How Large Are Lions?
Inducing Distributions over Quantitative Attributes

Yanai Elazar[*]
Bar Ilan Univesity
yanaiela@gmail.com

Abhijit Mahabal[†]
Pinterest
amahabal@gmail.com

Deepak Ramachandran
Google Research
ramachandrand@google.com

Tania Bedrax-Weiss
Google Research
tbedrax@google.com

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

Most current NLP systems have little knowledge about quantitative attributes of objects and events. We propose an unsupervised method for collecting quantitative information from large amounts of web data, and use it to create a new, very large resource consisting of distributions over physical quantities associated with objects, adjectives, and verbs which we call Distribution over Quantities (DoQ)[1]. This contrasts with recent work in this area which has focused on making only relative comparisons such as "Is a lion bigger than a wolf?". Our evaluation shows that DoQ compares favorably with state of the art results on existing datasets for relative comparisons of nouns and adjectives, and on a new dataset we introduce.

1 Introduction

How much does a lion weigh? How tall can they be? When do people typically eat breakfast? And, how long are concerts? Most people would know at least an approximate answer to these questions, many of which fall under the (somewhat ill-defined) notion of commonsense knowledge, and some (but certainly not all) of which exist in resources such as Wikipedia, or in knowledge graphs like Freebase (Bollacker et al., 2008). Natural Language Understanding systems should also know (at least approximately) the answers to these questions, to better support Question Answering and Textual Entailment (Dagan
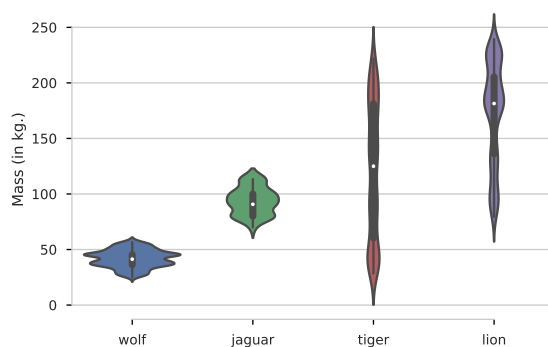
(a) Mass distributions for multiple animals.
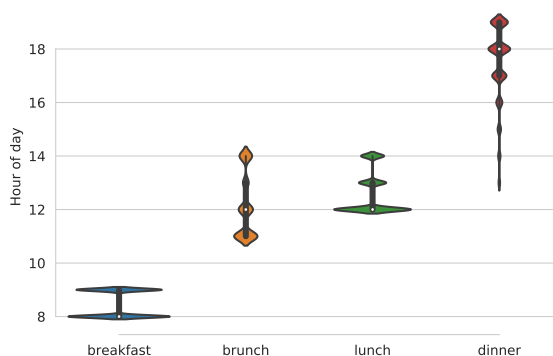
(b) Typical hours for different meals of the day.

Figure 1: Examples of different objects from DoQ in the MASS and TIME dimensions.

et al., 2013) and, more generally, in order to support reasoning about events described in natural language and converse with people naturally.

Acquiring commonsense knowledge from natural language text has been the subject of a lot of recent work. These approaches focus on facilitating comparisons between quantitative attributes of nouns (Bagherinezhad et al., 2016; Forbes and

---

Choi, 2017; Yang et al., 2018), intensity of adjectives (De Melo and Bansal, 2013; Cocos et al., 2018) and coarse classification of events duration and relative order (Gusev et al., 2011; Ning et al., 2018). However, they do not have complete coverage even for comparable objects, as a result of how they are acquired, and lack the ability to assign a numerical value to objects and events ("How hot is it in New York?"), which is often useful for reasoning, text generation, and other tasks.

In this work, we propose a method for acquiring distributions over ten dimensions: TIME, CURRENCY, LENGTH, AREA, VOLUME, MASS, TEMPERATURE, DURATION, SPEED, and VOLTAGE. We do this for nouns (e.g. elephant, airplane, NBA game), adjectives (e.g. cold, hot, lukewarm) and verbs (e.g. eating, walking, running). This results in a large resource we call Distribution over Quantities (DoQ) – over $350K$ triples each observed over 1000 times. Examples of entries in DoQ depicting MASS and TIME distributions are shown in Figures 1a and 1b.[2]

We develop DoQ by extracting and aggregating quantitative information from the web, in English, and collecting co-occurring objects from their surroundings. The quantitative information is normalized and associated with units to determine the relevant dimensions such as TEMPERATURE or MASS. As we show, despite the inherent noise in such an acquisition process due to extraction errors and reporting bias (Gordon and Van Durme, 2013), rather simple denoising methods result in a relatively clean resource, with very high coverage and good accuracy.

DoQ is significantly more comprehensive and accurate than any other related resource we know of. For each term, and each of its relevant dimensions we collected the actual numerical values associated with this pair. This gives us expressive distributional information about range, mean, median and other statistics. Moreover, since our resource is collected using only a few rules for detecting quantities and converting units, it can be extended to other languages easily.

We evaluate DoQ on several existing datasets and show that it compares favorably with existing methods that require more resources and have less coverage. In particular, we identify and correct problems with some of the existing datasets

resulting in new, cleaner, evaluation datasets.

Overall, we make the following contributions:

1. A new method for collecting expressive quantitative information about objects.

2. A large resource of distributions over quantitative attributes of nouns, adjectives, and verbs.

3. Strong results on existing datasets for noun and adjective comparison, refining and improving an existing dataset, and creating a new dataset for evaluating noun comparisons.

2 Related Work

There has been a lot of work trying to use Hearst-style patterns (Hearst, 1992) to extract relations between objects in large corpora (Tandon et al., 2014; Shivade et al., 2016). For example, from the sentence: *"Melons are bigger than apples"* they extract the relation: '*Melons*' > '*apples*'. These methods suffer from reporting bias and low coverage, since the precise patterns need to be found to make these inferences. Our method, which relies on co-occurring objects, is robust to this issue. Pattern-based methods were also used in the context of OpenIE, e.g., to extract event duration information (Gusev et al., 2011; Kozareva and Hovy, 2011), but were found to be highly brittle due to the dependence on finding specific pre-defined patterns.

