



Feng, D., Carstensdottir, E., El-Nasr, M. S. and Marsella, S. (2019) Exploring Improvisational Approaches to Social Knowledge Acquisition. In: 18th International Conference on Autonomous Agents and MultiAgent Systems, Montreal, QC, Canada, 13-17 May 2019, pp. 1060-1068. ISBN 9781450363099.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

© International Foundation for Autonomous Agents and Multiagent Systems 2019. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19), Montreal, QC, Canada, 13-17 May 2019, pp. 1060-1068. ISBN 9781450363099.

<http://eprints.gla.ac.uk/190276/>

Deposited on: 12 July 2019

Exploring improvisational approaches to social knowledge acquisition

Dan Feng
Northeastern University
Boston, Cambridge
danfeng@ccs.neu.edu

Magy Seif El-Nasr
Northeastern University
Boston, MA
magy@northeastern.edu

Elin Carstensdottir
Northeastern University
Boston, MA
elin@ccs.neu.edu

Stacy Marsella
University of Glasgow
Glasgow, UK
stacymarsella@gmail.com

ABSTRACT

To build agents that can engage user in more *open-ended* social contexts, more and more attention has been focused on data-driven approaches to reduce the requirement of extensive, hand-authored behavioral content creation. However, one fundamental challenge of data-driven approaches, is acquiring human social interaction data with sufficient variety to capture more open-ended social interactions, as well as their coherency. Previous work has attempted to extract such social knowledge using crowdsourced narratives.

This paper proposes an approach to acquire the knowledge of social interaction by integrating an improvisational theatre training technique into a crowdsourcing task aimed at collecting social narratives. The approach emphasizes theory of mind concepts, through an iterative *prompting* process about the mental states of characters in the narrative and *paired writing*, in order to encourage the authoring of diverse social interactions. To assess the effectiveness of integrating prompting and two-worker improvisation to the knowledge acquisition process, we systematically compare alternative ways to design the crowdsourcing task, including a) single worker vs. two workers authoring interaction between different characters in a given social context, and b) with or without prompts. Findings from 175 participants across two different social contexts show that the prompts and two-workers collaboration could significantly improve the diversity and the objective coherency of the narratives. The results presented in this paper can provide a rich set of diverse and coherent action sequences to inform the design of socially intelligent agents.

KEYWORDS

social knowledge acquisition; crowdsourcing; online collaboration

1 INTRODUCTION

Breakthroughs in Artificial Intelligence (AI) have shown that autonomous agents systems are effective at solving well-specified tasks, such as games like Poker [1] and Go [39]. However, as AI systems become ubiquitous, autonomous agents are being asked to solve more complex tasks that involve interacting with humans and other AI systems in more open-ended social contexts. Therefore, the development of agents capable of engaging in human social interaction has become critical. For example, agents are being used as social actors in applications designed to provide environments in which learners can practice and learn social skills[17].

Ideally, these systems should support a rich interactive experience for the users that gives them the freedom to explore different actions in various situations, which in turn requires the appropriate responses from socially intelligent agents (SIA) used in the systems. However, designing such SIAs with sufficient amount of actions and the capability to generate appropriate responses requires *extensive* behavioral content creation. Traditionally, this content is hand-authored, especially for social training systems [17]. Unfortunately, the amount of content produced using hand-authoring can be insufficient to sustain a flexible interactive experience that supports repeated interactions by a learner.

With advances in machine learning techniques, researchers are attempting to use data-driven approaches to design agent models [7]. One fundamental challenge of this data-driven approach is acquiring human social interaction data with sufficient variety to cover as much as possible of the interaction space, while maintaining the coherency of the interaction. One approach proposed to capture coherent human activities is by extracting knowledge, including human activity and dialogue, from crowdsourced narratives and stories that describes the interaction between different characters. Because narratives and stories represent the way people make sense of the world [9, 27]. Once such data is collected, machine learning techniques can be used to build an SIA model. For example, PIP [22] extracts the verbal and non-verbal behaviors from crowdsourced narratives using a semi-situated learning method. Feng et al. [5] presented an active learning approach to build a generative model trained on crowdsourced narrative data, which could then be used as an agent model to generate social actions.

The focus of this paper is on approaches to the crowdsourced data collection phase. Specifically, in this paper we explore an approach

to crowdsourcing narratives inspired in part by an improvisational theatre technique, Active Analysis (AA) [2], and systematically evaluate it against alternatives. AA is used by theatre actors to rehearse plays, through a process of having the actors improvise around a scene and exploring alternative mental states of the characters they are playing. According to [2], the actors could gradually obtain the inner incentives of the character’s actions during the process. As a result, the actor could better understand the character’s action and feel the character’s emotion, which is essential to the creative work on the stage. We believe AA suggests ways to generate diverse social interaction due to two significant attributes: 1) multiple actors dynamically improvise a scene with each of them playing a different character and 2) the actors iterate on the same scene with different prompts, given by a director, about the mental attitudes of the characters they play.

We can draw on these ideas to create a crowdsourcing task that a) uses paired writers improvising against each other and b) prompts them with different mental attitudes of the characters they are portraying. Since AA has been proved to be an effective way to help actors with their creativity, we thus hypothesize this AA-based approach will improve the diversity and coherency of narratives collected. On the other hand, these hypotheses may be incorrect, because multiple workers improvising their respective roles together can potentially negatively impact coherency, and because there is a possibility that the use of prompts might constrain the overall diversity of interaction across crowd workers.

Thus, to assess how well these techniques work with the knowledge acquisition process, we compare in this work alternative ways to design crowdsourcing tasks, specifically exploring these questions about diversity and coherency. The study compares social narratives containing an interaction between multiple characters for a specific social context, produced by a single writer vs. multiple writers. In addition, these conditions are explored with and without prompts.

