Learning Prototypical Functions for Physical Artifacts

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

Abstract

Humans create things for a reason. Ancient people created spears for hunting, knives for cutting meat, pots for preparing food, etc. The prototypical function of a physical artifact is a kind of commonsense knowledge that we rely on to understand natural language. For example, if someone says "She borrowed the book" then you would assume that she intends to read the book, or if someone asks "Can I use your knife?" then you would assume that they need to cut something. In this paper, we introduce a new NLP task of learning the prototypical uses for human-made physical objects. We use frames from FrameNet to represent a set of common functions for objects, and describe a manually annotated data set of physical objects labeled with their prototypical function. We also present experimental results for this task, including BERT-based models that use language model predictions from masked patterns as well as artifact sense definitions from WordNet and frame definitions from FrameNet.

1 Introduction

Humans are a creative species. New objects are invented by people every day, and most are created for a reason. Knives were created for cutting, bicycles were created for transportation, and telephones were created for communication. Some objects can perform multiple functions (e.g., smart phones) and humans are also creative at finding secondary uses for objects (e.g., heavy objects are often used as makeshift paperweights). But when we mention physical objects in conversation or in writing, people generally infer that the object will be used in the most prototypical way, unless they are told otherwise.

The *prototypical function* of human-made physical artifacts is a kind of commonsense knowledge that often plays a role in natural language understanding. Consider the following examples of inferences that arise from physical artifacts.

Example 1
a) He killed the mayor with a *gun*.
b) He killed the mayor with a *knife*.
c) He killed the mayor with a *bomb*.

Example 1 describes a killing with three different types of instruments. Most readers would assume that a) describes a shooting, b) describes a stabbing, and c) describes an explosion. But exactly how each instrument was used is implicit. We make different inferences about how they were used based on our knowledge of the objects.

Example 2
a) She finished the *cigarette*.
b) She finished the *puzzle*.
c) She finished the *movie*.

Example 2 illustrates how we infer different actions based on the object when the main action is elided (i.e., "finished" means that some action has ended but the action itself is implicit). Most people would assume that the cigarette was smoked, the puzzle was solved, and the movie was watched.

Example 3
a) She put the cake in the *box*.
b) She put the cake in the *oven*.
c) She put the cake in the *refrigerator*.

Example 3 illustrates second-order inferences that can follow from a sentence. The verb "put" means that the cake was placed somewhere, but the object of "in" leads to different inferences about intention. Putting a cake in an oven implies that it will be baked, but putting a cake in a refrigerator implies that it will be cooled.

Example 4

a) He ordered a *taxi*.
b) He ordered a *pizza*.
c) He ordered a *t-shirt*.

Example 4 reveals inferences about motivations and future plans. If someone orders a taxi then we infer that they need transportation, if they order a pizza then we expect they will eat it, and if they order a t-shirt then we assume it will be worn.

We believe that it is essential for NLP systems to "read between the lines" and make the same types of inferences that people do when reading these sentences. The goal of our research is to explore methods for learning the prototypical functions of human-made physical artifacts so that future NLP systems can benefit from this knowledge. First, we define a new NLP task to associate physical objects with frames from FrameNet as a canonical representation for their prototypical function. We introduce a gold standard data set of 938 physical artifacts that have each been labeled with a frame that represents its prototypical function based on human judgements. Second, we evaluate baseline models to assess how well existing resources and simple methods perform on this task. Third, we present transformer-based models for this task that exploit both masked sentence patterns and the definitions of physical artifacts and frames. Experiments show that our best model yields substantially better results than the baseline methods.

2 Related Work

Researchers have known for a long time that commonsense knowledge is essential for natural language understanding (Charniak, 1972; Schank and Abelson, 1977). Some of this work specifically argued that commonsense knowledge about physical objects, including functional knowledge, plays an important role in narrative text understanding (Burstein, 1979; Lehnert and Burstein, 1979).

These observations have led to considerable work toward constructing commonsense knowledge repositories. The Cyc project (Lenat, 1995) built a large ontology of commonsense concepts and facts over many years. More recently, ConceptNet (Speer et al., 2017) captures commonsense knowledge in the form of predefined relations expressed in natural language words and phrases. It was built from Open Mind Common Sense, a crowd-sourced knowledge project (Singh, 2002), and later enhanced with other sources such as Wiktionary and WordNet (Miller, 1995).

