$\mathcal{C}om\mathcal{F}act$: A Benchmark for Linking Contextual Commonsense Knowledge

Silin Gao[1], Jena D. Hwang[2], Saya Kanno[3], Hiromi Wakaki[3], Yuki Mitsufuji[3], Antoine Bosselut[1]

[1]NLP Lab, IC, EPFL, Switzerland [2]Allen Institute for AI, WA, USA
[3]Sony Group Corporation, Tokyo, Japan
{silin.gao,antoine.bosselut}@epfl.ch, jenah@allenai.org, {saya.kanno,hiromi.wakaki,yuhki.mitsufuji}@sony.com

Abstract

Understanding rich narratives, such as dialogues and stories, often requires natural language processing systems to access relevant knowledge from commonsense knowledge graphs. However, these systems typically retrieve facts from KGs using simple heuristics that disregard the complex challenges of identifying situationally-relevant commonsense knowledge (*e.g.*, contextualization, implicitness, ambiguity). In this work, we propose the new task of commonsense fact linking, where models are given contexts and trained to identify situationally-relevant commonsense knowledge from KGs. Our novel benchmark, $\mathcal{C}om\mathcal{F}act$, contains ∼293k in-context relevance annotations for commonsense triplets across four stylistically diverse dialogue and storytelling datasets. Experimental results confirm that heuristic fact linking approaches are imprecise knowledge extractors. Learned fact linking models demonstrate across-the-board performance improvements (∼34.6% F1) over these heuristics. Furthermore, improved knowledge retrieval yielded average downstream improvements of 9.8% for a dialogue response generation task. However, fact linking models still significantly underperform humans, suggesting our benchmark is a promising testbed for research in commonsense augmentation of NLP systems.

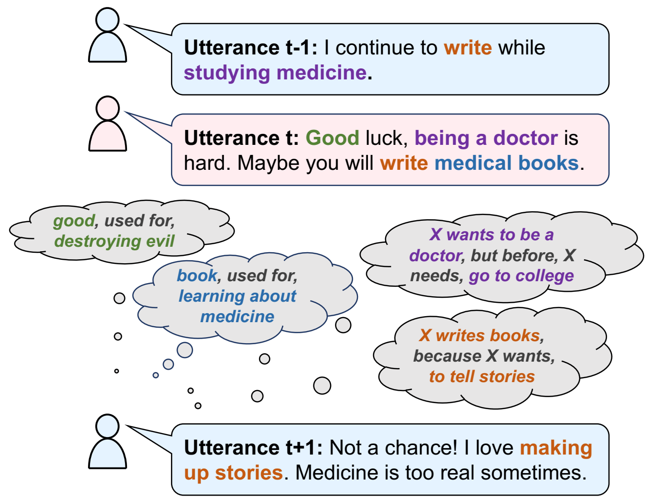

Figure 1: Commonsense fact linking in a conversation. Triples in bubbles represent linked facts. Words and phrases in green, blue, purple and orange illustrate four different linking relationships for facts.

Introduction

In conversations, stories, and other varieties of narratives, language users systematically elide information that readers (or listeners) reliably fill in with world knowledge. For example, in Figure 1, the speaker of utterance t (*i.e.*, pink) infers that their counterpart (cyan) wants to be a doctor because they are studying medicine, even though the cyan speaker does not explicitly mention their career goals. To reflect this ability, language understanding systems are often augmented with knowledge bases (KBs, *e.g.*, Speer, Chin, and Havasi 2017) that allow them to access relevant background knowledge.

Considerable research has examined how to construct large databases of world knowledge for this purpose (Lenat 1995; Suchanek, Kasneci, and Weikum 2007; Speer, Chin, and Havasi 2017), as well as how to design models that can reason over relevant subsets of this knowledge to form a richer understanding of language (*e.g.*, Lin et al. 2019). However, less work examines how to retrieve these inferences (or facts) from the KB in the first place. Current methods typically rely on pattern-based heuristics (Mihaylov and Frank 2018; Feng et al. 2020), unsupervised scoring using corpus statistics (Weissenborn, Kovcisk'y, and Dyer 2018) or neural re-rankers (Yasunaga et al. 2021), or combinations of these methods (Bauer, Wang, and Bansal 2018).

These simple methods produce computationally tractable knowledge representations, but frequently retrieve noisy information that is irrelevant to the narrative they are constructed to represent. Recent work demonstrates that models trained with heuristically-retrieved commonsense knowledge learn simplified reasoning patterns (Wang et al. 2021) and provide false notions of interpretability (Raman et al. 2021). We posit that inadequate retrieval from large-scale knowledge resources is a key contributor to the spurious rea-

Accepted to Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI 2023.
We release our data and code to the community at `https://github.com/Silin159/ComFact`

soning abilities learned by these systems.

Acknowledging the importance of retrieving relevant commonsense knowledge to augment models, we identify a set of challenges that commonsense knowledge retrievers must address. First, retrieved commonsense knowledge must be **contextually-relevant**, rather than generically related to the entities mentioned in the context. Second, relevant commonsense knowledge can often be **implicit**, *e.g.*, in Figure 1, writing may be a leisure hobby for the cyan speaker, explaining why they "love making up stories". Finally, knowledge may be **ambiguously** relevant to a context. The cyan speaker in Figure 1 may write as a relaxing hobby, or be thinking of quitting medical school to pursue a career as a writer. Without knowing the rest of the conversation, both inferences are potentially valid.

To more adequately address these challenges, we introduce the new task of commonsense fact linking, where models are given contexts and trained to identify situationally-relevant commonsense knowledge from KGs. For this task, we construct a **Com**monsense **Fact** linking dataset ($\mathcal{C}om\mathcal{F}act$) to benchmark the next generation of models designed to improve commonsense fact retrieval. $\mathcal{C}om\mathcal{F}act$ contains ∼293k contextual relevance annotations for four diverse dialogue and storytelling corpora. Our empirical analysis shows that heuristic methods over-retrieve many unrelated facts, yielding poor performance on the benchmark. Meanwhile, models trained on our resource are much more precise extractors with an average 34.6% absolute F1 boost (though still fall short of human). The knowledge retriever developed on our resource also brings an average 9.8% relative improvement on a downstream dialogue response generation task. These results demonstrate that $\mathcal{C}om\mathcal{F}act$ is a promising testbed for developing improved fact linkers that benefit downstream NLP applications.

Related Work

Commonsense Knowledge Graphs Commonsense knowledge graphs (KGs) are standard tools for providing background knowledge to models for various NLP tasks such as question answering (Talmor et al. 2019; Sap et al. 2019b) and text generation (Lin et al. 2020). ConceptNet (Liu and Singh 2004; Speer, Chin, and Havasi 2017), a commonly used commonsense KG, contains high-precision facts collected from crowdsourcing (Singh et al. 2002) and web ontologies (Miller 1995; Lehmann et al. 2015), but is generally limited to taxonomic, lexical and physical relationships (Davis and Marcus 2015; Sap et al. 2019a). ATOMIC (Sap et al. 2019a) and ANION (Jiang et al. 2021) are fully crowdsourced, and focus on representing knowledge about social interactions and events. ATOMIC$^{20}_{20}$ (Hwang et al. 2021) expands on ATOMIC by annotating additional event-centered relations and integrating the facts from ConceptNet that are not easily represented by language models, yielding a rich resource of complex entities. In this work, we construct our $\mathcal{C}om\mathcal{F}act$ dataset based on the most advanced ATOMIC$^{20}_{20}$ KG.

Commonsense Fact Linking Knowledge-intensive NLP tasks are often tackled using commonsense KGs to augment the input contexts provided by the dataset (Wang et al. 2019; Ye et al. 2019; Gajbhiye, Moubayed, and Bradley 2021; Yin et al. 2022). Models for various NLP applications benefit from this fact linking, including question answering (Feng et al. 2020; Zhang et al. 2022), dialogue modeling (Zhou et al. 2018; Wu et al. 2020) and story generation (Guan, Wang, and Huang 2019; Ji et al. 2020). All above works typically conduct fact linking using heuristic solutions.

Recent research explores unsupervised learning approaches for improving on the shortcomings of heuristic commonsense fact linking. Huang, He, and Liu (2021) and Zhou et al. (2022) use soft matching based on embedding similarity to link commonsense facts with implicit semantic relatedness. Guan et al. (2020) use knowledge-enhanced pretraining to implicitly incorporate commonsense facts into narrative systems, but their approach reduces the controllability and interpretability of knowledge integration. Finally, several works (Arabshahi et al. 2021; Bosselut, Bras, and Choi 2021; Peng et al. 2021a,b; Tu et al. 2022) use knowledge models (Bosselut et al. 2019; Da et al. 2021; West et al. 2022) to generate commonsense facts instead of linking from knowledge graphs. However, the contextual quality of generated facts from knowledge models is also underexplored in these application scenarios. In this paper, we conduct more rigorous study on commonsense fact linking.

$\mathcal{C}om\mathcal{F}act$ Construction

In this section, we give an overview of commonsense fact linking and its associated challenges, and describe our approach for building the $\mathcal{C}om\mathcal{F}act$ dataset centered around these challenges.

Overview

Notation We are given narrative samples $\mathcal{S}$ (*e.g.*, a dialogue or story snippet) containing multiple statements (or utterances for dialogues) $[U_1, U_2, ..., U_T]$. For the t-th statement U_t, the collections of statements that comprise its past and future context are defined as $U_{<t} = [U_{t-k}, ..., U_{t-1}]$ and $U_{>t} = [U_{t+1}, ..., U_{t+l}]$, respectively.

