The Interplay of Task Success and Dialogue Quality: An in-depth Evaluation in Task-Oriented Visual Dialogues

Alberto Testoni
DISI - University of Trento
Trento - Italy
alberto.testoni@unitn.it

Raffaella Bernardi
CIMeC and DISI - University of Trento
Rovereto (TN) - Italy
raffaella.bernardi@unitn.it

Abstract

When training a model on referential dialogue guessing games, the best model is usually chosen based on its task success. We show that in the popular end-to-end approach, this choice prevents the model from learning to generate linguistically richer dialogues, since the acquisition of language proficiency takes longer than learning the guessing task. By comparing models playing different games (GuessWhat, GuessWhich, and Mutual Friends), we show that this discrepancy is model- and task-agnostic. We investigate whether and when better language quality could lead to higher task success. We show that in GuessWhat, models could increase their accuracy if they learn to ground, encode, and decode also words that do not occur frequently in the training set.

1 Introduction

A good dialogue model should generate utterances that are indistinguishable from human dialogues (Liu et al., 2016; Li et al., 2017). This holds for both chit-chat, open-domain, and task-oriented dialogues. While chit-chat dialogue systems are usually evaluated by analysing the quality of their dialogues (Lowe et al., 2017; See et al., 2019), task-oriented dialogue models are evaluated on their task success and it is common practice to choose the best model based only on the task success metric. We explore whether this choice prevents the system from learning better linguistic skills.

Important progress has been made on the development of such conversational agents. The boost is mostly due to the introduction of the encoder-decoder framework (Sutskever et al., 2014) which allows learning directly from raw data to both understand and generate utterances. The framework has been found to be promising both for chit-chat (Vinyals and Le, 2015) and task-oriented dialogues (Lewis et al., 2017), and it has been further extended to develop agents that can communicate through natural language about visual content (Mostafazadeh et al., 2017; Das et al., 2017a; de Vries et al., 2017). Several dialogue tasks have been proposed as *referential guessing games* in which an agent (the Q-bot) asks questions to another agent (the A-bot) and has to guess the referent (e.g., a specific object depicted in the image) they have been speaking about (de Vries et al., 2017; Das et al., 2017b; He et al., 2017; Haber et al., 2019; Ilinykh et al., 2019; Udagawa and Aizawa, 2019). We are interested in understanding the interplay between the learning processes behind these two sub-tasks: generating questions and guessing the referent.

Shekhar et al. (2019) have compared models on GuessWhat and have shown that task success (TS) does not correlate with the quality of machine-generated dialogues. First of all, we check whether this result is task-agnostic by carrying out a comparative analysis of models playing different referential games. We choose a task in which visual grounding happens during question generation (GuessWhat, de Vries et al. 2017); a task in which it happens only in the guessing phase (GuessWhich, Das et al. 2017b), and a task that is only based on language (MutualFriends, He et al. 2017). We introduce a linguistic metric, Linguistic Divergence (LD), that, by assembling various metrics used in the literature (Shekhar et al., 2019; Murahari et al., 2019; van Miltenburg et al., 2019), measures how much the language generated by computational models differs, on the surface level, from the one used by humans. We consider LD to be a proxy of the quality of machine-generated dialogues.

For each task, we compare State-Of-The-Art (SOTA) models against their TS and LD. In the

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2071–2082
April 19 - 23, 2021.

core part of the paper, we study the relationship between the learning process behind TS and LD by comparing model performance across epochs and by downsizing the training set. Finally, we study whether and when a lower LD (i.e., the generated dialogues are more similar to humans) could help reach a higher TS.

Our results confirm that models performing similarly on TS differ quite a lot on their conversational skills, as claimed in Shekhar et al. (2019) for models evaluated on the GuessWhat game. Furthermore, we show that:

- SOTA models are much faster in achieving high performance on the guessing task compared to reaching a high dialogue quality (i.e., low LD). Hence, choosing the best model on task success prevents the model from reaching better conversational skills;
- SOTA models mostly use very frequent words; this limited vocabulary is sufficient for succeeding in a high number of games;
- in GusseWhat, a higher TS could be reached if the model learns to use also less frequent words.

2 Related Work

Task-oriented models can be evaluated based on their task success, but this is not enough to know whether the generated dialogues are human-like. The development of quantitative metrics to evaluate the quality of dialogues generated by conversational agents is a difficult challenge (Liu et al., 2016), and it is under investigation for chit-chat dialogue systems. For instance, Guo et al. (2017) study topic diversity in the conversational flow, which is rather important in chit-chat and open-domain dialogues, but less so for task-oriented ones; Kannan and Vinyals (2016), Li et al. (2017), Bruni and Fernández (2017) propose to use adversarial evaluation, whereas Lowe et al. (2017), See et al. (2019), and Hashimoto et al. (2019) propose automatic systems that build upon human evaluation. All these efforts are still preliminary and are not easily employable for new datasets or new models. Since no standard and unique metric has been proposed to evaluate the quality of task-oriented (grounded) conversational dialogues, we consider a mixture of metrics used independently in various studies, and we provide a comparative analysis across models and tasks based on the same set of linguistic metrics.

