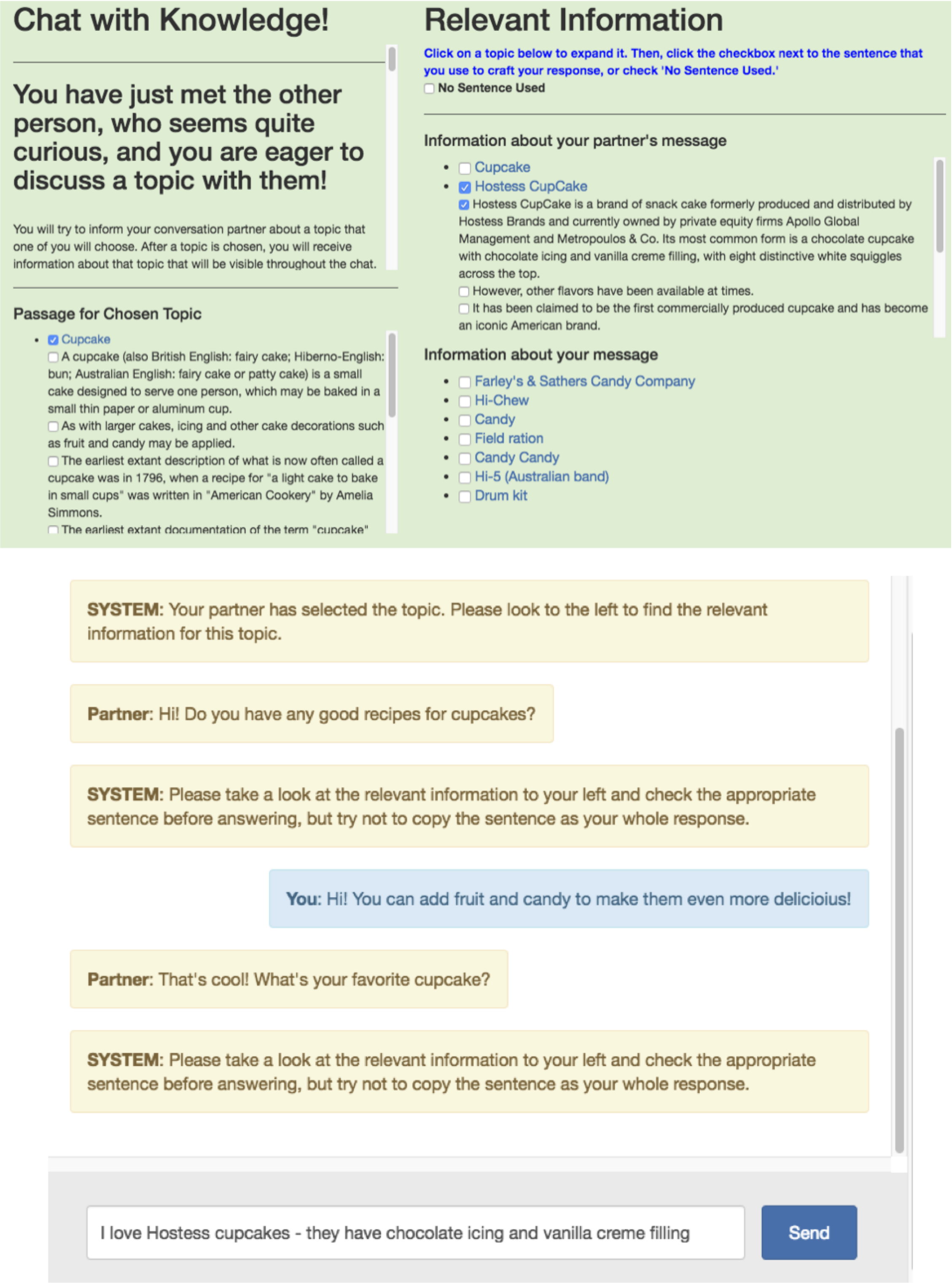
A Appendix: WoW Interface

Figure 5 shows the annotation interface used in curating Wizard of Wikipedia.

