Asking the Right Question: Inferring Advice-Seeking Intentions from Personal Narratives

Liye Fu
Cornell University
liye@cs.cornell.edu

Jonathan P. Chang
Cornell University
jpc362@cornell.edu

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

Abstract

People often share personal narratives in order to seek advice from others. To properly infer the narrator's intention, one needs to apply a certain degree of common sense and social intuition. To test the capabilities of NLP systems to recover such intuition, we introduce the new task of inferring what is the advice-seeking goal behind a personal narrative. We formulate this as a cloze test, where the goal is to identify which of two advice-seeking questions was removed from a given narrative.

The main challenge in constructing this task is finding pairs of semantically plausible advice-seeking questions for given narratives. To address this challenge, we devise a method that exploits commonalities in experiences people share online to automatically extract pairs of questions that are appropriate candidates for the cloze task. This results in a dataset of over 20,000 personal narratives, each matched with a pair of related advice-seeking questions: one actually intended by the narrator, and the other one not. The dataset covers a very broad array of human experiences, from dating, to career options, to stolen iPads. We use human annotation to determine the degree to which the task relies on common sense and social intuition in addition to a semantic understanding of the narrative. By introducing several baselines for this new task we demonstrate its feasibility and identify avenues for better modeling the intention of the narrator.

1 Introduction

> "Computers are useless.
> They can only give you answers." - Pablo Picasso

People often share their personal experiences to elicit advice from others. These personal narratives provide the necessary context for properly understanding the informational goals of the narrators. Endowing automated systems with the capability to infer these advice-seeking intentions could support personalized assistance and more empathetic human-computer interaction.

Personal narrative: I am generally a person who needs a lot of sleep, but today I was not able to sleep more than 6 hours and I am extremely tired. My eyes hurt and two hours later I have programming [lesson] so I have to be alert. I've already drunk a cup of coffee and although I rarely drink coffee, it had no effect on me. I am not at home so I have limited possibilities as for food. I don't want to do anything too unhealthy such as drinking 10 cups of coffee, tho I may consider drinking another one.

Which advice-seeking question is more likely to have been asked by the narrator:
Q1: Is it even possible to be addicted to coffee?
Q2: How can I energize myself?

Figure 1: An abbreviated instance from the ASQ dataset. A personal narrative is matched with two plausible advice-seeking questions, only one of which was actually asked by the narrator when sharing the story.

As humans, to properly distill the narrator's intention from the events and situations they describe, we need to apply a certain degree of social intuition (Conzelmann, 2012; Conzelmann et al., 2013; Baumgarten et al., 2015; Kehler and Rohde, 2017). As an example, consider the goals of a narrator sharing the personal story in Figure 1. We are presented with a wealth of information about the narrator's general sleep patterns, about a particular sleep deprivation situation and its physiological effects, about an upcoming lesson, about coffee intake, its effects, and potential health impacts, and about the current location of the narrator and its impact on food supply. Taking these facts separately, we can imagine providing advice on how to get more sleep, on whether to postpone the lesson,

	Task	**Desired output**
A	Question generation Reading comprehension	What do I need to do in 2 hours?
	Summarization	I must go for a lesson after getting little sleep.
B	Ending generation Narrative chains, story cloze	Lastly, I tried an energizing drink.
C	Event2Mind Desire fulfillment	to learn to code, to be educated
D	**Our task**	How can I energize myself?

Table 1: Contrast with desired outputs in related narrative understanding tasks that focus on the within-story (intradiegetic) aspects of narrative understanding. Tasks are grouped according to the categories discussed in Section 2.1 (which also includes corresponding references). We assumed the second sentence ("My eyes hurt and two hours later I have programming lesson so I have to be alert.") to be the answer span for question generation, and the input for Event2Mind (which operates at sentence level).

on how to get food delivered, or on the risks of caffeine intake. However, given how the narrative is constructed, we can intuit that the more likely goal of the narrator is to get advice on how to overcome the effects of sleep deprivation so that they can be alert for the upcoming programming lesson.

Importantly, the primary goal of our proposed task is not to understand details about the narrator's actions in the story ("Why is the narrator tired?", "When do they need to go to the lesson?"), but to infer the reason why the narrator is sharing this story (i.e., "To get advice on how to stay alert in the next few hours."). That is, we are not concerned with the *intradiegetic* aspects of the narrative, but with the *extradiegetic* intention of the narrator in sharing the story. While an understanding of the former is likely necessary for the latter, it is often not sufficient.

In this work, we introduce a task and a large dataset to evaluate the capabilities of automated systems to infer the narrator's (extradiegetic) intention in constructing and sharing an advice-seeking personal story. This complements existing narrative understanding tasks which focus on testing semantic understanding of events, actors and their (intradiegetic) intentions within the narrative itself. Table 1 contrasts the goals of these existing narrative understanding tasks with that of inferring a narrator's advice-seeking intention, in the context of our introductory example.

Formally, we implement the task as a binary choice cloze test, where the goal is to identify which of two candidate advice-seeking questions was actually asked by the narrator of a given personal narrative. Beyond collecting a large and diverse set of realistic personal stories that contain an advice-seeking question, the main challenge in constructing this task is finding a plausible alternative advice-seeking question for each given narrative. To address this challenge, we develop a methodology for identifying such questions by exploiting both the commonalities in experiences people share online and the diversity of possible advice-seeking intentions that can be tied to similar experiences.