A commonsense knowledge graph $\mathcal{G}$ is made up of a set of interconnected commonsense facts, each represented as a triple containing a head entity, a tail entity, and a relation connecting them, as depicted in Figure 1. The task in this work is to identify the subset of commonsense facts from $\mathcal{G}$ that may be relevant for understanding the situation described in the context $C_t = [U_{<t}, U_t, U_{>t}]$.

Challenges The task of commonsense fact linking poses several challenges:

- ***Contextualization***: many facts linked using simple heuristic methods, such as string-matching, are not actually relevant to the situation described in a context. For example, in Figure 1, the facts in bubbles are all

We follow prior naming convention for entity linking (Ling, Singh, and Weld 2015) and multilingual fact linking (Kolluru et al. 2021), though the task can also be viewed as information retrieval (IR) from a commonsense knowledge base.

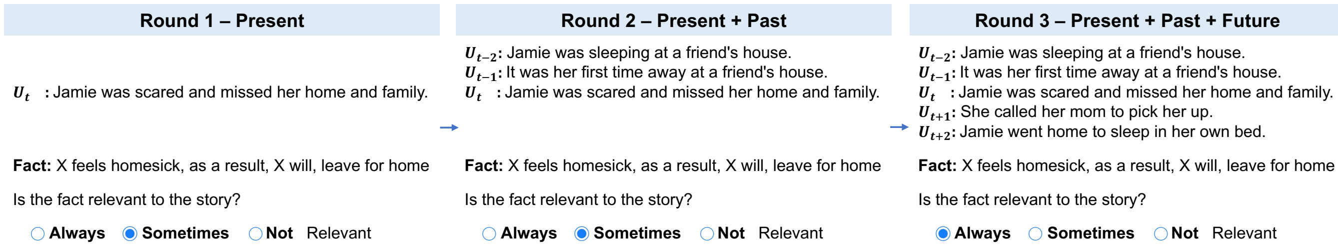

Figure 2: Illustration of our three-round fact candidate validation

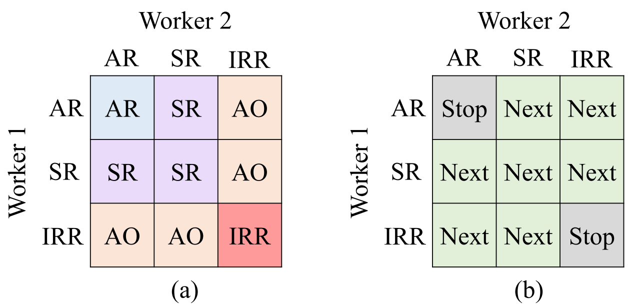

Figure 3: Summary of rules in fact candidate validation rounds. (a) Mapping from worker annotations to relevance labels: *always relevant* (AR), *sometimes relevant* (SR), *at odds* (AO) and *irrelevant* (IRR). (b) Mapping from worker annotations to action of the round: *evaluate in the next round* (Next) and *end validation* (Stop).

pattern-matched to the dialogue, but (*good*, *used for*, *destroying evil*) turns out to not be relevant to the situation when someone says *good luck*. Our study shows that only $\sim 25\%$ of facts linked through string matching end up being fully relevant to the context.

- ***Implicitness***: some facts are linked to the context in implicit ways. In Figure 1, the fact with *go to college* is implicitly linked to the phrase *studying medicine*, which makes it relevant to the context even though no direct reference to *college* is made in the dialogue, precluding it from being linked using string-matching.
- ***Ambiguity***: different observers can disagree on on whether a fact is relevant to reason about a situation, particularly if the future context of a narrative is unknown. For example, (*X writes books*, *because X wants*, *to tell stories*) in Figure 1 is relevant to the final produced utterance, but would not be if the final utterance had been about wanting to write scientific research papers instead (*n.b.*, the best use of writing skill).

While many methods have been proposed for linking facts in $\mathcal{G}$ to C_t, these methods typically rely on rule-based heuristics or unsupervised scoring methods, which do not adequately address the unique challenges of this task. In the following sections, we present our approach for building the $\mathcal{C}om\mathcal{F}act$ dataset that addresses the above challenges.

Fact Candidate Linking

Given $C_t = [U_{<t}, U_t, U_{>t}]$ from a natural language sample $\mathcal{S}$, we link an initial set of potentially relevant fact candidates from $\mathcal{G}$ using two approaches, one designed to extract explicit relevant facts and one for implicitly relevant facts.

Extracting Fact Candidates Similar to prior works (*e.g.*, Feng et al. 2020), we use surface-form pattern matching to retrieve head entities in $\mathcal{G}$ that are explicitly linked to U_t, and collect facts that contain the retrieved head entities as candidates. In particular, we lemmatize and part-of-speech (POS) tag U_t and every head entity in $\mathcal{G}$. Then, we match patterns between these sources that are words that are informative parts of speech (*e.g.*, nouns, verbs, adjectives, adverbs) or that correspond to n-grams in a master list of English idioms from Wiktionary. We retrieve head entities whose informative patterns all appear in the set of patterns from U_t.

However, pattern matching only extracts a set of fact candidates whose head entities can be explicitly recovered from the context U_t. To retrieve facts that may be semantically related to the context, but cannot be explicitly linked through patterns (*e.g.*, paraphrased facts), we use embedding similarity matching (Zhou et al. 2022). In particular, we use Sentence-BERT (Reimers and Gurevych 2019) to encode U_t along with every head entity in $\mathcal{G}$ as embedding vectors, and select the top-5 head entities whose embeddings have the highest cosine similarity with the embedding of U_t. Using this approach, we extend the sets of available candidates often retrieved by pattern matching methods and include implicit inferences in our candidate set.

Filtering Fact Candidates Head entities linked via pattern and embedding matching may connect to tail entities whose semantics are far different from that of C_t (*e.g.*, *destroying evil* in Figure 1). Consequently, we perform a first round of automatic filtering by pruning the tail entities of each head entity according to their similarity to C_t. Using Sentence-BERT, we encode each tail entity and C_t as embedding vectors. For each head entity, we keep its top-5 tail entities that have the highest embedding cosine similarity with that of C_t.

Crowdsourcing Relevance Judgements

We use the prior heuristics to over-sample a large initial set of knowledge (∼46 facts per example context), and then

https://en.wiktionary.org/w/index.php?title=Category:English_idioms

Method	PERSONA-ATOMIC					MUTUAL-ATOMIC					ROC-ATOMIC					MOVIE-ATOMIC				
	AR	SR	AO	IRR	κ	AR	SR	AO	IRR	κ	AR	SR	AO	IRR	κ	AR	SR	AO	IRR	κ
Explicit	8042 24%	772 2%	3731 11%	20940 63%	0.72	5352 32%	910 5%	4007 24%	6541 39%	0.52	8981 26%	862 2%	4449 13%	20320 59%	0.69	7352 17%	1883 5%	9578 22%	23957 56%	0.56
Implicit	2277 31%	224 3%	1076 15%	3813 51%	0.68	4206 35%	736 6%	2921 25%	4003 34%	0.49	6068 32%	653 3%	3177 17%	9234 48%	0.64	2582 22%	635 5%	2862 24%	5717 49%	0.55
Both	10319 25%	996 2%	4807 12%	24753 61%	0.71	9558 33%	1646 6%	6928 24%	10544 37%	0.51	15049 28%	1515 3%	7626 14%	29554 55%	0.67	9934 18%	2518 5%	12440 23%	29674 54%	0.56

Table 1: Relevance of fact candidates for different candidate extraction methods. **AR**: *always relevant*, **SR**: *sometimes relevant*, **AO**: *at odds*, **IRR**: *irrelevant*. κ denotes the Cohen's κ.

Relevance	PERSONA-ATOMIC				MUTUAL-ATOMIC				ROC-ATOMIC				MOVIE-ATOMIC			
	RPA	RPP	RPF	all	RPA	RPP	RPF	all	RPA	RPP	RPF	all	RPA	RPP	RPF	all
Always	8310 81%	1272 12%	738 7%	10320 100%	6678 70%	1728 18%	1152 12%	9558 100%	12048 80%	1562 10%	1439 10%	15049 100%	6495 65%	1681 17%	1758 18%	9934 100%
Sometimes	523 53%	130 13%	342 34%	995 100%	801 49%	316 19%	529 32%	1646 100%	734 48%	275 18%	506 34%	1515 100%	1132 45%	262 10%	1124 45%	2518 100%
Both	8833 78%	1402 12%	1080 10%	11315 100%	7479 67%	2044 18%	1681 15%	11204 100%	12782 77%	1837 11%	1945 12%	16564 100%	7627 61%	1943 16%	2882 23%	12452 100%

Table 2: Link type statistics of relevant facts on each data subset of $\mathcal{C}om\mathcal{F}act$.

evaluate the contextual relevance of these linked fact candidates using crowdworkers from Amazon Mechanical Turk.

We first task workers with validating the relevance of linked head entities with respect to the context. For each head entity, we show workers C_t and a head candidate associated with U_t, and ask them to judge whether the head candidate is relevant to U_t.

After curating a set of relevant head entities, workers then validate the relevance of the fact candidates associated with those head entities. To evaluate contextual relevance of facts in a fine-grained manner, we define a three-round task for workers shown in Figure 2.

In the first round, we show two workers U_t and the set of fact candidates, and independently ask them to judge whether the fact candidate is *always relevant*, *sometimes relevant*, or *irrelevant* to U_t. In the second round, we repeat this task, but show the past context along with U_t, namely $[U_{<t}, U_t]$. In the third round, we repeat the task again, but show the full context $C_t = [U_{<t}, U_t, U_{>t}]$.

