A Knowledge Base Schema

The attribute set $\mathcal{A}$ for the MutualFriends task contains name, school, major, company, hobby, time-of-day preference, and location preference. Each attribute a has a set of possible values (entities) $\mathcal{E}_a$. For name, school, major, company, and hobby, we collected a large set of values from various online sources.[14] We used three possible values (morning, afternoon, and evening) for the time-of-day preference, and two possible values (indoors and outdoors) for the location preference.

B Scenario Generation

We generate scenarios randomly to vary task complexity and elicit linguistic and strategic variants. A scenario S is characterized by the number of items (N_S), the attribute set ($\mathcal{A}_S$) whose size is M_S, and the values for each attribute $a \in \mathcal{A}_S$ in the two KBs.

A scenario is generated as follows.

1. Sample N_S and M_S uniformly from $\{5, \ldots, 12\}$ and $\{3, 4\}$ respectively.
2. Generate $\mathcal{A}_S$ by sampling M_S attributes without replacement from $\mathcal{A}$.
3. For each attribute $a \in \mathcal{A}_S$, sample the concentration parameter α_a uniformly from the set $\{0.3, 1, 3\}$.
4. Generate two KBs by sampling N_S values for each attribute a from a Dirichlet-multinomial distribution over the value set $\mathcal{E}_a$ with the concentration parameter α_a.

We repeat the last step until the two KBs have one unique common item.

C Chat Interface

In order to collect real-time dialogue between humans, we set up a web server and redirect AMT workers to our website. Visitors are randomly paired up as they arrive. For each pair, we choose a random scenario, and randomly assign a KB to each dialogue participant. We instruct people to play intelligently, to refrain from brute-force tactics (e.g., mentioning every attribute value), and to use grammatical sentences. To discourage random guessing, we prevent users from selecting a friend (item) more than once every 10 seconds. Each worker was paid $0.35 for a successful dialogue within a 5-minute time limit. We log each utterance in the dialogue along with timing information.

D Entity Linking and Realization

We use a rule-based lexicon to link text spans to entities. For every entity in the schema, we compute different variations of its canonical name, including acronyms, strings with a certain edit distance, prefixes, and morphological variants. Given a text span, a set of candidate entities is returned by string matching. A heuristic ranker then scores each candidate (e.g., considering whether the span is a substring of a candidate, the edit distance between the span and a candidate etc.). The highest-scoring candidate is returned.

A linked entity is considered as a single token and its surface form is ignored in all models. At generation time, we realize an entity by sampling from the empirical distribution of its surface forms in the training set.

E Utterance Categorization

We categorize utterances into inform, ask, answer, greeting, apology heuristically by pattern matching.

- An ask utterance asks for information regarding the partner's KB. We detect these utterances by checking for the presence of a '?' and/or a question word like "do", "does", "what", etc.
- An inform utterance provides information about the agent's KB. We define it as an utterances that mentions entities in the KB and is not an ask utterance.
- An answer utterance simply provides a positive/negative response to a question, containing words like "yes", "no", "nope", etc.
- A greeting utterance contains words like "hi" or "hello"; it often occurs at the beginning of a dialogue.

[14] Names: https://www.ssa.gov/oact/babynames/decades/century.html
Schools: http://doors.stanford.edu/~sr/universities.html
Majors: http://www.a2zcolleges.com/majors
Companies: https://en.wikipedia.org/wiki/List_of_companies_of_the_United_States
Hobbies: https://en.wikipedia.org/wiki/List_of_hobbies

- An apology utterance contains the word "sorry", which is typically associated with corrections and wrong selections.

See Table 2 and Table 1 for examples of these utterance types.

F Strategy

During scenario generation, we varied the number of attributes, the number of items in each KB, and the distribution of values for each attribute. We find that as the number of items and/or attributes grows, the dialogue length and the completion time also increase, indicating that the task becomes harder. We also anticipated that varying the value of α would impact the overall strategy (for example, the order in which attributes are mentioned) since α controls the skewness of the distribution of values for an attribute.