There is a line of work (Forbes and Choi, 2017; Yang et al., 2018) to determine the quantitative relation between two nouns on a specific scale. For adjectives (De Melo and Bansal, 2013; Kim and de Marneffe, 2013; Shivade et al., 2015; Cocos et al., 2018), comparisons were made only for relative intensities, i.e. '*freezing*' < '*cold*'. In contrast, we infer magnitudes as well, which make us robust to comparisons between different polarities of the same cluster (e.g. '*hot*' vs. '*cold*').

Spithourakis and Riedel (2018) propose several methods to represent numbers in language models (LMs) instead of using an out-of-vocabulary token, giving the LM more expressive ability to produce numbers. Spithourakis et al. (2016) showed that conditioning on numerical values in the LM can improve the consistency of the modeling for clinical reports. When using it along with a scorer for Semantic Error Correction (Dahlmeier and Ng, 2011), it makes more grounded suggestions, with realistic estimates of different measurements.

---

[2]The violin plots along the paper describe the probability density of the collected distribution at different values.
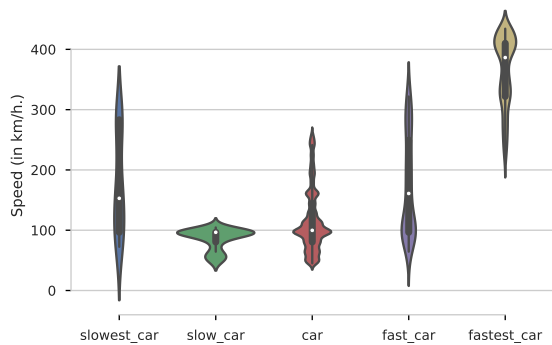
Figure 2: Car modifiers: Speed of cars, sidelong by different modifiers, which shift the cars speed distribution. Interesting to point out the high distribution of "slowest car" phrase (See the bias discussion in Sec. 6).

Our work overlaps with a number of approaches to ground textual objects by: achieving a common-sense understanding of numeric expressions (Chaganty and Liang, 2016), grounding adjectives into RGB colors (Winn and Muresan, 2018), grounding events duration (Pan et al., 2006; Gusev et al., 2011) and measurements' intensity within a given context (Narisawa et al., 2013).

Finally, our resource collection is in the line of work that uses counting across very large amounts of data (such as n-grams from books) to produce big resources (Lin et al., 2012; Goldberg and Orwant, 2013), which have had a significant impact on NLP Research.

3 Distribution over Quantities: Method

We propose a process for automatically extracting co-occurrences of objects and measurements from a large text corpus. Examples of the resulting output are the mass distributions of animals in Figure 1a, typical meal hours in Figure 1b and the car modifiers in Figure 2.

We first use a rule-based method for detecting and normalizing measurement mentions (Sec. 3.1). We then aggregate the detected measurements and objects that occurred in the nearby context (Sec. 3.2) and describe some simple heuristics for improving the resource accuracy (Sec. 3.3). Finally, Sec. 3.4 describes the resource produced in this process.

We note that the resource was built with the aim of keeping it as simple as possible, to test how accurate a simple approach can be. We believe it reflects the potential of transferring the process to other languages, where NLP resources are more sparse.

3.1 Measurement Identification and Normalization

Measurement identification uses a simple context-free grammar along with a mapping from units to dimensions. Thus, we know that *'inch'* is a unit in the LENGTH domain which is equal to 0.02524 meters, and that "acre foot" is a unit of VOLUME equal to 1233.48 standard units (here, cubic meters). Similar tables express SPEED in meters per second and TEMPERATURE in degree Kelvin.

If the unit is not expressed explicitly or recognized by the parser (for example, in the sentence "*New York was a scorching 110*"), we do not extract anything. There are occasional misparses caused by typographic shortcuts, such as "*17 C*" where Centigrade is meant, but is parsed as Coulombs. These show up as loss in coverage for us, since we deal with a limited set of dimensions in which charge is not included.

3.2 Object Collection

Object Extraction The main objects used in this work are 1-token words that are either nouns, adjectives or verbs. We also consider more complex phrases of these types (e.g. noun phrases). The complex phrases are collected enforcing minimum phrase spans. This way, for example, we collected the phrase "*race car*" and are able to compare its distribution to that of "*electric car*".

Object Head Along with each collected object, we also retrieve its syntactic head. For example, in the sentence: "*The fast car was driving 50 miles per hour*", collecting the adjective '*fast*' will also capture '*car*' as its head. With this information we are able to compare a "*fast car*" to a '*car*'. We note that this process is not possible for all languages and may result in less accurate extraction depending on the parser accuracy. Nonetheless, this phase is optional as it only adds the ability to compare more complex phrases and modifiers. A lot of information can still be collected without it.

Aggregation After identifying measurements in the sentence, we collect the objects that co-occur with these measurements within a certain context window. Using a bigger context size, we get broader coverage but also fewer occurrences. When reducing the context size, we get a sparser resource, but better attribution accuracy. More sophisticated collection methods are possible (e.g.

measuring parse-tree distances), but are left for future work.

Running the Entire Process We created the DOQ resource using the Flume framework (Chambers et al., 2010), to quickly processes billions of English webpages in parallel. First, we identified and normalized measurements (Sec. 3.1). Then, these sentences were parsed for POS tags and dependency trees (Andor et al., 2016) and the relevant objects gathered by identifying co-occurences (within sentence or distance threshold). The following step aggregated all of the objects with the same object-head-measurement tuple, creating a distribution of numbers (Sec. 3.2).

3.3 De-noising

The output of the described resource collection process is, as expected, quite noisy. It assumes a very simplified model of language, where co-occurring objects and numerical measurement are assumed equivalent to attribution, ignoring negations and reporting bias (Gordon and Van Durme, 2013). To address this, we employ de-noising filters focused on increasing precision. We get a cleaner resource at the expense of coverage, which is still valuable due to the high volume of data used.