After each round, we assign or update the relevance label of a fact candidate as: a) ***always relevant*** if both workers label it *always relevant*, b) ***sometimes relevant*** if one or both of the workers label it *sometimes* instead of *always relevant*, c) ***at odds*** if one worker chooses *always* or *sometimes relevant* and the other chooses *not relevant*, d) ***irrelevant*** if both workers select *not relevant* (as shown in Figure 3a). In practice, we find that including more context (*i.e.*, $U_{<t}$ or $U_{>t}$) rarely changes the validation of an initially *always relevant* or *irrelevant* fact. So after each round, if a fact candidate is labeled as *always relevant* or *irrelevant*, we do not evaluate it in the next round. Otherwise, there is relevance ambiguity over a fact, and we validate it again in the next round with additional context (as shown in Figure 3b). In the second and third rounds, if a worker annotates a fact candidate as *always* or *sometimes relevant*, we ask them to justify their selection by identifying which statement(s) in $U_{<t}$ or $U_{>t}$ make the fact relevant to the situation.

If a head entity is deemed ***irrelevant***, we assume that all fact candidates associated with it are *irrelevant* as well.

From feedback, we observe that crowdworkers prefer our fine-grained annotation scheme as it allows them to express uncertainty in the judgment compared to a binary choice.

Fine-grained Contextual Relevance

The three rounds of assessment allow us to perform a fine-grained annotation of the contextual relevance of a fact candidate. For each fact candidate, we map its relatedness to $C_t = [U_{<t}, U_t, U_{>t}]$ to one of the following four link types:

- **R**elevant to **P**resent **A**lone (**RPA**): the linked fact is directly relevant to U_t alone. For example, in Figure 1, the fact highlighted in blue, *medical books* are used for *learning about medicine*, is relevant to the concept of medical books mentioned in the utterance.
- **R**elevant to **P**resent given the **P**ast (**RPP**): the linked fact is not relevant to U_t alone, but relevant to U_t given $U_{<t}$. As shown in Figure 1, the purple fact helps interpret that *studying medicine* happens when someone *goes to college*, which is also a prerequisite of *being a doctor*.
- **R**elevant to **P**resent given the **F**uture (**RPF**): the linked fact is to U_t knowing $U_{<t}$ and $U_{>t}$. For example, the fact colored in orange in Figure 1 helps associate the action *Person X writes books* with the reason *making up stories*.
- **IRR**elevant to C_t (**IRR**): the linked fact is not relevant to the situation described in C_t. For example, as shown in Figure 1, the fact colored in green is irrelevant to help understand the context, although it is linked to *good luck* in surface form.

Context	Model	Setting	PERSONA-ATOMIC		MUTUAL-ATOMIC		ROC-ATOMIC		MOVIE-ATOMIC	
			Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
$U_{\leq t}$	Heuristic	none	0.211	0.348	0.290	0.450	0.231	0.375	0.205	0.340
	Head Linking		0.600	0.487	0.643	0.602	0.537	0.484	0.548	0.452
	LSTM	direct	0.805	0.471	0.749	0.573	0.761	0.457	0.769	0.417
	DistilBERT		0.840	0.626	0.792	0.649	0.811	0.608	0.800	0.512
	BERT (base)		0.859	0.659	0.801	0.668	0.841	0.669	0.818	0.547
	BERT (large)		0.859	0.660	0.819	0.689	0.848	0.690	0.835	0.595
	RoBERTa (base)		0.866	0.674	0.810	0.682	0.845	0.687	0.820	0.553
	RoBERTa (large)	direct	0.883	0.716	0.835	0.724	0.874	0.740	0.850	0.631
		pipeline	0.874	0.698	0.819	0.717	0.861	0.721	0.834	0.606
	DeBERTa (large)	direct	0.885	0.717	0.861	0.766	0.884	0.763	0.850	0.651
C_t	RoBERTa (large)	direct	0.882	0.721	0.838	0.740	0.879	0.748	0.851	0.635
		pipeline	0.874	0.693	0.825	0.722	0.867	0.731	0.830	0.603
	Human	none	**0.936**	**0.921**	**0.934**	**0.941**	**0.962**	**0.952**	**0.933**	**0.902**

Table 3: Fact linking results on the four data subsets of $\mathcal{C}om\mathcal{F}act$. We observe a substantial performance improvement by model-based fact linkers over heuristics typically used for fact linking. A large gap remains between the performance of best model-based fact linker (based on **DeBERTa**) and **Human** performance. Fact candidates labeled as *at odds* are not included.

If a fact candidate is finally labeled as *irrelevant*, we label its link type as **IRR**. If a fact candidate is finally labeled as *at odds*, we do not label its link type since its relevance is controversial. Otherwise, we further check the earliest assessment round where the fact's final relevance label comes out: we label the fact's link type as **RPA**, **RPP**, or **RPF** if its final relevance label first comes out at the first, second or third round, respectively.

Retaining Disagreements Facing the challenge of ambiguity in relevance (potentially due to inherent uncertainty in the facts being linked; Pavlick and Kwiatkowski 2019), we track disagreements between workers throughout our annotation pipeline, allowing us to measure the relevance controversy in commonsense fact linking. In particular, we record the disagreements of workers when: a) a head entity is ***relevant with half confidence***, b) a fact candidate is ***sometimes relevant***, c) a fact candidate is ***at odds*** in relevance. These rich annotations enable multiple modeling settings at different granularities for identifying relevant inferences, providing a rich set of potential label spaces for future work in granular fact linking.

$\mathcal{C}om\mathcal{F}act$ Analysis

We use ATOMIC$^{20}_{20}$ (Hwang et al. 2021) as the commonsense KG for building $\mathcal{C}om\mathcal{F}act$, which contains 1.33M complex facts covering physical objects, daily events and social interactions. ATOMIC$^{20}_{20}$ is a rich resource for building our dataset as it covers rich knowledge types, and is partially consolidated from other popular KGs including ConceptNet (Speer, Chin, and Havasi 2017)) and ATOMIC (Sap et al. 2019a), potentially offering better generalization for fact linking with these other resources.

We sample narrative contexts from four stylistically diverse English dialogue and storytelling datasets that involve elaborate contextual inference and understanding: PERSONA-CHAT (Zhang et al. 2018), MuTual (Cui et al. 2020), ROCStories (Mostafazadeh et al. 2016) and the CMU Movie Summary Corpus (Bamman, O'Connor, and Smith 2013). The context window size is set to 5 where $U_{<t} = [U_{t-2}, U_{t-1}]$ and $U_{>t} = [U_{t+1}, U_{t+2}]$. We denote the data portions collected from the four datasets as PERSONA-ATOMIC, MUTUAL-ATOMIC, ROC-ATOMIC and MOVIE-ATOMIC in $\mathcal{C}om\mathcal{F}act$.

Contextual Relevance Table 1 shows stratified statistics of the crowdsourced fact relevance annotations for the different candidate linking methods (*i.e.*, explicit pattern matching, implicit embedding matching). We observe that unsupervised fact linking methods, whether based on heuristics for explicit patterns or implicit matching mechanisms, often link irrelevant facts, introducing noise to any resulting extracted knowledge representation. Interestingly, once irrelevant head entities were removed, implicit fact candidates retrieved using embedding similarity were more likely to be judged relevant by human annotators, compared to pattern-matched fact candidates, showing the importance of generating a rich set of potentially relevant fact candidates.

To quantitatively measure the ambiguity of linked commonsense facts' relevance, we use Cohen's κ (Cohen 1960) to measure the agreement between workers that annotate the same facts. Most κ scores fall within the ranges that Cohen described as "moderate" (0.4 - 0.6) or "substantial" (0.6 - 0.8) agreement. We do observe that implicitly linked fact candidates have lower κ scores in their relevance validation, likely because they are linked to the context in a less straightforward way, causing more subjective relevance judgements.

For a fact candidate, we record judgements from the round where the fact's final relevance label first comes out.

See Appendix A for more data collection details.

Model	Acc.	Prec.	Recall	F1
Heuristic (explicit)	0.711	0.712	0.801	0.754
Heuristic (implicit)	0.289	0.291	0.199	0.236
Heuristic	0.553	0.553	**1.000***	0.712
RoBERTa (large)	**0.834**	**0.834**	0.874	**0.854**

Table 4: Head entity linking results on PERSONA-ATOMIC under the context window $U_{\leq t} = [U_{<t}, U_t]$. *Recall should be perfect here because the candidates for which we crowdsource relevance annotations are drawn from this heuristic.

Context	RPA	RPP	RPF	all
$U_{\leq t}$	0.735	0.634	0.561	0.698
C_t	0.749	0.653	0.651	0.720

Table 5: **Recall** of relevant facts by **RoBERTa (large)** on PERSONA-ATOMIC with respect to different link types and context windows, under the direct setting.

Link Types Table 2 shows statistics of the fine-grained link types (§) for *always* and *sometimes relevant* facts. We find that *always relevant* facts are mostly linked to the present statements alone (*i.e.*, **RPA**), while *sometimes relevant* facts are more often recognized with respect to larger context windows where the past and future statements are given (*i.e.*, **RPP**, **RPF**). Even though making up a relatively small total number, *sometimes relevant* facts may be critical inferences for imagining the future of the narrative, as they provide ambiguous hypotheses for where a narrative may be heading. In general, facts linked to the present alone occupy the largest proportion of relevant facts.

Experimental Methods

We evaluate our new benchmark using various baseline classification methods based on neural language models. All LMs are individually trained and evaluated on the four $\mathcal{C}om\mathcal{F}act$ datasets.

Approach Our models encode the concatenation of a narrative context with each of its fact candidates selected in our data collection. The output hidden states of the language models are then input to a binary classifier, which predicts whether the fact candidate is relevant to the context. We consider two context windows $U_{\leq t} = [U_{<t}, U_t]$ and $C_t = [U_{<t}, U_t, U_{>t}]$ in our experiments.

