# Leveraging Commonsense to Extract Implicit Knowledge from Dialogues

**Deepanway Ghosal**[†], **Pengfei Hong**[†], **Siqi Shen**[△],
**Navonil Majumder**[†], **Rada Mihalcea**[△], **Soujanya Poria**[†]

[†] Singapore University of Technology and Design, Singapore
[△] University of Michigan, USA
{deepanway_ghosal, pengfei_hong}@mymail.sutd.edu.sg
{navonil_majumder, sporia}@sutd.edu.sg
shensq,mihalcea@umich.edu

## Abstract

Extracting structured knowledge from text is a fundamental research problem in natural language processing. Human conversations are a rich source of both explicit and implicit knowledge as it requires contextual understanding, planning, inference, and several aspects of reasoning including causal, temporal, and commonsense reasoning. Extracting such knowledge from conversations is a challenging problem and could be conducive to improving several downstream applications. In this paper, we introduce DAIKE – a manually annotated dataset of explicit and implicit knowledge extracted from dyadic conversations. The annotated knowledge is categorized with respect to the presence of commonsense knowledge (e.g., causal, conditional, temporal). We setup three different tasks conditioned on the annotated dataset: Dialogue-level Natural Language Inference, Span Extraction, and Multi-choice Span Selection. Baseline results obtained with transformer-based models show that the task is especially difficult, paving the way for promising future research. The dataset and the baseline implementations are available at https://github.com/declare-lab/DAIKE.

## 1 Introduction

There has been much work on extracting structured knowledge from natural language text. However, there has been only little research to distinguish implicit knowledge from explicit knowledge present in the text. Explicit knowledge can be relatively easily parsed out using semantic parsing (Speer et al., 2017) and simple co-reference resolution (Joshi et al., 2019). Implicit knowledge, however, involves non-trivial inference, which becomes even more challenging on dialogue data due to the contextual interplay and latent background knowledge shared between the speakers. Extraction of both explicit and implicit knowledge could be conducive to improved question-answering systems and richer knowledge bases. To this end, we construct a dataset of Dialogues with Annotated Implicit KnowledgE (DAIKE), as illustrated in Fig. 1, which captures the relation between two textual spans appear in the dialogue. A span can constitute one or multiple entities, objects, actions, states, events, that can be extracted from the dialogue. The relations are commonsense based, as elaborated in §3.2. Each knowledge triplet is tagged either as explicit or implicit.

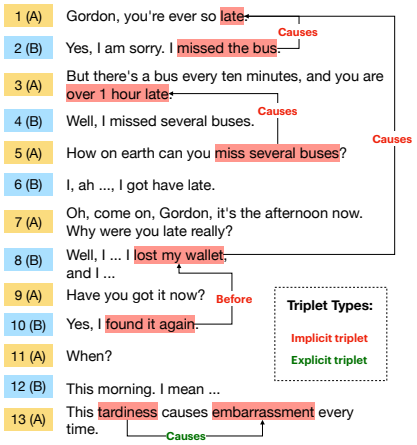



Figure 1: Example of various types of knowledge triplets in a dyadic dialogue; the red and green relations signify implicit and explicit triplets, respectively, whereas the italicized are commonsense.

We define three tasks on this dataset — (i) Dialogue-level Natural Language Inference, (ii) Span Extraction, and (iii) Multi-choice Span Selection — and setup baselines for each of them. Our baselines based on transformer language models found these tasks to be non-trivial and difficult to solve.

**The Importance of this Dataset:** The immediate aim of this research is to develop a rich knowledge base from a dialogue consisting of implicit and explicit knowledge, and then use that to perform inference and reasoning. We formulate non-trivial natural language inference (NLI) and question answering (QA) tasks that can be used to benchmark such reasoning capabilities of natural language processing models.

## 2 Related Work

Recently, language models have been scaled up a lot and have seen a performance improvement on various tasks (Brown et al., 2020; Raffel et al., 2020). However, it has been proved that declarative knowledge is still valuable, especially implicit relationships that are hardly acquired by the state of the art models (Hwang et al., 2020).

The commonsense knowledge bases widely used are of large scale and mainly based on crowdsourced effort. ConceptNet (Speer et al., 2017) is a semantic network with nodes composed of common words or phrases in their natural language form. It contains 34 relations, including taxonomic, temporal, and causal ones, such as *MotivatedByGoal* and *Causes*. However, the knowledge in ConceptNet is annotated solely based on the first entity without any other context, making it difficult to capture the long-tail knowledge outside of the most common ones. With a focus on inferential knowledge, ATOMIC (Sap et al., 2019) consists of nine relations, such as *xIntent* (the intent for personX's action) and *xEffect* (the effect of the event on personX). It covers knowledge around agents involved in the event for if-then reasoning, including subsequent events, mental state, and persona. In addition to being non-contextual, it also ignores causal relationships between events not carried out by a person. In contrast, our work captures relationships between spans across multiple turns in dialogues. Benefited from the dialogue aspect of our data, we also manage to cover implicit knowledge that requires context from conversations to make sense.

