Identifying Physical Object Use in Sentences

Tianyu Jiang and **Ellen Riloff**
School of Computing
University of Utah
Salt Lake City, UT 84112
{tianyu, riloff}@cs.utah.edu

Abstract

Commonsense knowledge about the typical functions of physical objects allows people to make inferences during sentence understanding. For example, we infer that *"Sam enjoyed the book"* means that Sam enjoyed *reading* the book, even though the action is implicit. Prior research has focused on learning the prototypical functions of physical objects in order to enable inferences about implicit actions. But many sentences refer to objects even when they are not used (e.g., *"The book fell"*). We argue that NLP systems need to recognize *whether* an object is being used before inferring *how* the object is used. We define a new task called *Object Use Classification* that determines whether a physical object mentioned in a sentence was used or likely will be used. We introduce a new dataset for this task and present a classification model that exploits data augmentation methods and FrameNet when fine-tuning a pre-trained language model. We also show that object use classification combined with knowledge about the prototypical functions of objects has the potential to yield very good inferences about implicit and anticipated actions.

1 Introduction

Physical objects play an important role in daily life. People use them for different purposes, for example we use knives for cutting, cars for transportation, and books for reading. Recent work by Jiang and Riloff (2021b) argued that most human-made physical artifacts were created for a specific purpose, and that commonsense knowledge about an object's *prototypical function* is essential for natural language understanding. For example, *"she finished the puzzle"* and *"she finished the cigarette"* implicitly refer to different actions associated with puzzles (solving) vs. cigarettes (smoking). Similarly, humans interpret *"he used a gun"* as a shooting but *"he used a knife"* as a stabbing based on our knowledge of guns and knives.

Jiang and Riloff (2021b) developed a method to learn the prototypical functions for physical artifacts from text corpora, with the goal of producing a commonsense knowledge resource for physical objects. However an open question is how to apply this knowledge for sentence understanding. It would be risky to assume that objects are always used in the most typical way because objects can be used in atypical ways too. For example, *"Max used the knife to open the bottle"* probably means that Max popped the top off the bottle with the knife, not that Max cut the bottle. But as we will discuss in Section 4, our study found that physical artifacts are used in the prototypical way most of the time (96%), so it is a very good assumption.

A much bigger problem for applying knowledge of prototypical functions is that physical objects are often mentioned when they are not used at all! For example, the sentences below mention a knife, but the knife is not being used:

(a) He put the knife in the dishwasher.
(b) She found a knife in the woods.
(c) The knife fell off the table.
(d) A good pocket knife costs $100.

In addition, some sentences suggest that a physical object *will be* used, although it has not been used yet. For example, Mary aims to acquire a knife through various actions in the sentences below:

(e) Mary got a knife from the drawer.
(f) Mary asked John for a knife.
(g) Mary purchased a chef's knife.

When reading these sentences, people naturally infer that Mary intends to use the knife, most likely in the typical way (i.e., to cut things). We believe that NLP systems should also make these predictive inferences to "read between the lines" during narrative text understanding. For example, consider the sentence *"The fish was too big for the freezer,*

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11362 - 11372
December 7-11, 2022 ©2022 Association for Computational Linguistics

so Mary got a knife.”. Human readers would assume that Mary used the knife to cut the fish into smaller pieces, even without any explicit mention of cutting.

We propose a new NLP task, *Object Use Classification*, to classify the usage status of physical objects mentioned in sentences with respect to three categories: *Used*, *Anticipated Use*, and *No Use*. Our first goal is to identify sentences that state or imply that an object was used (*Used*) to enable prototypical function inferences when the action is implicit. Our second goal is to identify sentences that describe actions which suggest someone’s probable intent to use the object (*Anticipated Use*) to enable second-order prototypical function inferences. Finally, identifying sentences where there is no use of a mentioned object (*No Use*) is important to recognize when prototypical function inferences should not be applied. We introduce a new object use dataset for this task with gold standard human annotations of sentences that mention physical objects. We found that all three use categories are common: our annotators labeled 45% of the sentences as *Used*, 28% as *Anticipated Use*, and 27% as *No Use*.