By applying our methodology to a large collection of online personal narratives, we construct a dataset of over 20,000 cloze test instances, covering a very broad spectrum of realistic advice-seeking situations.[1] Each instance contains a narrative that is matched with two advice-seeking questions, one of which is actually asked by the narrator (Q2 in our introductory example), and the other semantically related to the narrative (Q1).

We use human annotations to judge the relative difficulty of different subsets of the test instances and the type of reasoning necessary to solve them. We find that more than half of the instances contain pairs of questions that are not only semantically related to the narratives but also do not contain any explicit factual mismatches with the stories. These are thus unsolvable by pure logical reasoning and require some degree of common sense or social intuition. And indeed, simple

[1] The dataset is available at `https://github.com/CornellNLP/ASQ`.

baseline approaches perform worse on these types of instances, highlighting the need for more direct modeling of the intention of the narrator.

To summarize, in this work we:

- formulate the task of inferring advice-seeking intents from personal narratives (Section 2);
- develop a methodology to construct a large dataset of personal narratives matched with plausible options for **a**dvice-**s**eeking **q**uestions (the ASQ dataset) to be used for this task (Section 3);
- show the task is viable and evaluate the relative difficulty of its items (Sections 4 & 5).

We end by discussing the practical implications of endowing systems with the capability to infer advice-seeking intentions and use our results to identify avenues for developing better models.

2 Task formulation

To evaluate the capability of automated systems to infer advice-seeking intentions, we formulate a cloze-style binary choice test where the system is presented with a personal narrative and is required to choose between two plausible candidate questions: one actually asked by the narrator and the other one not (as exemplified in Figure 1).

We motivate the task by contrasting it with other (narrative) understanding tasks (Section 2.1), and provide the rationale for this particular formulation by discussing its advantages (Section 2.2).

2.1 Related narrative understanding tasks

There are many tasks involving reading comprehension in general, and story understanding in particular. Given a narrative, there are a few broad categories of questions that may be asked to test different types and degrees of understanding. Table 1 follows directly from the discussion below, by contrasting the goal of our task with those of (intradiegetic) narrative understanding tasks in the context of our introductory example.

A: What happened in the story? The most direct approach to test story understanding is to check whether the reader could comprehend the events and actions that occur within the story. This requires semantic understanding, but nothing more. This type of task can be set up in various forms, as the system can be asked to summarize the story (summarization, see Nenkova (2011); Allahyari et al. (2017) for surveys), generate a question that is answerable from the text (question generation (Du et al., 2017)), or answer a question for which the information can be retrieved or reasoned directly from the story (reading comprehension, see Chen (2018) for a survey; notable datasets include MCTest (Richardson, 2013) and NarrativeQA (Kočiský et al., 2018)).

B: What might happen next? While reading the story, people not only grasp and process the events that already occurred but also have some intuition of its likely trajectory. Related tasks include the narrative cloze task (Chambers and Jurafsky, 2008), the story cloze test (Mostafazadeh et al., 2016; Chaturvedi et al., 2016), and its generative versions (Guan et al., 2019). These tasks might require some common sense reasoning on top of semantic understanding; the fact that they aim to predict the future might require a deeper level of understanding than the previous tasks.

C: What can we infer about the characters? When people read a narrative, they not only grasp the facts explicitly stated in the story, but also make inferences about the actors' mental states, such as their attitudes and desires, as the story unfolds.[2] Oftentimes, such an understanding requires inference, either logical or based on common sense reasoning. Such tasks can aim to generate the likely intents and reactions from the actors involved in the events (Rashkin et al., 2018a,b), or to determine whether a given desire of the protagonist was fulfilled (Rahimtoroghi et al., 2017).

D: What is the intention of the narrator in sharing their story? While these prior tasks cover a wide range of angles to narrative understanding, they take an intradiegetic view by focusing on understanding the story itself. We propose another dimension to this line of work by taking an outside-the-story (extradiegetic) perspective[3] and aiming to understand *why* the story is shared by the narrator, potentially inferred from *how* the narrator decides to construct it. In particular, the task introduced here is to infer the *advice-seeking* intention of the narrator.[4]

[2] See Bratman (1987) for an account of the Belief-Desire-Intention model of human practical reasoning.

[3] Recognizing the importance of these two different perspectives for story understanding, Swanson et al. (2017) attempted to classify narrative clauses into intradiegetic vs. extradiegetic levels.

[4] Sharing personal stories can have other goals, e.g., therapeutic (Pennebaker, 1997; Pennebaker and Seagal, 1999).

We argue that solving this task requires not only the semantic understanding and common sense reasoning involved in prior tasks but also a certain degree of social intuition. To uncover the goals of the narrator, one needs to find cues in the narrative construction—what has been selectively included or emphasized, and what might have been purposefully omitted (Labov, 1972). In fact, such intention-understanding tasks are often included in "social intelligence" tests (Conzelmann et al., 2013; Baumgarten et al., 2015).

2.2 Advantages of cloze test formulation

To evaluate the capacity of NLP systems to solve this task, we consider a binary choice cloze test formulation for two main reasons. First, it allows natural ground-truth labels: often, when people share their personal experiences to seek advice, they add explicit requests for the information they are seeking. After removing these requests from the narratives, we can use them as proxies for the narrators' intentions. Second, the binary choice operationalization also has the advantage of non-ambiguity in evaluations and ease of comparisons between systems (as opposed to a generation task).