More recent work such as GLUCOSE (Mostafazadeh et al., 2020) captures implicit knowledge across multiple sentences. It is annotated based on ROCstories (Mostafazadeh et al., 2016), where each story consists of five short sentences. Our work instead annotates on dialogues, which have more complicated sentences and spoken conversational exchange.

## 3 Background

The primary impetus behind this dataset is the distillation of knowledge in the form of standard knowledge triplets that can be inferred only through commonsense reasoning. We focus on conversations as our data source, with the choice being motivated by the fact that part of the context in conversations is naturally implicit and interlocutor dependent (Grice, 1975). Commonsense knowledge is considered to be the set of all facts and knowledge about the everyday world which is assumed to be known by all humans (Davis, 2014). For this very reason, human-to-human dialogues – typically guided by the Gricean maxims of human interactions – tend to avoid explicit mentions of commonsense knowledge and the associated reasoning steps. It is thus reasonable to assume that conversations are generally likely to hold more context-specific inferable implicit knowledge than monologues. This ensures a rich dataset with plenty of contextual implicit knowledge and a reasonable amount of explicit knowledge.

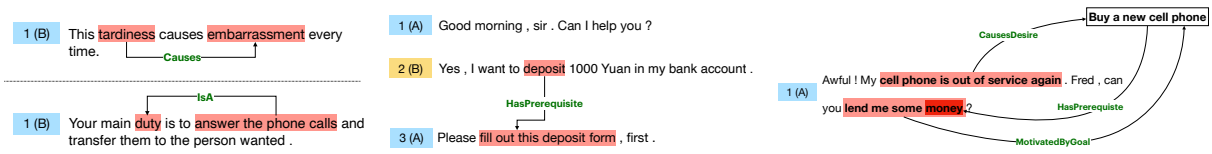
Two distinct spans (e.g., events, entities) in a dialogue may have an implicit connection that can be trivial for humans to interpret using commonsense knowledge and reasoning, but can be challenging for machines. Uncovering implicit knowledge has the potential to enable many important tasks, which we focus on later on. In this work, we propose a dataset that contains manually labeled implicit knowledge present in dyadic dialogues that require commonsense knowledge to infer. We use this dataset to evaluate pre-trained language models' ability for commonsense-based implicit knowledge inference.

The extracted triplets, of the form $(h, r, t)$ or alternatively $h \xrightarrow{r} t$, consist of a head ($h$) and a tail ($t$) span and the directed relation ($r$) between them. These spans are representative of some events, actions, objects, entities, and so on. The relation $r$ is directed and comes from a predefined set of relations $\mathcal{R}$ that describe the relationship between the head and tail spans within the context of the conversation — illustrated in Fig. 1 with the arrows between spans. Notably, the relation set $\mathcal{R}$ is intended to be generic in nature, rather than specifically factual or taxonomic, so as to accommodate wide categories of knowledge (§3.2) inferred from the context of the conversation.

### 3.1 Types of Triplets

The extracted triplets are either explicit or implicit as defined below:

*Explicit triplets* represent knowledge (see Fig. 2a) that is overtly expressed in an utterance in a dialogue. Fig. 1 illustrates one such annotated instance in utterance 13 — *tardiness* $\xrightarrow{Causes}$ *embarrassment* — where the triplet is worded verbatim in a head-relation-tail sequence. The head and tail span may contain some pronouns that can




(a) (b) (c)

Figure 2: (a) Explicit, (b) implicit knowledge extraction from dialogues. (c) Intermediate latent spans and triplets.

be decoded by simple co-reference resolution. In the presence of complex co-reference however the context suggests many possible candidates, and the triplet is implicit.

*Implicit triplets,* on the other hand, are not directly expressed in the dialogue and must be inferable through commonsense reasoning using the contextual information present in the dialogue. Instances of such triplets are shown in Fig. 1 and 2b with the relations in red font.