Distance Based Co-Occurrences When aggregating co-occurrences, we also record the token distance between the measurement and the object. This can be a good indication of the degree of relatedness of a word to its surroundings. We used two context distances in our experiments: (1) co-occurrence within the same sentence, (2) co-occurrence within a token distance k.[3] In our experiments, we explore the effectiveness of the resource with different distance thresholds.

Negation Negations can affect the precision of the resource and contribute a lot to the distribution tails, as in: "*The dimension of the car is not 50cm.*" We decided to simply discard all measurements that appear in the same sentence with a negation word.[4]

---

[3]In practice, we use $k = 3, 10$.

[4]Specifically, we used the following negation words: 'not', 'no', 'without', 'neither' and 'nor'.

Filter/Type	Nouns	Adjectives	Verbs
none	117,953,900	2,513,033	2,121,448
5	16,188,215	598,563	603,799
100	1,497,753	130,534	160,060
1000	266,655	40,518	51,625

Table 1: **Size:** Number of tuples of Nouns, Adjectives, and Verbs, coupled with dimension, in our resource, as a function of the number of occurrences in the web (more than 5, 100, 1000 occurrences).

3.4 Distribution over Quantities Statistics

The final resource contains 117,953,900 unique noun tuples, 2,513,033 unique adjective tuples and 2,121,448 unique verb tuples. The total number of triples in English are 122,588,381. Table 1 provides some more statistics.

4 Evaluation Data

In this section we describe the datasets we use for evaluation. For the dataset introduced in (Forbes and Choi, 2017), we highlight a few problems we identified in it and how we corrected them, resulting in a new, cleaned up version of the dataset (Sec. 4.1). Moreover, since DOQ is more fine-grained than previous approaches supported, we also describe a new dataset for noun comparisons that was annotated by human annotators. We then describe the evaluation used for comparing adjectives (Sec. 4.2), and finally, an intrinsic evaluation done directly on the resource quality (Sec. 4.3).

4.1 Commonsense Property Comparison

Forbes and Choi (2017) created a dataset consisting of 3,656 object pairs labeled by crowd workers. The annotators were asked to label the typical relation between two objects along five dimensions: SIZE, WEIGHT, STRENGTH, RIGIDITY and SPEED: whether the first object was typically greater than, lesser than, or equal to the second along each dimension. 38-59% of the annotations (depending on the dimension) yielded perfect agreement among all annotators; 90-95% of them had an identifiable majority label, and they chose to keep all of these. We refer to this dataset as ORIG F&C.

Ill-Defined Comparisons In preliminary experiments on ORIG F&C we observed low results relative to the 76% achieved by the current state-of-the-art (Yang et al., 2018). A close inspection of a sample of 100 pairs (20 from each dimen-

		Train	Dev	Test	All
All Labels	ORIG F&C	587	5,418	6,007	12,012
	NO-LEAK F&C	712	3,000	4,497	8,209
Subset Labels	ORIG F&C	361	3,311	3,650	7,322
	CLEAN F&C	173	1,268	1,523	2,964

Table 2: F&C dataset size. All Labels represent the original dataset with all the labels. Subset Labels are the subset labels which are inferable by the resource.

sion) revealed that only 57% of the examples were in agreement with the annotations in Forbes and Choi (2017) and 47% were not comparable. The most common reasons for disagreements were: (1) Broad objects: e.g. (*father*, *clothes*, *big*); (2) Abstract objects: e.g. (*seal*, *place*, *big*); (3) Ill-defined dimension: e.g. (*friend*, *bed*, *strong*).

Training Leakage and New Split Another problem we identified in ORIG F&C, which results in a biased evaluation, is *leakage* from the training set to the dev/test sets. We identified two types of leakage.

The first is a **Transitivity Leakage**: when the training set contains the tuples: (o_1, o_2, d) and (o_2, o_3, d), and the dev/test set contains the tuple (o_1, o_3, d). For example, the training set contains ('*person*', '*fox*', '*weight*', '*bigger*') and ('*fox*', '*goose*', '*weight*', '*bigger*'), and the dev set contains ('*person*', '*goose*', '*weight*', '*bigger*'). While transitivity is an inherent property of this data, success on the transitive closure of training examples does not reflect the ability of the algorithm to infer the correct relation between two unseen objects, and these examples should be removed from the evaluation data. We found 4.3% of the dev and 3.5% of the test data had transitive leakage.

The second type of leakage we identified is **Object Leakage**, where a certain object in the dev/test set already appeared in the training set. This happens in 94.8% and 95.7% of the examples in the dev/test sets, respectively. This means that success on these objects might not reflect the generalization abilities of the algorithm, but rather a memorization of the training data.

To address these concerns, we reorganized the train/dev/test sets, forming new splits, which we refer to as NO-LEAK F&C. The new split sizes can be found in Table 2. We re-ran the current models on NO-LEAK F&C and, as expected, we observe a drop of 5-6% in accuracy: from the original 76% accuracy on the dev/test sets, to 70% and 71% accuracy, respectively.

F&C Re-annotation Due to the ill-defined comparison we identified in the dataset, we re-annotated it using crowd-source workers, who were trained with specific instructions to attend to the validity of the comparison. We used 3 annotators per example and the majority vote was used as the final answer. Examples with no agreement, i.e., where each annotator chose a different option, were discarded. The inter annotator agreement yielded Fleiss kappa of $k = 89.8$. Out of 7322 tuples in the original dataset, 59.5% were discarded either because the objects were simply not comparable, or due to lack of agreement between the annotators. After removing the non-comparable examples, the kappa agreement was $k = 97.2$. We refer to this new dataset as CLEAN F&C. We also tested the agreement between the new labels, and the corresponding labels in the original dataset, achieving near-perfect agreement of $k = 90.2$, establishing the quality of the new annotations.