We explored several methods to tackle this task. First we applied prompting methods using two large language models to evaluate a zero-shot generalization approach, but the results were mediocre. Next, we fine-tuned a transformer-based model, which yielded much better performance. Finally, we added two data augmentation techniques, synonym replacement and back translation, and also provided exemplar sentences associated with the object’s prototypical function frame. Our experimental results show that the complete model achieves good performance for this task.

2 Related Work

Commonsense knowledge has long been recognized as an essential part of natural language understanding (Charniak, 1972; Woods, 1975; Schank and Abelson, 1977). Some work specifically argued that commonsense knowledge about physical objects is often used to make inferences and plays an important role in narrative text understanding (Burstein, 1979).

Recently, a variety of projects have focused on acquiring knowledge about physical objects, including relative physical knowledge (Forbes and Choi, 2017), relative spatial relations (Collell et al., 2018), location knowledge (Jiang and Riloff, 2018; Xu et al., 2018), and object affordance (Persiani and Hellström, 2019). Jiang and Riloff (2021b) developed a method to learn the most typical way that people use human-made physical artifacts, and they used FrameNet frames as a representation for common object functions. For their work, they created a dataset of physical objects annotated with their prototypical functions. Our research builds upon that work by developing a model for identifying the usage status of physical objects mentioned in a sentence, which we argue is a necessary precursor to applying prior knowledge about prototypical functions.

Recently, there have been efforts aimed at learning implicit information with pre-trained language models. Weir et al. (2020) explored using pre-trained masked language models to capture implicit knowledge elicited from humans, which are so-called stereotypical tacit assumptions. Geva et al. (2021) created a question answering dataset consisting of questions that require implicit multi-step reasoning skills, such as *“Did Aristotle Use a Laptop?”*. They show that a large language model fine-tuned on related datasets without retrieval of relevant knowledge performs far worse than humans, and high-quality retrieval makes the model more effective in the reasoning process. Talmor et al. (2020) trained language models with automatically generated data sampled from existing knowledge sources. They show that language models can combine implicit knowledge encoded in their parameters with explicit rules and facts, and further perform reasoning.

Our work also has ties to frame semantics (Fillmore, 1976, 1982), a theory of how we associate words and phrases with cognitive structures called frames. Frame semantic parsing (Baker et al., 2007; Swayamdipta et al., 2018) is the task of automatically extracting frame semantic structures from sentences, based on the Berkeley FrameNet project (Baker et al., 1998; Ruppenhofer et al., 2006). The process begins with identifying frame-evoking words in the sentence (*target identification*) and identifying the evoked frame for each target (*frame identification* (Botschen et al., 2018; Jiang and Riloff, 2021a)). However, one limitation of this setting is that it typically relies on the predicate to predict the frame. For example, the sentence *“Sam enjoyed the book”* would not trigger a reading frame because “enjoy” is not associated with reading in FrameNet. Our work strives to identify this implicit action by recognizing that the book was used and then applying the prototypical function associated with books.

3 Motivation

Actions involving physical objects are often left implicit in natural language. In many cases, these actions do not need to be explicitly stated because they can be easily inferred by people using our knowledge about physical objects. Early NLP research recognized this need for commonsense knowledge about physical objects (e.g., Burstein (1979)) and some efforts have been undertaken to compile such knowledge, including ConceptNet (Speer et al., 2017), which contains a "UsedFor" relation that captures possible uses for an object expressed in natural language, and recent work by Jiang and Riloff (2021b) that learns to associate physical objects with FrameNet frames describing their prototypical uses.

However, a crucial question is when to apply this knowledge in sentence understanding. In this paper, we claim that NLP systems must be able to distinguish between (1) sentences that mention a physical object and state or imply that the object was or will be used, and (2) sentences that mention a physical object but the object was not used. For example, the sentences *"Mary read the book"* and *"Mary enjoyed the book"* both imply that the book was used (read), but *"Mary dropped the book"* does not mean that the book was used, only that Mary was carrying it. We found that about 73% of sentences that mention a physical object in our data set (see Section 4) suggest that the object was (45%) or will be (28%) used. For the other 27% of sentences that mention a physical object, the object was not used at all. Consequently, we argue that an important task for understanding sentences about physical objects is *object use classification*.

