# Just ask a human? – Controlling Quality in Relational Similarity and Analogy Processing using the Crowd

Christoph Lofi

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
lofi@nii.ac.jp

**Abstract:** Advancing semantically meaningful and human-centered interaction paradigms for large information systems is one of the central challenges of current information system research. Here, systems which can capture different notions of 'similarity' between entities promise to be particularly interesting. While simple entity similarity has been addresses numerous times, relational similarity between entities and especially the closely related challenge of processing analogies remain hard to approach algorithmically due to the semantic ambiguity often involved in these tasks. In this paper, we will therefore employ human workers via crowd-sourcing to establish a performance baseline. Then, we further improve on this baseline by combining the feedback of multiple workers in a meaningful fashion. Due to the ambiguous nature of analogies and relational similarity, traditional crowd-sourcing quality control techniques are less effective and therefore we develop novel techniques paying respect to the intrinsic consensual nature of the task at hand. These works will further pave the way for building true hybrid systems with human workers and heuristic algorithms combining their individual strength.

## 1 Introduction

The increasing spread of the Internet and its multitude of information systems call for the development of novel interaction and query paradigms in order to keep up with the ever growing amount of information and users. These paradigms require more sophisticated techniques compared to established declarative SQL-style queries or IR-style keyword queries. Especially natural query paradigms, i.e. those query paradigms which try to mimic parts of natural human communication as for example questions answering [FB10], similarity browsing [Le10], or analogy queries require sophisticated semantic processing. One of these sematic processing steps is determining the *similarity* between two given entities. Broadly, similarity measures can be classified into two major categories: *attributional* similarity and *relational* similarity. Attributional similarity focuses only on the (usually explicitly provided) attributes values associated with entities, and is a well-researched problem with several efficient and high-quality algorithmic implementations, e.g., discovering similar vectors, or computing the similarity between database tuples. This allows for query-by-example interactions or similarity searches; for example, in an e-commerce setting, a user can simply provide an example object (e.g., a mobile phone), and the respective

information system can find other products with similar features and technical specifications [Lo10].

In contrast, *relational similarity* is a significantly more challenging problem. For example, there is a high relational similarity between the entity pairs (ostriches, bird) and (lions, cat) as ostriches are particularly large birds while lions are particularly large cats, i.e. the relations ("*x* is a very large *y*") between both is very similar. This type of similarity measurements is one of the central challenges of *analogy processing*, a core concept of human communication. Being able to reliably process and discover such similarities and analogies allows for a wide variety of new applications, e.g. more effective information extraction, analogy processing,, or certain subsets of question answering. However, actually assessing relational similarity is a difficult and extremely error-prone task for algorithms as well as for humans - current state-of-the-art algorithms achieve an average accuracy of "only" 56%.

Therefore, this paper explores the effectiveness of crowd-sourcing when applied to the challenging problem of relational similarity. The goal of this endeavor is establishing a baseline of human-based performance for further evaluations, and paving the way for later hybrid algorithms which combine algorithmic processing of relational similarity with on-demand crowd-sourcing for improved accuracy and reliability. It turns out that even human crowd-workers struggle with the complexity of assessing relational similarity correctly. This renders many established methods for quality control of crowd-sourcing tasks ineffective, i.e. Gold questions are hard to realize as even the performance of non-malicious workers is quite low (Gold questions: when using Gold questions, the correct answer to some tasks is known upfront. These tasks are transparently mixed into the set of real tasks and all workers which fail to answer the Gold questions correctly are excluded from the overall crowd-sourcing effort). Also, majority voting leads only to mediocre quality. Even more, for relational similarity or analogy, there is no clear "right" and "wrong" as the usefulness of a similarity statement basically depends on consensual agreement. Therefore, the central challenge approached in this paper is the problem of controlling quality of a crowd-sourcing task where individual worker responses are highly unreliable individually and no hard information on a solution's correctness is available during runtime.

