Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards

Shilad Sen
Macalester College
St. Paul, Minnesota
ssen@macalester.edu

Macademia Team: *
Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko Samuel Naden, Jesse Russell Zixiao "Ken" Wang
Macalester College
St. Paul, Minnesota

Brent Hecht
University of Minnesota
Minneapolis, Minnesota
bhecht@cs.umn.edu

ABSTRACT

In just a few years, crowdsourcing markets like Mechanical Turk have become the dominant mechanism for building "gold standard" datasets in areas of computer science ranging from natural language processing to audio transcription. The assumption behind this sea change — an assumption that is central to the approaches taken in hundreds of research projects — is that crowdsourced markets can accurately replicate the judgments of the general population for knowledge-oriented tasks. Focusing on the important domain of *semantic relatedness* algorithms and leveraging Clark's theory of common ground as a framework, we demonstrate that this assumption can be highly problematic. Using 7,921 semantic relatedness judgements from 72 scholars and 39 crowdworkers, we show that crowdworkers on Mechanical Turk produce significantly different semantic relatedness gold standard judgements than people from other communities. We also show that algorithms that perform well against Mechanical Turk gold standard datasets do significantly worse when evaluated against other communities' gold standards. Our results call into question the broad use of Mechanical Turk for the development of gold standard datasets and demonstrate the importance of understanding these datasets from a human-centered point-of-view. More generally, our findings problematize the notion that a universal gold standard dataset exists for all knowledge tasks.

Author Keywords

semantic relatedness; gold standard datasets; cultural communities; Amazon Mechanical Turk; user studies; natural language processing

*megiesel@gmail.com rebeccagold0@gmail.com bhillmann-@outlook.com mnlesicko@gmail.com samnaden@gmail.com jesse.p.russell@gmail.com zwang@macalester.edu

CSCW '15, March 14 – 18 2015, Vancouver, BC, Canada.

ACM 978-1-4503-2922-4/15/03$15.00.
http://dx.doi.org/10.1145/2675133.2675285

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Less than a decade after their inception, crowdsourcing markets like Amazon's Mechanical Turk[1] have transformed research and practice in computer science. While computer-supported cooperative work researchers have largely focused on understanding crowdsourcing markets, analyzing and developing crowdsourcing mechanisms, and finding new opportunities to apply these mechanisms, other areas of computer science have also embraced crowdsourcing. In natural language processing and many areas of artificial intelligence, crowdsourcing markets — and in particular Mechanical Turk — have become the de facto method of obtaining the critical resource known as *human-annotated "ground truth" datasets*.

Human-annotated ground truth datasets (henceforth referred to by their more common and simpler name: "gold standards") support a large variety of knowledge-oriented tasks. In general, these datasets capture human beings' "right answers" for tasks when no obvious answer for a given problem exists or can be algorithmically created. Gold standard data support algorithms in a wide range of problem spaces ranging from sentiment analysis [40], to audio transcription [8], to machine translation [5].

In 2008, Snow et al. marked a shift to Amazon's Mechanical Turk (AMT) for collecting gold standards [38]. They showed that AMT workers (known as"turkers") closely replicated five well-known gold standard datasets in the domain of natural language processing, but faster and much more cheaply. Out of this paper, a widely-accepted precept has arisen: AMT workers produce gold standard datasets for knowledge-oriented tasks that are more or less the same as those produced by other groups of people (traditionally domain experts, peers, or local undergraduates). We refer to this as the "turkers for all" precept.

Much work in the social sciences, however, would suggest that the "turkers for all" precept will often not hold. For instance, in Clark's definition of common ground, a cultural

[1] https://www.mturk.com

community — whether it is demarcated by professional, religious, ethnic, or other lines — is defined by the (mutually known) shared knowledge it has about a given set of concepts and their relationships [9]. It should follow then that the people who belong to different cultural communities would provide different answers to a variety of knowledge-oriented tasks. For instance, in a text categorization task, we might expect that scientists would categorize a document about Albert Einstein as a biography of a famous scientist, but a group of peace activists may also categorize it as a biography of a public figure against nuclear weapons, a group of Jewish scholars may want to categorize the text as a biography of a famous Jewish person, and so on [23]. We refer to the hypothesis that different communities produce different gold standards as the "communities matter" precept.

This paper analyzes the tension between the Snow et al. "turkers for all" and Clark "communities matter" precepts.

We focus on two broad questions at the core of the tension between these two precepts:

RQ1: Do different cultural communities produce different gold standards?

RQ2: Do algorithms perform differently on gold standards from different cultural communities?

We study these questions within the field of natural language processing, in the domain of semantic relatedness (SR) algorithms. Generally speaking, SR algorithms provide a single numeric estimate of the number and strength of relationships between any two concepts *a* and *b* (typically between 0 and 1) [20]. For instance, if concept *a* is "pyrite" and concept *b* is "iron", we might expect a good SR algorithm to output a high value for $SR(a, b)$. Conversely, if concept *b* is replaced with "socks", we might expect the opposite. SR has a nearly 50 year tradition of evaluation against gold standards, starting with the RG65 dataset collected by Rubenstein and Goodenough [36].

In this paper, we describe an experiment with 111 participants that shows that the "communities matter' hypothesis is supported for both RQ1 and RQ2. Specifically, we show that AMT workers, communities of general academics and subject-area experts each have their own "gold standard" for a series of SR problem instances. Broadly, as the cultural community's depth of knowledge increases, they seem to perceive more connections between concepts (instead of more differences). In addition, we show that SR algorithms determined to be "start-of-the-art" in their ability to replicate a gold standard dataset are indeed only "state-of-the-art" *for a given cultural community*. More specifically SR algorithms have difficulty matching the SR judgements of subject experts (particularly psychologists), even when trained on data from that same group.

We also find support for the "turkers for all" precept in some contexts. Mechanical Turk serves as a tremendously valuable resource offering fast, inexpensive data. Thus, we identify areas where turkers are more likely to provide gold standard judgements that are consistent with other communities. Specifically, for knowledge-oriented tasks for which all or most relevant information is in the common ground of a wide variety of cultural communities, using Mechanical Turk to develop gold standards is likely more appropriate.

To summarize, our work offers three main contributions:

1. We show that AMT-derived gold standards for knowledge-oriented tasks are often not representative of other communities. We explicitly probe when these differences occur.
2. We show that algorithms that perform well against an AMT-derived gold standard do not necessarily perform well against gold standards produced by other populations.
3. Our work problematizes the notion of the gold standard in general, highlighting types of tasks for which gold standards may be problematic and types of tasks for which which they are likely to be appropriate.

These findings have immediate implications for a number of constituencies, for instance researchers who use AMT and other crowdsourcing platforms, SR researchers, and practitioners designing systems that rely on the automatic completion of knowledge tasks, particularly systems for domain experts such as doctors, musicians, scholars, etc. More broadly, by problematizing the gold standard dataset, this work has implications for a methodology employed by researchers in a variety of areas of computer science.

Below, we first cover work related to this research and highlight our methodological approach. Following that, we report the results of our experiment. We then discuss our findings in more detail. Finally, we close by detailing the limitations of our study and highlighting future work in this area.

RELATED WORK

Semantic relatedness

The automatic estimation of the relatedness between two concepts has been an active area of research in artificial intelligence (AI) and natural language processing (NLP) for decades [36, 35, 6, 39, 14].

