Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5768–5777
Marseille, 20-25 June 2022

Strategy-level Entrainment of Dialogue System Users in a Creative Visual Reference Resolution Task

Deepthi Karkada[1], **Ramesh Manuvinakurike**[2],
Maike Paetzel-Prüsmann[3], **Kallirroi Georgila**[4]
[1]Intel Corp, USA, [2]Intel Labs, USA,
[3]University of Potsdam, Germany, [4]University of Southern California, USA
{deepthi.karkada, ramesh.manuvinakurike}@intel.com,
paetzel-pruesmann@uni-potsdam.de, kgeorgila@ict.usc.edu

Abstract

In this work, we study entrainment of users playing a creative reference resolution game with an autonomous dialogue system. The language understanding module in our dialogue system leverages annotated human-wizard conversational data, openly available knowledge graphs, and crowd-augmented data. Unlike previous entrainment work, our dialogue system does not attempt to make the human conversation partner adopt lexical items in their dialogue, but rather to adapt their descriptive strategy to one that is simpler to parse for our natural language understanding unit. By deploying this dialogue system through a crowd-sourced study, we show that users indeed entrain on a "strategy-level" without the change of strategy impinging on their creativity. Our work thus presents a promising future research direction for developing dialogue management systems that can strategically influence people's descriptive strategy to ease the system's language understanding in creative tasks.

Keywords: entrainment, dialogue systems, knowledge graphs

1. Introduction

Linguistic entrainment (also known as alignment, coordination, priming, convergence, accommodation, and adaptation) is the phenomenon in which interlocutors start speaking more similarly to each other (Brennan and Clark, 1996; Rahimi et al., 2017). Entrainment can happen on multiple levels such as prosody, speech rate, vocabulary usage, and syntactic and stylistic patterns (Iio et al., 2009; Levitan and Hirschberg, 2011; Levitan, 2013). Entrainment has been associated with dialogue and task success in human-human (Reitter and Moore, 2007; Friedberg et al., 2012) and human-agent (Fandrianto and Eskenazi, 2012; Lopes et al., 2013) interaction. So far entrainment has been studied in relatively constrained domains with a limited vocabulary size. However, in more complex dialogue system (DS) applications, where a large vocabulary size and variety or unpredictability of user responses would be challenging for automatic speech recognition (ASR) and natural language understanding (NLU), entrainment can potentially be even more beneficial.

In this paper, we focus on entrainment in the context of a pedagogical game called RDG-Map (Paetzel et al., 2020). RDG-Map is a collaborative game between two players, a Director and a Matcher, engaging in a goal-oriented conversation to identify a country on the world map. The Director gives descriptions for a country so that the Matcher can locate and select it on their map. The game is time-constrained to generate natural and fluid conversations similar to everyday dialogue. While directional descriptions (above USA, southwest of Zambia, etc.) and referring to the countries by name are the most unambiguous for a given target country, the knowledge of the average player about the location of countries on the map is very limited, which makes references to neighbouring countries difficult to resolve. Thus, people often use shape and size descriptions of a country to disambiguate between several countries in the same region. Consequently, an autonomous DS playing the role of the Matcher in the game needs visual reference resolution. Users can, however, be very imaginative in their descriptions (e.g., "[Germany] looks like a pac-man", "if Somalia had a nose it would blow towards the ocean") which makes this a creative task. For a machine learning model, these descriptions are difficult to infer because every human's description of shapes is distinct and subjective, and the reasoning and general knowledge required to resolve such complex references to scenes in the real world exceeds the capabilities of current AI systems.

In this work, we explore the use of entrainment to see if we can prime users to provide shape descriptions using simple geometric shapes such as rectangles, triangles, circles, and squares, and whether this helps with the NLU task (since it would potentially result in a more limited vocabulary used by the human interlocutor). Basically our goal is to influence the strategy used by users for providing shape descriptions. Instead of letting them use their imagination freely we guide them towards using simple geometric shapes, and thus we call this process "strategy-level entrainment". At the same time, we do not want to impinge on the users' creativity since creative descriptions can help distinguish between similarly shaped countries in the same neighbourhood and can also be an indicator of engagement.

To study entrainment, we use an autonomous DS with an NLU based on a knowledge graph that can play the role of the Matcher in the RDG-Map game. The DS is deployed on the web and paired with crowd-workers recruited on Amazon Mechanical Turk (AMT) to play the role of the Director and provide country descriptions to the system.

The aim of our study is to first understand whether the DS can actually influence people's descriptive strategy by suggesting geometric shape descriptions for a country in the form of a question to the human player. We hypothesize that *players entrain to general abstract concepts*, which means that in the post-entrainment phase they use more geometric shape descriptions than before the entrainment (H1). In an offline-analysis, we then feed the pre- and post-entrainment descriptions back to our knowledge graph to understand whether our NLU's recognition rate improves when people converge to using geometric shape descriptions. We hypothesize that *the entrainment does improve the NLU performance of the DS* because it can guide users to use a more well-defined vocabulary set (H2). Finally, we want to understand whether people's level of creativity decreases when entraining to the descriptive strategy suggested by the DS. We hypothesize that *despite the limited vocabulary the creativity in the descriptions is not harmed by the entrainment* (H3).