It is worth noting that our dataset is constructed in a way that allows easy modifications into other task formats if so desired. For instance, the methodology of identifying a plausible false choice for a given narrative could be applied multiple times to extend the task to a more difficult multiple-choice version. Similarly, by ignoring the incorrect question in each instance, our dataset can be used as a source for a new generation task, i.e., generating the advice-seeking question from the given narrative.

3 Task implementation

For a meaningful implementation of the proposed task, the collection of test instances must conform to several expectations, in terms of both the narratives and their (actual) advice-seeking questions. In what follows we outline these desiderata and our method for collecting instances that meet them (Section 3.1).

Furthermore, as with any multiple-choice cloze test formulation, the difficulty of each test instance largely depends on how plausible the alternative answers are. Yet, finding plausible (but not actually correct) alternatives automatically is challenging. Not surprisingly, many of the cloze-style multiple-choice datasets use humans to write these alternatives (Mostafazadeh et al., 2016; Xie et al., 2018), limiting their scalability.

We tackle this challenge by developing a methodology that exploits both the commonalities in human experiences shared online and the diversity in the types of advice needed for similar situations under different circumstances (Section 3.2).

3.1 Collection of candidate instances

Narratives desiderata. As a pre-requisite, we need to start from personal narratives containing advice-seeking needs that are explicitly expressed (as questions), and that can be removed to form the cloze test instances.[5] Ideally, these narratives would cover a broad range of topics, in order to be able to test how well a system can generalize to a diverse range of real-life scenarios, rather than apply only to restricted and artificial settings.

Question desiderata. Not all questions contained within an advice-seeking narrative are suitable for our task. Some of the questions might be too general, while others might be rhetorical. For instance, *Any advice?* holds no particular connection with the context of the narrative in which it appears. To contribute to meaningful test instances, questions need to meet a level of relevance and specificity such that (at least) humans could match them with the narratives from which they are extracted.

Data source. We start from a dataset of over 415,000 advice-seeking posts collected from the subreddit r/Advice, which self-defines as "a place where anyone can seek advice on any subject".[6] We only use publicly available data and will honor the authors' rights to remove their posts.

Applying cloze. For each post, we strip off all questions that appear in any position of the post, including the post title.[7]

We keep the remaining narratives as the cloze texts.[8] Figure 2 shows how the cloze transforma-

[5] An interesting future work avenue could be considering narratives that only have implicit advice-seeking intentions.

[6] We start from an existing collection of Reddit posts (Tan and Lee, 2015) which we supplement with The Baumgartner Reddit Corpus retrieved via Pushshift API on Nov. 21, 2018.

[7] To identify questions, we use the simple heuristic of looking for sentences that end with '?' or start with *why, how, am, is, are, do, does, did, can, could, should, would.*

[8] To ensure that the cloze text can provide sufficient context, yet are not overly verbose, we only consider cloze texts that are 50-300 tokens long. This is a choice we made prior to any experiments, and we do not claim it is the optimal range to set up the task.

Selected topics	**Question keywords**	**Example questions**
Housing	move live house city apartment roommate	What is it like living with **roommates**? Should I **move** to the **city**?
School	college school class degree study	Should I drop out of **college**? What's the best way for me to **study** for my biology tests?
Work	job boss quit work interview employer	Can I somehow ask to **work** from home? How do I explain during an **interview** why I left a **job**?
Relationships	girl date text tell guy think crush	Does it sound like this **girl** may like me? How can I think of a better greeting for online **dating**?
Personal finances	money car pay rent loan insurance	How do I afford a **car** in my situation? Am I stupid for wanting a student **loan**?
Family	parent convince let mom dad sister	How do I **convince** my **parents** to believe me? How can I try and make a better relationship with my **sister**?

Table 2: Selected narrative topics and example question keywords associated with each topic.

Title: How can I energize myself?

I am generally a person who needs a lot of sleep [...] I don't want to do anything too unhealthy such as drinking 10 cups of coffee, tho I may consider drinking another one. Help? What has worked for you?

Figure 2: Cloze application to the post from which we obtain the introductory test instance. After filtering out questions that are too general, only the title question remains as a candidate for representing the actual advice-seeking intention of the narrator.

tion is applied to the post containing our introductory example.

Selecting ground-truth test answers.[9] We select candidate ground-truth answers for the cloze test as the ?-ending sentences removed from narratives. In order to keep only well-formed information-seeking questions, we filter the candidate questions by keeping only those that start with interrogatives[10] or *any, anyone, help, advice, thoughts*. To further discard questions that are too general, we compute a simple specificity score $S(q)$ of a question q containing the set of words $\{w_1, w_2, \ldots, w_N\}$ as its maximum inverse document frequency (idf):

$$S(q) = S(\{w_1, w_2, \ldots, w_N\}) = \max_{i \in N} \mathrm{idf}(w_i),$$

and filter out questions for which $S(q) < 5$ or questions that have less than 5 words. At the end of this selection process, from the example post in Figure 2, *Help?* and *What has worked for you?* are discarded and the title question is kept as the ground-truth answer to this cloze instance. If multiple questions survive the filtering process, we select one at random.

