Do Dogs have Whiskers?
A New Knowledge Base of hasPart Relations

Anonymous authors

Abstract

We present a new knowledge-base (KB) of hasPart relationships, extracted from a large corpus of generic statements. Complementary to other resources available, it is the first which is all three of: accurate (90% precision), salient (covers relationships a person may mention), and has high coverage of common terms (approximated as within a 10 year old's vocabulary), as well as having several times more hasPart entries than in the popular ontologies ConceptNet and WordNet. In addition, it contains information about quantifiers, argument modifiers, and links the entities to appropriate concepts in Wikipedia and WordNet. Our contributions are both the resource itself and the construction method used, in particular the use of generic sentences as the source of knowledge. The KB is available at https://tinyurl.com/haspartkb (anonymized link).

1. Introduction

Meronymic (hasPart) relations are one of the most important and frequently used relationships in reasoning systems, perhaps second only to the generalization (“isa”) relationship. hasPart knowledge plays a role in multiple inference scenarios, for example:

- if X moves to Y, then X’s parts move to Y;
- if part of X is broken, then X is broken;
- to construct X, one needs all the parts of X;
- if X is ill, then some part of X may be the cause;

However, while there has been extensive research on mining hasPart relationships from text, e.g., [Girju et al., 2006, van Hage et al., 2006, Ling et al., 2013], there are only a few resources that have been made available. Two popular general resources, WordNet [Fellbaum, 1998] and ConceptNet [Speer et al., 2017], contain collections of only 9k and 13k hasPart relationships respectively. In addition, when restricted to hasPart relations between common terms, which we approximate as within the typical vocabulary of a Fifth Grader (age 10) [Stuart et al., 2003], these totals drop to ≈1k in each resource. Other resources have different limitations: Quasimodo [Romero et al., 2019] contains 18k partonomic relation-
ships, but only covering body parts rather than the general hasPart relationship; WebChild [Tandon et al., 2017] contains 256k hasPart relations, but a large proportion covers unusual concepts - only 9k are within a Fifth Grade vocabulary; and although the resource PWKB (part-whole KB) [Tandon et al., 2016] contains 6.5M hasPart relations, the large majority were computed by an inference-based expansion of a smaller set, resulting in many entries that a person would be unlikely to mention (low salience).

We contribute a complementary resource of hasPart knowledge, the first which is all three of: accurate (90% precision), salient (covers relationships a person may mention), and has high coverage of common terms (within a Fifth Grade vocabulary). While our main contribution is the resource itself, our approach to extraction is also novel: rather than extracting hasPart relations from arbitrary text, we only extract from generic sentences, i.e., statements about members of a category such as “Dogs have tails.”. Empirically, this results in a high yield of good quality extractions (Section 4), and significantly higher than a strong, prior extraction pipeline applied to the same corpus (Section 4.4). Our resulting hasPartKB contains 50k entries, including over 15k within a high-schooler’s vocabulary, each additionally annotated with information about quantifiers, argument modifiers, and links the entities to appropriate concepts in Wikipedia and WordNet.

Our work is targeted in two important ways. First, as there are several types of hasPart relationship (e.g., Winston et al. (1987) provide a commonly used taxonomy of six types), we bound the scope of our work to just the two most common types, namely “component/integral object” (e.g., a handle is part of a cup) and “stuff/object” (e.g., steel is part of a bike). Second, we target salient parts, which we informally define as those that a person might consider mentioning, and by implication more likely to be useful in an end-task. The restriction is important as, in a literal sense, many entities have millions of parts, making a complete enumeration both infeasible and unhelpful.

