Commonsense in Parts: Mining Part-Whole Relations from the Web and Image Tags

Niket Tandon[1] **Charles Hariman**[1] **Jacopo Urbani**[1,2] **Anna Rohrbach**[1] **Marcus Rohrbach**[3] **Gerhard Weikum**[1]
[1]Max Planck Institute for Informatics, Germany [2]VU University Amsterdam, Netherlands [3]UC Berkeley EECS and ICSI, USA

Abstract

Commonsense knowledge about part-whole relations (e.g., screen partOf notebook) is important for interpreting user input in web search and question answering, or for object detection in images. Prior work on knowledge base construction has compiled part-whole assertions, but with substantial limitations: i) semantically different kinds of part-whole relations are conflated into a single generic relation, ii) the arguments of a part-whole assertion are merely words with ambiguous meaning, iii) the assertions lack additional attributes like visibility (e.g., a nose is visible but a kidney is not) and cardinality information (e.g., a bird has two legs while a spider eight), iv) limited coverage of only tens of thousands of assertions.

This paper presents a new method for automatically acquiring part-whole commonsense from Web contents and image tags at an unprecedented scale, yielding many millions of assertions, while specifically addressing the four shortcomings of prior work. Our method combines pattern-based information extraction methods with logical reasoning. We carefully distinguish different relations: physicalPartOf, memberOf, substanceOf. We consistently map the arguments of all assertions onto WordNet senses, eliminating the ambiguity of word-level assertions. We identify whether the parts can be visually perceived, and infer cardinalities for the assertions. The resulting commonsense knowledge base has very high quality and high coverage, with an accuracy of 89% determined by extensive sampling, and is publicly available.

Introduction

Motivation and Problem. We all know that a thumb is part of a hand, and that a goalkeeper is part of a soccer or hockey team. For machines this kind of commonsense is not obvious at all, yet many modern computer tasks – like computer vision, Web search, question answering, or ads placement – require this kind of background knowledge to simulate human-like behavior and quality. For example, suppose a visual object detection algorithm has recognized two wheels, pedals and a chain in an image or video; a smart interpretation could then harness knowledge to infer that there is a bike in this scene. This would be a novel element and potential performance booster in computer vision (Rohrbach, Stark, and Schiele 2011). However, there is no comprehensive part-whole knowledge base available today.

There has been considerable research to automatically acquire part-whole knowledge in fields like linguistics and cognitive sciences (Winston, Chaffin, and Herrmann 1987), ontology engineering (Keet and Artale 2008), computer vision (Chen, Shrivastava, and Gupta 2013), and knowledge base construction (Girju, Badulescu, and Moldovan 2006; Havasi, Speer, and Alonso 2007; Speer and Havasi 2012; Tandon, de Melo, and Weikum 2011). Despite these efforts, however, the by far largest commonsense knowledge collections with part-whole relations are manually constructed and curated by experts like WordNet (Fellbaum 1998) or by crowdsourcing like ConceptNet (Havasi, Speer, and Alonso 2007; Speer and Havasi 2012). Their coverage is far from anywhere near being complete. Automated efforts like (Tandon, de Melo, and Weikum 2011) or NEIL (Chen, Shrivastava, and Gupta 2013) had to cope with fairly noisy inputs, like n-gram corpora or images; so their outputs for part-whole relations are quite inferior in quality compared to WordNet or ConceptNet. Other knowledge base (KB) projects, such as Cyc, DBpedia, Freebase, NELL, Probase, or YAGO, have focused on factual knowledge about individual entities, with commonsense limited to hypernymy/taxonomic relations.

Thus, prior part-whole KB's have major limitations:

i) The automated efforts to compile part-whole knowledge, such as (Chen, Shrivastava, and Gupta 2013) or (Tandon, de Melo, and Weikum 2011) conflate different kinds of part-whole relations into a single generic relation *partOf* and miss out on the semantic differences between *physicalPartOf* (e.g., wheel physicalPartOf bike), *memberOf* (e.g., cyclist memberOf team), or *substanceOf* (e.g., rubber substanceOf wheel).

ii) In all part-whole KB's except WordNet, the arguments of the relations (e.g., screen, notebook) are merely words with ambiguous meaning, whereas they should ideally be unique word senses, for example, by disambiguating them onto WordNet synsets.

iii) In all part-whole KB's, the assertions are merely qualitative; there is no information about either visibility or cardinality. Existing KB's lack the distinction between *visible* and *invisible* physicalPartOf (e.g., for an ordinary human, `nose` physicalPartOf `human` is visible, while `kidney` physicalPartOf `human` is invisible). Further, it

could be important to know that a bike has two wheels rather than three, and that a car has one steering wheel rather than two. These distinctions are crucial for visual applications.

iv) The coverage of part-whole knowledge is very limited. For example, ConceptNet contains only 1,086 instances of various part-whole relations in total. It has the notion of a memberOf relation and knows the concepts of a cyclist and sport team, yet does not have any memberOf information for these concepts.