Findings from 363 stories created by 175 participants across two different social contexts shows that the prompts and two workers improvising their own role can significantly improve the diversity and the objective coherency of the narratives collected from crowd workers. No significant effects are found on the subjective assessment of coherency. The results in this paper show that this approach with paired-worker improvisation and prompts of different mental attitudes is a viable methodology for collecting narratives suitable for social knowledge acquisition, and can therefore potentially inform the collection of data to build socially intelligent agents in flexible agent-based training systems such as [4, 5, 22].

2 RELATED WORK

2.1 Crowdsourcing for Knowledge Acquisition

Interest in human interaction and activities goes back to Schank and Abelson’s early work on gathering *scripts* [38] manually from procedural knowledge. More recently, such knowledge has been crowdsourced via Amazon Mechanical Turk, which has been used as an effective method of collecting common sense knowledge [24, 41, 44], creating creative content [43] and acquiring life experience [20, 21, 30, 35, 40, 45].

Weltman et al. [45] collected some highly detailed life experience descriptions and the causal relations between descriptions using sequences of comic frames. Regneri et al. [35] collected script-like event description to infer a temporal script graph (TSG) from 22 crowd-sourced stereotypical scenarios. However, their initial user study indicates the process to be complex and time-consuming. According to [45], people habitually omit obvious actions and states, which makes it difficult to model the sequence of common behaviors.

To simplify the crowdsourcing task, the Restaurant Game [32] asked crowd workers to play the role of a customer in a simulated virtual restaurant, then built a probabilistic model of their activity. The setting of the game is very well-defined and a set of actions (e.g., sit down, order, etc.) were a priori given to the crowd workers. However, it is labor-intensive to build a simulated environment manually for each domain to gather domain knowledge. Sina et al. [40] utilized semi-structured template to collect human activities and create alibis for the purpose of training police officers. Li et al. [21] asked the crowd to provide how real-world situations unfold using simple sentences.

Apart from the previous works, the proposed work focus more on the open-ended interactions including dialogue and physical actions between two or more persons rather than just individual actions or constrained scripts.

2.2 Crowdsourcing and Creativity

Efforts towards collecting sufficiently varied data of good quality include [42], where a ‘Taboo Words’ mechanism was used to prevent crowd workers from using words that had already been commonly agreed upon, in order to boost worker creativity and ensure data quality. KissKissBan [11] borrows these mechanics, and extends them by integrating both collaborative and competitive elements in a human computation game. [31] used music to prime for positive affect, in addition to affective pre-screening, in an attempt to boost creativity for crowdsourcing platforms.

The key difference in this work is the use of an Active Analysis inspired approach to improve the creativity of crowd workers in order to collect more diverse social interactions.

2.3 Collaborative Writing

Collaboratively writing high-quality stories has been explored using crowdsourcing techniques [15, 26]. Similar to traditional collaborative writing [14], the design of crowdsourcing writing tasks normally adopts strategies to 1) divide the complex task into micro-tasks that are mergeable after the parallel data collection [18, 19] or 2) iteratively crowdsourcing the data piece by piece [16].

Different from previous work on collaborative writing, the work reported here emphasizes the improvisation of the action in order to construct action space for agent design, as oppose to editing or commenting on the written stories.

2.4 Social Intelligent Agents

Previous research attempts to craft socially intelligent agent models rely mainly on the agent designer’s intuitions, their experience, or existing theories in economics and behavioral science [17, 29]. For

example, negotiation agents are normally built upon the assumption that the agent should behave rationally. Such agents consider what action to take from a pre-defined set of actions and make decisions solely in terms of its worth to themselves. However, these systems often turned out to be very complex, especially in open-ended negotiations, that requires extensive authoring efforts when modeling all domain knowledge [17].

With advances in machine learning techniques, more and more researchers attempt to use data-driven approaches to design agent models [5, 7, 8, 37]. One fundamental challenge of this data-driven approach is acquiring human social interaction data with sufficient variety, while maintaining the coherency of the interaction.

This work attempts to capture this variety by extracting social knowledge from crowdsourced narratives. Indeed, narratives have been used to train agents to exhibit believable behaviors using various machine learning techniques such as interactive machine learning [9], and active learning [5]. The goal of this work is exploring alternative ways to acquire data with diverse human behaviors that can be used by such learning techniques.

2.5 Active Analysis

The design of our crowdsourcing system was heavily influenced by Stanislavsky’s Active Analysis (AA) rehearsal technique [2] which was developed to help theatre actors rehearse a script and fine tune their performance. The overall script is composed into key events (i.e., short scenes) that actors rehearse and improvise under a director’s guidance.

Theoretically, AA consists of three phases: *Framing*, *Improvisation*, and *Performance Analysis*. These phases are then iterated upon by the rehearsal director who will often change the motivations and tactics the actors are using when playing out the scene.

Framing: The director determines what the overall context of the scene is, and what goals should occur as a product of the interaction in the scene. The director then helps the actors explore different motivations and approaches to their role by selectively providing information about mental attributes, or goals. This can include withholding information from other actors in the scene. The Framing process provides both context and a target to guide the actors as they improvise. This feature of AA is highly relevant to the work presented here, as it provides a methodology of guiding actors, crowd workers in this case, without explicitly constraining them.

Improvisation: Actors explore different tactics to achieve their goals through improvisation.

Performance Analysis: The director and actors evaluate the improvisation work.