2. Related Work

There has been substantial prior work on extracting hasPart knowledge, although the majority did not result in publicly available resources being released. Early work by Berland and Charniak (1999) used two Hearst patterns [Hearst, 1992] to extract part-wholes, but just covering a small number of objects. Girju et al. [Girju et al., 2006, 2003] developed a semi-automatic method for extracting part-whole relations, using a combination of hand-identified lexical patterns and machine-learned selectional constraints to optimally filter extractions, resulting in 10k extractions at 80% precision (although no public resource was released). Similarly, Ling et al. (2013) combined distant supervision with multi-instance learning for meronymy extraction, in particular aggregating evidence from multiple sentences together to reduce noise. However, the method was only applied to biology text and again no resource was released. More recently, Tandon et al. developed WebChild (2017), a large-scale resource including hasPart relations, and PWKB (part-whole KB) (2016) specif-
ically aimed at hasPart relations. WebChild includes 256k hasPart relations (comprising a subset of most reliable PWKB relations), while PWKB contains 337k core relations, expanded to 6.5M entries using rules for inheritance and transitivity - we compare our results to WebChild and PWKB in Section 4. PWKB’s relations were extracted by first finding part-whole lexical patterns (e.g., “noun of the noun”) using 1200 seed part-whole pairs from WordNet, applying those patterns to a large corpus with a novel scoring function, and finally expanding and filtering the results using inheritance and transitivity rules. Two other related, general resources are TupleKB [Mishra et al., 2017] and Quasimodo [Romero et al., 2019], both general KBs including hasPart knowledge. However, although TupleKB is large (280k entries), it contains less than 1000 hasPart entries (excluding those added directly from WordNet), so has limited hasPart coverage. Likewise, Quasimodo only includes the “has body part” relation in its extraction vocabulary, rather than general meronymic relationships.

In the last few years, general relation extraction has largely shifted to using neural techniques, e.g., [Lin et al., 2016, Kuang et al., 2019, Wang et al., 2019]. We similarly apply neural methods (using BERT [Devlin et al., 2018] and RoBERTa [Liu et al., 2019]), but specifically for hasPart, and do not claim any novelty in this aspect of our approach.

3. Approach

Our approach to hasPart extraction has five steps:
1. Extract generic sentences from a large corpus
2. Train and apply a RoBERTa model to identify hasPart relations in those sentences
3. Normalize the entity names
4. Aggregate and filter the entries
5. Link the hasPart arguments to Wikipedia pages and WordNet senses

We now describe each step in turn.

3.1 Step 1: Collecting Generic Sentences

Rather than extract knowledge from arbitrary text, we extract hasPart relations from generic sentences, e.g., “Dogs have tails.”, in order to bias the process towards extractions that are general (apply to most members of a category) and salient (notable enough to write down). Thus as a pre-processing step, we first collect generic sentences from a large corpus of text. For this we use the Waterloo corpus, a Web crawl of 1.7B sentences. This is the same corpus used to build the TupleKB [Mishra et al., 2017] and made available to us by the authors.

Generic sentences are identified in two steps. First, a small set of hand-authored lexicosyntactic rules were used to identify likely generic sentences, e.g., sentences starting with a plural noun-phrase, and remove sentences that have situational meaning, e.g., use of present participles such as “Pond snails are eating”. Second, a BERT classifier was trained on a crowdsourced dataset in which generics had been labeled as “standalone truths” or not, and applied to the candidate generics. The result was a corpus $GKB$ of 3.8M standalone generic sentences, to use for subsequent hasPart extraction.
3.2 Step 2: hasPart Extraction

To identify hasPart relationships in a sentence $S$, we first identify candidates and then train and apply a RoBERTa model to classify them, as we now describe.