The contributions of this paper can therefore be summarized as follows:

- We introduce the challenging problem *analogy processing* and *relational similarity*
- We provide a *baseline* for human performance when confronted with this problem
- We analyze the behavior of crowd-workers with respect to their *reliability* and *performance*
- We extensively discuss several aggregation techniques for improving and *controlling quality* of the final result, suitable for general crowd-sourcing problems relying on multiple choices and a community consensus
- We outline how human-based approaches to solving analogy challenges compare to currently researched techniques for automatically solving the problem performance-wise

## 2 Evaluation Scenario: Analogy Detection

Most human cognition is based on processing similarities of conceptual representations. During nearly all cognitive everyday tasks like e.g., visual perception, problem solving, or learning, we continuously perform *analogical inference* in order to deal with new information [GM97] in a flexible and cross-domain fashion. It's most striking feature is that analogical reasoning is performed on high-level relational or even perceptional structures and properties. Moreover, in contrast to formal reasoning, deduction, or formal problem solving, the use of analogies (and also analogical inference) appears to be easy and natural to people (contrary to needing a lot of training and experience). As analogical reasoning plays such an important role in many human cognitive abilities, it has been suggested that this ability is the "core of cognition" [Ho01] and the "thing that makes us smart" [Ge03]. Due to its ubiquity and importance, there is long-standing interest in researching the foundations and principles of analogies, mainly in the fields of philosophy, but later on also in linguistics and especially in the cognitive sciences. From here, research slowly spreads to computer science.

In general, an analogy is a cognitive process of transferring some high-level meaning from one particular subject (often called the *analogue* or the *source*) to another subject, usually called the *target.* When using analogies, one usually emphasizes that the "*essence"* of source and target is similar, i.e. their most discriminating and prototypical behaviors are perceived in a similar way. In this paper, we focus on the so-called *4-term analogy model* which is a simpler case of general analogies focusing closely on entities which behave similarly or have a similar role, i.e. those entities which are relational similar. For example, consider the following analogy question: "What is to the sky as is a ship to the ocean?" The obvious answer is 'an airplane', as airplanes are used to travel the sky as ships are used to travel the oceans. The actual differences in their physical properties (like the shape, color, or the material) and their non-prototypical relations to other concepts are ignored, as the ability of ships and airplanes to 'travel a certain domain' is highly similar and dominant. However, the "travels" relation in both term pairs is not the same, as traveling water and traveling air is different in many aspects. Still, by focusing on the important aspects of the relation and ignoring the others, the analogy can be drawn.

This simple type of analogy is also actively researched in linguistics, because many aspects of language evolution and the development of new words via neologisms are based on such simple analogies (think of the time when the word 'spaceship' appeared for the first time: while the word was new, many aspects of its meaning are immediately clear even to people who did not know what a spaceship actually is or how it exactly works). While the example of $ship : ocean :: airplace : sky$ (using Aristotele's notation) is quite illustrative, analogical reasoning can quickly become ambiguous when relational similarity between terms is either not strong enough or too many candidate pairs with similar relational strength exist (e.g. consider "What is to land as is a ship to the ocean?"; here, a correct answer is hard to determine without further information, should the answer be a car, a bus, a truck, an ox cart?). Although being just a simple subcase, the 4-term analogies play an important role in many real life analogies, and are thus very intriguing for computer science research.

### 2.1 Analogy and Similarity

The relation between analogy and similarity is by many people considered to be a confusing one. This is due to the fact that while analogy is not the same as similarity, analogical reasoning heavily relies on various 'flavors' of similarity. Therefore, sometimes it is claimed that there is the concept of *generic similarity* [Br89], for which the commonly used *property similarity* (what people usually mean when referring to 'similarity'), is just a special case. Other special cases of generic similarity are analogy in terms of relational similarity (e.g. 4-term analogy used in this paper) and structural similarity (e.g. complex analogy).

Like analogy, property similarity also establishes a certain relation between a source concept and a target concept. In this sense, when assuming knowledge provided in form of prepositional networks as in the structure mapping theory, 'normal' (property) similarity can be defined when most of the attributes/properties and the relations of the source are similar to those of the target [Ge83], e.g. if claiming that the "Kepler-30 star system is like our solar system[1]", this is a property similarity statement because both are star systems, both have similar suns, and both show similar planetary trajectories (albeit Kepler-30 has less planets with different properties). In contrast, claiming that "Atoms are like our solar system" is indeed an analogy, as atoms and the solar systems have no similar attributes, but do have similar relations between their related concepts. This difference between analogy and property similarity is quite significant, as similarity is a well-researched problem in information systems, with many efficient and mature implementations already published. Furthermore, similarity can be computed much easier as it mostly relies on attribute values, which are usually readily available. In contrast, computing relational similarity is a little researched problem, and also requires vast and diverse semantic knowledge bases which are difficult to obtain. Therefore, relational similarity is an entirely different challenge.