Within the NLP community, a variety of recent projects have focused on trying to acquire different types of commonsense knowledge, such as Forbes and Choi (2017); Collell et al. (2018); Rashkin et al. (2018); Yang et al. (2018). Sap et al. (2019) presented a crowd-sourced commonsense reasoning data set called ATOMIC that focuses on inferential knowledge related to events, which is organized as if-then relations. Bosselut et al. (2019) later proposed COMET, a transformer-based framework for automatic construction of commonsense knowledge bases that was trained from ATOMIC and ConceptNet. Both ConceptNet and COMET include a UsedFor relation that is relevant to our task, and we evaluate their performance on our data set in Section 6.

Of relevance to our work, Jiang and Riloff (2018) learned the prototypical "functions" of locations by identifying activities that represent a prototypical reason why people go to a location. For example, people go to restaurants to eat, airports to catch a flight, and churches to pray. They referred to the associated activity as a prototypical goal activity and presented a semi-supervised method to iteratively learn the goal activities.

Our work is also related to frame semantics, which studies how we associate words and phrases with conceptual structures called frames (Fillmore, 1976), which characterize an abstract scene or situation. The Berkeley FrameNet project (Baker et al., 1998; Ruppenhofer et al., 2016) provides an online lexical database for frame semantics and a corpus of annotated documents. There has been substantial work on frame semantic parsing (e.g., Das et al., 2014; Peng et al., 2018), which is the task of automatically extracting frame structures from sentences. Several efforts have enhanced FrameNet by mapping it to other lexicons, such as WordNet, PropBank and VerbNet (Shi and Mihalcea, 2005; Palmer, 2009; Ferrández et al., 2010). Pavlick et al. (2015) increased the lexical coverage of FrameNet through automatic paraphrasing and manual verification. Yatskar et al. (2016) introduced situation recognition, which is the problem of producing a concise summary of the situation that an image depicts. Similar to our work, they selected a subset of frames from FrameNet to represent possible situations depicted in an image. Our work uses a subset of frames from FrameNet to represent the

prototypical functions for human-made physical artifacts.

3 Motivation

Our work was motivated by observing sentences that mention physical objects and realizing that we often infer a richer meaning for these sentences than what they explicitly state. We came to appreciate that the prototypical function of an object was the basis for many of our inferences, but we also recognized that not all objects have a prototypical function. In particular, naturally occurring objects rarely have a prototypical function (e.g., *rock*, *snake*). In contrast, human-made physical objects usually do have a prototypical function because they were created for a purpose. Consequently, we limited the scope of our work to human-made artifacts. Of course, some objects are commonly used for multiple purposes, but in most cases there seems to be one use that is dominant, so for the sake of tractability we decided to assign a single (most) prototypical function to each artifact for this research. We had initially planned to include food items, but many foods are also naturally occurring plants or animals (e.g., *watermelon*, *shrimp*) so we omitted them. It may be worth re-examining these limitations in future work.

Another key decision that we had to make was how to represent the prototypical functions. Some recent work on commonsense knowledge acquisition has opted to generate words and phrases as expressions of a relation, such as ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019). As an example, ConceptNet includes a relation called UsedFor that lists the following phrases as uses for a knife: *stabbing, butter, cutting food, carving wood, slicing, boning*.

We chose to adopt a different approach. First, we wanted a canonical representation for each type of function that represents a general concept, rather than a list of phrases. This approach naturally captures clusters of objects (i.e., those assigned to the same frame) and avoids evaluation issues arising from differing phrases that may be learned for similar objects (e.g., *cut* vs. *carve* vs. *slice*). Second, we did not want to reinvent the wheel and develop a new taxonomy of action types ourselves. For these reasons, we chose to use the semantic frames in FrameNet as a canonical representation for our prototypical functions. Although FrameNet is not perfect nor complete, it contains many of the actions that we needed. Overall, it serves as an appropriate platform for our work.

This approach also opens up new avenues for research down the road. Although it is beyond the scope of this paper, we can imagine that sentences could trigger frames based on inferences originating from physical objects during semantic parsing. For example, *"She used a pencil"* should arguably be represented as a writing (*Text_Creation*) event. However we leave that challenge for future work. This paper focuses on the specific task of learning the prototypical functions for human-made physical artifacts using a subset of FrameNet frames as the set of function types.

4 Creating a Gold Standard Data Set

4.1 Artifact Selection

As explained in Section 3, our work focuses on artifacts that are 1) physical objects and 2) created by people. To acquire a list of objects that meet these criteria, we extracted all terms in synsets that are descendants of the *artifact.n.01* synset[1] in WordNet (Miller, 1995). We then removed a term from the list if the artifact sense was not its first sense definition.[2] This process produced 8,822 entries, many of which met our criteria except that the list still contained a lot of abstract terms (e.g., *vocabulary, modernism*).