On examining the data, we find that humans tend to first mention attributes with a more skewed (i.e., less uniform) distribution of values. Specifically, we rank the α values of all attributes in a scenario (see step 3 in Section B), and bin them into 3 distribution groups—least_uniform, medium, and most_uniform, according to the ranking where higher α values corresponds to more uniform distributions.[15] In Figure 4, we plot the histogram of the distribution group of the first-mentioned attribute in a dialogues, which shows that skewed attributes are mentioned much more frequently.

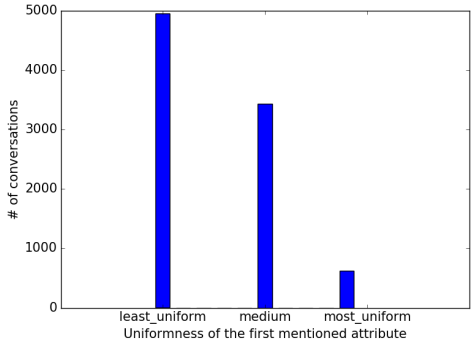

Figure 4: Histogram of the first attribute mentioned in a dialogue. People tend to first mention attributes from very skewed (non-uniform) distributions.

[15] For scenarios with 3 attributes, each group contains one attributes. For scenarios with 4 attributes, we put the two attributes with rankings in the middle to medium.

G Rule-based System

The rule-based bot takes the following actions: greeting, informing or asking about a set of entities, answering a question, and selecting an item. The set of entities to inform/ask is sampled randomly given the entity weights. Initially, each entity is weighted by its count in the KB. We then increment or decrement weights of entities mentioned by the partner and its related entities (in the same row or column), depending on whether the mention is positive or negative. A negative mention contains words like "no", "none", "n't" etc. Similarly, each item has an initial weight of 1, which is updated depending on the partner's mention of its attributes.

If there exists an item with weight larger than 1, the bot selects the highest-weighted item with probability 0.3. If a question is received, the bot informs facts of the entities being asked, e.g., "anyone went to columbia?", "I have 2 friends who went to columbia". Otherwise, the bot samples an entity set and randomly chooses between informing and asking about the entities.

All utterances are generated by sentence templates, and parsing of the partner's utterance is done by entity linking and pattern matching (Section E).

H Turn-taking Rules

Turn-taking is universal in human conversations and the bot needs to decide when to 'talk' (send an utterance). To prevent the bot from generating utterances continuously and forming a monologue, we allow it to send at most one utterance if the utterance contains any entity, and two utterances otherwise. When sending more than one utterance in a turn, the bot must wait for 1 to 2 seconds in between. In addition, after an utterance is generated by the model (almost instantly), the bot must hold on for some time to simulate message typing before sending. We used a typing speed of 7 chars / sec and added an additional random delay between 0 to 1.5s after 'typing'. The rules are applied to all models.

I Additional Human-Bot Dialogue

We show another set of human-bot/human chats in Table 8. In this scenario, the distribution of values are more uniform compared to Table 6. Nevertheless, we see that StanoNet and DynoNet

still learned to start from relatively high-frequency entities. They also appear more cooperative and mentions relevant entities in the dialogue context compared to Rule.

J Histograms of Ratings from Human Evaluations

The histograms of ratings from partner and third-party evaluations is shown in Figure 5 and Figure 6 respectively. As these figures show, there are some obvious discrepancies between the ratings made by agents who chatted with the bot and those made by an 'objective' third party. These ratings provide some interesting insights into how dialogue participants in this task setting perceive their partners, and what constitutes a 'human-like' or a 'fluent' partner.

K Example Comments from Partner and Third-party Evaluations

In Table 9, we show several pairs of ratings and comments on human-likeness for the same dialogue from both the partner evaluation and the third-party evaluation. As a conversation participant, the dialogue partner often judges from the cooperation and strategy perspective, whereas the third-party evaluator relies more on linguistic features (e.g., length, spelling, formality).

