

# It's not Rocket Science: Interpreting Figurative Language in Narratives

Tuhin Chakrabarty<sup>1\*</sup> Yejin Choi<sup>2,3</sup> Vered Shwartz<sup>4\*</sup>

<sup>1</sup>Columbia University, USA <sup>2</sup>Allen Institute for Artificial Intelligence, USA

<sup>3</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington, USA

<sup>4</sup>University of British Columbia, Canada

tuhin.chakr@cs.columbia.edu, yejinc@allenai.org, vshwartz@cs.ubc.ca

## Abstract

Figurative language is ubiquitous in English. Yet, the vast majority of NLP research focuses on literal language. Existing text representations by design rely on compositionality, while figurative language is often non-compositional. In this paper, we study the interpretation of two non-compositional figurative languages (idioms and similes). We collected datasets of fictional narratives containing a figurative expression along with crowd-sourced plausible and implausible continuations relying on the correct interpretation of the expression. We then trained models to choose or generate the plausible continuation. Our experiments show that models based solely on pre-trained language models perform substantially worse than humans on these tasks. We additionally propose knowledge-enhanced models, adopting human strategies for interpreting figurative language types: inferring meaning from the context and relying on the constituent words' literal meanings. The knowledge-enhanced models improve the performance on both the discriminative and generative tasks, further bridging the gap from human performance.

## 1 Introduction

Figurative language is a medium for making language expressive, communicating abstract ideas otherwise difficult to visualize, and provoking emotions (Roberts and Kreuz, 1994; Fussell and Moss, 1998). Despite the ubiquity of figurative language across various forms of speech and writing, the vast majority of NLP research focuses primarily on literal language. Figurative language is often more challenging due to its implicit nature and is seen as “a bottleneck in automatic text understanding” (Shutova, 2011).

In recent years, transformer-based language models (LMs) achieved substantial performance gains across various NLP tasks, however, they still struggle with figurative language. In particular, one of the challenges is that figurative expressions are often non-compositional, that is, the phrase meaning deviates from the literal meanings of its constituents. For instance, the idiom “chicken feed” in Figure 1 denotes “a ridiculously small sum of money” instead of “food for poultry”. By design, transformer-based LMs compute a word representation as a function of the representation of its context. LM-based phrase representations encode the meanings of the constituent words but hardly capture any meaning that is introduced by the composition itself (Yu and Ettinger, 2020). Even though LMs may recognize when a word is used non-literally, and potentially attend to it less, they still struggle to represent the implied, non-literal meaning of such phrases (Shwartz and Dagan, 2019).

While LMs potentially memorize familiar idioms, we can expect them to further struggle with similes, which are often created ad hoc (Carston and Wearing, 2011). For example, in Figure 1, the person is compared to “a high mountain lake without a wind stirring it” to imply calmness. Many such figurative expressions compose in a non-trivial way, and introduce implicit meaning that requires multiple reasoning steps to interpret.

In this paper we work on interpreting idioms and similes in narratives, where they are especially abundant. Existing work on narrative understanding focuses on literal stories, testing models on their ability to answer questions about a narrative (Kočiský et al., 2018) or continue an incomplete narrative (Story Cloze Test; Mostafazadeh et al., 2016). We follow the latter

\*Work done at the Allen Institute for AI.

Given These Narratives	The More Plausible Hypothesis is...
 <p>"I ... I'll ask you once more. How much did you make by a fighter taking that dive?" The boy hesitated. "A thousand dollars." Lucky Luciano laughed. "That's <b>chicken feed</b>".</p>	<p>Well I guess you made more for taking a dive than you could have as a fighter. <span style="color: red;">X</span></p> <p>Next time you sell your honor and dignity, try to hold out for more. <span style="color: green;">✓</span></p>
 <p>"What's the point of painting it?" It's hard to keep up with his mood changes. "You're scaring me," I say. He huffs and returns to the chair, lowering his voice, softening his eyes. His energy shifts once again. Now he feels like <b>a high mountain lake without a wind stirring it</b>.</p>	<p>He still had this rage pent up inside him and I could feel the intensity. <span style="color: red;">X</span></p> <p>He doesn't feel the same as he did, now he is more calm. <span style="color: green;">✓</span></p>

Figure 1: Example narratives from our datasets, containing an idiom (top) or a simile (bottom), along with human-written plausible and implausible continuations.

setup. We extracted short narratives from the Toronto Book corpus (Zhu et al., 2015), each containing a figurative expression, and crowd-sourced plausible and implausible continuations that rely on correct interpretation of the figurative expression. We defined two tasks: a discriminative setting, where the goal is to choose the plausible continuation among two candidates, and a generative setting, where the goal is to generate a plausible continuation that is coherent with the narrative and complies with the meaning of the figurative expression.

We report the performance of an extensive number of state-of-the-art LMs on both tasks, in zero-shot, few-shot, and supervised settings. Our results show that pre-trained LMs including GPT-3 (Brown et al., 2020) perform poorly in the zero-shot and few-shot settings. While the supervised model’s performance is closer to humans, the gap is still substantial: In the discriminative tasks, the gap from human performance was 10 and 14.6 points in accuracy for idioms and similes, respectively. In the generative tasks, there was a striking 24 and 28 points difference in human evaluation of the plausibility of generated continuations.

To further close this gap, we developed knowledge-enhanced models inspired by two human strategies for interpreting unknown idioms, as studied by Cooper (1999) and discussed in Shwartz and Dagan (2019). The first strategy is to infer the expression’s meaning from its *context*, for which we incorporate event-centered inferences from ParaCOMET (Gabriel et al., 2021b). The second relies on the *literal* meanings of the constituent words, using concept-centered knowledge from COMET-ConceptNET (Hwang et al., 2021). Additionally similes are often interpreted

by humans using the *literal* property of the vehicle or object of comparison and thus we use concept-centered knowledge here as well. The knowledge-enhanced models consistently outperformed other models on both datasets and settings, with a substantial gap on the generative tasks.