The goal and contribution of this paper is to overcome these limitations and build a comprehensive, semantically refined, high-quality part-whole KB, *PWKB* for short.

Approach and Contribution. We developed a complete knowledge-acquisition pipeline that combines statistical techniques with logical inference. Our method includes an extension of pattern-based extraction that substantially improves the extraction quality on large and noisy text corpora like the Wikipedia full text and the Google n-gram collection. Subsequently, we further eliminate false positives by devising rules for constraint checking, and we also infer additional assertions by logical deduction rules. For high coverage, these rules need to consider candidate assertions over multi-word noun phrases. To properly handle these, we have devised a new technique to integrate such phrases into WordNet. Finally, we developed novel techniques to enhance the assertions with visibility attribute values by tapping into image tags obtained from 100 Million Flickr images, and cardinality attribute values by tapping into Google-books n-grams from multiple languages.

We successfully tackle all four of the aforementioned limitations: i) distinguishing physicalPartOf, memberOf, and substanceOf, ii) mapping all arguments of our assertions to WordNet senses, thus eliminating ambiguity and redundancy, iii) inferring visibility and cardinality information for many instances of the various part-whole relations, and iv) building a large PWKB with about 6.75 million assertions – orders of magnitude larger than WordNet or ConceptNet while having similar of better quality. PWKB is publicly available at `http://tinyurl.com/partwholekb`.

We compare our PWKB against state-of-the-art baselines, most notably ConceptNet, by sampling the various relations and manually assessing their quality. As an extrinsic use-case, we show that our PWKB can contribute to the computer vision task of object classification in images.

Overview

Our method uses WordNet (Fellbaum 1998) to disambiguate concepts extracted from Web and image tags. WordNet is the most popular lexical database for English, which groups words into sets of synonyms called *synsets* or word *senses*. WordNet connects synsets by various relations. Relevant for us are hypernymy/hyponymy (type, subclass), which relate broader concepts to more specific ones, and three kinds of part-whole relationships : *(physical) partOf*, *memberOf*, and *substanceOf*.

Due to the nature of the part-whole relations, not every synset can be accepted as left argument (i.e., part – domain of the relation) or right argument (i.e., whole – range of the relation). For instance, physicalPartOf restricts both domain and range to be physical, memberOf restricts the range to be abstract, while substanceOf restricts the domain to be substance. Therefore, we first consider the synsets that are hyponyms of *Abstract Entity* (V_a) or *Physical Entity* (V_p). We assume $V_a \cap V_p = \emptyset$. WordNet has exceptions to this disjointness: around 1,000 synsets have hypernyms in both V_a and V_p (McCarthy 2001), e.g., `roller coaster`. For these we only use hypernyms in V_p.

Abstract entities include, for example, *teams*, *organizations*, *music*, *poems*, etc. Physical entities include everything that one can possibly touch, such as *bikes*, *cars*, *fingers*, *bones*, etc. Furthermore we distinguish the synsets under *Substance*, denoted as V_s, which is a hyponym of the physical entity, so that $V_s \subset V_p$. *Substance* synsets include for examples *iron*, *oxygen*, *clay*, *oil*, etc. Table 1 summarizes the part-whole relations with our type restrictions.

Table 1: Part-whole relations with type restriction

r	**domain**(r)	**range**(r)	**example**
$\prec_P$	V_p	V_p	wheel $\prec_P$ bike
$\prec_M$	$V_p \cup V_a$	V_a	cyclist $\prec_M$ team
$\prec_S$	V_s	V_p	rubber $\prec_S$ wheel

Our goal is to mine new assertions for the three WordNet part-whole relations and enrich them with two new attributes: *visibility* and *cardinality*. The first indicates whether the part can be visually perceived. The second attribute defines the number of parts in the whole.