Friends of A

ID	Major	Company	Hobby
1	Metallurgical Engineering	Gannett Company	Candle making
2	Business Education	Electronic Arts	Gunsmithing
3	Parks Administration	Kenworth	Water sports
4	Mathematics Education	Electronic Arts	Astronomy
5	Agricultural Mechanization	AVST	Field hockey
6	Mathematics Education	AVST	Shopping
7	Parks Administration	Adobe Systems	Foreign language learning
8	Agricultural Mechanization	Bronco Wine Company	Shopping
9	Metallurgical Engineering	Electronic Arts	Foreign language learning
10	Mathematics Education	Electronic Arts	Poi

Friends of B

ID	Major	Company	Hobby
1	Foreign Language Teacher Education	Gannett Company	Road biking
2	Mathematics Education	Electronic Arts	Astronomy
3	Petroleum Engineering	Western Sugar Cooperative	Candle making
4	Mathematics Education	American Broadcasting Company	Road biking
5	Petroleum Engineering	Western Sugar Cooperative	Road biking
6	Petroleum Engineering	A& W Restaurants	Golfing
7	Petroleum Engineering	American Broadcasting Company	Origami
8	Russian	The Walt Disney Company	Astronomy
9	Petroleum Engineering	The Walt Disney Company	Origami
10	Protestant Affiliation	Acme Brick	Astronomy

A: Human B: Human	A: Human B: DynoNet
A: hi B: hi A: Any friends work at AVST? B: petroleum engi?? || no A: no petroleum B: russian as major? A: Nope B: protestant afiil? || math? A: I have two math B: SELECT 2 A: one works at Electronic Arts || SELECT 10 B: SELECT 1 A: SELECT 4 B: SELECT 2	B: **hi** A: hey B: **i have three friends who like road biking** A: i have two friends who like foreign language learning B: **no foreign language learning here** A: I have no friends who like road biking B: **do they major in foreign language?** A: No || no B: **petroleum engineering**? A: No but I have four friends who work for the Electronic Arts company B: **SELECT 2** A: SELECT 4
A: StanoNet B: Human	**A: Human B: Rule**
A: **hi** || **do you have any friends that work at electronic arts?** B: One and they like astronomy A: **SELECT 10** B: SELECT 2 A: **SELECT 4**	B: **hiya** || **i have one foreign language and gannett** A: i have two avst B: **do you have any acme brick and astronomy?** A: many maths people || two are foreign language B: **do you have any petroleum engineering and american broadcasting company?** A: no || electronic arts B: **SELECT 1** A: avst B: **do you have any disney or restaurant?** ...

Table 8: Example human-bot chats. The mutual friend is highlighted in blue in each KB. Bots' utterances are in bold and selected items are represented by item IDs. Only the first half of the human-Rule chat is shown due to space limit. Multiple utterances of one agent is separated by ||.