To address this issue, we turned to Brysbaert et al. (2014) which presents concreteness ratings based on crowd sourcing for 37,058 English words and 2,896 two-word expressions. They used a 5-point rating scale ranging from abstract to concrete, so we extracted words with the part-of-speech "noun" and a rating ≥ 4.5, which produced a list of 3,462 concrete nouns. We then intersected this list with the terms extracted from WordNet, producing a set of 1,017 concrete physical artifacts.

4.2 Frame Selection

FrameNet 1.7 contains 1,221 frame definitions. However, not all of them are suitable for representing typical uses of physical artifacts, which should be actions that involve a physical object. For example, some frames are intended for abstract nominal categories (e.g., *Calendric_unit* for temporal terms), high-level abstractions (e.g., *Intentionally_act* which sits above more specific frames),

[1] Except we removed synsets for buildings and roads.

[2] Because the first sense definition in WordNet usually, though not always, represents the most common meaning.

Artifact Function Frames		
Wearing (145)	Light_movement (16)	Hunting (8)
Containing (76)	Building (15)	Cause_fluidic_motion (6)
Self_motion (69)	Dimension (15)	Eclipse (5)
Protecting (52)	Removing (14)	Inhibit_movement (5)
Supporting (49)	Closure (13)	Performing_arts (5)
Cause_harm (48)	Competition (13)	Setting_fire (5)
Perception_experience (44)	Create_representation (13)	Cause_to_fragment (4)
Make_noise (37)	Bringing (12)	Education_teaching (3)
Cause_motion (24)	Sleep (12)	Excreting (3)
Cutting (19)	Text_creation (12)	Cause_to_be_dry (2)
Cooking_creation (18)	Attaching (11)	Agriculture (1)
Ingestion (18)	Contacting (10)	Commercial_transaction (1)
Reading_activity (17)	Cure (9)	Residence (1)
Grooming (16)	Cause_temperature_change (8)	Rite (1)

Table 1: Frames for prototypical functions of physical artifacts. The frequency with which they occur in our gold standard data set is shown in parentheses.

and events or states that are not typically associated with physical artifacts (e.g., *Judgement*).

To focus on an appropriate subset of frames, we manually selected 42 frames in FrameNet that represent actions that are common functions of human-made physical artifacts. We intentionally didn't select frames that categorize nouns in a general way. For example, FrameNet contains an *Artifact* frame that includes *oven, phone* and *wheel* as its lexical units. This frame only serves to identify terms that represent physical objects, and we wanted frames that represent a function. The list of frames that we used is shown in Table 1 along with the frequency with which they occur in our gold standard data set, as described in the next section.

4.3 Human Annotation

To create a gold standard data set with frame assignments for the physical artifacts, we recruited 3 human annotators. We presented the annotators with the WordNet definition for each term and asked them to select one frame that captures the most prototypical use for the artifact. In addition to the 42 function frames, we also gave them a *None* option if none of the frames was a good match, and a *Not an artifact* option if the term was not in fact a human-made physical artifact (because our list extracted from WordNet and Brysbaert et al. (2014) was not perfect). To prepare the annotators, we asked them to read the definitions of all the frames beforehand and we gave them detailed annotation guidelines to familiarize them with the task. We randomly sorted the artifacts before presenting them to the annotators.

Frame	Artifact Examples
Wearing	*hat*, *shirt*
Containing	*basket*, *luggage*
Self_motion	*bicycle*, *yacht*
Protecting	*armor*, *helmet*
Supporting	*chair*, *scaffolding*
Cause_harm	*cannon*, *spear*
Perception_exp	*earphone*, *eyeglass*
Make_noise	*bell*, *violin*
Cause_motion	*engine*, *propeller*
Cutting	*knife*, *scissors*

Table 2: Examples of artifacts for the top 10 frames.

When the annotations were finished, we measured the pair-wise inter-annotator agreement (IAA) using Cohen's kappa. The IAA scores were 0.75, 0.72 and 0.69, with an average of $\kappa = 0.72$. Given the difficulty of this task (44 possible labels), we felt that the human agreement was relatively good.