First, for each sentence $S$ in GKB, we identify all noun chunks in the sentence using a noun chunker (spaCy’s Doc.noun_chunks). Each chunk is a candidate whole or part. Then, for each possible pair, we use a RoBERTa model to classify whether a hasPart relationship exists between them. The input sentence is presented to RoBERTa as a sequence of word piece tokens, with the start and end of the candidate hasPart arguments identified using special tokens, e.g.:

$$[CLS] [ARG1-B] \text{Some pond snails}[ARG1-E] \text{ have } [ARG2-B] \text{gills}[ARG2-E] \text{ to breathe in water.}$$

where $[ARG1/2-B/E]$ are special tokens denoting the argument boundaries. The $[CLS]$ token is projected to two class labels (hasPart/notHasPart), and a softmax layer is then applied, resulting in output probabilities for the class labels. We train with cross-entropy loss. We use RoBERTa-large (24 layers), each with a hidden size of 1024, and 16 attention heads, and a total of 355M parameters. We use the pre-trained weights available with the model and further fine-tune the model parameters by training on our labeled data for 15 epochs.

To train the model, we use a hand-annotated set of $\sim$2k examples. Given that sentences expressing hasPart information are sparse in GKB, collecting a representative sample of positive and negative examples to annotate is itself a challenge. To help with this, we proceeded as follows:

1. Train a similar model, distantly supervised using a subset of ConceptNet’s partOf relations:$^3$ GKB sentences mentioning both terms in a ConceptNet partOf relation are used as positive examples. Negative examples were generated by (a) reversing the arguments in positive examples, and (b) using GKB sentences mentioning arguments from other ConceptNet relations besides partOf. The result is a coarse-grained hasPart classifier based on ConceptNet’s data.
2. Apply this model to each GKB sentence, for each pair of noun chunks it contains, to find sentences that likely contain a hasPart relation. We treat these as good candidates to hand-annotate. We take a sample (380) of such sentences for this purpose.
3. For all noun chunk pairs in each sample sentence, we annotate each to indicate if a hasPart relationship holds or not. This process resulted in a final training set of 2,106 training examples, used to train the final RoBERTa-hasPart-classifier model.

After training the model, we run it over all sentences $S$ in GKB, and for all noun chunk pairs in $S$, to classify each pair as a hasPart relation. Of the (several million) classifications, we obtain a total of $\sim$127k hasPart examples (hasPart class score > 0.5) in the initial hasPart database. We now normalize, aggregate, filter, and link the this data, described next.

3.3 Step 3: Entity Normalization

The noun chunker sometimes identifies chunks that include quantifiers (e.g., “most”) and/or modifiers (e.g., “large”). To normalize the entity names, we remove these, but retain them

---

$^3$ ConceptNet has $\sim$13k partOf relations which are noisy. We used a combination of heuristics and manual filtering on this set to reduce it down to a set of $\sim$9k more reliable partOf relations.
Table 1: After thresholding, the final hasPart KB contains 50k entries, with precision of \(\approx 90\%\) (based on sampling).

as metadata. Quantifiers are identified simply by checking if the first word of the entity name is a quantification word (using a small list of quantifier words). Modifiers are identified by cross-referencing with Wikipedia titles: If the entire entity name is a Wikipedia title, it is retained. If not, the first word is removed as a possible modifier, and the shortened name is again checked against Wikipedia. This process is repeated until a Wikipedia name is found. If no Wikipedia name is found even after removing all words, the entire entity name is retained. In this way, entity names such as “large elephant” will become “elephant” modified by “large”, while “large intestine” will be retained as a single term.

3.4 Step 4: Aggregation and Thresholding
We post-processed these to aggregate duplicate tuples resulting from multiple sentences. The aggregated tuples are assigned an average-pooled score from the scores from individual sentences. To further improve precision (though at the expense of recall), we removed hasPart relations whose RoBERTa score was below a threshold (chosen to raise precision to 90%). The resulting yields and precisions (measured by sampling) are shown in Table 1.

3.5 Step 5: Entity Linking and Word Sense Disambiguation
Finally, the two entities in each hasPart entry are linked to both a Wikipedia page (using the Wikipedia title identified in the earlier step), and their WordNet word sense. If the Wikipedia title is ambiguous, we link to the primary page for its disambiguation. For assigning word senses, we re-implemented the GlossBERT word sense disambiguation model [Huang et al., 2019], which uses the associated sentence(s) as context to disambiguate the two hasPart arguments. In contrast to the original paper, however, we used the more recent RoBERTa encoder instead of BERT, and employed a modified ranking loss function\(^4\) [Richardson and Sabharwal, 2019] which we found achieves state-of-the-art performance on existing WSD benchmarks.