While our work builds on previous work on human-human and human-agent entrainment, it is *novel in its aim to make human interlocutors exchange their entire descriptive strategy instead of individual phrases in a creative task*. In addition, we *implement the entrainment into a fully-functioning DS generating the entrainment questions fully autonomously*, which allows us to study entrainment in an actual human-agent interaction setting.

2. Related Work

Entrainment in the context of dialogue systems has been extensively studied in the literature. Brennan and Clark (1996) investigated the effect of reducing lexical variability through conceptual pacts. They observed that people come to use the same terms when repeatedly referring to the same object in a conversation, and named this phenomenon lexical entrainment. Iio et al. (2009) is one of the few studies of entrainment in human-robot interaction which found two types of lexical entrainment: entrainment per term (users adopting a term of the robot) and entrainment per type of term (users adopting the same type of term as the robot). Brandstetter et al. (2017) studied lexical entrainment in the context of human-robot interaction and observed that the lexical convergence phenomenon persisted even beyond the interactions in the study. Mitchell et al. (2012) focused on entrainment in tutorial dialogue and found that convergence increased longitudinally and that it could be possible to leverage convergence to positively influence students' emotions. Mizukami et al. (2016) investigated the effects of entrainment on dialogue acts and on lexical choices given a dialogue act. Rahimi et al. (2017) studied the relationship between lexical and acoustic-prosodic entrainment in a multi-party setting. Rahimi and Litman (2020) proposed a novel graph-based vector representation of multi-party entrainment in the same domain.

With respect to dialogue success in human-human interaction, based on analysis of human-human dialogues, Reitter and Moore (2007) found that lexical and syntactic repetition can reliably predict task success. Friedberg et al. (2012) showed that the lexical entrainment of high performing student groups increased with time while entrainment of low performing student groups decreased with time. Nenkova et al. (2008) investigated how entrainment is associated with dialogue naturalness, flow, and task success.

In terms of human-agent interaction, Stoyanchev and Stent (2009) showed that users adapt to the system's lexical and syntactic choices and that system prompts can be used to guide users to produce utterances that are easier for the system to process resulting in higher task success. Porzel et al. (2006) showed that in certain situations it is important that multi-modal dialogue systems adapt linguistically to users. They analyzed human-human and human-agent (remote-controlled by a human Wizard of Oz) conversations. Fandrianto and Eskenazi (2012) studied entrainment strategies of two speaking styles, shouting and hyperarticulation. When shouting was detected, the system responded more softly and when hyperarticulation occurred, the system spoke faster. These strategies helped the user to return to more neutral utterances and explicit prompts and backing off to another objective resulted in modest task success improvements. Lopes et al. (2013) aimed to increase dialogue success by adapting the system's lexical choices to the user's lexical choices. This approach led to reduced error rates and number of turns. Levitan (2013) described the various forms of entrainment that exist such as lexical, syntactic, stylistic, acoustic-prosodic and phonetic entrainment and studied the effect of acoustic/prosodic entrainment in both human-human and human-agent (Wizard of Oz) interactions.

To our knowledge there is no previous work on the effect of entrainment in creative tasks. As summarized above, previous work has instead focused on relatively constrained tasks with a limited vocabulary where people replace their choice of words with simple substitutes. The work closest to ours is that of Bergqvist et al. (2020), also in the RDG-Map domain, where the system tries to prime the user with cardinal directions (e.g., when users said "above" the system would prime them to say "north of" instead of using a simple substitute word). This work showed convergence of the users' vocabulary towards the vocabulary used by the system even if the replacement term is non-trivial. In our work, the task is more creative and our entrainment strategies are much less restrictive. Users are primed to describe countries in simple geometric shape terms but there is still much variety and unpredictability in their descriptions, which, to the best of our knowledge, has not been done before. *Hence, the main contribution of our work is that we investigate if users can be entrained on general abstract concepts in a creative task.*

In this work, we also discuss how entrainment can be included into a fully autonomous DS that leverages a knowledge graph (KG) as part of the natural language understanding (NLU) module. In the past years, there has been a lot of work on utilizing KGs for question-answering, e.g., Sun et al. (2018), Lin et al. (2018), Huang et al. (2019), Hakkani-Tür et al. (2014), as well as in chatbots and task-oriented dialogue to integrate external knowledge and help learn more robust and coherent dialogue policies, e.g., He et al. (2017), Zhou et al. (2020), Jung et al. (2020), Yang et al. (2020b), Yang et al. (2020a), Xu et al. (2020). We are not aware of previous work on utilizing KGs from publicly available crowd-authored content (Wikidata) for integration into a reference resolution module of a DS. The work on Visual Genome (Krishna et al., 2017) seems to be the closest to ours as it captures objects in images as well as their relationships. But our domain is different from visual question-answering and focuses on entrainment and creativity, which provides a novel use-case for knowledge graphs in dialogue systems.