Semantic relatedness (SR) algorithms — the family of algorithms that perform this estimation – underly an enormous variety of applications. These applications range from low-level tasks like word sense disambiguation (e.g. [28]) and coreference resolution (e.g. [32]) to high-level technologies like search [6, 34] and information visualization systems [4, 37, 3].

SR algorithms are typically trained and evaluated against datasets of "gold standard" relatedness judgements from human participants (e.g. [12, 25]). The literature treats these judgements as by and large universally correct - it does not consider how the population of gold standard contributors (typically university students or AMT crowdworkers) aligns with that of the target application.

The vast majority of work in the semantic relatedness domain is dedicated to developing algorithms that can replicate the human judgements present in a small number of benchmark gold standard datasets. These algorithms are both diverse and

numerous, with approaches grounded in areas ranging from network analysis (e.g. [26]) to information theory (e.g. [35]) to information retrieval (e.g. [14, 34]) to the semantic web [13]. SR algorithms rely on a source of world knowledge, which is most commonly derived from Wikipedia (e.g. [14]) or WordNet (e.g. [35]).

WordSim353 [12] is the most widely-used benchmark dataset of SR judgements. The community(ies) from which the annotators of *WordSim353* were selected is not explicitly disclosed. More recently, other SR benchmark datasets have begun to appear in the literature, including *TSA287* [34] and *MTurk771* [18], all of which are made up of judgements from Mechanical Turk. Understanding the effects of this recent trend towards using Mechanical Turk as a source of human relatedness judgements is one of the key motivations behind our consideration of Mechanical Turk in this research.

Some algorithmic SR research has included sub-studies related to our research. For instance, while developing a new semantic relatedness algorithm designed specifically for the bioinformatics domain, Pedersen et al.[30], tangentially noted that SR judgements from doctors and medical coders differed, remarking that "by all means, more experimentation is necessary" in this area. In another primarily algorithmic paper, Pirró and Seco report on the effect of language ability on a small set of concept pairs. They found very high levels of agreement between native and non-native English speakers at the same university after excluding non-native outlier judgements [31]. Snow et al. [38] compared turkers' assessments to traditional SR annotators, also finding very high levels of agreement. Our work is distinguished from that above in scope, depth and, ultimately, outcome. We collect a dataset an order of magnitude larger than these previous datasets. We leverage this data to robustly probe the relationship between community membership and SR ratings. We also reveal divergence in intra- and interrater consistency across communities and demonstrate the effect of domain-related concepts versus general concepts.

In addition to its importance to a large number of research projects and applications in NLP and AI, semantic relatedness has also played a role in the HCI domain. Liesaputra and Witten have leveraged SR algorithms to create electronic books that improve performance on reading tasks [24] and Grieser et al. [17] did the same assess the relatedness of exhibits for museum visitors. Visualization has been a particularly active area of interest, with semantic relatedness being used, for instance, to cluster conversation topics in a system that highlights salient moments in live conversations [4], develop geospatial information that facilitates the development of spatial thinking skills [37], among other applications (e.g. [3]).

Cultural communities and online differences

To situate the results of our gold standard dataset analyses and evaluations of semantic relatedness algorithms, this work adapts Clark's definition of cultural communities from his theory of language as joint action [9]. As part of this theory, Clark defines a cultural community as "a group of people with shared expertise that other people lack." These groups can be delineated by geography, profession, language, hobbies, age, whether or not they are a part of an online community (e.g. AMT) and so on, but what they all share are unique sets of shared knowledge and beliefs. While Clark's theory is typically (and widely) applied in understanding how interlocutors work together to build a shared understanding of a given domain [15, 16], in this work, we are interested in the *pre-existing* shared knowledge that exists in cultural communities. We will show that Clark's formulation can help explain which gold standard datasets are liable to cultural variation and which ones will stay roughly constant across cultural groups (and why).

Cultural differences have been shown to influence annotations on tasks other than SR estimation. Dong and Fu demonstrated that European Americans tag images differently from people with a Chinese background [10] and a similar result has been identified across gender lines [33]. Similarly, Dong et al. found that culture has an effect not only on image tags themselves, but also annotators' reactions to tag suggestions [11]. However, this work has not examined cultural differences on gold standard datasets and has not sought to understand the effect of cultural bias on algorithm performance as we do here. Along the same lines, researchers have looked at the demographics of workers on AMT [22, 29], but have not focused on the effect these demographics have on gold standard datasets.

Crowdsourcing gold standards

A related area of research outside of the SR domain compares the performance of crowdworkers to that of local human annotators. Snow et al. established that "turkers" are able to closely replicate the results of local annotators on many labeling tasks in NLP in addition to SR [38]. Similar results have been seen in studies in psychology [7], translation [5], graphical perception [21], and a number of other areas. Our work adds to the literature that explores the relationship between "the crowd" and other annotators. However, unlike most research in this area, we find that judgements can differ substantially between these two groups of annotators. Importantly, our work also provides insight into the types of tasks for which "the crowd" and others will differ, and those for which this is not the case.

SURVEY METHODOLOGY

To measure the effects of cultural community on gold standards, we collected SR gold standards from different cultural communities[2] using an online survey. Before detailing the survey design and methodology, we provide an overall picture of its experimental and experiential design.

We recruited subjects from two cultural communities related to the "turkers for all" and "communities matter" precepts: AMT crowd workers and scholars. The survey collected *SR assessments* from subjects; a single SR assessment is a relatedness rating between 0 and 4 (inclusive) by a subject for a concept pair (e.g. a rating of 4 by turker number 12413 for the pair "movie", "film"). We followed common practice of SR

[2]The SR datasets in this paper are available online at http://shilad.com/pluraSR200.html

gold standards, which collect assessments from five to twenty subjects for each concept pair and report the mean SR rating for each pair.

To probe the applicability of each precept, we collected assessments for two type of concept pairs: *general knowledge* concept pairs (i.e. "television" and "admission"), and *domain-specific* concept pairs in history, biology, and psychology. With respect to Clark, we hypothesized that general knowledge concepts and their interrelationships would be relatively likely to be in the common ground of a wide variety of cultural groups. We hypothesized that domain-specific concepts in history, biology, and psychology and their interrelationships were *less* likely to be in a broadly-held common ground. In other words, people from different cultural communities (especially professionally delineated ones) would have a different understanding of each concept in these pairs, as well as the relationships between them.

All subjects completed an identically structured online survey. To summarize the survey's experimental design, it collected assessments for the two different types of concept pairs (general and domain-specific). Assessments came from three different cultural communities: AMT workers, scholars, and scholar-experts who were experts for a particular domain-specific assessment (e.g. a psychologist assessing "cognition" and "language"). Membership in each of the three communities is *question-specific* for researchers; a single researcher may be a scholar-expert for some concept pairs (e.g. psychology) and a scholar for others (e.g. history).

After consenting to the study, subjects entered basic demographic information (gender, education level) and indicated whether they conduct scholarly research. Those who did (scholars) provided their primary, secondary, and tertiary fields of study. Next, subjects provided 69 SR assessments spanning 6 pages. Subjects rated each concept pair on a 5 point scale ranging from 0 (not related) to 4 (strongly related) (Figure 1). Subjects could indicate that they "did not know a term" instead of providing an SR rating. After completing the 69 assessments, subjects were asked if they would like to complete a second round of assessments.