Models We use various pretrained language models as encoders for classifying facts, particularly the **BERT** (Devlin et al. 2019) and **RoBERTa** (Liu et al. 2019) model families. We test *base* and *large* sizes of these models, as well as more light-weight **DistilBERT** (Sanh et al. 2019) and more advanced **DeBERTa** (He et al. 2020). We also test the performance of a two-layer bi-directional **LSTM**. We evaluate each model on: a) **direct** prediction, where the model is trained to directly classify the fact candidates, and b)

See Appendix B for more data statistic results.

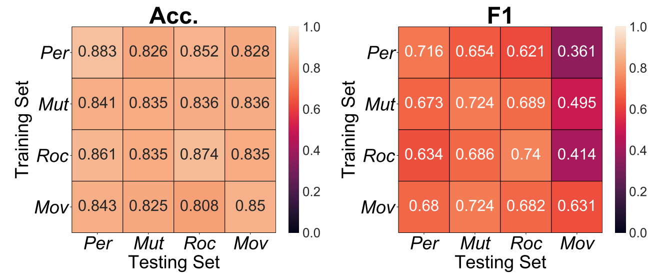

Figure 4: Fact linking results of **RoBERTa (large)** across the four data portions of $\mathcal{C}om\mathcal{F}act$: PERSONA-ATOMIC (*Per*), MUTUAL-ATOMIC (*Mut*), ROC-ATOMIC (*Roc*) and MOVIE-ATOMIC (*Mov*), under the direct setting and context window $U_{\leq t} = [U_{<t}, U_t]$.

pipeline prediction, where a first model is trained to classify head entities, and a second model classifies fact candidates associated with relevant head entities.

As a baseline, we report the performance of the same **Heuristic** used to generate candidates, which predicts all fact candidates retrieved in our data as relevant (*i.e.*, the typical linking approach for many methods). We also include a semi-heuristic baseline **Head Linking**, which finetunes RoBERTa (large) to only classify the head entities linked in our data, and predicts the relevance of each fact candidate as the relevance of its head. Finally, we run a **Human** study on a randomly sampled set of 200 contexts from each of the 4 test sets of $\mathcal{C}om\mathcal{F}act$, and ask crowdworkers to judge the relevance of 3 linked facts with respect to each context.

Experimental Results

We report the **Accuracy** and **F1** of fact candidate classification on $\mathcal{C}om\mathcal{F}act$ in Table 3. We find that trained fact linkers significantly outperform the **Heuristic** baselines, showing supervised neural classification on top of heuristically selected facts can significantly improve fact linking quality. Furthermore, the improvement over the **Head Linking** baseline demonstrates the importance of linking facts individually, rather than relying on coarse-grained entity linking. However, model-based fact linkers are still far from **Human** performance on $\mathcal{C}om\mathcal{F}act$, demonstrating that there is still considerable room for improvement. Interestingly, we find that directly predicting fact links outperforms a pipeline approach that first predicts relevant head entities and then only classifies fact candidates of relevant head entities.

Entity Linking Despite the error propagation, we see in Table 4 that a commonsense head entity linker trained on the PERSONA-ATOMIC dataset considerably outperforms **Heuristic** baselines for entity linking too. While the **Heuris-**

See Appendix C for more details of experimental settings.
Other unpresented results yield similar conclusions. The full evaluation results are included in Appendix D.
See Appendix E for full evaluation results of fact linking sub-tasks in pipeline setting.

Model	PPL	Distinct-1/2	BLEU-1/2/3/4	METEOR	ROUGE-L	CIDEr	SkipThoughts
CEM	**36.11**	**0.661**/2.998	0.133/0.063/0.034/0.021	0.071	0.163	0.160	0.478
CEM w/ $Com\mathcal{F}act$	36.14	0.655/**3.009**	**0.151/0.072/0.040/0.026**	**0.074**	**0.171**	**0.186**	**0.496**

Table 6: Downstream dialogue response generation results on the EmpatheticDialogues dataset.

tic model can be viewed as a **Recall** oracle in our setup, the precision score of this baseline ends up being considerably worse. For the purpose of this analysis, we also decompose the **Heuristic** baseline into its explicit and implicit components to investigate their individual entity linking performance. This decomposition corresponds to the two entity retrieval methods based on explicit pattern and implicit embedding described in Section . Due to the link type imbalance (*i.e.*, more explicit entity links), we find that Heuristic (explicit) recalls more of the head entities of the corpus, but still suffers in terms of precision, as many of these entities are irrelevant to the context. Heuristic (implicit) has low precision and recall, reinforcing the challenge of identifying relevant facts that are implicit.

Contextual Prediction Commonsense fact linking may be used to identify generic inferences in KGs that help augment full contexts, or to provide inferences that could help generate future portions of a narrative. To simulate fact linking for these two settings, we run experiments for two different input contexts: C_t, where the full context is given to the fact linker, and $U_{\leq t}$, where only the present and past context is given. As expected, our results show that models that receive the half-context window $U_{\leq t}$ perform worse than those that receive the full context window C_t as input, largely due to not recovering facts associated with the future context. In Table 5, we observe a decreasing recall from **RPA** to **RPP** to **RPF** facts, showing that linking is more difficult when a commonsense fact requires a more challenging contextual relevance judgment. The performance degradation supports our contention on the temporal *ambiguity* of commonsense facts: without knowing the future, many possible inferences may seem relevant or irrelevant in the present. However, facts that are relevant to the future (**RPF**) make up a small portion of the data (Table 2), so this shortcoming is not as clear in the overall reported performance.

Cross-Resource Generalization As annotating new commonsense fact links for all narrative datasets would be too expensive, fact linkers will need to generalize to new types of contexts with minimal performance loss. So we evaluate the performance of our fact linkers across different training and testing combinations of the four $Com\mathcal{F}act$ data portions, with results for RoBERTa (large) shown in Figure 4.

Optimistically, we find that the model trained on MOVIE-ATOMIC generalizes reasonably well to the other data portions. Models trained on the other data subsets still significantly beat the heuristic baselines on all testing subsets (and many of the other baselines from Table 3), but do not transfer as robustly. In particular, MOVIE-ATOMIC poses a challenging adaption problem, likely due to the relatively longer and more complex narratives found in the MovieSummaries corpus on which MOVIE-ATOMIC is annotated. Therefore, our trained fact linkers still have room to improve before they reliably scale to open-domain narrative corpora, making $Com\mathcal{F}act$ a promising testbed for further research on developing scalable fact linkers.

However, narrative generalization may not be enough for scaling fact linkers. As other knowledge resources are available (and new ones are constructed), models trained on our data should generalize to link to new commonsense knowledge resources. While more research is needed into cross-KG fact linking, we note that since $Com\mathcal{F}act$ is developed with ATOMIC$^{20}_{20}$, models trained on our benchmark learn to link rich physical, event-based, and social interaction commonsense inferences. In fact, as a portion of ATOMIC$^{20}_{20}$ includes a subset of ConceptNet (Speer, Chin, and Havasi 2017), we can stratify performance along the inferences found in ConceptNet (mainly physical relations), and see that our models actually perform better on this subset of the data than on social interactions from ATOMIC.

Downstream Application Our resource enables the development of improved fact linking models, which will provide more contextually-relevant commonsense knowledge to downstream task systems. To evaluate this hypothesis, we use the **CEM** model (Sabour, Zheng, and Huang 2022) trained on the EmpatheticDialogues dataset (Rashkin et al. 2019). **CEM** conditions on commonsense knowledge generated by COMET (Bosselut et al. 2019) to improve empathetic dialogue generation. Using their framework, we apply a $Com\mathcal{F}act$-trained DeBERTa (large) model to filter contextually-relevant fact subsets from the knowledge generated by COMET, and use this refined knowledge as the input to the **CEM** model (denoted as **CEM w/** $Com\mathcal{F}act$). Our results in Table 6 demonstrate that **CEM** w/ $Com\mathcal{F}act$ outperforms **CEM** on most metrics, hinting at $Com\mathcal{F}act$'s potential to benefit downstream NLP tasks by enabling improved commonsense knowledge retrieval.

Conclusion

In this work, we propose a general commonsense fact linking task that addresses the challenges of identifying relevant commonsense inferences for textual contexts: contextu-

Measuring recall in knowledge retrieval is a recurring challenge as it requires a gold set of relevant facts. In our case, we record recall with respect to a gold candidate set that our heuristics initially over-sample. We provide more discussion on this decision in Appendix C.

See Appendix G for data examples in $Com\mathcal{F}act$.

See Appendix F for more analysis on knowledge graph generalization of fact linkers.

See Appendix H for more downstream application details.

alization, implicitness, and ambiguity. To promote research into commonsense fact linking, we construct a new, challenging benchmark, $\mathcal{C}om\mathcal{F}act$, of over 293k contextually-linked commonsense facts. Our experimental results show that the predominant heuristic methods used to select relevant commonsense facts for downstream applications perform poorly, motivating the need for new methods that can predict contextually-relevant commonsense inferences.

References

Arabshahi, F.; Lee, J.; Bosselut, A.; Choi, Y.; and Mitchell, T. 2021. Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bamman, D.; O'Connor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 352–361.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bauer, L.; Wang, Y.; and Bansal, M. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. In *EMNLP*.

Bosselut, A.; Bras, R. L.; and Choi, Y. 2021. Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1406–1416.

Da, J.; Bras, R. L.; Lu, X.; Choi, Y.; and Bosselut, A. 2021. Analyzing Commonsense Emergence in Few-shot Knowledge Models. In *Proceedings of the Conference on Automated Knowledge Base Construction (AKBC)*.

Davis, E.; and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9): 92–103.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Feng, Y.; Chen, X.; Lin, B. Y.; Wang, P.; Yan, J.; and Ren, X. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1295–1309.