Finally, we created the gold standard data set[3] by using the majority label from the three human annotators. There were 72 artifacts with no majority label (i.e., the annotators assigned 3 different labels), and 7 terms with the majority label *Not an*

[3]The data set is available at: https://github.com/tyjiangU/physical_artifacts_function

artifact, so we discarded these 79 terms. Consequently, our gold standard data set contains 938 physical artifacts that are each labeled with a frame representing its most prototypical function, or labeled as *None* when none of our 42 frames was appropriate.[4] Table 2 shows the 10 most frequently assigned frames and a few examples of artifacts assigned to each frame.

5 Methods

We explored several approaches for learning the prototypical functions of human-made physical artifacts. To assess the difficulty of this task, we first present baseline models that 1) exploit information extracted from existing knowledge bases and 2) use co-occurrence information extracted from a text corpus. Next, we explore methods that use large neural language models. We describe a method that uses masked pattern predictions, and then present models that also incorporate artifact sense definitions and frame definitions.

5.1 Notation

We model our task as a multiclass classification problem. The artifacts and frames are denoted as a_i $(i = 1..m)$ and f_j $(j = 1..n)$. The task is to select the f_j that represents the most prototypical use for an artifact a_i. We will denote the set of lexical units for f_j in FrameNet as $LU_j = \{l_k | l_k \text{ evokes } f_j\}$.[5]

5.2 ConceptNet and COMET Baselines

ConceptNet (Speer et al., 2017) is a well-known commonsense knowledge resource that contains a UsedFor relation, which is potentially relevant to our task (though it should be noted that an object can be used in ways that are not prototypical, so our task of identifying the *prototypical* use is not exactly the same). COMET (Bosselut et al., 2019) is a framework that was trained on ConceptNet with the goal of improving upon its coverage. Our first experiments apply these resources to see how effective they can be for this task.

For each artifact in ConceptNet, we extract the first word from each phrase listed under its UsedFor relation. These are typically verbs that describe an action although sometimes they are nouns. For COMET, we use its *beam-10* setting to generate 10 phrases of the UsedFor relation for each artifact.

[4] 83 terms were assigned to the *None* category.

[5] We merged lexical units from similar frames in FrameNet. See details in Appendix A.

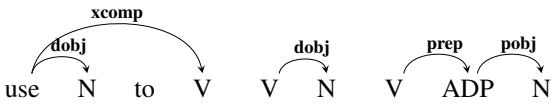

Figure 1: Dependency patterns used for co-occurrence.

Next, we want to use the extracted words to rank candidate frames. FrameNet defines *lexical units* that can evoke a specific frame. For example, *read* can trigger the *Reading_activity* frame. Suppose our artifact is a *book* and one of the extracted words is *read*, then *Reading_activity* is a candidate frame. We then score each frame based on the overlap between the words extracted from ConceptNet or COMET and the frame's lexical units. Specifically, we define $freq(a_i, w)$ as the count of a word w occurring in the UsedFor relation of artifact a_i, and $I(w, f_j) = 1$ if $w \in LU_j$ otherwise 0. Then our score for f_j is defined as:

$$S_{cn}(a_i, f_j) = \sum_{w \epsilon W} freq(a_i, w) * I(w, f_j), \quad (1)$$

where W is the set of extracted words. Finally, for each a_i, we select $f_{j'}$ such that $j' = \arg\max_j S_{cn}(a_i, f_j)$ as its prototypical function. If $S_{cn}(a_i, f'_j)$ equals zero, then we predict *None*.

5.3 Co-occurrence Baseline

An intuitive idea for potentially learning common functions associated with physical artifacts is to extract verbs that frequently co-occur with the artifact in a large text corpus. We assume that if a verb frequently co-occurs with an artifact, then the frames associated with the verb are plausible candidates for the artifact's prototypical function.

For this approach, we created 3 dependency parse patterns to extract <noun, verb> pairs, as depicted in Figure 1. The physical object is the noun represented by **N**. The activity is a verb (with an appended particle if one exists) represented by **V**. We included the verb-dobj pattern because some artifacts and their functions are expressed in this way, such as "*read book*" or "*wear jacket*". We used spaCy[6] to parse the whole English Wikipedia corpus (as of Feb 20, 2020) and extracted over 3.8 million <N, V> pairs (305,055 distinct pairs) for our 938 artifacts. We define the function $freq(a_i, v)$ as the co-occurrence count of artifact a_i and verb v in the corpus. Then we apply the same method described in Section 5.2 to assign a score to each