AA has two attributes especially relevant to social knowledge acquisition. First, AA requires the *collaboration* between different actors while each of them play a different character in the scene. Second, by providing different prompts at each iteration, AA is designed to foster an actor’s creativity and re-conceptualization of the beliefs, motivations and behavior of their own as well as other actors. Thus, AA’s engenders Theory of Mind (ToM) reasoning which can be critical in social interactions and creative work. ToM refers to the human ability to have, and use beliefs about the mental processes and states of others [46] which is implicated in a range of

social interaction constructs, e.g., cognitive empathy and emotional intelligence.

We argue that these AA techniques can be used to improve the crowdsourcing of social narratives. Specifically, supporting ToM constructs and inspiring crowd workers through prompts may, similar to AA, inspire creative thinking about the social interaction and thus allow us to collect content of rich, coherent interactive experiences more quickly given the same number crowd worker. Therefore, in this work, we focus on examining the impact of Framing and Improvisation. ‘Performance Analysis’ was not integrated in the current design.

3 METHOD

3.1 Crowdsourcing Task design

We transformed the AA process of actors repeatedly improving a scene with different prompts into a narrative writing task where crowd workers were asked to improvise the action of a character within a given scenario. To replicate the AA collaborative role-play between different actors, crowd workers worked in pairs on a scenario. Each was assigned a different character in the scenario and told to write their assigned character’s actions only. Distinct from traditional peer writing where writers critique each others work, the workers were asked to work on the actions of only their character. To simulate the iterative process, each task contained 3 stages. Each stage had the same initial scenario setting but the worker was asked to craft a new narrative. The crowdsourcing system played the role of the director, giving random prompts at different stages to the different workers in the scenarios. There was no explicit way for the workers to know the prompts given to other workers.

To examine the feasibility of this crowdsourcing technique, we designed a $2(\text{has prompts} / \text{no prompts}) \times 2(\text{single worker} / \text{two workers})$ between-subjects design experiment.

In each task, the crowd worker¹ was instructed to write the interaction between different characters. As previously noted, each task contains 3 stages where every stage has the description of the scenario, as shown in Appendix A. In the *has prompt* condition, each worker was given random prompts, such as ‘*Now Imagine : [CHARACTER NAME] is an ethical person*’, at the second and third stage, while in the no-prompt condition, no prompts were given to workers. They were asked to write a new story from scratch at each stage, and instructed that each story should have at least 6 turns of interactions.

In the two-worker condition, each worker was randomly assigned to play the role of one character in the scenario, e.g. Gerald or Anne. Workers were instructed to take turns to write the actions of the character they played. Note that here the worker could only write their own character’s action. No additional collaboration such as commenting on the history of actions or discussing the interface were allowed.

In this study, there are 2 scenarios in total. One is a realistic scenario adapted from a scene from the TV series House of Cards, and the second scenario is fictional and set in a stereotypical fantasy environment. Each worker only saw one scenario across the three

¹We used Amazon’s Mechanical Turk for all crowd-sourcing.

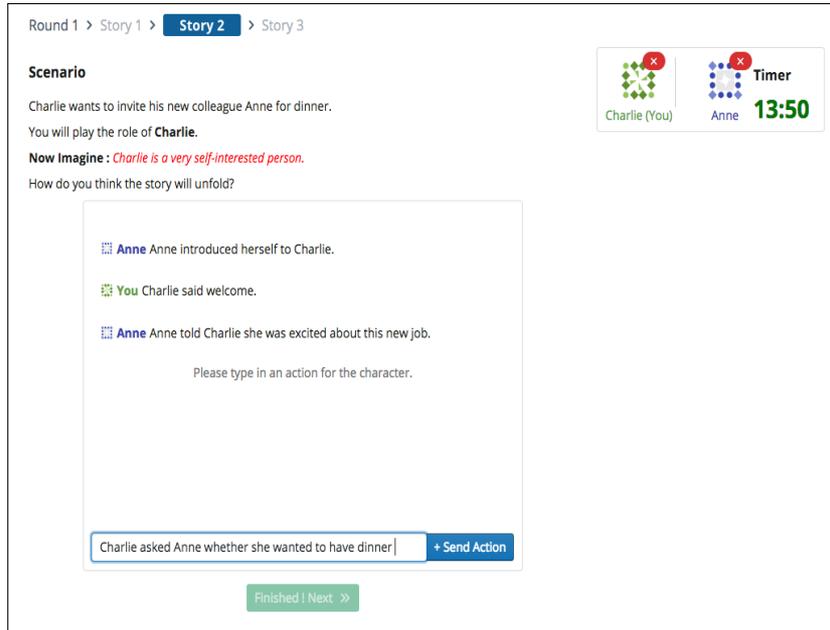


Figure 1: Example of Data Collection Interface. The same scenario was given to the worker at all stages. The prompt was highlighted in red color.

stages they participated in. The description of the scenarios and a full list of prompts can be found in the Appendix A.

Each worker was randomly assigned to one condition. If they were assigned to the two-worker condition and waited for more than 5 minutes without pairing with another worker, the single writer version would be presented to the worker.

The data collection was built based on Empirica, which is an open-source Real-Time, Synchronous, Virtual Lab Framework [33]. The interface, as shown in Fig. 1, contains the scenario setting, the role played by the worker, and the prompt. As introduced in Li et al. [21], the worker was primed to write using past tense verbs. To avoid focusing only on dialogue and conversation, and ignoring all the other types of actions such as physical movement, the workers were also instructed to write in third-person.

All tasks were posted on Amazon Mechanical Turk to United States workers with a minimum task approval rating of 98 and minimum approved tasks of 100. The reward of the workers includes base-payment (€50), waiting (€10 per minute, maximum waiting time is 5 minutes), writing (\$1.0 per story), and creativity (\$1) which was always given as long as the worker finished the task.