Diversity evaluation. To verify that the resulting data has broad topical diversity in both narratives and questions, we perform a two-step clustering analysis. First, we use singular value decomposition on tf-idf transformed narratives to obtain their vector representations, we then cluster similar narratives using k-means to surface underlying topics. Next, for each topic, we extract nouns and verbs from the questions attached to each narrative in the topic, and surface common question keywords as those with high document frequency within the topic, correcting for their global document frequency (via subtraction).

To provide a qualitative feel of the diversity of the data, Table 2 shows a selection of the resulting narrative topics and question keywords, together with example questions (corresponding narratives can be found in the data release). We find a

[9] As it happens, test answers are actually questions.

[10] We consider the following set of words as interrogatives: *what, when, why, where, which, who, whom, whose, how, am, is, are, was, were, do, does, did, has, have, had, can, could, shall, should, will, would, may, might, must*.

wide range of experiences represented in the narratives, from relationships to student life to apartment rentals. Furthermore, within each narrative topic, there is a variety of question types; for instance, questions related to housing could be about dealing with roommates, paying rent, or choosing a city to live in.

3.2 Finding alternative test answers

To find plausible alternative answer options for each candidate cloze test instance, one direct approach could be to find questions that are semantically related to the ground-truth question. However, there are two underlying problems with this approach. First, the task of finding semantically similar questions is itself very challenging (Haponchyk et al., 2018), given their terseness and lack of context. Second, semantic similarity is arguably a different concept from *plausibility* with respect to a narrative. For example, the two questions in the introductory example are semantically distant, but they are both plausible in the context of the narrative.

Our main intuition in solving this problem is that individuals who are in similar situations tend to have advice-seeking intentions that are related. For each candidate cloze test narrative instance, we can thus search for a similar narrative first (by exploiting commonalities in experiences people share online) and then select an advice-seeking question from that narrative as the *alternative* answer for the test.

Narrative pairing. To operationalize this intuition, we first find pairs of similar narratives based on the cosine similarity of their tf-idf representations.[11] A greedy search based on this similarity metric results in a set of pairs of related narratives ($\mathrm{N}_1, \mathrm{N}_2$) with their respective advice-seeking questions ($\mathrm{qn}_1, \mathrm{qn}_2$) identified in the previous step.

Narrative masking. At this point, the pair of advice-seeking questions could be used with either narrative to form a test instance. For example, Figure 3 shows the other possible cloze instance corresponding to the introductory example if we were to use the other narrative in the narrative pair. This, however, would arguably be a poor test instance since Q2 is hardly applicable to this other narrative. More generally, we want to ensure that our choice of which narrative (N_i) to include in the cloze test optimizes the plausibility of the question pair ($\mathrm{qn}_1, \mathrm{qn}_2$).

Masked narrative: I've noticed something, over the past few years I've gained a habit of drinking coffee. The average day is about six cups, but it can exceed that sometimes (8 or so). The only reason I question my habit is cause I'm up at 4AM right now cause I couldn't fall asleep. I honestly have a headache in the morning until I drink a cup of coffee. I'll have some for essentially no reason, I'll just make some out of a urge almost.

Q1: Is it even possible to be addicted to coffee?
Q2: How can I energize myself?

Figure 3: Alternative cloze test instance corresponding to the introductory example.

To achieve this, we compute the similarity between each narrative in the pair and each of the two respective questions,[12] and select the narrative that maximizes the minimum question-narrative similarity. Formally,

$$N_i = \arg\max_i \textsc{min}\{sim(N_i, qn_1), sim(N_i, qn_2)\}.$$

Importantly, this selection criterion is purposely symmetric with respect to the two questions in order to avoid introducing any unnatural preference between the two that a classifier (with no access to the masked narrative) could exploit.

As a final check, we ensure that in each cloze instance the two questions are neither too similar to each other (and thus indistinguishable) nor too dissimilar (which may indicate unsatisfactory narrative pairings). To this end, we discard instances in which the questions have extremely high or low surface similarity according to their InferSent (Conneau et al., 2017) sentence embeddings.[13]

This process leaves us with a total of 21,865 instances. A detailed account of the number of instances filtered at different stages of the construction process can be found in the Appendix.

[11] We consider both unigrams and bigrams, and set a minimum document frequency of 50. We also remove likely duplicates (cosine > 0.8) and cases for which the similarity between narratives is too low (cosine < 0.1). We have also experimented with embedding-based representations to compute cosine similarities from, but they do not seem to produce qualitatively better pairings upon inspection.

[12] To account for the terseness of the questions, we represent both narratives and questions with tf-idf weighted GloVe embeddings (Pennington et al., 2014) and compute the cosine similarity between them.

[13] We set a lower bound of 0.8 and an upper bound of 0.95. We choose this representation because questions are short and thus we anticipate tf-idf representation to be less informative.

4 Human performance

To understand the feasibility of the task, as well as the relative difficulty of the items in the dataset, eight non-author annotators labeled a random sample of 200 instances.[14] Each annotator is asked to choose first, out of the two candidate questions, which they consider to be *more likely* to have been asked by the narrator. Overall, human annotators achieve an accuracy of 90% (Cohen's κ = 0.79),[15] showing that humans can indeed recover the advice-seeking intentions of the narrators, and thus validating the feasibility of the task.[16]

We are also interested in understanding the types of skills needed to solve the task. In particular, we want to estimate the proportion of the task instances that can not be solved by mere factual reasoning. To this end, we ask humans to identify candidate questions that contain a factual mismatch with the narrative, making them **E**xplicitly incompatible; 57% of the annotated instances do not contain any such mismatches in any of the questions. Similarly, we want to estimate how many instances require common sense expectations about the behavior of the protagonist (within the story). So we ask annotators to mark questions as being **I**mplicitly incompatible if they do not contain any factual mismatches, but they are incompatible with what can be inferred implicitly about events and characters in the story.

