Evaluation of Commonsense Knowledge with Mechanical Turk

Jonathan Gordon
Dept. of Computer Science
University of Rochester
Rochester, NY, USA
jgordon@cs.rochester.edu

Benjamin Van Durme
HLTCOE
Johns Hopkins University
Baltimore, MD, USA
vandurme@cs.jhu.edu

Lenhart K. Schubert
Dept. of Computer Science
University of Rochester
Rochester, NY, USA
schubert@cs.rochester.edu

Abstract

Efforts to automatically acquire world knowledge from text suffer from the lack of an easy means of evaluating the resulting knowledge. We describe initial experiments using Mechanical Turk to crowdsource evaluation to non-experts for little cost, resulting in a collection of factoids with associated quality judgements. We describe the method of acquiring usable judgements from the public and the impact of such large-scale evaluation on the task of knowledge acquisition.

1 Introduction

The creation of intelligent artifacts that can achieve human-level performance at problems like question-answering ultimately depends on the availability of considerable knowledge. Specifically, what is needed is commonsense knowledge about the world in a form suitable for reasoning. *Open knowledge extraction* (Van Durme and Schubert, 2008) is the task of mining text corpora to create useful, high-quality collections of such knowledge.

Efforts to encode knowledge by hand, such as Cyc (Lenat, 1995), require expensive man-hours of labor by experts. Indeed, results from Project Halo (Friedland *et al.*, 2004) suggest that properly encoding the (domain-specific) knowledge from just one page of a textbook can cost $10,000. OKE, on the other hand, creates logical formulas automatically from existing stores of human knowledge, such as books, newspapers, and the Web. And while crowdsourced efforts to gather knowledge, such as Open Mind (Singh, 2002), learn factoids people come up with off the tops of their heads to contribute, OKE learns from what people normally write about and thus consider important. Open *knowledge* extraction differs from open *information* extraction (Banko *et al.*, 2007) in the focus on everyday, commonsense knowledge rather than specific facts, and on the logical interpretability of the outputs. While an OIE system might learn that Tolstoy wrote using a dip pen, an OKE system would prefer to learn that an author may write using a pen.

An example of an OKE effort is the KNEXT system[1] (Schubert, 2002), which uses compositional semantic interpretation rules to produce logical formulas from the knowledge implicit in parsed text. These formulas are then automatically expressed as English-like "factoids", such as 'A PHILOSOPHER MAY HAVE A CONVICTION' or 'NEGOTIATIONS CAN BE LIKELY TO GO ON FOR SOME HOURS'.

While it is expected that eventually sufficiently clean knowledge bases will be produced for inferences to be made about everyday things and events, currently the average quality of automatically acquired knowledge is not good enough to be used in traditional reasoning systems. An obstacle for knowledge extraction is the lack of an easy method for evaluating – and thus improving – the quality of results. Evaluation in acquisition systems is typically done by human judging of random samples of output, usually by the reporting authors themselves (*e.g.*, Lin and Pantel, 2002; Schubert and Tong, 2003; Banko *et al.*, 2007). This is time-consuming, and it has the potential for bias: it would be preferable to have people other than AI researchers label whether an output is commonsense knowledge or not. We explore the use of Amazon's Mechanical Turk service, an online labor market, as a means of acquiring many non-expert judgements for little cost.

2 Related Work

While Open Mind Commons (Speer, 2007) asks users to vote for or against commonsense statements contributed by others users in order to come to a consensus, we seek to evaluate an automatic system. Snow *et al.* (2008) compared the quality of labels produced by non-expert Turkers against those made by experts for a variety of NLP tasks and found that they required only four responses per item to emulate expert annotations. Kittur *et al.* (2008) describe the use and

[1] Public release of the basic KNEXT engine is forthcoming.

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 159–162,
Los Angeles, California, June 2010.

The statement above is a reasonably clear, entirely plausible, generic claim and seems neither too specific nor too general or vague to be useful:

- I agree.
- I lean towards agreement.
- I'm not sure.
- I lean towards disagreement.
- I disagree.

Figure 1: Rating instructions and answers.

necessity of verifiable questions in acquiring accurate ratings of Wikipedia articles from Mechanical Turk users. These results contribute to our methods below.