Selection of concept pairs for each subject:
Each subject provided 69 assessments of concept pairs. All subjects provided SR judgements for (a) 10 general knowledge assessments, (b) 50 domain-specific assessments chosen from the fields of biology, history, and psychology, (c) 4 validation assessments, and (d) 5 duplicate assessments. Each type of assessment is described in more detail below. The survey randomized the order of all concept pairs and ensured that each page spanned a variety of estimated relatedness values.

General knowledge concepts: General knowledge terms were chosen from *WordSim353*. The concept pairs in *WordSim353* consist of common nouns (e.g. "television", "admission") and a few widely-known named entities (e.g. countries, famous political figures). We randomly sampled 50 concept pairs from the dataset. The sample was stratified to ensure that the concepts captured a diverse set of relatedness values.

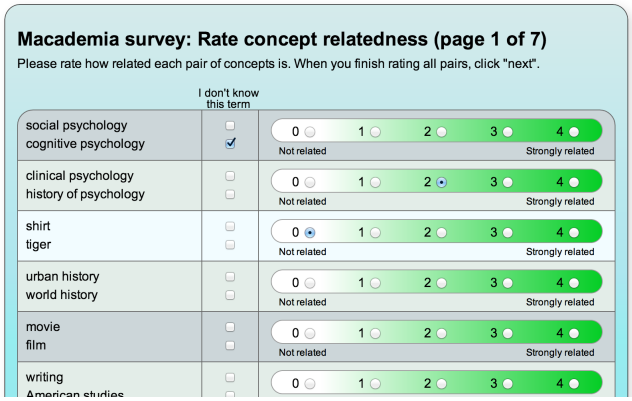

Figure 1. The rating page in the online survey. The subject has indicated that they do not understand the phrase "cognitive psychology.". The assessments on this subject's page include validation concepts ("movie", "film", 5 total), psychology concepts (25 total), and history concepts (25 total).

Domain-specific concepts: We chose 50 candidate concept pairs for each of the fields of biology, history, and psychology. Subjects who were biologists, historians, or psychologists provided 25 domain-expert assessments from their field and 25 assessments from a second target field (history, biology, or psychology). All other subjects provided 25 domain-specific questions from each of two target fields (two of history, biology, or psychology). Further rationale and and our methods for choosing these concept pairs is detailed in the following section.

Validation assessments: We added four validation assessments following the procedure of [18]. These assessments were intended to identify subjects who were not completing the survey in good faith. These concepts pairs contained the two most related ("female", "woman" and "film", "movie") and least related concept pairs ("shirt", "tiger" and "afternoon", "substance") from [18]. Subjects that did not accurately rate these pairs are excluded from our results. This validation test would exclude 94% of subjects who guess randomly.

Duplicate assessments: Subjects also completed five duplicate assessments to measure intra-rater agreement. All duplicate concept pairs were separated by at least one survey page.

Selecting domain-specific concept pairs:
The domain-specific concept pairs were harvested using the Macademia website[3] that visualizes research connections between scholars. Over 2000 users have created profiles on the website, which involves entering one's research interests as free-form text. We selected three diverse and popular fields (history, psychology, biology) as target knowledge areas. For each field, we selected the most common 16 interests, as we hypothesized that these were most likely to be within the common ground of members of each individual field[4] speci-

[3] http://macademia.macalester.edu
[4] We chose this cutoff because at least 16 interests were used three times in all three fields.

knowledge type	turker	scholar	scholar-expert
general	461	861	n/a
domain-specific	2,218	3,086	1,295

Table 1. Experimental manipulations in our study. Rows indicate knowledge type, and columns indicate community. Each cell contains the number of ratings in that condition. For example, 1,295 ratings have been provided for domain-specific concept pairs by scholar-experts (scholars with expertise in the field of the concepts in a pair).

fied by users in the field as candidate concepts. We randomly chose 50 concept pairs from the 16 candidate concepts.[5]

Subject recruitment and basic statistics:

As noted above, we recruited subjects via the Macademia website and Amazon's Mechanical Turk. We emailed invitations to a subset of Macademia users: all psychologists, biologists, and historians and a random sample of other users. We also hired "master" Mechanical Turk workers, who the AMT website describes as "an elite groups of Workers who have demonstrated accuracy on specific types of HITs on the Mechanical Turk marketplace[6]". We chose to recruit 45 turkers to match the number of scholar and scholar-expert subjects. All crowdworkers were paid at an average rate above the United States federal minimum wage.

In total, 145 subjects participated in the study. Twenty subjects did not complete the validation assessments accurately or did not finish the survey and are excluded from our results. To clearly delineate differences between turkers and scholars, we excluded users who were neither turkers nor scholars (10 subjects) and turkers who reported they were scholars (4 subjects).

Of the 111 final valid subjects, 39 were turkers, 42 were scholars in history, psychology, or biology, and the remaining 30 were scholars in some other field. 60% of valid subjects were female and 40% were male. All scholars indicated they held a graduate degree. Of the turkers, 13% had a graduate degree, 59% indicated their highest degree was a two- or four-year degree, and 28% indicated their highest degree was a G.E.D. or high school degree.

Table 1 shows the experimental manipulations we utilized in our study. The rows of the table indicate the knowledge type of the concept pairs (i.e. general or domain-specific). The columns of the table indicate the community of the subject. The numbers in each cell indicate the total number of assessments (not users) for each condition. For example, Table 1 reveals that we collected 861 total assessments of general knowledge concept pairs from all scholars (the *general/scholar* condition) and 1,295 assessments in the *domain-specific/scholar-expert* condition, which contains ratings from scholars who have domain expertise with regard to a given concept pair. Similarly, Table 1 shows that we received 461 assessments from turkers on general knowledge concept pairs (the *general/turker* condition) and 2,218 assessments in the *domain-specific/turker* condition.

Note that all subjects completed assessments that span multiple conditions: turkers assessed both specific and general concept pairs; biologists assessed some concept pairs in their field, some in either psychology or history, some from the general domain; and so on. The upper right column is not studied because the concepts in *WordSim353* are not associated with a specific domain of expertise.

Each concept pair was assessed by an average of 46 subjects. Subjects did not understand at least one concept in 1.6% of assessments. These assessments are not included in our results. Some analyses in later sections compare different groups' responses to each concept pair. To support these analyses, each group from Table 1 must have a reasonable number of responses to each concept pair. The biggest challenge to sample size occurs in the domain-specific/scholar-expert (scholar-expert responses for domain-specific concepts) because only a small subset of our population has domain expertise in a given concept. The mean number of scholar-expert responses per pair was 8.6 (median=9, min=6). Other conditions have more responses. For example, there are 14.8 (median=15, min=7) assessments per concept pair in the domain-specific/turker condition.

RQ1: EFFECTS ON GOLD STANDARDS

In this section, we analyze the judgements collected from the online survey to answer RQ1, which asks whether different cultural communities produce different gold standards. We study two characteristics of the human SR assessments that constitute a gold standard. First, we measure whether different communities coalesce around different numerical estimate values for a particular concept pair.

*RQ1a: Do different communities produce different semantic relatedness **estimates**?*

RQ1a studies whether a gold standard dataset must be matched to the audience of the system it serves. For example, if turkers and historians differ in their mean assessment of the relatedness of"sexuality" and "African history" (looking ahead: they do), a system serving historians may not be able to rely on a gold standard created by turkers.