New, More Conservative Dataset Due to the problems we identified in ORIG F&C and the fact that it became significantly smaller after filtering out ill-defined comparisons, we created a new dataset. We provided human annotators with more precise definitions and restricted comparisons to specific domains using only a subset of the dimensions – MASS, SPEED, CURRENCY and LENGTH. We further controlled the generation of comparable objects by using Category Builder (CB) (Mahabal et al., 2018), a method which can be used to expand a set of seed words into others in the same category. For each domain and dimension we fed an initial seed into CB, and used the top results as comparable pairs. Table 3 in the Appendix presents statistics and examples from each category from the new dataset. Note that the new dataset is only used as a test set and thus *leakage* is not applicable. Moreover, due to the controlled data generation process, we avoided some of the comparison issues we observed in ORIG F&C.

We used crowdsourcing to annotate the pairs, and obtained a substantial inter-annotator agreement of $k = 77.1$. Each example was annotated by three annotators and we used majority vote to determine the final labels. The final dataset discards examples with no agreement and Non-Comparable label, resulting in 4,773 examples.

Our method for determining a relation between two objects is unsupervised and does not require a training set. However, in order to compare with

		deMelo	Wilk-intense	Wilk-all
All Labels	#Objects	300	60	60
	#Pairs	682	60	133
Subset Labels	#Objects	58	21	22
	#Pairs	121	21	47

Table 3: Adjective dataset size from deMelo and Wilkinson. The All Labels row represent the original dataset, with all adjective clusters. The Subset Labels row describes the filtered partition on these datasets, with only measurable adjectives.

other methods, we split it into train/dev/test of sizes 326/1101/1101 respectively, while keeping the split free of leakage. (Note that the number of training examples is comparable to Yang et al. (2018). This split contains only comparisons from the SPEED, LENGTH and WEIGHT dimensions for fair comparison to previous work.)

The Relative Size Dataset Bagherinezhad et al. (2016) created a dataset of 486 object pairs between 41 physical objects. The dataset (which we refer to as RELATIVE) focuses solely on the physical size dimension. Nevertheless, we use it as another evaluation of DoQ.

4.2 Scalar Adjectives

Several test sets have been created to evaluate the intensity of adjectives. The dataset created by De Melo and Bansal (2013) uses adjective clusters based on the 'dumbbell' structure of adjectives in WordNet e.g. "$cold < frigid < frozen$". Wilkinson and Oates (2016) created another testset, by defining a total order between adjectives in the same cluster, spanning the entire scale range. For example, in the SIZE domain, the full cluster is: "$minuscule < tiny < small < big < large < huge < enormous < gigantic$". A total of 60 adjectives were collected across 12 clusters.

Since our method only handles measurable objects, we manually removed all of the non-measurable clusters (e.g., "$known < famous < legendary$" was removed) and evaluated on the rest. In this process we found that the new dataset by Cocos et al. (2018) contains only a small number of measurable clusters and some overlap with the other testsets, therefore we exclude this test set from our evaluation. The number of pair comparisons and unique objects are detailed in Table 3, both the original datasets and the subset we used in this work.

4.3 Intrinsic Evaluation

Lastly, since our resource is more expressive than what was done in this area before, we also conducted a novel intrinsic evaluation. We ran the evaluation as follows: Given an object and a dimension, we extracted the median of the distribution, expanded it into a range and then asked human raters whether this range overlaps with the range of the target object-dimension pair. For example, when evaluating the speed of a car, its median is 99.7 km/h. We then convert it to a range of 10-100 km/h by relaxing it to its nearest order of magnitude numbers, and asked annotators if this range corresponds to the typical speed of a car.

We collected a total of 1,271 examples from the same pool of comparisons used for our new dataset. Each example was evaluated by 3 annotators and labeled with the majority vote, discarding examples with no agreement.

5 Experimental Results

The object comparison task described in Sec. 4.1 is formulated as: Given objects o_1 and o_2 and dimension d, predict the relation $y \in \{<, =, >\}$.[5] To solve this task, we look up the set of measurements associated with each object-dimension pair in the object dictionary. For this evaluation, we aggregate all objects while ignoring their heads (as described in Sec. 3.4). We compare the two distributions obtained by their medians. If the object-dimension pair does not appear in DoQ, we assign it a 0 value.

Algorithm 1 Adjectives Comparison Inference

> **Input:** adjectives x, z, dimension d and object distributions H
> **Output:** comparison label
> **Procedure:**
> Initialize $\hat{y}$, the predictions per head
> $intersect \leftarrow$ findHeadIntersection(H, x, z, d)
>   ▷ the intersecting heads of x and z
> **for** $a_i, b_i \in intersect$ **do**
> $\hat{y}_i \leftarrow$ compare(a_i, b_i, d)
> **end for**
> Return majority($\hat{y}$)

Adjective comparisons require a different treatment. Earlier work was done mainly on intensities, and so comparisons are only across half the

Model/Dataset	F&C Clean		New Data	
	Dev	Test	Dev	Test
Majority	0.54	0.57	0.51	0.50
Yang et al. (PCE LSTM)	**0.86**	**0.87**	0.60	0.57
DoQ	0.78	0.77	**0.62**	**0.62**
DoQ + 10-distance	0.78	0.77	**0.62**	**0.62**
DoQ + 3-distance	0.81	0.80	**0.62**	0.61

Table 4: Results on the noun comparison datasets.

Model	Accuracy
Chance	0.5
Bagherinezhad et al.	0.835
Yang et al. (Transfer)	0.858
DoQ	0.872
DoQ + 10-distance	**0.877**
DoQ + 3-distance	0.858

Table 5: Results on the RELATIVE dataset. Yang et al. (2018) result was achieved by running their model on their training set, and using it as a transfer method on RELATIVE. Finally, we present our own predictions, with different thresholds, which surpass previous work.

scale (i.e., not on *hot* vs. *cold*, but on degrees of hot and degrees of cold separately). The dimension of the comparisons is not given explicitly; although it is possible to infer the most relevant dimension from DoQ it is not trivial and we leave this for future work. Instead, we manually label the dimension of each cluster used. For example, to the "*cold < frigid*" comparison, we assign the TEMPERATURE dimension.