A second question relates to the applicability of "prototypical" functions for objects: when an object *is* used, how often is it used in the prototypical way? In our dataset (Section 4), we found that when a sentence mentions or implies the use of an object, 96% of these sentences correspond to the prototypical use for the object. Only 4% of these sentences suggest that an object was used in an atypical way. These results indicate that an effective object use classifier can go a long way toward enabling NLP systems not only to infer *whether* an object was used, but also *how* an object was used, even when that action is not explicitly stated.

In this paper, we tackle the problem of object use classification and define three categories of object use: *Used*, *Anticipated Use*, and *No Use*. In the next section, we define these three categories and explain our motivation for them, and we present a new object use dataset for this task.

4 Dataset Creation

Since we are tackling a new task, we created a new *TOUCAN* (**T**extual **O**bject **U**se **C**l**A**ssificatio**N**) dataset[1] with gold standard human annotations. Our primary goal was to obtain human judgements for sentences that mention a physical object indicating whether the object was or will be used, or not. But a second goal was to better understand how often objects are used in a prototypical way, as opposed to an atypical way. So we obtained additional human judgements for the sentences in which an object was or will be used and asked the annotators to determine whether the use corresponds to the object's prototypical function. We leveraged the results of prior work that studied physical objects and their prototypical functions so as not to reinvent the wheel.

Physical Objects: We use the list of physical objects produced by Jiang and Riloff (2021b). They extracted human-made physical objects with sense definitions from WordNet (Miller, 1995), and used a concreteness dictionary (Brysbaert et al., 2014) to filter out abstract terms. Their freely available dataset[2] contains 938 human-made physical object terms.

Sentences: Then we extracted sentences containing these physical objects from the Spinn3r corpus (Burton et al., 2009), which consists of 44 million blog posts. In order to get a uniform distribution of different physical objects, we randomly sampled 4 sentences (or fewer if there are not enough) for each physical object. This produced a set of 2,460 sentences in total.

4.1 Human Annotation

4.1.1 Object Use Categories

First, we presented two people with a physical object term, a sentence that mentions the object, and

[1] The dataset can be found at: https://github.com/tyjiangU/toucan

[2] https://github.com/tyjiangU/physical_artifacts_function

Sentence	Use Category	Prototypical Function
(1) We took a **speedboat** up the river to the village. (2) He promptly walked over to his **mattress** and laid down. (3) I had sausage slices wrapped around olives, held together with a **toothpick**.	Used	Self_motion ✓ Sleep ✓ Removing ✗
(4) I quickly went to the bathroom and got more **ammo**. (5) All my new **cookware** will be put to use with these new recipes! (6) I got measured for my **tuxedo** for Dad 's wedding today.	Anticipated Use	Cause_harm ✓ Cooking_creation ✓ Wearing ✓
(7) Nope , it also hit my left **headlight** and broke it. (8) At one point , I saw a high heel **shoe**. (9) I promptly threw the **brochure** in a corner to collect dust.	No Use	- - -

Table 1: A sample of the annotated examples. The physical objects are marked in **red**. The second column shows the annotated use category. The third column shows the gold prototypical function frame for the object followed by ✓or ✗. The ✓means the frame is consistent with the use of the object in the sentence, otherwise ✗.

the WordNet definition of the object.[3] We asked the annotators to select one of these four categories:

Used: The sentence describes 1) an action in which the object is/was being used (by the writer or someone else), or 2) an action that directly resulted from the use of the object.

Anticipated Use: The sentence states that 1) the object will be used in the future, or 2) implies that someone will presumably use the object.

No Use: Neither *Used* nor *Anticipated Use*.

Wrong Sense: The given definition of the object term is different from its meaning in the sentence. (This option was provided to flag sentences in which the term's meaning is not its physical object sense. We do not include these sentences in our dataset.)

Table 1 shows some annotated examples. Sentences 1-3 are the *Used* cases, 4-6 show objects that have an *Anticipated Use*, and 7-9 are *No Use* examples.