Neural Networks have been shown to generate text that sounds unnatural due to the presence of repeated utterances, poor vocabulary, and inconsistency in word usage (Ray et al., 2019). Various improvements have been proposed to mitigate these weaknesses. To prevent the decoder from choosing words based simply on their frequency, Li et al. (2019) replace its maximum likelihood estimation objective, while others change the sampling search strategy (Holtzman et al., 2020; Wu et al., 2019; See et al., 2019); these changes aim to reduce the number of repeated questions, to increase the variety of words and their distribution. Attempts have been made to provide the conversational models with a reasoning module based on Bayesian inference (Abbasnejad et al., 2019) or Rational Speech Act (Shuklar et al., 2019) frameworks that should lead to more informative and coherent questions. Here, we do not propose new models, but rather aim to better understand the strengths and weaknesses of current models.

3 Games and Metrics

Our focus is on task-oriented dialogues. We consider a task that relies on grounding language into vision during question generation, i.e. GuessWhat (de Vries et al., 2017), a task that requires grounding only at the guessing phase, i.e. GuessWhich (Das et al., 2017b), and a task based only on language, i.e. MutualFriends, (He et al., 2017).

Games As illustrated by the snippets reported in Table 1, the three tasks also differ in the flexibility of the dialogues: GuessWhat and GuessWhich are both based on rigid turns in which an agent asks questions and the other answers, whereas MutualFriends has free-form dialogues. Moreover, GuessWhat consists only of Yes/No questions, while in GuessWhich this constraint does not apply. Relevant statistics of the three datasets are summarized in Table 1.

GuessWhat (de Vries et al., 2017) is an asymmetric game.[1] A Questioner (Q-Bot) has to ask Yes/No questions to guess which is the target object among a set of maximum 20 candidates; while asking questions, it sees the image containing the

[1]The dataset of human dialogues is available at https://guesswhat.ai/download.

	#dialogues training	testing	Vocab. size	#candidates	#turns	Examples
GuessWhat	108K	23K	4900	3-20	1-10	***A***: Is it a person? ***B***: No. ***A***: Is it a dog? ***B***: Yes. [***A*** guesses the target object]
GuessWhich	120K	2K	11321	2K	10	***A***: What color is the car? ***B***: Blue. ***A***: Who is driving it? ***B***: A man. [***A*** guesses the target image]
MutualFriends	8K	1K	5325	5-12	2-46	***A***: My friends work at Google. ***B***: None of mine do. [***A*** and ***B*** select a friend]

Table 1: Salient statistics of the human dialogues in the three data sets under consideration in this work. The last column reports samples of dialogues exchanged between two agents (A and B).

candidate objects and it has access to the dialogue history. The Answerer (A-Bot), who knows which is the target, provides the answers. The two bots learn to speak about the image by being trained on human dialogues, which have been collected by letting humans play the game. Humans could stop asking questions at any time (human dialogues contain on average 5.2 question-answer pairs), while models have to ask a fixed number of questions (8 in the setting we have considered).

GuessWhich (Das et al., 2017b) is also an asymmetric game. Unlike the task described above, the Q-Bot has to ask questions without seeing the candidate images, but it has access to captions describing the images. Q-Bot can ask any type of question; the target image has to be selected among 2K candidates at the end of the dialogue. The A-Bot instead sees both the caption and the target image. Human dialogues are from the VisDial dataset[2] and were collected as chit-chat dialogues (Das et al., 2017a). Both humans and models have to ask exactly 10 questions.

MutualFriends (He et al., 2017) is a symmetric game based only on text: two agents, each given a private list of friends described by a set of attributes/labels, try to identify their mutual friend based on the friend's attributes.

Metrics: Since we are interested in the interplay between the downstream task and the quality of the generated dialogues, we consider two types of metrics.

Task Success: We use the task success (TS) metrics used in the literature to evaluate models against these tasks, namely accuracy (ACC) for GuessWhat and MutualFriends, and Mean Percentile Rank (MPR) for GuessWhich. The latter is computed from the mean rank position (MR) of the target image among all the candidates. An MPR of e.g., 96% means that, on average, the target image is closer to the one chosen by the model than the 96% of the candidate images. Hence, in the VisDial test set with 2K candidates, 96% MPR corresponds to an MR of 80, and a difference of $\pm$ 1% MPR corresponds to $\mp$ 20 mean rank. The task success chance levels are: 5% accuracy (GuessWhat), 50% MPR (GuessWhich) and 11.76% accuracy (MutualFriends).