[6] https://spacy.io/

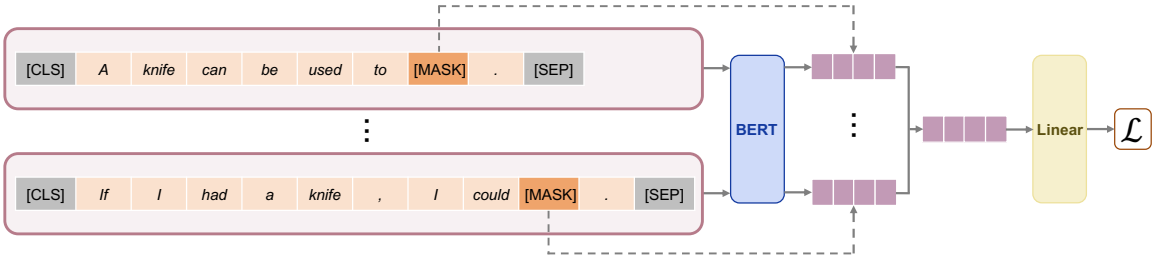

Figure 2: Overview of the PF$_{mask}$ model. Each pink block that is fed into BERT represents a sentence template for a given artifact.

frame based on the extracted verbs and select the best frame.

5.4 Masked Language Model (MLM) Baseline

Co-occurrence in text is a strong signal of correlation. But an activity that is highly correlated with an artifact may not be its prototypical use. For example, *cut* frequently co-occurs with *rope*, but the purpose of a *rope* is not to be *cut* – its prototypical use is for attaching things.

Recent work has successfully used masked language models to learn commonsense knowledge (Davison et al., 2019), so we explored whether masked language models could be beneficial for our task. We use the BERT (Devlin et al., 2019) masked language model to get prediction scores for every (a_i, l_k) pair, where a_i is one of our physical artifacts and l_k is a lexical unit linked to one of our 42 candidate frames. We defined 6 sentence templates that represent expressions describing what an object is used for, which are shown below. The first blank space is for artifact a_i and the second blank space is for action l_k.

(1) __ can be used to __ .
(2) I used __ to __ .
(3) __ can be used for __ .
(4) I used __ for __ .
(5) The purpose of __ is to __ .
(6) If I had __ , I could __ .

Next, we produced a probability distribution over all of the lexical units based on the second blank position. Specifically, for the t-th sentence template s_t, we obtain $Pr(l_k|s_t, a_i)$ by masking only the second blank space (a_i is inserted into the first blank) and we obtain $Pr(l_k|s_t)$ by masking both blank space. Then we define the score of l_k as the typical use of artifact a_i based on the t-th template as:

$$U(a_i, l_k, s_t) = \log Pr(l_k|s_t, a_i) - \log Pr(l_k|s_t). \quad (2)$$

The score $U(a_i, l_k)$ using all templates is computed as: $U(a_i, l_k) = \frac{1}{t}\sum_t U(a_i, l_k, s_t)$.

Finally, we define the score for f_j being the prototypical function for a_i as:

$$S_{mlm}(a_i, f_j) = \sum_{l_k \in LU_j} U(a_i, l_k). \quad (3)$$

We select $f_{j'}$ where $j' = \arg\max_j S_{mlm}(a_i, f_j)$ as the best frame. If $S_{mlm}(a_i, f'_j) \leq 0$, we predict *None*.

5.5 Learning from Masked Patterns

Our MLM baseline uses the discrete output of the masked language model (i.e., the prediction tokens from the vocabulary and their scores). In order to take advantage of a language model's fine-tuning capability, we use the same architecture as described in Section 5.4, except that instead of using the predicted lexical units and their probability $Pr(l_k|s_t, a_i)$, we retrieve the last hidden state vector for the `[MASK]` token as output. Since there are 6 masked templates, we have 6 output vectors for each artifact a_i. We compute the average of these vectors and pass it through a linear layer and a softmax layer to produce a probability distribution over all candidate frames plus *None*. Figure 2 shows the overview of this architecture, which we will call the PF$_{mask}$ model. We will refer to the final score for artifact a_i with respect to frame f_j as $S_{mask}(a_i, f_j)$. The loss function is defined as:

$$\mathcal{L} = -\sum_{i=1}^{n} \log S_{mask}(a_i, f_{j^*}), \quad (4)$$

where f_{j^*} is the gold label for a_i.