The questions that are neither explicitly nor implicitly incompatible would be labeled as being **C**ompatible, and as either **L**ikely or **U**nlikely to represent the narrators' intentions. Test items in our data forcing a choice between **C**ompatible questions are expected to be the hardest to solve, as they might require a certain degree of social intuition in addition to factual and common sense reasoning. Table 3 provides an example narrative and one representative question from each of the above-mentioned categories.[17]

Narrative: I asked a girl that I really like if she would like to get coffee sometime. She said she's really busy but that we'll see. I can't get her off my mind and I spend all day waiting for her to tell me she's free.

Explicitly incompatible (E):
How to deal with my roommate?

Implicitly incompatible (I):
What to do if I asked a girl out and now regret it?

Compatible (C) but unlikely (U):
Which coffee place would you recommend?

Compatible (C) and likely (L):
Would it seem desperate if I asked her again in a more direct way a week later?

Table 3: Example questions in each plausibility category for an example narrative.

Table 4 shows a human performance breakdown according to some of the most common types of instances in our data.[18] As expected, instances involving only compatible questions (C + C) are harder to solve,[19] as they might require some social intuition, whereas when explicit contradictions exist (C + E), they are perfectly solvable. We also note that humans can perfectly solve the subset of task instances (L + {U, I}) that exhibit perceived qualitative differences between the actual and the alternative questions, but nevertheless, require more than semantic understanding (and sometimes require social intuition).

Pair type	SIM	FT-LM	HUMAN	% in data
C + E	86%	88%	100%	38%
C + {C, I}	68%	74%	89%	46%
C + C	66%	73%	84%	32%
L + {U, I}	75%	75%	100%	30%
OVERALL	76%	80%	90%	

Table 4: Breakdown of performances on selected ACTUAL + ALTERNATIVE question pair types. For instance, the pair type C + E corresponds to instances where the ACTUAL question asked by the narrator is compatible and the ALTERNATIVE question is explicitly incompatible.

[14] See the Appendix for detailed annotation instructions.

[15] We obtained a second round of annotations on a subset of 75 task instances to compute agreement statistics.

[16] By construction, random accuracy is 50%.

[17] The example is adapted from our instructions to annotators, which includes further explanations for these categories. See the Appendix for details.

[18] See the Appendix for some representative examples for selected question pair types in our data.

[19] We also concede that some of the instances in this category may be unsolvable, e.g., when the wrong question fits the narrative just as well.

Model	Accuracy (held-out)
NARRATIVE-QN-SIM	73.4%
FINETUNED LM	**78.7%**

Table 5: Performance of different baselines.

5 Baseline systems performance

We divide our data into a 8,865-2,500 train-test split and have reserved 10,000 instances as a held-out set.[20] In Table 5 we report accuracy for the best-performing model on the (never-before-seen) held-out for a simple similarity-based method and for a deep learning method.

Narrative-question similarity. We expect that questions would show greater similarity to narratives they are removed from. We thus establish a narrative-question similarity baseline by considering features based on cosine similarities between narrative and questions, with text represented as tf-idf vectors, tf-idf weighted GloVe embeddings, averaged GloVe embeddings, as well as word overlap between content words, all combined in a logistic regression model.

Finetuned transformer LM. We also use a Finetuned Transformer LM model (Radford et al., 2018), which was shown to perform competitively on a diverse set of NLP tasks, achieving state-of-the-art results on the story cloze test.[21]

5.1 Error analysis

Required skills. As shown in Table 4, systems perform worst on items that do not exhibit any (implicit or explicit) mismatches (C + C), and thus might require some social intuition. Importantly, the largest gap between baseline and human performance (25%) is on the subset of items that can not be solved based solely on a semantic understanding (L + $\{$U, I$\}$). These results underline the need for models that can combine common sense reasoning about the events within the story with an intuition about the intention of the narrator.

Question concreteness. Questions may also differ in how concrete they are. In a preliminary analysis aimed at understanding how this property affects performance, we compare words used in ground-truth questions that the best-performing model predicts correctly with those used in questions that are classified incorrectly. We observe that questions that are predicted correctly have significantly higher average inverse document frequencies (t-test $p < 0.01$). Intuitively, these more specific questions may be more concrete in nature, making them easier to connect to the narratives to which they belong. We also find that some common interrogatives have skewed distributions. For instance, questions starting with *Is* are less likely to be classified correctly than those starting with *How*. A cursory manual investigation suggests that this can also be tied by concreteness, with the latter type of questions appearing to be more concrete than the former.

6 Further related work

One broad motivation behind our work is to eventually help better support personalized informational needs (Teevan et al., 2007). This connects to several related lines of work that were not previously discussed.

Query/question intents. Datasets and models are proposed for understanding user intents behind search queries (Radlinski et al., 2010; Fariha et al., 2018), or even more generally, user questions (Haponchyk et al., 2018). To complement this line of work that looks at user intents behind the explicit request, our task aims to uncover user intents when they are implied in personal narratives (without access to the explicit question).