#### 2.1.1 SAT Analogy Challenges

While not being very complex analogies, the SAT analogy challenges deserve some special attention due to their importance with respect to previous research in computer science. In the study in later sections of this paper, we will therefore also use the SAT dataset for our experiments. The SAT test is standardized test for general college admissions in the United States. It features major sections on analogy challenges to assess the prospective student's general analytical skills. These challenges loosely follow Kant's notion of analogy as relational similarity between two ordered pairs is required, and the challenges are therefore expressed as 4-term analogy problems (but with multiple choices answers): out of a choice of five word pairs, one pair has to be found which is analogous to a given word pair.

As an example, consider this challenge:
*legend* is to *map* as is
a) *subtitle* to *translation* b) *bar* to *graph* c) *figure* to *blueprint*
d) *key* to *chart* e) *footnote* to *information*

[1] http://web.mit.edu/newsoffice/2012/far-off-solar-system-0725.html

Here, the correct answer is d) as a key helps to interpret the symbols in a chart as does the legend with the symbols of a map. While it is easy to see that this answer is correct when the solution is provided, actually solving these challenges seems to be a quite difficult task for aspiring high school students as the correctness rates of the analogy section of SAT tests is usually reported to be around 57%.

Currently, many research works which deal with analogies in computer science aim at solving the SAT challenges, such as [Bo09, Da08]. This is due to the fact that solving these challenges is significantly easier than dealing with general analogies: basically, the relational similarity for each of the answer choices to the source term pair is computed, and the most similar one is picked. But still, this task is very difficult and far from solved in a satisfied fashion. We outline the current progress of these efforts in comparison to human performance in section 3.3.

## 3 Crowd-Sourcing Analogy Problems

In the following, we will study the performance of crowd-sourcing in the context of analogy and relational similarity processing. A major focus is on result reliability, a major issue in crowd-sourcing due to malicious or simply incompetent workers. In most simple cases, these concerns can be addressed effectively with quality control measures like majority voting or Gold sampling [Lo12]. In previous studies on crowd sourcing it has been shown that within certain bounds, missing values in database tuples can be elicited with reliable efficiency and quality as long as the information is generally available. Especially for factual data that can be looked-up on the Web without requiring expert knowledge (e.g., product specifications, telephone numbers, addresses, etc.), the expected data quality is high with only a moderate amount of quality control (e.g., majority votes). For example, [Lo12] reports that crowd-sourced manual look-ups of movie genres in IMDB.com are correct in ~95% of all cases with costs of $0.03 per tuple (including quality assurance). Unfortunately, solving analogy problems falls into the difficult complexity class of consensual problems requiring competent workers with certain non-ubiquitous skills. This kind of problem is difficult to approach with simple quality control techniques [Se12].

The problem of controlling quality in this case stems from the very nature of analogies (and is comparable to problems faced in other perceptual tasks like measuring property similarity, finding prototypes, identifying essential properties, etc.). In particular, this is because there is no clear "right" or "wrong" for analogies, but the "correctness" of an analogy statement depends on the general perception and consensus of the people using it and may change over time or between different groups of people.

For example consider: *lions* is to *mammals* as is
a) *crocodile* to *reptiles* b) *shark* to *fish* c) *Queen Elisabeth* to *the English*

Every one of these answers could be argued for, and none is inherently more correct than the others. Therefore, when crowd-sourcing analogy challenges one does not look for correct answerers (which usually do not exist), but for those which people perceive to be the better answer. This unfortunately renders Gold questions, one of the most powerful quality

management tools for factual information unusable as this would wrongfully punish people to voice their opinion on an ambiguous problem. Furthermore, this also means that all other quality control techniques developed for crowd-sourcing relying on the existence of clear truth values as for example probabilistic error models [Li12,Wh09] are inherently unsuitable for analogies, similarity and other consensus-based problems. Therefore, the techniques developed in this paper are heavily based on inter-worker agreements in order to filter out malicious or incompetent workers, and therefore leading to high-quality judgments.