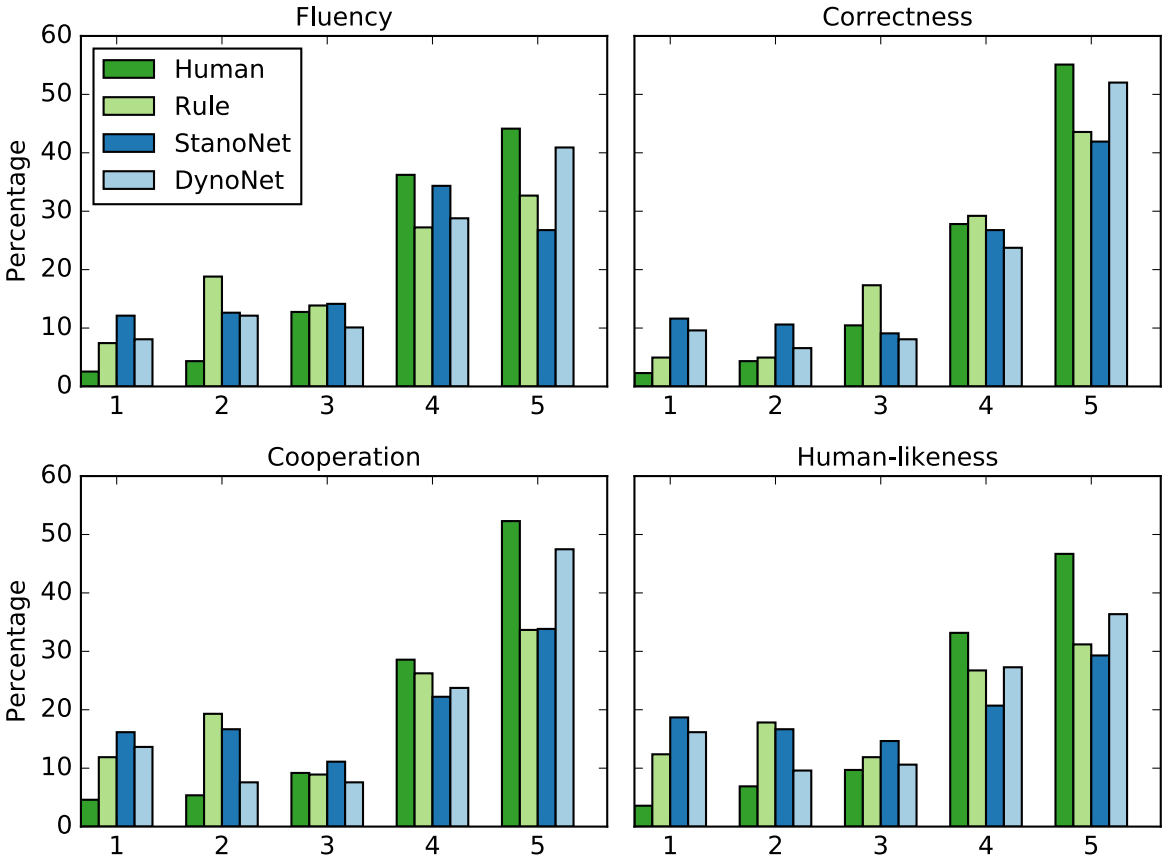

Figure 5: Histogram of ratings (higher is better) from dialogue partners. DynoNet is better than all other systems, especially in cooperation.