B Appendix: Topical-Chat

Considering the absence of sentence-level gold labels in Topical-Chat, we exploit the response-knowledge overlap and employ a pre-trained sentence embedding model to estimate the gold knowledge sentence within the gold section. More specifically, we use the `all-mpnet-base-v2` model from the 'sentence-transformers' library (Reimers and Gurevych, 2019) which shows the highest performance on benchmarks, and pick the sentence which has the highest cosine similarity with the response. Manually checking the performance on a small subset (500 grounded samples) shows an error rate of 18% (accuracy =~82%) of which 10% is due to incorrect gold section labels. Enforcing an acceptance similarity threshold of 0.2, filters out 13% of samples including 88% of errors, which improves the accuracy to 93%. We apply this setting to the train set[8], and –to keep conversations in reasonable lengths (and therefore less likely to be damaged by the filtering)–, we remove dialogs with less than 80% of accepted utterances. This, results in a more reliable subset of 7922 (out of 8,628) conversations, with 150564 utterances.

[8] We consider the first 20 utterances in each dialog, which is the minimum required length during crowd-sourcing.



Chat with Knowledge!

You have just met the other person, who seems quite curious, and you are eager to discuss a topic with them!

You will try to inform your conversation partner about a topic that one of you will choose. After a topic is chosen, you will receive information about that topic that will be visible throughout the chat.

Passage for Chosen Topic

- ☑ Cupcake
  ☐ A cupcake (also British English: fairy cake; Hiberno-English: bun; Australian English: fairy cake or patty cake) is a small cake designed to serve one person, which may be baked in a small thin paper or aluminum cup.
  ☐ As with larger cakes, icing and other cake decorations such as fruit and candy may be applied.
  ☐ The earliest extant description of what is now often called a cupcake was in 1796, when a recipe for "a light cake to bake in small cups" was written in "American Cookery" by Amelia Simmons.
  ☐ The earliest extant documentation of the term "cupcake"

Relevant Information

Click on a topic below to expand it. Then, click the checkbox next to the sentence that you use to craft your response, or check 'No Sentence Used.'
☐ No Sentence Used

Information about your partner's message

- ☐ Cupcake
- ☑ Hostess CupCake
  ☑ Hostess CupCake is a brand of snack cake formerly produced and distributed by Hostess Brands and currently owned by private equity firms Apollo Global Management and Metropoulos & Co. Its most common form is a chocolate cupcake with chocolate icing and vanilla creme filling, with eight distinctive white squiggles across the top.
  ☐ However, other flavors have been available at times.
  ☐ It has been claimed to be the first commercially produced cupcake and has become an iconic American brand.

Information about your message

- ☐ Farley's & Sathers Candy Company
- ☐ Hi-Chew
- ☐ Candy
- ☐ Field ration
- ☐ Candy Candy
- ☐ Hi-5 (Australian band)
- ☐ Drum kit

SYSTEM: Your partner has selected the topic. Please look to the left to find the relevant information for this topic.

Partner: Hi! Do you have any good recipes for cupcakes?

SYSTEM: Please take a look at the relevant information to your left and check the appropriate sentence before answering, but try not to copy the sentence as your whole response.

You: Hi! You can add fruit and candy to make them even more delicioius!

Partner: That's cool! What's your favorite cupcake?

SYSTEM: Please take a look at the relevant information to your left and check the appropriate sentence before answering, but try not to copy the sentence as your whole response.

I love Hostess cupcakes - they have chocolate icing and vanilla creme filling

Figure 5: Annotation interface for the Wizard of Wikipedia dataset (from (Dinan et al., 2018))