Conversational search/QA. One way to better satisfy user intents is by making such processes collaborative (Morris and Horvitz, 2007; Morris, 2013), or conversational (Radlinski and Craswell, 2017). Conversational QA datasets (Choi et al., 2018; Reddy et al., 2019) have been introduced to help develop systems with such capability.

Social QA. Some questions posed by users are inherently more social in nature, and require more nuanced contextual understanding (Harabagiu, 2008). The social nature may affect how people ask questions (Dahiya and Talukdar, 2016; Rao and Daumé III, 2018), and pose challenges for identifying appropriate answers (Shtok et al., 2012; Zhang et al., 2017).

[20] The set annotated by humans is disjoint.

[21] We fine-tune with our training set on top of the pre-trained transformer language model, using the implementation from `https://github.com/huggingface/pytorch-openai-transformer-lm`.

7 Discussion

In this work, we introduce the new task of inferring advice-seeking intentions from personal narratives, a methodology for creating appropriate test instances for this task and the ASQ dataset. This task complements existing (intradiegetic) narrative understanding tasks by focusing on extradiegetic aspects of the narrative: in order to understand "Why is the narrator sharing this?", we often need to apply a certain degree of common sense and social intuition.

From a practical perspective, this extradiegetic capability is a prerequisite to properly address personalized information needs that are constrained by personal circumstances described as free-form personal stories. Currently, to address these types of information needs, people seek (or even hire) other individuals with relevant experience or expertise. As with conversational search (Radlinski and Craswell, 2017), we can envision systems that can more directly address complex information needs by better understanding the circumstances and intentions of the user.

Our analysis of the human and baseline performance on different types of test instances points to interesting avenues for future work, both in terms of designing better-performing systems and in terms of constructing better test data. We envision that (intradiegetic) narrative understanding could help identify the components of the narrative that are most relevant to the advice-seeking goal. For example, identifying the narrator's intentions and desires within the story (Rashkin et al., 2018b), and whether these desires are fulfilled (Rahimtoroghi et al., 2017) could help focus the attention of the model, especially when dealing with less concrete questions. Furthermore, a better representation of the structure of the narrative (Ouyang and McKeown, 2014), in terms of discourse acts (Elson, 2012) and sentiment flow (Ouyang and McKeown, 2015), could also help distinguish between spurious and essential circumstances of the narratives.

In terms of improving the task itself and the methodology for creating testing instances that better approximate the inferential task, we note a few possible directions. Firstly, better narrative modeling could lead to higher quality matching. Similarly, better representation of the questions can help select more appropriate candidate options (e.g., currently 6% of the questions are deemed by the annotators to be too general). In addition, the generative version of the task, when appropriately evaluated, could be a closer operationalization for intention inference, and also offer more potential for practical uses.

Finally, future work could expand on our methodology to formulate other more general tasks aiming to understand the reasons why a person is sharing a personal story. While we have focused on narratives shared with the intention of seeking advice, people may also share stories to express emotions, to entertain or educate others. A better understanding of these different (explicit or implicit) intentions could lead to more personalized and empathetic human-computer interaction.

Acknowledgments. The authors thank Tom Davidson, Tsung-Yu Hou, Qian Huang, Hajin Lim, Laure Thompson, Andrew Wang, Xiaozhi Wang and Justine Zhang for helping with the annotations. We are grateful to Thorsten Joachims, Avery Quinn Smith and Todd Cullen for helping us when our server crashed on the day of the deadline while testing the model on the held-out set, to Lillian Lee, Andrew Wang, Justine Zhang and the anonymous reviewers for their helpful comments, and to Fernando Pereira for the early discussions that inspired this research direction. This work is supported in part by NSF CAREER award IIS-1750615 and NSF Grant SES-1741441.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. *arXiv:1707.02268v3*.

Melanie Baumgarten, Heinz-Martin Süß, and Susanne Weis. 2015. The Cue Is the Key: The Relevance of Cues and Contextual Information in the Social Understanding Tasks of the Magdeburg Test of Social Intelligence. *European Journal of Psychological Assessment*, 31(1).

Michael Bratman. 1987. *Intention, Plans, and Practical Reason.*

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL.*

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daume III. 2016. Ask, and Shall You Receive? Understanding Desire Fulfillment in Natural Language Text. In *Proceedings of AAAI.*

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of EMNLP*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of EMNLP*.

Kristin Conzelmann. 2012. *Social Intelligence and Auditory Intelligence – Useful Constructs?* Ph.D. thesis, The Otto von Guericke University Magdeburg.

Kristin Conzelmann, Susanne Weis, and Heinz-Martin Süß. 2013. New Findings About Social Intelligence: Development and Application of the Magdeburg Test of Social Intelligence (MTSI). *Journal of Individual Differences*, 34(3).

Yogesh Dahiya and Partha Talukdar. 2016. Discovering Response-Eliciting Factors in Social Question Answering: A Reddit Inspired Study. *Proceedings of ICWSM*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proceedings of ACL*.

David K. Elson. 2012. *Modeling Narrative Discourse*. Ph.D. thesis, Columbia University.

Anna Fariha, Sheikh M. Sarwar, and Alexandra Meliou. 2018. SQuID: Semantic Similarity-Aware Query Intent Discovery. In *Proceedings of SIGMOD*.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story Ending Generation with Incremental Encoding and Commonsense Knowledge. *Proceedings of AAAI*.