**Why Focus on Implicit Triplets?** As pointed out earlier, extracting explicit knowledge from a conversation or any natural language text is relatively straightforward and has been studied in much detail in the literature (Auer et al., 2007; Carlson et al., 2010; Speer et al., 2017). The much more challenging problem, however, is to extract or distill implicit knowledge. For example, in Fig. 1 the triplet *miss several buses* $\xrightarrow{Causes}$ *over 1 hour late* requires commonsense reasoning and knowledge about the world. Similarly, another triplet *lost my wallet* $\xrightarrow{Causes}$ *late* requires multi-utterance reasoning with contextual understanding to be extracted. Such distillation is not covered by the explicit-knowledge extraction framework.

The extraction of such implicit knowledge also requires contextual understanding and complex commonsense reasoning involving multiple steps and utterances. Thus, the extraction of implicit knowledge is challenging and a focus of this work.

**Latent Spans and Differences with GLUCOSE (Mostafazadeh et al., 2020):** As argued earlier, annotating implicit triplets often requires multi-step reasoning. In such cases, one or more intermediate spans (which may not be present in the dialogue) may be required to explain the relation between the constituting spans. Readers are urged to check Fig. 2c for one such example. Annotators were given the freedom to identify such intermediate steps when they deemed so. However, such cases are infrequent in our dataset and, thus, we have chosen to omit the intermediate spans in our experimental studies for the sake of simplicity. We leave the intermediate step modelling as a direction for future work.

In this context, it is also important to highlight the fundamental differences between our dataset and GLUCOSE (Mostafazadeh et al., 2020): (1) In our dataset, the knowledge represented by the spans and the relation connecting them is true (valid) given the context, but establishing this connection using an explicit relation requires complex commonsense inference and understanding of the discourse. The resulting triplet is thus valid in the context and grounded by the context. This is similar to the deductive commonsense reasoning as defined by (Davis, 2014). GLUCOSE however focuses on abductive commonsense inference, where given an event/state and its context, the annotators provided inferred speculative causal explanations of the event (state) according to *their* world and commonsense knowledge. These explanations, although they may fit in the given context, may not always be entailed by it. As a consequence, GLUCOSE is conducive to generative modeling, whereas our dataset leads to extractive modeling. (2) GLUCOSE has a limited set of relations, where inference is only performed across the following dimensions: *cause*, *enable*, and *result in*. In contrast, we have a much more diverse set of relations, which we describe in §3.2. (3) We construct our dataset based on conversations between two humans, while GLUCOSE is built using monologue-like stories that have significant differences with respect to the discourse structure and semantics of dialogues.

### 3.2 Types of Relations

Our proposed DAIKE dataset contains 25 main and 6 negated relations. Among the main 25 relations, 19 have been adopted from ConceptNet (Speer et al., 2017). We introduce 6 new relations to cover some aspects that are not covered by ConceptNet. Brief explanations, examples, and the new relations we introduce are shown in Table 1.
We categorize the different relations as follows:

| Category | Relation | Explanation | Example |
|---|---|---|---|
| Attribution | Capable Of | Something that A can typically do is B. | knife → cut |
| | Depends On* | A depends on B. | postage fee → weight of the post |
| | Has A | B belongs to A, either as an inherent part or due to a social construct of possession. | bird → wing; pen → ink; gearshift → car |
| | Has Property | A has B as a property; A can be described as B. | ice → cold |
| | Has Subevent | A and B are events, and B happens as a subevent of A. | eating → chewing |
| | Is A | A is a subtype or a specific instance of B; every A is a B. | car → vehicle; Chicago → city |
| | Manner Of | A is a specific way to do B. Similar to "Is A", but for verbs. | auction → sale |
| Causal | Causes | A causes B to happen. | exercise → sweat |
| | Causes Desire | A makes someone want B. | having no food → buy food |
| | Implies* | A implies B. | wet cloth → caught in rain |
| Comparison | Antonym | A and B are opposites in some relevant way, such as being opposite ends of a scale, or fundamentally similar things with a key difference between them. Counter-intuitively, two concepts must be quite similar before people consider them antonyms. | black ↔ white; hot ↔ cold |
| | Distinct From | A and B are distinct member of a set; something that is A is not B. | red ↔ blue; August ↔ September |
| | Similar To | A is similar to B. | mixer ↔ food processor |
| | Synonym | A and B have very similar meanings. They may be translations of each other in different languages. | sunlight ↔ sunshine |
| Conditional | Has Prerequisite | In order for A to happen, B needs to happen; B is a dependency of A. | dream → sleep |
| Intentional | Desires | A is a conscious entity that typically wants B. Many assertions of this type use the appropriate language's word for "person" as A. | person → love |
| | Motivated By Goal | Someone does A because they want result B; A is a step toward accomplishing the goal B. | compete → win |
| | Obstructed By | A is a goal that can be prevented by B; B is an obstacle in the way of A. | sleep → noise |
| | Used For | A is used for B; the purpose of A is B. | bridge → cross water |
| Social | Social Rule* | A is the social norm for when B happens or during B. | apology → late |
| Spatial | At Location | A happens at location B, or B is a typical location for A. | try clothes → changing room |
| | Located Near | A and B are typically found near each other. | table → chairs |
| Temporal | Before* | A starts/ends before B. | brush teeth → go to bed |
| | Happens On* | A happens during B. | celebration → birthday |
| | Simultaneous* | A and B happens at the same time. | heavy sports → heavy breath |