The inference method for adjectives is also more subtle. As adjectives can describe a wide range of objects, their variance is higher than that of nouns. Therefore, our inference method makes use of an aggregation of individual objects: For each pair of adjectives we wish to compare, we query DoQ for every noun that both adjectives are seen to modify. For each such noun, we compare the distributions along the specified dimension, and choose the majority comparison over all such nouns as the prediction for the adjective pair. This process is outlined in Algorithm 1.

For the experiments using DoQ we used all three distance-based versions (sentence distance, 10 and 3 words distance). We found that the sentence-based has the higher coverage, but lower precision, whereas the lower distance-based has less coverage but higher precision.

5.1 Comparative Evaluation

Noun Comparison The left column of Table 4 presents results for the cleaned version of the Forbes and Choi (2017) dataset. The current state-of-the-art model achieves a total accuracy on the test set of 87%, while our best method achieves 80%. First, we note that the accuracies are significantly higher than those on the original dataset, for all methods. Second, we still observe lower accuracy for our method compared to Yang et al. (2018). We can attribute this gap to two reasons. First, they fine-tune their model on a training set, and although the training set size isn't large, it is necessary for achieving these results. Secondly, they are able to exploit similarities and

capture synonym information through pre-trained word embeddings, which our method cannot. For example, the development set contains the comparison: ('*lady*', '*step*', '*size*') and ('*wife*', '*ship*', '*size*'). While these comparison are valid, they are less intuitive, and can be solved by embedding methods due to their proximity in the embedding space to similar words, such as '*person*'. And indeed, when using the word '*person*' in our method instead of '*lady*' and '*wife*', our method makes the correct prediction.[6]

Results on the new objects comparison dataset we created are shown in the rightmost column of Table 4. Although our method doesn't benefit from a split into train/dev/test, we split it nevertheless to compare to previous work. This split is created such that there is no leakage from the train to the dev/test sets. We get better results than previous methods on this dataset: 63% and 61% accuracy on the dev/test sets compared to 60% and 57%. These relatively low results on this new dataset indicate that it is more challenging.

The last evaluation of noun comparatives is on RELATIVE (Bagherinezhad et al., 2016), presented in Table 5. We report the results of the original work, where the best score used a combination of visual and textual signals, achieving 83.5% accuracy. We also tested the method by Yang et al. (2018) on this dataset. Since the dataset is small, we did not split it, and instead used the training set from Forbes and Choi (2017). This can be viewed as a transfer learning evaluation. The accuracy achieved by this method is 85.8%, surpassing the previous method by more than 2 points. We evaluated our method on this dataset, achieving a new state-of-the-art result of 87.7% accuracy with $k = 10$ as a filter method.

---

[6]While it is possible to augment DoQ and access it via word embeddings, we chose not to do it in our experiments to better estimate the quality of the resource itself.

Model	deMelo	Wilk-intense	Wilk-all
Global Ranking	**0.642**	0.818	-
Cocos et al.	0.620	**0.841**	-
DoQ	0.617	0.700	0.870
DoQ + 10-distance	0.608	0.750	**0.891**
DoQ + 3-distance	0.567	0.500	0.761

Table 6: Results on scalar adjectives datasets.

5.2 Adjective Comparison

For the scalar adjective datasets, we present an evaluation on the deMelo dataset (De Melo and Bansal, 2013), and the Wilkinson dataset (Wilkinson and Oates, 2016). Previous work is limited by the patterns used for extraction to comparing adjectives from the same half-cluster. As Wilkinson data contains the full scalar range, we also present results on the full range. We compare to De Melo and Bansal (2013), using the re-implementation of Cocos et al. (2018) for global ranking. We also use the new method of Cocos et al. (2018) to evaluate. This work is not entirely comparable, as the coverage of the data depends on the exact method used i.e. the combination of patterns, lexicon-based evidence and paraphrasing. Therefore, for each dataset, we used the method that obtained the highest coverage. For pairs with no coverage, we chose random labels with uniform distribution. The method of De Melo and Bansal (2013) outperforms the rest for their dataset, while the method of Cocos et al. (2018) performs best on the Wilkinson data. Our method does get comparable results on De Melo and Bansal (2013), while on Wilkinson (Wilkinson and Oates, 2016) we lag behind by 9.1 points. Finally, we do achieve good results when evaluating on the full range scale of Wilkinson - 89.1% accuracy. All of the errors by our method on this dataset evaluation are indeed on the intensity level, and not between the extremes. We therefore conclude that our method is good at differentiating between the adjectives on the two tips of the scale.

In the Adjective comparison, we also observe the highest variance as a function of the context window size k. While DoQ with $k = 10$ achieves the best results on two of the three datasets, when $k = 3$ the results suffer from a big drop in performance. We hypothesize that this performance gap is due to the higher variance in the use of adjectives vs. nouns, and our inference method that is based not on the adjective itself, but on all its modifying objects.

Method/Data	Mass	Length	Speed	Currency	All
Indian Annotators	0.61	0.79	0.77	0.58	0.69
US Annotators	-	-	-	0.76	-

Table 7: Intrinsic Evaluation. Accuracy of the number of objects which our proposed median fall into range of the object, given the dimension.

5.3 Intrinsic Evaluation

We perform the following intrinsic evaluation to assess the distribution quality of the resource. The results of the intrinsic evaluation on a sample of DoQ are shown in Table 7. The total agreement is 69%, while the specific agreements for MASS, LENGTH, SPEED and CURRENCY are 61%, 79%, 77% and 58% respectively. Originally, these annotations were performed by annotators from India and, while inspecting the annotation, we found cultural differences in the perceived prices of items. We re-annotated the samples in the currency category with annotators from the U.S. and found a much higher agreement score: 76%. For example, Indian annotators reported that a suit could not cost between 1K-10K$, while U.S-based annotators all reported it was possible.