The second research question measures whether different cultural communities exhibit different levels of agreement within their respective communities.

*RQ1b: Do different communities exhibit different levels of **agreement** in their SR ratings?*

If the answer to this question is yes, some cultural communities will need more contributors than others to obtain a desired level of gold standard sample error [1].

For each of four different analysis (two each for RQ1a and RQ1b), we report differences between the five conditions in

[5]We followed the procedure of Radinksy et al. [34] to ensure that the 50 pairs selected for each field captured a diverse set of relatedness values. Radinksy and colleagues randomly sampled concept pairs stratified by pointwise mutual information calculated using a New York Times corpus. We followed the same procedure, but used a domain-specific corpus for each field, each of which contained 800 scholarly publications in the field chosen by querying the top 50 documents in Google Scholar for each of the 16 interests in the field.

[6]https://www.mturk.com/mturk/help?helpPage=worker

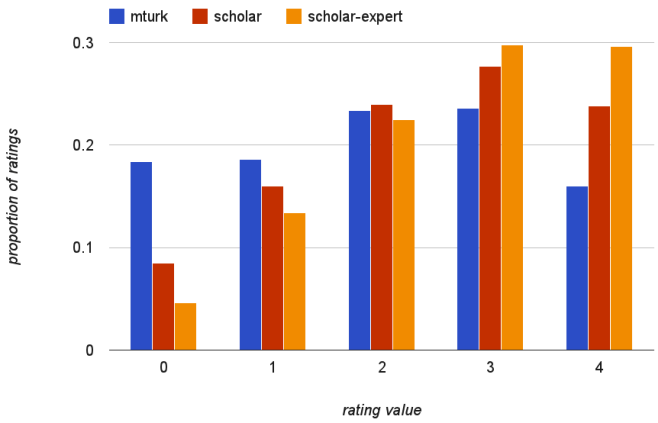

Figure 2. The distribution of ratings for domain-specific knowledge for turkers, scholars, and scholar-experts. In general, turkers judge concepts as less related than scholars and scholar-experts. For example, turkers are four times more likely to rate a domain-specific pair a "0" than scholar-experts.

Table 1 corresponding to the two knowledge types (general, domain-specific) and three cultural contexts (turker, scholar, scholar-expert).

RQ1a: Distribution of assessments

We begin by analyzing the overall distribution of SR ratings for each knowledge type and community. For general knowledge, turkers and scholars generate a similar distribution of ratings. There were no significant differences in mean SR rating between turkers ($\mu = 2.24, \sigma = 1.47, n = 461$) and scholars ($\mu = 3.36, \sigma = 1.4, n = 887$). However, an ANOVA that controls for concept pair means does indeed reveal significant differences between turkers and scholars ($p < 0.05$). We return to this point in the next section (Table 2).

The distribution of ratings for domain-specific knowledge (Figure 2) show marked differences. In aggregate, turkers rate pairs the lowest (mean 2.0), followed by scholars (2.42), and scholar-experts (2.67). Chi-square tests shows these effects to be significant ($\chi^2 = 279, p < 0.0001$). These differences are most apparent at the scale extremes. Turkers assign 18% of judgements a relatedness of 0 compared to 9% of scholars and 4% of scholar-experts. On the opposite end of the spectrum, scholar-experts assign 30% of judgements a 4 compared to 24% and 15% of judgements from scholars and turkers respectively.

In these results, we see our first evidence that both the 'turkers for all' and "communities matter" precepts hold in certain contexts. As Clark's notion of cultural communities suggests, scholar-experts perceive stronger relationships between concepts in their common ground when compared to turkers assessing concepts not in their common ground. The same is true to a lesser degree for scholars (non-experts), although they appear to share more common ground with scholar-experts, which is to be expected given their shared environment and experience.

concept 1	concept 2	turker mean	scholar mean	p-value
Arafat	peace	0.33	2.38	***
hardware	network	1.75	3.15	***
energy	consumer	1.09	2.45	**
plane	car	1.71	2.79	**
dollar	yen	2.89	3.78	**

Table 2. General knowledge concepts with the largest difference between turker and scholar means (* = $p < 0.001$, ** = $p < 0.01$ according to a two-tailed t-test).**

RQ1a: Correlation between community consensus ratings

In this section, we probe differences between average community ratings for both general and domain specific knowledge. For each concept pair, we calculate the mean at the community-level and combine these values to create community-level "consensus lists."

For general knowledge concepts, the Spearman correlation between consensus lists for turkers and scholars ($\rho_s = 0.91, n = 50$) approaches the estimated within-community correlation ($\rho_s = 0.94$).[7] While the between-community correlation for general concepts is high, there are certain concept pairs that are large outliers. Table 2 shows the general knowledge concept pairs that displayed the largest differences between turker and scholar ratings. As a reminder, assessments use a zero to four scale. Most notably, while turkers judge the relatedness between "Arafat" and "peace" to be 0.33 on average, or basically not related at all, the mean scholar rating for the pair is a moderate 2.38. As noted above, the ("Arafat", "peace") pair comes from the *WordSim353* dataset, which has been used to evaluate dozens of SR algorithms. The raters of *WordSim353* also gave the pair a moderate score (mean of 6.73 on a continuous 10-point scale). This suggests that SR algorithms have been evaluated (and trained, in many cases) using a point of view closer to scholars than turkers on this controversial subject. The reasons behind the significant divergence on this pair are unclear, but this result certainly raises the prospect of controversy (i.e. relationship valence) having an effect on semantic relatedness judgements across communities.

For domain-specific concepts, we find broader differences between communities' consensus ratings. While the correlation between turkers and scholars ($\rho_s = 0.88$, $n = 150$ questions) nearly matches the within-condition correlation estimates ($\rho_s = 0.89$, $n = 150$), scholar-experts and turkers are much farther apart in their consensus lists ($\rho_s = 0.76$, $n = 150$). Scholars and scholar-experts are in-between ($\rho_s = 0.82$, $n = 150$). As with the general knowledge analysis, the concept pairs with largest differences provide insight into the dynamics of community disagreement (Table 3). This list appears to favor broad, complex concepts ("sexuality", "African history", "popular religion"). One hypothesis is that scholar-experts' deep domain knowledge includes

[7]The use of Spearman's correlation coefficient is considered to be a best practice in the SR literature because it does not assume interval scales, does not make any assumptions about the distribution of ratings, and for a number of other reasons [42]. Within-condition correlations were estimated using a bootstrap procedure.

group 1	group 2	concept a	concept b	group 1 mean	group 2 mean	p-value
mturk	**scholar**	psychophysiology	aging	1.30	3.15	***
		introductory biology	statistics	0.80	2.11	***
		research methods	linguistics	1.23	2.51	***
mturk	**scholar-expert**	sexuality	African history	0.60	2.80	***
		cognition	acculturation	0.73	2.83	***
		modern European history	women	1.56	3.62	***
scholar	**scholar-expert**	sexuality	historical memory	2.07	3.50	**
		South Asia	popular religion	2.31	3.70	**
		collective memory	women	2.58	3.92	**

Table 3. Domain-specific concepts with the largest difference between turker and scholar means (* indicates $p < 0.001$, ** = $p < 0.01$ according to a two-tailed t-test).**

specific relationships that link concepts that may appear unrelated to non-experts (scholars and turkers). This would also explain scholar-experts' higher absolute SR scores.