4.1.2 Prototypical Use Annotations

Jiang and Riloff (2021b) proposed that most human-made physical artifacts have a prototypical function (i.e., the intended purpose of the object). They selected 42 frames from Framenet v1.7 (Ruppenhofer et al., 2006) to represent actions that are common functions of physical artifacts. Table 2 shows a few physical objects and their prototypical function frames.

To better understand how often objects are used in a prototypical way as opposed to an atypical way, we collected additional human judgements.

Frame	Physical Objects
Wearing	*hat*, *shirt*
Containing	*basket*, *luggage*
Self_motion	*bicycle*, *yacht*
Protecting	*armor*, *helmet*
Supporting	*chair*, *scaffolding*

Table 2: Objects and their prototypical function frames.

If an annotator selected *Used* or *Anticipated Use*, we also asked the annotator whether the use of the object most likely corresponds to its prototypical function (based on the gold frame in Jiang & Riloff's dataset). The annotator was shown the prototypical function frame for the object, and asked to select *Yes* or *No* as to whether the frame correctly characterizes the use of the object in the sentence. The last column in Table 1 shows the prototypical function frames, followed by the annotated *Yes* (✓) or *No* (✗). For example, in sentence (3) of Table 1, a toothpick is typically used to remove food that is stuck between our teeth, but here it is used to hold sausage and olives together so this sentence represents an atypical use for a toothpick.[4]

To prepare the annotators, we provided them with detailed annotation guidelines to familiarize them with the task. Deciding whether the prototypical function frame is correct requires knowledge of FrameNet frames, so we also asked them to read FrameNet's definitions and exemplar sentences for each relevant frame. We randomly shuffled the physical objects before presenting them to the annotators. The pairwise inter-annotator agreement

[3] We manually identified the WordNet definition corresponding to the physical object sense of the term.

[4] One could argue that toothpicks have multiple common functions, but Jiang & Riloff defined only one prototypical function for each object in their work.

using Cohen's kappa for the 3 object use categories was 0.71, and the simple agreement rate (percentage of agreement) for the Yes/No prototypical function question was 0.92.

To create the final set of gold standard labels, we had the annotators adjudicate their disagreements. This process produced 2,123 sentences annotated with one of the *Used*, *Anticipated Use* or *No Use* label. The **All Cases** row of Table 3 shows the distribution of the 3 categories. The Used category accounts for 45% of the sentences, with Anticipated Use and No Use each making up about 27% of the data.

The **Prototypical Use** row shows the prototypical use results. The annotators determined that objects were used in their most prototypical way in 97% (935/964) of the Used sentences and in 96% (560/583) of the Anticipated Use sentences. This data suggests that if we had a perfect object use classifier, we could infer *how* an object was used with 96.6% accuracy simply by assuming its prototypical function.

	Used	**Anticipated Use**	**No Use**
All Cases	964	583	576
Prototypical Use	935	560	-

Table 3: Annotated data statistics.

5 Object Use Classification Models

We explored several approaches to tackle this task. We first present a transformer-based model fine-tuned solely on our gold standard training sentences. Then we present a method that takes advantage of two commonly used data augmentation techniques, synonym replacement and back translation. Finally, we also show that our model further benefits from prior knowledge of the object's prototypical function by incorporating the exemplar sentences associated with its function frame.

5.1 Task Definition and Base Model

We model our task as a 3-class classification problem. Given a sentence, and an object mentioned in the sentence, the task is to determine if the object has been used, has an anticipated use in the future, or has no stated or implied use.

We build our model based on RoBERTa (Liu et al., 2019). For our base model, we use the sentence as the input sequence into the model, and use the last hidden vector representing the object (if there is more than one token, we compute the average of all tokens) as output, and then pass it through a linear classifier to predict the label.

5.2 Synonym & Hyponym Replacement

Our physical object list originated from WordNet. To increase the number of training sentences, we created copies of each original training sentence where the object term is replaced by one of its synonyms or hyponyms in WordNet. Specifically, for each object that has a WordNet synset, we first extract all the lemmas belonging to the same synset and also traverse one level down in WordNet's hierarchy to extract the lemmas of its direct hyponyms. For example, the furniture term *sofa* belongs to the synset `sofa.n.01`, which also contains *couch* and *lounge*. Suppose its direct hyponyms are *daybed*, *divan*, and *loveseat*. Then we have a list of 5 new object terms. For each training sentence that mentions a *sofa*, we replace the word *sofa* with its synonyms or hyponyms, generating 5 new training examples with the same label. In general, we use all of the synonyms and up to 5 hyponyms (if there are more than five then we randomly select five).