However, as measuring performance and quality of any heuristic for an ambiguous and consensus-based task is difficult to realize, the following study is based on the previously mentioned SAT dataset. Besides containing only simple and easier to handle 4-term analogy problems, this dataset has another unique and exceptional property: as the dataset is used in college admission test, every task is specifically and carefully hand-tailored in such a way that there is only one single non-ambiguous correct answer to every challenge. Therefore, if a worker understands the tasks correctly and knows all involved entities and their relations, he then can indeed derive a correct and undisputable answer. This property is not representative for real-life analogies, but allows us in this study to observe worker performance in a meaningful way.

### 3.1 Crowd Worker Baseline Performance

The following study is based on a crowd-sourcing experiment on CrowdFlower.com encompassing all 353 challenges from the SAT dataset. For each challenge, 8 judgments from workers which are recruited from the Amazon Mechanical Turk worker pool are elicited. As solving SAT challenges requires a very good understanding of language and vocabulary, we restricted the worker pool to only native speakers from the English-speaking countries Great Britain, United States (from where the SAT test originates, i.e. there

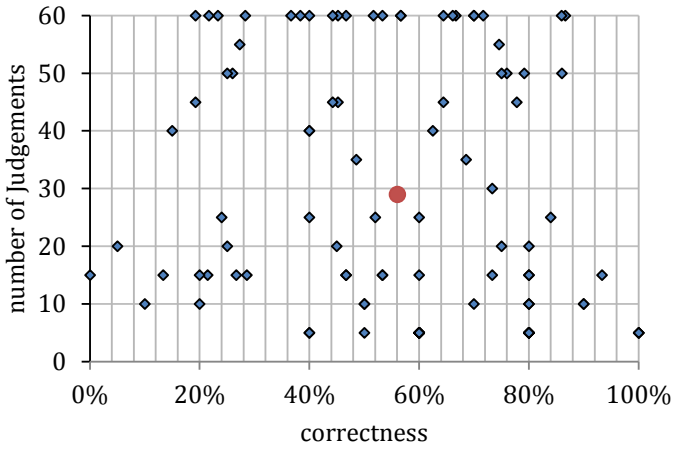


Figure 1. Worker analysis: Number of judgments vs. average correctness for each worker (99 workers overall, marked dot is the average:56% and 29 judgments )

is a high chance that some workers specifically trained solving these challenges), and Australia. Each HIT (smallest work unit) issued to workers consists of 5 analogy challenges for which we paid $0.15 overall (this amount is rather high compared to other crowd-sourcing tasks on Amazon Mechanical Turk). This results in an overall cost of $84.72 paid directly to the workers plus additional platform overhead fees. In order to enforce a higher worker churn and discouraging a single worker solving all challenges alone (and therefore also limiting the effects a malicious worker can have on the results), we restricted each worker to solving only 60 challenges (12 HITS). In accordance to the above observations that generally analogies are consensual, no Gold questions are used.

This resulted in 99 workers participating in this study. In average, each worker solved ~29 SAT challenges, i.e. most workers decided not continuing to work on our task after slightly less than 6 HITs in average. This can be attributed to the fact that in comparison to many common crowd-sourcing tasks, solving analogies is rather difficult and requires some deeper though in order to solve the tasks reliably. When observing the correctness of the answers of each individual worker (which can be measured due to the special nature of the SAT dataset), then the average correctness of all judgments is 56% - a number closely matching the reported performance of college applicants performing the SAT test for real. There is also no significant correlation between performance and number of judgments (Pearson correlation of 0.18 towards that workers providing many judgments have lower accuracy). This means that in general, workers tried to solve the tasks and did not simply and blatantly cheat by providing 60 random judgments. A visualization of general worker behavior is given in Figure 1.