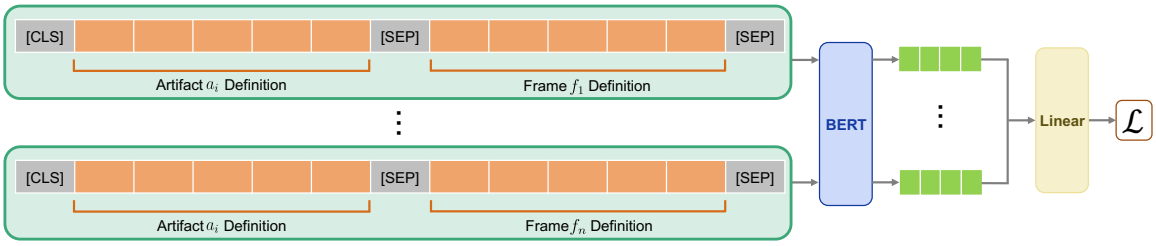

Figure 3: Overview of the PF_{def} model. Each green block that is fed into BERT represents an artifact and one of the candidate frames.

5.6 Learning from Definitions

The challenge for our task is obtaining information about the intended function of a physical artifact. We observed that this information is often described in the dictionary definition of an artifact, although it can be expressed in many different ways. For example, the first sense definition in WordNet for *knife* is *"edge tool used as a cutting instrument..."*, and for *bus* it is *"a vehicle carrying many passengers..."*. The definition often provides a short and precise sentence that describes what the artifact is as well as what it is typically used for.

FrameNet also provides a definition for each frame. For example, the definition of the *Cutting* frame is *"An Agent cuts a Item into Pieces using an Instrument"*. Jiang and Riloff (2021) exploited both frame and lexical unit definitions for the frame identification task in a model that assesses the semantic coherence between the meaning of a target word in a sentence and a candidate frame. Similarly, we hypothesized that a model could potentially learn the semantic relatedness between the definitions of a physical artifact and the frame that describes its typical function.

To investigate this idea, we used the BERT model (Devlin et al., 2019) as the base of our architecture and fine-tuned BERT for our task using both dictionary definitions of artifacts and frame definitions from FrameNet. Figure 3 shows the overview of this architecture, which we call the PF_{def} model. Each large green block represents an artifact a_i paired with one of the candidate frames. We encode WordNet's definition of the artifact as the first input sequence and the frame's definition from FrameNet as the second input sequence to BERT. We use the last hidden vector of the `[CLS]` token as the output. For each artifact a_i, we have $n+1$ such pairs where n is the number of candidate frames and 1 refers to the *None* option. On top of BERT's output, we apply a linear and a softmax layer to produce a probability distribution

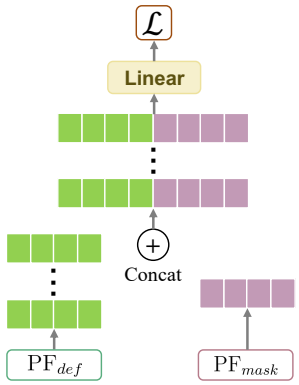

Figure 4: Overview of the $\text{PF}_{def+mask}$ model.

over all candidate frames. We will refer to the final score for artifact a_i with respect to frame f_j as $S_{def}(a_i, f_j)$. The loss function is defined as:

$$\mathcal{L} = -\sum_{i=1}^{n} \log S_{def}(a_i, f_{j^*}), \qquad (5)$$

where f_{j^*} is the gold label for a_i.

5.7 Learning from Definitions plus Masked Patterns

Our final model combines the idea of using both definitions and masked sentence patterns. Figure 4 depicts the combined $\text{P}_{def+mask}$ model. The left part is the PF_{def} model which estimates the relatedness between artifact and frame definitions. Its output is a matrix of dimension (*# of frames, hidden vector size*). The right part is the PF_{mask} model, which predicts the most probable frame for an artifact using our masked patterns. It produces a single output vector of dimension (*1, hidden vector size*). We broadcast it across the rows to have the same dimension as (*# of frames, hidden vector size*) and then we concatenate the matrices of both models to pass through a linear layer before computing the loss. The model uses fine-tuning to jointly learn all parameters so that information from both models will optimally contribute to the final prediction.

Model	Acc	Pre	Rec	F1
ConceptNet	17.5	33.6	13.5	16.4
Co-occurrence	31.9	24.1	23.9	19.9
COMET	30.7	29.7	35.6	28.2
MLM	42.8	29.5	33.8	28.2
PF_{mask}	58.5	35.7	36.5	35.4
PF_{def}	74.7	63.5	57.6	59.3
$PF_{def+mask}$	**76.8**	**65.2**	**61.1**	**62.4**

Table 3: Experimental results for different models.