5.3 Back Translation

Back translation (Sennrich et al., 2016) is a widely used data augmentation technique, which automatically generates new training examples by translating a sentence to another language and then translating it back, aiming to produce diverse paraphrases of the original sentence. Its effectiveness has been shown for downstream tasks such as text classification (Xie et al., 2020) and question answering (Longpre et al., 2019).

Table 3 shows that the label distribution in our dataset is roughly 2:1:1. Though this imbalance reflects the actual distribution of these categories, we hypothesized that the model would perform better with a more balanced distribution of class labels. So we performed back translation on the training examples labeled with *Anticipated Use* and *No Use*, augmenting these categories to be roughly the same size as the *Used* category. For back translation, we use the Helsinki-NLP English to Chinese and Chinese to English transformer-based machine translation system (Tiedemann, 2020).

5.4 FrameNet Exemplars

Our annotation results suggested that when objects are used, they are almost always used in the

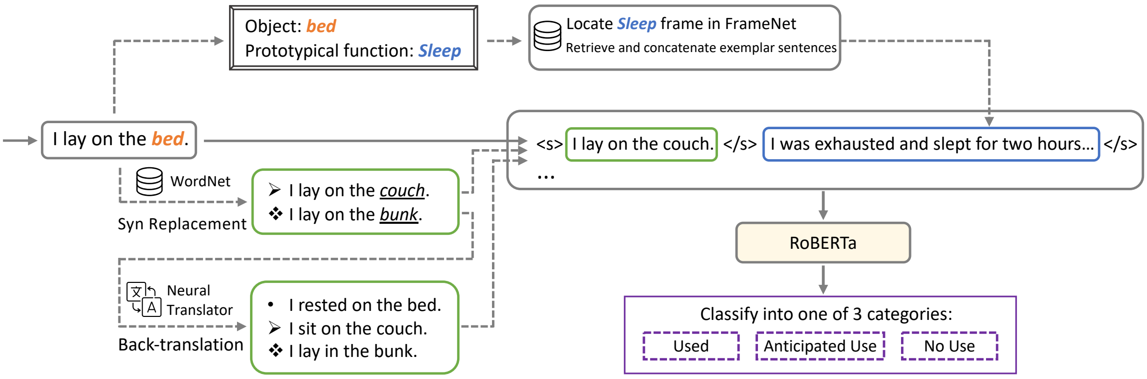

Figure 1: An illustration of the TOUCAN object use classification model.

prototypical way. So for each object, we utilize the prototypical function frame in the gold standard dataset produced by Jiang and Riloff (2021b) as prior knowledge to recognize whether the sentence describes a relevant situation. Specifically, we extract the frame's *exemplar sentences* from FrameNet, which are the annotated sentences associated with a lexical unit that triggers the frame. For example, the *Sleep* frame contains examplar sentences such as *"I was exhausted, and slept for two hours"*, *"Well, better get some shut-eye"*, etc. We concatenate all of the exemplar sentences for the frame as one sequence, then pair it with the sentence containing this object as the input to the RoBERTa model.

5.5 Complete Model Architecture

Figure 1 shows an overview of the full architecture for our learning process. For simplicity, we call the model TOUCAN as well. First we use synonym/hyponym replacement to augment the original training set. Then we apply back translation to all of the sentences labeled as *Anticipated Use* or *No Use* to generate more sentences, and add them to the training set. When applying back translation, for each sentence, we generate only one new sentence from the translator. Finally, for each sentence in the training set, we extract the exemplar sentences from FrameNet corresponding to the object's prototypical function frame. The exemplar sentences are concatenated and given to RoBERTa along with the original sentence as input. Then we send the last hidden vector of the object into a linear classifier on top of the RoBERTa model.[5] If there are multiple tokens, we compute their average.