Next, we observe the individual judgments themselves, especially with respect to the ambiguousness or unequivocalness of the crowd workers. As it turns out, workers are rarely in agreement on which answer choice is the correct one. This is shown in Figure 3. Out of 5 possible answers choices, most challenges (>67%) received 3 or 4 different answers overall by just 8 crowd-workers; for 9% of all challenges even every single answer choice was selected at least once. Only 3.9% off all challenges are unequivocal, and one of those is even unequivocally wrong. Just 19.5% of all challenges received rather focused answers with just two different result choices.

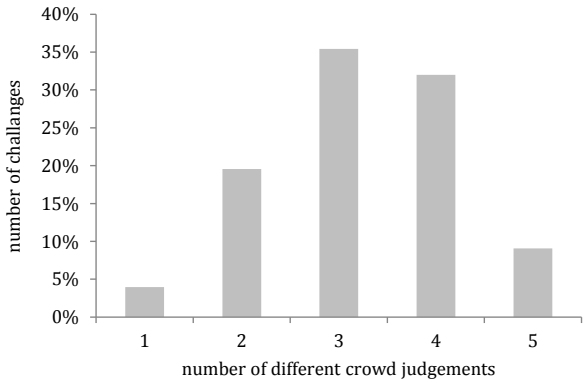

Figure 2. Different answer choices (out of 5) provided by 8 crowd workers for all SAT challenges

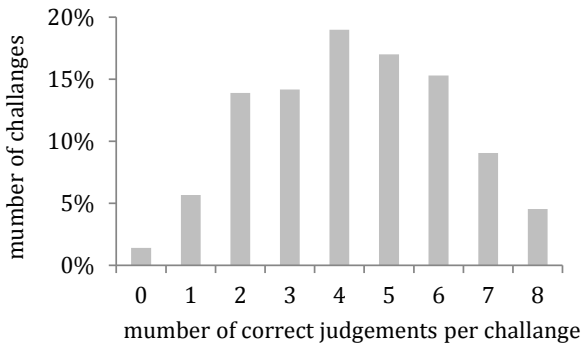


Figure 3. Analogy challenges: out of 8 judgments per challenge, how many judgments are correct? (353 challenges)

As each SAT challenge does indeed have a single carefully engineered correct answer, we can in the following focus on how reliably the workers choose this correct choice. In average only 4.2 workers out of 8 judge vote for the correct answer for each analogy challenge. A full overview of correct judgements is shown in Figure 3. Only very few challenges have an unequivocal and correct result (3.6%) while most challenges can only obtain 4 correct votes. Therefore, the base performance of our crowd-workers is rather low showing many disagreements between the workers themselves. Still, when simply performing a majority vote, this leads to an overall accuracy of 69%, which is clearly better than the performance of a single worker but not satisfying from a general perspective

### 3.2 Improving Quality and Capturing the Crowd Consensus

In this section, we will introduce heuristics modifying the basic majority vote in such a way that the overall result quality achieved by combining the feedback of the crowd workers further increases. Here, we aim at minimizing the impact of malicious and or non-competent workers while giving more power to workers which have proven themselves trustworthy and valuable. In contrast to the probabilistic approaches as e.g. in [Li12, Wh09], which rely on strict truth values for calibrating the heuristics, our heuristic aims at capturing the overall worker *consensus* and are therefore better suited for tasks like solving analogy problems or establishing similarity.

Our basic technique for deriving the best answer to an analogy challenge is a weighted majority vote, with the weight representing the credibility and reputation of a given worker. We will visit different definitions for the reputation weighting factor $r_i$ later in this section. The generic definition for our basic voting heuristic is then given by:

Definition 1 (Weighted Majority Vote): Given a multiple-choice crowd sourcing task $t$ with possible choices $C_t = \{c_{t,1}, \dots, c_{t,N_c}\}$, a set of workers $W = \{w_1, \dots, w_{N_W}\}$, and a set of corresponding worker judgements $J \subset (W \times C_t)$. Then the aggregated judgement $c_v$ can be derived by:

$$v = \max_{j=1}^{N_C} \sum_{i=1}^{N_W} r_i \delta_{ij}$$

with $r_i$ being the reputation of worker $w_i$ and

$$\delta_{ij} = \begin{cases} 1 & \text{if } w_i \text{ votes for choice } c_{t,j}; i.e. \;\; if \; (w_i, c_{t,j}) \in J \\ 0 & \text{otherwise} \end{cases}$$