Gajbhiye, A.; Moubayed, N. A.; and Bradley, S. 2021. ExBERT: An External Knowledge Enhanced BERT for Natural Language Inference. In *International Conference on Artificial Neural Networks*, 460–472. Springer.

Guan, J.; Huang, F.; Zhao, Z.; Zhu, X.; and Huang, M. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics*, 8: 93–108.

Guan, J.; Wang, Y.; and Huang, M. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6473–6480.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Huang, C.; He, W.; and Liu, Y. 2021. Improving Unsupervised Commonsense Reasoning Using Knowledge-Enabled Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4875–4885.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6384–6392.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; Zhu, X.; and Huang, M. 2020. Language Generation with Multi-hop Reasoning on Commonsense Knowledge Graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 725–736.

Jiang, L.; Bosselut, A.; Bhagavatula, C.; and Choi, Y. 2021. "I'm Not Mad": Commonsense Implications of Negation and Contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4380–4397. Online: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Kolluru, K.; Rezk, M.; Verga, P.; Cohen, W.; and Talukdar, P. P. 2021. Multilingual Fact Linking. *ArXiv*, abs/2109.14364.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2): 167–195.

Lenat, D. B. 1995. CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38: 32–38.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, W. B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. 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)*, 2829–2839.

Lin, B. Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; and Ren, X. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1823–1840.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Ling, X.; Singh, S.; and Weld, D. S. 2015. Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics*, 3: 315–328.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4): 211–226.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mihaylov, T.; and Frank, A. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 821–832. Melbourne, Australia: Association for Computational Linguistics.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 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*, 839–849.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pavlick, E.; and Kwiatkowski, T. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7: 677–694.

Peng, X.; Li, S.; Wiegreffe, S.; and Riedl, M. 2021a. Inferring the reader: Guiding automated story generation with commonsense reasoning. *arXiv preprint arXiv:2105.01311*.

Peng, X.; Xie, K.; Alabdulkarim, A.; Kayam, H.; Dani, S.; and Riedl, M. O. 2021b. Guiding Neural Story Generation with Reader Models. *arXiv preprint arXiv:2112.08596*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Raman, M.; Agarwal, S.; Wang, P.; Chan, A.; Wang, H.; Kim, S.; Rossi, R. A.; Zhao, H.; Lipka, N.; and Ren, X. 2021. Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation. *ArXiv*, abs/2010.12872.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, 3982–3992.

Sabour, S.; Zheng, C.; and Huang, M. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11229–11237.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. 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)*, 4463–4473.

Singh, P.; Lin, T.; Mueller, E. T.; Lim, G.; Perkins, T.; and Zhu, W. L. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 1223–1237. Springer.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: a core of semantic knowledge. In *WWW '07*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158.

Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.-R.; and Yan, R. 2022. MISC: A MIxed Strategy-Aware Model Integrating COMET for Emotional Support Conversation. *arXiv preprint arXiv:2203.13560*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, K.; Zhang, Y.; Yang, D.; Song, L.; and Qin, T. 2021. GNN is a Counter? Revisiting GNN for Question Answering. *ArXiv*, abs/2110.03192.

Wang, X.; Kapanipathi, P.; Musa, R.; Yu, M.; Talamadupula, K.; Abdelaziz, I.; Chang, M.; Fokoue, A.; Makni, B.; Mattei, N.; et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7208–7215.

Weissenborn, D.; Kovcisk'y, T.; and Dyer, C. 2018. Dynamic Integration of Background Knowledge in Neural NLU Systems. *arXiv: Computation and Language*.

West, P.; Bhagavatula, C.; Hessel, J.; Hwang, J. D.; Jiang, L.; Bras, R. L.; Lu, X.; Welleck, S.; and Choi, Y. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wu, S.; Li, Y.; Zhang, D.; Zhou, Y.; and Wu, Z. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5811–5820.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 535–546.

Ye, Z.-X.; Chen, Q.; Wang, W.; and Ling, Z.-H. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *arXiv preprint arXiv:1908.06725*.

Yin, D.; Dong, L.; Cheng, H.; Liu, X.; Chang, K.-W.; Wei, F.; and Gao, J. 2022. A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models. *arXiv preprint arXiv:2202.08772*.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213.

Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C. D.; and Leskovec, J. 2022. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. In *Proceedings of the 10th International Conference for Learning Representations (ICLR)*.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4623–4629.

Zhou, P.; Gopalakrishnan, K.; Hedayatnia, B.; Kim, S.; Pujara, J.; Ren, X.; Liu, Y.; and Hakkani-Tur, D. 2022. Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1237–1252.

Appendix

A. Data Collection Details

Dataset Selection The narrative contexts in $\mathcal{C}om\mathcal{F}act$ are sampled from four stylistically diverse dialogue and storytelling datasets. Specifically, PERSONA-CHAT (Zhang et al. 2018) contains a rich amount of consistent chit-chat dialogues crowdsourced with additional persona profiles. MuTual (Cui et al. 2020) contains more reasoning-focused dialogues from English listening comprehension exams. ROCStories (Mostafazadeh et al. 2016) and the CMU Movie Summary Corpus (Bamman, O'Connor, and Smith 2013) are commonly used storytelling datasets. All above datasets involve elaborate contextual inference and understanding.

For PERSONA-CHAT, MuTual and ROCStories, we sample dialogues or stories which (potentially) involve the richest commonsense knowledge. In particular, we conduct fact candidate linking (as described in Sec.) on all dialogues and stories in these datasets, and then select the dialogues and stories that have the most fact candidates. For the CMU Movie Summary Corpus, we sample movie summaries that belong to the genre of *slice of life story*, *childhood drama*, *children's* and/or *family*, which are supposed to involve more commonsense inferences, and meanwhile remove movie summaries that also belong to non-commonsensical genres, *e.g.*, *fantasy*, *supernatural*, *mystery*, etc. The total number of our sampled dialogues/stories, statements, and linked head entities and fact candidates are summarized in Table 7.

Knowledge Graph We use ATOMIC$^{20}_{20}$ (Hwang et al. 2021) as the commonsense knowledge graph for building $\mathcal{C}om\mathcal{F}act$. This advanced knowledge graph contains 1.33M everyday inferential facts covering a rich variety of complex entities, where 0.21M facts are about physical objects, 0.20M facts are centered on daily events, and other 0.92M facts involve social interactions.

Crowdsourcing Details For validating the relevance of head entities, similar to the validation of fact candidates, we ask two workers to annotate each head candidate independently. Head candidates are labeled as: a) ***relevant*** **with full confidence** if both workers identify the head entity as relevant, b) ***relevant*** **with half confidence** if only one of the workers choose relevant, or c) ***irrelevant*** if neither of the workers choose relevant.

We also conduct worker qualifications for our crowdsourcing relevance judgements described in Sec. . Specifically, for head entity validation, we test workers with 5 narrative contexts, each with 4 linked head entities, and choose workers who can annotate 19 or more (*i.e.*, $\geq 95\%$) head entities reasonably. For fact candidate validation, we

Statistic	PERSONA-ATOMIC	MUTUAL-ATOMIC	ROC-ATOMIC	MOVIE-ATOMIC
Dialogue/Story Samples	123	237	328	81
Statements	1740	1554	1640	1476
Linked Head Entities	17421	13088	17553	20051
Linked Fact Candidates	72003	53120	81928	86083

Table 7: Number of sampled contexts and linked fact candidates on the four data portions of $\mathcal{C}om\mathcal{F}act$. The context window size is set to 5 where past context $U_{<t} = [U_{t-2}, U_{t-1}]$ and future context $U_{>t} = [U_{t+1}, U_{t+2}]$.

Method	PERSONA-ATOMIC				MUTUAL-ATOMIC				ROC-ATOMIC				MOVIE-ATOMIC			
	RFC	RHC	IRR	κ	RFC	RHC	IRR	κ	RFC	RHC	IRR	κ	RFC	RHC	IRR	κ
Explicit	5652 54%	1846 18%	2963 28%	0.62	2233 42%	1692 32%	1393 26%	0.38	5172 55%	2143 23%	2038 22%	0.48	6932 55%	2853 22%	2892 23%	0.50
Implicit	900 13%	1066 15%	4994 72%	0.53	1281 17%	1697 22%	4792 61%	0.45	2530 31%	1587 19%	4083 50%	0.59	1481 20%	1326 18%	4567 62%	0.56
Both	6552 38%	2912 17%	7957 46%	0.66	3514 27%	3389 26%	6185 47%	0.47	7702 44%	3730 21%	6121 35%	0.57	8413 42%	4179 21%	7459 37%	0.58

Table 8: Relevance validation results of head entities with respect to different fact candidate linking methods. "RFC", "RHC" and "IRR" denote *relevant with full confidence*, *relevant with half confidence* and *irrelevant*, respectively. κ denotes the Cohen's κ.

still test workers with 5 narrative contexts, each with 4 linked fact candidates, and choose workers who can annotate 18 or more (*i.e.*, $\geq$ 90%) fact candidates reasonably. The number of workers that we choose as qualified for head entity and fact candidate validation are 54 and 106, respectively. For PERSONA-ATOMIC, MUTUAL-ATOMIC and ROC-ATOMIC, we pay each worker $1.20 for every 60 annotations in the head entity validation, and $1.00, $1.60 and $2.00 for every 60 annotations in the three rounds of fact candidate validation, respectively. For MOVIE-ATOMIC, which involves more complex narratives, we pay each worker $1.80 for every 60 annotations in the head entity validation, and $1.50, $2.50 and $3.50 for every 60 annotations in the three rounds of fact candidate validation, respectively. The average hourly wage for each worker is about $25.00.