These results show that the "communities matter" precept extends beyond individuals in communities to the gold standards that aggregate individuals' assessments. As with the previous analysis, scholars agree more often with scholar-experts than turkers, both in rating distribution and consensus rank order. This may be a sign of interdisciplinarity. Some scholars' specialized knowledge may reach beyond their core field of study. For example, a computer scientist who studies social computing may have some expertise in the SR judgement for the concept pair "personality" and "social psychology". On the other hand, these results may transcend domain knowledge and reflect other cultural commonalities shared by scholars.

RQ1b: Inter-rater agreement

We next examine the inter-rater agreement of judgements in different conditions by measuring the absolute difference between judgments for two raters in the same community.[8]

To calculate absolute differences, for each concept pair we compared responses from all pairs of human raters within a community. We report the mean absolute error (MAE) between the responses and the percentage of responses that disagreed by more than one rating point (e.g. 2 vs 4, or 2 vs 5). We determined confidence intervals for MAE and percentage significant disagreement using a bootstrapping procedure [27] and report 95% confidence intervals.

For general knowledge concepts, we found no meaningful differences in the agreement of within-group ratings. Turkers and scholars exhibited low absolute errors between ratings (turker MAE = 0.92$\pm$0.013, scholar MAE = 0.89$\pm$0.025) and few significant disagreements (22% $\pm$ 1.0 for both). For domain-specific concepts, we found significant differences in inter-rater agreement. Turkers disagreed most often, followed by scholars, followed by scholar-experts. These results held for both MAE (turker = 1.06 $\pm$ 0.007, scholar = 0.94 $\pm$0.005, scholar-expert = 0.86$\pm$0.007) and percent significant disagreements (turker = 28% $\pm$ 0.4, scholar = 23% $\pm$ 0.3, scholar-expert = 19% $\pm$ 0.6).

[8]We also calculated Pearson correlations between individual raters and found consistent results.

The difference between general knowledge and domain-specific agreement is particularly striking, and begs further study into the mechanisms underlying judgements of domain-specific concepts. The similar levels of agreement in general knowledge judgements suggest that agreement is not simply a function of professional background (scholar vs turker), but is also related to a subject's expertise in the topics of a judgement.

RQ1b: Intra-rater agreement

Finally, we examine the internal agreement within individual subjects. Recall that each subject provided five duplicate ratings for concept pairs, where the duplicating rating was separated from the original rating by at least one page. We report the average difference (MAE) between all subjects' duplicate ratings within a community. Since this represents a paired experimental design (unlike the previous analysis), we can calculate significance using straightforward t-tests.

Overall, the intra-rater agreement of ratings was high (MAE=0.38). Subjects exhibited higher intra-rater agreement for general concepts (MAE=0.237, σ=0.52, n=114) than domain-specific concepts (MAE=0.406, n=554, σ=0.610) These differences are significant (two-sample t-test, $p <= 0.05$).

As with inter-rater agreement, we saw no significant differences in the intra-rater agreement of general knowledge concepts for turkers (MAE=0.24, σ=0.532, n=42) and scholars (mae=0.224, σ=0.517, n=67). We did find significant differences in the intra-rater agreement of domain-specific concepts. Scholar-experts agreed with themselves most (MAE=0.260, σ=0.441, n=96), followed by scholars (MAE=0.382, σ=0.57, n=251), followed by turkers (MAE=0.537, σ=0.738, n=164). These pairwise differences are significant (two-tailed t-test, $p <= 0.05$).

The results for intra-rater agreement parallel the results for inter-rater agreement. We observe no significant community differences in agreement for general knowledge terms, but we do find significant differences for domain-specific knowledge. Scholar-experts agree most strongly, followed by scholars, followed by turkers.

RQ1: Summary

In summary, we find support for the "turkers for all" precept for general knowledge tasks, and "communities matter"

for domain-specific concepts. We observe community differences for domain-specific concepts across several measurements: differences in individual ratings, differences in aggregate consensus ratings, and agreement within and between raters.

These finding have direct implications for practitioners and researchers collecting domain-specific gold standards. When collecting a gold standard researchers and practitioners must consider the audience of the gold standard, the system or algorithm that uses it, and the type of knowledge. In addition, to achieve a desired level of sample error in a domain-specific gold standard, more data is needed from people without expert knowledge than with [1]. Because domain expertise is often valued in the marketplace, this finding exposes a monetary tradeoff: more data that is inexpensive from general workers, or less data that is expensive from domain experts.

RQ2: EFFECTS ON ALGORITHMS

We have shown that under some important conditions, different communities do indeed produce different gold standards. Next, we analyze whether the algorithmic "communities matter" precept also holds. That is, we ask:

RQ2: Do algorithms perform differently on gold standards from different cultural communities?

For example, given a particular algorithm and set of problem instances (e.g. concept pairs), do we see different results when the algorithm is trained and evaluated on data from biologists compared to turkers? If the algorithmic "communities matter' precept holds, system designers need to consider the relationship between three entities: the audience they serve, the communities producing their gold standard, and the algorithms that power their system.

We study this question by measuring the performance of SR algorithms on gold standards produced by each cultural community. We evaluate three state-of-the-art algorithms that have been shown to perform well on common gold standards:

1. **MW**, an algorithm developed by Milne and Witten that fist disambiguates each of the two phrases to Wikipedia articles (i.e. the phrase "Apple" $\rightarrow$ the article "Apple_Inc.") and then measures the overlap in links to and from the two articles [41] (MW).

2. **ESA**, Gabrilovich et al.'s Explicit Semantic Analysis algorithm creates a "concept vector" for each phrase by searching for Wikipedia articles whose text closely match the phrase [14]. The similarity between the phrases is the cosine similarity between the two concept vectors.

3. **ENS**, an ensemble algorithm that linearly combines a variety of SR algorithms as described by Hecht at al. [19].

We developed implementations of each algorithm available in the WikiBrain software framework [9], calibrated to perform similarly to published accuracy measurements [19]. Although SR research typically tunes, trains, and tests algorithms on the same dataset, we employ cross-validation because it is common practice in the CSCW and machine learning communities. Cross-validation results in slightly lower accuracy measurements than described in [19], but constitutes a more realistic measurement.

[9]http://wikibrainapi.org

	MW	ESA	ENS
Spearman's correlation, ρ_s	0.59	0.69	0.76

Table 4. Results for the three algorithms on the WordSim353 dataset. The three algorithms perform relatively similarly on other widely used datasets. Our use of cross-validation leads to slightly lower correlations than reported in the SR literature, but a more robust evaluation.

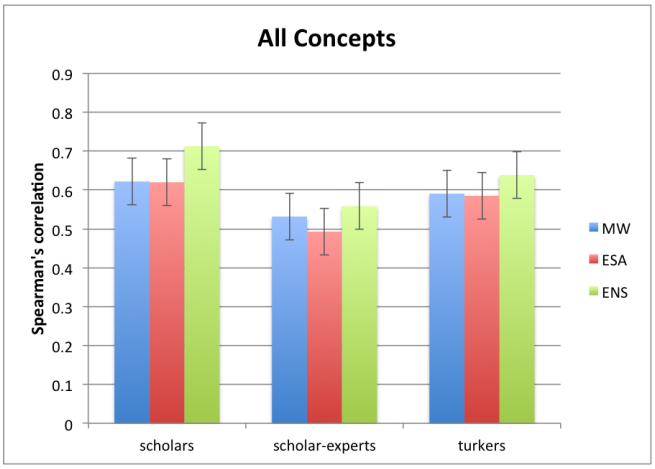

Figure 3. The correlation between algorithmic estimates and gold standard datasets for each of the gold standards and algorithms. The linear ensemble performs best across all groups, followed by MW and ESA. All algorithms are less accurate for scholar-experts than other cultural communities.