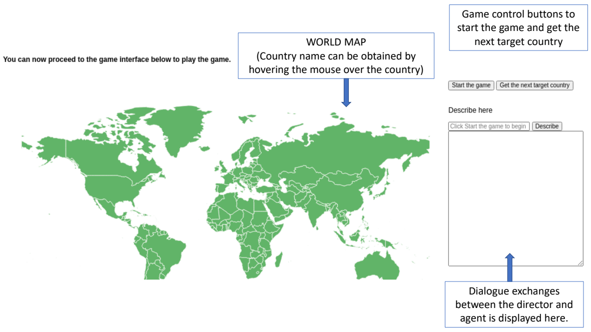
As mentioned above, we are not aware of previous work on the effect of entrainment on creativity. Previous works have used metrics such as Eigenvector Centrality (EVC) (Bonacich, 2007; Chen et al., 2020), Novelty metric based on Cosine Similarity (NCS), and Novelty metric based on Jaccard Similarity (NJS) (Tsai et al., 2010) as a measure of creativity in a user's utterances. We leverage these metrics to understand whether the user's creativity decreases when entraining to the descriptive strategy suggested by our system.

3. The Dialogue System

While the RDG-Map game (Paetzel et al., 2020) can be played between two human players, in this work, we explore an online setup of the game in which a human Director and an autonomous agent Matcher play via textual exchanges (see Figure 1). In this section, we describe the general setup of the DS, before discussing how we included entrainment into the dialogue manager in Section 4

3.1. Corpora

The human Directors adopt various strategies to describe the target country, e.g., naming the continent ("it's in Africa"), giving directional relations to other countries ("it's above Egypt"), a landmark ("this is near the Mediterranean sea"), size ("biggest one in Africa"), or describing the shape ("almost looks like a pac-man"). In order to build an autonomous Matcher, the NLU model needs to understand all of these complex descriptions and hence requires large amounts of data for training. Since collecting human-human and human-agent conversations is time-consuming and costly, we augment the data with publicly available information from Wikidata and a focused online data collection for subjective shape descriptions. The data sources used for training the agent Matcher hence consist of three parts (see Table 1 for an overview): (i) Spoken human-wizard conversations (D_{spoken}), (ii) Wikidata for the location and directional relations between countries (D_{wiki}), and (iii) additional written shape descriptions for countries collected on AMT (D_{shapes}).



You can now proceed to the game interface below to play the game.

WORLD MAP
(Country name can be obtained by hovering the mouse over the country)

Game control buttons to start the game and get the next target country

Dialogue exchanges between the director and agent is displayed here.

Director	Matcher
Pakistan.	
	I don't know where that is.
It's in Asia.	
	I need more information.
It's northwest of India.	
	What does it look like?
It looks like a T-Rex.	
	Got it.

Figure 1: The experiment setup along with a sample conversation between the Director and Matcher.

Spoken Human-Wizard Conversations (D_{spoken}) This part of the dataset contains the human-wizard annotated conversations collected by Paetzel et al. (2020) between a human Director and a remote-controlled (wizarded) agent Matcher. The data contains spoken dialogue interactions which are first automatically transcribed using Sonix.ai[1] and then manually corrected. The data is then annotated by experts according to the annotation scheme detailed in (Paetzel et al., 2020). In total, the dataset contains the dialogues of 80 users and is split into a training (64 users) and test set (16 users).

Wikidata (D_{wiki}) Wikidata (Vrandečić and Krötzsch, 2014) is an openly available knowledge base that we use to augment the RDG-Map data with objective information about the world map. The KG extracted from Wikidata consists of *Entities* (i.e., countries, areas, capitals, major geographical landmarks, continents, neighbours, latitude and longitude) and *Relations* which capture the relationships between the *Entities*. The relations in the KG can be mapped directly to the target description strategies in the human-wizard conversations data, e.g., "it's in Asia" results in the mapping *Target - InContinent - Asia*. A sample section of the KG is shown in Figure 2.

[1] https://sonix.ai/

D_{spoken}	# Users	80
	# Dialogues	980
	# Turns	3989
	# Tokens	24726
	# Shape descriptions	905
D_{wiki}	# Entities (KG)	2868
	# Relations (KG)	38
D_{shapes}	# Tokens	6496
	# Shape descriptions	1563

Table 1: Statistics of the datasets used in this work.

Shape Descriptions (D_{shapes}) The RDG-Map corpus contains a limited number of shape descriptions (see Table 1). Such creative shape descriptions are also not available from objective datasets like Wikidata, and since such descriptions are a challenge for the NLU, we hence augment the dataset by collecting additional shapes data using crowd-sourcing[2]. Workers on AMT were shown the world map with one of the countries highlighted similar to the Director's screen from the RDG-Map setting. They were then asked to provide three creative shape descriptions for each country. We paid \$0.03 per description. On average, each description has a length of 4.15 words. We collected data from 619 participants for all countries visually identifiable on the world map. The data was filtered by the authors to remove non-shape descriptions. Each description provided by the crowd-workers is appended to the KG constructed in D_{wiki} with a *'HasShape'* relation.

3.2. Dialogue System Implementation

The core of the artificial agent autonomously playing the role of the Matcher in the RDG-Map game is the language understanding based on knowledge graphs. The NLU receives written text as input and provides the most likely country the Director is describing along with confidence scores across all entities as the output. These values are then fed into the dialogue manager, which decides whether the confidence is high enough to confirm the current selection, ask questions or request that the Director should move on to the next target.