### 3.2.1 Agreement-based Reputation Weighting

In this first version of our reputation-weighting heuristic, worker reputation is based on their overall agreement with the predominant opinion of the work force. The rationale behind this technique is that while the predominant opinion (which can be derived by a simple majority vote) might be adulterated by malicious or incompetent workers, it is still a good indicator hinting at the best solution. Therefore, any worker who regularly agrees with the community opinion can be considered more valuable than a worker who is frequently disagreeing (and is most likely malicious, incompetent, or just has exotic and therefore generally unhelpful opinions). When combined with definition 1, this results in a 2-stage algorithm in which first all judgments are aggregated with simple majority votes in order to derive the worker reputation, and then are again aggregated using weighted majority votes incorporation said reputation to compute the final crowd judgment.

These considerations lead to following definition of worker reputation:

Definition 2 (Agreement-based Reputation): Given a set of crowd-sourcing tasks $T = \{t_1, \dots, t_{N_T}\}$ and according choices $C_t$ and workers $W$ as in definition 1, and worker judgements $J \subset \left(W \times \bigcup_{t \in T} C_t\right)$. Then, agreement-based reputation can be computed as follows:

$$r_i = \left(\frac{1}{N_T}\right) \sum_{t \in T} \alpha_{t,i}$$

with

$$\alpha_{t,i} = \begin{cases} 1 & \text{if } \left(w_{i,}\, c_{t,v_{majority}}\right) \in J \\ 0 & \text{otherwise} \end{cases}$$

whereas $c_{t,v_{majority}}$ is the result of a simple majority vote of all judgements of task $t$

### 3.2.2 PageRank-based Reputation Weighting

During this study, we also tried to apply established reputation management algorithms from other domains. Especially, the challenge of reputation management for crowd-sourcing can also be modeled as a graph: here, each worker is a node, and every time one worker decides for the same solution choice as a given second worker for a crowd-sourcing task, this can be seen as a positive vote of the first worker towards the second one (as each worker believes his choice is correct, he also vouches for other workers sharing his beliefs). Each of these votes then results in a directed edge in the graph. This setup allows

us to employ common link analysis algorithms as for example PageRank [BP98] to determine the importance and reputation coefficient $r_i$ for all workers .

Definition 3 (PageRank-based Reputation): Given tasks $T$, according choices $C_t$, workers $W$, and respective judgements $J$ as in definition 2, then an agreement reputation graph can be defined as follows:

Let $RG$ be a directed graph with $RG = (W, E)$ with $E$ being the set of agreement edges given by:

$$E = \{(w_i, w_j) \in W \times W \mid \exists c_{t,x} \in J: ((w_i, c_{t,x}) \in J \wedge (w_j, c_{t,x}) \in J)$$

The final reputation weights for all workers $W$ can be obtained by applying the PageRank algorithm on RG.

### 3.2.3 Hard Reputation Limits

The previous two techniques for capturing worker reputation can be further modified to exclude all workers whose reputation is below a certain threshold. When the threshold is well chosen, this tightly limits the impact of malicious workers. The effect of different thresholds is examined in the next section.

## 3.3 Evaluation

In the following, we will evaluate our aggregation heuristics on the judgments obtained in the study presented in section 3.1, and compare the results to currently established computer-based techniques for solving SAT challenges. This comparison allows for a rough assessment of the potential of different approaches and is summarized in Table 1 (page 12).

Regarding human performance, unanimous votes are close to meaningless with correctness of just 3.6%. Average worker performance is at 56%, with 69% being the accuracy achieved by simple majority votes. Our proposed weighted majority votes significantly improve this initial result quality. By using PageRank-based reputation weights, the performance is pushed to 74.8% and can be further improved to 75.4% by excluding the workers whose reputation is below 40% of the maximal reputation. However, the simpler agreement-based heuristics fares significantly better with a baseline of 76.2% and an accuracy of 81.5% when excluding workers with a reputation below 50% of the maximal reputation.