3.2 Hypotheses

- (1) The intervention of a prompt will facilitate creativity. By prompting the crowd, workers provide more diverse actions without decreasing the coherency.
- (2) By collaborating with another writer, workers will produce more diverse and more coherent results.

3.3 Evaluation Metrics

To evaluate the collected data, we adapted some of the features proposed by [36] that were used in story generation, and in evaluating writing quality. The features listed were used to evaluate both the coherency and the diversity of the collected data.

3.4 Structural Evaluation Metric

3.4.1 Number of interactions. The number of interactions was evaluated by the number of turn-takings in the stories. In this work, the workers were asked to write at least 6 turns of interactions together. However, there was no constraint placed on the maximum length of the story. Therefore, the length of the interaction can be used as an indication of engagement, and how much effort and time the crowd workers were willing to put in the task.

3.4.2 Number of sentences per turn. The number of sentence was computed by averaging number of sentences per turn;

3.4.3 Word Frequency. : Word frequency has been found to correlate with writing quality [3]. Here, word frequency refers to the average frequency of all the words, excluding stop words² in a story using the 100 million frequency obtained from Exquisite Corpus³.

3.5 Coherency

3.5.1 Subjective Coherency. : To assess the *coherency* and *consistency* of the collected stories, another group of crowd workers was tasked with evaluating them, using five questionnaire items

²stop words refer to the most common words without specific semantic meaning such as “the”, “at”, etc. In this work, the names of the characters were also excluded.

³<https://github.com/LuminosoInsight/wordfreq>

adapted from [6] including: “I understand the story”, “The story makes sense”, “Every sentence in the story fits into the overall plot”, “The characters’ behaviors are consistent with their goals and beliefs” “The characters’ interaction is believable”.

Each item was evaluated using a 5-point Likert scale ranging from “Strongly disagree” to “Strongly agree”. For ease of analysis, the items were then mapped to a numerical *score* from -2 to 2 . Each story was assessed by 3 workers. The subjective coherency score is the average score across all 5 questions.

3.5.2 Semantic Similarity. Semantic continuity is indispensable for coherent text. In this work, the semantic similarity was evaluated using the average cosine similarity of adjacent sentence vectors. The vectors were obtained using GloVe word embedding trained on Wikipedia 2014, and Gigaword 5[34]. The sentence vector was computed using a Bag-of-word approach by averaging a sentence’s word vectors.

3.5.3 Entity Co-reference. Co-reference between common entities is very important for establishing the coherency of a story. For example, in the sentence “*Anne greeted Charlie. Charlie asked her what did she want.*”, “her” and “she” both refer to “Anne”. The entity co-reference score attempts to capture the connection between different sentences. The score was computed by the number of co-reference chains extracted using Stanford CoreNLP tool [25].

3.6 Diversity

3.6.1 Sentiment Diversity. : The sentiment score can be used to predict the consistency of the writing task. The sentiment score was computed using Stanford CoreNLP tool which predicated the sentiment label of a sentence from a set of “very negative”, “negative”, “neutral”, “positive”, and “very positive”. The sentiment score was then converted to an integer number from “-2” to “2”. The sentiment variance was evaluated by the variance of the sentiment score for all the sentences in a single story.

3.6.2 Lexical Diversity. : The syntactic complexity is another important indication of writing quality [28]. We examined this feature in terms of the number of unique lexical phrases, such as *noun phrases (NP)*, *verb phrases (VP)*, as well as *named entities (NE)* in the collected sentences. All the variables were normalized by the number of turns.

3.6.3 Action Diversity. : One of the main goals of this crowdsourcing technique is to construct the action space for the agent to act in open-ended social situations. In this work, the action was represented by the predicates of a sentence since it contains verbs, and main actors who execute the action. Semantic role labelling (SRL) tool [10] was used to extract predicates from sentences. The number of predicates was normalized by the number of turns. For example, in one of the collected stories, we obtained the following sentence “*Charlie reached for his phone, called 911, and explained that a woman had broken into his home.*”. After the SRL labeling, 4 events/actions represented by predicates were obtained as following:

- **Predicate:** reached(1), **A0:** Charlie, **V:** reached, **A1:** for his phone, called 911,
- **Predicate:** called(5), **A0:** Charlie, **A1:** his phone, **V:** called, **A2:** 911

Table 1: Number of stories in each condition

	No Prompt	Has Prompt
Single worker	111	96
Two workers	75	81

- **Predicate:** explained(8), **A0:** Charlie, **V:** explained, **A1:** that a woman had broken into his home
- **Predicate:** broken(13), **A1:** a woman, **V:** broken, **AM-DIR:** into his home

4 RESULTS AND DISCUSSION

240 participants ($Mean_{Age} = 36.9, SD = 13.3$) were recruited from Amazon Mechanical Turk to write the stories.

4.1 Pre-processing

Participants who dropped out in the middle of the task, or did not finish all 3 stages with at least 6 interaction turns, were excluded. 65 participants were excluded based on this criteria, with 175 participants remaining in the set. This left 363 stories collected from 121 trails for later analysis as shown in Table. 1. Examples of stories in the corpus can be found in Appendix B.

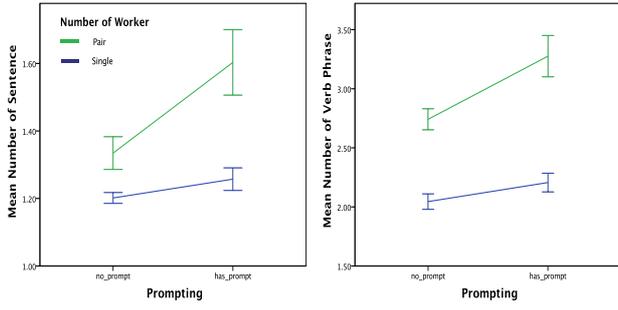
In addition, another group of 90 participants was recruited to evaluate the coherency of the stories using the five questions discussed in section 3.5.1. Each participants was asked to evaluate 5 stories. Of those, 4 were randomly selected from the collected stories, and 1 was randomly selected from a set of expert-assessed set of stories, which served as a golden standard, as an assessment test. If the crowd worker failed the test, their responses were excluded from later analysis. The stories included in the coherency analysis were evaluated by at least 2 raters, this excluded 26 stories from the set.