Our method proceeds in three phases:

- *Phase 1 – KB Construction:* We extend statistical techniques for pattern-based extraction by introducing weighted seeds in candidate scoring to improve the output quality on large and noisy text corpora. This phase gives us candidate assertions for part-whole relations, and we map the arguments of the candidate assertions to WordNet senses.
- *Phase 2 – KB Enrichment:* The candidate assertions from Phase 1 still contain many false positives. We devise logical inference rules to enforce consistency and obtain cleaner assertions. Additionally, we propose deduction rules for deriving additional assertions, enlarging the PWKB. A key novelty here is that these rules apply to assertions over multi-word noun phrases. We have developed techniques to handle these by carefully extending the WordNet taxonomy.
- *Phase 3 – KB Enhancement:* We enhance the part-whole relations by two new attributes *visibility* and *cardinality*. We develop a novel technique that exploits image tags to detect the visibility of the part in the whole. For cardinality, we exploit the grammatical structure of German and Italian to handle cases which cannot be easily dealt with in English.

Phase 1: KB Construction

We construct our PWKB (Part-Whole Knowledge Base) by introducing novel extensions of the state-of-the-art pattern based extraction techniques (with a new scoring model) and disambiguation techniques (extending from words to phrases).

Extraction of $\prec_P$,$\prec_M$,$\prec_S$. We use a pattern-based information extraction approach, following (Tandon, de Melo, and Weikum 2011), to obtain candidate patterns from text. This method requires a small number of high-quality seed assertions to bootstrap the identification of extraction patterns. As the text source, we use the full text of Wikipedia and the Google-Web n-grams. As seeds, we pick 1,200 instances of the physicalPartOf, memberOf, and substanceOf relations of WordNet. Patterns are automatically obtained by matching the seed pairs in our input corpora, and extracting the essential words between the two concepts (i.e., considering only words with certain part-of-speech tags). For example, the seed `goalkeeper` $\prec_M$ `team` leads to the extraction pattern <Noun> *of the* <Noun>.

Scoring Model for Candidate Ranking. The quality of patterns varies widely. We identify good patterns regarding two aspects: i) patterns should co-occur with many distinct seeds (not just very frequently with some seeds), and ii) patterns should discriminate between the three part-whole relations that we aim to populate. The Specificity Ranker (SR) of (Tandon, de Melo, and Weikum 2011) already takes the first aspect into account. However, we improve this prior model by introducing a notion of weighted support and by considering the second aspect.

Let $\sigma_{SR}(p_i)$ denote the score that SR assigns to pattern p_i, using all seeds for all relations, and let $\sigma_{SR}(p_i|R_j)$ be the score if only seeds for relation R_j (e.g., $\prec_M$) are used. We leverage these SR scores as weights for scoring the candidate assertions that result from the obtained patterns. The *weighted support* of candidate assertion a_k is

$$supp(a_k) = \sum_{p_i} \sigma_{SR}(p_i)\delta(p_i, a_k)$$

where $\delta(p_i, a_k)$ is 1 if p_i co-occurs with a_k and 0 otherwise. Analogously, we define the *R_j-specific weighted support* for a_k as

$$supp(a_k|R_j) = \sum_{p_i} \sigma_{SR}(p_i|R_j)\delta(p_i, a_k)$$

This is the basis for defining the *discriminative strength* of a_k for R_j:

$$str(a_k|R_j) = \sum_{\nu \neq j} \left(\frac{supp(a_k|R_j)}{supp(R_j)} - \frac{supp(a_k|R_\nu)}{supp(R_\nu)} \right)$$

where $supp(R_j) = \sum_{a_k} supp(a_k|R_j)$.

Finally, we normalize both support and strength, to yield values between 0 and 1, and combine them into the overall score of assertion candidate a_k:

$$\sigma(a_k) = \frac{e^{supp(a_k)}}{1 + e^{supp(a_k)}} \; \frac{e^{str(a_k)}}{1 + e^{str(a_k)}}$$

Thus, we can rank candidates and apply thresholding to reduce false positives.

Mapping Words and Phrases to Senses. The selected assertions are word pairs and hence ambiguous. We extend the IMS (ItMakesSense) tool (Zhong and Ng 2010) to disambiguate words onto WordNet senses. IMS operates at a word level and can thus not handle multi-word noun phrases. Our novel contribution is to add a new layer on top of IMS to solve this problem. First, we perform noun phrase chunking on the input sentence where the assertion occurs. We use the widely used OpenNLP Chunker (`opennlp.apache.org`). Next, for every noun phrase, we identify and disambiguate its lexical head using IMS (e.g., the out of WordNet phrase `the electrical plant` to `plant#1`). This yields canonicalized assertions for our part-whole relations, with unique senses and free of redundancy. This also enables us to apply type-restrictions based on the domain and range of the relations (see Table 1) to further filter the assertions.

Phase 2: KB Enrichment

We enrich the PWKB by proposing logical inference rules for deduction and consistency.