Furthermore, different strategies were favored for each case: The generative context model performed well on idioms, in line with Cooper’s findings, while the literal model was favored for similes, which are by design based on a constituent’s literal attribute (e.g., calm lake). The knowledge-enhanced models leave room for improvement on our dataset. We hope that future work will use additional techniques inspired by the properties of figurative language and human processing of it. Our code and data are available at <https://github.com/tuhinjubcse/FigurativeNarrativeBenchmark> and our leaderboard is available at <https://leaderboard.allenai.org/idiom-simile/>.

## 2 Background

### 2.1 Idioms

Idioms are figurative expressions with a non-literal meaning. For instance, “break a leg” is a good luck greeting before a performance and shouldn’t be taken literally as wishing someone to injure themselves. Idioms are typically non-compositional (i.e., the meaning of an idiom is not derived from the meanings of its constituents) and fixed (i.e., allowing little variance in syntax and lexical choice).<sup>1</sup> Idiomatic expressions include

<sup>1</sup>The meaning of some idioms may be derived from the non-literal meanings of their constituents. For example, in “spill the beans”, the non-literal meaning of spill is “reveal” and the beans signify the secret (Sag et al., 2002).

proverbs (“actions speak louder than words”), clichés (“what goes around comes around”), euphemisms (“rest in peace”), and more.

Prior work on idioms largely focused on identifying the idiomticity of a multi-word expression. This is a classification task, defined either at the token-level (is the phrase idiomatic within a given context?), or the type-level (may the phrase be idiomatic in some context?) (Fazly et al., 2009; Li and Sporleder, 2009; Verma and Vuppuluri, 2015; Peng and Feldman, 2016; Salton et al., 2016; Liu and Hwa, 2017). Compared to identification, the interpretation of idioms has been less explored. Approaches for representing idiomatic expressions include substituting idioms with literal paraphrases (Liu and Hwa, 2016; Zhou et al., 2021), representing them as a single token, or learning to compose them at the character level rather than the word level (Liu et al., 2017).

With the rising popularity of pre-trained LMs, several recent papers studied their capacity to accurately represent idioms. Shwartz and Dagan (2019) found that while LMs excelled at detecting non-literal word usage (e.g., “flea” in “flea market”), the representation of idiomatic expressions was of lower quality than that of literal ones. Yu and Ettinger (2020) showed that LMs encode the words that appear in a given text, but capture little information regarding phrase meaning. Finally, Garcia et al. (2021) studied the compositionality of noun compounds in English and Portuguese, and found that LM-based models did not perform well on detecting compositionality, and represented idiomticity differently from humans.

## 2.2 Similes

Similes are a figure of speech that compares two things, usually with the intent to make the description more emphatic or vivid, and spark the reader’s imagination (Paul et al., 1970). Similes may either be explicit, namely, specify the topic, vehicle, and similarity property, as in “The house was cold like Antarctica” (where the topic is “house”, the vehicle is “Antarctica” and the property of comparison is “cold”), or implicit, namely, omitting the property, as in “the house was like Antarctica” (Section 3.2). Most work in NLP has focused on simile detection, that is, distinguishing literal from figurative comparisons. Earlier work relied on semantic and

syntactic characteristics, namely, higher semantic similarity between the topic and the vehicle in literal comparisons than in figurative comparisons (Niculae and Danescu-Niculescu-Mizil, 2014; Qadir et al., 2015; Mpouli, 2017), and dictionary definitions (Qadir et al., 2016), while more recent work is based on neural methods (Liu et al., 2018; Zeng et al., 2020). Simile interpretation focused on inferring the implicit property (Qadir et al., 2016). In other lines of work, Chakrabarty et al. (2020b) and Zhang et al. (2021) proposed methods for generating similes from their literal counterparts, while Chakrabarty et al. (2021a) showed that state-of-the-art NLI models fail on pragmatic inferences involving similes.

## 2.3 Human Processing of Figurative Language

The ways in which humans process figurative language may inspire computational work on figurative language interpretation. Cooper (1999) studied how L2 English speakers interpret unfamiliar English idioms. He found that the leading strategy was to infer the meaning from the given context, which led to successful interpretation 57% of the time, followed by relying on the literal meaning of the constituent words (22% success rate). For example, a participant asked to interpret “robbing the cradle” in the context “Robert knew that he was robbing the cradle by dating a sixteen-year-old girl” used the literal meaning of cradle to associate the meaning with babies and indirectly with young age, and along with the context inferred that it meant to “date a very young person”. Asl (2013) repeated the same experiment with stories, and concluded that longer contexts improved people’s ability to interpret unknown idioms. With respect to novel similes and metaphors, they are interpreted through shared literal attributes between the topic and vehicle (e.g., “Antarctica is cold, can a house also be cold?”) (Wolff and Gentner, 2000; Carston and Wearing, 2011).

## 2.4 Narrative Understanding

Early computational work on narrative understanding extracted chains of subevents and their participants from narratives (Chambers and Jurafsky, 2009). An alternative task is machine reading comprehension, that is, answering multiple-choice questions based on a narrative,

such as MCTest (Richardson et al., 2013) and NarrativeQA (Kočiský et al., 2018).

The most commonly used benchmark for narrative understanding today is ROCStories (Mostafazadeh et al., 2016), a collection of 50k five-sentence commonsense stories pertaining to everyday life. The story cloze task requires models to identify the plausible continuation sentence among two candidate continuations in its discriminative form, or generate a plausible sentence, in its generative form. Since the release of this dataset, many computational approaches for the task have been developed (Chaturvedi et al., 2017; Schwartz et al., 2017b; Cai et al., 2017; Srinivasan et al., 2018; Li et al., 2019; Cui et al., 2020; Brown et al., 2020, *inter alia*). In this paper, we follow the story cloze benchmark setup, and collect benchmarks particularly aimed at testing the comprehension of figurative language in narratives.