However, finding a suitable limit for excluding low-reputation workers is non-trivial as the optimal cut-off thresholds varies with the chosen heuristic, dataset, and worker pool. The effects of different thresholds are visualized in Figure 4, and show clearly that choosing the threshold too low or too high will yield sub-optimal results. Furthermore, it also shows that limiting the majority vote to include only high reputation workers does not necessarily lead to a better result. This is due to the fact

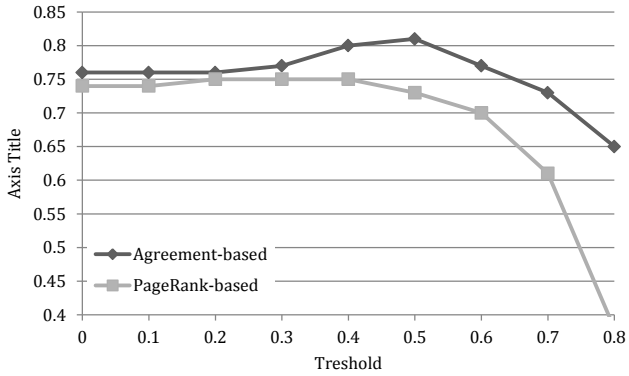


Figure 4. SAT Correctness with increasing reputation thresholds

that reputation for both heuristics is heavily based on agreement among workers, and captures real worker performance only approximately (please note that even though it might be intriguing to design an aggregation heuristics for analogies which is self-tuning by analyzing which workers frequently answer correctly, this approach is not suitable in general for non-SAT analogies or similar problems as usually there are no 'correct' answer which could serve as a Gold standard).

## 4 Summary and Outlook

In this paper, we examined crowd-sourcing techniques for approaching the challenge of analogy processing and relational similarity. Due to the consensual and ambiguous nature of these tasks, worker's average performance is rather low, rendering traditional quality control techniques like simple majority votes or Gold questions less effective. We approached this issue by weighted majority votes which balance each user judgment with the respective user's reputation. For capturing the notion of reputation, we showcased two heuristics both based on inter-worker agreement. By using these techniques, the measured result quality in the study we performed on the well-known SAT analogy dataset could be raised from 69% for simple majority votes to 81.5%.

In future works, one central challenge will be to further minimize the costs for obtaining crowd judgments. Especially, our approach can be modified in such a way that it respects the ambiguity of a given challenge. Therefore, for each challenge an optimal number of crowd judgments can be elicited (more judgments for ambiguous challenges, less judgments for seemingly easier challenges with more unequivocal results). Finally, this research should lead to dynamically combining the outputs of available machine-based techniques with focused crowd-sourcing in order to build true hybrid systems. Most of the system should rely on automatic algorithms, but in case of doubt, crowd-sourcing can be used to provide additional input.

| Algorithm / Approach | Score (%) |
|---|---|
| Random guessing | 20.0 |
| Jiang and Conrath [Tu06] | 27.3 |
| Lin [Tu06] | 27.3 |
| Leacock and Chodrow [Tu06] | 31.3 |
| Hirst and St.-Onge [Tu06] | 32.1 |
| Resnik [Tu06] | 33.2 |
| PMI-IR [Tu06] | 35.0 |
| SVM [Bo08] | 40.1 |
| LSA + Predictation [Bo12] | 42.0 |
| WordNet [Ve04] | 43.0 |
| Bicici and Yuret [BY06] | 44.0 |
| VSM [TL05] | 47.1 |
| Combined [Bo12] | 49.2 |
| Pair-Classifier [Tu08] | 52.1 |
| RELSIM [Bo09] | 51.1 |
| Pertinence [Tu06-2] | 53.5 |
| LRA [Tu06] | 56.1 |
| *High-school Students (historic SAT data)* | 57.0 |
| *Average Mechanical Turk Worker* | 56.0 |
| *MTurk Unanimous Vote (8 Jugements)* | 3.6 |
| *MTurk Majority Vote (8 Jugements)* | 69.0 |
| *MTurk Weighted Vote (8 Jugements, PageRank)* | 74.8 |
| *MTurk Weighted Vote (8 Jugements, PageRank, Limit 0.4)* | 75.4 |
| *MTurk Weighted Vote (8 Jugements, Agreement )* | 76.2 |
| *MTurk Weighted Vote (8 Jugements, Agreement, Limit 0.5)* | 81.5 |

Table 1. Reported correctness scores for computer-based and human-based approaches for solving SAT challenges

## Funding Acknowledgements

This work was supported by funds provided by the DAAD FIT program (http://www.daad.de/fit/).