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised Clustering of Questions into Intents for Dialog System Applications. In *Proceedings of EMNLP*.

Sanda M. Harabagiu. 2008. Questions and Intentions. In *Advances in Open Domain Question Answering*.

Andrew Kehler and Hannah Rohde. 2017. Evaluating an Expectation-Driven Question-Under-Discussion Model of Discourse Interpretation. *Discourse Processes*, 54(3).

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl M. Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *TACL*.

William Labov. 1972. The Transformation of Experience in Narrative Syntax. *Language in the Inner City*.

Meredith R. Morris. 2013. Collaborative Search Revisited. In *Proceedings of CSCW*.

Meredith R. Morris and Eric Horvitz. 2007. SearchTogether: An Interface for Collaborative Web Search. In *Proceedings of UIST*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of NAACL*.

Ani Nenkova. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2).

Jessica Ouyang and Kathleen McKeown. 2014. Towards Automatic Detection of Narrative Structure. In *Proceedings the LREC*.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling Reportable Events as Turning Points in Narrative. In *Proceedings of EMNLP*.

James W. Pennebaker. 1997. Writing About Emotional Experiences as a Therapeutic Process. *Psychological Science*, 8(3).

James W. Pennebaker and Janel D. Seagal. 1999. Forming a Story: The Health Benefits of Narrative. *Journal of Clinical Psychology*, 55(10).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training. *Preprint*.

Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of CHIIR*.

Filip Radlinski, Martin Szummer, and Nick Craswell. 2010. Inferring Query Intent from Reformulations and Clicks. In *Proceedings of WWW*.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling Protagonist Goals and Desires in First-Person Narrative. In *Proceedings of SIGDIAL*.

Sudha Rao and Hal Daumé III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *Proceedings of ACL*.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling Naive Psychology of Characters in Simple Commonsense Stories. *Proceedings of ACL*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018b. Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of ACL*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *TACL.*

Matthew Richardson. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of EMNLP.*

Anna Shtok, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. Learning from the Past: Answering New Questions with Past Answers. In *Proceedings of WWW.*

Reid W. Swanson, Andrew S. Gordon, Peter Khooshabeh, Kenji Sagae, Richard Huskey, Michael Mangus, Ori Amir, and Rene Weber. 2017. An Empirical Analysis of Subjectivity and Narrative Levels in Weblog Storytelling Across Cultures. *Dialogue & Discourse*, 8(2).

Chenhao Tan and Lillian Lee. 2015. All Who Wander: On the Prevalence and Characteristics of Multicommunity Engagement. In *Proceedings of WWW.*

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2007. Characterizing the Value of Personalizing Search. In *Proceedings of SIGIR.*

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-Scale Cloze Test Dataset Created by Teachers. In *Proceedings of EMNLP.*

Wei E. Zhang, Quan Z. Sheng, Jey Han Lau, and Ermyas Abebe. 2017. Detecting Duplicate Posts in Programming QA Communities via Latent Semantics and Association Rules. In *Proceedings of WWW.*

Appendix

A1 Instructions to human annotators

As described in Section 4 of the main paper, we obtained human annotations on a small subset of our data for validation purposes. Annotators were shown instructions which included the definitions for different question-narrative plausibility types, together with two examples to help further clarify the task and the definitions. The exact instructions are reproduced in Table A8, while the examples provided are shown in Table A9.

A2 Distribution of plausibility categories

Table A6 shows the distribution of plausibility categories given out by our human annotators, for both the actual questions which belong to the given narrative (Column 2) and the paired alternative questions (Column 3).

Question type	**% in actual**	**% in altern.**
Compatible and **L**ikely	81%	15%
Compatible but **U**nlikely	10%	21%
Incompatible (**I**mplicit)	5%	16%
Incompatible (**E**xplicit)	2%	41%
Very **G**eneral	3%	8%

Table A6: Data distribution estimated from the human annotated subset. For more than half of the cases, the wrong answer (i.e., the alternative question) could not be simply discarded based on factual mismatches, and the task instance would require additional common sense or social intuition to solve.

# of unique post ids	415,693
# of posts with narrative bodies	339,815
# of narratives with questions	262,721
# of narratives after filtering for length	151,418
# of narratives with specific questions	89,527
# of narratives paired	43,730

Table A7: Counts for narrative instances at different stages of dataset construction.

A3 Further processing details

Table A7 provides the number of instances remaining at each of the processing steps. After masking one narrative from each narrative pair, we have a total of 21,865 narratives, each successfully paired with two plausible candidate questions.

A4 Examples of pair types

In our error analysis, we find that the performance of both our baselines, as well as that of our human annotators, vary depending on question pair type, where pair type is defined as the human-judged plausibility of the ground-truth question and that of the alternative question. To give a better sense of what these pair types look like in practice, Table A10 shows example instances for a few selected pair types.

You will be presented with one narrative and two advice-seeking questions (qn1 and qn2).

Firstly, you will need to indicate which of these questions is **more likely** to be asked by the narrator in the context of the narrative (Column D, in yellow). Use the dropdown menu to select the more likely question among the two. (You must pick one.)