As a frame of reference, Table 4 shows the accuracy for the SR three algorithms on the widely used WordSim353 dataset [12]. The table shows Spearman's correlation, ρ_s, between SR estimates and the actual SR values in the gold standard. Spearman's correlation is the most commonly used evaluation measure in the SR community. The linear ensemble ENS performs best, followed by ESA and MW. These algorithmic results (ENS best, ESA second, MW third) generally hold across other reference datasets [19].

We first evaluate each algorithm's performance for each cultural community. Figure 3 shows Spearman's correlation for all three groups' datasets and all three algorithms. As expected, ENS performs best on all three datasets. However, for the domain-specific dataset, MW bests ESA for each group, contradicting years of SR results on datasets such as WordSim353. The differences in performance for ESA and MW may arise from the particular knowledge domains we consider. MW relies on link overlap, and concepts within the same domain likely contain links that overlap more often, creating a stronger signal for MW. In addition, ESA relies on textual search and may perform better for single words than the specific multi-word phrases that commonly appear in our dataset.

The differences in relative SR accuracy between WordSim353, a general knowledge gold standard, and our domain-

focused gold standard show that an algorithm that performs well on one gold standard need not perform well on others. They also point to the "community matters" hypothesis for algorithmic performance; MW may be more suitable for an audience of historians, biologists, or psychologists than in other settings.

To provide deeper insight into the algorithmic "community matters" hypothesis, we next focus on algorithmic peformance for domain specific questions, where we saw the largest community differences in previous analyses. As with previous analyses, each algorithm is trained and tested on one community (e.g. historians) by performing cross-validation against that community's gold standard. Figure 4 shows the results for the 150 domain-specific questions, broken out by community. The domain-specific results for the three top-level communities (scholars, scholar-experts, and turkers) show the same trend as all concepts (Figure 3), but with larger effects. The three scholar-expert sub-communities exhibit distinctive SR performance on the 50 concept pairs in their domain. All three algorithms perform roughly the same for historians, while ESA slightly outperforms MW for biologists. Psychologists emerge as a clear outlier, with ENS performing substantially worse than both MW and ENS.

The poor performance of ENS for psychologists is statistically significant, unique, and supports the "community matters' algorithmic hypothesis. The Spearman's correlation for ENS on the 50 psychology pairs is only 0.22 when trained and tested on the psychology dataset, compared to 0.67 for the turker dataset.[10] A standard two-sided confidence interval on this correlation falls just short of significant ($p = 0.06$). However, a non-parametric statistical evaluation that uses the Wilcoxon signed-rank test to leverage the paired structure of the observation finds them to be very significant ($p = 1 \times 10^{-5}$, details described in the Appendix). We next investigate this difference more deeply.

To understand why ENS performed poorly for psychologists, we next consider whether any consistents patterns arise in SR estimates for psychologists. The poor performance primarily arises from highly related concepts that are predicted as moderately related or unrelated. Figure 5 visualizes the five psychology concept pairs most responsible for the decline in algorithmic performance differences between the two communities. The x axis shows actual human relatedness responses (right is more related in the gold standard). The y axis shows algorithmic predicted relatedness (up is predicted to be more related). Arrows and colors visualize the change in actual and predicted relatedness from turkers (blue) to psychologists (red). The green diagonal line x = y shows accurate predictions (predicted rank = actual rank).

To provide intuition for the visualization, we describe the "health psychology", "child development" pair appearing in the upper left. First, consider the blue starting point of the line. The turker gold standard ranked the pair as the 29th most similar (X axis). The turker-trained algorithm predicted

[10]The Spearman's correlation of 0.71 in Figure 4 is calculated on all 150 pairs. When pruning to the 50 psychology pairs the correlation drops slightly to 0.67.

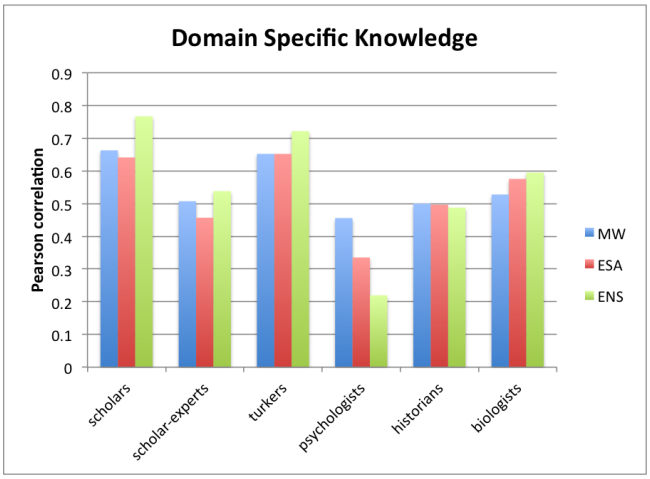

Figure 4. The correlation between algorithmic estimates and gold standard datasets for each of the gold standards and algorithms on the domain-specific questions.

the concept pair would be the 26th most similar pair of the 50 psychology terms (Y axis). Since this is a relatively accurate ranking (26 is close to 29), the blue starting point appears near the green line. The "red shift" to the upper left indicates that performance on the psychology gold standard worsened. Psychologists assessed it as the 37th most related pair (relatively unrelated), while the algorithm trained on the psychology gold standard yielded a predicted rank of 8 (quite related)/

Several patterns emerge from Figure 5. First, all five concept pairs show a decrease in performance for psychologists (i.e. a shift away from the green diagonal line). This is by design; this visualization includes the pairs most responsible for the decline in psychology performance. Second, all shifts except for "health psychology" / "child development" arise from an increase in actual gold standard relatedness coupled with a decrease in predicted relatedness. Third, many of the pairs include a relatively broad psychology concept ("cognition", "language", "neuroscience", "child development", "research methods").

These final two observations — 1) that broad concepts account for the most egregious scholar-expert errors, and 2) that algorithms typically *under*-predict the relatedness of broad concepts for scholar experts — are consistent with our earlier findings. Recall that broad concepts accounted for the largest community differences in human SR estimates (Table 3); scholar experts viewed broad concepts as more related than other cultural communities. Our findings also suggest that the knowledge encoded within Wikipedia, and the algorithms derived from this knowledge, may more closely reflect the perspectives of turkers than domain experts. If true, Wikipedia-based algorithms trained on data from turkers will not only perform differently for domain experts — they will likely perform worse. This may introduce a dangerous scenario in which an algorithm developer targeting domain experts may over-estimate the algorithm's performance if they test (incorrectly) on a turker gold-standard.