## 2.5 Commonsense Knowledge Models

Many language tasks require relying on implicit commonsense knowledge that is never mentioned explicitly because it is assumed to be known by everyone. To that end, commonsense knowledge bases (KBs) record such facts. Notably, ConceptNet (Speer et al., 2017) is a large-scale concept-centric KB, while ATOMIC (Sap et al., 2019) contains event-centric knowledge about causes, effects, and the mental states of the participants. To overcome the sparsity of KBs, knowledge models such as COMET (Bosselut et al., 2019; Hwang et al., 2021) fine-tuned an LM on structured KB triplets. COMET is capable of providing inferences for new events or concepts. ParaCOMET (Gabriel et al., 2021a) is an extension of ATOMIC-COMET that works at the paragraph level and generates discourse-aware commonsense knowledge. Recently, several works have used such commonsense knowledge models for improved natural language understanding or generation such as Bhagavatula et al. (2019) for abductive reasoning, Shwartz et al. (2020) for QA, Guan et al. (2019), and Ammanabrolu et al. (2020) for story generation, Majumder et al. (2020) for dialog generation, and Chakrabarty et al. (2020a; 2020b; 2021b) for creative text generation.

In our work we use the knowledge models COMET (Hwang et al., 2021) and ParaCOMET (Gabriel et al., 2021a), respectively, to provide

Idioms	Similes
any port in a storm	like a psychic whirlpool
been there, done that	like a moth-eaten curtain
slap on the wrist	like a first date
no time like the present	like a train barreling of control
lay a finger on	like a sodden landscape of melting snow
walk the plank	like a Bunsen burner flame
curry favour	like a moldy old basement
not to be sneezed at	like a street-bought Rolex
no peace for the wicked	like an endless string of rosary beads

Table 1: Examples of idioms and similes present in the narratives in our datasets.

more information about the literal meaning of constituent words or the narrative context useful to infer the figurative expressions meaning.

## 3 Data

We build datasets aimed at testing the understanding of figurative language in narratives, focusing on idioms (Section 3.1) and similes (Section 3.2). We posit that a model that truly understands the meaning of a figurative expression, like humans do, should be able to infer or decide what happens next in the context of a narrative. Thus, we construct a dataset in the form of the story-cloze test.

### 3.1 Idioms

We compile a list of idioms, automatically find narratives containing these idioms, and then elicit plausible and implausible continuations from crowdsourcing workers, as follows.

**Collecting Idioms.** We compile a list of 554 English idioms along with their definitions from online idiom lexicons.<sup>2</sup> Table 1 presents a sample of the collected idioms.

**Collecting Narratives.** We use the Toronto Book corpus (Zhu et al., 2015), a collection of 11,038 indie ebooks extracted from smashwords.com. We extract sentences from the corpus containing an idiom from our list, and prepend the 4 preceding sentences to create a narrative. We manually discarded paragraphs that did not form a coherent narrative. We extracted 1,455 narratives with an average length of 80 words, spanning 554 distinct idioms.

<sup>2</sup>[www.theidioms.com](http://www.theidioms.com), [idioms.thefreedictionary.com](http://idioms.thefreedictionary.com).

**Collecting Continuations.** We collected plausible and implausible continuations to the narrative. We used Amazon Mechanical Turk to recruit 117 workers. We provided these workers with the narrative along with the idiom definition, and instructed them to write plausible and implausible continuations that are pertinent to the context, depend on the correct interpretation of the idiom, but that don’t explicitly give away the meaning of the idiom. We collected continuations from 3 to 4 workers for each narrative. The average plausible continuation contained 12 words, while the implausible continuations contained 11 words.

To ensure the quality of annotations, we required that workers have an acceptance rate of at least 99% for 10,000 prior HITs (Amazon Mechanical Turk tasks), and pass a qualification test. We then manually inspected the annotations to identify workers who performed poorly in the initial batches, disqualified them from further working on the task, and discarded their annotations.

Our automatic approach for collecting narratives does not account for expressions that may be used figuratively in some contexts but literally in others. For example, the idiom “run a mile” (i.e., avoiding something in any way possible) may also be used literally to denote running a distance of one mile. To avoid including literal usages, we instructed the workers to flag such examples, which we discard from the dataset. We further manually verified all the collected data. Overall, we removed 12 such narratives.

The final idiom dataset contains 5,101 <narrative, continuation> tuples, exemplified in the top part of Figure 1. We split the examples to train (3,204), validation (355), and test (1,542) sets. To test models’ ability to generalize to unseen idioms, we split the data such that there are no overlaps in idioms between train and test.

### 3.2 Similes

A simile is a figure of speech that usually consists of a topic and a vehicle (typically noun phrases) that are compared along a certain property using comparators such as “like” or “as” (Hanks, 2013; Niculae and Danescu-Niculescu-Mizil, 2014). The property may be mentioned (*explicit* simile) or hidden and left for the reader to infer (*implicit* simile). We focus on implicit similes, which are less trivial to interpret than their ex-

plicit counterparts (Qadir et al., 2016), and test a model’s ability to recover the implicit property.

**Collecting Similes.** Because there are no reliable methods for automatically detecting implicit similes, we first identify explicit similes based on syntactic cues, and then deterministically convert them to implicit similes. We look for sentences in the Toronto Book corpus containing one of the syntactic structures “as ADJ/ADV as” or “ADJ/ADV like” as a heuristic for identifying explicit similes. We additionally add the constraint of the vehicle being a noun phrase to avoid examples like “I worked as hard as him”. We remove the adjectival property to convert the simile to implicit, as demonstrated below:

<b>Explicit:</b>
He feels <b>calm</b> , like a high mountain lake without a wind stirring it.
He feels as <b>calm</b> as a high mountain lake without a wind stirring it.
<b>Implicit:</b>
He feels like a high mountain lake without a wind stirring it.

We collected 520 similes along with their associated property. We asked workers to flag any expression that was not a simile, and manually verified all the collected data. Table 1 presents a sample of the collected similes. Many of the similes are original, such as “like a street-bought Rolex” which implies that the subject is fake or cheap.

**Collecting Narratives.** Once we identified the explicit simile and converted it to its implicit form, we similarly prepend the 4 previous sentences to form narratives. The average length of the narrative was 80 words.

**Collecting Continuations.** We repeat the same crowdsourcing setup as for idioms, providing the explicit simile property as the definition. Each narrative was annotated by 10 workers. The average length of continuations was identical to the idiom dataset (12 for plausible and 11 for implausible).

The simile dataset contains 4,996 <narrative, continuation> tuples, exemplified in the bottom part of Figure 1. We split the examples to train (3,100), validation (376), and test (1,520) sets with no simile overlaps between the different sets.