B. Data Statistics Details

Table 8 shows stratified statistics of the crowdsourced head entity relevance annotations for the different candidate linking methods described in Sec. (*i.e.*, explicit pattern matching, implicit embedding matching). For contextual relevance, we draw similar conclusions as the statistics of fact relevance annotations described in Sec. . While in terms of implicitness, different from the statistics of fact relevance annotations, we find that implicit head entities retrieved using embedding similarity are less likely to be relevant to the context than pattern-matched head entities. This shows that implicitly related head entities are more difficult to be retrieved than the explicitly related head entities. We also observe that relevance validation on head entities has overall lower Cohen's κ than that of fact candidates shown in Table 1. This indicates that linked head entities contain more relevance controversy compared to their fact candidates, since a head node is more vague than the whole fact it relates to, which provides less information that increases the ambiguity in relevance. Besides, the Cohen's κ ranking of explicit and implicit (and both) methods in head entity linking seems to change randomly across different data portions in $\mathcal{C}om\mathcal{F}act$. This implies that the difference of ambiguity between explicit and implicit linking narrows down as the linked object becomes simpler (*i.e.*, from a whole fact triple to a head node in it).

We also investigate the coverage of *always* and *sometimes relevant* facts in $\mathcal{C}om\mathcal{F}act$ on the relations of ATOMIC$^{20}_{20}$ knowledge graph. The statistical results are shown in Table 9, where different ATOMIC$^{20}_{20}$ relations are associated with different commonsense knowledge types. Besides the simplest knowledge of ObjectUse, we find that social interaction of "PersonX" (xNeed, xWant, xIntent, xReact, xEffect and xAttr) occupy a large proportion of relevant facts. And in general, more than half of the relevant facts are beyond simple physical knowledge, which involve more complicated daily events or social interactions. This shows that context inference and understanding widely involve complicated daily events and social knowledge, which are difficult to be retrieved by simple heuristics. This reveals the necessity of improving commonsense fact linking methods in NLP systems.

Complex Structures Even though our dataset contains annotations for *individual* linked facts, we find that complex graphical structure emerges among these annotated facts. Each narrative sample we annotate results in an average of 101 *bridge* paths where two relevant facts share the same tail entity. Such paths form potentially explanatory multi-hop reasoning chains among facts relevant to the narrative sample. We also find an average of 492 bridge paths among irrelevant facts and 143 bridge paths between relevant and irrelevant facts (*i.e.*, likely invalid reasoning chains), demonstrating the importance of precisely retrieving facts to avoid spurious explanations (Raman et al. 2021).

Type	Relation	PERSONA-ATOMIC		MUTUAL-ATOMIC		ROC-ATOMIC		MOVIE-ATOMIC	
		Count	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
Physical	ObjectUse	3641	32.2%	4567	40.8%	4747	28.7%	3257	26.2%
	HasProperty	834	7.4%	484	4.3%	435	2.6%	535	4.3%
	CapableOf	424	3.7%	299	2.7%	391	2.4%	614	4.9%
	AtLocation	394	3.5%	634	5.7%	375	2.3%	503	4.0%
	MadeUpOf	238	2.1%	336	3.0%	131	0.8%	200	1.6%
	Desires	34	0.3%	16	0.1%	8	0.05%	31	0.2%
	NotDesires	32	0.3%	16	0.1%	5	0.03%	62	0.5%
	Total	5597	49.5%	6352	56.7%	6092	36.8%	5202	41.8%
Event	HasSubEvent	680	6.0%	530	4.7%	629	3.8%	374	3.0%
	HinderedBy	245	2.2%	184	1.6%	498	3.0%	358	2.9%
	xReason	72	0.6%	28	0.2%	35	0.2%	20	0.2%
	Causes	29	0.3%	24	0.2%	36	0.2%	75	0.6%
	isFilledBy	26	0.2%	25	0.2%	43	0.3%	29	0.2%
	isBefore	21	0.2%	39	0.3%	109	0.7%	94	0.8%
	isAfter	14	0.1%	27	0.2%	113	0.7%	98	0.8%
	Total	1087	9.6%	857	7.6%	1463	8.8%	1048	8.4%
Social	xNeed	1470	13.0%	1306	11.7%	2883	17.4%	1280	10.3%
	xWant	973	8.6%	766	6.8%	1982	12.0%	1143	9.2%
	xIntent	876	7.7%	538	4.8%	1560	9.4%	764	6.1%
	xReact	368	3.3%	253	2.3%	411	2.5%	425	3.4%
	xEffect	361	3.2%	329	2.9%	951	5.7%	912	7.3%
	xAttr	352	3.1%	360	3.2%	369	2.2%	512	4.1%
	oWant	111	1.0%	258	2.3%	430	2.6%	440	3.5%
	oEffect	64	0.6%	102	0.9%	287	1.7%	474	3.8%
	oReact	56	0.5%	83	0.7%	136	0.8%	252	2.0%
	Total	4631	40.9%	3995	35.7%	9009	54.4%	6202	49.8%
Total		11315	100.0%	11204	100.0%	16564	100.0%	12452	100.0%

Table 9: ATOMIC$^{20}_{20}$ relation coverage of relevant facts in $\mathcal{C}om\mathcal{F}act$.

Split	PERSONA-ATOMIC				MUTUAL-ATOMIC				ROC-ATOMIC				MOVIE-ATOMIC			
	# N/D	# S/U	# HE	# FC	# N/D	# S/U	# HE	# FC	# N/D	# S/U	# HE	# FC	# N/D	# S/U	# HE	# FC
Train	90	1296	12789	49526	170	1147	9613	33968	235	1175	12520	53045	58	1047	14205	52690
Valid	15	194	1985	7912	33	230	1778	5896	46	230	2434	10403	11	177	2481	9307
Test	18	250	2647	9758	34	177	1697	6328	47	235	2599	10854	12	252	3365	11646
Total	123	1740	17421	67196	237	1554	13088	46192	328	1640	17553	74302	81	1476	20051	73643

Table 10: Split of narratives or dialogues (N/D) on the four data portions of $\mathcal{C}om\mathcal{F}act$, with their contained number of statements or utterances (S/U), labeled head entities (HE) and labeled fact candidates (FC).

C. Experimental Details

Data Preprocessing For all head entities labeled as *relevant* in our data (regardless of the confidence), we combine fact candidates labeled as *always* and *sometimes relevant* as positive samples, and keep fact candidates labeled as *irrelevant* as negative samples. Fact candidates labeled as *at odds* are not included in this evaluation, though we release them as part of the dataset. For head entities initially labeled as *irrelevant*, all their fact candidates are negative samples.

Table 10 shows our split of training, development and testing sets on the four data portions of $\mathcal{C}om\mathcal{F}act$. Note that the total number of labeled fact candidates does not match up with Table 1 because we remove at odds facts and include facts of irrelevant head entities which are not validated in crowdsourcing.

For the fact linker based on two-layer bidirectional LSTM, we use GloVe (Pennington, Socher, and Manning 2014) to initialize the embedding matrix, and set embedding size, hidden size and vocabulary size as 300, 300 and 10000, respectively. The LSTM encoder is combined with an MLP classifier on top of the output hidden state concatenation of the start token "¡bos¿" and the end token "¡eos¿", where the MLP inner layer size is set as 1200. We set dropout rate as 0.5 and use Adam optimizer (Kingma and Ba 2014) with a learning rate of $2e^{-4}$, which is selected via grid search from $\{5e^{-5}, 1e^{-4}, 2e^{-4}, 5e^{-4}, 1e^{-3}\}$. For the fact linkers based on pretrained language models, we use the default model settings on Hugging Face. All pretrained language models are combined with a linear classifier on top of the output hidden states of the start token ("[CLS]" for BERT and DistilBERT, "¡s¿" for RoBERTa). Adam optimizer is still used and the learning rates are set as $2e^{-6}$ for DistilBERT, BERT (base) and RoBERTa (base), and $5e^{-7}$ for BERT (large) and RoBERTa (large), selecting via grid search from $\{1e^{-7}, 2e^{-7}, 5e^{-7}, 1e^{-6}, 2e^{-6}, 5e^{-6}, 1e^{-5}\}$.