3 Experiments

Previous evaluations of KNEXT output have tried to judge the relative quality of knowledge learned from different sources and by different techniques. Here the goal is simply to see whether the means of evaluation can be made to work reasonably, including at what scale it can be done for limited cost. For these experiments, we relied on $100 in credit provided by Amazon as part of the workshop shared task. This amount was used for several small experiments in order to empirically estimate what $100 could achieve, given a tuned method of presentation and evaluation.

We took a random selection of factoids generated from the British National Corpus (BNC Consortium, 2001), split into sets of 20, and removed those most easily filtered out as probably being of low quality or malformed. We skipped the more stringent filters (originally created for dealing with noisy Web text), leaving more variety in the quality of the factoids Turkers were asked to rate.

The first evaluation followed the format of previous, offline ratings. For each factoid, Turkers were given the instructions and choices in Fig. 1, where the options correspond in our analysis to the numbers 1–5, with 1 being agreement. To help Turkers make such judgements, they were given a brief background statement: "*We're gathering the sort of everyday, commonsense knowledge an intelligent computer system should know. You're asked to rate several possible statements based on how well you think they meet this goal.*" Mason and Watts (2009) suggest that while money may increase the number and speed of responses, other motivations such as wanting to help with something worthwhile or interesting are more likely to lead to high-quality responses.

Examples of *good* statements:

- A SONG CAN BE POPULAR
- A PERSON MAY HAVE A HEAD
- MANEUVERS MAY BE HOLD -ED IN SECRET
  It's fine if verb conjugations are not attached or are a bit unnatural, *e.g.* "hold -ed" instead of "held".

Examples of *bad* statements:

- A THING MAY SEEK A WAY
  This is *too vague*. What sort of thing? A way for/to what?
- A COCKTAIL PARTY CAN BE AT SCOTCH_PLAINS_COUNTRY_CLUB
  This is *too specific*. We want to know that a cocktail party can be at a country club, not at this particular one. The underscores are not a problem.
- A PIG MAY FLY
  This is *not literally true* even though it happens to be an expression.
- A WORD MAY MEAN
  This is *missing information*. What might a word mean?

Figure 2: The provided examples of good and bad factoids.

Participants were then shown the examples and explanations in Fig. 2. Note that while they are told some categories that bad factoids can fall into, the Turkers are not asked to make such classifications themselves, as this is a task where even experts have low agreement (Van Durme and Schubert, 2008).

An initial experiment (Round 1) only required Turkers to have a high (90%) approval rate. Under these conditions, out of 100 HITs[2], 60 were completed by participants whose IP addresses indicated they were in India, 38 from the United States, and 2 from Australia. The average Pearson correlation between the ratings of different Indian Turkers answering the same questions was a very weak 0.065, and between the Indian responders and those from the US and Australia was 0.132. On the other hand, the average correlation among non-Indian Turkers was 0.508, which is close to the 0.6–0.8 range seen between the authors in the past, and which can be taken as an upper bound on agreement for the task.

Given the sometimes subtle judgements of meaning required, being a native English speaker has previously been assumed to be a prerequisite. This difference in raters' agreements may thus be due to levels of language understanding, or perhaps to different levels of attentiveness to the task. However, it does not seem to be the case that the Indian respondents rushed: They took a median time of 201.5 seconds (249.18 avg. with a high standard deviation of 256.3 s – some took more than a minute per factoid). The non-Indian responders took a median time of just 115.5 s (124.5 avg., 49.2 std dev.).

Regardless of the cause, given these results, we restricted the availability of all following experiments to Turkers in the US.Ideally we would include other English-speaking countries, but there is no straight-

[2] Human Intelligence Tasks – Mechanical Turk assignments. In this case, each HIT was a set of twenty factoids to be rated.