## 4 Discriminative Task

The first task we derive from our dataset is of discriminative nature in the setup of the story cloze

task. Given a narrative  $N$  and two candidate continuations  $\{C_1, C_2\}$ , the goal is to choose which of the continuations is more plausible.

#### 4.1 Methods

For both idioms and similes, we report the performance of several zero-shot, few-shot, and supervised methods as outlined below. Most of our experiments were implemented using the transformers package (Wolf et al., 2020).

**Zero-shot.** The first type of zero-shot models is based on standard language model score as a proxy for plausibility. We use GPT-2 XL (Radford et al., 2019) and GPT-3 (Brown et al., 2020) to compute the normalized log-likelihood score of each continuation given the narrative, predicting the continuation with the highest probability:  $\text{argmax}_i P_{LM}(C_i|N)$ .

We also use UnifiedQA (Khashabi et al., 2020), a T5-3B model (Raffel et al., 2020) trained on 20 QA datasets in diverse formats. We don’t fine-tune it on our dataset, but instead use it in a zero-shot manner, with the assumption that the model’s familiarity with QA format and with the narrative domain through training on the NarrativeQA dataset (Kočiský et al., 2018) would be useful. To cast our task as a QA problem we format the input such that the question is “Which is more plausible between the two based on the context?”.

**Few-shot.** Language models like GPT-3 have shown impressive performance after being prompted with a small number of labelled examples. A prompting example in which the correct continuation is the first is given in the following format:  $Q: N \ (1) \ C_1 \ (2) \ C_2 \ A: (1)$ .

We provided the model with as many prompting examples as possible within the GPT-3 API limit of 2,048 tokens, which is 6 examples. The test examples are provided without the answer and the model is expected to generate (1) or (2).

We also use the recently proposed Pattern Exploiting Training model (PET; Schick and Schütze, 2021). PET reformulates the tasks as a cloze question and fine-tunes smaller masked LMs to solve it using a few training examples.<sup>3</sup>

---

<sup>3</sup>Specifically, it uses ALBERT XXL V2 (Lan et al., 2020), which is 784 times smaller than GPT-3.

We use the following input pattern: “ $N. \ C_1. \ You \ are \ _$ ” for idioms and “ $N. \ C_1. \ That \ was \ _$ ” for similes. PET predicts the masked token and maps it to the label inventory using the verbalizer {“right”, “wrong”} for idioms and {“expected”, “unexpected”} for similes respectively mapping them to {TRUE, FALSE}.<sup>4</sup> We provide each model 100 training examples, train it for 3 epochs, and select the model that yields the best validation accuracy.

**Supervised.** We fine-tune RoBERTa-large (Liu et al., 2019) as a multiple-choice model. For a given instance, we feed each combination of the narrative and a continuation separately to the model in the following format:  $N < s/ > C_i$ .

We pool the representation of the start token to get a single vector representing each continuation, and feed it into a classifier that predicts the continuation score. The model predicts the continuation with the higher score. We fine-tune the model for 10 epochs with a learning rate of  $1e-5$  and a batch size of 8, and save the best checkpoint based on validation accuracy.

**Knowledge-Enhanced.** Inspired by how humans process figurative language, we develop RoBERTa-based models enhanced with common-sense knowledge. We develop two models: The first model obtains additional knowledge to better understand the narrative (*context*), while the second seeks knowledge pertaining to the *literal* meaning of the constituents of the figurative expression (Section 2.3). In both cases, in addition to the narrative and candidate continuations, the model is also provided with a set of inferences:  $\{\text{Inf}_1, \dots, \text{Inf}_n\}$  that follow from the narrative, as detailed below and demonstrated in Figure 2.

The literal model uses the COMET model (Hwang et al., 2021), a BART-based language model trained to complete incomplete tuples from ConceptNet. As opposed to extracting knowledge from ConceptNet directly, COMET can generate inferences on demand for any textual input. For an idiom, we retrieve knowledge pertaining to the content words among its constituents,

---

<sup>4</sup>We also experimented with the pattern and verbalizer used by Schick and Schütze (2021) for MultiRC (Khashabi et al., 2018), with the pattern: “ $N. \ Question: \ Based \ on \ the \ previous \ passage \ is \ C_1 \ a \ plausible \ next \ sentence? \ _$ ” and the verbalizer {“yes”, “no”}, but it performed worse.

## ① Knowledge Extraction

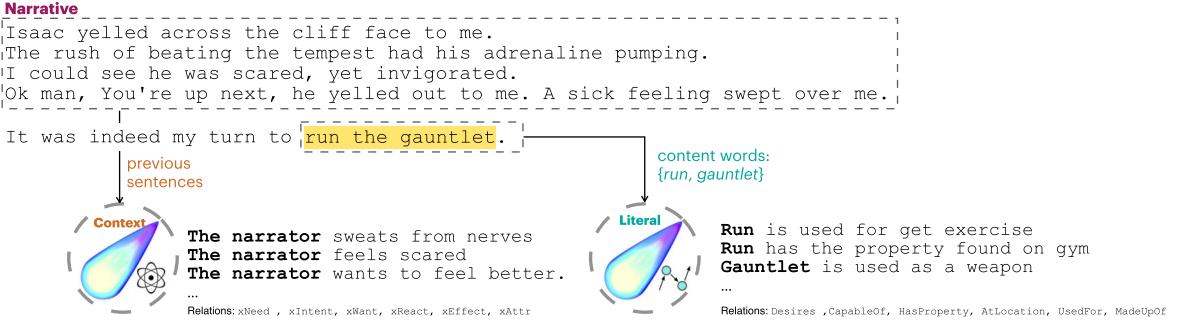


Figure 2: Extracting inferences from COMET regarding the context (previous sentences in the narrative) and the literal meaning of the content words among the idiom constituents.

## ② Knowledge Integration (discriminative)

### Candidate Continuations

- (1) I took a deep breath, convinced it was time, and ran full-speed **away** from the cliff.  
(2) I took a deep breath, convinced it was time, and ran full-speed **towards** the edge of the cliff.