https://huggingface.co

Context	Model	Setting	PERSONA-ATOMIC				ROC-ATOMIC			
			Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
$U_{\leq t}$	Heuristic (explicit)	none	0.451	0.247	0.779	0.375	0.501	0.258	0.616	0.364
	Heuristic (implicit)		0.549	0.140	0.221	0.171	0.499	0.198	0.384	0.261
	Heuristic		0.211	0.211	**1.000**	0.348	0.231	0.231	**1.000**	0.375
	Head Linking		0.600	0.334	0.899	0.487	0.537	0.326	0.936	0.484
	LSTM	direct	0.805	0.549	0.413	0.471	0.761	0.450	0.464	0.457
		pipeline	0.796	0.523	0.404	0.456	0.745	0.479	0.386	0.428
	DistilBERT	direct	0.840	0.617	0.636	0.626	0.811	0.584	0.635	0.608
		pipeline	0.834	0.615	0.568	0.591	0.791	0.540	0.657	0.593
	BERT (base)	direct	0.859	0.672	0.647	0.659	0.841	0.646	0.693	0.669
		pipeline	0.855	0.668	0.622	0.644	0.819	0.595	0.687	0.638
	BERT (large)	direct	0.859	0.670	0.650	0.660	0.848	0.653	0.732	0.690
		pipeline	0.856	0.673	0.620	0.645	0.828	0.606	0.728	0.661
	RoBERTa (base)	direct	0.866	0.691	0.657	0.674	0.845	0.645	0.734	0.687
		pipeline	0.859	0.663	0.679	0.671	0.831	0.609	0.748	0.671
	RoBERTa (large)	direct	0.883	0.735	0.698	0.716	0.874	0.709	0.774	0.740
		pipeline	0.874	0.706	0.690	0.698	0.861	0.673	0.776	0.721
	DeBERTa (base)	direct	0.869	0.702	0.664	0.682	0.871	0.699	0.781	0.737
		pipeline	0.864	0.683	0.661	0.672	0.848	0.637	0.793	0.706
	DeBERTa (large)	direct	**0.885**	**0.743**	0.693	**0.717**	**0.884**	**0.725**	0.806	**0.763**
		pipeline	0.873	0.691	0.722	0.706	0.868	0.678	0.816	0.741
C_t	Head Linking	none	0.595	0.330	0.892	0.482	0.543	0.328	0.931	0.485
	LSTM	direct	0.818	0.615	0.369	0.461	0.766	0.491	0.365	0.419
		pipeline	0.813	0.605	0.350	0.443	0.756	0.470	0.436	0.452
	DistilBERT	direct	0.844	0.630	0.627	0.628	0.808	0.572	0.668	0.616
		pipeline	0.842	0.633	0.600	0.616	0.791	0.537	0.703	0.609
	BERT (base)	direct	0.862	0.682	0.651	0.666	0.845	0.650	0.716	0.681
		pipeline	0.853	0.661	0.624	0.642	0.823	0.596	0.724	0.654
	BERT (large)	direct	0.869	0.704	0.657	0.680	0.855	0.667	0.743	0.703
		pipeline	0.854	0.666	0.622	0.643	0.833	0.610	0.767	0.680
	RoBERTa (base)	direct	0.871	0.703	0.671	0.687	0.845	0.642	0.748	0.691
		pipeline	0.862	0.670	0.682	0.676	0.839	0.627	0.750	0.683
	RoBERTa (large)	direct	0.882	0.721	0.720	**0.721**	0.879	**0.720**	0.779	0.748
		pipeline	0.874	0.716	0.691	0.703	0.867	0.687	0.781	0.731
	DeBERTa (base)	direct	0.871	0.701	0.676	0.688	0.859	0.661	0.799	0.723
		pipeline	0.864	0.683	0.660	0.671	0.853	0.644	0.809	0.718
	DeBERTa (large)	direct	**0.884**	**0.734**	0.707	0.720	**0.884**	0.717	0.826	**0.768**
		pipeline	0.880	0.730	0.684	0.707	0.880	0.712	0.806	0.756

Table 11: Fact linking results on PERSONA-ATOMIC and ROC-ATOMIC.

All models are trained using binary cross entropy loss.

On each data portion of $\mathcal{C}om\mathcal{F}act$, we train each fact linker for 20 epochs and test the performance of the fact linker from the epoch where it achieves the best F1 score on the validation set. The training and evaluation batch sizes are both set as 8 for LSTM, DistilBERT, BERT (base) and RoBERTa (base), and 2 for BERT (large) and RoBERTa (large). Model training and evaluation is performed on four NVIDIA TITAN X Pascal GPUs.

Recall Measurement Finally, we note that our F1 scores depend on a faithful measurement of recall. This measurement poses a recurring challenge in open-ended retrieval tasks as measuring recall requires approximating of the "true" number of relevant facts to a narrative context using a suitable gold set of facts. In our case, the full ATOMIC$^{20}_{20}$ KG would be the most expansive candidate set, but using it as a gold set would require annotating 1M+ facts for each narrative context, which is not scalable. Instead, we record recall with respect to a concrete set of candidates that heuristics initially over-sample, $\sim$41 facts per example context (excluding the facts which are annotated as *at odds* by crowdworkers), producing a diverse set of initial facts with which to measure recall (*i.e.*, models would not score highly on F1 simply by making few predictions). Conceptually, we also note that commonsense KGs only provide a very limited snapshot of commonsense in the world. Therefore, even the full ATOMIC$^{20}_{20}$ KG would not provide a complete picture of relevant commonsense knowledge for measuring recall.

D. Full Results of Fact Linking

Table 11 and 12 shows the full evaluation results of our fact linking baselines on $\mathcal{C}om\mathcal{F}act$. In terms of the comparisons

Context	Model	Setting	MUTUAL-ATOMIC				MOVIE-ATOMIC			
			Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
$U_{\leq t}$	Heuristic (explicit)	none	0.555	0.344	0.583	0.433	0.487	0.247	0.738	0.370
	Heuristic (implicit)		0.445	0.239	0.417	0.304	0.513	0.138	0.262	0.181
	Heuristic		0.290	0.290	**1.000**	0.450	0.205	0.205	**1.000**	0.340
	Head Linking		0.643	0.445	0.931	0.602	0.548	0.301	0.912	0.452
	LSTM	direct	0.749	0.566	0.581	0.573	0.769	0.431	0.404	0.417
		pipeline	0.745	0.557	0.596	0.576	0.761	0.409	0.374	0.391
	DistilBERT	direct	0.792	0.636	0.663	0.649	0.800	0.510	0.514	0.512
		pipeline	0.773	0.602	0.651	0.625	0.796	0.501	0.509	0.505
	BERT (base)	direct	0.801	0.648	0.690	0.668	0.818	0.558	0.537	0.547
		pipeline	0.781	0.616	0.656	0.635	0.814	0.547	0.534	0.540
	BERT (large)	direct	0.819	0.672	0.706	0.689	0.835	0.598	0.593	0.595
		pipeline	0.797	0.634	0.711	0.670	0.827	0.575	0.588	0.582
	RoBERTa (base)	direct	0.810	0.663	0.703	0.682	0.820	0.561	0.546	0.553
		pipeline	0.794	0.631	0.700	0.664	0.818	0.551	0.543	0.547
	RoBERTa (large)	direct	0.835	0.702	0.748	0.724	**0.850**	**0.635**	0.628	0.631
		pipeline	0.819	0.658	0.787	0.717	0.834	0.590	0.623	0.606
	DeBERTa (base)	direct	0.836	0.694	0.779	0.734	0.820	0.562	0.551	0.556
		pipeline	0.822	0.665	0.781	0.719	0.817	0.553	0.549	0.551
	DeBERTa (large)	direct	**0.861**	**0.751**	0.781	**0.766**	**0.850**	0.621	0.683	**0.651**
		pipeline	0.835	0.671	0.847	0.749	0.839	0.596	0.656	0.625
C_t	Head Linking	none	0.647	0.447	0.913	0.600	0.564	0.307	0.901	0.458
	LSTM	direct	0.755	0.577	0.585	0.581	0.772	0.437	0.393	0.414
		pipeline	0.756	0.581	0.579	0.580	0.771	0.434	0.388	0.409
	DistilBERT	direct	0.795	0.641	0.667	0.654	0.807	0.527	0.536	0.531
		pipeline	0.781	0.620	0.638	0.629	0.801	0.514	0.513	0.514
	BERT (base)	direct	0.804	0.653	0.694	0.673	0.829	0.587	0.562	0.574
		pipeline	0.795	0.638	0.682	0.659	0.814	0.544	0.555	0.550
	BERT (large)	direct	0.817	0.670	0.727	0.697	0.839	0.607	0.604	0.605
		pipeline	0.797	0.639	0.688	0.663	0.824	0.567	0.597	0.582
	RoBERTa (base)	direct	0.811	0.662	0.716	0.688	0.832	0.590	0.588	0.589
		pipeline	0.795	0.638	0.682	0.659	0.809	0.530	0.576	0.552
	RoBERTa (large)	direct	0.838	0.694	0.792	0.740	**0.851**	**0.637**	0.633	0.635
		pipeline	0.825	0.671	0.780	0.722	0.830	0.578	0.631	0.603
	DeBERTa (base)	direct	0.842	0.712	0.766	0.738	0.842	0.618	0.601	0.609
		pipeline	0.824	0.661	0.805	0.726	0.805	0.520	0.619	0.565
	DeBERTa (large)	direct	**0.859**	**0.726**	0.826	**0.773**	0.848	0.620	0.657	**0.638**
		pipeline	0.848	0.705	0.817	0.757	0.839	0.595	0.666	0.628

Table 12: Fact linking results on MUTUAL-ATOMIC and MOVIE-ATOMIC.

between different models, between direct and pipeline settings, and between context windows $U_{\leq t}$ and C_t, we draw similar conclusions as described in Sec. .

E. Results of Fact Linking Sub-Tasks in Pipeline Setting

Table 13 and 14 show the evaluation results of the fact linking sub-tasks in the pipeline prediction setting, including head entity linking and fact linking of head entities classified as relevant. Experiments are conducted on the PERSONA-ATOMIC and ROC-ATOMIC data portions of $\mathcal{C}om\mathcal{F}act$. For both sub-tasks, we find that all language models achieve overall higher evaluation results compared to their results in the direct prediction setting shown in Table 3. This shows that the pipeline prediction successfully divides the whole fact linking task into two simpler steps. However, as described in Sec. , this does not make the pipeline prediction finally outperforms the direct prediction, as error propagation exits between the two sub-tasks.