3.6 Final hasPartKB
The resulting hasPart database contains over 50k entries with a (sampled) precision of \(\approx 90\%\). In addition, each entry contains information about quantifiers, argument modifiers, and links the entities to appropriate concepts in Wikipedia and WordNet.

4. Evaluation
We now evaluate hasPartKB along three dimensions: precision, coverage, and salience, and compare these to several existing resources containing hasPart information, specifically

\(^4\) Details are provided in Appendix A1.
Table 2: Precision of hasPart entries in the differing resources. All show good ($\approx 80\%+$) precision for their entries.

<table>
<thead>
<tr>
<th>Source</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordnet 3.0</td>
<td>100% (from [Tandon et al., 2016])</td>
</tr>
<tr>
<td>ConceptNet 5.6.0</td>
<td>91% (sampled)</td>
</tr>
<tr>
<td>Quasimodo 1.2</td>
<td>78% (sampled)</td>
</tr>
<tr>
<td>TupleKB</td>
<td>80% (from [Mishra et al., 2017])</td>
</tr>
<tr>
<td>Webchild</td>
<td>82% (from [Tandon et al., 2017])</td>
</tr>
<tr>
<td>PWKB</td>
<td>89% (from [Tandon et al., 2016])</td>
</tr>
<tr>
<td>hasPartKB (ours)</td>
<td>90% (sampled)</td>
</tr>
</tbody>
</table>

WordNet [Fellbaum, 1998], ConceptNet [Speer et al., 2017], Quasimodo [Romero et al., 2019], TupleKB [Mishra et al., 2017], WebChild [Tandon et al., 2017], and PWKB [Tandon et al., 2016]. Note that these three metrics interact, thus no single measure should be taken in isolation.

4.1 Precision

Table 2 shows the (approximate) precision of our KB and other resources, using either the published precision figures, or human judgements over a small (200) sample of randomly selected entries. The main observation is that all the resources have good precision ($\approx 80\%+$), reflecting the respective care that has gone into their construction.

4.2 Coverage

We now evaluate the coverage of our hasPartKB and the other resources. We also conduct a small case study on coverage of six selected concepts, using independently authored parts lists for each.

4.2.1 Database Sizes (Yield)

To what extent does the KB comprehensively tell us the parts of entities? While the notion of coverage is hard to define, we use two approximations: First, what is the overall yield (size) of the database, and second, what is the yield (size) when constrained to “common” concepts, which we approximate as those with names within the vocabulary of a Fifth Grader. In addition, for the Fifth Grade subset, we count how many distinct wholes and how many distinct parts are mentioned. Although these are approximate measures, they provide some indicators of coverage.

Table 3 show the comparative sizes of the full hasPart databases, and the subset within a Fifth Grade vocabulary. We observe that:

- PWKB has the overall largest coverage. However, this is largely due to its inference-based construction process, resulting in most entries being obscure (low salience, Section 4.3).
- Of the remainder, although WebChild has the largest general coverage, our hasPartKB has 50% greater coverage of hasPart relations between common terms (Fifth Grade
Table 3: hasPart Relation Yield. Our hasPartKB has greatest yield over common (5th Grade) terms, with the exception of PWKB. However, PWKB suffers from low salience (Section 4.3).