Table 1: Annotated relations in our dataset. * indicates new relations introduced by us that are not present in ConceptNet. ↔ in the examples indicate symmetric relations. In addition to the above, we also have a few negation relations as illustrated in §3.3.

**Attribution.** Relations that indicate attributes, properties, and definitions of concepts: (i) *Capable Of*, (ii) *Depends On*, (iii) *Has A*, (iv) *Has Property*, (v) *Has Subevent*, (vi) *Is A*, and (vii) *Manner Of*.

**Causal.** Relations that indicate cause and effect of events: (i) *Causes*, (ii) *Causes Desire*, and (iii) *Implies*.

**Comparison.** Relations that indicate comparison, similarity, or dissimilarity between concepts: (i) *Antonym*, (ii) *Distinct From*, (iii) *Similar To*, and (iv) *Synonym*.

**Conditional.** This category, having only one relation *Has Prerequisite*, indicates dependency of one event on the other.

**Intentional.** Relations that indicate the intent or usage of an entity or a person: (i) *Desires*, (ii) *Motivated By Goal*, (iii) *Obstructed By*, and (iv) *Used For*.

**Social.** The category involves social commonsense relations specifying social rules, conventions, norms, and suggestions. The relation in this category is: (i) *Social Rule*.

**Spatial.** This category encompasses relations which signifies spatial properties, such as location of events, entities, actions. The relations include: (i) *At Location*, and (ii) *Located Near*.

**Temporal.** This category involves the idea of time considering the start, end, duration, and order of events. The constituent relations are: (i) *Before*, (ii) *Happens On*, and (iii) *Simultaneous.*

### 3.3 Negative and Symmetric Relations

Apart from the relations in Table 1, the negations of some of these relations are necessary to form the triplets during annotation. These negated relations are (i) *Not Causes*, (ii) *Not Causes Desire*, (iii) *Not Has Property*, (iv) *Not Implies*, (v) *Not Is A*, and (vi) *Not Motivated By Goal*.

It should be noted that there are some symmetric relations[1] in our relation set. The set of symmetric relations $\mathcal{R}^{\mathcal{S}}$ contains (i) *Antonym*, (ii) *Distinct From*, (iii) *Similar To*, (iv) *Synonym*, (v) *Located Near*, and (vi) *Simultaneous*.

## 4 Dataset Construction

### 4.1 Source Datasets of Dialogues

The annotation is performed on the following datasets containing dyadic dialogues:

[1] A relation $R$ is considered symmetric if the validity of $A \xrightarrow{R} B$ implies the validity of $B \xrightarrow{R} A$ and vice versa.

**DailyDialog** (Li et al., 2017) is aimed towards emotion and dialogue-act classification at utterance level. The conversations cover various topics ranging from ordinary life, work, and relationships, to tourism, finance and politics.

**MuTual** (Cui et al., 2020) is a manually annotated dataset for multi-turn dialogue reasoning. It was introduced to evaluate several aspects of dialogue-level reasoning in terms of next utterance prediction given a dialogue history. These aspects include attitude reasoning, intent prediction, situation reasoning, multi-fact reasoning, and others.

**DREAM** (Sun et al., 2019) is a dialogue-based multiple-choice reading-comprehension dataset collected from English as foreign language exams. This dataset presents several challenges as it contains non-extractive answers that require commonsense reasoning beyond a single sentence.

In total, we sampled 807 dialogues from the three datasets. Each sampled dialogue has 5 to 12 utterances, and each constituent utterance has no more than 30 words.

### 4.2 Annotation Process

**Annotation guidelines.** The annotators are instructed to identify either explicit and implicit knowledge in a dialogue and represent them in terms of triplets (§3.1). Such a triplet consists of a pair of spans, say $A$ and $B$, and an appropriate relation $R$ between them, denoted as $A \xrightarrow{R} B$. A *span* is defined as a word, phrase, or a sub-sentence unit of an utterance that represents some concept such as an entity, event, action. The annotators are instructed to meet the following constraints during the annotation:

- The extracted triplets must be entailed by the conversation to be valid.
- The spans of a triplet should be as short and concise as possible. Also, a triplet may connect a pair of spans from distinct utterances in a dialogue.
- Multiple distinct valid relations between the same pair of spans are allowed. All these relations correspond to distinct triplets.