Linguistic metrics: It has been shown that the quality of the dialogues generated by computational agents is not satisfactory. The main weaknesses of these models consist of poor lexical diversity, a high number of repetitions, and the use of a limited vocabulary. To evaluate the quality of the generated dialogues (defined as the closeness to human dialogues according to surface-level cues), we use several metrics that have been proposed in the literature. As in He et al. (2017), we compute *unigram entropy (H)*, which measures the entropy of unique unigrams in the generated dialogues normalized by the total number of tokens used by the model. From Murahari et al. (2019), we take the *Mutual Overlap (MO)* metric, which evaluates the question diversity within a dialogue by computing the average of the BLEU-4 score obtained by comparing each question with the other questions within the same dialogue.[3]

[2] VisDial is available from https://visualdialog.org/data.

[3] A high number of novel questions and low mutual overlap cannot be taken per se as a sign of high quality of the dialogues: a model could ask a question never seen in training or with very little overlap with the other questions but completely out of scope. To rule out this possibility, we compute the cosine similarity of each question marked as novel and with a low mutual overlap with the dialogue they occur in, and compare it with the similarity between the latter and

Moreover, following Shekhar et al. (2019), we report the percentage of games with *one question repeated verbatim (GRQ)* within a dialogue. Finally, we compare models with respect to their ability on lexical acquisition by calculating the *Global Recall (GR)* introduced by van Miltenburg et al. (2019) to evaluate image captioning: it is defined as the overall percentage of learnable words (from the training set) that the models recall (use) during generation. Furthermore, taking inspiration from the Local Recall introduced in the same work, we propose a similar metric tailored to dialogues, i.e., *Local Recall-d (LRd)*, which measures how many content words the generated dialogue shares with the corresponding human dialogue for the same game. Given a human dialogue D_h about an image and a generated dialogue D_g about the same image, we compute LRd as the normalized lexical overlap (considering only content words) between D_h and D_g.

We sum up all these linguistic metrics used in the literature so far into one which we take as a proxy of the quality of dialogues: it shows the *linguistic divergence (LD)* of the dialogues generated by a model from human dialogues. To this end, we normalize each metric so that all values lie between 0 and 1: 0 stands for human performance for the "lower is better" metrics and 1 stands for human performance for "higher is better" metrics. We compute LD by averaging all the scaled values V for each model; we take $1 - V$ for "higher is better" metrics to obtain a "divergence" value. All the metrics are equally weighted. By definition, LD is 0 for human dialogues. LD captures three main surface-level aspects: overall vocabulary usage (H, GR), diversity of questions/phrases within a dialogue (MO, GRQ), and similarity of content word usage with respect to human dialogues (LRd). There could be some correlation between metrics capturing similar aspects of language quality, but this does not affect the validity of the proposed LD metric.

4 Models

For both visual dialogue games, GuessWhat and GuessWhich, supervised learning has been compared with other learning paradigms. After the introduction of the supervised baseline model (de Vries et al., 2017), several models have been proposed for GuessWhat. They exploit either reinforcement learning (Sang-Woo et al., 2019; Zhang et al., 2018b,a; Zhao and Tresp, 2018; Gan et al., 2019; Pang and Wang, 2020) or cooperative learning (Shekhar et al., 2019; Pang and Wang, 2020); in both cases, the model is first trained with the supervised learning regime and then the new paradigm is applied. This two-step process has been shown to reach higher task success than the supervised approach. For GuessWhich, after the supervised model introduced in Das et al. (2017a), new models based on reinforcement learning have been proposed, too (Das et al., 2017b; Murahari et al., 2019; Zhou et al., 2019), but their task success is comparable if not lower than the one achieved by using only supervised learning (see Testoni et al. 2019). Below, we briefly describe the models we have compared in our analysis. For each task, we have chosen generative models trained with different learning paradigms and for which the code is available; for each paradigm, we have tried to choose the best performing ones or those that obtain a task success near to state-of-the art and could help better understand the interplay between task success and dialogue quality.

GuessWhat We use the A-Bot introduced in de Vries et al. (2017), which is trained in a supervised learning (SL) fashion. For the Q-Bot, we compare models based on different learning paradigms: supervised and cooperative learning (GDSE-SL and GDSE-CL, respectively) proposed in Shekhar et al. (2019) and reinforcement learning (RL) proposed in Strub et al. (2017). In RL, the reinforce paradigm used aims at optimizing the task accuracy of the game. Besides using different learning paradigms, these models differ in their architecture. In particular, while in RL the Question Generator (QGen) and the Guesser are trained independently, in GDSE a common visually-grounded dialogue state encoder is used and the two modules are trained jointly. In both cases, the Guesser receives as input the candidate object's categories and their spatial coordinates, and during training it is updated only at the end of the dialogue.[4]

random questions taken from other dialogues. Embeddings are obtained by using Universal Sentence Encoder-USE (Cer et al., 2018). We found that novel and low-MO questions are more similar to their dialogue than the random ones, confirming the effectiveness of these metrics.