6 Evaluation

We split our gold standard data set into roughly 70% for training, 15% for development and 15% for testing. We also made sure that the objects in the test set do not appear in the training set. Our fine-tuning framework is based on the RoBERTa-base model (Liu et al., 2019). For the hyperparameters, we used a max sequence length of 192, a batch size of 8, learning rate initialized as 2e-5, and train for 15 epochs. Each result is averaged over three runs with different random seeds. We report overall accuracy as well as precision, recall and F1 scores macro-averaged over the 3 classes.

6.1 Prompting Baseline

Recent advances in pre-trained language models have demonstrated their ability to attain zero-shot generalization on different downstream tasks (Brown et al., 2020). Specifically, prompting has become a widely used technique in natural language processing. It works by recasting NLP tasks in the form of a natural language response to a natural language input. To see how well this approach can work for our task, we explore prompting with two language models: GPT-2 (Radford et al., 2019) and T0++ (Sanh et al., 2021). T0++ is an encoder-decoder model that has been trained on a collection of downstream tasks such as question answering and summarization, with multiple prompts per dataset. We cast our problem as a textual entailment task and use the same set of prompts in (Sanh et al., 2021). For example, one template is:

```
Suppose [premise]. Can we infer that
"[hypothesis]"? Yes, or no?
```

Since our task is to distinguish between three different categories (*Used*, *Anticipated Use*, and *No Use*), we created a two-template pipeline to obtain the prediction. As an example, consider the

[5] In rare cases, the object no longer exists in the sentence after back translation. In this case we use the vector of the first token in the sentence.

	Use			Anticipated Use			No Use		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
T0++	64.9	40.2	48.4	**73.6**	23.3	34.4	34.6	**76.3**	47.2
TOUCAN$_{base}$	71.8	74.1	72.9	57.1	72.2	63.7	65.5	46.0	54.1
+Synonyms	72.9	78.4	75.5	64.6	72.2	68.2	70.6	54.0	61.2
+BackTrans	**76.6**	82.1	79.3	62.8	**77.0**	69.1	**75.5**	50.7	60.6
+Syn&BackTrans	74.0	80.6	77.2	67.2	73.8	**70.3**	73.3	55.8	63.3
+Exemplar	75.8	**84.1**	**79.8**	61.0	71.0	65.6	71.2	47.5	56.9
TOUCAN	75.1	80.8	77.9	66.1	74.2	69.9	74.8	56.9	**64.6**

Table 5: Results breakdown for each label.

Model	Acc	Pre	Rec	F1
random	31.0	30.5	30.2	30.0
GPT-2	30.1	24.6	31.9	18.5
T0++	46.0	57.7	46.6	43.3
TOUCAN$_{base}$	65.8	64.8	64.1	63.6
+Synonyms	70.0	69.4	68.2	68.3
+BackTrans	72.0	71.6	69.9	69.7
+Syn&BackTrans	71.9	71.5	70.1	70.3
+Exemplar	70.5	69.4	67.5	67.4
TOUCAN	**72.4**	**72.0**	**70.6**	**70.8**

Table 4: Object use results across models.

sentence *"John finished the watermelon with the spoon"*, where the spoon is the object in question. Two templates would be generated:

T1: `Suppose John finished the watermelon with the spoon. Can we infer that "The spoon has been used"? Yes, or no?`

T2: `Suppose John finished the watermelon with the spoon. Can we infer that "The spoon will be used in the future"? Yes, or no?`

If the output for T1 is *Yes*, it means the prediction is *Used*. If the output for T1 is *No* but for T2 is *Yes*, it means the prediction is *Anticipated Use*. Otherwise the prediction is *No Use*. Since the T0++ model has been fine-tuned with the prompt templates, it will always predict *Yes* or *No* as the output. However the GPT-2 model can predict other tokens as the next word. So for GPT-2, we compare the probability score for the *Yes* and *No* tokens and choose the one that is higher. We report results averaged over all templates.

6.2 Results

Table 4 shows our experimental results. As a baseline, the first row shows that random labeling (assigning each label with the probability of 1/3) achieves 30.0% F1. The next two rows show the results for the prompt-based methods. The GPT-2 model predicts *No* much more frequently than *Yes* and most predictions fall into the *No Use* category. It very rarely predicts *Anticipated Use*. This produces a low F1 score of 18.5%. T0++ also suffers from low recall for *Anticipated Use*, but it is substantially better than GPT-2, achieving 46.0% accuracy and 43.3% F1 score.