We used a web-based tool called BRAT (Stenetorp et al., 2012) for the annotation. The annotators are three PhD students who have thorough knowledge about the task. They were first briefed about the annotation rules, followed by a trial with a few samples to evaluate their understanding of the annotation guidelines and ability to extract both explicit and implicit triplets. Although annotators extract both types, they were instructed to focus more on annotating implicit triplets since extracting those are more challenging. The trial stage was conducted to ensure that annotators are well-versed in annotating high quality triplets in the final phase.

### 4.3 Annotation Verification and Agreement

Each dialogue is primarily annotated by a single annotator. We then verify the validity of the annotated triplets using the following strategy:

1. All extracted triplets are independently validated by two other validation annotators, in terms of their inferability from their source dialogues.
2. Unanimously agreed-upon valid triplets are kept, while unanimously agreed-upon invalid triplets are discarded. In the case of a disagreement, we bring in a third annotator to break the tie.
3. The final set of valid triplets is labelled as being explicit or implicit by the same two annotators as step (1) . The majority vote is assigned as the final label. Similar to the previous step, in case of a disagreement, we bring in a third annotator to break the tie.

After this stage, we obtained a Cohen's Kappa inter-validation-annotator agreement of 0.91 for triplet verification and 0.93 for knowledge type labelling. We found that the number of explicit triplets in the final annotated dataset is significantly lesser than implicit triplets since the informal nature of the source datasets' conversations enables the extraction of much more frequent implicit triplets than explicit ones. Statistics of the annotated dataset are shown in Table 2.

## 5 Experimental Setup and Results

We formulate three distinct tasks from DAIKE dataset: 1) Dialogue-level Natural Language Inference, 2) Span Extraction, and 3) Multi-choice Span Selection.

### 5.1 Dialogue-level Cross Validation

We consider a dialogue-level cross-validation strategy to benchmark our models. We partition the annotated dialogues into five disjoint and roughly equal-sized folds. Per cross-validation round, the triplets from four folds are considered for training, and the remaining one fold is used for test.

| Description | Instances |
|---|---|
| # Dialogues/# triplets in DailyDialog | 245/1286 |
| # Dialogues/# triplets in MuTual | 182/658 |
| # Dialogues/# triplets in DREAM | 380/2595 |
| # Dialogues/# triplets Total | 807/4539 |
| # Dialogues with # triplets $< 3$ | 142 |
| # Dialogues with # triplets between 3-5 | 312 |
| # Dialogues with # triplets between 5-10 | 281 |
| # Dialogues with # triplets $> 10$ | 72 |
| Average # triplets per dialogue | 5.62 |
| # Explicit triplets | 204 |
| # Implicit triplets | 4335 |
| # Triplets with spans from Utt. distance = 0 | 1009 |
| # Triplets with spans from Utt. distance = 1 | 1490 |
| # Triplets with spans from Utt. distance between 2-5 | 1501 |
| # Triplets with spans from Utt. distance between 6-8 | 401 |
| # Triplets with spans from Utt. distance $> 8$ | 138 |
| # Triplets having spans from same speaker | 2475 |
| # Triplets having spans from different speakers | 2064 |
| # Span pairs with single relation | 4203 |
| # Span pairs with multiple relations | 164 |

Table 2: Statistics on our dataset DAIKE. Please refer to the appendix for frequency statistics of the relations.

### 5.2 Task 1: Dialogue-level Natural Language Inference (DNLI)

Natural language inference (NLI) is the task of identifying if a "hypothesis" is true (entailment), false (contradiction), or undetermined (independent) given a "premise". We extend this definition to conversations and propose *Dialogue-level Natural Language Inference* (DNLI), which is the task of determining whether a knowledge triplet (hypothesis) is true or false given a dialogue (premise).

It should be noted that most NLI datasets such as SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2017), SciTail (Khot et al., 2018) consist of a single sentence hypothesis and premise, whereas for DNLI the hypothesis and the premise are a triplet and a conversation, respectively.

For our experiments, the *hypothesis* is formed by concatenating the elements of the triplet $h \xrightarrow{r} t$ in $h, r, t$ order. Similarly, the *premise* is formed by concatenating the utterances of the dialogue.