[4]The code of the A-Bot and of RL is available at https://github.com/GuessWhatGame/guesswhat. The code of GDSE at: https://github.com/shekharRavi/

GuessWhich We use the A-Bot introduced in Das et al. (2017b). For the Q-bot, we compare Diverse (Murahari et al., 2019) and ReCap (Testoni et al., 2019). Diverse and ReCap have similar architectures: several encoders incrementally process the linguistic inputs to produce the dialogue hidden state. This state is used to condition a decoder that generates a new question at each turn, and a guesser that is trained to produce the visual representation of the target image through a feature regression module. The two models differ in the encoders used and in the training paradigm. While Diverse encodes the caption together with the dialogue history through a Hierarchical LSTM, ReCap has two independent LSTMs that produce the linguistic features of the caption and of the dialogue history, merged together to produce the dialogue hidden state. Secondly, in Diverse, an auxiliary objective on the dialogue state embedding (Huber loss) is used to incentivize the bot to ask more diverse questions with respect to the immediate previous turn. In ReCap, the Guesser sees the ground-truth image only at the end of the game while in Diverse the Guesser is updated at each turn. ReCap has been trained only by SL, Diverse both by SL (D-SL) and SL plus RL (D-RL). Further details can be found in the respective papers (Murahari et al., 2019; Testoni et al., 2019).[5]

MutualFriends We evaluate the model proposed in He et al. (2017), DynoNet (Dynamic Knowledge Graph Network), in which entities are structured as a knowledge graph and the utterance generation is driven by an attention mechanism over the node embeddings of such graph. The model is trained via supervised learning and at test time it plays with itself. DynoNet consists of three components: a dynamic knowledge graph (which represents the agent's private KB and shared dialogue history as a graph), and two LSTMs that map the graph embedding over the nodes and generate utterances or guess the entity.[6]

Beyond-Task-Success-NAACL2019.

[5]The code for the A-Bot model and for D-SL and D-RL is available at https://github.com/vmurahari3/visdial-diversity; it is not specified how the best models are chosen. For ReCap: we have obtained the code by the authors and trained the model; we have chosen the model whose MPR does not increase for the subsequent 5 epochs.

[6]The code is available at https://github.com/stanfordnlp/cocoa/tree/mutualfriends.

5 Experiments and Results

Shekhar et al. (2019) has shown that in GuessWhat task success (TS) does not correlate with the quality of the dialogues. First of all, we check to what extent this result is task and model agnostic by taking GuessWhat, GuessWhich and MutualFriends as case-studies and compare the behaviour of the models described above.

First, we evaluate the impact of the number of epochs and the size of the training set; then, we study whether and when a lower LD could help to reach a higher TS.

5.1 Task Success and Linguistic Divergence

We evaluate all models described above, in their supervised, cooperative, or reinforcement learning version, aiming to test whether some patterns can be found irrespectively of the model and data explored. Our results confirm what has been shown in Shekhar et al. (2019) for GuessWhat: **TS does not correlate with the quality of the generated dialogues**; models with similar TS generate dialogues that vary greatly with respect to the linguistic metrics. We run a Spearman's analysis and found a very weak correlation between LD and TS (coefficient < 0.15, p-value < 0.05). Our comparison across tasks shows that both in GuessWhich and in GuessWhat the vocabulary used by humans while playing the games in the training and testing set is rather similar (resp., 91% and 84% of words are in common between training and testing sets). Yet in both visual tasks models reach an LRd of around 42%. Specifically, the average mean rank of words they fail to use is 7000 (over 11321) for GuessWhich and 3016 (over 4900) for GuessWhat. **Hence, models mostly use very frequent words.** Details on the metrics for each task and model are reported in Table 2.

5.2 Learning Processes behind TS and LD

We aim to understand the relation between TS and LD. To this end, we compare the two metrics during the training processes across epochs and by downsizing the training set. For each task, we consider the models trained in a SL fashion since those trained with other paradigms build on them. Hence, we focus on ReCap, GDSE-SL, and DynoNet.

Comparison Across Epochs We study for *how long* a model has to be trained to reach its best performance on guessing the target referent and

	GuessWhich				GuessWhat				MutualFriends	
	D-SL	D-RL	ReCap-SL	Hum	GDSE-SL	GDSE-CL	RL	Hum	DynoNet-SL	H
TS ↑	95.2	94.89	96.76	-	48.21	59.14	56.3	84.62	0.98	0.82
GR ↑	6.46	9.04	14.4	27.69	34.73	36.35	12.67	72.98	51.15	65.2
LRd ↑	39.93	41.83	42.76	-	42.1	42.41	34.51	-	-	-
MO ↓	0.51	0.41	0.23	0.07	0.39	0.23	0.46	0.03	-	-
GRQ ↓	93.01	81.17	55.37	0.78	64.96	36.79	96.54	0.8	-	-
H ↑	4.03	3.92	4.19	4.55	3.52	3.66	2.42	4.21	3.91	4.57
LD ↓	0.58	0.52	0.38	-	0.46	0.36	0.67	-	0.18	-

Table 2: Comparative analysis of different models on several tasks and datasets.TS: task success. GR: global recall. LRd: local recall. MO: mutual overlap. GRQ: games with repeated questions. H: unigram entropy. LD: linguistic divergence. ↑: higher is better. ↓: lower is better.