6 Conclusion and Discussion

This paper develops an unsupervised method for collecting quantitative information from a large web corpus, and uses it to create DoQ, a very large resource consisting of distributions over physical quantities associated with nouns, adjectives, and events. We have evaluated DoQ on multiple existing and new datasets and showed that it compares favorably with other methods that require more resources and lack coverage relative to DoQ. Below, we discuss a few interesting issues brought up by the data collection process that should be addressed in future work.

Reporting Bias and Exaggeration Although reporting bias (Gordon and Van Durme, 2013) would seem to be a problem for a corpus-driven approach, in practice, DoQ is quite resilient to it due to the usage of very big web corpora and the collection method. As we do not rely on explicit comparisons between objects, but only on co-occurrences with numeric measurements, we can automatically infer relationships post-facto. One form of reporting bias we observe is that people are more likely to discuss objects when they are exceptional, or they exaggerate measurements

(a) Measured average temperature

(b) Induced average temperature

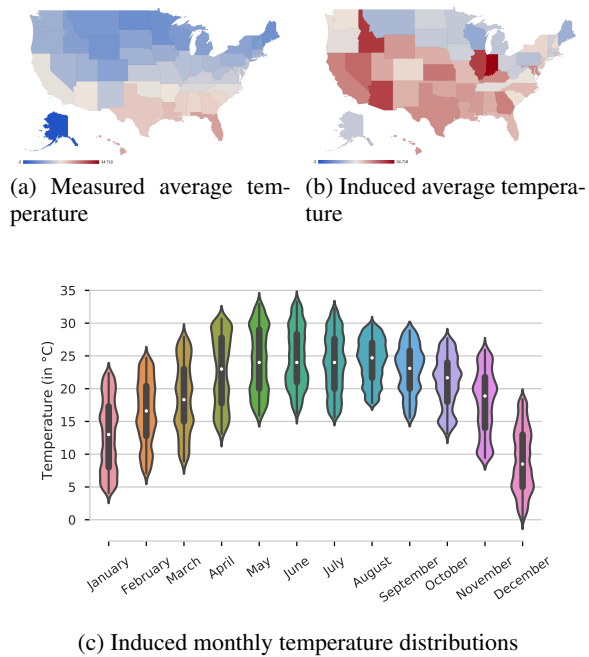
(c) Induced monthly temperature distributions

Figure 3: Reporting bias: The observed average temperatures of U.S. states tend to extremes (Real average temperature (3a) vs. the induced one (3b)). Another sort of bias is exemplified in Figure 3c, by a bias towards the northern hemisphere.

Figure 4: Reporting bias can also be seen in this example where alfalfa's weight is induced in tons whereas in reality alfafa's weight is measure in grams.

for rhetorical effect, leading to long tails for some distributions (see "*slowest car*" in Figure 2 and extreme temperatures in Figure 3). It is interesting to note that in the case of temperatures, both in the U.S states case (Figure 3) and the world case (Figure 2 in the Appendix), the exaggeration is towards hot temperatures, and not cold ones.

A somewhat different bias is shown in Figure 3c; although the temperatures are an adequate representation of the cyclic year, it is highly biased towards the northern hemisphere, a result of the English web source data.

A more subtle form of bias is due to attribution. For example, when comparing the size of alfalfa with the size of watermelons as shown in Figure 4, we see that alfalfa is mostly talked about in quantities in which it is harvested (order of tons) rather than individual units (grams). This kind of bias cannot be identified as easily as the attribution bias discussed in Sec. 3.3.

Polysemy We have not systematically explored how our resource performs on polysemous words and their senses, although our overall results indicate that in most cases the relatively biased distribution of polysemous senses render this a non-
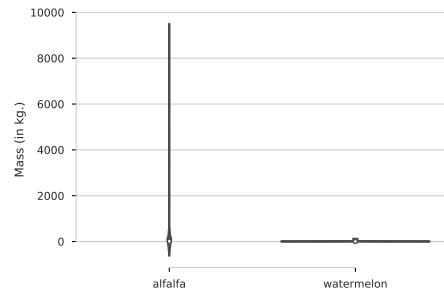
problem. We have also observed that in some cases the data itself can help disambiguate between different word senses. For example, 'bat' can refer to the animal, a baseball bat or a cricket bat. Figure 1 in the Appendix shows the induced distributions of length for these three senses of bat. While the distributions for "Baseball bat" (which measure about 1m) and "Cricket bat" (which may be no more than 956mm) are correct, the distribution for 'bat' is probably a consolidation of these, the animal bat that can measure from 15cm to almost 1.7m in length, and some attribution noise (e.g. the distance the bat flew).

In conclusion, we developed and studied an unsupervised method for collecting quantitative information from large amounts of web data, and used it to create Distribution over Quantities (DoQ), a new, very large resource consisting of distributions over physical quantities associated with nouns, adjectives, and verbs. The histogram version of the resource, as well as the new created dataset and evaluation code are available at https://github.com/google-research-datasets/distribution-over-quantities.

Acknowledgments

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2442–2452.

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are elephants bigger than butterflies? Reasoning about sizes of objects. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Arun Chaganty and Percy Liang. 2016. How much is 131 million dollars? Putting numbers in perspective with compositional descriptions. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 578–587.

Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R. Henry, Robert Bradshaw, and Nathan Weizenbaum. 2010. Flumejava: Easy, efficient data-parallel pipelines. In *Proceedings of the 31st ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '10, pages 363–375, New York, NY, USA. ACM.

Anne Cocos, Veronica Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning scalar adjective intensity from paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Correcting semantic collocation errors with l1-induced paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 107–117. Association for Computational Linguistics.

Gerard De Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

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 (Volume 1: Long Papers)*, pages 266–276.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. *Atlanta, Georgia, USA*, page 241.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30. ACM.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the ninth international conference on computational semantics*, pages 145–154. Association for Computational Linguistics.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630.