	All		High Corr. (> 0.3)	
Round	*Avg.*	*Std. Dev.*	*Avg.*	*Std. Dev.*
1 (BNC)	2.59	1.55	2.71	1.64
3 (BNC)	2.80	1.66	2.83	1.68
4 (BNC)	2.61	1.64	2.62	1.64
5 (BNC)	2.76	1.61	2.89	1.68
6 (Weblogs)	2.83	1.67	2.85	1.67
7 (Wikipedia)	2.75	1.64	2.75	1.64

Table 1: Average ratings for all responses and for highly correlated responses. to other responses. Lower numbers are more positive. Round 2 was withdrawn without being completed.

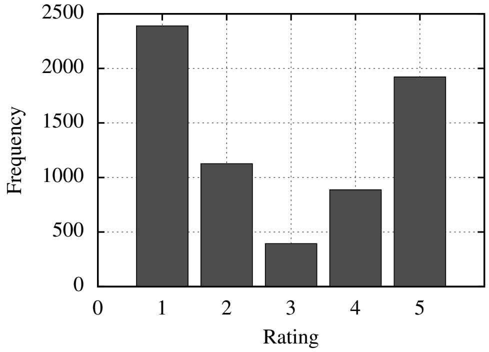

Figure 3: Frequency of ratings in the high-corr. results of Round 3.

forward way to set multiple allowable countries on Mechanical Turk.When Round 2 was posted with a larger set of factoids to be rated and the location requirement, responses fell off sharply, leading us to abort and repost with a higher payrate (7¢ for 20 factoids *vs* 5¢ originally) in Round 3.

To avoid inaccurate ratings, we rejected submissions that were improbably quick or were strongly uncorrelated with other Turkers' responses. We collected five Turkers' ratings for each set of factoids, and for each persons' response to a HIT computed the average of their three highest correlations with others' responses. We then rejected if the correlations were so low as to indicate random responses. The scores serve a second purpose of identifying a more trustworthy subset of the responses. (A cut-off score of 0.3 was chosen based on hand-examination.) In Table 1, we can see that these more strongly correlated responses rate factoids as slightly worse overall, possibly because those who either casual or uncertain are more likely to judge favorably on the assumption that this is what the task authors would prefer, or they are simply more likely to select the top-most option, which was "I agree".

An example of a factoid that was labeled incorrectly by one of the filtered out users is 'A PERSON MAY LOOK AT SOME THING-REFERRED-TO OF PRESS RELEASES', for which a Turker from Madras in Round 1 selected "I agree". Factoids containing the vague 'THING-REFERRED-TO' are often filtered out of our results automatically, but leaving them in gave us some obviously bad inputs for checking Turkers' responses. Another (US) Turker chose "I agree" when told 'TES MAY HAVE 1991ES' but "I disagree" when shown 'A TRIP CAN BE TO A SUPERMARKET'.

We are interested not only in whether there is a general consensus to be found among the Turkers but also how that consensus correlates with the judgements of AI researchers. To this end, one of the authors rated five sets (100 factoids) presented in Round 3, which yielded an average correlation between all the Turkers and the author of 0.507, which rises slightly to 0.532 if we only count those Turkers considered "highly correlated" as described above.

As another test of agreement, for ten of the sets in Round 3, two factoids were designated as fixpoints – the single best and worst factoid in the set, assigned ratings 1 and 5 respectively. From the Turkers who rated these factoids, 65 of the 100 ratings matched the researchers' designations and 77 were within one point of the chosen rating.[3]

A few of the Turkers who participated had fairly strong negative correlations to the other Turkers, suggesting that they may have misunderstood the task and were rating backwards.[4] Furthermore, one Turker commented that she was unsure whether the statement she was being asked to agree with (Fig. 1) "was a positive or negative". To see how it would affect the results, we ran (as Round 4) twenty sets of factoids, asking simplified question "Do you agree this is a good statement of general knowledge?" The choices were also reversed in order, running from "I disagree" to "I agree" and color-coded, with agree being green and disagree red. This corresponded to the coloring of the good and bad examples at the top of the page, which the Turkers were told to reread when they were halfway through the HIT. The average correlation for responses in Round 4 was 0.47, which is an improvement over the 0.34 avg. correlation of Round 3.

Using the same format as Round 4, we ran factoids from two other corpora. Round 6 consisted of 300 random factoids taken from running KNEXT on weblog data (Gordon *et al.*, 2009) and Round 7 300 random factoids taken from running KNEXT on Wikipedia.