The next section of Table 4 shows the results for our fine-tuned models. Using the sentence alone (TOUCAN$_{base}$) achieves 65.8% accuracy and 63.6% F1. Each of the following rows adds one new component to the architecture to evaluate its contribution independently (not cumulatively). The *+Synonyms* and *+BackTrans* rows show results for data augmentation using synonym/hyponym replacement and back translation respectively on top of the TOUCAN$_{base}$ model. We see that synonym/hyponym replacement increases the F1 score to 68.3%, and back translation performs even better at 69.7%. When using both synonym/hyponym replacement and back translation (row *+Syn&BackTrans*), the model achieves over 70% F1 score.

The *+Exemplar* row shows the results when giving the sentence as well as FrameNet's exemplar sentences as a sequence pair to RoBERTa. Note that this model requires gold information about an object's prototypical function. Compared to TOUCAN$_{base}$, adding the exemplar sentences increases the F1 score from 63.6% to 67.4%. The last row (TOUCAN) shows the results when combining all of the elements together, which yields the highest accuracy score of 72.4% and highest F1 score of 70.8%.

Table 5 shows the performance breakdown for

Sentences
i. The couch was crammed under the window with the tv in the corner.
ii. Today I dropped my spectacles in the dog kennel again while getting my crazy dog out.
iii. She laid out the smock on the wardrobe and moved over to me.
iv. I am now debating taking the cabinet back to Target and exchanging it.
v. Duo took a step back and leaned against the workbench.

Table 6: A sample of *No Use* cases that were predicted incorrectly by the system. Our TOUCAN model predicted *Used* for i., ii., and iii., and *Anticipated Use* for iv. and v.

Implicit Type	Count
Underspecified verb	16
Light verb	16
Metonymic verb	4
Prepositional phrase	9
Misc	13

Table 7: Manual analysis on how common implicit actions are in a random sample of 200 sentences.

Explicit vs. Implicit Frames
1) She had this spunky , schoolgirl-theme outfit complete with ammo backpack and skirt. had → Possession outfit → Wearing
2) I grabbed my 7x50 binoculars but the coyote has run away. grabbed → Manipulation binoculars → Perception_exp
3) You stand on a street with a guitar and a crowd will come. stand → Posture guitar → Make_noise
4) When the aids came in and said she had to use the bedpan, she threw a fit. use → Using bedpan → Excreting

Table 8: The predicate (target) for frame identification is red. The physical objects are blue. The red frame represents the action explicitly indicated by the predicate. The blue frame represents the prototypical action associated with the object, which people would infer.

each label. Here we only show T0++ for comparison with the fine-tuned models. A clear difference between T0++ and the fine-tuned models is that T0++ labels far too many instances as *No Use*. The fine-tuned models do a much better job at distinguishing the 3 classes. We can also see that both data augmentation methods help improve recall, especially for the *Used* and *No Use* categories.

6.3 Analysis

Performance on the *No Use* category is lower than on the other categories, so we did some manual investigation to better understand why. Table 6 shows some *No Use* examples that were incorrectly labeled by TOUCAN. We see some clues that seem potentially useful, such as prepositional phrases indicating that the object is not the main focus (e.g., under the window, against the workbench). And "dropped" implies that the object was passive (i.e., something happened to it). But we saw many different types of *No Use* contexts. Focusing on this category could be an interesting direction for future research.

We also conducted a manual analysis to see how common truly implicit actions are. We randomly sampled 200 examples from the *Used* or *Anticipated Use* sentences in our dataset and judged whether the main predicate explicitly described the action involving the object. This was an informal study, but we judged nearly 30% (58) as having implicit actions. Table 8 shows some sentences with implicit actions. We noticed a few common categories and showed their frequencies in Table 7.

In 16 sentences, the main verb was underspecified, such as "use". There were 16 light verb constructions (Tu and Roth, 2011), in which the verb has little semantic content of its own. There were 4 metonymic verbs (Lapata and Lascarides, 2003; Utt et al., 2013) such as "finish" and "start". In 9 additional cases, the main predicate did not describe the action, but it could be inferred from a prepositional phrase (e.g., sentence 3 in Table 8). There are also 13 implicit examples that do not fall into any of these categories.