The input utterances received by the NLU are first converted to an embedding vector of size 256 and then passed through a classifier which consists of two LSTM layers (each containing 128 hidden units) and a fully connected layer with sigmoid activation. We use a dropout ratio of 0.1 for the outputs of the LSTM layers. This relations classifier is trained for 25 epochs using the training data from D_{spoken} and tested on the held-out data. The classifier accuracy is 89.33%. It is important to note that, while a better relations classifier can be created and further hyperparameter tuning can be performed, that is not the focus of this paper.

The knowledge graph is a ConvE (Dettmers et al., 2018) model constructed using D_{spoken} and D_{wiki} that requires the source entity and the relations as input. It encodes the entities and the relations as embeddings which are then passed through the convolutional neural network to predict the target entities. For identifying the source entity, we initially extract the names of the countries utilizing an off-the-shelf BERT-based Named Entity Recognizer (NER) (Peters et al., 2017). Since the subjective shape descriptions cannot be identified by the NER, we also use an entity-linking approach to identify the source entity for the KG inference in case of the *HasShape* relation. For entity linking, we obtain the embeddings for the user descriptions from the ConvE model and then choose the source entity by linking the entity with the highest cosine similarity.

Algorithm 1: Agent dialogue management

```
Data: $tgt \leftarrow$ Current target
$MAX_{turns}$, $Select_{threshold}$
$Q = \{(r_k, q_k)\}$, Dictionary mapping the relations to query
$U = \{u_0, u_1, \ldots, u_n\}$, 'n' number of user utterances for the current target
$C = \{c_0, c_1, \ldots, c_m\}$, 'm' number of possible targets
$Sn = \{s_{c0}, s_{c1}, \ldots, s_{cm}\}$, Confidence scores for each of the 'm' possible targets after processing 'n' user utterances for the current target
Result: $Resp$, System response;
$i \leftarrow 0$;
$R \leftarrow \{\}$ ;
for $u_i$ in $U$ do
    if $i \leq MAX_{turns}$ then
        if $max(S_n) \geq Select_{threshold}$ then
            $Resp \leftarrow$ Got it & Return ;
        else
            $Resp \leftarrow$ Let's skip & Return;
        end
    else
        $e_i \leftarrow Entity\_Extraction(u_i)$ ;
        $r_i \leftarrow Relations\_classifier(u_i)$ ;
        $KG_{scores} \leftarrow CONVE(e_i, r_i)$ ;
        $S_n \leftarrow Aggregate(S_n, KG_{scores})$ ;
        Add $r_i$ to $R$ ;
    end
    $i \leftarrow i + 1$;
end
if $HasShape \notin R$ And Entrainment Phase then
    $Resp \leftarrow ENTRAIN(tgt)$ ;
else
    $Resp \leftarrow Random(tgt, Q)$ ;
end
```

The target description from the Director spans multiple turns and thus the confidence scores from each turn need to be aggregated before the agent can make a target country selection. The KG generates a vector of confidence scores over all the predicted entities for each turn. The dialogue manager aggregates these con-

[2] We will make this corpus publicly available in the future.

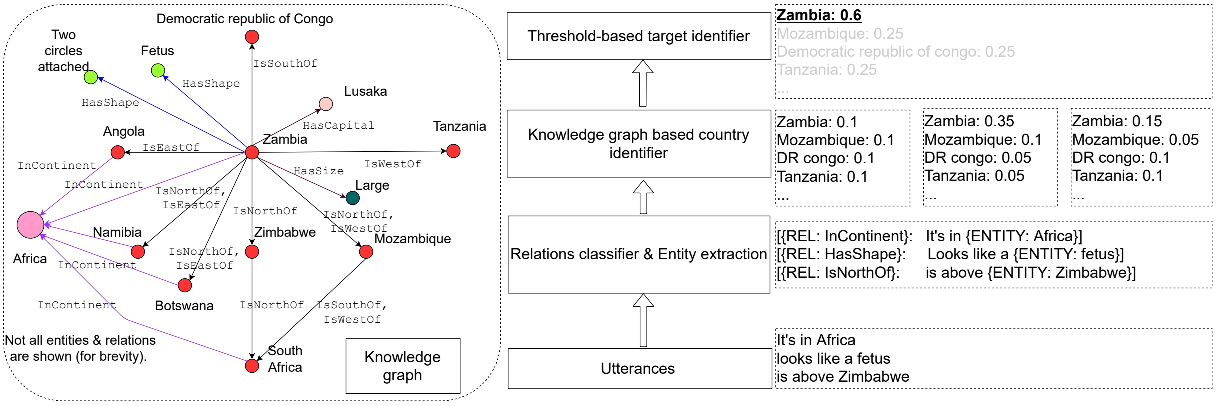


Figure 2: Part of the knowledge graph (left) and the inference pipeline followed by the agent in this work (right).

fidence scores and generates a single confidence score for each country on the map. The agent Matcher always selects the country with the current highest aggregated confidence. This selection is, however, not visible to the human Director, who requires verbal confirmation to move on. If the aggregated score is greater than a predefined target selection threshold ($Select_{threshold}$ = 0.5) the agent says 'got it' indicating to the human user that it is confident enough in its selection. If at the end of MAX_{turns} the confidence score is less than the threshold the agent "skips" the country by requesting the user to move on to the next target. Figure 2 shows the agent architecture along with the entities and the relations captured by the KG representation. Algorithm 1 shows the agent's dialogue manager design.