Zornitsa Kozareva and Eduard Hovy. 2011. Learning temporal information for states and events. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 424–429. IEEE.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. Association for Computational Linguistics.

Abhijit A. Mahabal, Dan Roth, and Sid Mittal. 2018. Robust handling of polysemy via sparse representations. In **SEM*.

Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. 2013. Is a 204 cm man tall or small? Acquisition of numerical common sense from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 382–391.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the 2018 North American Chapter of the Association for Computational Linguistics (NAACL2018)*, pages 841–851.

Feng Pan, Rutu Mulkar, and Jerry R Hobbs. 2006. An annotated corpus of typical durations of events. In *LREC*, pages 77–82. Citeseer.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2016. Identification, characterization, and grounding of gradable terms in clinical text. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 17–26.

Georgios Spithourakis, Isabelle Augenstein, and Sebastian Riedel. 2016. Numerically grounded language models for semantic error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 987–992.

Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2014. Acquiring comparative commonsense knowledge from the web. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Bryan Wilkinson and Tim Oates. 2016. A gold standard for scalar adjectives. In *LREC*.

Olivia Winn and Smaranda Muresan. 2018. Lighter can still be dark: Modeling comparative color descriptions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 790–795.

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 (Volume 2: Short Papers)*, volume 2, pages 644–649.

A   Supplementary Material

In Figure 5 we present length distributions for the word 'bat' and two of its senses, showing a clear distinction between, what we believe to be, baseball bats and cricket bats. In Figure 6, we show the DoQ-induced and real average temperatures of countries. Figure 7 shows some additional measurement distributions from DoQ for various objects types and dimensions, as well as some examples of systematic errors caused by the collection/aggregation process.

In Table 8 we present specific example sentences used to extract correct measurements for objects along the ten most informative dimensions, while Table 9 shows examples where incorrect measurement attributions were extracted. Finally, Table 10 describes the different measurements and domains used in the new dataset we built, along with inter-annotator agreements, a representative example and the labeled relation.
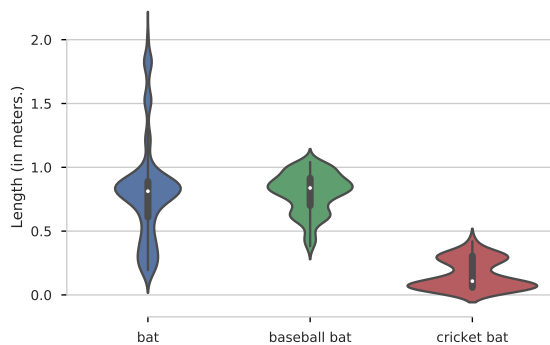
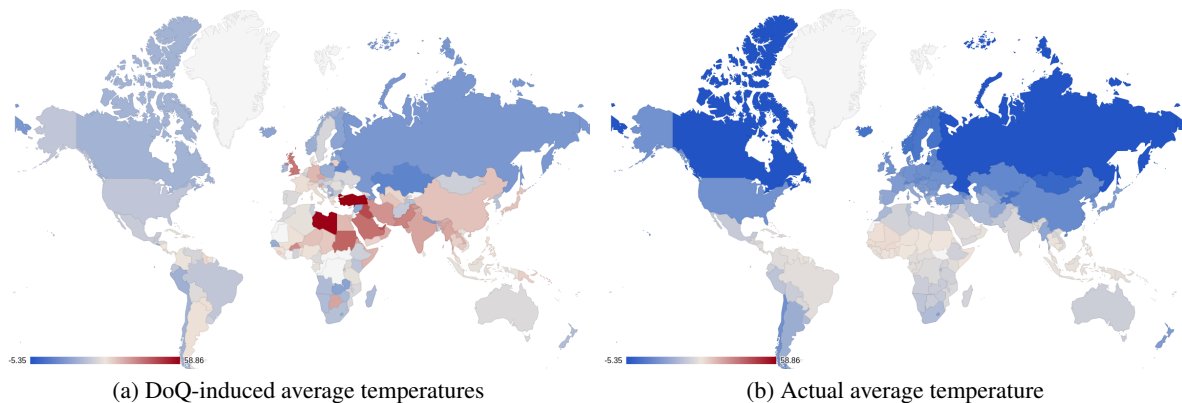Figure 5: Induced distributions for the various polysemous meanings of 'bat'.

(a) DoQ-induced average temperatures   (b) Actual average temperature

Figure 6: Average world temperatures, predicted by DoQ (Figure 6a), and the ground truth (Figure 6b, taken from `https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly_temperature`). Interestingly, predicted temperatures are consistently higher than true ones, with the two hottest countries being Turkey (58.8°C) and Libya (57.7°C).

| Dimension	Example
TIME	On 24 September drivers struck between 7.30 am and 8.30 am, the middle of the morning **rush hour**.
CURRENCY	McKay picked up $2,000 (the biggest winning **cheque** of her career) for winning the competition.
LENGTH	The railway line has 140 bridges and 25 km of **tunnels**
AREA	With over 36,000 acres, **Smoky Hill** is the Air National Guard's largest weapons range.
VOLUME	The traditional Bordeaux barrique is an **oak barrel** with a capacity of 225 litres.
MASS	The **men**'s 73 kg competition of the 2014 World Judo Championships was held on 27 August.
TEMPERATURE	Places as far north as **New York City** reached 70 F (21 C) on Christmas Eve.
DURATION	Formerly the **market** lasted for 14 days.
SPEED	Dora was the strongest **storm** of the year, peaking at 155 mph, just short of Category 5 status.
VOLTAGE	The motorcycle used a 24 volt electric starter **motor** from a Douglas A-4B fighter plane.