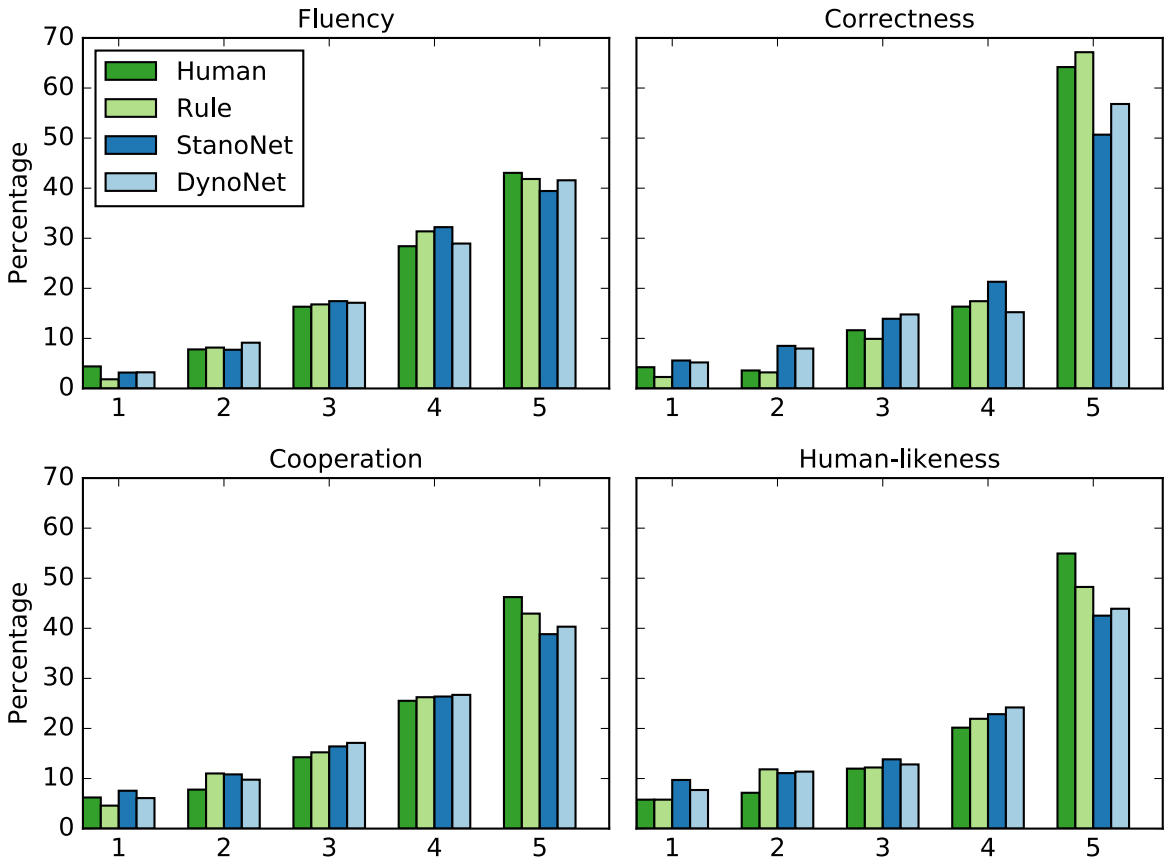

Figure 6: Histogram of ratings (higher is better) from third-party evaluators. Differences between systems are less significant.

System	Partner evaluation (1 per dialogue)		Third-party evaluation (5 per dialogue)	
	Human	Comments	Human	Justifications
Human	4	Good partner. Easy to work with	4.6	- you have any friends who went to monmouth? - The flow was nice and they were able to discern the correct answers. - human like because of interaction talking - Answers are human like, not robotic. Uses ”hiya” to begin conversation, more of a warm tone. - more human than computer Agent 2: hiya Agent 1: Hey
Rule	2	Didn’t listen to me	4	- agent 2 looked human to me - definitely human - A2 could be replaced with a robot without noticeable difference. - They spoke and behaved as I or any human would in this situation. - The agent just seems to be going through the motions, which gives me the idea that the agent doesn’t exbit humanlike characteristics.
StanoNet	5	Took forever and didn’t really respond correctly to questions.	3.5	- No djarum – This doesn’t make sense in this context, so doesn’t seem to be written by a human. - human like because of slight mispellingss - Can tell they are likely human but just not very verbose - Their terse conversion leans to thinking they were either not paying attention or not human. - The short vague sentences are very human like mistakes.
DynoNet	4	I replied twice that I only had indoor friends and was ignored.	3.8	- Agent 1 is very human like based on the way they typed and the fact that they were being deceiving. - Pretty responsive and logical progression, but it’s very stilted sounding - i donot have a jose - Agent gives normal human responses, “no angela i don’t” - agent 1 was looking like a humanlike

Table 9: Comparison of ratings and comments on human-likeness from partners and third-party evaluators. Each row contains results for the same dialogue. For the partner evaluation, we ask the human partner to provide a single, optional comment at the end of the conversation. For the third-party evaluation, we ask five Turkers to rate each dialogue and report the mean score; they must provide justification for ratings in each aspect. From the comments, we see that dialogue partners focus more on cooperation and effectiveness, whereas third-party evaluators focus more on linguistic features such as verbosity and informality.