#### 5.2.1 Creating Negative Examples

Let $C$ be a conversation, $T$ be the set of all valid triplets in $C$, and $A \xrightarrow{R} B$ be one such valid triplet in $T$. We denote $\boldsymbol{\mathcal{R}}$: set of all relations; $\boldsymbol{\mathcal{R}^S}$: set of symmetric relations. The samples with valid triplets as hypotheses are termed as positive examples. The contradicting triplets/hypotheses for the negative samples are created from $T$ as follows:

**Reverse Relation Direction.** In $A \xrightarrow{R} B$, if $R \notin \boldsymbol{\mathcal{R}^S}$, then $B \xrightarrow{R} A$ is a contradicting hypothesis.

**Substitute Relation Type.** For $A \xrightarrow{R} B$, another relation $Q$ is randomly sampled from $\boldsymbol{\mathcal{R}} \setminus \{R\}$ and $A \xrightarrow{Q} B$ is considered a contradicting hypothesis.

**Substitute Span.** For $A \xrightarrow{R} B$, either $A$ or $B$ is replaced with another random span $X$ from the other triplets in set $T$. $X \xrightarrow{R} B$ or $A \xrightarrow{R} X$ is then considered a contradicting hypothesis.

**Combination of All.** A combination of the above three strategies can also be used to create the contradicting hypothesis. We ensure that the contrived contradicting hypotheses do not appear in the set of annotated triplets $T$.

The above strategies allow us to create multiple negative samples from a positive sample. In our experiments, we had two and eight negative samples per positive sample in the training and test split, respectively. We intentionally keep fewer negative samples in the training data to evaluate the generalization capacity of the models on a more diverse range of negative samples in the test data. Fold-wise statistics are shown in Table 3.

| Split | Label | Fold1 | Fold2 | Fold3 | Fold4 | Fold5 |
|---|---|---|---|---|---|---|
| Train | Positive | 3627 | 3630 | 3631 | 3630 | 3638 |
| | Negative | 6441 | 6469 | 6527 | 6492 | 6470 |
| Test | Positive | 912 | 909 | 908 | 909 | 901 |
| | Negative | 7425 | 6876 | 6989 | 7061 | 7187 |

Table 3: Cross validation fold statistics for Task 1: DNLI.

#### 5.2.2 Baseline

**RoBERTa-large Fine-tuned on MNLI.** We use the pretrained `roberta-large-mnli` model (Liu et al., 2019) to benchmark this task. The input to the model is: `<CLS> Premise <SEP> Hypothesis <SEP>`. The classification is performed on the `<CLS>` token vector from the final layer. We choose this model as it has been fine-tuned on the MNLI dataset and shows impressive performance on a number of NLI tasks.

#### 5.2.3 Results

The performance of the RoBERTa-MNLI model is reported in Table 4. As DNLI is a classification task, we report macro F1, weighted F1, and precision and recall over the positive examples (with valid triplets). We notice that the metrics are quite consistent across the five different folds and thus we report our conclusion against the average score. We obtained an average weighted F1 score of **85.78%**. However, the macro F1 score is noticeably lower at **69.83%**, suggesting that the model performs poorly on the less-frequent positive examples. The recall score suggests that **76.85%** of the valid hypotheses are correctly identified by the

| Metric | Fold1 | Fold2 | Fold3 | Fold4 | Fold5 | Avg. |
|---|---|---|---|---|---|---|
| Macro F1 | 69.15 | 71.07 | 68.14 | 71.29 | 69.49 | 69.83 |
| Weighted F1 | 86.76 | 85.48 | 84.07 | 86.42 | 86.17 | 85.78 |
| Precision Positive | 35.79 | 39.18 | 34.87 | 39.37 | 37.05 | 37.25 |
| Recall Positive | 77.55 | 78.54 | 77.56 | 78.16 | 72.45 | 76.85 |

Table 4: Results for the RoBERTa-MNLI model in Task 1: Dialogue-level Natural Language Inference (DNLI).

model. However, the precision score is quite low at **37.25%**, suggesting that almost $2/3$-rd of the predicted valid hypothesis are in-fact invalid. Without fine-tuning, the model produces much lower macro F1 of **17.76%**, precision of **15.06%**, and recall of **47.4%**. The state-of-the-art RoBERTa MNLI model is thus not very capable of correctly identifying triplets entailed by the conversation. We conclude that knowledge inference from conversational context is not straightforward for pretrained language models.

### 5.3 Task 2: Span Extraction

*Span Extraction* is defined as identifying the tail span $B$, given the head span $A$, the relation $R$ between $A$ and $B$, and the conversation $C$ where $A \xrightarrow{R} B$ is encoded. It is analogous to the task of node prediction in knowledge bases, where the missing tail node $B$ in $A \xrightarrow{R}?$ is to be predicted.