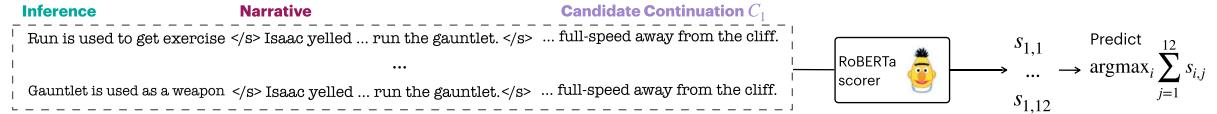


Figure 3: Integrating commonsense inferences into a RoBERTa-based discriminative model.

focusing on the following relations: `UsedFor`, `Desires`, `HasProperty`, `MadeUpOf`, `AtLocation`, and `CapableOf`. For each content word, we extract the top 2 inferences for each relation using beam search. For example, given the idiom “run the gauntlet”, we obtain inferences for “run” and “gauntlet”. We convert the inferences to natural language format based on the templates in Guan et al. (2019). Given the nature of the simile task, we focused solely on the vehicle’s `HasProperty` relation and obtain the top 12 inferences. For example, given the simile “like a psychic whirlpool”, we obtain inferences for the phrase “psychic whirlpool”.

The context model is enhanced with knowledge from ParaCOMET (Gabriel et al., 2021a), trained on ATOMIC. We feed into ParaCOMET all but the last sentence from the narrative, excluding the sentence containing the figurative expression. We generate inferences along ATOMIC dimensions pertaining to the narrator (`PersonX`), namely: `xIntent`, `xNeed`, `xAttr`, `xWant`, `xEffect`, and `xReact`. Again, we extract the top 2 inferences for every relation using beam search.

In both models, as demonstrated in Figure 3, the input format  $X_{i,j}$  for continuation  $C_i$  and inference  $\text{Inf}_j$  is:  $\text{Inf}_j < s /> N < s /> C_i$ .

We compute the score of each of these statements separately, and sum the scores across inferences to get a continuation score:

$$s_i = \sum_{j=1}^{12} s_{i,j} = \sum_{j=1}^{12} \text{scorer}(\text{RoBERTa}(X_{i,j}))$$

where `scorer` is a dropout layer with dropout probability of 0.1 followed by a linear classifier. Finally, the model predicts the continuations with the higher score. We fine-tune the context and literal models for 10 epochs with a learning rate of  $1e-5$  and an effective batch size of 16 for idioms and 64 for similes, and save the best checkpoint based on validation accuracy.

## 4.2 Results

Table 2 shows the performance of all models on the discriminative tasks. For both similes and idioms, supervised models perform substantially better than few-shot and zero-shot models, but still leave a gap of several points of accuracy behind human performance. Human performance is the average accuracy of two native English speakers on the task. We did not provide them with the idiom definition, and we assume they were familiar with the more common idioms. The models performed somewhat better on idioms

<b>Method</b>	<b>Model</b>	<b>Idiom</b>	<b>Simile</b>
Majority		50.0	50.8
Zero-shot	GPT2-XL	53.6	53.7
	GPT3	60.2	62.4
	UnifiedQA	67.7	60.6
Few-shot	GPT3	54.1	51.7
	PET	66.1	55.2
Supervised	RoBERTa	82.0	80.4
	-narrative	65.0	67.9
Knowledge Enhanced	Context Literal	82.8 <b>83.5*</b>	79.9 <b>80.6</b>
Human Performance		<b>92.0</b>	<b>95.0</b>

Table 2: Model performance (accuracy) on the idiom and simile discriminative tasks. \* Difference is significant ( $\alpha < 0.07$ ) between the supervised and knowledge-enhanced models via t-test.

than on similes, possibly due to the LMs’ familiarity with some common idioms as opposed to the novel similes.

Among the zero-shot models, GPT-2 performed worse than GPT-3 and UnifiedQA, each of which performed best on one of the tasks. In particular, UnifiedQA performed well on idioms, likely thanks to its familiarity with the QA format and with the narrative domain.

In the idiom task, PET outperformed few-shot GPT-3 by a large margin of 12 points in accuracy for idioms and 3.5 points for simile, which we conjecture is attributed to the different number of training examples: 6 for GPT-3 vs. 100 for PET. The small number of examples used to prompt GPT-3 is a result of the API limit on the number of tokens (2,048) as well as the setup in which all prompting examples are concatenated as a single input.

Overall, few-shot models performed worse than zero-shot models on both datasets. We conjecture that this is due to two advantages of the zero-shot models. First, the GPT-2 and GPT-3 models performed better than the majority baseline thanks to the similarity between the task (determining which continuation is more plausible) and the language model objective (guessing the next word). Second, the UnifiedQA model performed particularly well thanks to its relevant training. At the same time, both few-shot models had to learn a new task from just a few examples.

The supervised models leave some room for improvement, and the knowledge-enhanced mod-

els narrow the gap for idioms. For similes we see a minor drop in the context model and nearly comparable performance for the literal model.

**Annotation Artifacts.** Human-elicited texts often contain stylistic attributes (e.g., sentiment, lexical choice) that make it easy for models to distinguish correct from incorrect answers without solving the actual task (Schwartz et al., 2017a; Cai et al., 2017; Gururangan et al., 2018; Poliak et al., 2018). Following previous work, we trained a continuation-only baseline, which is a RoBERTa-based supervised model that was trained only on the candidate continuations without the narrative. The results in Table 2 (-narrative) show that the performance is above majority baseline, indicating the existence of *some* bias. However, the performance of this baseline is still substantially worse than the supervised baseline that has access to the full input, with a gap of 17 points for idioms and 12 points for similes, indicating that this bias alone is not enough for solving the task.

### 4.3 Analysis

The knowledge-enhanced models provide various types of inferences corresponding to different relations in ConceptNet and ATOMIC. We are interested in understanding the source of improvements from the knowledge-enhanced models over the supervised baseline, by identifying the relations that were more helpful than others. To that end, we analyze the test examples that were incorrectly predicted by the supervised baseline but correctly predicted by each of the knowledge-enhanced models. We split the examples such that every example consists of a single inference, and feed the following input into the model to predict the plausible continuation: `Inf <s/> N <s/> C`. We focus on the idiom dataset, since for the literal model for similes the only used relation was `HasProperty` and the context model performed slightly worse than the baseline.