4.2 Multivariate Analyses

To determine if there is an interaction between the two independent variables (IVs) on the dependent variables (DVs) introduced in section 3.5, we first conducted a two-way MANOVAs (Multivariate Analysis of Variance) of Prompting (*has-prompt, no-prompt*) and Number-of-workers (*single, pair/two-worker*) on all DVs in question (Structural metrics, Coherency, and Diversity).

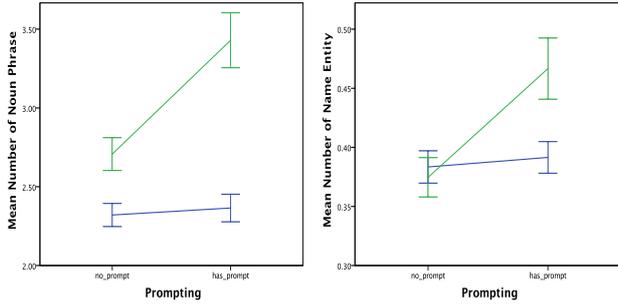
4.2.1 Structural Metrics. There was a significant interaction between Prompting and Number-of-workers ($F(3, 357) = 2.78, p = .041$; *Pillai’s Trace* = .02) on the average number of sentence per turn ($F(1, 359) = 4.14, p = .04$). As shown in fig. 2a, a simple main effect analysis shows in the two-worker condition, using prompts could significantly improve the number of sentences ($F(1, 359) = 9.63, p = .002$). Main effects of the number of workers were found on the number of interactions ($F(1, 332) = 4.49, p = .035$), and average word frequency ($F(1, 332) = 20.01, p < .001$).

4.2.2 Coherency. No interaction effects were found on any coherency metrics. A main effect of Prompting was found on the number of co-referenced entities $F(1, 346) = 5.02, p = .03$. A main effect of the number of the workers was also found on semantic



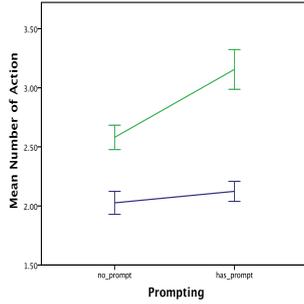
(a) Number of Sentence

(b) Unique Verb Phrase



(c) Unique Noun Phrase

(d) Unique Named Entity



(e) Action

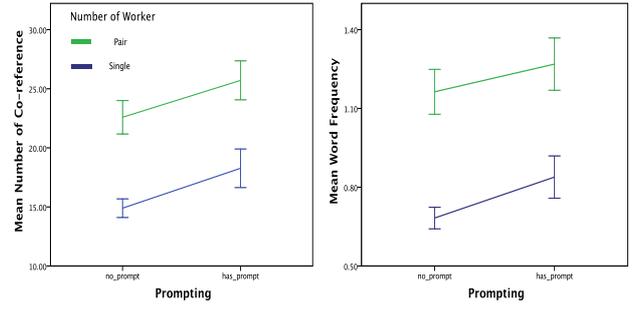
Figure 2: Interaction effect of two IVs. The error bar represents ± 1 SD

	<i>F</i> Score	<i>p</i> - value
Structural Metrics		
Number of Sentence	4.14	.04
Diversity		
Unique Noun Phrase	10.13	.002
Unique Named Entity	5.4	.02
Unique Verb Phrase	4.68	.031
Action	3.99	.047

Table 2: Interaction Effects of two independent variables

similarity ($F(1, 346) = 60.05, p < .001$) and number of co-referenced entities ($F(1, 346) = 29.72, p < .001$).

4.2.3 **Diversity.** There was a significant interaction effect between Prompting and Number-of-workers ($F(5, 301) = 2.42, p = .036$; *Pillai's Trace* = .036) on the number of noun phrases, named entities and verb phrases. A simple main effect test shows using



(a) Entity Co-reference

(b) Word Frequency

Figure 3: Main effects of the two IVs. The error bar represents ± 1 SD

Metrics	M_{single}	SD_{single}	M_{pair}	SD_{pair}	<i>T</i>
Structural Metrics					
Number of Interactions*	8.8	5.2	9.9	3.9	-2.23
Number of Sentence**	1.22	.26	1.47	.70	-4.18
Word Frequency**	0.76	.63	1.22	.82	-5.86
Coherency					
Semantic Similarity**	.62	.10	.69	.08	-7.3
entities Co-reference**	16.46	12.6	24.2	13.89	-5.58
Diversity					
Sentiment Diversity*	.51	.30	.58	.31	-2.07
Unique VP**	2.12	.73	3.02	1.28	-7.85
Unique NP**	2.34	.82	3.09	1.35	-6.11
Action**	2.07	.90	2.88	1.25	-6.64

Table 3: Metrics that have significant difference between *single* worker v.s. *pair* workers. * denotes $p < 0.05$, ** denotes $p < 0.01$

prompts in the two-worker condition could significantly improve the number of noun phrases as shown in fig. 2c, verb phrase as shown in fig. 2d and the number of actions as shown in fig. 2e.