In this paper, *Span Extraction* is formulated as a Machine Reading Comprehension (MRC) task similar to SQuAD (Rajpurkar et al., 2016) where a question is to be answered from a given passage of text or more generally context. The equivalencies with MRC are defined as follows:

**Context.** The entire conversation $C$ is treated as the context, as the span $B$ in the triplet $A \xrightarrow{R} B$ can come from any utterance of $C$.

**Question and Answer.** For each relation type $R$, we create a question template that includes a placeholder for span $A$ and asks for span $B$ as the answer. The templates are filled with the appropriate valid triplets to generate the question-answer pairs. Please refer to the question template in the supplementary material.

#### 5.3.1 Baselines

We use two pretrained transformer-based models to benchmark the *Span Extraction* task. The methodology described in BERT QA models (Devlin et al., 2019) is used to extract the tail-spans/answers.

**RoBERTa Base.** We use the `roberta-base` model (Liu et al., 2019) as a baseline model.

**SpanBERT Fine-tuned on SQuAD.** We use SpanBERT (Joshi et al., 2020) fine-tuned on SQuAD 2.0 dataset as the other baseline model.

#### 5.3.2 Evaluation Metrics

**EM (Exact Match).** % of the predicted answers that are identical to the gold answers. **NM (No Match).** % of the predicted answers that bear no match with the gold answer. **F1:** The F1 score introduced by Rajpurkar et al. (2016) to evaluate word-level overlap of predictions with the gold answers for extractive QA models.

#### 5.3.3 Results

| Model | Metric | Fold1 | Fold2 | Fold3 | Fold4 | Fold5 | Avg. |
|---|---|---|---|---|---|---|---|
| SpanBERT | EM | 29.2 | 28.35 | 26.57 | 31.54 | 26.37 | 28.41 |
| | NM | 46.47 | 48.71 | 52.91 | 47.48 | 50.0 | 49.11 |
| | F1 | 43.72 | 42.27 | 39.31 | 44.22 | 40.77 | 42.06 |
| RoBERTa | EM | 15.87 | 13.18 | 12.1 | 15.12 | 13.48 | 13.95 |
| | NM | 57.36 | 56.71 | 61.57 | 53.22 | 57.4 | 57.25 |
| | F1 | 31.31 | 30.83 | 28.93 | 34.38 | 31.86 | 31.46 |

Table 5: Results for Span Extraction task. Higher EM, F1, and lower NM scores are better.

The results for this task is reported in Table 5. We notice that the SpanBERT model performs significantly better than the RoBERTa model. This is expected as SpanBERT has been pretrained with a different objective function and it particularly excels at span extraction tasks, such as, question answering. However, the EM score of **28.41%** and the F1 score of **42.06%** for the superior SpanBERT model is still subpar. The EM score suggests that the model extracts the exact correct answer less than $1/3$-rd of the time. The NM score also indicates that the extracted answer and the actual answer have no overlap around half of the time. Without fine-tuning, the SpanBERT model produces an EM score of **7.96%** and a F1 score of **20.78%**, much lesser than the fine-tuned model. We conclude that the state-of-the-art pretrained language models struggle with extracting missing spans.

### 5.4 Task 3: Multi-choice Span Selection

*Multi-choice Span Selection* is motivated by the SWAG commonsense inference task (Zellers et al., 2018). In SWAG, given a partial description of a situation, the appropriate ending is to be selected from a given list of choices using commonsense inference. In our case, *Multi-choice Span Selection* is formulated as a multiple-choice question answering task. Similar to the previous task, given a conversation $C$ and partial information about a triplet $A \xrightarrow{R}?$, the goal is to predict the missing span $B$ as an answer to a question created from $A$ and $R$. However, in contrast to task 2, the missing span $B$ has to be selected from a list of four

possible answers $S = \{s_1, ..., s_4\}$. The context, question, and answers are created as follows:

**Context and Question:** Both the context and the question construction follow §5.3.

**Correct and Confounding Options:** The options include the target answer and the three confounding options that are extracted from the same context .

### 5.4.1 Creating Confounding Options

To mitigate the stylistic artifacts that could give away the target answer (Gururangan et al., 2018; Poliak et al., 2018), the confounding options are generated in an adversarial fashion.

**Generating Confounding-option Candidates.** We first select a large number of spans from $C$ to form a confounding-option collection $\mathcal{N}$ by leveraging the SpanBERT fine-tuned on the samples of Task 2 (§5.3). We feed each individual utterance as the context, and the question created from $A$ and $R$ to the Task-2 fine-tuned SpanBERT. This leads to one or two candidate answers (spans) per contextual utterance per question, averaging around 30 confounding spans per question. We discard the spans that form a valid triplet with $A$ and $R$.