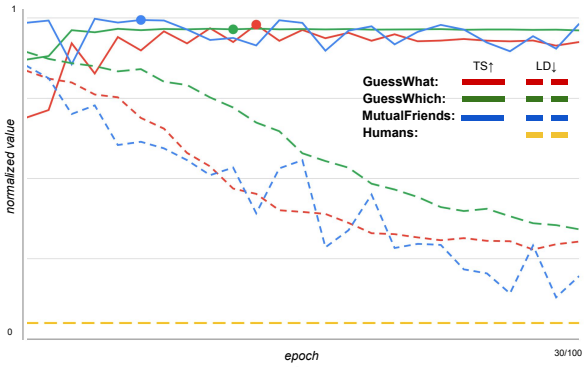


Figure 1: Comparison of Task Success (TS, solid lines) and Linguistic Divergence (LD, dashed lines) for GDSE-SL (GuessWhat, trained for 100 epochs), ReCap (GuessWhich, trained for 100 epochs), DynoNet (MutualFriends, trained for 30 epochs); Humans LD lower-bound metric in yellow. The LD of the generated dialogues keeps decreasing (moving close to human level) even though we no longer notice improvements in TS, whose highest value is reached well before (marked by bullets).

generating human-like dialogues. Figure 1 reports the TS and the LD of three models trained on the three tasks under examination. Each line is normalized w.r.t the highest value for each metric, so that it is possible to see different trends on the same plot. As we can see from the figure, the highest TS (marked by bullets) is reached earlier in GuessWhich and in MutualFriends than in GuessWhat. More interestingly, for all the tasks, the LD of the generated dialogues keeps decreasing (moving close to human level) even though we no longer notice improvements in TS, whose highest value is reached well before. Figure 2 (solid lines) reports the details of the linguistic metrics used to compute LD. We see that for all tasks a high entropy is reached already after a few epochs; this means that though the number of words used is small, models learn to distribute their use well. All the other metrics improve through the epochs quite a lot. For MutualFriends, we do not compute MO and GRQ since the model trained on it, DynoNet, asks questions referring to different attributes and hence, by design, it generates very few repetitions. From the results of this first experiment, it emerges fairly clearly that in all referential games we have considered, **models learn to perform well on the task quite quickly.** On the one hand, this means that **choosing the best model purely on the basis of its TS prevents the model from developing better linguistic skills**, on the other hand, that **the higher quality of the dialogues does not help reach a higher TS.** This result holds in all cases despite the target being an entity in a graph described by linguistic attributes (MutualFriends), an object (GuessWhat) or an image (GuessWhich).

Comparison by Downsizing the Training Set To understand whether the relation between TS and LD is related to the size of the training set, we compare models trained on datasets of decreasing size. We evaluate the models by training them with 50% and 25% of the standard GuessWhat and GuessWhich datasets. For MutualFriends, we have not run the downsizing analysis since the dataset is too small. For readability reasons, in Figure 2 we report only the results obtained with the 25% setting since they represent the observed pattern well enough. The y-axis reports the metrics scaled between 0 and 1. In GuessWhich the TS (yellow lines) does not decrease by downsizing the dataset: when using just 25% of the full dataset (dotted line) it gets very close to the highest MPR obtained by the model trained on the full dataset (solid line) already after the first 5 epochs. Interestingly, the linguistic metrics do not get worse ei-