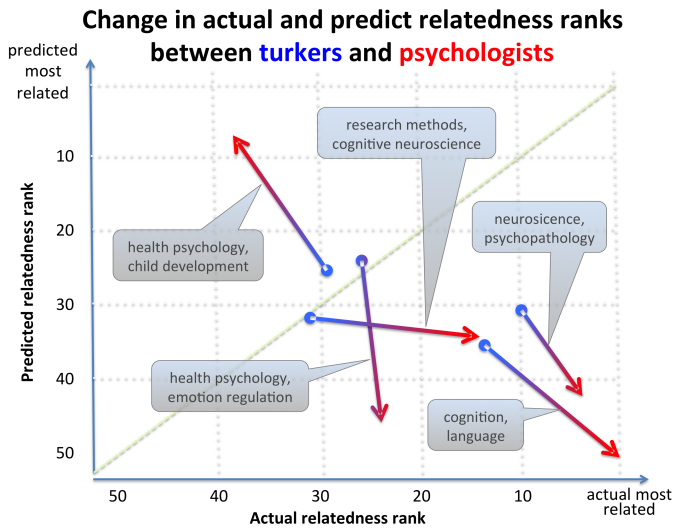

Figure 5. The five biggest shifts in ENS's predicted relatedness (X axis) and actual relatedness (Y axis) between turkers and psychologists. Arrows show change from turkers (blue) to psychologists (red). The green diagonal line shows accurate predictions (predicted rank = actual rank).

To summarize our findings for RQ2, we find support for the algorithmic "communities matters" precept in contexts when domain expertise is relevant. Algorithmic performance differences arise from differences in both community judgements and algorithmic estimates, and performance declines most when these two shifts occur in opposite directions. It seems that as with human assessments, broad abstract concepts lead to the largest community differences in algorithmic SR estimates. Our findings also may suggest that sophisticated algorithms with many tunable parameters (such as ENS) are more susceptible to "community confusion" than relatively simple algorithms with no tunable parameters.

DISCUSSION

This paper investigates the relationship between cultural communities, gold standards, and the algorithms that rely on them. Using 7,921 judgements from 111 subjects, we examine how SR judgements differ across three cultural communities: Mechanical Turk workers, scholars, and scholars who are experts in the field of an SR judgement. We show that different cultural communities can produce different gold standard judgements, and these differences can shape algorithmic accuracy.

The results reported in this paper have important implications for the use of Mechanical Turk in the development of gold standards. Specifically, they show that while workers can produce widely applicable gold standards for some knowledge tasks (i.e. the "turkers for all" precept), there are many others for which this is not the case (i.e. the "community matters" precept).

The work of Clark [9] provides a useful lens to help distinguish these two types of tasks. We saw that the "turkers for all" precept tended to hold when only considering general knowledge concept pairs, but broke down when the concept pairs were more domain-specific. Clark's notion of shared expertise helps explain this difference. For general knowledge concept pairs like those in WordSim353 (e.g. "war" and "troops"), the concepts and their predominant relationships can be understood to be defined similarly in a wide range of cultural communities' shared expertise. Given that SR has been defined as the number and strength of relationships between two concepts [20], broadly shared knowledge may allow communities that have little overlap in member — e.g. turkers and scholars — to provide similar SR ratings of general knowledge concept pairs. In this light, our findings support the use for Mechanical Turk in generating gold standards for knowledge tasks that involve widely-shared and agreed-upon knowledge.

However, our work also suggests that if a knowledge task does not satisfy the above condition — if it requires knowledge that may differ across cultural communities' — the use of AMT in gold standard development is problematic. We saw this occur with domain-specific concept pairs. The concepts included in these pairs are likely understood differently across the shared expertise of different professional cultural communities (i.e. domain experts vs. turkers). As a result, the SR estimates from these two groups differed significantly.

Mechanical Turk has generated gold standards for an enormous variety of knowledge tasks, some of which likely rely exclusively on general knowledge and others of which likely rely heavily on knowledge that varies across cultural communities. For example, simple information extraction tasks — e.g. validating a mailing address that was extracted from a web page — seem to fall in the former category. The knowledge required to validate a mailing address falls in many cultural communities' shared expertise (though certainly not all). On the other hand, other types of tasks like semantic relatedness result in dubious data and unreliable assessments of algorithmic performance. For example, image tagging and sentiment analysis may also represent cultural-specific knowledge tasks; previous research has pointed to cultural differences as a limiting factor in the accuracy of sentiment analysis [2] and it has shown that image tags can vary across national cultures [11].

The implications of this work extend beyond the use of AMT in gold standard development to problematizing the use of gold standards as a whole. While AMT has become the de facto gold standard generation engine in many areas of computer science, our work suggests that many of the same concerns with regard to shared knowledge may apply to the development of gold standards using other means (e.g. undergraduate students). Researchers and practitioners should carefully consider their specific knowledge tasks and the audience of their research and systems when deciding how to develop gold standards. Failing to do so can lead to misunderstandings of algorithmic performance, as we saw when we observed that a single SR algorithm trained on gold standards from two different communities can achieve dramatically different performance. For psychology SR questions, for example, state of the art algorithms are three times better at predicting turker assessments than psychologists for the same concept pairs.

Our findings also have specific implications for the way in which SR algorithms are evaluated and for the entire task of semantic relatedness estimation, a task that has spawned an enormous literature. Namely, our results suggest that SR algorithms should no longer be evaluated in terms of how well they predict contextless SR values, but rather on how well they predict SR values for a given community. In other words, the findings above imply that we need to move from a semantic relatedness defined by $SR(a, b)$ to one that is defined by $SR(a, b, community)$.

To support this change, we are publicly releasing our dataset of 7,921 estimates split across three communities and two types of knowledge (general and domain-specific)[11]. Using this dataset, the many researchers working to develop better SR algorithms can understand the performance of their algorithms according to each of the five conditions we consider in our experiment. Instead of reporting, for instance, "our algorithm's output had an ρ_s = 0.63 correlation with human judgements", SR researchers will be able to report "our algorithm's output had an ρ_s = 0.52 with turkers' judgements on general knowledge concept pairs, an ρ_s = 0.65 with scholar-expert's judgements" (and so on).

Finally, our work also provides insight into the underlying processes by which people make SR estimates. At a high-level, we have seen that differences in SR estimates can be understood through the lens of Clark's definition of cultural communities as the result of one group's shared knowledge that is lacked by other groups. However, Clark's definition is part of a theory of language as joint action, and future work should investigate whether a subject's beliefs about how their judgements will be used affect their responses. For instance, if a participant in our study assumed that her SR judgements would be seen by a very broad audience, there might be a tendency to "regress to the mean" in her assessments (e.g. ignoring a relationship between two concepts well-known in her field, but not widely). On the other hand, if the participant assumed that the judgements were only going to be seen or used by people in her direct academic field, it could have an opposite effect. We are currently working to develop a survey that uses prompts to test this hypothesis.

Our findings suggest a variety of additional new directions. Examining communities beyond turkers and scholars would enrich our understanding of the relationship between community membership and gold standard creation. Would we see the same effects across age groups? National cultures? Personality types? Home towns? Future research should also study whether our findings hold for other knowledge tasks beyond SR (e.g. sentiment analysis, image labeling).

APPENDIX: COMPARING SR ALGORITHMS

When analyzing RQ2, we wanted to know whether the large difference in Spearman's correlation for the ensemble SR algorithm train on turkers (ρ_s = 0.67) and psychologists (ρ_s = 0.22) was significant. A traditional two-sided confidence interval on the pair of correlations finds them to be just short of significant. However, this statistical test incorrectly assumes that the observations across samples are independent. In fact, the turker and psychology samples have a single observation for each of the 50 concept pairs.