Further, a main effect of the prompts was observed on the number of noun phrases ($F(1, 305) = 10.17, p < .002$), verb phrase ($F(1, 305) = 9.04, p < .003$), number of named entities ($F(1, 305) = 7.67, p < .006$) and the number of actions ($F(1, 305) = 10.58, p = .001$). Main effects of the number of workers were also observed on the number of noun phrases ($F(1, 305) = 31.43, p < .001$), verb phrase ($F(1, 305) = 51.16, p < .001$) and the number of actions ($F(1, 305) = 57.12, p < .001$).

4.3 Result of Number of the workers

An independent sample t-test shows that there are significant differences between single worker condition and two-worker conditions. Using two workers has a positive effect, increasing diversity and coherency of the collected data as shown in Table. 3. No significant differences were found for subjective coherency.

4.4 Result of Prompting

As shown in Table. 4, an independent sample t-test shows that there are significant differences in no-prompt and has-prompt condition on the lexical diversity and objective coherency. No significant difference was found on subjective coherency.

Metrics	M_{no}	SD_{no}	M_{has}	SD_{has}	T
Structural Metrics					
Number of Sentence**	1.26	.30	1.42	.66	-2.96
Word Frequency*	0.88	.62	1.04	.86	-2.03
Coherency					
Entities Co-reference*	17.99	10.76	21.67	15.9	-2.57
Diversity					
Unique NP**	2.48	.85	2.86	1.36	-3.25
Unique VP**	2.33	.80	2.70	1.32	-3.21
Unique NE**	.38	.14	.42	.19	-2.63
Action*	2.24	.97	2.61	1.27	-2.93

Table 4: Metrics that have significant different between *has_prompt* worker v.s. *no_prompt* conditions. * denotes $p < 0.05$, ** denotes $p < 0.01$

5 CONCLUSIONS

This paper presents and evaluates a novel crowdsourcing technique, based on theatre training, to acquire social knowledge, i.e., stories describing interaction sequences between different characters in a specific scenario. The key components of this approach include the dynamic improvisation of separate workers for each character, and random prompts at different stages.

Comparing stories created by a single worker vs. two workers, we found that in general, stories created in the two-worker condition were significantly better than those produced in the single worker condition, both in terms of the coherency and diversity of the collected stories. This could be an indication that it is more challenging for a single worker to create good quality stories, where the worker has to think about actions taken by different characters. One worker mentioned this in the end survey saying

"I wish I was able to create a story with another person. I had to write the story by myself (no one else was in the lobby after waiting for a long time) so I had to pretend to be all characters and think of how they would react."

This description is telling, as the single worker condition requires the writer to think from different perspectives, back and forth during the task. This is part of the Theory of Mind (ToM) element, and is an important component of the process. Studies have shown that although people typically develop ToM at an early age, even adults with a fully formed capacity for ToM often fail to employ it [13, 23]. Therefore, reasoning about and writing from different perspectives requires significant cognitive effort, as it can result in recursive reasoning across chains of actions for multiple characters. Asking a single worker to reason in such a way will therefore most likely require higher cognitive effort, compared to two workers where each worker would only have to reason about one character. The higher cognitive effort on a single worker for this task may explain the decrease in diversity and coherency for that condition.

Another factor that may be impacting objective coherency is that a single writer may assume there exists common knowledge, common ground, between characters which may not be explicit in the character’s behavior and therefore not explicit for a reader. On the contrary, two workers need to convey this knowledge explicitly to each other through their respective characters actions and dialog during the improvisational process. In fact, in the two-worker condition, more high frequency words were used, suggesting that

two workers use more common words to reduce the complexity of communication.

When it comes to prompts, we found that giving random prompts to workers significantly improved diversity without decreasing the coherency of the stories. This was especially true for the two-worker conditions, where giving different mental attitudes to different workers could significantly improve the diversity and length of the interaction. This indicates that two-worker writing tasks and prompts, are an efficient way to simulate social interaction through text, and give direction to workers that will increase variety while maintaining coherency of the stories produced.

In addition to collaboration producing better quality stories, our findings seem to indicate that crowd workers enjoyed the joint work on a creative task like the one they were presented with in this study. When asked through an open-ended question about their over-all experience after finishing the writing process, many workers mentioned they enjoyed the HIT⁴, as listed below, which indicates a feasibility of collecting data with such a task design.

"I actually enjoyed doing this, which surprised me."

"Had a great time with this writing partner. Super fun hit."

"I thoroughly enjoyed this task. It was creative and made feel more attentive and interested. Great concept!"

These comments suggest that one factor that may be influencing the diversity in the two worker condition, and perhaps the prompt condition, is that they lead to more positive emotional states in performing the task. Such positive states have been associated with increased creativity [12]. Compare and contrast to the work discussed earlier, [31] where music was similarly employed to alter mood. In that case the music was incidental to the task, while here the design of the task itself may be influencing the affective state.

As discussed in previous work, social intelligent agents rely in many cases on hand crafted social knowledge and creating such a corpus of social knowledge is both time-consuming and requires a lot of labor. A hand crafted knowledge base might exclude information that its authors might not consider relevant due to their own social context and perspective. In addition, it will likely require an extensive amount of work to add to, or re-build, when the agent needs to be applied to new domains. Using crowdsourced stories that contain social knowledge provides an alternative to crafting such a complex corpus. However, crowdsourcing has limitations and the quality of the data collected depends on a variety of factors like the task design, the worker population and their engagement. By integrating the improvisation training techniques to a crowdsourcing task, we demonstrate how crowdsourced stories can be further refined to increase both variety and output of the social knowledge collected. As a result, this study presents a number of implications for how social knowledge, meant for social intelligent agents, can be acquired. Our current work compared different narratives based on their diversity on the level of phrases, actions and sentiments. While these components are important for building the social agents, future work will explore other metrics. For example, the diversity of adjacent action pairs could also be compared to examine the differences at plot and narrative arc levels. In addition,

⁴HIT refers to Human Intelligence Task on Amazon Mechanical Turk.

future work will focus on using these crowdsourced data to extract rich and coherent action sequences that can be used for a generative model, that will then be used to drive the behaviors of a social intelligent agent. Specifically, we plan to cluster the predicates extracted from the stories, based on their syntactic and semantic similarities. The clusters obtained can then serve as sets of action spaces (i.e. actions sets), and the transition probability between clusters could then be used to build a stochastic generative model. The agent built from this data will also be compared with hand-crafted agents.