4. Entrainment

To investigate whether the agent can entrain users to generate shape descriptions using a pre-defined strategy, we divided the game into three phases, the *pre-entrainment phase*, the *entrainment phase*, and the *post-entrainment* phase. In each of the phases the users cover four unique target countries. The target countries used for each of the phases are shuffled continuously. This is to avoid affinitization of the use of geometric shapes with only certain countries. The division into the phases was unbeknown to the players. In the pre- and post-entrainment phases, the agent always asks the same questions inquiring about the location, shape, continent and the landmark. Questions regarding the shape of the country are phrased as an open question, i.e., "What does it look like?". In the entrainment phase, the only difference is that the shape question is a yes/no question querying about a specific geometric shape description linked in the knowledge graph. For instance, if the target hypothesis is 'Turkey', the agent asks: "Does it look like a rectangle?".

We hypothesize that by entraining the users to provide creative shape descriptions but limit their vocabulary to common geometric shapes, descriptions between cluegivers would be more similar and it would hence be easier for an NLU module to identify the target.

5. Experimental Setup and Analyses

We designed an experiment to answer three main research questions linked to our hypotheses stated in the introduction: (RQ1) Can users be entrained to use a specific strategy when giving subjective descriptions of the country's shape? (RQ2) Can entrainment help the dialogue system with the target country identification? (RQ3) Does entrainment impact the Director's creativity in describing the target country?

5.1. Experimental Setup

We recruited 31 participants on Amazon Mechanical Turk to play the game with the agent Matcher. Each user was paid $3.00 for a 10 minute game. The users were required to be native English speakers with a qualification criteria of greater than 85% approval rate from at least 100 tasks. The experiment interface is shown in Figure 1. The agent with pre-trained models was deployed on a server with Intel Xeon CPU and no GPU with the model providing inference in real time. We excluded data from 3 users, one having repeated the study and the other 2 providing content copied from the web.

5.2. Analyses

Entrainment Analysis (RQ1) To answer the question if the users entrained on the geometric shapes, the utterances obtained from the pre-entrainment and post-entrainment phases were annotated by expert annotators. For each utterance, the annotators marked whether it contains a geometric shape description (yes/no label). A pre-defined list of geometric shapes was obtained from Wikipedia (Wikipedia, 2021a; Wikipedia, 2021b) (containing Triangle, Rectangle, Circle, Oval, Trapezium, Square, Line, Cube, Crescent, Pentagon, Hexagon, Cone, Sphere) and used for annotating the utterances to ensure consistency between the annotators. Descriptions in which geometric shapes were not explicitly named (e.g., "elongated shape from north to south" - which implicitly describes a vertical bar) were not marked as a geometric shape. We found Cohen's kappa = 0.96 implying almost perfect agreement between the human annotators.

Agent Performance (RQ2) We first measure the general target selection NLU accuracy of our agent pipeline on the test data from D_{spoken}. In order to compare the performance of our NLU with the language understanding abilities of human Matchers, we conducted a second evaluation on AMT. In this, we showed crowd-workers (same qualification criteria as mentioned in D_{shapes} applied; pay=\$0.10 per task) descriptions given by the human Directors in D_{spoken} and asked them to provide their top-1 and top-3 guesses of the target country. For the agent Matcher, we measure the NLU accuracy by checking if the target country is present in the top-1, top-3, or top-5 predictions. Since, the crowd-workers have provided their top three guesses for the descriptions, we can compute the human accuracy only for their top-1 and top-3 choices.

In a second step, we measure if entrainment could help improve the target selection accuracy of our NLU. The target country selection by the DS depends not only on the shape descriptions but also the descriptions received in other turns. To answer RQ2, we need to measure if the shape descriptions from the post-entrainment phase increase the accuracy of the target country selection compared to the shape descriptions received before the entrainment. We hence artificially generate (simulate) three non-shape target country descriptions and then append these with the shapes descriptions from the pre- and post-entrainment phases. An example of this is shown below:

Pre-entainment dialogue sample Human description from User 1 in the pre-entrainment phase:

——

Target: Zambia

——

DS: Which continent is it in?
User; Turn 1: It's in Africa.
{Rel: InContinent, Ent: Africa}
DS: Tell me more about the location.
User; Turn 2: It's above Zimbabwe.
{Rel: IsNorthOf, Ent: Zimbabwe}
DS: Is it next to a body of water?
User; Turn 3: It's land-locked.
{Rel: HasLocation, Ent: land-locked}
DS: What shape does it look like?
User; Turn 4: It looks like a spectacle.
{Rel: HasShape, Ent: a spectacle}

Agent prediction: Incorrect

Post-entainment dialogue sample Human description from User 2 in the post-entrainment phase:

——

Target: Zambia

——

DS: Which continent is it in?
User; Turn 1: It's in Africa.
{Rel: InContinent, Ent: Africa}
DS: Tell me more about the location.
User; Turn 2: It's above Zimbabwe.
{Rel: IsNorthOf, Ent: Zimbabwe}
DS: Is it next to a body of water?
User; Turn 3: It's below Democratic Republic of Congo.
{Rel: IsSouthOf, Ent: DRC}
DS: What shape does it look like?
User; Turn 4: It looks like two circles connected.
{Rel: HasShape, Ent: two circles connected}

Agent prediction: Correct

There are two main reasons for generating the Director's descriptions instead of only using the real descriptions obtained in our experiment. The first reason is to make the comparisons of the NLU performance between the pre-entrainment and post-entrainment utterances fair while measuring any potential gain in NLU performance due to entrainment. More complex utterances encountered in either the pre-entrainment or the post-entrainment phases could create confounding factors making the comparison unfair. Second, misclassification of relations and entities (due to disfluencies and typos) in the user descriptions could impact measuring the gains from the entrainment alone.