F. Knowledge Graph Generalization

Table 15 shows the fine-grained fact linking results of RoBERTa (large) with respect to different ATOMIC$^{20}_{20}$ fact types. In particular, we evaluate RoBERTa (large) model (finetuned on universal training set) on different test subsets including physical facts, event-based and social (event&social) facts, or all facts (*i.e.*, universal test set). The promising performance of RoBERTa (large) on linking physical knowledge reveals that fact linkers developed on $\mathcal{C}om\mathcal{F}act$ has the potential of generalizing to other knowledge graphs, *e.g.*, ConceptNet whose physical facts make up part of the ATOMIC$^{20}_{20}$ contents. We also observe that evalu-

Context	Model	PERSONA-ATOMIC				ROC-ATOMIC			
		Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
$U_{\leq t}$	LSTM	0.683	0.767	0.612	0.681	0.704	0.752	0.795	0.773
	DistilBERT	0.787	0.779	0.856	0.816	0.764	0.780	0.875	0.825
	BERT (base)	0.802	0.803	0.852	0.827	0.766	0.782	0.874	0.825
	BERT (large)	0.794	0.796	0.845	0.820	0.772	0.782	0.888	0.832
	RoBERTa (base)	0.823	0.830	0.854	0.842	0.783	0.802	0.871	0.835
	RoBERTa (large)	**0.834**	0.834	0.874	**0.854**	0.819	0.830	**0.898**	**0.863**
C_t	LSTM	0.688	0.759	0.638	0.693	0.703	0.744	0.809	0.775
	DistilBERT	0.794	0.783	0.867	0.823	0.758	0.771	0.879	0.821
	BERT (base)	0.793	0.791	0.852	0.820	0.773	0.787	0.879	0.830
	BERT (large)	0.797	0.804	0.837	0.820	0.780	0.792	0.885	0.836
	RoBERTa (base)	0.825	**0.850**	0.829	0.839	0.791	0.817	0.863	0.839
	RoBERTa (large)	0.832	0.830	**0.875**	0.852	**0.821**	**0.836**	0.892	**0.863**

Table 13: Head entity linking results on PERSONA-ATOMIC and ROC-ATOMIC, given different context windows $U_{\leq t} = [U_{<t}, U_t]$ or $C_t = [U_{<t}, U_t, U_{>t}]$.

Context	Model	PERSONA-ATOMIC				ROC-ATOMIC			
		Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
$U_{\leq t}$	LSTM	0.710	0.719	0.470	0.568	0.694	0.622	0.554	0.586
	DistilBERT	0.789	0.779	0.671	0.721	0.774	0.703	0.730	0.716
	BERT (base)	0.813	0.810	0.705	0.754	0.811	0.746	0.783	0.764
	BERT (large)	0.815	0.811	0.706	0.755	0.824	0.766	0.795	0.780
	RoBERTa (base)	0.836	0.803	0.789	0.796	0.823	0.754	0.812	0.782
	RoBERTa (large)	0.846	0.845	0.760	0.800	0.853	0.803	0.827	0.815
C_t	LSTM	0.718	0.753	0.456	0.568	0.695	0.636	0.514	0.569
	DistilBERT	0.797	0.792	0.678	0.731	0.782	0.702	0.769	0.734
	BERT (base)	0.813	0.811	0.705	0.754	0.823	0.757	0.804	0.780
	BERT (large)	0.819	0.821	0.707	0.760	0.832	0.764	0.826	0.794
	RoBERTa (base)	0.833	0.794	**0.796**	0.795	0.832	0.763	0.829	0.795
	RoBERTa (large)	**0.850**	**0.846**	0.770	**0.806**	**0.864**	**0.819**	**0.838**	**0.828**

Table 14: Fact linking results of head entities classified as relevant on PERSONA-ATOMIC and ROC-ATOMIC, given given different context windows $U_{\leq t} = [U_{<t}, U_t]$ or $C_t = [U_{<t}, U_t, U_{>t}]$.

Fact Type	Acc.	Prec.	Recall	F1
physical	0.888	0.775	0.753	0.764
event&social	0.879	0.691	0.639	0.664
all	0.883	0.735	0.698	0.716

Table 15: Fact linking results of **RoBERTa (large)** on PERSONA-ATOMIC with respect to different fact types, under the direct setting and window $U_{\leq t} = [U_{<t}, U_t]$.

ation scores of linking event&social facts are overall lower than linking physical facts. This indicates that more elaborate research is needed to make our trained fact linkers generalize to event-based and social knowledge graphs, since they typically involve more complex contents and are often linked to the context in more implicit ways.

G. Data Examples

Table 16 shows examples from the four data portions of $\mathcal{C}om\mathcal{F}act$, including a piece of context from each data portion and its linked facts with different link types. As shown in examples, MOVIE-ATOMIC has more complex narrative contexts, which contains longer statements compared to the other three data portions.

H. Downstream Application Details

We use the CEM (Sabour, Zheng, and Huang 2022) model trained on the EmpatheticDialogues (Rashkin et al. 2019) dataset as our downstream framework. EmpatheticDialogues is a large-scale multi-turn dialogue dataset containing 25K empathetic conversations between crowdworkers. The task of dialogue models on this dataset is to play the role of a listener and generate empathetic responses to a speaker. To augment empathetic response generation with commonsense knowledge, the CEM model first appends five kinds of ATOMIC (Sap et al. 2019a) relation tokens (*xReact*, *xIntent*, *xNeed*, *xEffect* and *xWant*) to each dialogue context. Based on the context and appended tokens, it then uses COMET (Bosselut et al. 2019) to generate five commonsense inferences (*i.e.*, tail entities) for each relation, whose purpose is to help the model generate more empathetic responses.

However, the commonsense inferences generated by COMET may not all be relevant to the dialogue context and helpful for generating more empathetic responses. Therefore, we train neural fact linkers on $\mathcal{C}om\mathcal{F}act$ to refine the knowledge generated by COMET. Specifically, we train a DeBERTa (large) fact linker on the union of all four data portions of $\mathcal{C}om\mathcal{F}act$. To adapt our fact linker to the CEM setting, where tail entities are generated without given head entities, we remove the head entity of each fact out of the input when training the fact linker. We also only use the facts whose relations are one of the five CEM appended relations to build our training samples. Finally, we follow the same hyper-parameter settings suggested by CEM to train the dialogue model (*i.e.*, we do not re-tune the model for our setting), and filter out the COMET inferences which are classified as irrelevant by our trained fact linker, resulting in $\sim$38.5% generated facts being removed from the CEM input. For evaluation, we use the same metrics Perplexity (PPL) and Distinct-n (Li et al. 2016) from Sabour, Zheng, and Huang (2022), and also include commonly used metrics BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SkipThoughts (Kiros et al. 2015). Table 17 shows examples of our downstream dialogue response generation results.

	Persona-Atomic
Context	U_{t-2}: I like cooking macrobiotic and healthy food and working out at the gym. U_{t-1}: What is macrobiotic food? My best friend is my mother. U_t : Things like whole grains. I drink at bars so I have to stay healthy. U_{t+1}: You should not drink a lot, it's bad for you. U_{t+2}: Well that is where I meet women, at bars. So I end up drinking.
Fact	**RPA**: stay healthy, HasSubEvent, eat healthy foods (*always relevant*) **RPP**: stay healthy, xNeed, exercise and eat balanced meals (*always relevant*) **RPF**: bar, ObjectUse, take their friends to (*sometimes relevant*) **IRR** : PersonX likes to drink, xAttr, thirsty (*irrelevant*)
	Mutual-Atomic
Context	U_{t-2}: It's $800 in all, sir. Do you want to pay in cash? U_{t-1}: Well, can I use my check please? U_t : Sorry, sir. We don't take checks. You can pay by credit card. U_{t+1}: OK. Here's my credit card. U_{t+2}: Thank you, sir. Here you go with your credit card and the receipt.
Fact	**RPA**: credit, ObjectUse, pay for the food (*sometimes relevant*) **RPP**: pay by check, HasSubEvent, know amount of check (*always relevant*) **RPF**: card, ObjectUse, give to clerk (*always relevant*) **IRR** : personal check, ObjectUse, pay someone back (*irrelevant*)
	Roc-Atomic
Context	U_{t-2}: Jamie was sleeping at a friend's house. U_{t-1}: It was her first time away at a friend's house. U_t : Jamie was scared and missed her home and family. U_{t+1}: She called her mom to pick her up. U_{t+2}: Jamie went home to sleep in her own bed.
Fact	**RPA**: PersonX misses PersonX's parents, xNeed, to be away from the parents (*always relevant*) **RPP**: family, AtLocation, house (*always relevant*) **RPF**: PersonX feels homesick, xEffect, take leave to go home (*always relevant*) **IRR** : PersonX misses home, xWant, to watch a movie about home (*irrelevant*)
	Movie-Atomic
Context	U_{t-2}: Fred is surprised at how real Wilma's tears are during a major scene, and soon learns she refuses to speak to him for forgetting pebbles. U_{t-1}: Fred soon starts living the play and realizes that Christmas isn't about greed, but about happiness and love. U_t : He soon apologizes to everyone he was rude to, including Wilma, who is still mad at him for his thoughtlessness, and everyone in bedrock truly has a very merry Christmas. U_{t+1}: But then Fred gets sick however because of the bedrock bug. U_{t+2}: This is further complicated when Wilma assures him that he will recover just in time to attend his mother-in-law's Christmas dinner.
Fact	**RPA**: apologize, xIntent, made mistake (*always relevant*) **RPP**: PersonX rude to PersonY, oEffect, starts crying (*always relevant*) **RPF**: PersonX tells Fred, xIntent, Fred to know (*always relevant*) **IRR** : Christmas, HasProperty, celebrated by Christians (*irrelevant*)

Table 16: Data examples in $\mathcal{C}om\mathcal{F}act$.

Context	I could not wait to go to the concert.
CEM	What did you do?
CEM w/ $\mathcal{C}om\mathcal{F}act$	What band is it?
Human	Which concert?
Context	My landlady was really understanding when I told her that I would be late with rent this month.
CEM	That is good, did you find out who did it?
CEM w/ $\mathcal{C}om\mathcal{F}act$	That is so nice of her.
Human	That is nice. I could get that here.
Context	Running my first (and maybe only!) marathon felt like such a huge accomplishment!
CEM	I am sure you will do great!
CEM w/ $\mathcal{C}om\mathcal{F}act$	That is great! Congratulations!
Human	Wow, that is an amazing accomplishment! Congratulations!

Table 17: Examples of downstream dialogue response generation results on the EmpatheticDialogues dataset.