ACKNOWLEDGMENTS

Funding for this research was provided by the National Science Foundation Cyber-Human Systems under Grant No. 1526275.

A SCENARIO SETTING AND PROMPTS

A.1 Scenario I

*Anne, an aspiring reporter, wants to interview Bob at his house. Bob employs **Charlie** to make sure that nobody is able to bother him. Charlie guards the door. Anne is in front of Bob's house and is attempting to interview him.*

A.2 Scenario II

*A small village has almost been destroyed by a monstrous dragon. All the treasures the villagers possessed were taken. **The knight, K**, in the village seeks to defeat the dragon to rebuild the village. But first, K needs a magic sword which can only be built using the raw steel from **Gerald**, who is a merchant in the village.*

A.3 Random Prompts

The 6 prompts used in this study: "*ambitious*", "*ethical*", "*dedicated*", "*manipulative*", "*self-interested*", "*arrogant*".

B SAMPLE OF COLLECTED STORIES

B.1 Sample Story of Scenario I

B.1.1 Prompts : **None**.

- Charlie looked at the woman and wondered what she wanted.
- Anne introduced herself to Charlie.
- Charlie smiled at Anne and firmly told her no guests were being accepted at the present time.
- Anne explained that she only wanted a few minutes of Bob's time.
- Charlie nicely told Anne she would need to schedule an appointment for another day.
- Anne thanked Charlie and asked him for Bob's contact information.

B.1.2 Prompts : **Anne : ethical, Charlie : ambitious**.

- "Excuse me kind sir." said Anne, "I have been scheduled to interview Bob."
- "I'm sorry, ma'am." said Charlie, "I've been instructed: no visitors while Bob works."
- "I understand, but I assure you I am expected." she replied.
- Charlie said "Well, you're not expected by me. And I have authority here."
- Anne's disappointment showed clearly on her young face.

- Seeing that she was desperate, Charlie thought up a scheme. "Look. I can see this interview means a lot to you. Perhaps if you dropped me a little donation, say a hundred bucks, I could look the other way" he said with a sneer.
- Anne looked startled. "I'm sorry, sir, but I would much rather do things the right way" she said. "I'll make another appointment, in which I'll be sure to notify Bob just what kind of man you are."
- "Good luck, lady. NO ONE sees Bob without going through me!" Charlie growled.
- Just then, Bob emerged from the front door. "Thanks for your help, Anne" said Bob.
- Charlie turned around with a confused look on his immense face.
- "No problem, sir" said Anne smartly. "We have to be sure of who we employ for your safety, and it looks like your hunch was right."
- Charlie stands stunned....a few days later...in the unemployment line.

B.2 Sample Story of Scenario II

B.2.1 Prompts : **None**.

- K ran to Gerald's shop.
- Gerald asked K what he needed.
- K explained that only the raw steel Gerald had could make a magic sword.
- Gerald was flattered.
- K implored Gerald to work quickly, before the dragon came for them, too.
- K and Gerald suddenly smelled something awful.
- The dragon was right outside the window. The dragon roared fire.
- K and Gerald died.

B.2.2 Prompts : **K : ethical, Gerald : dedicated**.

- K approaches Gerald to purchase a sword in order to help the villagers.
- Gerald asks what specifics the sword needs.
- K mentions that he requires a steel sword that has been enchanted to defeat a dragon.
- Gerald asks K how soon he needs the completed sword.
- K explains that he needs it as soon as possible in order to help the village and what's left of it's occupants.
- Gerald promises to complete the sword as quickly as possible and with incredible quality.
- K graciously thanks Gerald for his help.

REFERENCES

- [1] Noam Brown and Tuomas Sandholm. 2017. Libratus: The superhuman ai for no-limit poker. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- [2] Sharon Marie Carnicke. 2008. *Stanislavsky in focus: An acting master for the twenty-first century*. Routledge.
- [3] Scott A Crossley, Jennifer L Weston, Susan T McLain Sullivan, and Danielle S McNamara. 2011. The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication* 28, 3 (2011), 282–311.
- [4] Dan Feng, Elin Carstensdottir, Sharon Marie Carnicke, Magy Seif El-Nasr, and Stacy Marsella. 2016. An Active Analysis and Crowd Sourced Approach to Social Training. In *International Conference on Interactive Digital Storytelling*. Springer, 156–167.