Increasing Coverage

We improve the PWKB coverage by applying the following two deduction rules:

C1. Transitivity: $(a \prec b \wedge b \prec c) \Rightarrow a \prec c$

C2. Inheritance: $(a \prec b \wedge c$ `hyponymOf` $b) \Rightarrow a \prec c$

We exploit the fact that physicalPartOf and substanceOf are transitive (Keet and Artale 2008) and perform a 2-step transitive closure. We do not consider the full transitive closure as it tends to produce too many trivial assertions (e.g., `atom` $\prec_P$ `matter`). We propose $C2$ to propagate part-whole relations to hyponyms of the whole. For example, having the knowledge: `wheel` $\prec_P$ `bike` and `mountain bike` `hyponymOf` `bike`, we infer: `wheel` $\prec_P$ `mountain bike`.

Such an enrichment will fail if a concept is absent in WordNet (e.g., `racing bike`). Thus, we extend WordNet by extracting multi-word noun phrases and mapping them to the proper WordNet hypernyms. We select candidate multi-word noun phrases from bigrams and trigrams in the Google n-grams corpora (Brants and Franz 2006), as noun phrases are usually upto three tokens. We restrict these by their part-of-speech tags to consider only nouns and adjectives as a prefix (e.g., *mountain bike*, *lightweight racing bike*).

These restrictions alone are insufficient and we still get many noisy phrases. So, we rank candidate phrases by NPMI (normalized pointwise mutual information) (Bouma 2009) and also by occurrence frequency. Neither of these criteria alone works robustly, so we propose to accept noun phrases only if *both* measures are above specified thresholds.

Specifically, we compute histograms for the NPMI values and for the log-frequencies. A noun phrase is retained only if its NPMI is above the 90% quantile (i.e., very high), or, its NPMI is above the 10% quantile (i.e., above noise level) and the log-frequency is above the median (i.e., substantial).

For each of the retained noun phrases, we perform sense disambiguation based on the phrase's head word, using the technique of Phase 1. These additional phrases extend WordNet, with hypernymy links between phrases and their head-word senses, enabling us to perform Rule $C2$ for phrases absent in WordNet.

While $C2$ is useful in many cases (e.g., deducing that mountain bikes have wheels, too), it also comes with the risk of generating false assertions, e.g., that mountain bikes have headlights. Here we rely on the pragmatic assumption that WordNet's hyponymy links induce subsumptions between the sets of instances for the respective synsets/classes. Our experimental evaluation reports on the benefits and risks of the deduction rules.

Improving Quality

We improve the PWKB quality by checking for inconsistencies and eliminating false assertions, using the following two constraints:

Q1. Irreflexivity: $\neg\,(a \prec a)$

Q2. Acyclicity: $\neg\,(a \prec b \wedge b \prec a)$

We drop assertions that violate the first type of inconsistency. For the second type, we detect all cycles of length $\leqslant 3$ and break each cycle by dropping the assertion with the lowest score computed in Phase 1.

Phase 3: KB Enhancement

We enhance the PWKB assertions by introducing two new attributes: visibility and cardinality.

Visibility Attribute

Our goal is to determine which physical parts of a whole are visible (for an ordinary human, e.g., not a mechanic or surgeon). If a and b co-occur in an image and we have the knowledge that $a \prec_P b$, then a is visible. The superscript V is used for $\prec_P$ to denote visibility (e.g. `license plate` $\prec_P^V$ `car`) while NV denotes non-visibility (e.g. `automatic brake system` $\prec_P^{NV}$ `car`).

We could consider obtaining visibility information directly from images, or alternatively, from annotations of images like captions and tags. To compare these two approaches, we computed co-occurrence statistics from i) running a visual object detector (LSDA (Hoffman et al. 2014)) versus ii) user-provided tags that annotate Yahoo! Flickr images (Shamma 2014). We compared both results against the already compiled $\prec_P$ assertions for a sample of 100K Flickr images. We obtained ca. 12,000 positive matches with LSDA object detections versus ca. 26,000 with tags. Thus, image annotations give better coverage.

We thus used Flickr tags to compute $\prec_P^V$ at large scale. We set the visibility of $a \prec_P b$ to true, if a and b co-occur as tags of at least a certain number of Flickr images. In the experiments we set this co-occurrence threshold to two.