Table 3 shows the percents of successful test set predictions for each relation type. The relations in the context model perform similarly, with the best relation `xReact` performing as well as all of the relations (Table 2). In the literal model, it seems that the combination of all relations is beneficial, whereas the best relation, `CapableOf`, performs slightly worse than the full model. For

Literal	Context		
HasProperty	82.3	xNeed	82.2
CapableOf	<b>83.2</b>	xIntent	82.6
Desires	82.5	xWant	82.2
AtLocation	82.7	xReact	<b>82.8</b>
UsedFor	82.4	xEffect	82.5
MadeUpOf	82.8	xAttr	82.5

Table 3: Percents of successful predictions for each relation type for the test set examples.

a narrative snippet “Since Dominic isn’t *up for grabs* anymore, I figure that I will concentrate on something else, Carmen declares”, the inference “grabs is capable of hold on to” was compliant with the meaning of “up for grabs” (available or obtainable), and led to the correct prediction of the plausible continuation “The good news is that there are many other available bachelors out there”. Conversely, the inference corresponding to the Desires relation was “grab desires making money” which was irrelevant and led to an incorrect prediction.

## 5 Generative Task

In the generative task, given a narrative N, the goal is to generate a plausible next sentence that is coherent with the context and consistent with the meaning of the figurative expression. Each instance consists of a reference plausible continuation C.

### 5.1 Methods

We similarly experiment with zero-shot, few-shot, and supervised models.

**Zero-shot.** We use standard LMs, GPT-2 XL and GPT-3, to generate the next sentence following the narrative. We let the models generate up to 20 tokens, stopping when an end of sentence token was generated. Following preliminary experiments, for GPT-2 XL and the rest of the models we use top-k sampling (Fan et al., 2018) as the decoding strategy with  $k = 5$  and a softmax temperature of 0.7, while for GPT-3 we use the method provided in the API which is nucleus sampling (Holtzman et al., 2020) with a cumulative probability of  $p = 0.9$ .

**Few-shot.** We prompt GPT-3 with 4 training examples of the form Q: N A: C followed

by each individual test example, and decode the answer.

**Supervised.** We fine-tune GPT-2 XL with a language model objective for 3 epochs with a batch size of 2. We also trained T5 large (Raffel et al., 2020) and BART large (Lewis et al., 2020) as encoder-decoder models. Both were trained for 5 epochs for idioms and 20 epochs for similes, with an effective batch size of 64. For each model, we kept the best checkpoint based on the validation set perplexity, and used top-k decoding with  $k = 5$  and a temperature of 0.7.

**Knowledge-Enhanced.** We followed the same intuition and inferences we used for the knowledge-enhanced discriminative models (Section 4.1). We fine-tune the models for one epoch as the effective data size is multiplied by the number of inferences per sample. The overall architecture of the generative knowledge-enhanced model is depicted in Figure 4. The models are based on GPT-2 XL and trained with a language model objective to predict the next sentence given the narrative and *a single inference*. The input format for inference  $\text{Inf}_j$  is:  $\text{Inf}_j \text{ <sep1> } N \text{ <sep2>}$ , where  $\text{<sep1>}$  and  $\text{<sep2>}$  are special tokens, and the expected output is the plausible continuation C. During inference, we combine the generations from all inferences pertaining to a given narrative. Inspired by Liu et al. (2021), who ensemble logits from multiple LMs, we ensemble the logits predicted for multiple input prompts using the same model.

A standard decoding process gets at each time step an input prompt text  $x_{<t}$  of length  $t - 1$ . The prompt is encoded and the model outputs the logits for the next ( $t^{th}$ ) token, denoted by  $z_t \in \mathbb{R}^{|V|}$ , where V is the vocabulary. To get a discrete next token,  $z_t$  is normalized and exponentiated to resemble a probability distribution over the vocabulary:  $P(X_t|x_{<t}) = \text{softmax}(z_t)$ , and the next token  $x_t$  is sampled from  $P(X_t|x_{<t})$ . This token is then appended to the prompt and the process iteratively continues until a predefined length or until an end of sentence token had been generated.

Our decoding process differs in that at time step t, we compute the logits  $z_{t,j=1}^{12}$  corresponding to the prompts derived from each of the inferences:  $\text{Inf}_j \text{ <sep1> } N \text{ <sep2>}$  for  $j = 1 \dots 12$ . We sum the logits vectors to obtain  $z_t = \sum_{j=1}^{12} z_{t,j}$ , from which we decode the next token as usual.

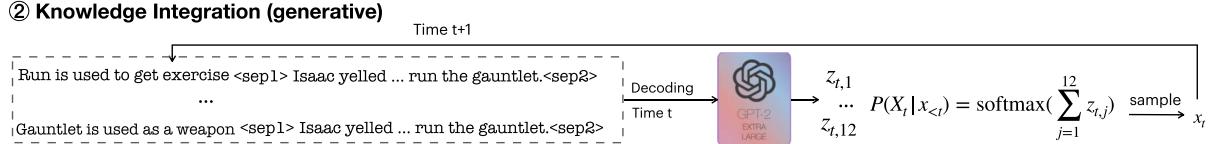


Figure 4: Integrating commonsense inferences into a GPT2-based generative model.

Method	Model	Idiom		Simile	
		R-L	B-S	R-L	B-S
Zero-shot	GPT2-XL	6.2	40.2	17.0	47.7
	GPT3	8.2	33.6	13.9	40.2
Few-shot	GPT3	12.8	51.2	23.1	56.1
	GPT2-XL	<b>15.9</b>	<b>54.2</b>	26.2	59.0
Supervised	T5-large	12.9	51.0	22.9	54.9
	BART-large	12.4	48.8	26.7	58.4
Knowledge	Context	15.4	52.6	20.5	55.1
Enhanced	Literal	13.6	51.4	<b>28.9</b>	<b>59.1</b>

Table 4: Model performance on the generative tasks in terms of automatic metrics. R-L denotes Rouge-L and B-S denotes BERT-Score.