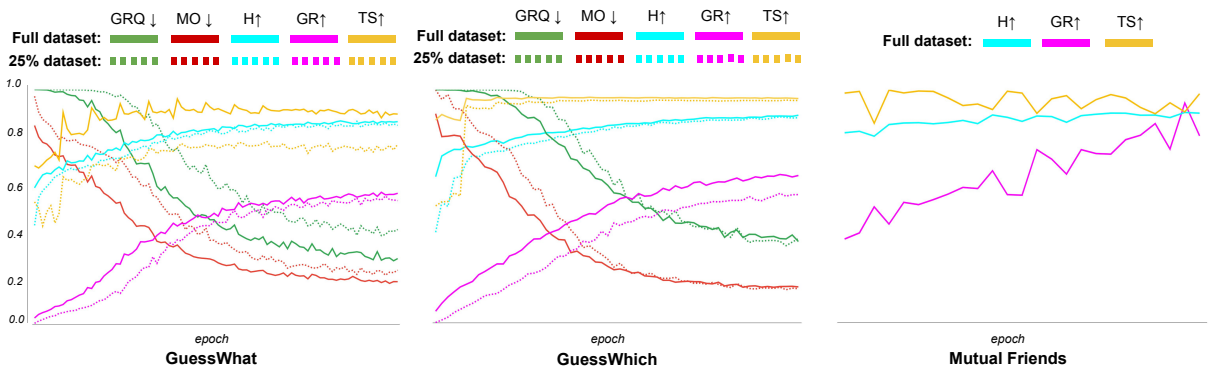


Figure 2: Comparison across epochs and by downsizing the training data using the following metrics: Task Success (TS), Games with Repeated Questions (GRQ), Mutual Overlap (MO), Unigram Entropy (H) and Global Recall (GR); all metrics are scaled between 0 and 1 (y-axis). **Left:** GDSE-SL on GuessWhat. **Middle**: ReCap on GuessWhich. **Right**: DynoNet on MutualFriends. Downsizing the training data has a higher impact, both for TS and some linguistic metrics, in GuessWhat than in GuessWhich. Among the linguistic metrics, entropy is the most stable and GR increases through the epochs in all tasks. For readability, we have not reported Local Recall-d since its pattern is very close to GR. The dataset of MutualFriends is too small to analyse the effect of downsizing it.

ther, with the only exception of GR. However, in GuessWhat the TS decreases when downsizing the training data (again, yellow solid vs. dotted lines) and dialogues quality is affected too (with the exception of entropy and GR). This result shows that **in GuessWhat *how well* the model learns to ground language plays an important role and affects the TS.** In the next experiment, we aim to further understand the difference between the two visual tasks and when LD could impact TS in GuessWhat.

5.3 When Could Language Quality Impact Task Success?

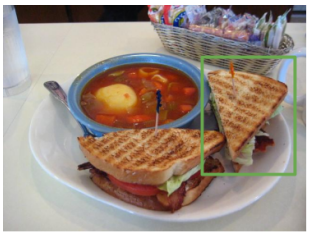
First of all, we check the extent to which the dialogue is used by the Guesser module to pick the correct target. Secondly, we evaluate whether the quality of the dialogues could lead to higher task success. Finally, we pinpoint when a lower LD could contribute to succeed in the task.

The role of the dialogues on TS We run a by-turn evaluation checking whether the information incrementally gained through the dialogue brings increased performance. We evaluate ReCap on GuessWhich and GDSE-SL on GuessWhat. We find that the performance of ReCap is flat across the dialogue turns, confirming results reported in Murahari et al. (2019) for other models. Instead, the performance of GDSE-SL keeps on increasing at each turn from the beginning till the end of the dialogue, though the increase in the first 3 turns is higher than in the later ones (details in Appendix A). **This suggests that in GuessWhich the role of the dialogue is rather limited.** This might be due to the highly informative image caption that GuessWhich models receive together with the dialogue to solve the guessing task (Testoni et al., 2019). Instead, in **GuessWhat dialogues do play a major role in the guessing task**. Hence, we focus on this dataset to understand whether and when the quality of the dialogue could lead to a higher task success.

Impact of the quality of dialogues on TS To check whether the Guesser could profit from dialogues of better quality, we evaluate GDSE-SL using human dialogues. When given human dialogues, the model reaches an accuracy of 60.6%, which is +8.5% higher than the one it achieves with the dialogues generated by its decoder (52.1%). One hypothesis could be that this higher TS is due to the mistakes produced by the A-bot when using instead the generated dialogues, but this is not the case: we have evaluated the model when receiving human questions paired with the A-bot's answers for each question and the accuracy drops of only 2.5%. This experiment suggests that **a lower LD could indeed lead to a higher TS.**

The role of less frequent words As we have observed above, models mostly use very frequent words. Here, we aim to understand to what extent this penalizes GuessWhat models. In this dataset, more than half (55%) of the words in the vocabulary are used less than 15 times in the training set. We refer to this set of words as "*rare*" words: most

Human dialogues		Generated dialogues	
Questioner	**Answerer**	**Questioner**	**Answerer**
1. Is it a man?	No	1. Is it a person?	No
2. Is it food?	Yes	2. Is it food?	Yes
3. Is *pancake*?	No	3. Is it pizza?	Yes
4. Is egg?	Yes	4. Is it the pizza in front?	Yes
		5. . . .	
⇝ **model suceeds guessing**		⇝ **model fails guessing**	



Human dialogues		Generated dialogues	
Questioner	**Answerer**	**Questioner**	**Answerer**
1.Is it edible?	Yes	1. Is it food?	Yes
2.Is it a sandwich?	Yes	2. Is it a sandwich?	Yes
3. Does it have an orange *toothpick* in it?	Yes	3. Is it on the right?	Yes
		. . .	
⇝ **model fails guessing**		⇝ **model succeeds guessing**	

Figure 3: Examples of GuessWhat games in which humans use "rare" words (rare words in italic) and the corresponding generated dialogues. The failure of the model could be due to the inability to generate (top) or to encode (bottom) rare words.