In addition, for **each** question separately, you will need to provide a rating on how plausible the question is in the context of the narrative by choosing one of the following options (dropdown menu in Columns E and F, in green):

1. **Very general**: i.e., this question could follow most narratives, and it's not in any way specific to this narrative.
2. **Compatible and likely**: i.e., the question follows naturally from this narrative.
3. **Compatible but unlikely**: i.e., while there is no direct contradiction (either explicit or implicit) with the narrative, it seems unlikely that the narrator's intention was to ask this question.
4. **Incompatible (explicit)**: i.e., the question is incompatible due to clear factual mismatches with the information explicitly contained in the narrative, or it is completely irrelevant.
5. **Incompatible (implicit)**: i.e., the question is incompatible due to mismatches with something that you can indirectly infer from the narrative.

You should judge each question **separately** when selecting a category. It is possible for both questions to fall into the same category.

Now you need to read the example narratives, questions and explanations in the adjacent cells (B2 and C2) to get a feel of each of the categories, after which you could proceed to the sub-sheet **Items to annotate (see tab at the bottom of this page)** to complete the annotation task.

Optionally, you can also provide comments for each item (scroll to the right to see the comment column): Did you find an item particularly challenging or interesting? Was one of the questions not really asking for an advice? Do share your thoughts with us.

Table A8: Instruction text shown to the annotators.

Example 1	**Example 2**
Narrative: “I am a freshman in college and I have a group of friends that I have been hanging out with for the past couple months. I feel like we have a good time when we hang out, but a lot of the time, the rest of the group will go out and do stuff together, but I won’t be included.”	Narrative: “I asked a girl that I really like if she would like to get coffee sometime. She said she’s really busy but that we’ll see. I can’t get her off my mind and I spend all day waiting for her to tell me she’s free.”
Example questions in each category [and explanations where appropriate]:	Example questions in each category [and explanations where appropriate]:
a) **Very general**: Any advice?	a) **Very general**: What should I do?
b) **Compatible and likely**: Can I ask to be included?	b) **Compatible and likely**: Would it seem desperate if I asked her again in a more direct way a week later ?
c) **Compatible but unlikely**: How to make new friends in college? [explanation: it is more likely that the narrator is trying to be more included in the current group of friends, rather than giving up entirely on them and look for replacement.]	c) **Compatible but unlikely**: Which coffee place would you recommend? [Explanation: While the narrator is trying to invite the girl for coffee, the main concern seems to be whether the attempt would be successful rather the choices between coffee places.]
d) **Incompatible (explicit)**: Any advice for finding new friends for a senior in college? [Explanation: The narrator is a freshman in college, not a senior. This constitutes a clear factual mismatch between the question and the narrative.]	d) **Incompatible (explicit)**: How to deal with my roommate? [Explanation: This question is completely irrelevant to the narrative (no roommate is mentioned).]
e) **Incompatible (implicit)**: What are some excuses to not hang out with them? [Explanation: We can imply from the narrative that the narrator wants to hang out with the group. This is incompatible with a question asking how NOT to do that.]	e) **Incompatible (implicit)**: What to do if I asked a girl out and now regret it? [Explanation: We can infer that the narrator is looking forward to the potential date, which contradicts with the feeling of regret in the question.]

Table A9: Example narratives and questions shown to the annotators.

Pair type	Narrative	Actual question	Alternative question
L + L	Hey everyone I have a bit of a dilemma. It's the first week of school and I am talking three advanced classes, AP world history II, English honors II and Chemistry honors. I am pretty sure that I can handle it but; I am falling behind in chemistry honors and it is the first week. I don't have the mathematical background as the other students. They have taken physics and geometry. I am in a special Algebra class which means I am a year behind in math and science.	Should I drop chemistry honors?	How much do honors courses matter?
L + U	My college roommate/one of my best friends is getting married Saturday. I'm a groomsman, as is our third roommate. Our third roommate gave he and his betrothed their wedding gifts early today: an Xbox One and a crystal decorative bowl from Tiffany. I'm an assistant manager at a sporting goods store making $8.50 an hour, and between rent, utilities, groceries, gas, and my student loan payments, I usually either barely break even every month or have to borrow money from my parents until my next paycheck. I've checked their registry, and even the less expensive gifts are outside what I can afford ($30 can make or break me right now).	What to do about a wedding gift if I'm broke?	Where can I buy food that's cheap, and it'll last me until then?
C + I	I just recently switched schools this school year. I'm pretty okay with how it's going so far academics wise but I have no idea how to put myself out there. Everyone has seemed to have made friend groups already or they already know everyone from previous years. I used to be in a private school so no one really knows me from this school except for my close friends that I've known for a long time.	Is there any way that I could gain any popularity before it's too late?	Is it too early to tell if I want to drop out?
L + E	So there is a dream job which is PERFECT for me and of course I really want it. I called the employer last week and she said she was going to call candidates for interviews that week. Then I called this week and she said she was going to call for interviews this week. And please, no advice telling me 'don't call'. I have nothing to lose, so I'm going to call, I would just really appreciate some advice as to how to ask for an interview appropriately - Thank you all!	How do I call an employer asking for an interview?	When I went for the interview she did seem busy so maybe she was too busy to call?
L + G	I'm in a relationship with an amazing girl and feel very happy with her. Recently though I've been having intrusive thoughts about her ex-boyfriends (her having sex with them, etc) which are leading to feelings of jealousy and it's really disrupting my ability to enjoy my time with her.	How to deal with feelings of jealousy?	Is this "normal" – in the sense of, do other people experience this?

Table A10: Example task instances for different ACTUAL + ALTERNATIVE question pair types.