## 5.2 Results

**Automatic Evaluation.** Table 4 shows the performance of all the models on the generative tasks in terms of automatic metrics. We report the performance of the recall-oriented n-gram overlap metric Rouge-L (Lin, 2004), typically used for summarization tasks, and the similarity-based BERT-Score (Zhang et al., 2019). We use the latest implementation to date, which replaces BERT with `deberta-large-mnli`—a DeBERTa model (He et al., 2021) fine-tuned on MNLI (Williams et al., 2018). In terms of automatic evaluation, the best-performing knowledge-enhanced model (context for idioms and literal for similes) performs similarly to the GPT-2 XL supervised baseline, with slight preference to the baseline for idioms and to the knowledge-enhanced model for similes. Both types of supervised models outperform the zero-shot and few-shot models.

**Human Evaluation.** Although automatic metrics provides an estimate of relative model performance, these metrics were often found to have very little correlation with human judgments (Novikova et al., 2017; Krishna et al., 2021). To account for this we also performed human evaluation of the generated texts for a sample of the

Model	Absolute		Comparative	
	Idiom	Simile	Idiom	Simile
GPT2-XL	56	60	15	18.6
+Context	68	68	<b>45</b>	16
+Literal	48	76	13	<b>46.7</b>
Human	<b>80</b>	<b>88</b>	—	—
All	—	—	8	12
Neither	—	—	17	6.7

Table 5: Percent of times that the generation from each of the models and human-written references was chosen as plausible (absolute) or preferred (comparative) by the majority of workers.

test narratives. The human judgments were collected using Amazon Mechanical Turk. Workers were shown a narrative, the meaning of the idiom (or the property of the simile), and a list of 3 generated continuations, one from each of the supervised GPT-2 model, the context model, and the literal model. We performed two types of evaluations. In the absolute evaluation, we randomly sampled 50 narratives for each task, and asked workers to determine for each of the generated continuations along with the human references whether it is plausible or not. In the comparative evaluation, we randomly sampled 100 narratives for idioms and 75 for similes, and presented the workers with a randomly shuffled list of continuations, asking them to choose the most plausible one (or indicate that “neither of the generations were good” or “all are equally good”). In both evaluations, workers were instructed to consider whether the generation is sensible, coherent, follows the narrative, and consistent with the meaning of the figurative expression. Each example was judged by 3 workers and aggregated using majority voting. The inter-annotator agreement was moderate with Krippendorff’s  $\alpha = 0.68$  and  $\alpha = 0.63$  for the absolute and comparative evaluations, respectively (Krippendorff, 2011).

In both absolute and comparative performance, Table 5 shows that for each of the tasks,

Given These Narratives		System Generated Continuations	
	<p>"I ... I'll ask you once more. How much did you make by a fighter taking that dive?" The boy hesitated. "A thousand dollars." Lucky Luciano laughed. "That's <b>chicken feed</b>".</p> <p>She wanted to know if it could do more damage. "It's so hard to say," he answered. "But I honestly don't know how it could do more damage. Maybe it could deflect the attention to him, where it should be." "it's fine," she said, her monotone voice an echo of his sister. She was like <b>a high-definition TV</b>.</p>	<p>He was lucky he made so little <span style="float: right;">✗ GPT-2</span></p> <p>Lucky Luciano made much more money by doing other things such as fixing cars or selling drugs <span style="float: right;">✓+context</span></p> <p>It was hard to tell what she was thinking <span style="float: right;">✗ GPT-2</span></p> <p>Her voice was so smooth, you could hear every word she said <span style="float: right;">✓+literal</span></p>	
			

Figure 5: Narratives ending in an idiom (top) or a simile (bottom) with the continuations generated by the baseline GPT-2 model and a knowledge-enhanced model, as preferred by human judges.

Model	Category	Example
Literal (Simile)	①	Is that how you got those bruises on your face? They colored her pure white skin <b>like a sloppily designed ink blot test</b> <i>It was as if her skin was smudged with a brownie</i> .
	②	We stand there munching next to the evening sun-swept grass and whispering trees of Duboce Park, where a bunch of happy-ass dogs are sprinting around in circles while their owners stand there <b>like a scarecrow convention</b> . <i>The humans were having a great time, jumping and barking and running around</i> .
Context (Idiom)	①	She could hold some of her thoughts while she was a wolf. Better than the rest of us. Not as well as Sam, but better than me. She and Derrick were <b>thick as thieves</b> <i>They would always be looking for opportunities to steal</i>
	③	That's why I packed the heat. Just in case I needed some motivation for him to calm down should he be inclined to <b>go ballistic</b> on me. : <i>because I was thinking of ways to solve this problem in a peaceful way</i>

Table 6: An example for each error category. Each example consists of a narrative, with the figurative expression in bold and the continuation in italic.

a knowledge-enhanced model outperformed the baseline GPT-2 model. What makes a more compelling case is that the context model was favored for idioms while the literal model was favored for similes, complying with prior theoretical grounding on these figurative language types. Figure 5 shows examples generated by the baseline and the best model for each task. We note that 80% of the human-written continuations for idioms and 88% of those in the simile task were judged as plausible. Based on our analysis, the gap from 100% may be explained by the ambiguity of the narratives that leaves room for subjective interpretation.

### 5.3 Error Analysis

We analyze the continuations labeled as implausible by the annotators, for the best model in each task: context for idioms and literal for similes. We found the following error categories, with percent details in Table 7 and exemplified in Table 6:

Cat.	Literal (Simile)	Context (Idioms)
①	50	72
②	33	14
③	17	14

Table 7: Error categories along with their proportion (in %) among the implausible continuations.

**① Inconsistent with the figurative expression:** The continuation is inconsistent or contradictory to the figurative expression. For instance, the simile in the first row in Table 6 is ‘like a sloppily designed ink blot test’, for which the property of comparison is “a pattern of dark blue, purple, and black”, but the generated continuation mentions brownie, which has a *brown* color. Similarly for the idiom “thick as thieves” the model generates a literal continuation without

understanding its actual meaning “closest of friends”.