To generate these simulated descriptions for the target country we query the knowledge graph to retrieve three non-shape relation-entity pairs and convert these into natural language (e.g., *InContinent-Africa* is converted to "it's in Africa") by using templates, append them to the shape descriptions and measure the target country selection accuracy of our NLU.

Creativity (RQ3) Creativity in our work refers to the novelty in the shape descriptions, which is vital since it helps the NLU select the target country between similarly shaped countries. Note that creativity in our domain can relate to both the choice of vocabulary / descriptive strategy as well as the creativity in providing unique descriptions for a country given the same descriptive strategy. Our aim is to limit the first, i.e., entrain users to converge to the same descriptive strategy, while still ensuring that the resulting descriptions are unique for each country.

Measuring creativity in human-machine interactions is challenging, and in our work, we hence adopt three methods which have been used in previous works: (i) Eigenvector Centrality (EVC), (ii) a Novelty metric based on Cosine Similarity (NCS), (iii) a Novelty metric based on Jaccard Similarity (NJS). We adopt all three metrics to measure the creativity of the utterances obtained from the pre-entrainment phase of our experiments and compare it to the creativity of the utterances obtained from the post-entrainment phase.

EVC indicates connectedness of a node in a graph which is an indicator of creativity by representing natural language responses as nodes in a graph and measuring its centrality (Bonacich, 2007; Chen et al., 2020). We construct a separate lexical graph for the pre- and post-entrainment phases with each of the nodes represented by a sentence embedding vector constructed using pre-trained BERT embeddings (Devlin et al., 2019) and the edges are weighted by the cosine similarity between the nodes as given by Algorithm 2.

Algorithm 2: EVC computation

```
Data: $U = \{u_0, u_1, \ldots, u_n\}$, 'n' number of user
        utterances in the dataset
nx_graph = Empty networkx graph
Result: EVC score
for $u_i$ in $U$ do
    for $u_j$ in $U$ do
        $k_i \leftarrow BERT\_Embeddings(u_i)$
        $q_j \leftarrow BERT\_Embeddings(u_j)$
        $cos\_sim \leftarrow cosine\_similarity(k_i, q_j)$
        $nx\_graph \leftarrow add\_edge(i, j, weight = cos\_sim)$
    end
end
EVC = eigenvector_centrality(nx_graph)
```

The lexical graph is constructed in such a way that each node/utterance is connected to every other node/utterance similar to Chen et al. (2020) with one modification. While Chen et al. (2020) use TF-IDF to represent the node, we utilize BERT-embeddings. We then calculate the EVC for each node using the networkx library (Hagberg et al., 2008). A higher value of EVC indicates higher similarity to other nodes and consequently less creativity. As in Chen et al. (2020), we hence report the negative mean of EVC values of all the nodes in the graph as the overall creativity score.
Cosine similarity (NCS) and Jaccard similarity (NJS) measure similarity between vectors (Tsai et al., 2010). We calculate the pairwise cosine similarity and Jaccard similarity between all the utterances obtained in the pre-entrainment phase and post-entrainment phase separately. The novelty or creativity metric is then defined as 1 minus the cosine similarity or Jaccard similarity. Thus higher NJS/NCS indicates higher creativity.

6. Results

6.1. Entrainment Analysis (RQ1)

Out of the 28 unique users interacting with our agent, we find 36 descriptions containing geometric shapes in the pre-entrainment phase and 53 descriptions containing geometric shapes in the post-entrainment phase. A One Sample t-test reveals that our participants indeed use significantly more geometric shapes in the post-entrainment compared to the pre-entrainment phase, t = 4.5323, p $<$ 0.001. *We can hence conclude that the agent was indeed able to get users to adapt their descriptive strategy after the entrainment phase.*

6.2. Agent Performance (RQ2)

In the first step we analyze the overall NLU performance of our agent Matcher based on the test user dialogues obtained from the RDG-Map domain (D_{spoken}) and compare it with the performance of human Matchers provided with the same descriptions. In Table 2 we can see that the human crowd-workers can identify the target countries better than the KG-based pipeline. Interestingly, we see that the accuracy for the human Matchers is at the same level no matter if they can name their one or three most likely candidates. On the contrary, the agent Matcher's performance increases substantially when it can give three instead of one top prediction. This shows that the correct target country is often ranked high, but not highest in the KG, which presents an avenue for future work to improve upon.

Model	Top-1	Top-3	Top-5
Humans	0.62	0.63	-
Agent	0.29	0.47	0.57

Table 2: The results (accuracy) of the target country selection by humans and the agent developed in this work when asked to provide the top-1, top-3, and top-5 (agent only) predictions.

	Top-1	Top-3	Top-5
Baseline	0.673	**0.854**	0.873
Pre-entrainment	0.673	0.836	0.864
Post-entrainment	**0.682**	0.846	**0.882**

Table 3: Target selection accuracy measured separately for the top-1, top-3, and top-5 predictions of the agent using the utterances without shape (baseline) and with shape descriptions provided in the pre- and post-entrainment stages.