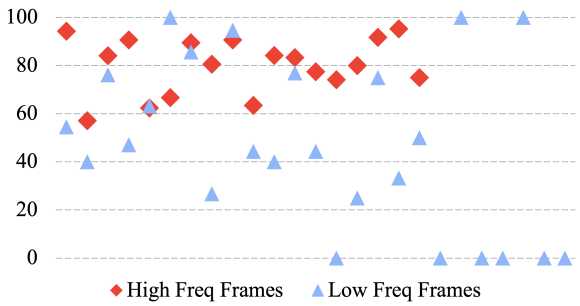

Figure 5: F1 scores for high & low frequency frames.

6 Evaluation

6.1 Experiment Settings

Our gold standard data set contains 938 artifacts that are each paired with one frame that represents its most prototypical use. We set aside 20% (188) of the data as a development set and used 80% (750) as the test set. We evaluated all of the learning models by performing 5-fold cross validation on the test set. We use the pre-trained uncased BERT-base model with the same settings as Devlin et al. (2019) and fine-tuned BERT on the training data. We set the max sequence length as 200, batch size as 1, learning rate started at 2e-5, and train for 10 epochs. All reported results are averaged over 3 runs. We report overall accuracy as well as precision, recall and F1 scores macro-averaged over the 43 class labels (42 frames + *None*).

6.2 Results

The first four rows in Table 3 show the performance of our four baseline methods. ConceptNet and the Co-occurrence model produced the lowest F1 scores. We see that ConceptNet has better precision but low recall because only about 1/3 of the artifacts in our data set has a UsedFor relation defined in ConceptNet. We also tried adding the CapableOf relation, which is defined as what an item can do, but it is even more sparse than UsedFor and combining both relations only marginally increased recall. The performance of COMET shows that COMET does indeed improve upon the coverage of ConceptNet, although it sacrifices some precision. We also tried using the *beam-5* and *greedy* settings of COMET, which produced higher precision but lower recall and F1 scores.

Compared to COMET, the Co-occurrence baseline has higher accuracy but a much lower F1 score. The explanation is that the Co-occurrence model performs much better on frames that are associated with artifacts that are frequently mentioned in the corpus than for frames associated with less frequent artifacts. This is intuitive because, in general, we expect to extract a more representative sample of activities when we have more data. This phenomenon (accuracy much higher than F1) can also be observed in the MLM model which uses a pre-trained language model that learns from large corpora, so it is not surprising that Co-occurrence and the MLM model behave similarly. In contrast, ConceptNet and COMET behave more consistently across the set of frames.

The bottom section of Table 3 shows the results for our new models, which were trained specifically for this task. The PF_{mask} model achieves 58.5% accuracy and a 35.4% F1 score, which outperforms all of the baselines. The PF_{def} model performs substantially better, achieving 74.7% accuracy and a 59.3% F1 score. This result demonstrates that the definitions of the artifacts and the frames provide valuable information that a learner can benefit from. The last row shows the performance of the combined model, which performed better than the individual models. This model saw additional gains in both precision and recall, increasing the accuracy from 74.7% to 76.8% and the F1 score from 59.3% to 62.4%.

6.3 Analysis

To understand the degree to which the number of training instances for each frame correlated with performance, we divided the frames into two sets: high frequency frames assigned to ≥ 15 artifacts and low frequency frames assigned to < 15 artifacts. The results are shown in Figure 5 with the F1 scores from the $PF_{def+mask}$ model displayed on the Y-axis. We conclude that frames with more training instances generally showed better performance, so our model would likely further improve

	ID	Artifact	PF_{mask}	PF_{def}
MASK ✓ DEF ✓	1	scissors	**Cutting**	**Cutting**
MASK ✓ DEF ✗	2	hydrant	**Cause_fluidic_motion**	Cause_temperature_change
MASK ✗ DEF ✓	3	bed	Supporting	**Sleep**
MASK ✗ DEF ✓	4	helmet	Wearing	**Protecting**
MASK ✗ DEF ✗	5	snowplow	Hunting	Self_Motion

Table 4: Sample output of PF_{def} and PF_{mask} models. The correct predictions are in bold.

given more training data.

Table 4 shows some examples of output from the PF_{mask} and PF_{def} models to compare their behavior. The correct predictions appear in bold. Both models are correct for example 1. For example 2, only the PF_{mask} model is right, which indicates that the masked pattern can be more useful than the definition sometimes. For examples 3 and 4, PF_{def} was correct and PF_{mask} was wrong. The PF_{mask} model sometimes generates frames representing functions that are true but tangential. For example, beds do support us and helmets are worn, but these functions do not sufficiently characterize the objects (e.g., chairs also support us but are not typically used for sleeping, and jewelry is also worn but not used for protection). For example 5, both models are wrong – the correct frame is *Removing*. Though both are wrong, the PF_{def} model produces a more reasonable answer than the PF_{mask} model.[7] We also observed that the MLM baseline sometimes produces seemingly random answers that are hard to explain.