Table 8: Examples of sentences from the web corpus used to infer a correct measurement of the corresponding object for different dimensions.

| Measurement	Example
TIME	The largest **aftershock** measured 5.6 and struck nine minutes later at 1.51 pm.
CURRENCY	Tuncay's move was confirmed the following **day**, with Stoke paying 5 million for the Turkish player.
LENGTH	Everest from nine miles away in Rongbuk valley.
AREA	The **plant** is spread over 25 acres of land.
VOLUME	It was created for **cars** with engine size below 2 Litres.
MASS	In recent years, the reported average annual **peanut** production lies around 828,000 tons (95% for oil).
TEMPERATURE	The corresponding **phosphorine** and benzene undergo the analogous reaction at 100 C and 200 C, respectively.
DURATION	Its **surface** remains frozen for around 155 days per year.
SPEED	Both trains covered the 466 mile **route** at an average pace of 49 mph.
VOLTAGE	The value of the shunt ensures a potential difference of 0.5 volt across it at the maximum **generator** load.

Table 9: Examples of sentences from the web corpus from which an incorrect measurement was extracted, due to attribution/reporting bias or other reasons.
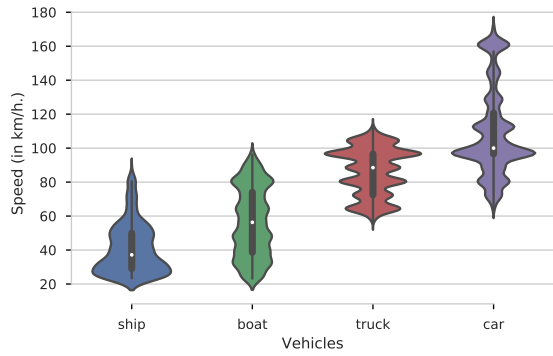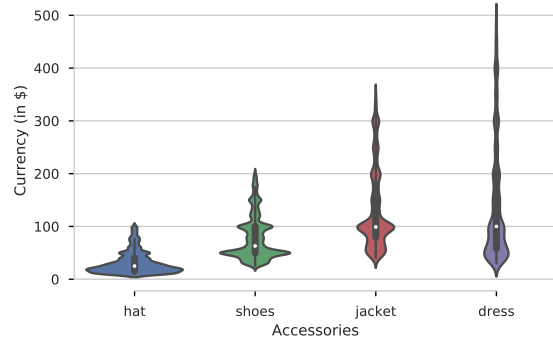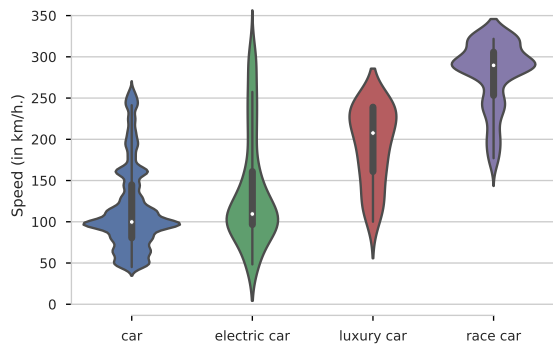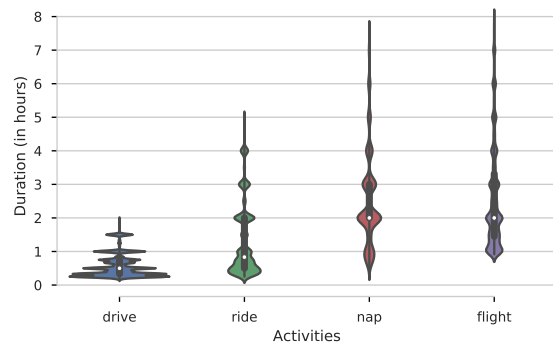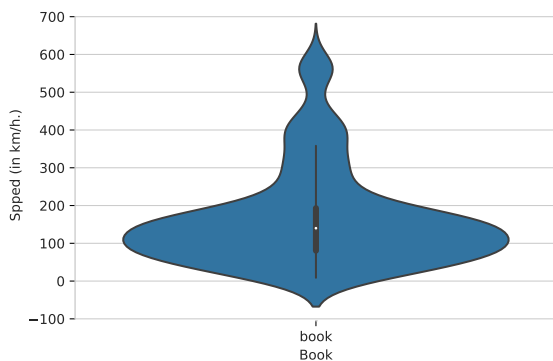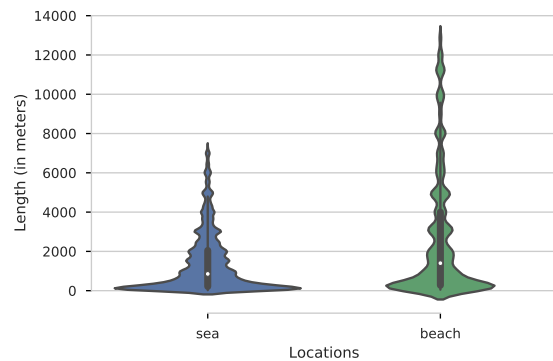
Figure 7: Comparisons of measurements extracted from DoQ for different categories of objects along their most salient dimensions. Note that (e) and (f) show spurious measurements for the speed of books (caused by attribution bias) and the lengths of seas and beaches (caused by confusion with sea levels).

Measurement	Domain	Count	Inter-Annotator Agreement	Example	Relation
	Vegtables & Fruits	500	76.9	(alfalfa, watermelon)	<
Mass	Animals	501	86.0	(skunk, lemur)	<
	Technology	500	80.2	(modem, microphone)	=
	Animals	500	83.3	(doll, dragon)	N.A.
Speed	Vehicles	500	78.8	(tractor, truck)	<
	Activities	500	70.3	(barreling, traversing)	>
	Dishes	500	74.3	(tiramisu, gnocchi)	>
Currency	Accessories	500	68.9	(bra, bikini)	<
	Technology	500	76.2	(hose, sensor)	>
	Places	500	76.8	(bungalow, arena)	N.A.
Length	House Object	500	72.6	(terrace, railing)	N.A.
	Organs	499	73.2	(arm, ovary)	>
All	All	6,000	77.1	-	-

Table 10: Statistics for the new dataset, with the number of examples in each domain, their inter-annotator agreement, a representative example and the majority label assigned by annotators.