With our second analysis we aim to understand how the performance of our dialogue system changes between the shape descriptions obtained in the pre-entrainment and post-entrainment phase. The results are summarized in Table 3. The baseline in the table refers to the target clues without any shape descriptions present. The pre- and post-entrainment refers to the conversations with shape descriptions obtained in the pre- and post-entrainment phase respectively. We can observe two important points in these results: First, we can infer that *the presence of shape descriptions is not always beneficial for the NLU performance* (see the drop in accuracy in the top-3 predictions when shape descriptions are included). We discuss this in more detail in Section 7. Second, for the top-1 and top-5 predictions, we observe *a slight improvement in performance in the NLU in the post-entrainment phase*. This improvement, however, is not significant as a One Sample t-test reveals, t=1.6, p=0.2.

6.3. Creativity (RQ3)

The creativity scores for each of the three methods applied to the spoken conversations from D_{spoken} and our pre- and post-entrainment clues from the main experiment are shown in Table 4. A higher score indicates that the utterances were more creative. We

Condition	EVC	NJS	NCS
Spoken dialogue	**-0.187**	**0.861**	**0.528**
Pre-entrainment	-0.329	0.607	0.363
Post-entrainment	-0.334	0.640	0.370

Table 4: The creativity scores from spoken dialogue (from D_{spoken}), and the pre- and post-entrainment stages in the main experiment. (higher is better)

find no statistically significant difference between the creativity of the descriptions given in the pre- and post-entrainment stages using NJS (Student's t-test, p=0.207) and NCS (Student's t-test, p=0.76) metrics. Computing a significance value at sample level for EVC is not possible. *In other words, entrainment did not impinge on users' creativity.*

7. Discussion & Future Work

Entrainment (RQ1, H1) We hypothesized that players entrain to general abstract concepts, which means that in the post-entrainment phase they use more geometric shape descriptions than before (H1). Indeed, we found the use of geometric shape descriptions to increase significantly between the pre- and post-entrainment phases. *This provides first evidence that artificial agents can not only entrain human users on the use of a particular wording, but even prime them on a descriptive strategy resembling an abstract concept rather than a simple replacement of words.* An important area for future work resulting from our findings are the ethical implications resulting from a dialogue system that can not only influence the user's choice of vocabulary, but also their entire dialogue strategy.

Agent Performance (RQ2, H2) Our overall aim of priming users to use a specific descriptive strategy was to help our NLU unit by narrowing down the set of descriptive strategies and related choice of vocabulary given by people. We hypothesized that our NLU performance would increase if people were using a well-defined vocabulary instead of freely associating the country's shape (H2). The analysis we performed could, however, not support this hypothesis. One potential reason could be that the training data for the knowledge graph was only composed of the written descriptions provided on AMT, not on spoken dialogue from real game interactions. Our creativity measures revealed that the human-wizard spoken descriptions scored much higher in creativity compared to the written descriptions obtained on AMT. This effect is likely worsened by the size of the shape corpus used for training the knowledge graph, which is still comparatively small. Collecting a larger dataset, particularly tailored towards geometric shapes in combination with using entrainment as presented in this paper would likely increase the NLU performance.

Moreover, extending the corpus by gathering more spoken interactions would potentially give us more creative descriptions, which could further help the NLU. In the future, we plan to enlarge the training set for geometric shapes for the knowledge graph and then conduct a study in which we particularly compare the performance of an autonomous Matcher priming users for geometric shape descriptions with the performance of a Matcher only asking for generic descriptions. *We anticipate that the effect of the entrainment in this setup will indeed lead to a significant increase in NLU performance.*

Interestingly, our data also revealed that in some cases the shape descriptions can even harm the language understanding of the dialogue system. This is, however, not surprising given that shape descriptions are difficult to resolve even for humans (e.g., using "looks like a seahorse on its back" alone cannot be used to identify the target but needs to occur with other descriptions narrowing down the region of the country first). *Shape descriptions hence serve as a 'nudge' rather than being the main driver for target country selection.* In the future, we aim to incorporate this knowledge into our dialogue system by choosing the order of questions strategically instead of randomly and to explore recent advancements in reinforcement learning (Manuvinakurike et al., 2017) to learn better response selection strategies. In addition, we plan to utilize recent advancements in the field of transformer-based embeddings in knowledge graphs as in (Sun et al., 2021) to explore the promise of such approaches.

Creativity (RQ3, H3) Our last hypothesis concerns the creativity of descriptions. Preserving the creativity in the descriptions is not only important for being able to distinguish the descriptions of different countries, we also consider it as an implicit measure of engagement in the task. If participants become bored, they would likely become less creative in their descriptions which would likely lower their performance in the game. We hypothesized that even though the vocabulary could be limited after the strategy-based entrainment, the creativity in the descriptions is not impaired (H3). Indeed, we did not find a significant difference in the creativity of the descriptions. *This indicates that priming users for a specific strategy in a goal-based interaction as the RDG-Map domain does not impair the creativity of their descriptions, suggesting that the quality of descriptions given and the level of engagement in the task is not affected by the entrainment.*

8. Conclusion

In this paper, we studied and discussed the effect of strategy-level entrainment on the performance of a dialogue system by integrating entrainment into an autonomous knowledge-graph based dialogue system for a visual reference resolution task. We found that the system was indeed able to influence the strategy people use to describe the objects in the visual scene and that this did not decrease the overall level of creativity in their descriptions.