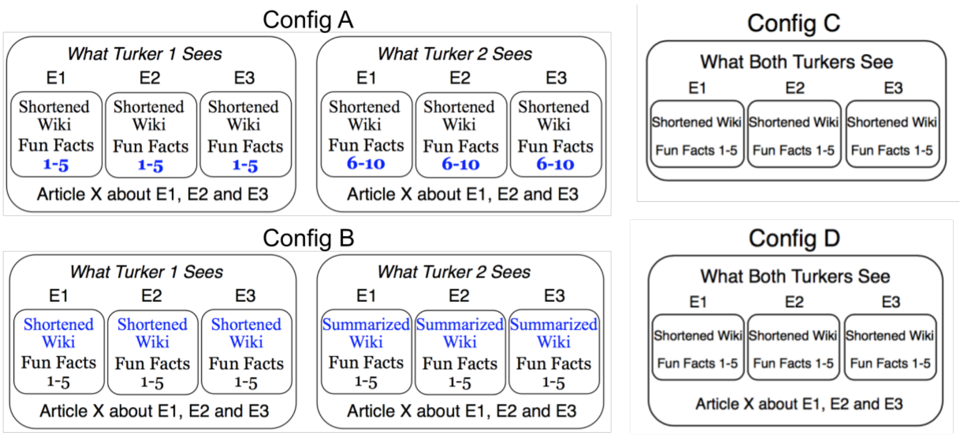
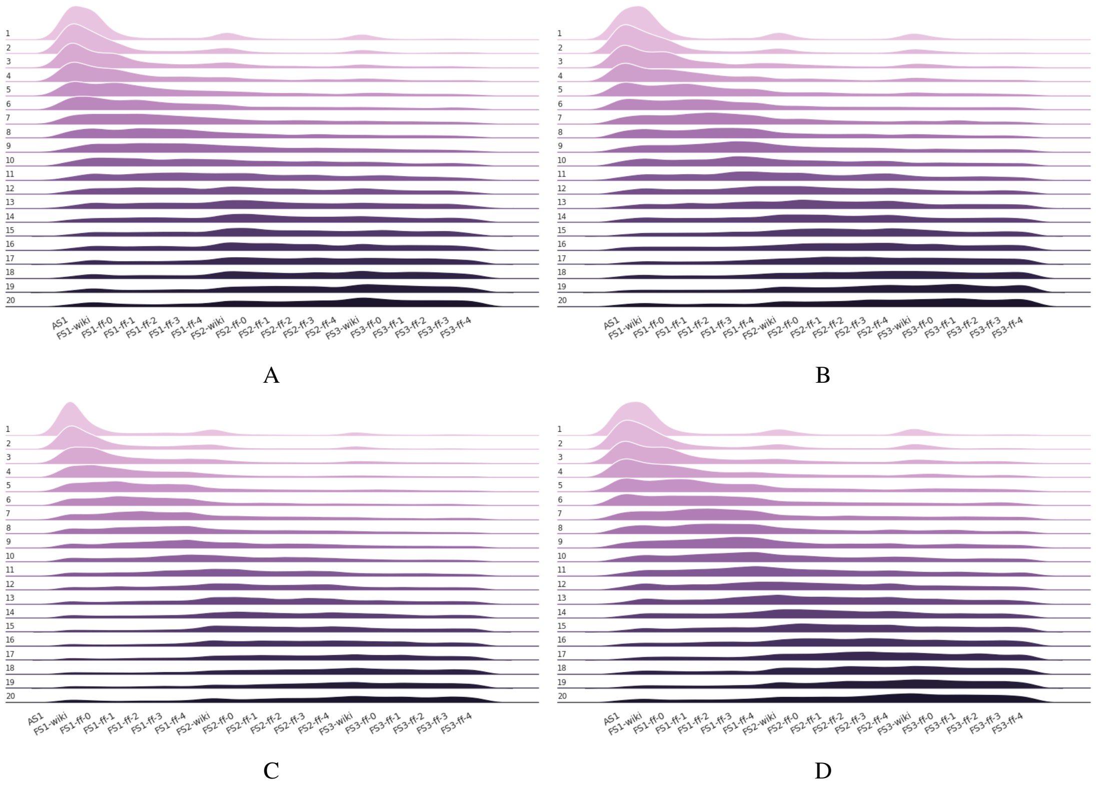

Config A
What Turker 1 Sees
E1 E2 E3
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Article X about E1, E2 and E3
What Turker 2 Sees
E1 E2 E3
Shortened Wiki Fun Facts 6-10
Shortened Wiki Fun Facts 6-10
Shortened Wiki Fun Facts 6-10
Article X about E1, E2 and E3
Config C
What Both Turkers See
E1 E2 E3
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Config B
What Turker 1 Sees
E1 E2 E3
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Article X about E1, E2 and E3
What Turker 2 Sees
E1 E2 E3
Summarized Wiki Fun Facts 1-5
Summarized Wiki Fun Facts 1-5
Summarized Wiki Fun Facts 1-5
Article X about E1, E2 and E3
Config D
What Both Turkers See
E1 E2 E3
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Shortened Wiki Fun Facts 1-5
Article X about E1, E2 and E3

Figure 6: The four knowledge configurations in TopicalChat (from Gopalakrishnan et al. (2019). E stands for Entity.



A

B

C

D

Figure 7: Turn-wise grounding distribution over different knowledge parts (x-axis) for different configurations (A, B, C, D) in TopicalChat train set.