<table>
<thead>
<tr>
<th>Source</th>
<th>All</th>
<th>in 5th Grade Vocab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordnet 3.0</td>
<td>9098</td>
<td>1399</td>
</tr>
<tr>
<td>ConceptNet 5.6.0</td>
<td>12988</td>
<td>2164</td>
</tr>
<tr>
<td>Quasimodo 1.2</td>
<td>3197</td>
<td>1546</td>
</tr>
<tr>
<td>TupleKB</td>
<td>898</td>
<td>645</td>
</tr>
<tr>
<td>WebChild</td>
<td>256k</td>
<td>10566</td>
</tr>
<tr>
<td>PWKB</td>
<td>6.5M</td>
<td>336k</td>
</tr>
<tr>
<td>hasPartKB (ours)</td>
<td>50752</td>
<td>15200</td>
</tr>
</tbody>
</table>

Table 4: The number of distinct Wholes and Parts, for hasPart entries within a Fifth Grade Vocabulary.

<table>
<thead>
<tr>
<th>KB</th>
<th># wholes</th>
<th># parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordnet 3.0</td>
<td>787</td>
<td>1068</td>
</tr>
<tr>
<td>ConceptNet 5.6.0</td>
<td>1125</td>
<td>1497</td>
</tr>
<tr>
<td>Quasimodo 1.2</td>
<td>836</td>
<td>119</td>
</tr>
<tr>
<td>TupleKB</td>
<td>109</td>
<td>58</td>
</tr>
<tr>
<td>Webchild</td>
<td>2061</td>
<td>496</td>
</tr>
<tr>
<td>PWKB</td>
<td>12179</td>
<td>995</td>
</tr>
<tr>
<td>hasPartKB (ours)</td>
<td>3294</td>
<td>3304</td>
</tr>
</tbody>
</table>

Table 4 shows the number of distinct terms in the Fifth Grade Vocabulary subset. By this metric, the results show hasPartKB has the greatest coverage of parts, and, apart from PWKB, also for wholes within this core vocabulary.

4.2.2 Case Study: Parts Coverage of Six Entities

As a narrower case study of coverage, we randomly sampled six different entities present in the 5th Grade vocabulary that returned a manually authored list of parts via Google Search. The selection process was:

- Randomly sample the vocabulary to locate a term X that is a physical entity
- Submit a query "What are the parts of X?" to Google
- If Google does not return a Featured Snippet that directly lists the parts, discard the term and repeat. If not:
  - Record the parts from the page(s) linked in the featured snippets.
  - Check for coverage of the entire list of parts in our KB and other KBs we compare against

We measured coverage using both strict (exact match) and loose (head noun match) measures. The results are shown in Table 5, and suggest that, at least for this small sample,
Table 5: Percent coverage of parts mentioned in an independent parts list, for six randomly chosen Fifth Grade concepts.

the hasPartKB has a higher coverage of parts in the sample parts list (78% average) compared to the other sources. Although this is a small study, it provides a second indicator of hasPartKB’s good coverage. For the list of parts we evaluated against for each entity, see Appendix A2.

4.3 Salience

Many objects have thousands, or even millions, of parts (including electrons and quarks), making a complete enumeration both infeasible and unhelpful. Rather, we wish to collect hasPart relationships that are likely to be useful. We refer to such parts as salient parts. As a rather approximate indicator of salience, we consider an entry salient if it is one that “someone might reasonably consider mentioning.” For example, “a tail is part of a dog” is salient, but “a vacuole is part of a queen consort” is not. We can weakly operationalize this by, given hasPart\((x, y)\), checking whether there is a sentence mentioning both \(x\) and \(y\) in a large corpus, i.e., the relationship has (likely) been mentioned.

In fact, by this criterion, all but one of our resources have high salience. WordNet and ConceptNet were either hand-built, thus, by our definition, all the entries are salient as someone thought to mention the relation. The remainder, bar PWKB, contain entries extracted from at least one sentence, thus again by definition, someone wrote down the relationship. The one exception is PWKB, where the large majority (over 90%) of the contents were inferred through inheritance and transitivity of the hasPart relationship, rather than directly extracted. To assess salience in PWKB, we queried a large corpus (using the 1.7B Waterloo corpus, described earlier in Section 3.1) for sentences mentioning both entities in a PWKB hasPart relationship, for a random sample of 1000 entries. We find that only 7.2% pass this “salience” test. This suggests that PWKB, although large, contains mainly obscure relationships. The results from the first 10 are shown in Figure 1 to illustrate this.