7 Conclusion

We introduced a new NLP task, *object use classification*, which identifies whether an object mentioned in a sentence has been used or likely will be used. We introduced a gold standard dataset for this task and showed that all 3 categories (*Used*, *Anticipated Use*, and *No Use*) are common in real sentences. Then we presented a transformer-based architecture for this task that uses two types of data augmentation techniques (synonym/hyponym replacement and back translation) and also exploits exemplar sentences from FrameNet that correspond to an object's prototypical function. The resulting classification model achieves reasonably good performance for this task, although there is room for improvement that we hope will inspire future work on this problem.

8 Future Work

Our research was motivated by earlier work that introduced methods to automatically learn the prototypical goal activities for locations (Jiang and Riloff, 2018) and prototypical functions associated with human-made physical artifacts (Jiang and Riloff, 2021b). Table 8 illustrates the potential for combining our new object use classification model with commonsense knowledge about the prototypical functions of objects in order to improve sentence understanding. Current NLP systems would typically characterize these sentences based on the actions shown in Red, but we argue that the actions shown in Blue are the inferences that humans make when reading these sentences. In future work, we hope to put these pieces together to fully capture both the explicit and implicit meaning behind sentences and the commonsense inferences that people naturally make when reading sentences.

9 Limitations

Our dataset consists of 2,123 annotated sentences, which is relatively small. This is mainly due to the fact that our manual annotation effort not only required annotators to select among the object use categories but also required them to understand and label the prototypical function frames from FrameNet, which requires training and is time-consuming. It would be valuable to expand the object use data set in future work, both with more objects and more sentence contexts. Another limitation is that this paper does not evaluate the benefits of object use identification in downstream application tasks, which is a very interesting avenue for future research. This paper also focused exclusively on human-made physical objects because they usually have prototypical functions and our motivation was to infer the implicit use of these objects. However an open question is whether object use classification models could perform a similar task for natural objects (e.g., plants and rocks).

Acknowledgments

We thank the Utah NLP group for their constructive comments. We also thank the anonymous ARR reviewers for their valuable suggestions and feedback.

References

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007)*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING 1998)*.

Teresa Botschen, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly-Sergieh, and Stefan Roth. 2018. Multimodal frame identification with multilingual evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Mark H. Burstein. 1979. The use of object-specific knowledge in natural language processing. In *17th Annual Meeting of the Association for Computational Linguistics (ACL 1979)*.

Kevin Burton, Akshay Java, Ian Soboroff, et al. 2009. The icwsm 2009 spinn3r dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Eugene Charniak. 1972. *Toward a model of children's story comprehension*. Ph.D. thesis, MIT.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Charles J Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280 (1), pages 20–32.

Charles J Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm, The Linguistic Society of Korea (ed.)*, pages 111–137. Seoul: Hanshin Publishing Company.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Tianyu Jiang and Ellen Riloff. 2018. Learning prototypical goal activities for locations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

Tianyu Jiang and Ellen Riloff. 2021a. Exploiting definitions for frame identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*.

Tianyu Jiang and Ellen Riloff. 2021b. Learning prototypical functions for physical artifacts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Maria Lapata and Alex Lascarides. 2003. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

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*.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Michele Persiani and Thomas Hellström. 2019. Unsupervised inference of object affordance from text corpora. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-first AAAI conference on artificial intelligence (AAAI 2017)*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*.

Jason Utt, Alessandro Lenci, Sebastian Padó, and Alessandra Zarcone. 2013. The curious case of metonymic verbs: A distributional characterization. In *Proceedings of the IWCS 2013 Workshop Towards a Formal Distributional Semantics*.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. *arXiv preprint arXiv:2004.04877*.

William A Woods. 1975. What's in a link: Foundations for semantic networks. In *Representation and understanding*, pages 35–82. Elsevier.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the*

Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

Frank F. Xu, Bill Yuchen Lin, and Kenny Zhu. 2018. Automatic extraction of commonsense LocatedNear knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.