Cardinality Attribute

Consider the computer vision task of recognizing different types of cycles (unicycle, bicycle, tricyle). Knowing that a unicycle has one wheel, bicycle has two, whereas tricycle has three wheels, will help the object detector. This motivates us to further extend the PWKB by cardinality information, where we distinguish the cases *1,2,3+* and *uncountable* denoted as ω. The uncountable case applies, e.g., to the fur of a dog or pebbles of a beach. We represent the cardinality as an attribute that we add to the $\prec_P$ and $\prec_M$ relations, and denote it by a superscript c; e.g., `wheels` $\prec_P^{\{2,V\}}$ `bike` denoting that a bike has two visible wheels. The method for inferring c in $a \prec^c_{r\in\{P,M\}} b$ has three steps:

1) Determine whether a and b are countable. We use `wiktionary.org` to look up if a word is countable.
2) If the dictionary does not have that information for a, then we compute the frequencies $f_{sin}(a)$ and $f_{plu}(a)$ of the occurrences of a in singular and plural form within a text corpus, using standard grammar rules. If $f_{sin}(a) \gg f_{plu}(a)$ or $f_{plu}(a) \gg f_{sin}(a)$, then we consider a to be uncountable. The threshold for these comparisons is determined from a set of known uncountable concepts.
3) We compare the grammatical forms of a and b. If the majority of a and b occurrences in the same sentence is in the form {singular, singular} (e.g., {handle, bike}), then we set $c = 1$. If the majority of occurrences has the form {plural, singular} (e.g., {wheels, bike}), and the supporting patterns include a numeric token (e.g., 2,3, ...), numbers in text forms ("two", "three", ...), or cues such as "both" or "couple of", then we set $c = 2$ for patterns indicating 2, and $c = 3+$ for all others.

   For the remaining cases where a, b co-occur in the forms {singular, plural} or {plural,plural}, we use default settings: $c = 1$ for $\prec_P$ and $c = 3+$ for $\prec_M$.

As English articles and determiners ("the", "some", "any", etc.) do not easily discriminate singular and plural, Step 3 is error-prone. We thus tapped German and Italian corpora (Google-books n-grams) where plural forms are more easily detectable by variants of articles and inflections of nouns. For the resulting assertions, we use Wiktionary to map the German or Italian words back to English.

Results and Experimental Comparisons

Input Data. We construct PWKB from the following:

i) *Google Web 1T N-gram* Dataset Version 1 (Brants and Franz 2006) which contains frequencies of n-grams (n=1,..,5) for English web pages;

ii) *Wikipedia* (2010 snapshot) (Shaoul 2010) which contains all English Wikipedia articles as of April 2010;

iii) *Google books n-grams* in English, Italian and German (2010 snapshot) which contains POS-tagged 4-grams and 5-grams from millions of books;

iv) *Yahoo! Flickr* images (Shamma 2014), which contains 100 million images from `www.flickr.com` with title, description, and tags.

Baselines. We consider two types of baselines: *KB baselines* and *methodology baselines*.

Table 2: Precision (first line) and coverage (second line)

	$\prec_P$	$\prec_M$	$\prec_S$	vis.	card.	overall
WN	1.00	1.00	1.00	1.00	-	1.00
	12892	3714	609	1304	-	17215
SR	0.19	0.20	0.23	0.16	-	0.20
	0.49M	0.49M	0.15M	0.15M	-	1.13M
CN	0.82	0.45	0.43	0.85	-	0.68
	921	516	56	665	-	1493
NEIL	0.15	-	-	0.15	-	0.15
	68	0	0	68	-	68
PWKB	0.89	0.96	0.71	0.98	0.80	0.89
	6.65M	0.04M	0.06M	0.74M	6.69M	6.75M

As KB baselines, we consider the manually constructed *WordNet (WN)*, the recall-oriented text-based *Specificity Ranker (SR)* of (Tandon, de Melo, and Weikum 2011), the image-based *NEIL* (Chen, Shrivastava, and Gupta 2013), and the crowdsourcing-based *ConceptNet (CN)* of (Havasi, Speer, and Alonso 2007). The part-whole relations of SR and CN are not refined into the more specific relations that PWKB has. To make a fair comparison with SR and CN, we partitioned its assertions into the relations $\prec_P$, $\prec_M$, $\prec_S$ by domain-range type restriction (see Table 1), and set the visual attribute in case the arguments of $\prec_P$ map to Flickr tags (identically to our method). SR and CN contain many part-whole assertions that are encyclopedic rather than commonsense (e.g., Castro-district partOf California), in addition to noise (e.g., misspellings). Such concepts are not mappable to WordNet, so we drop them. Further, for SR and CN, we optimized the score thresholds for coverage. This explains the difference in numbers from original papers on CN and SR versus our setting.

As methodology baselines for scoring assertions, we include the widely used *Espresso* (Pantel and Pennacchiotti 2006) and SR, both run on our input data. For word disambiguation, we compare against the widely used and strong *Most Frequent WordNet Sense (MFS)* heuristic. For the quality of noun phrases, *NPMI* alone is used as a baseline.