---

5. Of course, the sentence may be describing some other relationship besides hasPart, and there may be cases where a hasPart relationship is expressed over multiple sentences. This measure is thus only approximate.
hasPart(chiropodist,liver): (no sentence found)
hasPart(nuncio,hip): (no sentence found)
**hasPart(kapok,bark):** Found: "Tree bark from the Kapok tree."
hasPart(Kansa,uveoscleral pathway): (no sentence found)
hasPart(krigia,section): (no sentence found)
hasPart(recruit,musculus sphincter ani internus): (no sentence found)
hasPart(Paiute,lacrimal bone): (no sentence found)
hasPart(herring,nerve fiber): (no sentence found)
hasPart(Kickapoo,Golgi’s cell): (no sentence found)
hasPart(ayatollah,musculus adductor magnus): (no sentence found)

Figure 1: A random selection of 10 entries in PWKB, plus a retrieved sentence containing both entities (if present) in the Waterloo corpus. We use the existence of a retrieved sentence as a weak indicator of the salience of the hasPart relationship.

4.4 Same-Corpus Comparison

As we have used the same source corpus as for the TupleKB (the Waterloo corpus, Section 3.1), we have the unusual opportunity to directly compare the different extraction techniques used, given the same input. Most importantly, we observe hasPartKB has both a significantly higher overall yield of hasPart relations (Table 3) at higher precision (Table 2), given the same input corpus. Although the TupleKB was targeting a wide variety of relations, rather than just hasPart, this provides an indication that our use of generic sentences, rather than an extraction pipeline over all sentences, has yielded an advantage, at least for hasPart extraction.

5. Limitations and Discussion

Our hasPartKB complements existing resources, and is the first one that is all three of: accurate (≈90%), high coverage of common terms (within a Fifth Grade vocabulary); and salient (covers relationships a person may mention). However, the extractor still makes errors. From a random sample of incorrect extractions, we identify the following categories of error:

5.1 Ambiguous Relations (≈35% of cases)

In some cases the linguistic expression of hasPart is ambiguous, e.g., “of” can denote multiple relations, not just meronymy. Ideally, the trained model will correctly distinguish when hasPart is intended, but in practice errors occur. For example, from:

“Inflammatory cells consist of lymphoid cells, as well as mast cells, ...”

we incorrectly extract hasPart(“inflammatory cells”,”lymphoid cells”). Here, the model has taken “consist” to indicate meronymy. Additional training may help alleviate such errors.

5.2 Incorrect Pairing (≈30%)

Sometimes the model incorrectly identifies a hasPart relationship between two distant spans, for example, from:
Slugs belong to families which include snails with shells.

we incorrectly extract \texttt{hasPart(“family”, ”shell”). Again, additional training data may help alleviate such mistakes.

\section{5.3 Contextual Relationships ($\approx$25\%)}

For $\approx$25\% of the errors, an over-general, contextual term was extracted, for example, from:

\textit{Most species have specialized breathing siphons.}

the extractor finds the over-general \texttt{hasPart(“species”, ”breathing siphon”). In this context, “species” is not referring all species, but species of an organism mentioned in the previous sentence. Even without cross-sentence contextualization, such errors can arise, e.g., from \textit{Birds are animals with beaks and feathers}. we extract \texttt{hasPart(“animal”, ”beak”), an over-general extraction.}

\section{5.4 Metonymy and Factual Errors ($\approx$10\%)}

In some cases, the original sentence is incorrect from a literal reading, either due to a factual error or (more commonly) metonymy [Fass et al., 1997]. For example, given:

\textit{Spider monkeys have no thumbs, so their grasping is done with four fingers.”}

We incorrectly extract \texttt{hasPart(“spider monkey”, ”four finger”). In fact, the sentence is metonymically referring to the (unstated) hand of a spider monkey as having four fingers.