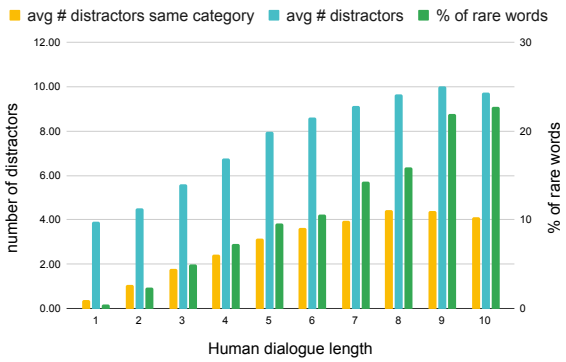


Figure 4: In GuessWhat, longer human dialogues contain more rare words, more distractors and more distractors of the same category of the target object.

of them are nouns (79%) or verbs (11%) (e.g., "*feline*", "*forest*", "*compute*", "*highlight*").

We check whether there is a relation between rare words and difficult games. Human dialogue length is a good proxy of the difficulty of the games, both for humans and models. Figure 4 illustrates some statistics about human dialogues: games for which humans ask more questions are about images with a higher number of distractors, with a higher number of distractors of the same category of the target, and with a higher number of "rare" words. In 10% of the games in the test set, humans have used at least one rare word. These dialogues are longer than those that do not contain rare words (resp., 7.8 vs. 4.7 turns on average). Interestingly, the accuracy of the model on these games is lower than the overall accuracy: -13.8% (48.7% vs. 62.5%) when evaluating it with human dialogues and -8.3% (45% vs. 53.3%) when using the dialogues generated by the model itself. Moreover, the accuracy reached by the model in the latter setting is lower when comparing games for which humans have used a higher number of rare words. Overall, we found that 65% of the rare words in the human test set show up in games that the model is not able to solve correctly.

Figure 3 shows some examples of games in which humans have used a rare word. It illustrates the human vs. generated dialogue and whether the model succeeds in guessing the target object when receiving the former or the latter. The failure of the model in guessing the target object could be due to its inability to generate or encode rare words. The example on top shows that if the model fails to generate an appropriate word (e.g. the rare word "*pancake*") this can have a domino effect on the next words and the next questions it generates. On the other hand, the model can fail to encode rare words, e.g., "*toothpick*" in Figure 3-bottom. The inability to generate rare words could be mitigated by developing dialogue strategies that produce less natural but still informative dialogues. For instance, in the example at the bottom, the model avoids using "*toothpick*" by asking a spatial question ("*Is it on the right?*") which is rather informative for the Guesser since it has the coordinates of each candidate object. These observations show that current models fail to properly ground and use

rare words and suggest that, in some contexts, **the use of only frequent words could be behind the failure in reaching the communication goal.**

6 Conclusion

Our work highlights the different complexity of two sub-tasks involved in referential guessing (visual) games: guessing the target and asking questions. We have shown that while learning to win the game can be reached in a rather short time (with a few epochs), learning to generate rich human-like dialogues takes much longer. This holds for all three tasks and models we have scrutinized independently of the size of the vocabulary, the task, and the learning paradigm used. Therefore, choosing the best model only on the base of the task success could prevent the model from generating more human-like dialogues. We have shown that in GuessWhich decreasing the size of the training set does not bring a drop in either TS (task success, higher is better) or in LD (linguistic divergence, lower is better) and, moreover, the dialogues play a minor role on TS. Instead, for GuessWhat, decreasing the size of the training dataset brings a decrease in TS and an increase in LD, and, through dialogues, models accumulate information to succeed in the task. Hence, we have focused our in-depth analysis on GuessWhat. Furthermore, we have investigated whether and when higher language quality could lead to higher task success. We have shown that if models are given human dialogues, they can reach a higher TS. Hence, LD could boost TS. We have shown that this boost could help more in difficult games, i.e. those for which humans ask longer dialogues. These games contain images with more distractors and humans use less frequent words while playing them. Hence, we claim that in GuessWhat models could increase their accuracy if they learn to ground, encode and decode words that do not occur frequently occur in the training set.

In the paper, we propose the LD metric that, despite its limitations (i.e., being based only on surface cues) represents a proxy of the quality of dialogues. We believe LD effectively captures the most common deficiencies of current models and it allows a straightforward comparison between different models. As future work, LD can be used as a training signal to improve the quality of generated dialogues. Moreover, a comparison between human quality judgments and LD may shed some light on the strengths and weaknesses of this metric. Further work is needed to design new metrics that capture more fine-grained phenomena and better evaluate the quality of generated dialogues.

Acknowledgements

The authors kindly acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research. We are grateful to SAP for supporting the work. We would like to thank the following people for their suggestions and comments: Luciana Benotti, Guillem Collell, Stella Frank, Claudio Greco, Aurelie Herbelot, Sandro Pezzelle, and Barbara Plank. Finally, we thank the anonymous reviewers for the insightful feedback.

References

Ehsan Abbasnejad, Qi Wu, Javen Shi, and Anton van den Hengel. 2019. What's to know? Uncertainty as a guide to asking goal-oriented questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4155–4164.

Elia Bruni and Raquel Fernández. 2017. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*, pages 169–174.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *2017 IEEE International Conference on Computer Vision*, pages 2951–2960.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474.

Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram.

2017. Topic-based evaluation for conversational bots. In *Proceedings ofthe 31st Conference on Neural Information Processing Systems (NIPS)*.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701.