PWKB Statistics and Evaluation. In total, PWKB contains 6.75 Million assertions for the three fine-grained part-whole relations, with disambiguated arguments, and, to some extent, with the two additional attributes. To evaluate the quality of PWKB, we compiled a random sample of 1000 assertions from $\prec_P$,$\prec_M$,$\prec_S$, with at least 200 assertions from each relation. We relied on human annotators to judge each assertion. An assertion was marked as correct if the judge stated that the disambiguation of the arguments was correct and the part-whole relation was correct.

For the baselines, we generously evaluated the assertions based on their surface forms as the baselines do not have disambiguated arguments. We compute the precision as $\frac{c}{c+i}$, where c and i are the counts of correct and incorrect assertions, respectively. For statistical significance, we computed Wilson score intervals for $\alpha = 95\%$ (Brown, Cai, and DasGupta 2001). The inter-annotator agreement for three judges in terms of Fleiss' κ was 0.78. We used majority voting to decide on the gold-standard labels.

The per-relation results are reported in Table 2. PWKB clearly outperforms all baselines in terms of coverage. In terms of quality, PWKB has an overall average precision of 89%, which seems good enough for many downstream applications (e.g., in computer vision) where commonsense can be used for distant supervision or as a prior in probabilistic computations. Such applications need to cope with uncertain inputs anyway, so ∼90% precision is useful.

Table 3: PWKB anecdotal examples

mouth#1 $\prec_P^{\{1,V\}}$ man#1	electron#1 $\prec_P^{\{3+,NV\}}$ atom#1
sheep#1 $\prec_M^{3+}$ herd#1	musician#2 $\prec_M^{2}$ duet#2
fibre#1 $\prec_S$ cloth#1	steel#1 $\prec_S$ boiler#1

PWKB is much larger than all prior KB's while having higher precision than all except the manually curated WordNet. This holds also for the visibility assertions, where we outperform NEIL, constructed from 2M images, by an order of magnitude. Table 3 shows anecdotal results from PWKB.

Evaluating the PWKB Construction Pipeline. We evaluated the performance of the components of the three-phase PWKB pipeline. For each phase, we had three judges assess the output. For statistical significance, we again computed Wilson score intervals for $\alpha = 95\%$.

For the first phase – the initial construction of $\prec_P$,$\prec_M$,$\prec_S$, assertion ranking is the most important component which in turn relies on the ranking of patterns. Our assertion ranking model (0.85±0.05) outperforms the baseline Espresso ranking (0.34±0.07) and also the Specificity Ranker (0.55 ±0.06) by a large margin. For the disambiguation of arguments, our IMS-based method (0.80±0.07) achieves substantially better precision than the MFS baseline (0.70±0.07). Table 5 lists some prominent patterns for the three part-whole relations. Note that some of them are of mixed quality: good for recall, but poor in precision – for example, "y's x" for $x \prec_P y$, which would be matched by "Alice's husband". Note, however, that the patterns are further restricted by the domain and range of the relations (Table 1). Candidates such as (x=Alice, y=husband) are rejected because Alice is an instance rather than a concept of WordNet type "physical entity".

For the second phase – enrichment, our noun phrase ranker (0.60±0.04) significantly outperforms the baseline NPMI (0.25 ± 0.05), yielding 36,498 high quality noun phrases that we attach to WordNet. We performed an ablation study on the influence of the logical rules; Table 4 shows the results. The C rules for deduction increased the coverage from ca. 382K assertions to 6.75M. The Q rules for constraint checking, on the other hand, were able to remove nearly 150K false assertions that exhibited inconsistencies. For coverage, each of the two rules individually yields a major increase in the size of the PWKB; their combined effect boosts the size even more. Note, though, that even without any logical rules at all, PWKB with 382K assertations is already an order of magnitude larger than WordNet or any other prior KB of similarly high quality.

In the third phase – cardinality inference, our method achieved a precision of 0.80±0.07, significantly improving upon relying solely on English (0.61±0.09). As for the cardinality values, we found that we achieve high precision for

Table 5: Prominent patterns for PWKB relations

$\prec_P$	$\prec_M$	$\prec_S$
y have x	x (be) member of y	y (be) made of/from x
y 's x	x be in y	x found in y
x be part of y	x be of y	y (be) composed of x

cardinalities 1 and 2. However, we did not compute many assertions with 3+. This was because our heuristic method preferred a cardinality of ω (uncountable) in many cases.

Use Case: Image Classification. In this experiment we evaluate the benefit of PWKB for image classification. The task is to recognize unseen image categories by transferring knowledge from known categories. For example, being able to recognize *wheels* of cars and *seats* of chairs might allow us to recognize a *wheelchair* even if we have no training image for *wheelchair*. This "zero-shot recognition" is crucial as many categories have no (or very sparse) training data.