\section{5.5 The Semantics of hasPart at the Boundaries}

As well as the specific error categories, above, we note that the semantics of “having a part” itself has some ambiguity for boundary cases along three dimensions:

(a) What exactly constitutes a part: For example, is wallpaper part of a room or contained in the room?
(b) What exactly constitutes the entities being related: For example, does a person’s hand have four fingers or five?
(c) What quantification is sufficient: For example, does an airplane have a propeller?

Although these issues only affect boundary cases, they are important to note for future development.

\section{6. Conclusion}

Meronymic relations are one of the most important relationships between entities. To complement existing resources, we have presented a new knowledge-base of hasPart relationships, constructed in a novel way by using generic sentences as a source of knowledge. Empirically, the approach has yielded the first resource that is all three of: accurate ($\approx$90\%), salient (covers relationships a person may have), and has high coverage of common terms. In addition, it contains information about quantifiers, argument modifiers, and links the entities to appropriate concepts in Wikipedia and WordNet. The KB is available for the community at https://tinyurl.com/haspartkb (anonymized link).
References


Appendix: Supplemental Materials

A1. Word Sense Disambiguation Model

Our word sense disambiguation model is a variant of the GlossBERT model of Huang et al. [2019], which uses synset glosses in WordNet to disambiguate words. More specifically, given a sentence \( s \) and a set of target words \( \{w_1, w_2, ..., w_n\} \) in that sentence, the BERT encoder is used to score each \( w_j \) against a set of candidate WordNet glosses \( \{g_1^{(w_j)}, g_2^{(w_j)}, ..., g_m^{(w_j)}\} \) using the following format:

\[
s^{(w_j)} := \text{[CLS]} \ s \ \text{[SEP]} \ w_j : g_0^{(w_j)} \ \text{[SEP]}.
\]

This model is trained on the SEMCOR 3.0 training corpus [Raganato et al., 2017] using a standard sentence-pair classification loss over the hidden state of the classifier token for (denoted as \( c_{s,w_j}^{(w_j)} = \text{BERT}(s^{(w_j)} \in \mathbb{R}^{512}) \) and an additional classification layer \( (c(\cdot)) \).

In our version, the more recent RoBERTa encoder is used in place of BERT. We also employ an alternative multiple-choice ranking loss and format, following that used in [Richardson and Sabharwal, 2019] where the loss of the correct gloss \( g^* \) is computed over all other glosses (i.e., given \( c_{s,w_j}^{(w_j)} \) and the probability \( p_{w,s}^{(w_j)} \propto \exp(c_{s,w_j}^{(w_j)}) \) of \( g^* \) over all alternative glosses, our loss over our training dataset \( D \) is given as \( \mathcal{L} = \sum_{(w,s) \in D} - \log p_{w,s}^{(w_j)} \).

A2. Reference Parts Lists for Coverage Case Study

The following are the independently authored parts lists used in the coverage case-study (Section 4.2.2):

- **Dog**: eye, cheek, tongue, neck, shoulder, chest, elbow, forelimb, wrist, claw, paw, toes, hock, tail, abdomen(belly), thigh, wither, back, nose, muzzle, ear, mouth, whiskers
- **Ribosome**: small subunit, large subunit, A site, P site, E site
- **Snail**: eyespots, foot, shell, head, mouth, respiratory pore, tentacles
- **Maple**: leaf, flower, seeds, twig, trunk, bark, root, branch, crown
- **Hippo**: toes, eyes, nostrils, skin, legs, tail
- **Fish**: caudal fin, anal fin, pelvic fin, pectoral fin, gill, gill cover, mouth, eye, nostril, dorsal fins, scales, lateral line