**Adversarial Filtering.** Once we have the collection $\mathcal{N}$, we follow Zellers et al. (2018) to filter the confounding options generated in §5.4.1.

Check Appendix Section A for details. We use `roberta-base` model to filter out stylistic patterns. During the filtering process, discriminator prediction accuracy decreased from 0.55 to 0.27, suggesting the method's effectiveness in removing easy confounding candidates with stylistic patterns.

### 5.4.2 Baseline

We experiment with `bert-base-uncased` and `roberta-base` on the adversarially created dataset. The input to the models is the concatenation of conversation $C$, question $Q$, and candidate answers $A_j, j \in \{1, ..., 4\}$: `<CLS> C <SEP> Q <SEP> A_j <SEP>`. Each score is predicted from the corresponding `<CLS>` token vector and the highest scoring one is selected as answer.

### 5.4.3 Results

The results reported in Table 6 indicate the importance of contextual information in improving models' performance. Our human verifiers could also predict the answers significantly more accurately when contextual information was available. It is worth noting that all the pre-trained language models perform poorly in this task and the obtained results are far from reaching the human-level performance. Besides, the accuracy score for `bert-base-uncased` and `roberta-base` without fine-tuning are 25.60% and 26.22% respectively which is similar to a random baseline (25.00%), confirming the conclusion in Task 2 (§5.3) that current language models have difficulties in predicting the missing span.

| Model | Setting | Fold1 | Fold2 | Fold3 | Fold4 | Fold5 | Avg. |
|---|---|---|---|---|---|---|---|
| BERT | C&Q | 60.35 | 58.96 | 51.84 | 61.62 | 60.55 | 58.66 |
| | Q | 47.21 | 50.89 | 51.25 | 54.46 | 47.84 | 50.33 |
| RoBERTa | C&Q | 61.16 | 51.05 | 65.28 | 73.31 | 62.04 | 62.57 |
| | Q | 51.05 | 62.04 | 56.60 | 58.92 | 55.76 | 56.87 |
| Human | C&Q | 89.90 | 82.69 | 83.02 | 80.77 | 80.78 | 83.43 |
| | Q | 69.39 | 67.31 | 60.00 | 65.38 | 71.15 | 66.45 |

Table 6: Results for Multi-choice Span Selection task. C&Q → model input is the Context and the Question; Q → model input is only the Question.

**Performance across Relation Categories.** We report the results across different relation categories for each task with the corresponding best performing models in Table 7. We notice that *Spatial* is one of the top-performing categories across all three tasks. Performance in *Attribution* and *Temporal* category are also reasonably well in Task 1 and Task 1, 2 respectively. Interestingly, the result of *Temporal* category in Task 3 is the worst. The performance in *Causal* and *Conditional* category is around the average mark across all three tasks. This implies that pretrained language models find it difficult to understand the concept of causal events or dependent events. Finally, we observe that the performance in *Social* category is the worst or among the worst for all the tasks, suggesting that the models find it very challenging to reason about social norms, rules, and conventions.

| Task | Attribution | Causal | Comparison | Conditional | Intentional | Social | Spatial | Temporal |
|---|---|---|---|---|---|---|---|---|
| 1 | 74.97 | 67.26 | 68.75 | 68.51 | 70.49 | 58.97 | 79.06 | 71.56 |
| 2 | 43.34 | 38.04 | 36.78 | 38.97 | 46.70 | 28.34 | 57.41 | 54.26 |
| 3 | 64.64 | 61.20 | 58.76 | 55.72 | 63.34 | 58.00 | 71.20 | 54.53 |

Table 7: Average five-fold Macro-F1, F1, and Accuracy score over the relation categories. We report results for RoBERTa-MNLI, SpanBERT and RoBERTa models for the three tasks.

## 6 Conclusion

In this work, we introduced a new dataset DAIKE that primarily focuses on commonsense-based implicit knowledge extraction from dialogues. The dataset consists of more than 4,500 manually annotated knowledge triplets from over 800 dialogues. We also introduced dialogue-level NLI and QA tasks, along with baselines to evaluate the inference and reasoning capabilities of pretrained transformer-based models.

## References


Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam Hruschka, and Tom Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ernest Davis. 2014. *Representations of commonsense knowledge*. Morgan Kaufmann.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Herbert P. Grice. 1975. Logic and conversation. *Speech acts*, pages 41–58.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S Weld. 2019. Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5807–5812.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.