**② Inconsistent with the narrative:** The continuation is inconsistent or contradictory to the flow of the narrative. For instance, the narrative in the second row in Table 6 states that “the owners who are humans are *standing*”, while the continuation states they are jumping. The model further predicts that the *humans* are barking, instead of the *dogs*. In general, across multiple examples we have found that models tend to confuse the various characters in the narrative.

**③ Spelling or grammar errors:** Some generations contained spelling mistakes or introduced grammar errors such as starting with a punctuation or having extra blank spaces. Although we instructed the crowdsourcing workers to ignore such errors, they may have affected their plausibility judgments.

## 6 Conclusion

We introduced a narrative understanding benchmark focused on interpreting figurative language, specifically idioms and similes. Following the story cloze test, we designed discriminative and generative tasks with the goal of continuing a narrative. We found that pre-trained LMs irrespective of their size struggle to perform well in zero-shot and few-shot setting, and that the supervised models while competitive are still behind human performance by a significant margin. We further bridged some of this gap with knowledge-enhanced models that are inspired by the way humans interpret figurative expressions. Our analysis reassessed known findings that although LMs generate grammatical human-like texts, they are often inconsistent and the model’s ability to distinguish characters in a story is limited. We hope this work will spark additional interest in the research community to further advance the representations and modeling of figurative language, which is too common to ignore.

## Acknowledgment

This research was supported in part by DARPA under the MCS program through NIWC Pacific (N66001-19-2- 4031), Google Cloud computing, and the Allen Institute for AI (AI2).

## References

- Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. *arXiv preprint arXiv:2009.00829*.
- Fatemeh Mohamadi Asl. 2013. The impact of context on learning idioms in EFL classes. *TESOL Journal*, 37(1):2.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1470>
- 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*.
- Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622, Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-2097>
- Robyn Carston and Catherine Wearing. 2011. Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, 3(2):283–312. <https://doi.org/10.1515/langcog.2011.010>
- Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020a. R<sup>3</sup>: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In

- Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7976–7986, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.711>
- Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021a. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-acl.297>
- Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020b. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.524>
- Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. MER-MAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.336>
- Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics. <https://doi.org/10.3115/1690219.1690231>
- Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1168>
- Thomas C. Cooper. 1999. Processing of idioms by L2 learners of english. *TESOL Quarterly*, 33(2):233–262. <https://doi.org/10.2307/3587719>
- Yiming Cui, Wanxiang Che, Wei-Nan Zhang, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Discriminative sentence modeling for story ending prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7602–7609. <https://doi.org/10.1609/aaai.v34i05.6260>
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103. <https://doi.org/10.1162/coli.08-010-R1-07-048>
- Susan R. Fussell and Mallie M. Moss. 1998. Figurative language in emotional communication. *Social and Cognitive Approaches to Interpersonal Communication*, pages 113–141.
- Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021a. Paragraph-level common-sense transformers with recurrent memory. In *AAAI*.
- Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021b. Discourse understanding and factual consistency in abstractive summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-main.34>
- Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio.

2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.212>
- Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480. <https://doi.org/10.1609/aaai.v33i01.33016473>
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-2017>
- Patrick Hanks. 2013. *Lexical analysis: Norms and exploitations*. MIT Press. <https://doi.org/10.7551/mitpress/9780262018579.001.0001>
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1023>
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.171>
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328. [https://doi.org/10.1162/tacl\\_a\\_00023](https://doi.org/10.1162/tacl_a_00023)
- Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Computing*, 1:25–2011.
- Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.393>
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self supervised learning of language representations. <https://arxiv.org/abs/1909.11942v3>.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation,

- translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323, Singapore. Association for Computational Linguistics. <https://doi.org/10.3115/1699510.1699552>
- Zhongyang Li, Xiao Ding, and Ting Liu. 2019. Story ending prediction by transferable BERT. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1800–1806. International Joint Conferences on Artificial Intelligence Organization.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.522>
- Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.
- Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Brussels, Belgium. Association for Computational Linguistics.
- Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.
- 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*.
- Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? You probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.
- 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. <https://doi.org/10.18653/v1/N16-1098>
- Suzanne Mpouli. 2017. Annotating similes in literary texts. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.
- Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold:

- Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018, Doha, Qatar. Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1215>
- Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1238>
- Anthony M. Paul, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 1970. Figurative language. In *Philosophy & Rhetoric*, pages 225–248.
- Jing Peng and Anna Feldman. 2016. Experiments in idiom recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2752–2761, Osaka, Japan. The COLING 2016 Organizing Committee.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/S18-2023>
- Ashequl Qadir, Ellen Riloff, and Marilyn Walker. 2015. Learning to recognize affective polarity in similes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 190–200, Lisbon, Portugal. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D15-1019>
- Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2016. Automatically inferring implicit properties in similes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1223–1232, San Diego, California. Association for Computational Linguistics.
- <https://doi.org/10.18653/v1/N16-1146>
- Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. In *Language models are unsupervised multitask learners*.
- 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*, 21:1–67.
- Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.
- Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163. <https://doi.org/10.1111/j.1467-9280.1994.tb00653.x>
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer. [https://doi.org/10.1007/3-540-45715-1\\_1](https://doi.org/10.1007/3-540-45715-1_1)
- Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P16-1019>
- 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035. <https://doi.org/10.1609/aaai.v33i01.33013027>

- Timo Schick and Hinrich Schütze. 2021. It’s not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.185>
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017a. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-1004>
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017b. Story cloze task: UW NLP system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55, Valencia, Spain. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-0907>
- Ekaterina V. Shutova. 2011. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory. [https://doi.org/10.1162/tacl\\_a\\_00277](https://doi.org/10.1162/tacl_a_00277)
- Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419. <https://doi.org/10.18653/v1/2020.emnlp-main.373>
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI Conference on Artificial Intelligence*.
- Siddarth Srinivasan, Richa Arora, and Mark Riedl. 2018. A simple and effective approach to the story cloze test. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 92–96, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-2015>
- Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 681–687, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1101>
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- P. Wolff and D. Gentner. 2000. Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 26 2:529–41.

- Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.
- Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9515–9522. <https://doi.org/10.1609/aaai.v34i05.6496>
- Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Ya-Long Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and a neural approach. In *AAAI*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.
- Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.mwe-1.5>
- Yukun Zhu, Jamie Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27. <https://doi.org/10.1109/ICCV.2015.11>