To increase statistical power, we instead use a statistical procedure that leverages the paired structure of the observations. For each concept pair (e.g. "cognition", "language") we computed the rank error for both turkers (for the example, predicted rank = 13, actual rank = 36, rank error = 23) and psychologists (for the example, predicted rank = 1, actual rank = 48, rank error = 47). We then compute the difference in rank errors (23 - 47 = -24). If the errors are negative, the turker prediction was closer. If they are positive, the psychologist prediction was closer Of the 50 concept pairs, the turker-trained SR algorithm performed best on 36 (72%), the psychologist-trained SR aglorithm performed best on 11 (22%), and they tied on the remaining 3 (6%).

Finally, we performed a Wilcoxon signed rank test on the difference in rank errors to determine whether it statistically favored turkers or psychologists. A Wilcoxon test is appropriate because, like Spearman's correlation, it is non-parametric and tells us whether the observed differences differ significantly from zero. The p-value reported by the Wilcoxon test for these differences was $p = 1 \times 10^{-5}$.

[11] http://shilad.com/pluraSR200.html

ACKNOWLEDGMENTS

This research has been generously supported by Macalester College and the National Science Foundation (grants IIS-0964697 and IIS-0808692). We would also like to thank the Macademia users and Mechanical Turk workers who completed our online survey. This research would not be possible without their human SR judgements.

REFERENCES

1. Babbie, E. R., et al. *Survey research methods*. Wadsworth Belmont, CA, 1990.
2. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202* (2013).
3. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. Omnipedia: Bridging the wikipedia language gap. In *CHI '12* (2012).
4. Bergstrom, T., and Karahalios, K. Conversation clusters: grouping conversation topics through human-computer dialog. In *CHI '09* (Boston, MA, 2009), 2349–2352.
5. Bloodgood, M., and Callison-Burch, C. Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (2010).
6. Budanitsky, A., and Hirst, G. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics 32*, 1 (2006), 13–47.
7. Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's mechanical turk a new source of inexpensive,

yet high-quality, data? *Perspectives on Psychological Science 6*, 1 (Jan. 2011), 3–5.

8. Callison-Burch, C., and Dredze, M. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics (2010), 1–12.
9. Clark, H. H. *Using Language*. Cambridge University Press, May 1996.
10. Dong, W., and Fu, W.-T. Cultural difference in image tagging. In *CHI '10* (Atlanta, Georgia, USA, 2010), 981.
11. Dong, Z., Shi, C., Sen, S., Terveen, L., and Riedl, J. War versus inspirational in forrest gump: Cultural effects in tagging communities. In *ICWSM '12* (May 2012).
12. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. Placing search in context: The concept revisited. *ACM Transactions on Information Systems 20*, 1 (2002), 116–131.
13. Freitas, A., Oliveira, J. G., O'Riain, S., da Silva, J. C., and Curry, E. Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data & Knowledge Engineering 88*, 0 (2013), 126 – 141.
14. Gabrilovich, E., and Markovitch, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI '07* (Hyberabad, India, 2007).
15. Gergle, D., Kraut, R. E., and Fussell, S. R. Action as language in a shared visual space. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, ACM (New York, NY, USA, 2004), 487–496.
16. Gergle, D., Millen, D. R., Kraut, R. E., and Fussell, S. R. Persistence matters: Making the most of chat in tightly-coupled work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, ACM (New York, NY, USA, 2004), 431–438.
17. Grieser, K., Baldwin, T., Bohnert, F., and Sonenberg, L. Using ontological and document similarity to estimate museum exhibit relatedness. 10:110:20. Cited by 0013.
18. Halawi, G., Dror, G., Gabrilovich, E., and Koren, Y. Large-scale learning of word relatedness with constraints. In *KDD '12*, ACM (New York, NY, USA, 2012), 14061414.
19. Hecht, B., Carton, S. H., Quaderi, M., Schöning, J., Raubal, M., Gergle, D., and Downey, D. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ACM (2012), 415–424.
20. Hecht, B., and Gergle, D. The tower of babel meets web 2.0: User-generated content and its applications in a multilingual context. In *CHI '10*, ACM (Atlanta, GA, 2010), 291300. ACM ID: 1753370.
21. Heer, J., and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *CHI '10* (2010), 203212.
22. Ipeirotis, P. G. Demographics of mechanical turk.
23. Kittur, A., Chi, E. H., and Suh, B. What's in wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, ACM (New York, NY, USA, 2009), 1509–1512.
24. Liesaputra, V., and Witten, I. H. Realistic electronic books. *International Journal of Human-Computer Studies 70*, 9 (Sept. 2012), 588–610. Cited by 0002.
25. Miller, G. A., and Charles, W. G. Contextual correlates of semantic similarity. 1–28.
26. Milne, D., and Witten, I. H. Learning to link with wikipedia. In *CIKM '08* (Napa Valley, California, USA, 2008), 509518. ACM ID: 1458150.
27. Mooney, C. Z., Duval, R. D., and Duvall, R. *Bootstrapping: A nonparametric approach to statistical inference*. Sage, 1993.
28. Patwardhan, S., Banerjee, S., and Pedersen, T. Using measures of semantic relatedness for word sense disambiguation. In *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Springer Berlin Heidelberg, Jan. 2003, 241–257.
29. Pavlick, E., Post, M., Irvine, A., Kachaev, D., and Callison-Burch, C. The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics 2* (2014), 79–92.
30. Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics 40*, 3 (2006), 288–299.
31. Pirró, G., and Seco, N. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *On the Move to Meaningful Internet Systems: OTM 2008*, R. Meersman and Z. Tari, Eds., no. 5332 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Jan. 2008, 1271–1288.
32. Ponzetto, S. P., and Strube, M. Exploiting semantic role labeling, WordNet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (2006), 192199.
33. Popescu, A., and Grefenstette, G. Mining user home location and gender from flickr tags. In *ICSWM '10* (2010).

34. Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. A word at a time: Computing word relatedness using temporal semantic analysis. In *WWW '11* (Hyberabad, India, 2011), 337–346.

35. Resnick, P. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI '95* (Montreal, Quebec, Canada, 1995), 448–453.

36. Rubenstein, H., and Goodenough, J. B. Contextual correlates of synonymy. *Communications of the ACM 8*, 10 (Oct. 1965), 627633.

37. Schöning, J., Hecht, B., Raubal, M., Krger, A., Marsh, M., and Rohs, M. Improving interaction with virtual globes through spatial thinking: Helping users ask Why?. In *IUI '08* (Masapalomas, Gran Canaria, Spain, 2008), 129–138.

38. Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fastbut is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08* (2008), 254263.

39. Strube, M., and Ponzetto, S. P. WikiRelate! computing semantic relatedness using wikipedia. In *AAAI '06* (Boston, MA, 2006), 1419–1424.

40. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. Lexicon-based methods for sentiment analysis. *Computational linguistics 37*, 2 (2011), 267–307.

41. Witten, I., and Milne, D. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA* (2008), 25–30.

42. Zesch, T., and Gurevych, I. Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words. *Natural Language Engineering 16*, 1 (2010), 25.