## References

[Bo08] Bollegala, D. et al.: Www sits the sat: Measuring relational similarity on the web. Europ. Conf. on Artificial Intelligence (ECAI), Patras, Greece, 2008.

[Bo09] Bollegala, D.T. et al.: Measuring the similarity between implicit semantic relations from the web. 18th Int. Conf. on World Wide Web (WWW), Madrid, Spain, 2009.

[Bo12] Bollegala, D. et al.: Improving Relational Similarity Measurement using Symmetries in Proportional Word Analogies, Information Processing & Management, 2012.

[BP98] Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 30, 1-7, 107–117, 1998.

[Br89] Brown, W.R.: Two Traditions of Analogy. Informal Logic. 11, 3, 160–172, 1989.

[BY06] Bicici, E., Yuret, D.: Clustering word pairs to answer analogy questions. TAINN, 2006.

[Da08] Davidov, D.: Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions. Ass. for Computational Linguistics: Human Language Technologies (ACL:HLT). , Columbus, Ohio, USA, 2008.

[FB10] Ferrucci, David; Brown, Eric; Chu-Carroll, J. et al.: Building Watson: An Overview of the DeepQA Project. AI Magazine. 31, 3, 59–79, 2010.

[Ge83] Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive science. 7, 155–170, 1983.

[Ge97] Gentner, D., Markman, A.B.: Structure mapping in analogy and similarity. American Psychologist. 52, 45–56, 1997.

[Ge03] Gentner, D.: Why We're So Smart. Language in Mind: Advances in the Study of Language and Thought. pp. 195–235 MIT Press, 2003.

[Ho01] Hofstadter, D.R.: Analogy as the Core of Cognition. The Analogical Mind. pp. 499–538, 2001.

[Le10] Lee, J. et al.: Product EntityCube: A Recommendation and Navigation System for Product Search. 26th IEEE International Conference on Data Engineering (ICDE, Long Beach, California, USA, 2010.

[Li12] Lin, C.H. et al.: Crowdsourcing Control: Moving Beyond Multiple Choice. Human Computation Workshops at AAAI Conference on Artificial Intelligence, Ontario, Canada, 2012.

[Lo10] Lofi, C. et al.: Mobile Product Browsing Using Bayesian Retrieval. 12th IEEE Conference on Commerce and Enterprise Computing (CEC). , Shanghai, China, 2010.

[Lo12] Lofi, C. et al.: Information Extraction Meets Crowdsourcing: A Promising Couple. Datenbank-Spektrum. 12, 1, 2012.

[Se12] Selke, J. et al.: Pushing the Boundaries of Crowd-Enabled Databases with Query-Driven Schema Expansion. 38th Int. Conf. on Very Large Data Bases (VLDB). pp. 538–549 in PVLDB 5(2), Istanbul, Turkey, 2012.

[TL05] Turney, P., Littman, M.: Corpus-based learning of analogies and semantic relations. Machine Learning. 60, 251–278, 2005.

[Tu06] Turney, P.: Expressing Implicit Semantic Relations without Supervision. Int. Conf. on Computational Linguistics (COLING). , Sydney, Australia, 2006.

[Tu06-2] Turney, P.D.: Similarity of semantic relations. Computational Linguistics. 32, 3, 379–416, 2006.
[Tu08] Turney, P.D.: A uniform approach to analogies, synonyms, antonyms, and associations. Int. Conf. on Computational Linguistics (COLING). , Manchester, UK, 2008.
[Ve04] Veale, T.: Wordnet sits the sat: a knowledge-based approach to lexical analogy. Europ. Conf. on Artificial Intelligence (ECAI), Valencia, Spain, 2004.
[Wh09] Whitehill, J. et al.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, Advances in Neural Information Processing Systems. pp. 2035–2043, Vancouver, Canada, 2009.