[3] If we only look at the highly correlated responses, this increases slightly to 68% exact match, 82% within one point.

[4] This was true for one Turker who completed many HITs, a problem that might be prevented by accepting/rejecting HITs as soon as all scores for that set of factoids were available rather than waiting for the entire experiment to finish.

The average ratings for factoids from these sources are lower than for the BNC, reflecting the noisy nature of much writing on weblogs and the many overly specific or esoteric factoids learned from Wikipedia.

The results achieved can be quite sensitive to the display of the task. For instance, the frequency of ratings in Fig. 3 shows that Turkers tended toward the extremes: "I agree" and "I disagree" but rarely "I'm not sure". This option might have a negative connotation ("Waffling is undesirable") that another phrasing would not. As an alternative presentation of the task (Round 5), for 300 factoids, we asked Turkers to first decide whether a factoid was "incoherent (not understandable)" and, otherwise, whether it was "bad", "not very good", "so-so", "not so bad", or "good" commonsense knowledge. Turkers indicated factoids were incoherent 14% of the time, with a corresponding reduction in the number rated as "bad", but no real increase in middle ratings. The average ratings for the "coherent" factoids are in Table 1.

4 Uses of Results

Beyond exploring the potential of Mechanical Turk as a mechanism for evaluating the output of KNEXT and other open knowledge extraction systems, these experiments have two useful outcomes:

First, they give us a large collection of almost 3000 factoids that have associated average ratings and allow for the release of the subset of those factoids that are believed to probably be good (rated 1–2). This data set is being publicly released at `http://www.cs.rochester.edu/research/knext`, and it includes a wide range of factoids, such as 'A REPRESENTATION MAY SHOW REALITY' and 'DEMONSTRATIONS MAY MARK AN ANNIVERSARY OF AN UPRISING'.

Second, the factoids rated from Round 2 onward were associated with the KNEXT extraction rules used to generate them: The factoids generated by different rules have average ratings from 1.6 to 4.8. We hope in future to use this data to improve KNEXT's extraction methods, improving or eliminating rules that often produce factoids judged to be bad. Inexpensive, fast evaluation of output on Mechanical Turk could be a way to measure incremental improvements in output quality coming from the same source.

5 Conclusions

These initial experiments have shown that untrained Turkers evaluating the natural-language verbalizations of an open knowledge extraction system will generally give ratings that correlate strongly with those of AI researchers. Some simple methods were described to find those responses that are likely to be accurate. This work shows promise for cheap and quick means of measuring the quality of automatically constructed knowledge bases and thus improving the tools that create them.

Acknowledgements

This work was supported by NSF grants IIS-0535105 and IIS-0916599.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the Web. In *Proc. of IJCAI-07*.

BNC Consortium. 2001. The British National Corpus, v.2. Dist. by Oxford University Computing Services.

Noah S. Friedland *et al.*. 2004. Project Halo: Towards a digital Aristotle. *AI Magazine*, 25(4).

Jonathan Gordon, Benjamin Van Durme, and Lenhart K. Schubert. 2009. Weblogs as a source for extracting general world knowledge. In *Proc. of K-CAP-09*.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proc. of CHI '08*.

Douglas B. Lenat. 1995. Cyc: A Large-scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11):33–48.

Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proc. of COLING-02*.

Winter Mason and Duncan J. Watts. 2009. Financial incentives and the "performance of crowds". In *Proc. of HCOMP '09*.

Lenhart K. Schubert and Matthew H. Tong. 2003. Extracting and evaluating general world knowledge from the Brown corpus. In *Proc. of the HLT-NAACL Workshop on Text Meaning*.

Lenhart K. Schubert. 2002. Can we derive general world knowledge from texts? In *Proc. of HLT-02*.

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

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast – but is it good? In *Proc. of EMNLP-08*.

Robert Speer. 2007. Open mind commons: An inquisitive approach to learning common sense. In *Workshop on Common Sense and Intelligent User Interfaces*.

Benjamin Van Durme and Lenhart K. Schubert. 2008. Open knowledge extraction through compositional language processing. In *Proc. of STEP 2008*.