Finally, we investigated the 83 instances that were labeled as *None* to see what kind of artifacts fell into this category. The biggest cluster of related artifacts were 17 types of fabric, such as *linen, silk* and *canvas*. FrameNet does not include a frame for materials of this kind, probably because they are an ingredient for making clothes rather than tools themselves. Artifacts like *toy* were also labeled as *None* presumably because toys are used in a general way (for play). This category also included some artifacts not tied to a single prototypical function but commonly used for many purposes (e.g., *computer, laptop*).

7 Conclusion

We introduced the new task of learning prototypical functions for human-made physical artifacts, and used a subset of frames from FrameNet to represent the set of common functions. We also presented a manually annotated data set of 938 physical artifacts for this task. Our experiments showed that a transformer-based model using both artifact and frame definitions as well as masked pattern predictions outperforms several baseline methods. In future work, we hope to show the value of functional knowledge about objects for sentence-level understanding tasks as well as narrative document understanding.

[7] In fact, snowplow can also refer to a skiing action, although WordNet does not contain that word sense.

Acknowledgments

We thank Yuan Zhuang for his helpful comments on our work. We also thank the anonymous reviewers for their valuable suggestions and feedback.

References

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)*.

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 (ACL 2019)*.

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 *Proceeding of the 17th annual meeting on Association for Computational Linguistics (ACL 1979)*.

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)*.

Dipanjan Das, Desai Chen, André F.T. Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Óscar Ferrández, Michael Ellsworth, Rafael Muñoz, and Collin F. Baker. 2010. Aligning FrameNet and WordNet based on semantic neighborhoods. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 10)*.

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.

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)*.

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. 2021. Exploiting Definitions for Frame Identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*.

Wendy G. Lehnert and Mark H. Burstein. 1979. The Role of Object Primitives in Natural Language Processing. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI 1979)*.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

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

Martha Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the generative lexicon conference*, pages 9–15. GenLex-09, Pisa, Italy.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. FrameNet+: Fast paraphrastic tripling of FrameNet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. 2018. Learning joint semantic parsers from disjoint data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

Josef Ruppenhofer, Michael Ellsworth, Myriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*.

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

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

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *International conference on intelligent text processing and computational linguistics*.

Push Singh. 2002. The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*.

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

Yiben Yang, Larry Birnbaum, Ji-Ping Wang, and Doug Downey. 2018. Extracting Commonsense Properties from Embeddings with Limited Human Guidance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

A Appendix

When selecting frames to represent the prototypical functions of physical artifacts, we observed that some frames in FrameNet share similar meanings (e.g., *Reading_activity* and *Reading_perception*) or related functions (e.g., *Create_representation* and *Recording*). However, these frames often have complementary sets of lexical units.

Since our baselines (ConceptNet, COMET, Co-occurrence, and MLM) rely on the lexical units of frames to make predictions, increasing the coverage of lexical units can be beneficial. So we manually clustered frames that share a related definition with our 42 chosen frames and merged their lexical units. The table below shows the cluster for which the lexical units are merged. Our experiments showed that merging lexical units from these frames improved both the precision and recall.

Primary Frame	**Clustered Frames**
Agriculture	Food_gathering
	Growing_food
	Planting
Attaching	Connectors
Cause_fluidic_motion	Cause_to_be_wet
Cause_harm	Attack
	Weapon
Cause_motion	Cause_to_move_in-_place
Cause_to_fragment	Grinding
Commercial_transaction	Commerce_buy
	Commerce_sell
Competition	Exercising
Containing	Containers
Cooking_creation	Apply_heat
Create_representation	Recording
Cure	Recovery
Hunting	Taking_captive
	Trap
Inhibit_movement	Immobilization
Light_movement	Location_of_light
Make_noise	Cause_to_make_noise
	Noise_makers
Perception_experience	Perception_active
	Cause_to_perceive
	Information_display
Reading_activity	Reading_perception
Removing	Emptying
Self_motion	Vehicle
	Ride_vehicle
	Operate_vehicle
Supporting	Posture
Wearing	Body_decoration
	Clothing
	Accoutrements
	Clothing_parts