- [5] Dan Feng, Pedro Sequeira, Elin Carstendottir, Magy Seif El-Nasr, and Stacy Marsella. 2018. Learning Generative Models of Social Interactions with Humans-in-the-Loop. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 509–516.
- [6] Matthew Guzdial, Brent Harrison, Boyang Li, and Mark O. Riedl. 2015. Crowdsourcing Open Interactive Narrative. In *The 10th Int. Conf. on the Foundations of Digital Games*.
- [7] Galit Haim, Dor Nisim, and Marian Tsatkin. 2016. Human-Computer Agent Negotiation Using Cross Culture Reliability Models. In *International Workshop on Conflict Resolution in Decision Making*. Springer, 118–133.
- [8] Brent Harrison, Siddhartha Banerjee, and Mark O Riedl. 2016. Learning from stories: using natural communication to train believable agents. In *IJCAI 2016 Workshop on Interactive Machine Learning*. New York.
- [9] Brent Harrison and Mark O Riedl. 2016. Towards Learning From Stories: An Approach to Interactive Machine Learning. In *AAAI Workshop: Symbiotic Cognitive Systems*.
- [10] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep Semantic Role Labeling: What Works and What’s Next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [11] Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-jen Hsu, and Kuan-Ta Chen. 2010. KissKissBan: a competitive human computation game for image annotation. *ACM SIGKDD Explorations Newsletter* 12, 1 (2010), 21–24.
- [12] Alice M Isen, Kimberly A Daubman, and Gary P Nowicki. 1987. Positive affect facilitates creative problem solving. *Journal of personality and social psychology* 52, 6 (1987), 1122.
- [13] Boaz Keysar, Shuhong Lin, and Dale J Barr. 2003. Limits on theory of mind use in adults. *Cognition* 89, 1 (2003), 25–41.
- [14] Hee-Cheol Ezra Kim and Kerstin Severinson Eklundh. 2001. Reviewing practices in collaborative writing. *Computer Supported Cooperative Work (CSCW)* 10, 2 (2001), 247–259.
- [15] Joy Kim, Justin Cheng, and Michael S Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 745–755.
- [16] Joy Kim, Sarah Stermann, Allegra Argent Beal Cohen, and Michael S Bernstein. 2017. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 233–245.
- [17] Julia M. Kim, Jr Hill, Paula J. Durlach, H. Chad Lane, Eric Forbell, Mark Core, Stacy Marsella, David Pynadath, and John Hart. 2009. BiLAT: A Game-Based Environment for Practicing Negotiation in a Cultural Context. *IJAIED* 19, 3 (2009), 289–308. <http://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai19-3-03>
- [18] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.
- [19] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. ACM, 1003–1012.
- [20] Iolanda Leite, André Pereira, Allison Funkhouser, Boyang Li, and Jill Fain Lehman. 2016. Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 13–20.
- [21] Boyang Li, Stephen Lee-Urban, Darren Scott Appling, and Mark O Riedl. 2012. Crowdsourcing narrative intelligence. *Advances in Cognitive Systems* 2, 1 (2012).
- [22] Boyang Li, Mohini Thakkar, Yijie Wang, and Mark O Riedl. 2014. Storytelling with adjustable narrator styles and sentiments. In *International Conference on Interactive Digital Storytelling*. Springer, 1–12.
- [23] Shuhong Lin, Boaz Keysar, and Nicholas Epley. 2010. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology* 46, 3 (2010), 551–556.
- [24] Hugo Liu and Push Singh. 2002. MAKEBELIEVE: Using commonsense knowledge to generate stories. In *AAAI/IAAI*. 957–958.
- [25] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. <http://www.aclweb.org/anthology/P/P14/P14-5010>
- [26] Bruce Mason and Sue Thomas. 2008. A million penguins research report. *Institute of Creative Technologies, De Montfort University, Leicester, United Kingdom* (2008).
- [27] Michael Mateas and Phoebe Sengers. 2003. *Narrative intelligence*. J. Benjamins Pub.
- [28] Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication* 27, 1 (2010), 57–86.
- [29] Johnathan Mell, Gale M Lucas, and Jonathan Gratch. 2018. Welcome to the Real World: How Agent Strategy Increases Human Willingness to Deceive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1250–1257.
- [30] Margaret Mitchell, Dan Bohus, and Ece Kamar. 2014. Crowdsourcing language generation templates for dialogue systems. *Proceedings of the INLG and SIGDIAL 2014 Joint Session* (2014), 172–180.
- [31] Robert R Morris, Mira Dontcheva, Adam Finkelstein, and Elizabeth Gerber. 2013. Affect and creative performance on crowdsourcing platforms. In *Affective computing and intelligent interaction (ACII), 2013 humane association conference on*. IEEE, 67–72.
- [32] Jeffrey David Orkin. 2013. *Collective artificial intelligence: simulated role-playing from crowdsourced data*. Ph.D. Dissertation. Massachusetts Institute of Technology.
- [33] Nicolas Paton and Abdullah Almaatouq. 2018. Empirica: Open-Source, Real-Time, Synchronous, Virtual Lab Framework. (Nov. 2018). <https://doi.org/10.5281/zenodo.1488413>
- [34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. <http://www.aclweb.org/anthology/D14-1162>
- [35] Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 979–988.
- [36] Melissa Roemmele, Andrew S Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*.
- [37] Avi Rosenfeld, Inon Zuckerman, Erel Segal-Halevi, Osnat Drein, and Sarit Kraus. 2016. NegoChat-A: a chat-based negotiation agent with bounded rationality. *Autonomous Agents and Multi-Agent Systems* 30, 1 (2016), 60–81.
- [38] Roger C Schank and Robert Abelson. 1977. Scripts, goals, plans, and understanding. (1977).
- [39] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.
- [40] Sigal Sina, Avi Rosenfeld, and Sarit Kraus. 2014. Generating Content for Scenario-Based Serious-Games Using CrowdSourcing.. In *AAAI*. 522–529.
- [41] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, 1223–1237.
- [42] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.
- [43] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.
- [44] Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 75–78.
- [45] Jerry S Weltman, S Sitharama Iyengar, and Michael Hegarty. 2013. Mind the gap: Collecting commonsense data about simple experiences. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 179–190.
- [46] Andrew Whiten. 1991. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*. Basil Blackwell Oxford.