9. Bibliographical References

Bergqvist, A., Manuvinakurike, R., Karkada, D., and Paetzel, M. (2020). Nontrivial lexical convergence in a geography-themed game. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 209–214.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555–564.

Brandstetter, J., Beckner, C., Sandoval, E. B., and Bartneck, C. (2017). Persistent lexical entrainment in HRI. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 63–72, Vienna, Austria.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Chen, X. L., Levitan, S. I., Levine, M., Mandic, M., and Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, 8:199–214.

Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota, USA.

Fandrianto, A. and Eskenazi, M. (2012). Prosodic entrainment in an information-driven dialog system. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 342–345, Portland, Oregon, USA.

Friedberg, H., Litman, D., and Paletz, S. B. F. (2012). Lexical entrainment and success in student engineering groups. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 404–409, Miami, Florida, USA.

Hagberg, A., Swart, P., and Chult, D. S. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab (LANL), Los Alamos, New Mexico, USA.

Hakkani-Tür, D., Celikyilmaz, A., Heck, L., Tur, G., and Zweig, G. (2014). Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2113–2117, Singapore.

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

Huang, X., Zhang, J., Li, D., and Li, P. (2019). Knowledge graph embedding based question answering. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 105–113, Melbourne, Australia.

Iio, T., Shiomi, M., Shinozawa, K., Miyashita, T., Akimoto, T., and Hagita, N. (2009). Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary? In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3727–3734, St. Louis, Missouri, USA.

Jung, J., Son, B., and Lyu, S. (2020). AttnIO: Knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3484–3497.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D., Bernstein, M. S., and Fei-Fei, L. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Levitan, R. and Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3081–3084, Florence, Italy.

Levitan, R. (2013). Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *Proceedings of the NAACL-HLT Student Research Workshop*, pages 84–90, Atlanta, Georgia, USA.

Lin, X. V., Socher, R., and Xiong, C. (2018). Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3253, Brussels, Belgium.

Lopes, J., Eskenazi, M., and Trancoso, I. (2013). Automated two-way entrainment to improve spoken dialog system performance. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8372–8376, Vancouver, Canada.

Manuvinakurike, R., DeVault, D., and Georgila, K. (2017). Using reinforcement learning to model incrementality in a fast-paced dialogue game. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 331–341, Saarbrücken, Germany.

Mitchell, C. M., Boyer, K. E., and Lester, J. C. (2012).

From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. In *Proceedings of the Annual Meeting pf the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 94–98, Seoul, South Korea.

Mizukami, M., Yoshino, K., Neubig, G., Traum, D., and Nakamura, S. (2016). Analyzing the effect of entrainment on dialogue acts. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 310–318, Los Angeles, California, USA.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, pages 169–172, Columbus, Ohio, USA.

Paetzel, M., Karkada, D., and Manuvinakurike, R. (2020). RDG-Map: A multimodal corpus of pedagogical human-agent spoken interactions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 600–609, Marseille, France.

Peters, M., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Porzel, R., Scheffler, A., and Malaka, R. (2006). How entrainment increases dialogical effectiveness. In *Proceedings of the IUI Workshop on Effective Multimodal Dialogue Interfaces*, Sydney, Australia.

Rahimi, Z. and Litman, D. (2020). Entrainment2Vec: Embedding entrainment for multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8681–8688, New York, New York, USA.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., and Yu, M. (2017). Entrainment in multi-party spoken dialogues at multiple linguistic levels. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1696–1700, Stockholm, Sweden.

Reitter, D. and Moore, J. D. (2007). Predicting success in dialogue. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 808–815, Prague, Czech Republic.

Stoyanchev, S. and Stent, A. (2009). Lexical and syntactic priming and their impact in deployed spoken dialog systems. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT): Short Papers*, pages 189–192, Boulder, Colorado, USA.

Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., and Cohen, W. (2018). Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4231–4242, Brussels, Belgium.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Tsai, F. S., Tang, W., and Chan, K. L. (2010). Evaluation of novelty metrics for sentence-level novelty mining. *Information Sciences*, 180(12):2359–2374.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep.

Wikipedia. (2021a). List of two-dimensional geometric shapes — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=List%20of%20two-dimensional%20geometric%20shapes&oldid=1015871338`. [Online; accessed 14-August-2021].

Wikipedia. (2021b). Mathematical object — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Mathematical%20object&oldid=1034158734`. [Online; accessed 02-August-2021].

Xu, J., Wang, H., Niu, Z., Wu, H., and Che, W. (2020). Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9338–9345, New York, New York, USA.

Yang, K., Kong, X., Wang, Y., Zhang, J., and De Melo, G. (2020a). Reinforcement learning over knowledge graphs for explainable dialogue intent mining. *IEEE Access*, 8:85348–85358.

Yang, S., Zhang, R., and Erfani, S. (2020b). Graph-Dialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888.

Zhou, H., Zheng, C., Huang, K., Huang, M., and Zhu, X. (2020). KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7098–7108.