Association for the Advancement of Artificial Intelligence
HCOMP 2016

Best Paper Award

Matthew Lease

Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments

Tyler McDonnell
Dept. of Computer Science
University of Texas at Austin
tyler@cs.utexas.edu

Matthew Lease
School of Information
University of Texas at Austin
ml@utexas.edu

Mucahid Kutlu, Tamer Elsayed
Dept. of Computer Science and Engineering
Qatar University
{mucahidkutlu,telsayed}@qu.edu.qa

Abstract

When collecting subjective human ratings of items, it can be difficult to measure and enforce data quality due to task subjectivity and lack of insight into how judges' arrive at each rating decision. To address this, we propose requiring judges to provide a specific type of *rationale* underlying each rating decision. We evaluate this approach in the domain of Information Retrieval, where human judges rate the relevance of Webpages to search queries. Cost-benefit analysis over 10,000 judgments collected on Mechanical Turk suggests a win-win: *experienced* crowd workers provide rationales with almost no increase in task completion time while providing a multitude of further benefits, including more reliable judgments and greater transparency for evaluating both human raters and their judgments. Further benefits include reduced need for expert gold, the opportunity for *dual-supervision* from ratings and rationales, and added value from the rationales themselves.

1 Introduction

Ensuring data quality remains a significant challenge in crowdsourcing (Kittur et al. 2013), especially with paid microtask platforms such as Mechanical Turk (MTurk) in which inexpert, remote, unknown annotators are provided only rudimentary communication channels and training. The annotation process is largely opaque, with only the final labels being observable. Such factors do little to inspire trust between parties and faith in the overall paradigm. Risks may be seen to outweigh potential benefits, limiting the scale and complexity of tasks for which crowdsourcing is considered viable, and thereby the number of jobs made available to workers. When the accepted practice to ensure data quality requires posting a task redundantly to multiple workers, the cost of data collection increases and worker wages suffer.

We propose that *annotator rationales* (Zaidan, Eisner, and Piatko 2007) offer a new opportunity for traction on the above problems. The key idea of rationales is to ask human annotators to provide justifications for their labeling decisions in a particular, constrained form. As with Zaidan, Eisner, and Piatko (2007), we emphasize that the idea of rationales generalizes beyond the particular annotation task or form of rationale used (e.g., Donahue and Grauman (2011) investigate rationales for imagery tasks). However, while rationales were originally conceived merely to support a specific machine learning goal (and pursued with trusted annotators), we hypothesize that rationales offer far broader applicability and benefits to be realized (Section 2).

We ground our investigation of annotator rationales in the specific Information Retrieval (IR) task of *relevance assessment*, which calls on human judges to rate the relevance of *documents* (e.g., Webpages) to search queries. Unlike simple labeling tasks, describing relevance criteria precisely is difficult. Consequently, annotator agreement is typically low, even with trusted judges (Voorhees 2000; Bailey et al. 2008). While crowdsourcing's potential for more efficient relevance judging has sparked great interest (Alonso, Rose, and Stewart 2008), its use has tended to only further exacerbate issues of low annotator agreement.

In this work, we ask *assessors* to provide a rationale for each judgment by copy-and-pasting a short document excerpt (2-3 sentences) supporting their judgment. Table 4 shows examples. To collect relevance judgments, we created three task designs, iteratively refined through pilot experiments (Section 4). Our Standard Task collects relevance judgments without rationales. While intended as a baseline, it slightly outperforms careful task design of prior work (Hosseini et al. 2012), without any use of Honey-Pot questions or platform-specific worker filtering mechanisms. Our Rationale Task achieves further improvement, and remarkably, does so entirely from asking judges to provide rationales; the submitted rationales themselves are completely ignored. Moreover, we find that *experienced* workers (completing 20 or more tasks) are able to complete the Rationale Task with almost no increase in average task completion time (29 vs. 27 seconds). Finally, our Two-Stage Task asks one judge to complete the Rationale Task, then a second *reviewer* to verify or fix that judgment. With the same number of workers and task cost, the Two-Stage Task yields further improvement in quality over the Rationale Task.

Whereas Zaidan, Eisner, and Piatko (2007) motivate annotator rationales solely to support *dual-supervision* over collected rationales and labels, so far we have only discussed data quality improvement we achieve while ignoring the collected rationales. To derive further benefit from the rationales themselves, we hypothesize that our task design will

Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP), pages 139-148, 2016. https://doi.org/10.1609/hcomp.v4i1.13287