For this task, we repeated the experiment of (Rohrbach, Stark, and Schiele 2011), who trained classifiers for 811 part categories to recognize unseen categories. To associate the unseen categories with the parts, part-whole patterns (Berland and Charniak 1999) were retrieved with Yahoo search. For comparability, we used the same visual features and the same image classification architecture as in the original study. We solely replaced the original part-whole relation with the relations from PWKB.

On the zero-shot task of recognizing 200 unseen categories, the top-5 accuracy increases from 23.8% (best single part-whole variant *Yahoo Snippets*) to 25.5% by using PWKB. We note that (Rohrbach, Stark, and Schiele 2011) achieved better performance, up to 35% accuracy, with a hierarchy-based transfer or combining multiple measures, which is orthogonal to the use of our part-whole knowledge. We could combine the PWKB asset with this technique. Note that this task is inherently difficult; we are not aware of any methods that achieve more than 40% accuracy.

Related Work

Part-whole relations are studied in several disciplines.

Philosophy. In *mereology*, there is wide consensus that the part-whole relation should be modeled as a weak partial ordering, i.e., a property that is *reflexive*, *transitive*, and *antisymmetric* (Varzi 2014). (Winston, Chaffin, and Herrmann 1987) and (Keet and Artale 2008) discuss semantic variants of part-whole relations in natural languages. (Smith et al. 2005) discusses the specific setting of biomedical ontologies. Our work, just like WordNet, follows the conceptual framework of (Winston, Chaffin, and Herrmann 1987).

Computational Linguistics. In contrast to the extensive work on lexico-syntactic patterns for hyponymy/hypernymy and taxonomy induction, there is relatively little work on extracting meronymy/holonymys concept pairs. (Berland and Charniak 1999) used two Hearst patterns, on genitive forms, to extract candidate pairs and used statistical measures for ranking. However, the high ambiguity of genitive forms ('s, of) led to very limited results. (Girju, Badulescu, and Moldovan 2003; 2006) extended and generalized this approach by using additional, still handcrafted, patterns and adding constraints about the lexical hypernyms (in WordNet) that concept pairs need to be in a meaningful part-whole relation. The method achieved a precision of ca. 80% on a few 10,000 sampled sentences from news corpora.

Table 4: Ablation study on the logical rules of Phase 2

	No Rule	+Rule 1	+Rule 2	+Rule 1,2
C rules	382K	+55K	+700K	+6.4M
Q rules	+6.4M	-476	-146K	-146K

(Pantel and Pennacchiotti 2006) developed the Espresso algorithm that extended prior work on seed-based pattern induction (such as (Ravichandran and Hovy 2002)) by introducing PMI-based pattern rankings. This resulted in a precision of 80% for part-whole extractions from benchmark corpora. The output pairs were not sense-disambiguated and the output size was small. (Ruiz-Casado, Alfonseca, and Castells 2007) harnessed Wikipedia and patterns near hyperlinks, and achieved a precision for meronymy/holonymy $\leqslant 70\%$ in small-scale experiments.

Recent works on acquisition of lexical relations include (Tandon, de Melo, and Weikum 2011) and (Ling, Clark, and Weld 2013). The former addressed a wide variety of commonsense relations without specific concern for part-whole, whereas the latter was geared for meronyms among biological concepts. (Ittoo and Bouma 2010; 2013) studied refined classes of part-whole relations, based on the taxonomy of (Keet and Artale 2008). They extended prior work by using different seed sets for different part-whole relations extracted from Wikipedia texts, and achieved an overall precision of ca. 80% for an output of ca. 10,000 concept pairs.

None of these prior works was designed for constructing a large, fine-grained and disambiguated part-whole KB.

Knowledge Acquisition. Commonsense acquisition projects like Cyc (Lenat 1995; Matuszek et al. 2005), ConceptNet (Havasi, Speer, and Alonso 2007; Speer and Havasi 2012), NELL (Carlson et al. 2010; Mitchell et al. 2015), and WebChild (Tandon et al. 2014) have compiled large amounts of commonsense knowledge. Among these, only Cyc and ConceptNet contain a sizable number of instances of part-whole relations. Cyc has relied on manual expert input, which is expensive and does not scale. ConceptNet is based on crowdsourcing, but lacks argument disambiguation and semantic refinement.

The NEIL project (Chen, Shrivastava, and Gupta 2014) has embarked on discovering part-whole and other commonsense relations about scenes by analyzing a large number of images. So far the project has acquired around a hundred instances of a generic part-whole relation.