H. He, A. Balakrishnan, M. Eric, and P. Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1766–1776.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *to appear in Proceedings of ICLR 2020*.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell Me More: A Dataset of Visual Scene Description Sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157.

Anjuli Kannan and Oriol Vinyals. 2016. Adversarial evaluation of dialogue models. In *NIPS 2016 Workshop on Adversarial Training*.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2019. Don't say that! Making inconsistent dialogue unlikely with unlikelihood training. ArXiv:1911.03860.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2019. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 462–472.

Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1449–1454.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *Proceedings of 34th AAAI Conference on Artificial Intelligence*.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5863–5868, Hong Kong, China. Association for Computational Linguistics.

Lee Sang-Woo, Gao Tong, Yang Sohee, Yao Jaejun, and Ha Jung-Woo. 2019. Large-scale answerer in questioner's mind for visual dialog question generation. In *Proceedings of International Conference on Learning Representations, ICLR*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and

Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587.

Pushkar Shuklar, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, William Yang Wang, and Matthew Turk. 2019. What should I ask? Using conversationally informative rewards for goal-oriented visual dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451.

Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2765–2771.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Alberto Testoni, Ravi Shekhar, Raquel Fernández, and Raffaella Bernardi. 2019. The devil is in the detail: A magnifying glass for the GuessWhich visual dialogue game. In *Proceedings of the 23rd SemDial Workshop on the Semantics and Pragmatics of Dialogue (LondonLogue)*, pages 15–24.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *ICML Deep Learning Workshop*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Self-supervised dialogue learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3857–3867.

Jiaping Zhang, Tiancheng Zhao, and Zhou Yu. 2018a. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 140–150.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018b. Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)*, pages 186–201.

Rui Zhao and Volker Tresp. 2018. Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient. In *Proceedings of IJCAI*.

Mingyang Zhou, Josh Arnold, and Zhou Yu. 2019. Building task-oriented visual dialog systems through alternative optimization between dialog policy and language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 143–153.

A Appendix A

Figure 5 reports the token frequency curve for human dialogues and generated dialogues on the GuessWhat test set (Zipf's law). Human dialogues are clearly more rich and diverse compared to generated dialogues.

Figure 6 and Figure 7 show the per-turn accuracy of GDSE-SL for GuessWhat and ReCap for GuessWhich, respectively. For GuessWhat, we report the simple task accuracy on the game, while for GuessWhich we use the Mean Percentile Rank; please refer to the main paper for additional details. For GuessWhat, the accuracy keeps increasing while new turns are given as input to the model. For GuessWhich, on the other hand, the Mean Percentile Rank (MPR) is pretty stable after very few turns and it is already high at turn 0, i.e. when only the caption is provided without any dialogue history.

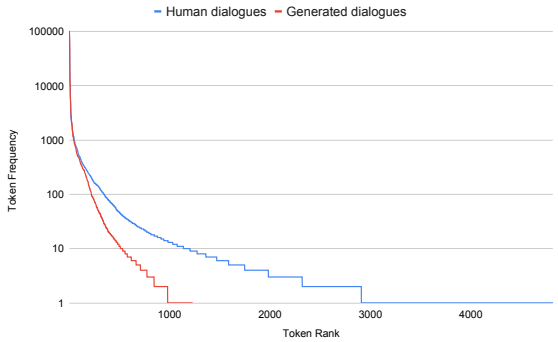


Figure 5: Token frequency plot (Zipf's Law curve) for human vs. generated dialogues

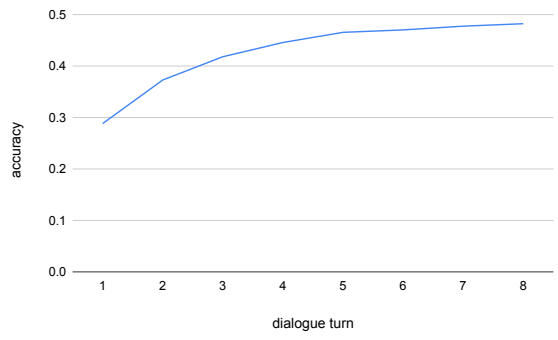


Figure 6: GDSE-SL per-turn accuracy on the GuessWhat game.

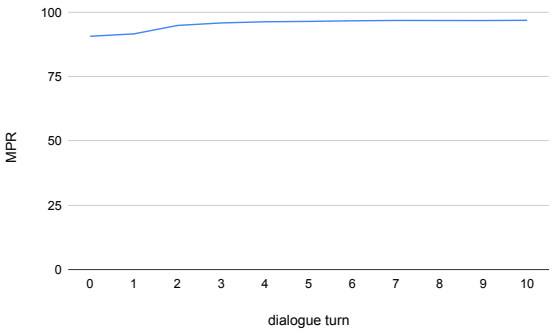


Figure 7: ReCap per-turn Mean Percentile Rank (MPR) on the GuessWhich game