Conclusions

We presented the methodology for automatically constructing a large, high-quality KB of part-whole relations. We improve the state of the art in several ways: i) by capturing many instances for the refined relations *physicalPartOf*,

memberOf, and *substanceOf*, ii) by disambiguating the arguments of assertions onto WordNet senses, iii) by additionally inferring visibility and cardinality information for part-whole instances, and iv) by doing all this at very large scale using a novel combination of statistical pattern-based extraction and logical reasoning. The resulting KB contains more than 6.75 million assertions; sample-based manual assessment shows that this output is of high quality. As future work, we aim to extract facts from additional languages and an even wider range of Web contents.

Acknowledgments. Marcus Rohrbach was supported by a fellowship within the FITweltweit-Program of the DAAD. Jacopo Urbani was partly funded by the NWO VENI project 639.021.335.

References

Berland, M., and Charniak, E. 1999. Finding parts in very large corpora. In *ACL*.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 31–40.

Brants, T., and Franz, A. 2006. {Web 1T 5-gram Version 1}.

Brown, L. D.; Cai, T. T.; and DasGupta, A. 2001. Interval estimation for a binomial proportion. *Statistical Science*.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Jr., E. H.; and Mitchell, T. 2010. Toward an architecture for never-ending language learning. In *AAAI*, 1306–1313. AAAI Press.

Chen, X.; Shrivastava, A.; and Gupta, A. 2013. Neil: Extracting visual knowledge from web data. In *CVPR*.

Chen, X.; Shrivastava, A.; and Gupta, A. 2014. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Girju, R.; Badulescu, A.; and Moldovan, D. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *NAACL*.

Girju, R.; Badulescu, A.; and Moldovan, D. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*.

Havasi, C.; Speer, R.; and Alonso, J. 2007. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *RANLP*.

Hoffman, J.; Guadarrama, S.; Tzeng, E.; Donahue, J.; Girshick, R.; Darrell, T.; and Saenko, K. 2014. LSDA: Large scale detection through adaptation. In *NIPS*.

Ittoo, A., and Bouma, G. 2010. On learning subtypes of the part-whole relation: do not mix your seeds. In *ACL*.

Ittoo, A., and Bouma, G. 2013. Minimally-supervised extraction of domain-specific part–whole relations using wikipedia as knowledge-base. *Data & Knowledge Engineering* 85:57–79.

Keet, C. M., and Artale, A. 2008. Representing and Reasoning over a Taxonomy of Part-Whole Relations. *Applied Ontology*.

Lenat, D. B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*.

Ling, X.; Clark, P.; and Weld, D. S. 2013. Extracting meronyms for a biology knowledge base using distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 7–12. ACM.

Matuszek, C.; Witbrock, M.; Kahlert, R.; Cabral, J.; Schneider, D.; Shah, P.; and Lenat, D. 2005. Searching for common sense: Populating Cyc from the Web. In *Proc. AAAI 1999*.

McCarthy, D. 2001. Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences. PhD. thesis.

Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; and Welling, J. 2015. Never-ending learning. In *AAAI*.

Pantel, P., and Pennacchiotti, M. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *ACL*.

Ravichandran, D., and Hovy, E. 2002. Learning surface text patterns for a question answering system. 41–47. ACL.

Rohrbach, M.; Stark, M.; and Schiele, B. 2011. Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting. In *CVPR*.

Ruiz-Casado, M.; Alfonseca, E.; and Castells, P. 2007. Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data & Knowledge Engineering* 61(3):484–499.

Shamma, D. 2014. One hundred million Creative Commons Flickr images for research. `http://labs.yahoo.com/news/yfcc100m/`.

Shaoul, C. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta*.

Smith, B.; Ceusters, W.; Klagges, B.; Köhler, J.; Kumar, A.; Lomax, J.; Mungall, C.; Neuhaus, F.; Rector, A. L.; and Rosse, C. 2005. Relations in biomedical ontologies. *Genome Biology* 6(R46).

Speer, R., and Havasi, C. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*.

Tandon, N.; de Melo, G.; Suchanek, F. M.; and Weikum, G. 2014. Webchild: harvesting and organizing commonsense knowledge from the web. In *WSDM*.

Tandon, N.; de Melo, G.; and Weikum, G. 2011. Deriving a web-scale common sense fact database. In *AAAI*.

Varzi, A. C. 2014. Mereology. In Zalta, E. N., ed., *Stanford Encyclopedia of Philosophy*.

Winston, M. E.; Chaffin, R.; and Herrmann, D. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science*.

Zhong, Z., and Ng, H. T. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *ACL*.